Facultad de Ciencias

MODELIZACIÓN DE LA VEGETACIÓN DE CANTABRIA A PARTIR DE IMÁGENES DE SATÉLITE

(MODELLING OF THE VEGETATION OF CANTABRIA USING SATELLITE IMAGES)

Trabajo de Fin de Máster
para acceder al

MÁSTER EN MATEMÁTICAS Y COMPUTACIÓN

Autor: Juan Antonio Romillo Barquín

Director\es: Daniel San Martín Segura y José Manuel Gutierrez Llorente

Julio - 2015

Agradecimientos

En primer lugar deseo expresar mi agradecimiento a mis directores de este Trabajo de Fin de Máster, Daniel San Martín Segura y José Manuel Gutiérrez Llorente por su dedicación e implicación en la realización de este trabajo, por su paciencia y comprensión.

A Max Tuni, compañero en la empresa *Predictia Intelligent Data Solutions SL*, por haber estado siempre dispuesto a echar una mano, por su orientación y atención en momentos de duda.

A Juan Busquets, trabajador del Centro de Investigación y Formación Agrarias del Gobierno de Cantabria, por su colaboración en la realización de este trabajo, sobretodo en la confección de la cartografía de vegetación.

A la Consejería de Agricultura, Ganadería, Pesca y Aguas del Gobierno de Cantabria por la cesión de datos sin los cuales el correcto desarrollo de este trabajo habría sido imposible.

Resumen

Este trabajo se enmarca en un proyecto que tiene como objetivo la generación de un mapa de vegetación de Cantabria con técnicas de teledetección. Para ello se utilizará una cartografía de algunas zonas de Cantabria realizada por la Consejería de Agricultura, Ganadería, Pesca y Aguas del Gobierno de Cantabria así como de imágenes de satélite RapidEye de diferentes fechas de 2012 con una resolución 5x5 metros.

El objetivo de este trabajo es el desarrollo de una metodología de clasificación multicategoría mediante la combinación de modelos de clasificación binarios probabilísticos. Para ello, se hace uso de diferentes técnicas de aprendizaje automático como son *MaxEnt* y los árboles de decisión.

Summary

This work is part of a project that aims to generate a vegetation map of Cantabria using remote sensing techniques. For this purpose, we will use a cartography of some parts of Cantabria produced by the Consejería de Agricultura, Ganadería, Pesca y Aguas del Gobierno de Cantabria and RapidEye satellite images of different dates from 2012 with a resolution of 5x5 meters.

The aim of this work is the development of a multi-category classification methodology by combining probabilistic models of binary classification. For this purpose, we make use of different machine learning techniques such as MaxEnt and decision trees.

Palabras clave

Teledetección, aprendizaje automático, clasificación, *MaxEnt*, árboles de decisión.

Key words

Remote sensing, machine learning, classification, *MaxEnt*, decision trees.

Índice

1 Introducción

En este apartado se va a exponer en qué consiste el problema que se pretende resolver, cuáles son los objetivos que se quieren alcanzar y cuáles son las circunstancias que rodean la realización de este.

1.1 Confidencialidad

Dado el carácter confidencial de los datos utilizados en este trabajo, cualquier parte del mismo sólo podrá ser difundido previo consentimiento del autor, de la empresa *Predictia Intelligent Data Solutions SL*, de la Consejería de Agricultura, Ganadería, Pesca y Aguas del Gobierno de Cantabria y del Centro de Investigación y Formación Agrarias del Gobierno de Cantabria.

1.2 Descripción del problema

El objetivo de este trabajo consiste en desarrollar una metodología de clasificación de vegetación con técnicas de teledetección que pueda ser aplicada para generar un mapa de vegetación de Cantabria. Se pretende lograr que este mapa de vegetación sea lo más completo posible, es decir, que contenga al mayor número de clases de vegetación sin perder de vista que, al mismo tiempo, esas clases de vegetación han de estar clasificadas satisfactoriamente. De este modo, se podrá emplear para diversas aplicaciones relacionadas con la gestión y ordenación del territorio. Por ejemplo, una de estas aplicaciones, que motivó precisamente el desarrollo de este trabajo, es facilitar el cálculo del Coeficiente de Admisibilidad de Pastos (CAP) por parte de la Consejería de Agricultura, Ganadería, Pesca y Aguas del Gobierno de Cantabria para la aplicación de la Política Agrícola Común de la Unión Europea (PAC), que gestiona las subvenciones que se otorgan a la producción agrícola en la Unión Europea[1].

El trabajo ha sido realizado por el alumno en la empresa *Predictia Intelligent Data Solutions SL.*. Este trabajo parte de los resultados previos obtenidos en el trabajo de fin de Máster [1] en el que se analiza el uso de diferentes técnicas de aprendizaje automático en el campo de la teledetección para la clasificación de tipos de vegetación. En este trabajo extenderemos el trabajo de fin de Máster anteriormente mencionado profundizando en las diferentes técnicas de clasificación multicategoría a partir de clasificadores binarios.

1 http://ec.europa.eu/agriculture/index_es.htm

1.3 Teledetección

En este subapartado vamos a realizar una introducción a la teledetección: en qué consiste, cuál es su historia y cuáles sus principales aplicaciones en la actualidad.

La teledetección es la técnica de adquisición de datos de la superficie terrestre desde sensores instalados en plataformas espaciales. La interacción electromagnética entre el terreno y el sensor genera una serie de datos que son procesados posteriormente para obtener información interpretable de la Tierra [2].

La teledetección por satélite se viene practicando desde 1972. Se empezaron utilizando canales groseros y de resolución espacial moderada como el satélite Landsat MSS[2]. En la década de 1980, los satélites Landsat TM y SPOT proporcionaron una notable mejora sobre la resolución espacial y espectral.

Es en estas últimas décadas cuando se ha producido un auge en el uso de las imágenes por satélite. Se han empleado diferentes enfoques para la predicción espacial de movimientos en masas superficiales y se han aplicado técnicas de análisis multiespectrales para identificar los tipos de cobertura del suelo [3]. Destaca el desarrollo producido en la modelización de la radiación óptica y térmica, y las mejoras en algoritmos para la estimación de la temperatura y la emisividad de la superficie terrestre [4].

La teledetección proporciona innumerables ventajas. Una de ellas es que los instrumentos de teledetección permiten detectar longitudes de onda que el ojo humano no puede percibir lo cual da acceso a una información adicional que puede llegar a ser extremadamente valiosa. Otra de ellas es la capacidad de acceder a información de lugares peligrosos o inaccesibles.

Las ventajas que proporciona el uso de la teledetección hacen que los satélites de observación de la Tierra tengan una importancia creciente en áreas cada vez más diversas como puede ser observar las consecuencias del cambio climático [5], detectar incendios forestales [6] o conocer cuánta energía solar absorben las plantas mediante la fotosíntesis [7].

La teledetección a partir de imágenes de satélite también ofrece la posibilidad de describir algunos aspectos de los sistemas ecológicos. Por ejemplo, se pueden discriminar clases de vegetación, tipos de bosque o incluso individuos de distintas especies en bosques mixtos, lo que convierte a la teledetección en una importante herramienta para la gestión y conservación de la biodiversidad [8]. Varios estudios han demostrado que muchas especies vegetales tienen diferentes respuestas en el espectro electromagnético y pueden ser diferenciadas [9].

La teledetección se ha utilizado en multitud de estudios para generar cartografías de vegetación de diferentes tipos [10]. Antes de la aparición de la teledetección las cartografías de vegetación se generaban a partir de la información proporcionada por expertos y datos obtenidos con trabajo de campo. La aplicación de la teledetección a este problema supone por tanto una reducción enorme de los costes si la zona cartografiada es de grandes proporciones.

2 http://landsat.gsfc.nasa.gov/

1.4 Aprendizaje automático

En las últimas décadas la cantidad de datos que existe en el mundo ha ido creciendo desmesuradamente y, además, el crecimiento del volumen de datos se hace cada vez mayor. Estos sucesos nos han conducido a la situación en la que la nos encontramos en la actualidad, donde la cantidad de datos presente en nuestras vidas es desbordante.

A pesar de que se dispone de más datos, este hecho no necesariamente se traduce en una mejor comprensión de ellos. De hecho, a medida que la proporción de datos que tenemos a nuestro alcance aumenta, la proporción de ellos que la gente es capaz de comprender decrece. En los datos puede encontrarse información útil y valiosa (patrones de consumidores, patrones de votantes, patrones en la migración de especies animales, etc.). Sin embargo, el acceso a esa información es complicado debido al volumen de datos disponibles.

La minería de datos (en inglés *data mining*) trata de resolver problemas analizando los datos disponibles, es decir, trata de enfrentar la desbordante cantidad de datos para reducirlos y comprenderlos. El aprendizaje automático (en inglés *machine learning*) es una parte de la minería de datos que se centra en el estudio y desarrollo de técnicas automáticas que permitan extraer información a partir de datos.

El aprendizaje automático pretende en gran medida "aprender" de los datos disponibles y ser capaz de aplicar ese conocimiento "aprendido" a otros datos distintos, lo que se conoce como capacidad de generalización. Para lograr esta meta no hay un único camino sino diversos, por ello surgen distintos métodos con distintas soluciones igualmente válidas para un mismo problema.

Existen diversas causas por las que un método no es capaz de generalizar adecuadamente:

- Una de ellas es el sobreajuste a los datos, que consiste en "aprender" los datos hasta el punto de no ser capaz de aplicar el conocimiento aprendido a unos datos que no sean iguales a los datos de entrenamiento (los datos con los que se "aprende" el modelo). Un ejemplo de sobreajuste sería aprenderse de memoria los exámenes de años anteriores de una determinada asignatura. Hay varias formas de lidiar con el problema del sobreajuste, o más bien, de saber si un modelo está sobreajustado a los datos. Lo más habitual es dividir el conjunto de datos en dos subconjuntos: el de entrenamiento y el de validación. El primero se emplea para que los métodos "aprendan" los datos y el segundo para evaluar si lo aprendido por los métodos sufre de sobreajuste.

- Otra de las causas es el no tener una muestra representativa en los datos de la realidad que queremos "aprender". Un ejemplo, aplicable al problema que queremos resolver en este trabajo, sería si los modelos "aprendiesen" los datos de la ciudad de Santander y después se pretendiese aplicar el conocimiento aprendido por los modelos a toda Cantabria. Como la vegetación de Cantabria no está bien representada por la vegetación de Santander, la generalización sería completamente insatisfactoria.

El aprendizaje automático se emplea para resolver muy diferentes problemas como la clasificación, la detección de patrones o la predicción de series temporales. En lo referente a la clasificación, puede diferenciarse la clasificación supervisada y la no supervisada. En la primera de

ellas, además de disponer de los datos también se dispone de la clase o etiqueta de cada una de las muestras mientras que en el segundo caso no. Por ejemplo, si queremos clasificar el género de una determinada especie animal en función de la altura y el peso, y tenemos a nuestra disposición el género de cada muestra se trataría de clasificación supervisada. En cambio, si no disponemos del género de las muestras y, simplemente, queremos realizar varios grupos se trataría de aprendizaje no supervisado. Se podría dar el caso de que al aplicar técnicas de ambos tipos de aprendizaje automático los resultados fuesen similares si el género influye sobre la altura y el peso de la especie animal que se está analizando.

Como el objetivo principal de la clasificación supervisada es lograr acertar la etiqueta de los datos, existen diferentes tipos de clasificación supervisada en función del número de etiquetas de los datos: binaria (2 etiquetas) y multicategoría (más de 2). Además de la división anterior de la clasificación supervisada, también podemos encontrar cada una de estas divisiones a su vez divididas en clasificación probabilística o determinista. En la primera de ellas cada muestra obtiene una probabilidad de pertenencia a las distintas clases y en la segunda, se obtiene la clase en la que será etiquetada. En un problema de clasificación supervisada binario, es muy fácil pasar las probabilidades a etiquetas (sin más que establecer un umbral que delimite ambas clases). En cambio, para el caso de la clasificación multicategoría existen numerosas aproximaciones para obtener las etiquetas de las muestras.

Si se desea ampliar conocimientos en materia de aprendizaje automático, se recomienda los siguientes libros: [11], [12] y [13].

2 Datos utilizados

En este subapartado vamos a describir cuáles han sido los datos empleados en la realización de este trabajo. Más concretamente, se describirán tanto las imágenes de satélite empleadas como la cartografía de vegetación que se ha utilizado.

El objetivo de este trabajo es construir modelos que nos permitan clasificar las clases de vegetación a partir de imágenes de satélite. Para poder entrenar estos modelos necesitaremos unas imágenes que nos ofrezcan la suficiente información como para diferenciar las distintas clases de vegetación. Por otro lado, necesitaremos unos datos de salida que, necesariamente, serán las clases de vegetación correspondientes a esas zonas del espacio dadas por las imágenes de satélite. De este modo podremos generar unos modelos que "aprendan" los datos, incluyendo la clase de vegetació. Por tanto, se tratará de un problema de aprendizaje supervisado.

2.1 Imágenes de satélite

Las imágenes de satélite empleadas son de la compañía alemana *RapidEye AG*[3], que se dedica a la gestión de toma de decisiones en base a sus propias imágenes satelitales. Más concretamente, se han utilizado imágenes de satélite del año 2012 de resolución espacial de 5x5 metros. Estas imágenes fueron adquiridas y cedidas para la realización de este trabajo por la Consejería de Agricultura, Ganadería, Pesca y Aguas del Gobierno de Cantabria. En cada zona de la que se dispone de imágenes de satélite, tenemos respuesta espectral de 5 bandas diferentes. Las respuestas espectrales de las bandas vienen enumeradas en la Tabla 2.1.

Banda	Descripción	Longitud de onda	Resolución
B1	Azul (espectro visible)	440-510 nm	5x5 m
B2	Verde (espectro visible)	520-590 nm	5x5 m
B3	Rojo (espectro visible)	630-685 nm	5x5 m
B4	*Red Edge*	690-730 nm	5x5 m
B5	Infrarojo cercano	760-850 nm	5x5 m

Tabla 2.1 En esta tabla se muestra la descripción de la bandas de satélite RapidEye.

La disponibilidad de estas imágenes depende de varios factores entre los que destaca la meteorología. Por ejemplo, han de ser captadas en un momento en el que no haya nubosidad, lo que implica que la fecha en la que son tomadas no tiene por qué ser la ideal para la consecución de nuestro objetivo, que es el de ser capaces de diferenciar entre clases de vegetación. Lo idóneo sería disponer de imágenes de satélite que cubriesen toda Cantabria en las mismas fechas. Sin embargo,

3 www.blackbridge.com/rapideye/

esto no es posible pues sería indispensable la ausencia de nubes en toda la región en el momento de la toma de las imágenes satelitales y, además, el tamaño de tesela del satélite *RapidEye* no permite cubrir toda la superficie de Cantabria de forma simultánea. Dado que estamos restringidos a los datos de los que disponemos, la solución planteada ha sido la de dividir Cantabria en varias zonas. Por tanto, se dispone de imágenes de satélite de distintas fechas en cada una de las zonas en las que se ha divido Cantabria.

Otro problema que surge es que en lugares donde la pendiente es muy elevada, las sombras hacen que la respuesta en frecuencia de las bandas se vea afectada, distorsionando la realidad. Para reducir este efecto se ha aplicado una corrección topográfica utilizando la técnica de Minnaert [14].

Además, siempre que ha sido posible se han utilizado dos imágenes de cada banda en fechas diferentes en cada zona. La razón se debe a que se quiere capturar diferentes huellas espectrales de la vegetación a lo largo de ciclo fenológico, el cual evoluciona a lo largo del año. La aparición y maduración de frutos, la caída de las hojas, la floración, etc. influyen enormemente en la respuesta espectral obtenida Se ha de considerar también que la fenología de distintas variedades de plantas cambian en períodos de tiempo diferentes. Por todo lo expuesto, el hecho de seleccionar imágenes de satélite en distintas épocas del año nos permite tener a nuestra disposición una valiosa información que ayudará a discriminar las clases de vegetación.

A continuación se va a mostrar cómo se ha divido Cantabria y en la Tabla 2.2 cuáles son las fechas de las imágenes satelitales seleccionadas en cada una de las zonas.

En las ilustraciones que van desde la Ilustración 2.2 hasta la Ilustración 2.9 se puede observar qué regiones corresponden cada una de las 8 zonas en las que se dividió Cantabria. En las ilustraciones se muestra el mapa de vegetación que se generó en cada una de las zonas, además de algunos puntos fuera de las correspondientes zonas que representan la cartografía de vegetación de la que hablaremos un poco más adelante. Los distintos colores de las ilustraciones representan las clases de vegetación predichas, la leyenda de las clases de vegetación asociadas a esos colores se puede ver en la Ilustración 2.1.

Zona	Número de imágenes por banda	Fechas
A	2	24/09/2012 y 23/07/2012
B	2	23/07/2012 y 23/05/2012
C	1	23/05/12
D	1	01/10/12
E	2	20/09/2012 y 16/08/2012
F	2	01/10/2012 y 20/09/2012
G	2	01/10/2012 y 18/08/2012
H	2	01/10/2012 y 18/08/2012

Tabla 2.2 En esta tabla se muestra la descripción de la bandas de satélite *RapidEye*.

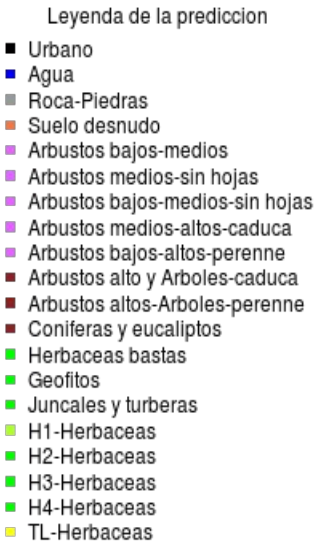

Ilustración 2.1 En esta ilustración se muestra la leyenda que se ha empleado en las ilustraciones que van desde la Ilustración 2.2 hasta la Ilustración 2.9.

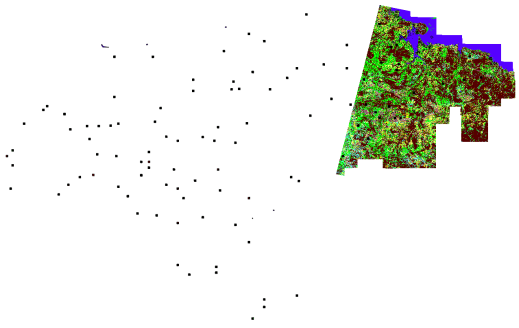

Ilustración 2.2 En esta ilustración se muestra la zona A.

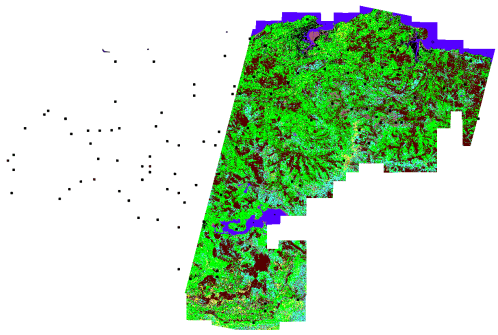

Ilustración 2.3 En esta ilustración se muestra la zona B.

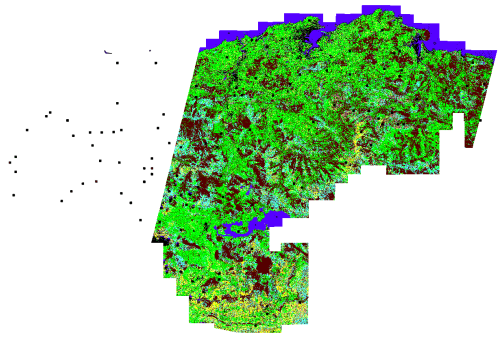

Ilustración 2.4 En esta ilustración se muestra la zona C.

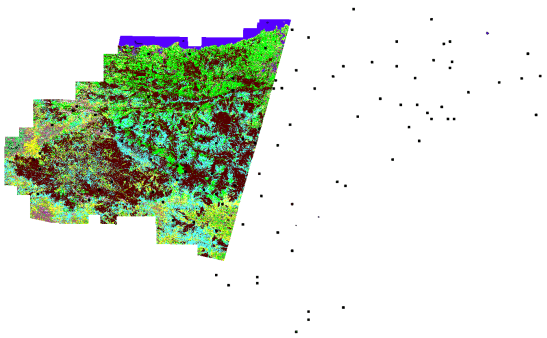

Ilustración 2.5 En esta ilustración se muestra la zona D.

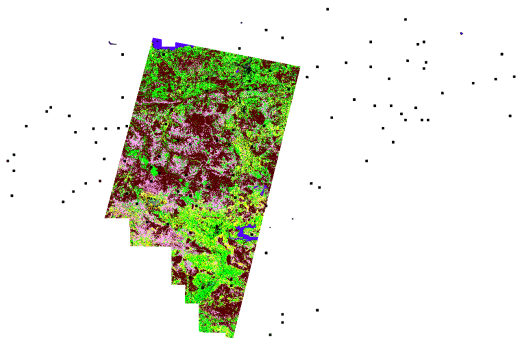

Ilustración 2.6 En esta ilustración se muestra la zona E.

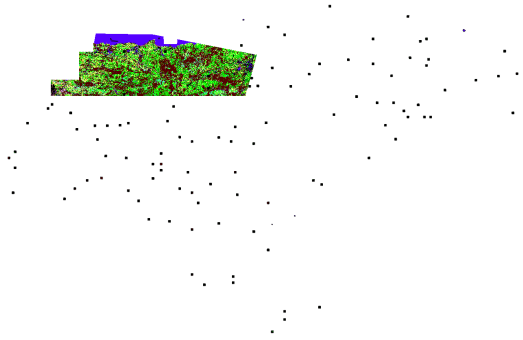

Ilustración 2.7 En esta ilustración se muestra la zona F.

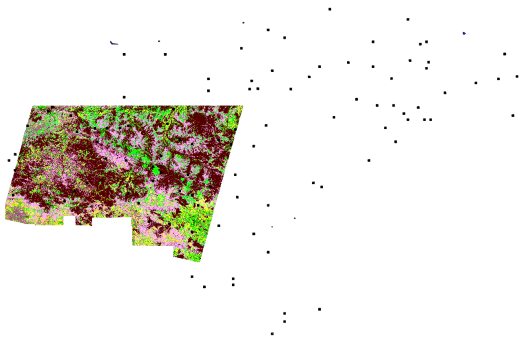

Ilustración 2.8 En esta ilustración se muestra la zona G.

Ilustración 2.9 En esta ilustración se muestra la zona H.

Además de las 5 bandas antes mencionadas, se generó el Índice de vegetación de diferencia normalizada (*Normalized Difference Vegetation Index* [15]), también conocido como *NDVI* por sus siglas en inglés. Este índice es usado como estimador de la cantidad, calidad y desarrollo de la vegetación en base a la intensidad de la radiación de ciertas bandas del espectro electromagnético que la vegetación emite o refleja.

El *NDVI* se define como $\frac{(IR\ cercano - Rojo)}{(IR\ cercano + Rojo)}$. Es decir, que hemos utilizado dos de las bandas que teníamos disponibles para generarlo.

El *NDVI* puede tomar valores entre -1 y 1. Es uno de los índices de vegetación más usados, esto se debe a que permite diferenciar con mucha facilidad y con bastante exactitud si en un determinado lugar se da o no la presencia de vegetación, cómo es la calidad de esta vegetación y cuánta hay. En aquellos puntos en los que se obtiene un valor por debajo de 0.1 suele haber presentes gran cantidad de rocas, tierra, arena o nieve. En cambio, en aquellos en los que toman valores entre 0.2 y 0.5, suele haber arbustos, plantas bajas o vegetación dispersa. Por último, los puntos en los que la banda del *NDVI* toma valores entre 0.6 y 0.9 se suelen corresponder con vegetación muy densa o con vegetación en el pico de su desarrollo.

En la Ilustración 2.10 se muestra el *NDVI* de la imagen obtenida el 23/05/2012. Como puede apreciarse tanto el embalse del Ebro como el mar aparecen con un *NDVI* muy bajo, contrariamente a otros lugares con abundante vegetación. Recordemos que si el *NDVI* toma valores altos es un indicador de la presencia de vegetación.

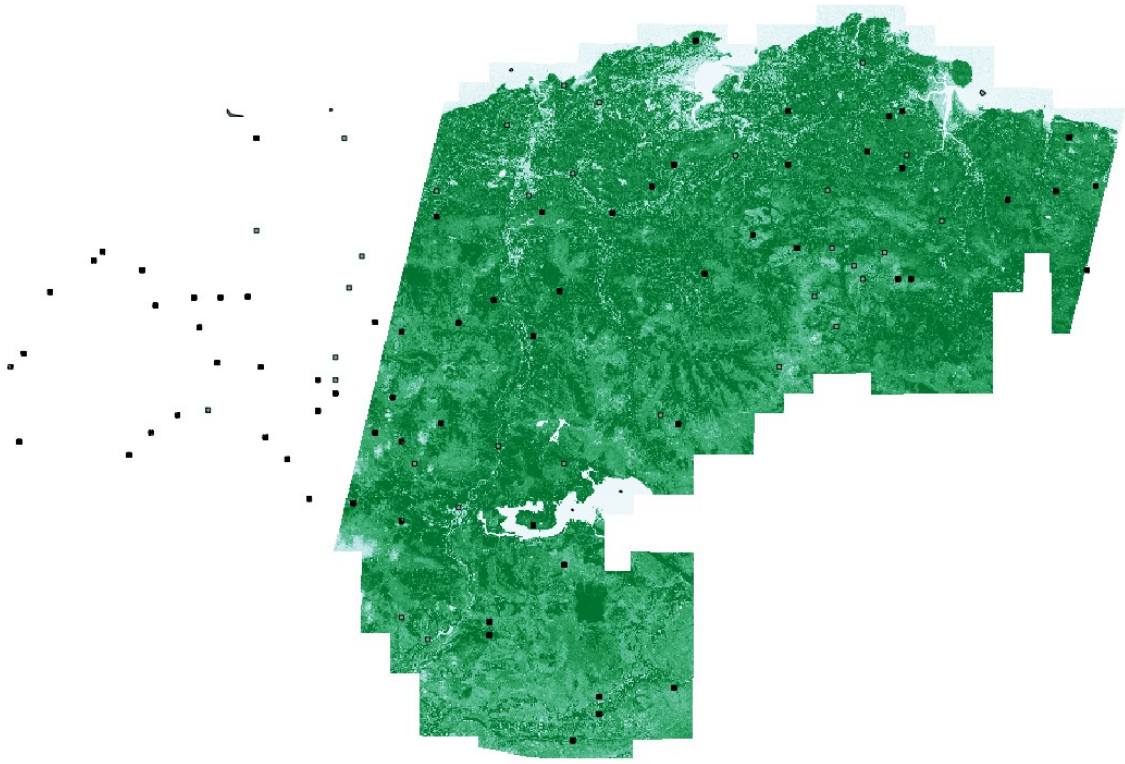

Ilustración 2.10 En esta ilustración se muestra el *NDVI* en una escala que va del blanco (*NDVI* bajo) al verde (*NDVI* alto).

Para terminar con este subapartado se van a mostrar unos diagramas de cajas que nos permitan apreciar cómo quedan representadas cada una de las clases de vegetación con las que estamos trabajando (que serán presentadas en el próximo subapartado Cartografía de vegetación) en cada una de las bandas del satélite disponibles. De este modo, podremos observar cuáles de estas clases son similares y cuáles son separables a simple vista. Sólo se van a mostrar los diagramas de cajas asociados a la zona *B*, recordemos que en esta zona se dispone de bandas en dos meses distintos del año 2012, una de ellas del mes de mayo y la otra del mes de julio.

La leyenda de las clases de vegetación mostradas en todas las ilustraciones desde la Ilustración 2.11 a la Ilustración 2.16 es la siguiente:

- 1: Urbano
- 2: Agua
- 3: Roca-Piedras
- 4: Suelo desnudo
- 5: Arbustos bajos-medios
- 6: Arbustos medios-sin hojas
- 7: Arbustos bajos-medios-sin hojas
- 8: Arbustos medios-altos-caduca
- 9: Arbustos bajos-altos-perenne
- 10: Arbustos alto y Arboles-caduca
- 11: Arbustos altos-Arboles-perenne
- 12: Coníferas y eucaliptos
- 13: Herbáceas bastas
- 14: Geófitos
- 15: Juncales y turberas
- 16: TL-Herbáceas
- 17: H1-Herbáceas
- 18: H2-Herbáceas
- 19: H3-Herbáceas
- 20: H4-Herbáceas.

Como se puede observar en las ilustraciones (desde la Ilustración 2.11 a la Ilustración 2.16), existen clases de vegetación que son realmente similares y que, por lo tanto, su diferenciación parece ser una tarea complicada como ocurre con la clase Arbustos bajos-medios y la clase Arbustos bajos-medios-sin hojas, que corresponden a los números 5 y 7 de la leyenda.

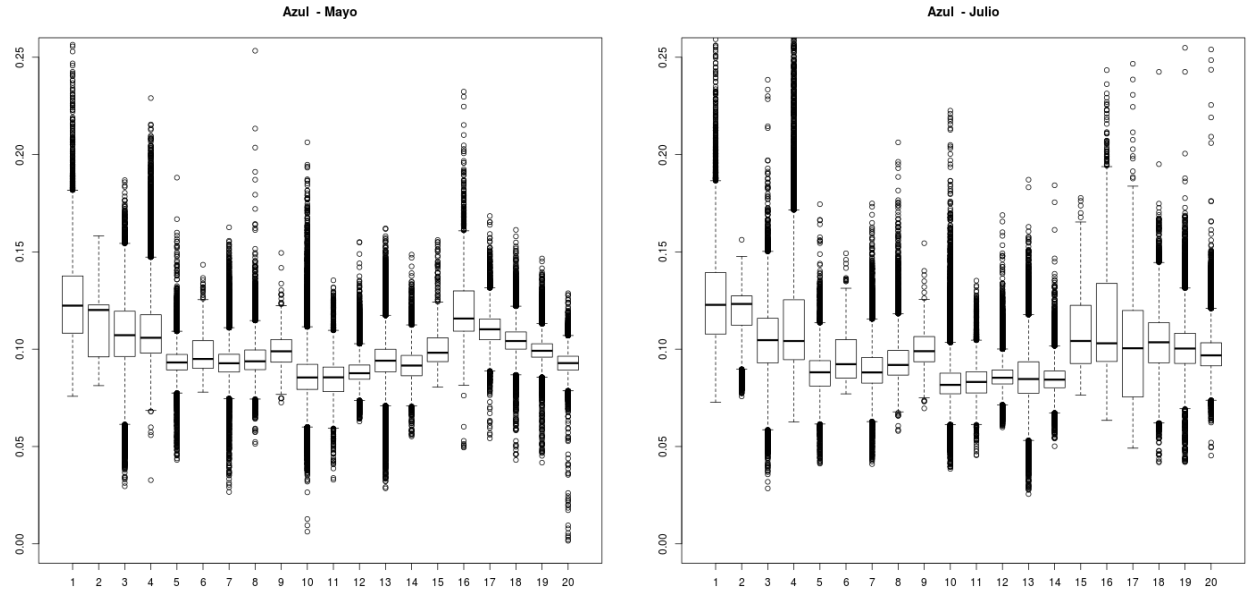

Ilustración 2.11 En esta ilustración se muestran dos diagramas de cajas. En el eje Y, el valor de la banda del azul y en el eje X, las clases de vegetación. A la izquierda, los diagramas de cajas con las bandas de mayo y, a la derecha, los diagramas de cajas con las bandas de julio.

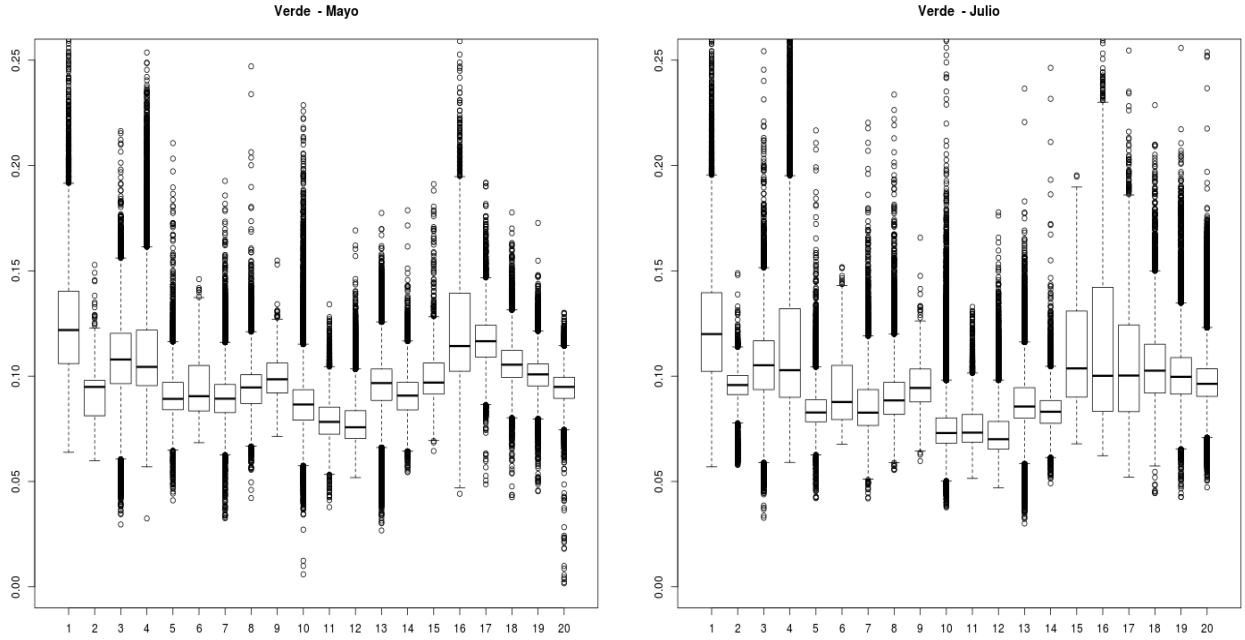

Ilustración 2.12 En esta ilustración se muestran dos diagramas de cajas. En el eje Y, el valor de la banda del verde y en el eje X, las clases de vegetación. A la izquierda, los diagramas de cajas con las bandas de mayo y, a la derecha, los diagramas de cajas con las bandas de julio.

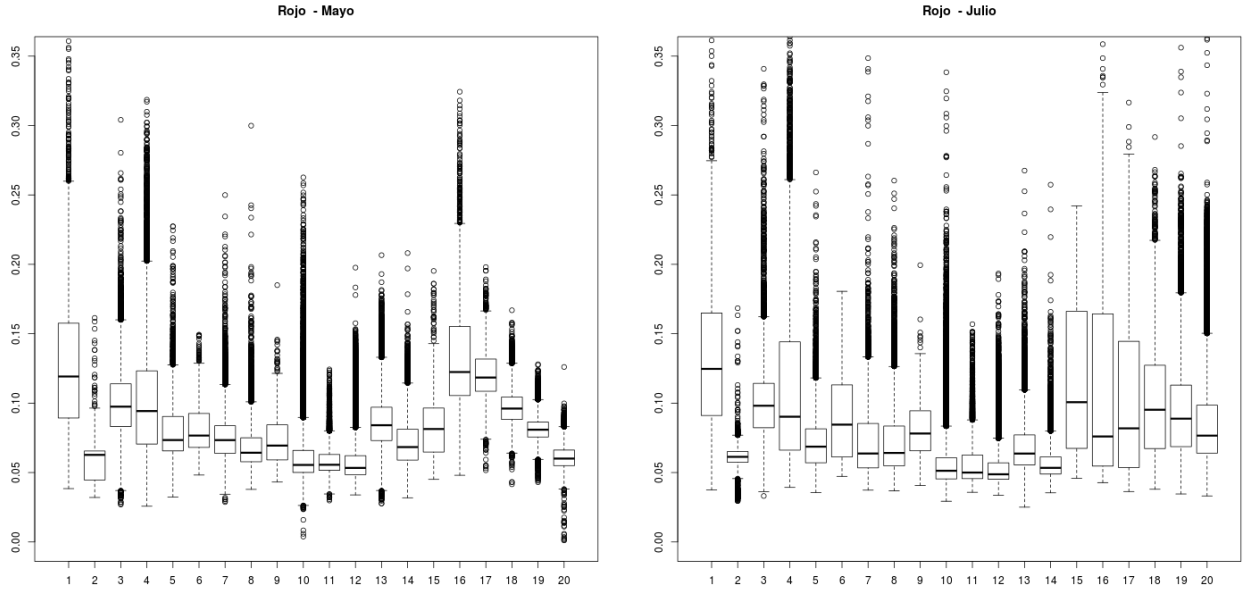

Ilustración 2.13 En esta ilustración se muestran dos diagramas de cajas. En el eje Y, el valor de la banda del rojo y en el eje X, las clases de vegetación. A la izquierda, los diagramas de cajas con las bandas de mayo y, a la derecha, los diagramas de cajas con las bandas de julio.

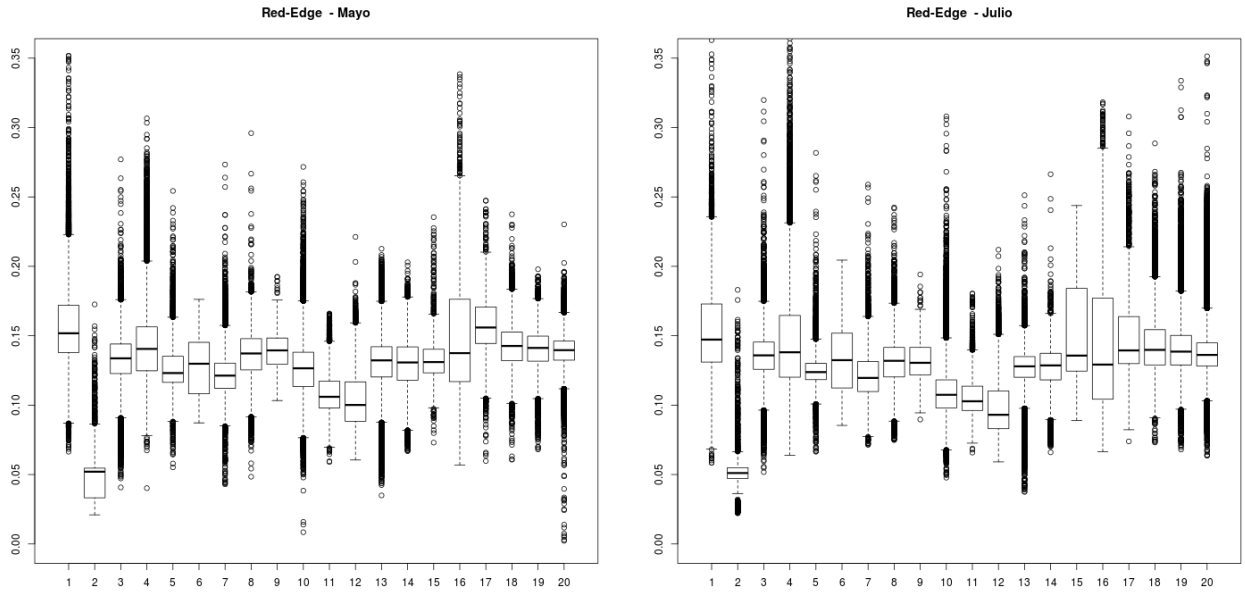

Ilustración 2.14 En esta ilustración se muestran dos diagramas de cajas. En el eje Y, el valor de la banda del *Red Edge* y en el eje X, las clases de vegetación. A la izquierda, los diagramas de cajas con las bandas de mayo y, a la derecha, los diagramas de cajas con las bandas de julio.

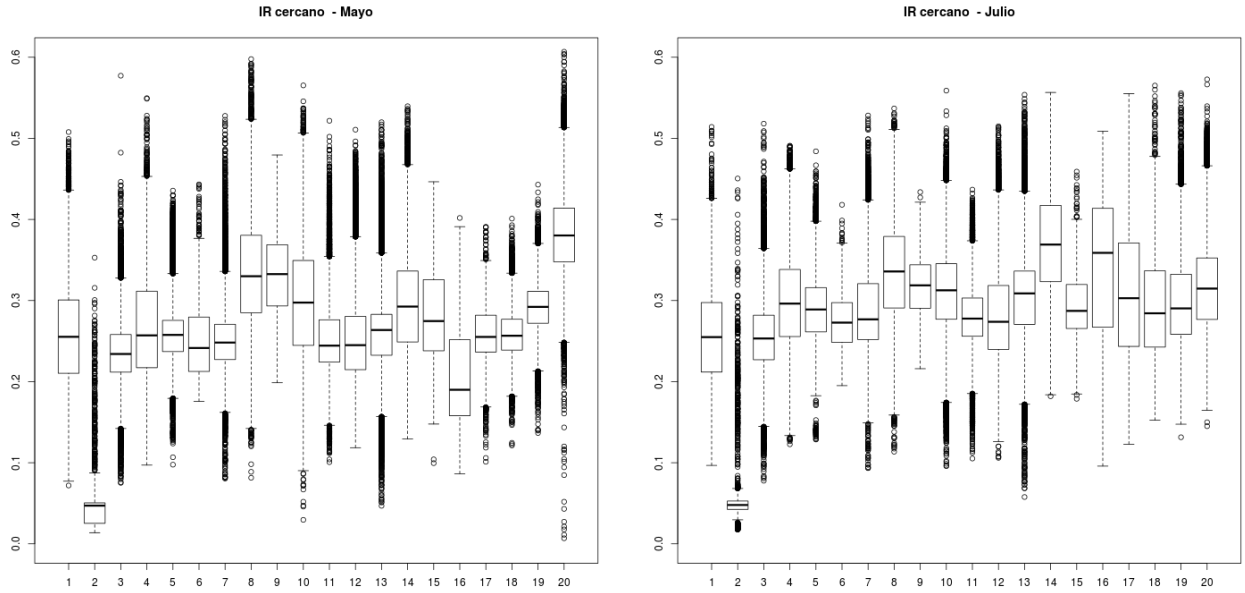

Ilustración 2.15 En esta ilustración se muestran dos diagramas de cajas. En el eje Y, el valor de la banda del *IR* cercano y en el eje X, las clases de vegetación. A la izquierda, los diagramas de cajas con las bandas de mayo y, a la derecha, los diagramas de cajas con las bandas de julio.

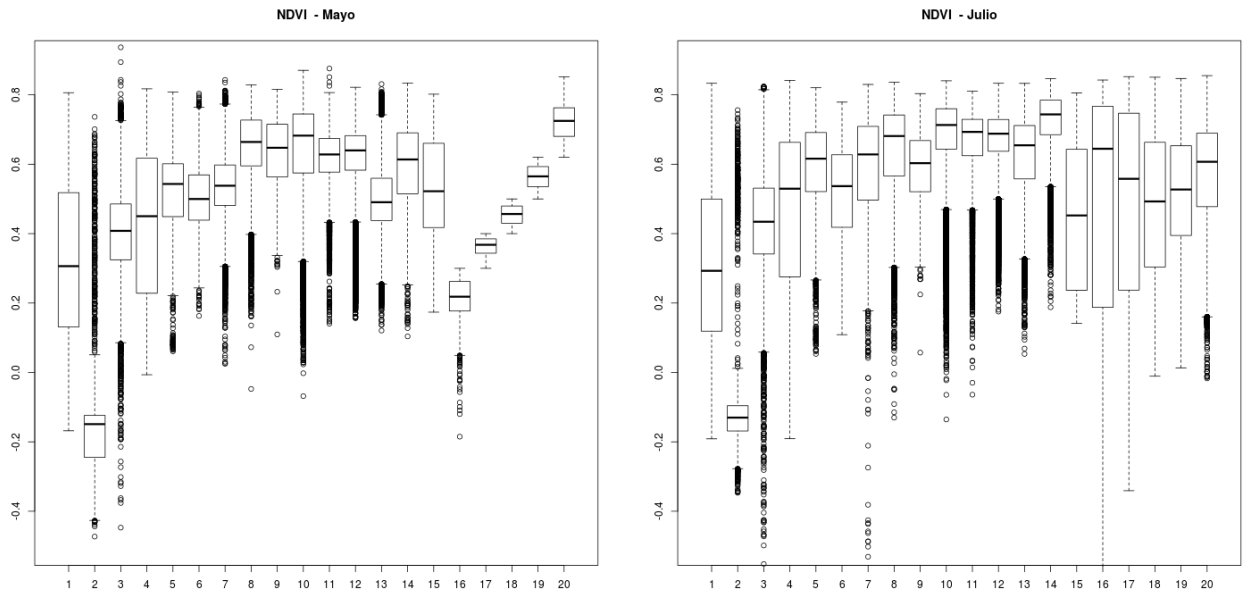

Ilustración 2.16 En esta ilustración se muestran dos diagramas de cajas. En el eje Y, el valor de la banda del *NDVI* y en el eje X, las clases de vegetación. A la izquierda, los diagramas de cajas con las bandas de mayo y, a la derecha, los diagramas de cajas con las bandas de julio.

2.2 Cartografía de vegetación

Además de las imágenes de satélite, también utilizaremos una cartografía de vegetación realizada por el CIFA en el año 2013. Esta cartografía de vegetación ha sido realizada de forma que tenga la misma resolución espacial que las imágenes de satélite que disponemos. De este modo, para cada recinto o píxel de 5x5 metros que esté cartografiado dispondremos de los valores de las bandas, 6 ó 12 valores dependiendo de en que zona nos hallemos (con una única imagen de satélite o dos), y de la clase de vegetación predominante en ese píxel. En cambio, en aquellos recintos no cartografiados tan sólo dispondremos de los valores de las bandas del satélite.

La metodología seguida por el CIFA para generar esta cartografía se describe a continuación. Primero se realizó una estratificación del territorio por variables como el suelo, clima, altitud, etc. Posteriormente en cada una de los estratos en los que se había dividido el territorio los lugares seleccionados para ser cartografiados fueron elegidos aleatoriamente del siguiente modo: primero se seleccionaba un punto de la rejilla de resolución 5x5 metros de Cantabria, posteriormente se designaba tal punto como vértice superior izquierdo de un cuadrado con área previamente fijada, de este modo tendríamos seleccionado el primer cuadrado, se repitió este proceso asegurándose de que ningún cuadrado nuevo solapara con alguno de los antiguos hasta disponer de 105 cuadrados. Después de haber seleccionado los cuadrados, se realizó el trabajo de campo correspondiente para asignar a cada píxel de 5x5 metros de cada uno de los 105 cuadrados la clase de vegetación predominante y la segunda con mayor presencia. En el caso de que tan sólo se diese la presencia de una única clase en algún píxel, ésta aparecería en ambas: en la clase primaria y en la secundaria.

En un principio, la cartografía de vegetación de la que se disponía contaba con alrededor de 50 clases de vegetación. Sin embargo, hubo que agruparlas debido a la escasez de muestras de

algunas de ellas o la extrema dificultad de diferenciar algunas de estas clases iniciales entre ellas. En la Tabla 2.3 se describirán cuáles eran las clases de vegetación con las que contaba la cartografía de vegetación empleada para la realización de este trabajo. Vamos a mostrar una imagen (Ilustración 2.17) en la que podremos observar dónde se encuentran los cuadrados seleccionados para la cartografía de vegetación.

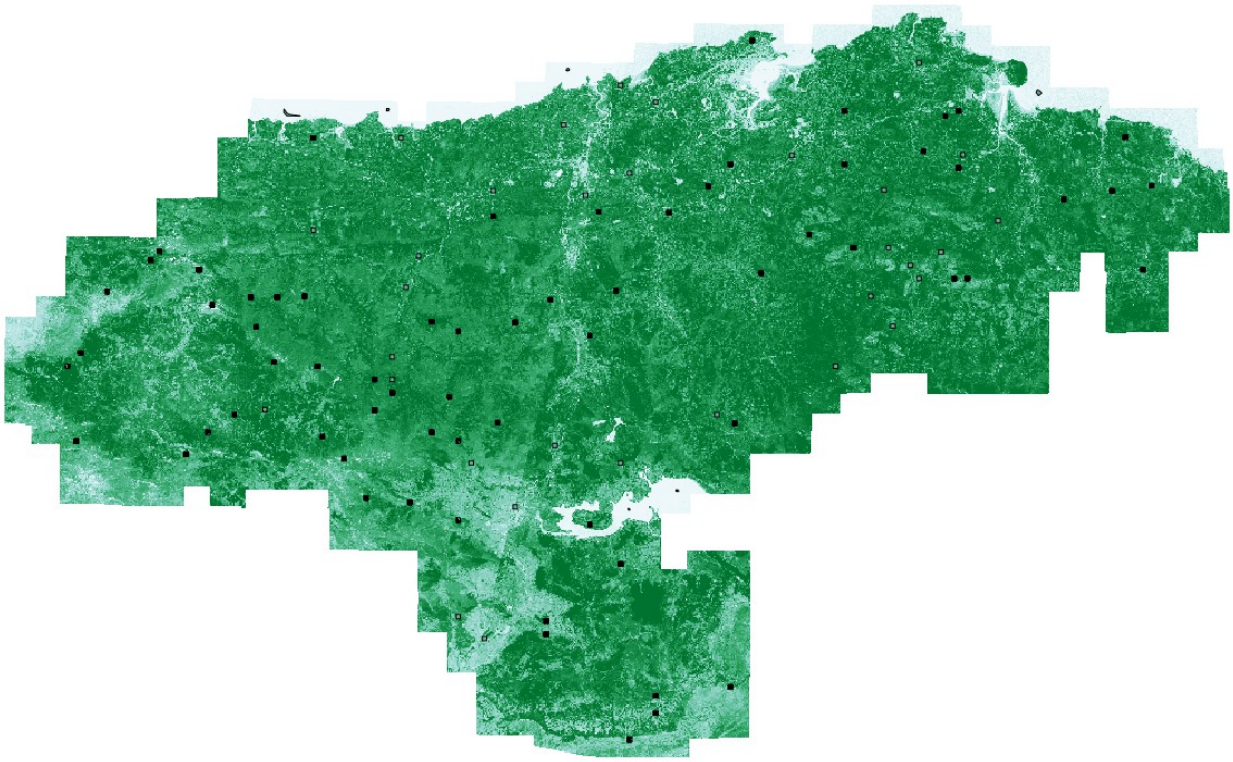

Ilustración 2.17 En esta ilustración se muestra el *NDVI* en toda Cantabria en una escala que va del blanco (*NDVI* bajo) al verde (*NDVI* alto). Además, se pueden apreciar los cuadrados que se han seleccionado para la cartografía de vegetación (en negro).

Nombre de la clase de vegetación	Descripción	Número de puntos en la zona B
Urbano	Urbano.	14881
Agua	Vegetación nula o rala. Acuática.	12765
Roca-Piedras	Vegetación nula o rala. Roca-Piedras.	17728
Suelo desnudo	Vegetación nula o rala. Suelo desnudo.	10802
Arbustos bajos-medios	Vegetación leñosa. Arbustos bajos-medios. Ericáceas y similares.	15445
Arbustos medios-sin hojas	Vegetación leñosa. Arbustos medios (60-200cm). Sin hojas aparentes. Escobas.	656
Arbustos bajos-medios-sin hojas	Vegetación leñosa. Arbustos bajos-medios (30-200cm). Sin hojas aparentes. Leguminosas espinosos.	36628
Arbustos medios-altos-caduca	Vegetación leñosa. Arbustos medios-altos (60-500cm). Hoja caduca. Predominantemente Rosáceas.	10860
Arbustos bajos-altos-perenne	Vegetación leñosa. Arbustos bajos-altos (30cm-5m). Hoja perenne. Marismas.	495
Arbustos alto y Arboles-caduca	Vegetación leñosa. Arbustos altos y árboles. Hoja caduca.	102364
Arbustos altos-Arboles-perenne	Vegetación leñosa. Epífitos-Arbustos altos-Árboles (2-40m). Hoja perenne.	14900
Coníferas y eucaliptos	Vegetación leñosa. Coníferas y eucaliptos.	44356
Herbáceas	Herbáceas. Buen valor forrajero.	168840
Herbáceas bastas	Herbáceas bastas.	30575
Geófitos	Herbáceas. Geófitos.	19425
Juncales y turberas	Herbáceas. Juncales y turberas.	1691

Tabla 2.3 En esta tabla se muestra la descripción de las clases de vegetación presentes en la cartografía de vegetación utilizada y el número de puntos de cada clase de vegetación en algunas de las zonas en las que se dividió Cantabria.

Después de estudiar detenidamente la cartografía de vegetación, se observó la escasez de muestras de la clase agua. Por lo tanto, se generó manualmente una nueva cartografía de vegetación en la que se añadieron representantes de esta clase a la cartografía de vegetación anterior. Este paso se realizó seleccionando algunos polígonos en el Mar Cantábrico y otros en el embalse del Ebro.

3 Metodología

En este apartado se va a exponer en qué consisten las técnicas que se han empleado para resolver este problema, además de las herramientas de las que nos hemos servido para poder evaluar cómo de satisfactorios son los resultados obtenidos.

Recordemos que lo que se pretende conseguir es crear una metodología que permita determinar, a partir de los valores que tomen las bandas del espectro electromagnético en un determinado píxel, la clase de vegetación mayoritaria o primaria presente en ese píxel.

Como se avanzaba en el capítulo de Aprendizaje automático para resolver este problema existen distintas aproximaciones a la clasificación supervisada multicategoría. Por ejemplo, se podría hacer uso de alguna técnica de clasificación que devolviese como valor de salida un número entero entre 1 y el número de clases de vegetación disponibles en la cartografía para cada uno de los píxeles. Sin embargo, se comprobó experimentalmente que empleando este tipo de técnicas los resultados obtenidos no eran los deseados. Por lo que se siguió la aproximación más habitual, que consiste en dividir el problema multicategoría en varios problemas más sencillos que puedan ser resueltos con modelos binarios más específicos.

Para abordar un problema multicategoría con modelos binarios pueden considerarse, entre otras, las siguientes aproximaciones para entrenar los modelos binatios:

- *One Vs One.* Consiste en entrenar $N(N-1)/2$ modelos, donde N es el número de clases, del siguiente modo: se seleccionan los datos asociados a dos clases distintas y se entrena un modelo que permita diferenciar estas dos clases. Este proceso se sigue hasta que se entrenan los $N(N-1)/2$ modelos asociados a todas las posibles parejas de clases distintas.
- *One Vs All.* Consiste en entrenar un modelo para cada clase del siguiente modo: a los datos correspondientes a la clase seleccionada se le asigna la etiqueta 1 y al resto de los datos, asociados al resto de clases, se les asigna la etiqueta 0. De este modo, se dispone de N modelos capaces de distinguir entre su clase asociada y todas las demás.
- *Error Correcting Output Coding (ECOC)*: Consiste en separar el espacio de las clases en dos subconjuntos, uno de ellos con la etiqueta 1 y otro con la etiqueta 0, y entrenar los modelos con esos nuevos datos. Sin embargo, entrenar los modelos para todos los subconjuntos posibles (exponencial en el número de clases) puede ser muy costoso, por lo que en problemas que cuenten con un gran tamaño de datos se suele seleccionar una parte de esos subconjuntos. Es claro que esta metodología incluye a las otras dos anteriores.

Independientemente de por cuál de las metodologías de entrenamiento se opte, se ha de juntar de algún modo toda la información suministrada por los modelos binarios para poder discriminar cuál de todas las clases de vegetación es la predominante en cada píxel. Además, si se quisiese juntar la información proveniente de más de un clasificador binario, el problema se complicaría. A tal fin, se emplearon técnicas de *Majority Voting*, *Weighted Majority Voting, Behavior Knowledge Space Method*, *Dempster-Shafer Combination* [16], *The Product Rule* [17], etc.

Sin embargo, con ninguna de estas técnicas se obtuvo unos resultados que justificasen el

entrenar otro clasificador binario aparte de *MaxEnt*, que era el clasificador binario que mejores resultados obtenía como ya se había mostrado en [1].

Por último, cabe mencionar que se optó por la metodología de *One Vs All* debido a que era computacionalmente más eficiente que la metodología de *One Vs One* y *ECOC*, además, los resultados obtenidos por la estrategia *One Vs All* eran superiores a los obtenidos por la estrategia de *One Vs One*. Además, también se comprobó que los clasificadores binarios de tipo probabilístico tenían rangos de probabilidades muy dispares, por lo que no se podían comparar directamente. Por ello, se hizo uso los árboles de decisión para juntar la información suministrada por los *N* modelos de *MaxEnt* debido a los buenos resultados obtenidos y a su bajo coste computacional.

3.1 Técnicas empleadas

En este subapartado se van a describir las técnicas empleadas en la elaboración del mapa de vegetación de Cantabria. Como comentábamos en la introducción del capítulo de Metodología se han empleado *MaxEnt* y árboles de decisión.

3.1.1 MaxEnt

MaxEnt [18] es un clasificador binario probabilista que surgió en 2004 como un modelo de distribución de especies (en inglés, *Species Distribution Model, SDMs*). Estos modelos estiman la relación existente entre los lugares donde se han encontrado las especies (animales , vegetales, etc.) y las características de esos lugares. Una de las peculiaridades de estos métodos son el tipo de datos que emplean: datos de sólo presencia (en inglés, *presence-only data*), que consisten en datos donde se conocen los lugares donde se han observado las especies y en el resto se asume que se desconoce la especie presente, por lo general no hace falta asumirlo sino que los datos tienen esa estructura. A parte de disponer de la presencia de las especies, también se dispone de una serie de variables ambientales que caracterizan esos lugares donde se han presenciado las especies como pueden ser la temperatura media, la cantidad media de lluvia, la media de horas de sol diarias, la altitud del terreno, la respuesta en frecuencia (como en este problema), etc. El objetivo que se marcan los *SDMs* es el de predecir cuáles son los lugares, en base a las características donde se han observado las especies, donde es más probable que las especies se hallen.

Como comentábamos en el párrafo anterior, *MaxEnt* toma como entrada tanto los datos de sólo presencia como las variables que caracterizan esos lugares que suelen ser celdas a una determinada resolución. Además, *MaxEnt* selecciona de todos los datos, incluyendo de los datos de sólo presencia los datos donde se ha presenciado las especies, una muestra que toma como datos de fondo (en inglés, *background data*). Estos datos son empleados para contrastar los lugares donde se ha presenciado la especie frente a los datos de fondo, donde se asume el desconocimiento de la presencia de la especie.

Cuando el tamaño de la población total de las especies es conocido, *MaxEnt* es capaz de predecir la proporción de ocurrencia de las especies en las celdas [19]. Sin embargo, el tamaño de la

población total de las especies suele ser desconocido, por lo que la salida que proporciona *MaxEnt* se conoce como proporción relativa de ocurrencia (en inglés, *relative occurrence rate, ROR*) [19]. El *ROR* mide la probabilidad relativa de que una celda esté contenida en el conjunto de muestras de datos de sólo presencia.

Al utilizar *MaxEnt*, como con otros *SDMs* y otras técnicas que emplean datos de sólo presencia, se ha de tener muy en cuenta cómo se ha realizado el muestreo de los datos. Generalmente, se tiene el siguiente problema: o, bien, se asume que las especies se han muestreado aleatoriamente (esto suele ser falso por regla general porque implicaría el conocimiento de dónde se hallan todos los individuos de la especie) o, bien, se asume que las celdas han sido muestreadas aleatoriamente. La primera es una hipótesis conduce a que la salida dada por *MaxEnt,* el *ROR,* coincida con la probabilidad de que se dé la presencia de las especies. Por otro lado, la segunda hipótesis conduce a que la salida dada por *MaxEnt* es un índice que mide cómo de adecuado son las celdas para que se dé la presencia de las especies, es decir, no se puede emplear como probabilidad de presencia [20].Sin embargo, si tan sólo se pretende emplear *MaxEnt* como una técnica que permita valorar cómo de adecuadas son las distintas celdas para que se dé la presencia de las especies, entonces no es necesario asumir las hipótesis antes descritas sobre la forma de muestrear [21].

MaxEnt predice el *ROR* como una función de las variables ambientales que caracterizan las celdas: $P(z(x_i))=\mathrm{e}^{\lambda z(x_i)}/\sum_i \mathrm{e}^{\lambda z(x_i)}$, donde z es el vector que incluye todas las transformaciones de las variables ambientales que emplea *MaxEnt* que caracterizan la celda x_i *y* λ es un vector de coeficientes de regresión de los cuales hablaremos un poco más adelante. Si en la fórmula anterior se desarrolla $\lambda\, z(x_i)$ queda $\lambda_1\, z_1(x_i) + \lambda_2\, z_2(x_i) + \ldots + \lambda_J\, z_J(x_i)$, donde J es el número de variables que emplea *MaxEnt*. El término del denominador es un término de normalización que consigue que la suma en todas las celdas sea 1.

MaxEnt puede generar modelos sumamente complejos porque toma las variables ambientales que caracterizan las celdas y las transforma (linealmente, cuadráticamente, realizando el producto de varias de ellas, categorizándolas, etc.). Después de realizar estas transformaciones de las variables iniciales y seleccionar aquellas que permiten distinguir entre las celdas donde se ha observado la especie y las celdas tomadas como datos de fondo, reescala las nuevas variables al intervalo [0,1] de modo que todas ellas tengan el mismo peso y, posteriormente, ajusta el vector de coeficientes de regresión λ que otorga el peso adecuado a cada una de las nuevas variables.

Para seleccionar las transformaciones que *MaxEnt* realiza, se maximiza una función de ganancia en la que se favorece los modelos que mejor diferencian las celdas donde se encuentran los individuos de la especie frente a los datos de fondo y que, además, tengan una baja complejidad de transformaciones. De este modo, el modelo que se obtiene es capaz de discriminar sin sobreajustarse a los datos. La función de ganancia que emplea *MaxEnt* es la siguiente: $\frac{1}{M}\sum_{i=1}^{M}\lambda\, z(x_i) - \log\sum_{i=1}^{N} Q(x_i)\mathrm{e}^{\lambda z(x_i)} - \sum_{j=1}^{J}|\lambda_j|\beta\sqrt{s^2(z_j)/M}$, donde M es el número de datos de sólo presencia donde se ha presenciado la especie, N es el número de datos de fondo, β es una constante que regula la complejidad de las transformaciones, $s^2(z_j)$ es la varianza de la transformación j en los datos de sólo presencia y la $Q(x_i)$ es la distribución a priori de los datos de fondo que toma el valor 1 en las celdas que con toda probabilidad no se espera encontrar la especie y un 0 si la especie se encuentra en esa celda x_i. El primer término favorece aquellos modelos que obtienen un *ROR* alto en los datos de sólo presencia, el segundo término penaliza los modelos que obtienen un *ROR* alto en los datos de fondo y el tercer término penaliza la complejidad de las transformaciones llevadas a

cabo por *MaxEnt*. Este tercer término se asegura de hacer que muchos coeficientes sean 0, reduciendo el número de variables seleccionadas por *MaxEnt* [22] y, al mismo tiempo, se permite introducir más variables a medida que se dispone de más datos debido a que este término está dividido por la raíz cuadrada de *M*.

Por último, es importante resaltar que se ha de realizar un muestreo adecuado al problema que se quiera resolver es fundamental. Seleccionar los datos de fondo de un modo adecuado es esencial para que *MaxEnt* se convierta en una herramienta que sea capaz de discriminar correctamente. También se ha de comentar que hay tres tipos de salidas del método *MaxEnt*: el *ROR* del que se ha hablado; el acumulado, que en una determinada celda toma la suma de todos aquellas celdas que tengan un *ROR* igual o menor que el de esa celda; y, el logístico, que consiste en una transformación del *ROR*. De estos tres, el logístico suele dar unos resultados mejores[22].

En este trabajo se ha hecho uso del paquete m*axent*[4] implementado en *R-cran* para entrenar los modelos de *MaxEnt*.

3.1.2 Arboles de decisión

Un árbol de decisión, o en inglés *decision trees*, es una técnica muy utilizada en minería de datos para explicar las relaciones entre algunas variables de entrada con una variable objetivo o para predecir el valor que toma esta última a partir de las primeras. Los árboles de decisión no son sólo empleados en minería de datos, de hecho su origen procede de la lógica y la estadística. Hoy en día se emplean en minería de texto, en la extracción de información, en aprendizaje supervisado y reconocimiento de patrones. Esto es debido a todos los beneficios que poseen:

- Son versátiles. Pueden ser utilizados tanto en clasificación, en regresión, en *clustering*, en selección de variables, etc.
- Son fáciles de entender y de explicar. Incluso, si los árboles son de gran tamaño, su explicación no es compleja sino extensa.
- Son flexibles a la hora de tratar diversos tipos de datos. Pueden ser nominales, numéricos o textuales.
- Su adaptabilidad ante la falta de datos o errores en los datos es buena.
- Tienen una buena capacidad predictiva y, además, su coste computacional es relativamente bajo.
- Poseen un buen comportamiento en conjuntos de datos muy grandes.

En minería de datos, un árbol de decisión puede ser utilizado para representar tanto un modelo de clasificación como uno de regresión. En el primer caso se le llama árbol de clasificación y en el segundo árbol de regresión. En nuestro caso, tan sólo estamos interesados en estudiar los árboles de clasificación.

Los árboles de clasificación son empleados en campos como las finanzas, el *marketing*, la ingeniería y la medicina debido a su buen uso como técnica exploratoria. Sin embargo, no se ha de olvidar el hecho de que existen muchas otras técnicas, como las máquinas de vector de soporte o las redes neuronales, para abordar los problemas de clasificación o predicción y que esta herramienta

4 http://cran.r-project.org/web/packages/maxent/maxent.pdf

no pretende reemplazarlas sino ser una más.

Un árbol de decisión es un clasificador expresado como una partición del espacio de instancias, que es el conjunto de todos los posibles combinaciones de casos diferentes que pudiera haber en los datos del problema en cuestión. El árbol de decisión consiste en una serie de nodos que forman un árbol con raíz. Un árbol con raíz consiste en un árbol dirigido que tiene un nodo llamado raíz al cual no llegan ninguna arista. Además, a todos los otros nodos tan sólo les llega una arista. Un nodo del que no salga ninguna arista se conoce como nodo interno, nodo test o como hoja, éstos son los nodos en los que se especifica cuál es la instancia. Todos los otros nodos con llamados ramas, aunque también podemos llamarlos nodos de decisión por lo que veremos más adelante.

En un árbol de decisión, cada nodo interno divide el espacio de instancias en 2 o más subespacios en base a una función discreta a la que llegan valores de las variables de entrada, también llamadas atributos. En el caso más sencillo y más frecuente, en cada uno de los nodos de decisión tan sólo se considera un único atributo, de modo que el espacio de instancias se divide en función del valor de este atributo. Evidentemente, en el caso de valores numéricos se establecen rangos para poder dividir el espacio. Además, un mismo atributo puede ser requerido en varios nodos de decisión a lo largo de la construcción del árbol de decisión.

Los datos son clasificados sin más que recorrer el árbol desde la raíz hasta cualquier hoja, la hoja a la que llegue será decidida en base a los resultados obtenidos en los nodos de decisión por los que haya pasado.

A continuación vamos a exponer un ejemplo sencillo extraído de wikipedia que muestre cómo podemos dividir el espacio de instancias. Supongamos que disponemos algunos datos de los alumnos de cierta asignatura y queremos saber en qué grupo de estudiantes se encuentran: aquellos que se han de presentar al examen final, aquellos que están exentos de él o aquellos que han de ir directamente al examen extraordinario. Para ello, disponemos de 2 atributos: el aprovechamiento de la asignatura por un lado y la puntualidad y asistencia por otro. En función del valor que un

Ilustración 3.1 En esta ilustración se muestra un ejemplo de un árbol de decisión con tres clases y dos atributos. Recuperado el 10 de julio de 2015, de https://es.wikipedia.org/wiki/%C3%81rbol_de_decisi%C3%B3n_(modelo_de_clasificaci%C3%B3n_ID3).

estudiante obtenga en estos dos atributos será clasificado en uno de los tres grupos de estudiantes. En la Ilustración 3.1 se puede apreciar cuáles han sido los criterios para poder separar el espacio de instancias.

El hecho de que podamos dividir el espacio de instancias tantas veces como queramos nos lleva a pensar que esto podría producir un enorme sobreajuste y que, por lo tanto, esta técnica no generalizaría bien. Este es uno de los motivos por los que a la hora de construir los árboles de decisión se tenga que tener en cuenta el tamaño de este. Generalmente la complejidad de árbol de decisión se mide teniendo en cuenta alguno de los siguientes parámetros: el número total de nodos, el número total de hojas, la profundidad del árbol (longitud del camino más largo desde la raíz) y el número de atributos usados. La complejidad del árbol de decisión es controlada acotando cada uno de estos parámetros o alguno de ellos.

Los algoritmos que tratan de construir un árbol de decisión a partir de unos determinados datos están basados, en general, en minimizar los errores de clasificación sujetos a ciertas restricciones, como pueden ser limitar la complejidad del árbol teniendo en cuenta alguno de los parámetros del párrafo anterior.

El hecho de que no se busque el árbol de decisión óptimo, el de menor complejidad, y que permita separar perfectamente los datos dados es porque es considerado como un problema complicado. De hecho, está probado que encontrar el mínimo árbol de decisión consiste en un problema *NP* duro. Por lo tanto, encontrar el árbol de decisión óptimo no es un problema abordable. Esto nos conduce a una situación en la que hemos de emplear técnicas heurísticas. Estos algoritmos suelen ser *greedy* y suelen considerar cuál es el atributo que mejor divide el espacio en cada uno de los nodos. En algunos algoritmos heurísticos, aparte de dejar que el árbol crezca, también se eliminan algunas hojas. Esta operación es conocida como poda, en inglés *pruning*. Estas operaciones de dejar crecer el árbol y podarlo continúan hasta que alguno de los criterios de parada son satisfechos.

Por último, los algoritmos de árboles de decisión más conocidos son los siguientes: *ID3 (Iterative Dichotomiser 3), C4.5, CART (Classification and Regression Tree)* y *CHAID (CHi-squared Automatic Interaction Detector)*.

En este trabajo se ha hecho uso del algoritmo de Quinlan's C5.0 implementado en *R-cran* en el paquete C5.0[5].

Si el lector quiere ampliar lo visto en este subapartado se le invita a seguir el libro [23].

3.2 Evaluación de la calidad de los modelos

En este subapartado se van a describir cuáles han sido las herramientas utilizadas para poder valorar si los clasificadores o los modelos que se han generado son satisfactorios o no. El AUC servirá como un indicador de cómo de bueno es un clasificador binario probabilístico y la metodología de *cross validation*, se empleará con el objeto de tener evidencias de que los clasificadores funcionarán razonablemente bien en zonas en las que no se disponga de cartografía,

5 http://cran.r-project.org/web/packages/C50/C50.pdf

es decir, que son capaces de generalizar. También se describirán las tablas de contingencia que se han empleado para evaluar los resultados obtenidos por un clasificador multecategoría.

3.2.1 Curvas ROC. AUC

Una curva *ROC* (acrónimo de *Receiver Operating Characteristic*, en castellano Característica Operativa del Receptor) es una técnica empleada para visualizar cómo de bueno es un determinado clasificador basándonos en cómo haya sido su comportamiento.

Nos centraremos en analizar el caso en el que tan sólo disponemos de dos clases *{p, n}*, positivos y negativos respectivamente. Los clasificadores binarios tratan de predecir correctamente estas dos clases. Para distinguir entre la clase real y la clase predicha por los clasificadores, denotaremos a estas últimas por *{Y, No}*. La respuesta dada por estos clasificadores no ha de ser necesariamente *Y* o *No* inicialmente, debido a que estos clasificadores pueden ser probabilistas, lo cual significa que podemos tener como resultado del clasificador un valor continuo. Más adelante se explicará cómo podemos actuar en esa situación y convertir ese valor continuo en *Y* o *No*.

Dado un clasificador y una instancia, se tienen cuatro casos distintos. Si la instancia es positiva y ha sido clasificada como positiva, este caso se cuenta como un verdadero positivo; si en cambio ha sido clasificada como negativa, en ese caso se cuenta como un falso negativo. Si la instancia es negativa y ha sido clasificada como negativa, este caso se cuenta como un verdadero negativo; si en cambio ha sido clasificada como positiva siendo negativa se cuenta como un falso positivo. Todo lo que se ha descrito en este párrafo puede verse representado en la siguiente tabla de contingencia, también llamada matriz de confusiones.

		Clase verdadera	
		p	n
Clase predicha	Y	Verdaderos positivos	Falsos positivos
	No	Falsos negativos	Verdaderos negativos
Número de instancias		P	N

Tabla 3.1 Esta es la tabla de contingencia descrita en el párrafo anterior. Además, en la última se ha añadido el número de instancias de la clase p y el número de instancias de la clase n.

Aquellos casos bien clasificados se corresponden con los verdaderos positivos y con los verdaderos negativos, que coinciden con los elementos de la diagonal de esta tabla de contingencia 2x2 y aquellos elementos fuera de la diagonal son los errores, las confusiones del clasificador.

A partir, de la Tabla 3.1 se pueden calcular algunos términos interesantes y de mucha utilidad. Siendo *VP*, los verdaderos positivos; *FP*, los falsos positivos; *FN*, los falsos negativos; *VN*, los verdaderos negativos.

- Precisión o exactitud (*accuracy*): $(VP+VN)/(P+N)$.
- Tasa de falsos positivos o tasa de falsas alarmas *(false positive rate* o *false alarm*

rate): FP/N .

- Tasa de verdaderos positivos *(true positive rate* o *hit rate*): VP/P . Que coincide coincide con un término sumamente importante como es la sensibilidad (*sensitivity)*. Más adelante veremos que juega un papel fundamental en el cálculo del *AUC* y cuando expliquemos en qué consiste las cuvas *ROC*.
- Especificidad (*specificity*): VN/N . Que a su vez puede ser escrita como 1 menos la tasa de falsas alarmas.

En este momento disponemos de todos los términos necesarios para poder hablar de las curvas ROC. Las curvas *ROC* son gráficas en dos dimensiones donde la tasa de verdaderos positivos o la sensibilidad se dibuja en el eje *Y* y la tasa de falsos positivos (1 menos la especificidad), en el eje *X*. Esto puede observarse en la siguiente ilustración.

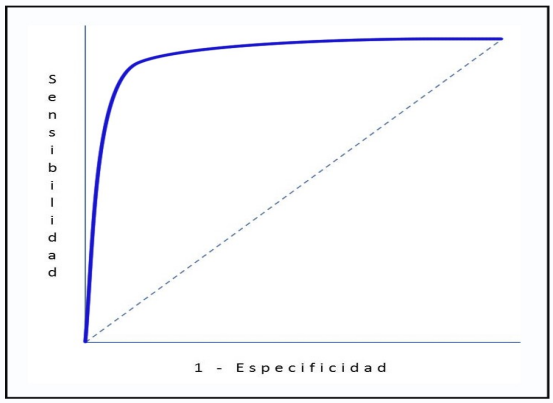

Ilustración 3.2 En esta ilustración se muestra una curva *ROC*, en el eje *X* está dibujada 1 menos la especificidad y en el eje *Y*, la sensibilidad. Recuperado el 10 de julio de 2015, de http://anestesiar.org/2014/el-dilema-del-vigilante-sensibilidad-y-especificidad/.

Hay varios puntos del espacio *ROC* que tienen una gran importancia. El punto (0, 0) representa la estrategia de nunca dar una clasificación positiva, es decir, que no hay verdaderos positivos pero tampoco falsos positivos. La estrategia contraria, siempre dar una clasificación positiva, nos lleva al otro extremo, el punto (1, 1). El punto que representa la clasificación perfecta es el (0, 1). Hablando un poco informalmente se podría decir que es mejor que un punto en el espacio *ROC* esté cercano al punto (0, 1) y es malo cuanto más se acerca a la diagonal. Cualquier punto que se encuentre sobre la diagonal implica que el clasificador asociado a ese punto sigue una estrategia de predicción azarosa, es decir, que independientemente de la información de la que dispongamos la predicción es realizada de manera aleatoria (con cierta probabilidad *p* predigo positivo y con probabilidad 1 – *p* predigo negativo). Es claro que, encontrarse sobre la diagonal no implica que el clasificador sea aleatorio, sin embargo, nos da pistas de que no se está explotando la información disponible del modo adecuado o todo lo bien que se debería. Hay casos incluso peores que encontrarse sobre la diagonal y es encontrarse por debajo de ella, esto significa que estamos interpretando la información en el sentido equivocado y que sin más que invertir las predicción

dadas por el clasificador, es decir, invertir los resultados positivos y los negativos, se consigue un clasificador que se encontraría por encima de la diagonal. Esto es debido a que los verdaderos positivos del 'mal' clasificador se convierten en los falsos negativos del nuevo clasificador y los falsos negativos del 'mal' clasificador se convierten en los verdaderos positivos.

Hay que comentar que en el espacio *ROC* el hecho de que un clasificador consiga situarse ligeramente por encima de la diagonal puede entenderse como un clasificador aleatorio o bien como un clasificador que está haciendo poco uso de la información disponible. Sin embargo, no está claro en cual de ambos casos nos encontramos.

Hablando informalmente, aquellos clasificadores que se sitúan en la zona izquierda del espacio *ROC* son clasificadores conservativos, sólo realizan predicciones positivas cuando tienen fuertes evidencias, esto les lleva a cometer pocos errores y, al mismo, tiempo a mantener una tasa de verdaderos positivos baja. En cambio, aquellos clasificadores que se encuentran en la zona de arriba a la derecha son considerados clasificadores liberales, éstos se caracterizan por tener una alta sensibilidad, sin embargo esta alta tasa de verdaderos positivos suele ser alcanzada al realizar predicciones positivas sin la evidencia suficiente, por lo que les conduce a tener, a su vez, una alta tasa de falsas alarmas. La elección entre varios tipos diferentes de clasificadores está condicionada en muchas ocasiones al problema que se pretenda resolver. Por ejemplo, si lo que se pretende es detectar qué personas padecen una enfermedad grave podría quererse tener una alta tasa de verdaderos positivos, aunque esto conllevase a su vez una alta tasa de falsas alarmas ya que no queremos arriesgarnos a un posible contagio posterior.

En los párrafos anteriores hemos estado comentando que los clasificadores tienen asociados un punto en el espacio *ROC* en función de cuál sea su sensibilidad y su especificidad. Sin embargo, esto depende del tipo de clasificador con el que estemos tratando. Si el clasificador es binario no probabilista, es decir, que sus resultados son *Y* o *No*, al representar este clasificador en el espacio *ROC* se verá como un único punto. En cambio, si el clasificador es probabilista y nos proporciona como resultado un valor numérico entonces su representación en el espacio *ROC* no es un punto sino una curva, la curva *ROC*. Estos valores numéricos no necesariamente son probabilidades, de hecho, en general, son valores no calibrados con la propiedad de que un valor mayor implica una mayor probabilidad. Es claro que a partir de un clasificador probabilista podemos construir un clasificador binario del siguiente modo: se fija un valor *a priori* llamado umbral, si el resultado dado por el clasificador probabilista es mayor que ese umbral, la clase predicha será una de las dos o bien *Y* o *No*, y si es menor o igual que dicho umbral, entonces será la otra. Esto nos lleva a que para cada valor del umbral se disponen de unos resultados del clasificador u otros, por lo que la representación en el espacio *ROC* de este clasificador es una curva en función de los diferentes umbrales posibles que varían entre 0 y 1.

En este momento estamos en condiciones de explicar lo que es el *AUC*. El *AUC* es un valor que nos permite saber cómo de bueno ha sido el comportamiento de cierto clasificador. Para calcular este valor hay que hallar el área que se encuentra por debajo de curva *ROC*. Es claro a partir de su definición que este valor está acotado inferiormente por 0 y superiormente por 1. Sin embargo, como ya explicábamos anteriormente, ningún clasificador debería situarse por debajo de la diagonal, por lo que el *AUC* más bajo debería ser 0.5 y no 0.

El *AUC* tiene una propiedad estadística muy interesante: el *AUC* de un clasificador equivale a la probabilidad de que el clasificador puntúe o valore más alto una instancia positiva elegida aleatoriamente que a una instancia negativa elegida aleatoriamente.

El hecho de que un clasificador *A* obtenga un *AUC* mayor que otro *B*, como puede apreciarse en la Ilustración 3.3 no implica que la curva *ROC* del clasificador *A* se encuentre en todo momento por encima que el *B*, sin embargo el obtener un mayor *AUC* podría ser tomado como una prueba de que el clasificador *A* es mejor clasificador que el *B*. Se considera que el *AUC* se comporta generalmente bien, por ello se usa a menudo a la hora de comparar clasificadores o a la hora de valorar cómo de bueno es un determinado clasificador.

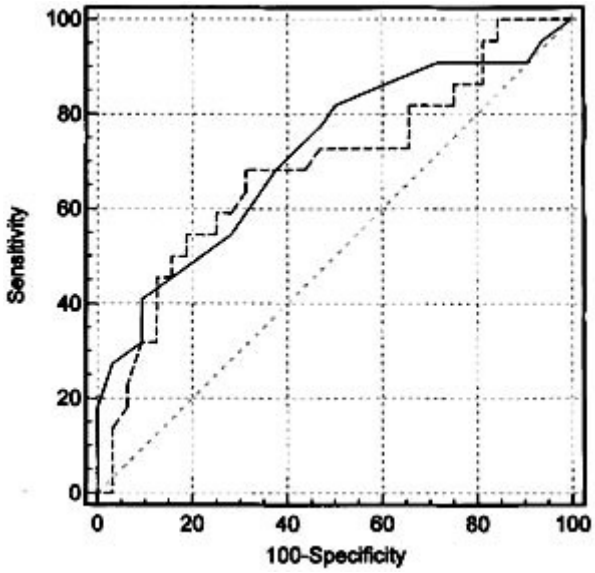

Ilustración 3.3 En esta ilustración se muestra una 2 curvas *ROC*, en el eje *X* está dibujada 100 menos la especificidad y en el eje *Y*, la sensibilidad, ambas medidas en tanto por ciento. Recuperado el 10 de julio de 2015, de http://www.scielo.cl/scielo.php?pid=S0034-98872006000300006&script=sci_arttext.

Si el lector está interesado en conocer más sobre las curvas *ROC*: qué tipo de algoritmos se pueden usar para construir estas curvas, qué algoritmos se emplean para estimar el *AUC* o conocer cómo se comportan estos términos en casos más complejos, como puede ser en el caso de que no tengamos únicamente dos clases (*Y* o *No*) sino varias, le invitamos a seguir la lectura de [24] donde podrá encontrar las respuestas a todos estos interrogantes.

3.2.2 Validación Cruzada

En estadística y en minería de datos, una cuestión bastante común es el de tratar de que un modelo, por ejemplo un modelo de regresión o uno de clasificación, aprenda a partir de unos datos. Pero, al tratar de hacer uso de un modelo que se ha comportado razonablemente bien en los datos de entrenamiento sobre unos datos de prueba, puede ocurrir que la predictibilidad mostrada por el

modelo no sea la esperada, obteniéndose unos resultados claramente peores. La validación cruzada, en inglés *cross-validation*, es un procedimiento que nos permite estimar cómo de bien pasará a comportarse un determinado modelo sobre unos datos desconocidos respecto a cómo se ha comportado en unos datos de entrenamiento. La idea de la validación cruzada surgió en los años 30. Actualmente, esta técnica está ampliamente aceptada y es frecuentemente empleada como un procedimiento estándar para estimar el buen comportamiento de modelos y poder elegir de entre varios de ellos.

Después de esta breve introducción, se va a explicar en qué consiste la validación cruzada exactamente. Se trata de un método estadístico que permite evaluar y comparar modelos al dividir los datos en dos grupos: uno de ellos usado para entrenar el modelo y el otro, para validarlo. Existen distintas formas de realizar la validación cruzada: validación cruzada de k iteraciones, validación cruzada aleatoria, validación cruzada dejando uno fuera, etc. Nos centraremos en explicar en qué consiste la validación cruzada de k iteraciones pues ésta será la que emplearemos a lo largo de la realización del trabajo.

En la validación cruzada de k iteraciones el conjunto de los datos es dividido aleatoriamente en k subconjuntos iguales a ser posible, si no fuese posible (el número de datos no es divisible por k) serían divididos en subconjuntos lo más parecidos en tamaño. En la Ilustración 3.4 se puede observar cómo es la división que se acaba de describir. Después de haber realizado el primer paso, se llevarán a cabo k iteraciones, en la iteración i se tomará el subconjunto i como la muestra de validación y los otros $k - 1$ subconjuntos como el conjunto de entrenamiento sobre el que se trabajará para entrenar el modelo. Por lo tanto, se tendrán que generar k modelos diferentes que serán validados en el correspondiente subconjunto de validación. Es decir, que el coste computacional de llevar a la práctica esta metodología es mayor cuanto más grande sea el k, bajo la hipótesis de que el tamaño de los datos es grande y k, pequeño. En la literatura, se suele tomar k igual a 10, aunque, evidentemente, esta elección depende en gran medida del tamaño de los datos y del coste computacional que tenga la generación de los modelos. En muchas ocasiones en las que no se dispone de suficientes muestras, los datos son estratificados al emplear esta técnica. La estratificación consiste en reordenar los datos de modo que nos aseguremos de que cada uno de los k subconjuntos es un buen representante del total de los datos.

Después de haber aplicado la técnica anterior a un conjunto de datos, se ha disponer de una medida de precisión fijada *a priori* (*AUC*, *hit*, …) que nos muestre cómo de bien se ha comportado cada uno de los k modelos entrenados. Por lo que tendremos a nuestra disposición k medidas de precisión, una para cada uno de los modelos. Se puede hacer uso de diferentes metodologías de agregación para poder comparar los resultados obtenidos entre dos algoritmos distintos sobre el mismo conjunto de datos y ser capaces de lanzar la hipótesis sobre si uno de ellos es mejor que el otro.

A continuación, vamos a enunciar dos de los posibles objetivos que se podrían tener al aplicar una validación cruzada:

- Estimar cómo de bien pasará a comportarse un determinado modelo sobre unos datos desconocidos. Es decir, para aseverar la generalización del modelo.
- Para comparar varios algoritmos o modelos y poder decidir cuál de ellos se comporta mejor en los datos dados.

Para ampliar los conocimientos de este subapartado puede seguirse [25].

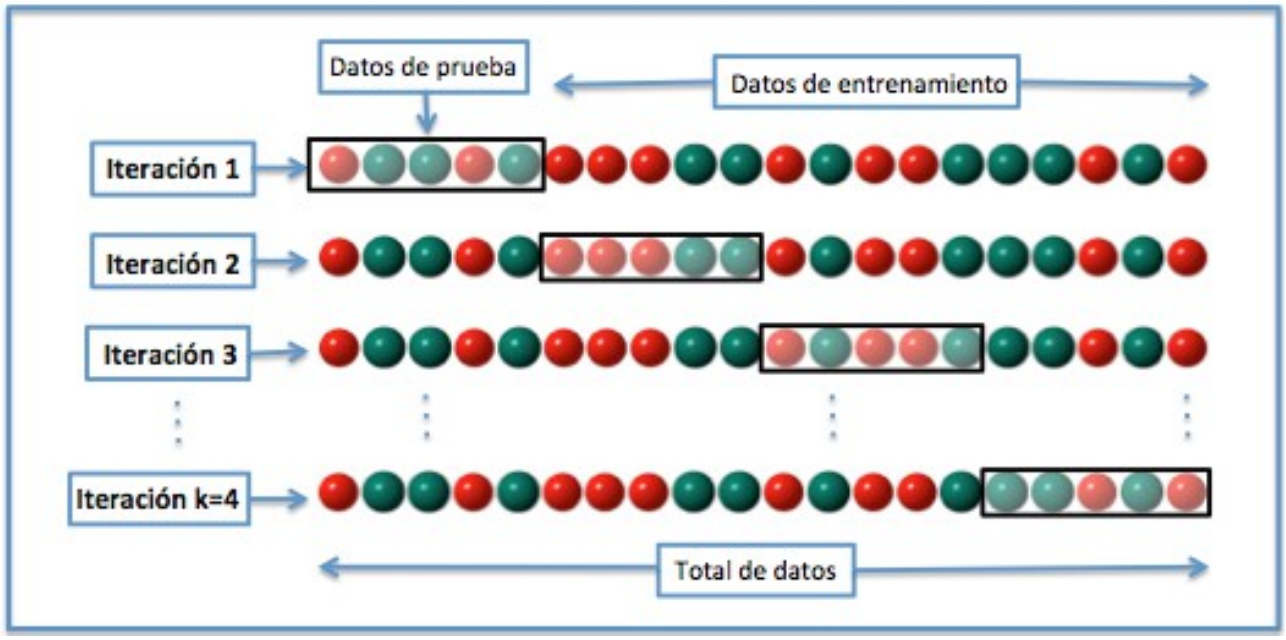

Ilustración 3.4 En esta ilustración se muestra cómo es la división realizada en una validación cruzada de 4 iteraciones. Recuperado el 10 de julio de 2015, de https://es.wikipedia.org/wiki/Validaci%C3%B3n_cruzada.

4 Análisis y resultados

En esta sección se describirán con todo detalle todos los pasos descritos en la introducción realizada en la sección de Metodología. Además, también se expondrán cuáles han sido los resultados al aplicar la metodología descrita.

Todo la implementación y los resultados obtenidos al aplicar la metodología que se describe en esta sección ha sido realizada haciendo uso de *R-cran.*

4.1 Planteamiento del problema

En esta subsección se van a describir todos los pasos seguidos en la resolución del problema.

Recordemos que el problema que se quiere resolver consiste en determinar la clase de vegetación con mayor presencia en un punto a partir de los valores que tomen las bandas del espectro electromagnético en ese punto concreto.

Para lograr clasificar satisfactoriamente las clases de vegetación en puntos no cartografiados, es necesario encontrar una o varias técnicas de clasificación que al emplear la validación cruzada de *k* iteraciones obtengan unos resultados satisfactorios tanto en los conjuntos de entrenamiento como en los conjuntos de validación y que, además, esos resultados obtenidos en ambos conjuntos sean similares. De este modo, tendremos evidencias que nos permitirán afirmar que los modelos que se han construido son generalizables.

Como ya se avanzó en la sección de Metodología, la primera y más sencilla idea con la que se trabajó fue la de utilizar técnicas de clasificación que nos devolviesen como valor de salida un número entero entre 1 y el número de clases de vegetación disponibles en la cartografía. Sin embargo, se comprobó experimentalmente que los resultados obtenidos no eran los deseados, por lo que hubo que buscar una nueva forma de abordar el problema.

Observando que una estrategia tan directa como la anterior no funcionaba, lo que se trató de hacer fue dividir el problema en otros más simples y más fáciles de resolver. Es decir, seguir una estrategia de *divide y vencerás*.

En vez de entrenar un único modelo de clasificación multicategoría, se entrenaron *N* modelos diferentes, donde *N* es el número de clases de vegetación de la cartografía, siguiendo la estrategia de *ONE vs ALL* descrita en la sección de Metodología.

Se probaron diferentes técnicas de clasificación binaria (*MaxEnt*, *MARS*, *Suppor Vector Machines*, *Boosting*, *Bagging*, *Random Forest*...). Al analizar los resultados obtenidos, se observó que *MaxEnt* lograba mejores resultados que las demás técnicas empleadas, coincidiendo con lo visto en [1]. A pesar de ello y con la esperanza de que se pudiese utilizar la información generada por las otras técnicas, se emplearon métodos (ver introducción de la sección de Metodología) que

nos permitiesen juntar los resultados dados por todas ellas. Sin embargo, la mejora no era apreciable con respecto a los resultados obtenidos por *MaxEnt*, por lo tanto se decidió descartar el resto de modelos para no aumentar el coste computacional de la resolución del problema.

Antes de proseguir con los pasos seguidos para resolver este problema se ha de explicar brevemente ciertos aspectos sobre la generación de los modelos de *MaxEnt*.

Cuando se generan los modelos *MaxEnt*, se calculan varios umbrales, o *thresholds* en inglés, que son unos valores que permiten decidir si la especie asociada al modelo está presente en los diversos puntos. Estos umbrales son calculados de varias formas: buscando maximizar el porcentaje de aciertos, buscando que la proporción de puntos donde se da la presencia de la clase de vegetación coincida con la proporción de puntos que hay de esa clase de vegetación en el conjunto de entrenamiento, etc. En el caso de que la salida dada por *MaxEnt* en un punto sea superior al valor del umbral seleccionado, en ese punto se detectaría la presencia de la clase de vegetación asociada a ese modelo de *MaxEnt.*

Como se dispone de *N* modelos de *MaxEnt,* cada uno de ellos asociados a una clase de vegetación, puede ocurrir que en un punto varios de los modelos detecten a la vez la presencia de sus clases de vegetación asociadas. Sin embargo lo que uno desearía es que en cada punto tan sólo uno de los *N* modelos detectase la presencia de su clase de vegetación asociada. De este modo, tendríamos una situación donde clasificaríamos todos los puntos fácilmente, sin más que asociarle a cada punto la etiqueta de esa única clase de vegetación que se ha detectado.

Sin embargo, la realidad chocaba con el caso ideal descrito en el párrafo anterior. Lo que ocurría era que en un 74% de los puntos, nos encontrábamos en la situación ideal descrita en el párrafo anterior; en un 24% de los puntos, todos los modelos 'negaban' la presencia de su especie asociada, es decir, ningún modelo tenía suficiente evidencia como para 'afirmar' que se diese la presencia su clase de vegetación; y, en un 2% de los puntos, la situación era que había varios modelos que detectaban la presencia de sus clases de vegetación, lo que podría interpretarse como la coexistencia de varios tipos de vegetación en un recinto de 25 metros cuadrados.

En las siguientes ilustraciones: Ilustración 4.1, Ilustración 4.2, Ilustración 4.3 e Ilustración 4.4 se muestran cómo son las distribuciones de las probabilidades de las predicciones dadas por 8 de los modelos. No se han mostrado todos los modelos en esta subsección, el resto de ellos se pueden encontrar en el apéndice, más concretamente, en las ilustraciones desde la Ilustración 7.2 hasta la Ilustración 7.6.

Lo que cabría esperar al ver estos histogramas con frecuencias absolutas es que tuviesen todos la forma del modelo de agua en la Ilustración 4.1. Es decir que tuviesen una especie de valle entre dos montañas, una de ellas cerca del mínimo valor dado por el modelo y la otra cerca del máximo valor dado por el modelo. Además, también cabría esperar que el umbral seleccionado por el modelo de esa clase de vegetación se situase entre ambas montañas. Sin embargo, este comportamiento deseable no es el que se aprecia en la mayor parte de los modelos haciendo complicado juntar toda la información proporcionada por los modelos.

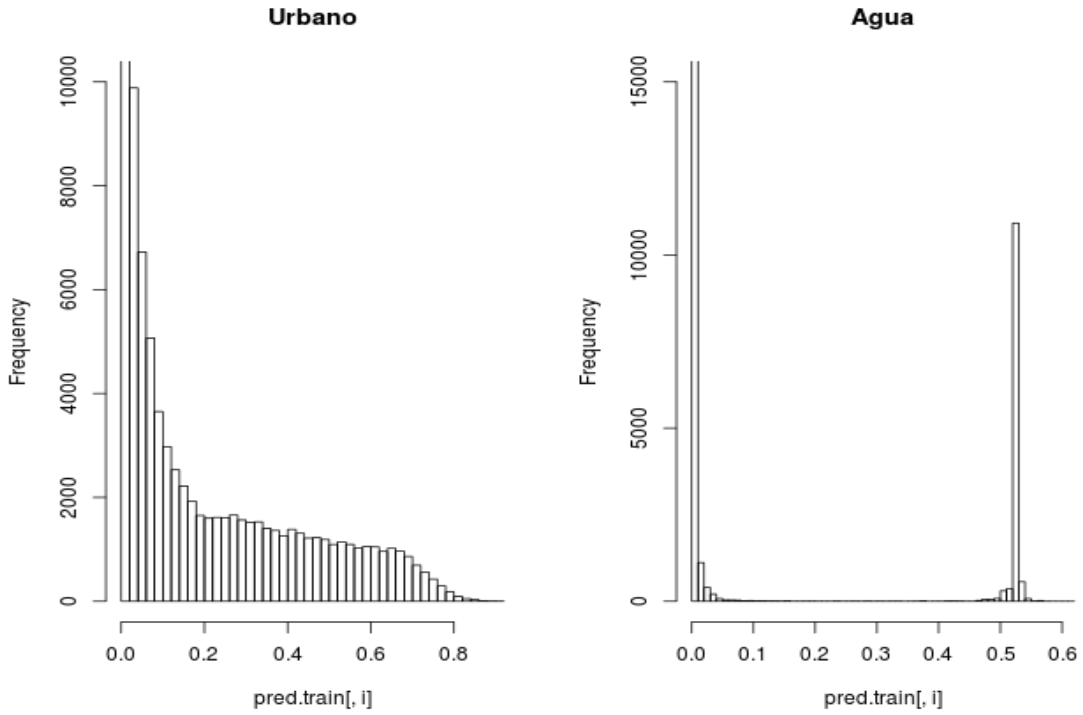

Ilustración 4.1 En este ilustración se muestra en el eje Y el número de píxeles y en el eje X los valores de la predicción dada por el modelo de la clase urbano, a la izquierda, y por la clase agua, a la derecha.

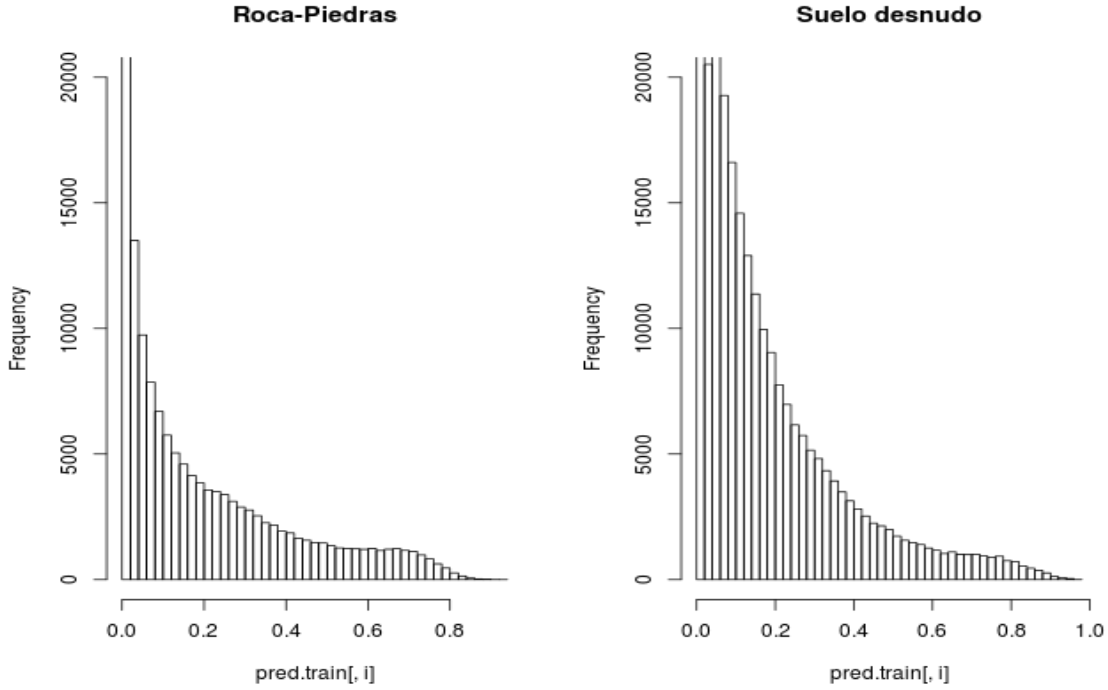

Ilustración 4.2 En esta ilustración se muestra en el eje Y el número de píxeles y en el eje X los valores de la predicción dada por el modelo de la clase roca-piedras, a la izquierda, y por la clase suelo desnudo, a la derecha.

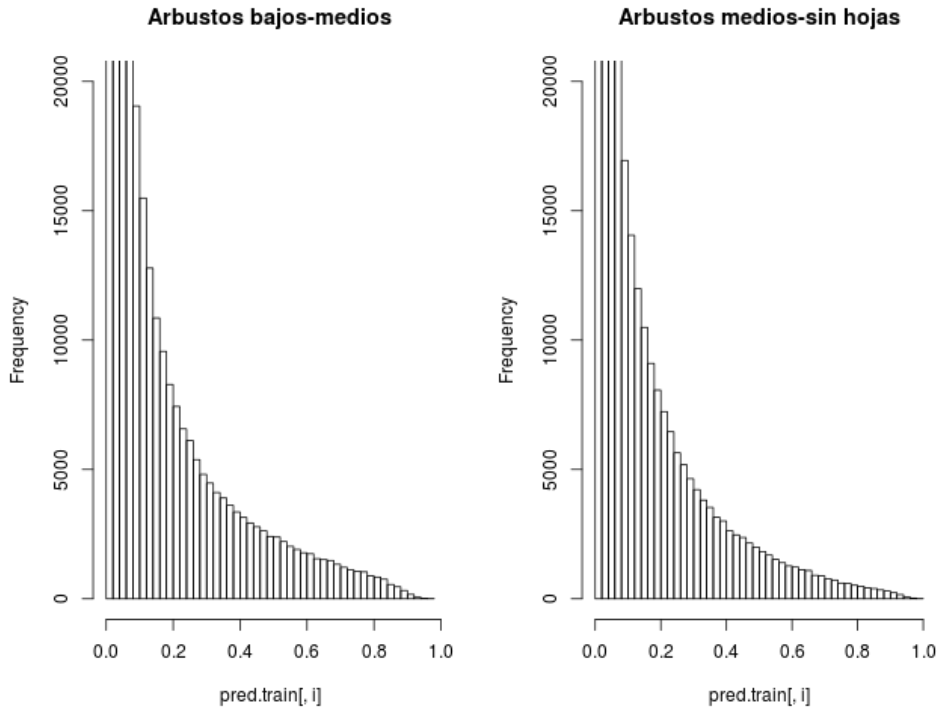

Ilustración 4.3 En esta ilustración se muestra en el eje Y el número de píxeles y en el eje X los valores de la predicción dada por el modelo de la clase arbustos bajos-medios, a la izquierda, y por la clase arbustos medios-sin hojas, a la derecha.

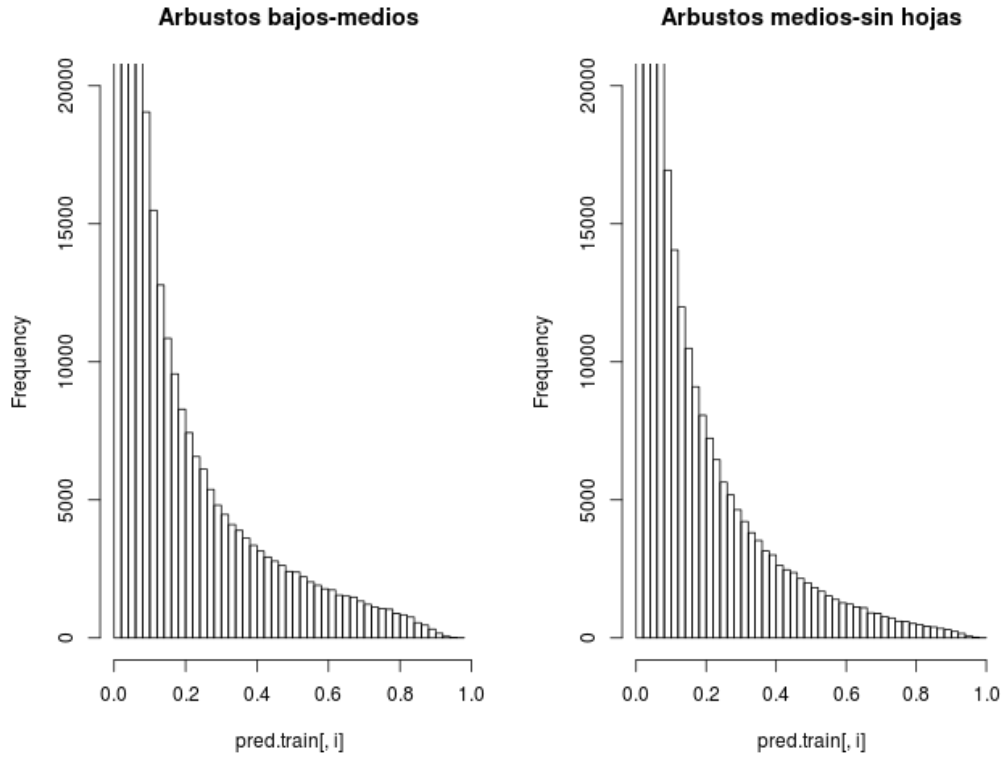

Ilustración 4.4 En esta ilustración se muestra en el eje Y el número de píxeles y en el eje X los valores de la predicción dada por el modelo de la clase arbustos bajos-medios-sin hojas, a la izquierda, y por la clase arbustos medios-altos-caduca, a la derecha.

Como una primera aproximación para juntar la información obtenida por los modelos, parecía razonable asumir que el 74% de los puntos en los que sólo una clase de vegetación era detectada fuesen etiquetados por esa clase de vegetación y dar un trato especial al resto de los puntos. Sin embargo, después de probar varias metodologías heurísticas basadas en cómo de cerca estuviésemos de los umbrales dados por *MaxEnt* y, por otro lado, de aplicar algunas técnicas basadas en *decision templates* [16], decidimos construir un árbol de decisión, más concretamente un árbol de clasificación. Este árbol de clasificación recibía como entrada en cada uno de los puntos los *N* valores dados por los modelos y tenía como valor de salida la etiqueta de la clase de vegetación como sería clasificado ese punto. Un dato que muestra hasta qué punto se está de satisfecho con los resultados obtenidos por el árbol de clasificación es el siguiente: en el 74% de los puntos donde la clasificación parecía sencilla (tan sólo un modelo detectaba su clase de vegetación asociada), el árbol de clasificación obtiene una precisión *(accuracy)* más elevada que la obtenida al clasificar esos puntos con la clase de vegetación detectada. Esto puede deberse a que el árbol de decisión tiene en cuenta cuáles son los modelos cuya información es más fiable, es decir, si se da el caso de que un modelo realiza una clasificación perfecta entonces el árbol de decisión consideraría dividir el espacio de instancias usando la información proporcionada por ese modelo antes que cualquier otro. Además de los buenos resultados obtenidos por el árbol de clasificación se ha de destacar también el poco coste computacional que añadía al proceso.

Además de todos los pasos para resolver el problema descritos hasta ahora, también se va a hablar del trato previo que se le dieron a los datos disponibles.

Se detectaron algunos errores en la cartografía debidos a que la fecha de las imágenes de satélite no es la misma que la fecha de la realización de la cartografía. Los errores que pudieron ser tratados *a posteriori* son aquellos que pueden ser detectados a partir del *NDVI*. Es sabido que las clases de vegetación que engloban a las plantas tienen un valor en la banda del *NDVI* elevado, en cambio, en aquellas como urbano, las rocas o suelos desnudos es todo lo contrario. Al revisar la cartografía se encontraron clases de plantas con un valor en la banda del *NDVI* sorprendentemente bajos y clases de improductivo con un valor en la banda del *NDVI* demasiado alto. Por ello se decidió descartar trabajar con todos aquellos puntos que fuesen de muy dudosa calidad.

También se decidió dividir una de las clases de vegetación originales, la clase de herbáceas con buen valor forrajero. Esto fue motivado porque no se le encontraba ninguna explicación razonable a las confusiones de esta clase de vegetación con las demás. Lo que se quería comprobar era si al dividir esta clase en varias se podían explicar los errores cometidos por separado. Esta división fue realizada en base al valor que tomaban los píxeles cartografiados como herbáceas con buen valor forrajero en la banda del *NDVI*. En la Tabla 4.1 se muestra cuál fue la división final de esta clase de vegetación. Cabe mencionar que de esta clase de vegetación no fueron descartados ningún punto de dudosa calidad sino que aquellos que iban a ser descartados fueron catalogados como una de las nuevas clases de vegetación, más concretamente, como aquellas que representan las herbáceas con buen valor forrajero con valores en la banda del *NDVI* más bajos.

Nombre de la clase de vegetación	Descripción
TL-herbáceas	Herbáceas. Buen valor forrajero. *NDVI* muy bajo.
H1-herbáceas	Herbáceas. Buen valor forrajero. *NDVI* bajo.
H2-herbáceas	Herbáceas. Buen valor forrajero. *NDVI* normal.
H3-herbáceas	Herbáceas. Buen valor forrajero. *NDVI* alto.
H4-herbáceas	Herbáceas. Buen valor forrajero. *NDVI* muy alto.

Tabla 4.1 En esta tabla se muestra la descripción de las nuevas de vegetación en las que se dividió las herbáceas con buen valor forrajero.

4.2 Resultados

En esta subsección vamos a mostrar los resultados obtenidos al llevar a cabo la metodología descrita tanto en la sección de Metodología como en la subsección de Planteamiento del problema.

Recordemos que se ha dividido Cantabria en 8 zonas debido a las limitaciones de los datos con los que se trabaja. Debido a que la información que se podría llegar a generar de todas estas zonas podría llegar a ser abrumadora, tan sólo se van a presentar los resultados de una de las zonas, más concretamente, de la denominada zona *B* (véase la Ilustración 2.3). Esta es una de las zonas en las que se cuenta con más puntos cartografiados, aproximadamente unos 600000 píxeles de 5x5 metros, es decir, alrededor de 15 kilómetros cuadrados de superficie cartografiada. A pesar de que sólo se muestren los resultados de esta zona, en las demás zonas se siguió un desarrollo muy similar exceptuando algunas pequeñas diferencias que dependen de aspectos como el mes en el que fueron tomadas las imágenes de satélite en esa zona, la ausencia de algunas especies en esas zonas, etc.

El primer pasó de todos fue aplicar la validación cruzada de 10 iteraciones a las predicciones realizadas por *MaxEnt* en la cartografía de la zona *B*.

Se ha de comentar que, en general, los modelos de *MaxEnt* generados tenían un buen comportamiento en términos de *AUC* y que, además, no se apreciaban grandes diferencias entre el *AUC* obtenido en el los subconjuntos de entrenamiento y los subconjuntos de validación. Más concretamente, salvo 3 ó 4 de los 20 modelos generados el resto tenían un *AUC* superior a 0.9, es decir, eran modelos de muy buena calidad. Además, aproximadamente la mitad de ellos tenían un *AUC* superior a 0.95. Y aquellos que no tenían un comportamiento tan bueno se encontraban por encima del 0.85 de *AUC*. Con lo que los modelos de *MaxEnt* parecen satisfactorios.

Después de entrenar los modelos de *MaxEnt*, el siguiente paso fue entrenar y validar el árbol de clasificación. Para ello, también se siguió la técnica de la validación cruzada de 10 iteraciones.

En las siguientes tablas se van a mostrar la sensibilidad y la especificidad de cada clase de vegetación obtenidas a partir de las clasificaciones realizadas por el árbol de clasificación. Ambas medidas han sido calculadas para cada clase de vegetación asumiendo que los puntos que están cartografiados como esa clase de vegetación son positivos y el resto, negativos. Tan sólo vamos a mostrar 2 tablas, de la Tabla 4.2 a la Tabla 4.3. En el apéndice hemos mostrado otras 4, de la Tabla 7.1 a la Tabla 7.4.

En ninguno de los 10 conjuntos de validación se ha encontrado alguna clase de vegetación que muestre un claro sobreajuste. Las clases de vegetación que muestran mayores diferencias entre los resultados obtenidos en el conjunto de entrenamiento y los obtenidos en el conjunto de validación son las que tienen menos puntos que las demás.

Dado que en ninguna de las dos fases *(MaxEnt* y árbol de clasificación*)* se han detectado indicios de un posible sobreajuste tenemos evidencias de que el método es generalizable y, por tanto, se pueden generar las predicciones de la zona *B* completa.

Como dato adicional, se ha de decir que la precisión *(accuracy)* global obtenida en los conjuntos de entrenamiento es del 74.3% y en los conjuntos de validación, el 72.2%. Además, el rango de la precisión global de los 10 conjuntos de validación no era mayor del 1%, es decir que la dispersión era pequeña.

Clase de vegetación	Sensibilidad	Especificidad
Urbano	0,5996503497	0,9877040378
Agua	0,9935760171	0,9989020373
Roca-Piedras	0,1446540881	0,9928560172
Suelo desnudo	0,2011173184	0,9848179212
Arbustos bajos-medios	0,582010582	0,9828633406
Arbustos medios-sin hojas	0,08	0,9988002712
Arbustos bajos-medios-sin hojas	0,5715396579	0,9692350642
Arbustos medios-altos-caduca	0,1261682243	0,9800724638
Arbustos bajos-altos-perenne	0,25	0,9996873046
Arbustos alto y Arboles-caduca	0,646528404	0,9538225939
Arbustos altos-Arboles-perenne	0,413592233	0,9838338419
Coníferas y eucaliptos	0,644295302	0,9600787995
Herbáceas bastas	0,4566787004	0,9838536638
Geófitos	0,4843601896	0,9700126785
Juncales y turberas	0,1125	0,9962858339
TL-Herbáceas	0,7900552486	0,9980015777
H1-Herbáceas	0,6186770428	0,9948254924
H2-Herbáceas	0,8663990826	0,9866490947
H3-Herbáceas	0,9010989011	0,989640884
H4-Herbáceas	0,944285129	0,99188302

Tabla 4.2 En esta tabla se muestra la sensibilidad y la especificidad de cada clase de vegetación del primer conjunto de validación.

Clase de vegetación	Sensibilidad	Especificidad
Urbano	0,5939942411	0,9882433519
Agua	0,9954797198	0,9992510677
Roca-Piedras	0,2279360667	0,9935153623
Suelo desnudo	0,1918117344	0,9852238234
Arbustos bajos-medios	0,6321494794	0,9844409855
Arbustos medios-sin hojas	0,0661478599	0,9986087763
Arbustos bajos-medios-sin hojas	0,6049288618	0,9710174641
Arbustos medios-altos-caduca	0,172048703	0,9814899195
Arbustos bajos-altos-perenne	0,3835616438	0,9994786266
Arbustos alto y Arboles-caduca	0,6606593188	0,9552170352
Arbustos altos-Arboles-perenne	0,4536755111	0,9850626453
Coníferas y eucaliptos	0,7011460818	0,9669811927
Herbáceas bastas	0,4697286013	0,9848788822
Geófitos	0,5221912521	0,972729887
Juncales y turberas	0,1779661017	0,9966174394
TL-Herbáceas	0,8295724466	0,9983224517
H1-Herbáceas	0,6356460324	0,994852078
H2-Herbáceas	0,8803798624	0,9874312396
H3-Herbáceas	0,9068478128	0,9902208464
H4-Herbáceas	0,9491885539	0,9925387996

Tabla 4.3 En esta tabla se muestra la sensibilidad y la especificidad de cada clase de vegetación del primer conjunto de entrenamiento.

Hasta ahora tan sólo nos hemos preocupado por generar unos modelos que obtengan los mayores valores de *AUC* y que tengan una buena generalización, pero no se ha tenido en cuenta el hecho de que existen algunas clases de vegetación que no querríamos confundir en ningún caso. Por ejemplo, no tiene el mismo impacto confundir un árbol con otro tipo de árboles que confundirlo con la clase de agua.

Con el objeto de detectar cuáles son los errores más habituales que se están cometiendo, se va a mostrar una tabla en la que en las columnas se encuentra la clase de vegetación real y en las filas, la clase de vegetación predicha. Esta tabla sería una tabla de contingencia si no se hubiese dividido cada uno de los valores de la tabla entre el número de puntos de la clase de vegetación de la columna correspondiente, es decir, la suma de todos los elementos de cada columna es 100. Por lo tanto, si en la celda de la fila *i* y la columna *j* aparece un 10 significa que el 10% de los puntos de la especie *j* se han clasificado como la especie *i*.

La leyenda de las clases de vegetación empleada en esta tabla es la misma que la utilizada en

la subsección de Imágenes de satélite.

Clases de Vegetación	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	59	0	4	6	0	1	0	2	2	1	0	0	1	0	3	13	17	0	0	0
2	0	100	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
3	1	0	23	1	0	0	1	0	0	0	0	0	0	0	5	0	1	0	0	0
4	1	0	1	23	0	1	0	0	1	0	0	0	1	0	1	4	2	0	0	0
5	0	0	4	1	64	2	4	3	0	1	1	1	1	2	1	0	0	4	1	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	2	0	33	3	7	48	60	5	7	4	5	5	18	9	11	0	5	3	1	0
8	0	0	0	0	1	0	0	16	2	0	1	1	0	1	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	53	0	0	0	0	0	0	0	0	0	0	0
10	1	0	2	2	4	3	4	8	5	66	22	7	4	6	3	0	0	0	3	4
11	0	0	2	0	0	1	1	1	7	2	46	2	0	0	0	0	0	0	0	0
12	3	0	1	12	2	10	10	13	20	7	16	72	1	5	4	0	0	2	2	0
13	0	0	2	0	5	21	6	0	0	0	0	0	47	1	1	0	2	0	0	0
14	0	0	3	1	2	8	5	5	0	2	1	1	7	53	4	1	8	3	2	0
15	0	0	2	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0
16	1	0	0	4	0	0	0	0	0	0	0	0	0	0	1	81	0	0	0	0
17	3	0	4	6	0	2	1	1	0	0	0	0	3	1	3	0	64	0	0	0
18	23	0	14	22	5	2	4	12	3	4	2	4	8	8	8	0	0	87	0	0
19	4	0	4	13	5	0	2	14	1	6	3	4	6	7	19	0	0	0	90	0
20	1	0	0	6	3	0	1	18	0	6	3	2	2	6	11	0	0	0	0	95

Tabla 4.4 En esta tabla se muestran las confusiones entre las clases de vegetación. Se pueden ver 5 cuadrados dentro de la tabla, estos delimitan las 5 zonas donde se deberían concentrar los errores de las clases de vegetación: el gris representa las clases improductivas; el turquesa, las clases de arbusto; el rojo, las clases de árboles; el verde, algunas clases herbáceas; y, el amarillo, la clase de herbáceas con buen valor forrajero (evidentemente, no ha de haber confusión entre estas pues son conjuntos separables a partir del *NDVI*). Los números que aparecen en la tabla han sido redondeados.

Vamos a comentar los aspectos más relevantes de la Tabla 4.4: se ha logrado detectar con qué clases de herbáceas de buen valor forrajero (las de menor *NDVI*) se confunden las clases

improductivas, lo cual era uno de los motivos por los que se había dividido la clase de herbáceas de buen valor forrajero; algunas clases arbustivas se confunden a menudo con arbustos altos y árboles-caduca y con herbáceas bastas; las clases de árboles se confunden entre ellas, con herbáceas con altos valores en la banda del *NDVI* y con arbustos medios-altos-caduca; y, las herbáceas se confunden entre ellas y con arbustos medios-altos-caduca. En resumen, salvo por dos de las clases de arbustivo y algunas de las clases improductivas se está bastante satisfecho con los resultados obtenidos.

Se van a mostrar las predicciones realizadas en los lugares de los cuales se dispone de cartografía de vegetación. De este modo, podremos comparar a simple vista qué clases de vegetación están relativamente bien clasificadas y cuáles no son tan satisfactorias.

Con el objetivo de no confundir al lector con las 20 clases de vegetación se han realizado agrupaciones de algunas de las clases de vegetación. Más específicamente, se han agrupado todas las clases arbustivas en un único grupo; las clases arbóreas, en otro; y, todas las clases de herbáceas salvo las dos clases de herbáceas con buen valor forrajero que tomaban los valores más bajos en la banda del *NDVI*, en otro. En la Ilustración 4.5 se pueden ver los colores que se han asignado a cada una de las clases de vegetación. Como se puede observar, en la leyenda de la cartografía no se dispone de la subdivisión realizada de las herbáceas con buen valor forrajero, lo que se mostrará será la cartografía original.

Ilustración 4.5 En esta ilustración se muestran las leyendas que se emplearán en las ilustraciones posteriores.

Antes de mostrar las imágenes se ha de recordar que los resultados obtenidos tanto en los

conjuntos de validación como en los de entrenamiento son muy similares, casi no apreciándose diferencias entre uno y otro. Para entrenar los modelos con los que se han generado las predicciones que se observan en las imágenes que se mostrarán a continuación se seleccionaron la mitad de los puntos y se dejaron la otra mitad en el conjunto de validación, así que la mitad de las predicciones están realizadas sobre el conjunto de entrenamiento y la otra mitad sobre el conjunto de validación. Además, es posible que, en apariencia, los resultados sean mejores que lo expuesto en la Tabla 4.4 ya que al agrupar algunas clases de vegetación no es posible discernir entre ellas y, por lo tanto, los errores entre esas clases han sido obviados. Por lo tanto, no es de extrañar que parezca que la precisión global sea mayor que el 73% que se ha alcanzado en esta zona. Se ha de comentar que si en la ilustración de las predicciones aparecen algunos recintos en blanco, eso significa que en esos recintos no se disponía de imágenes de satélite.

Se disponen de unas 50 imágenes para comparar visualmente la cartografía frente a las predicciones realizadas. En esta sección se van a mostrar diez de ellas que van desde la Ilustración 4.6 a la Ilustración 4.15, el resto de ellas pueden encontrarse en el apéndice, desde la Ilustración 7.7 a la Ilustración 7.49.

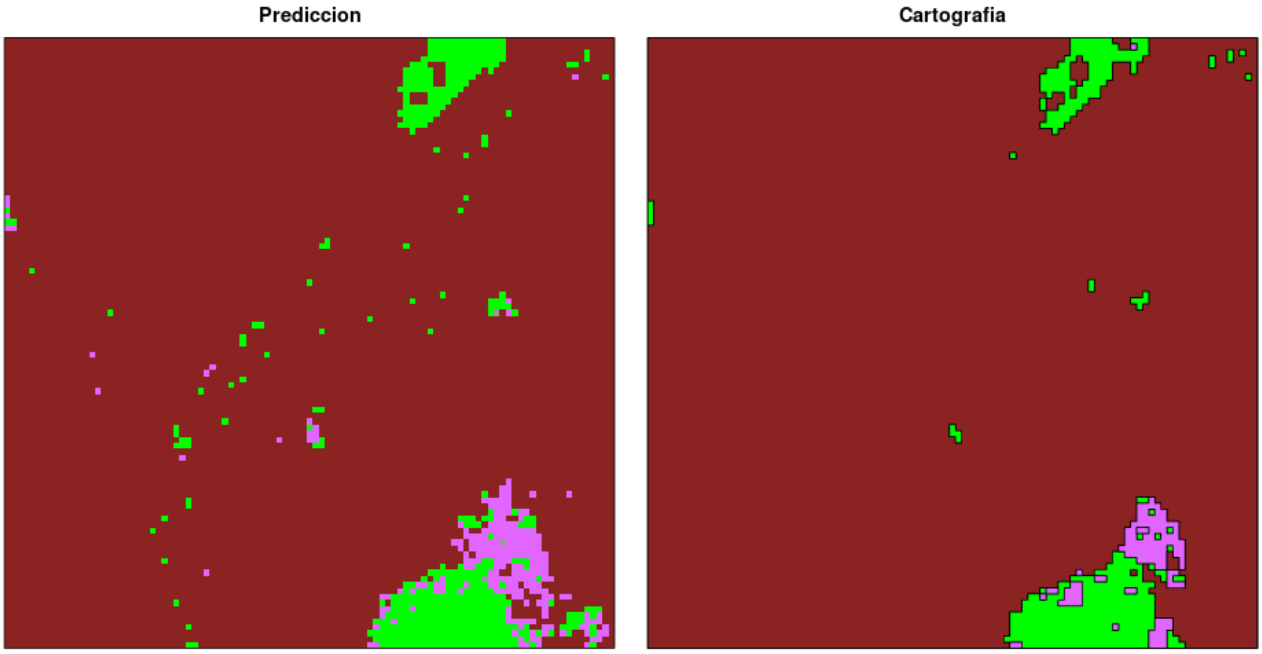

Ilustración 4.6 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

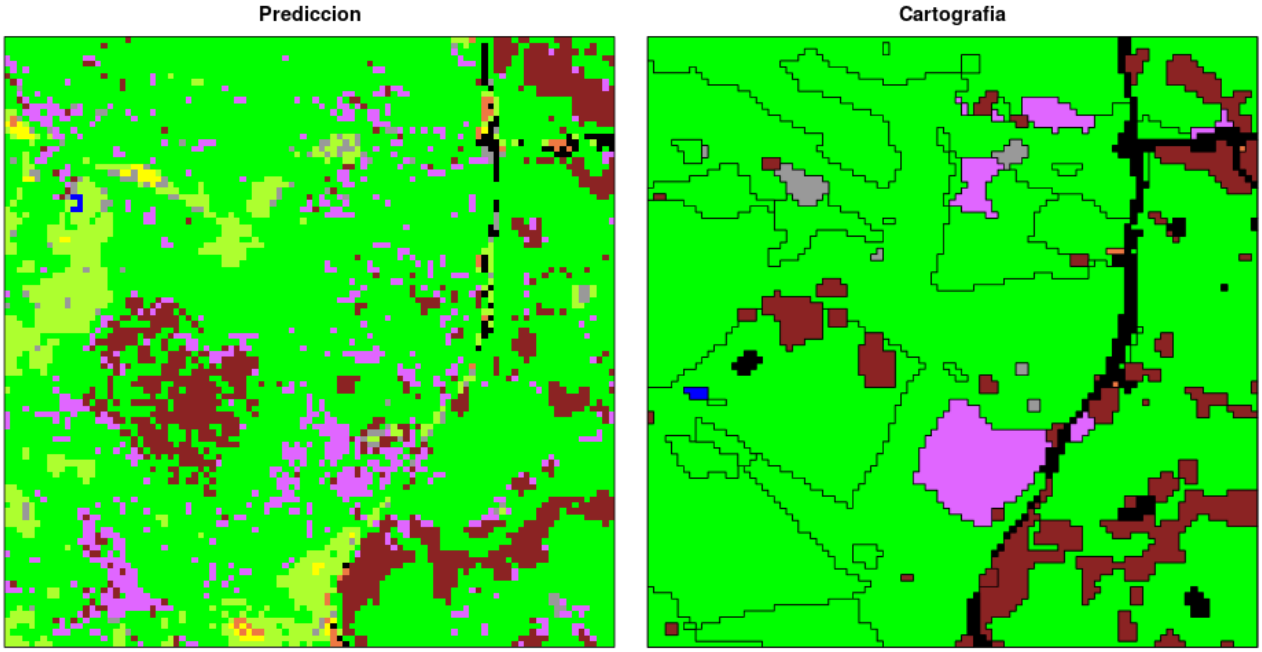

Ilustración 4.7 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

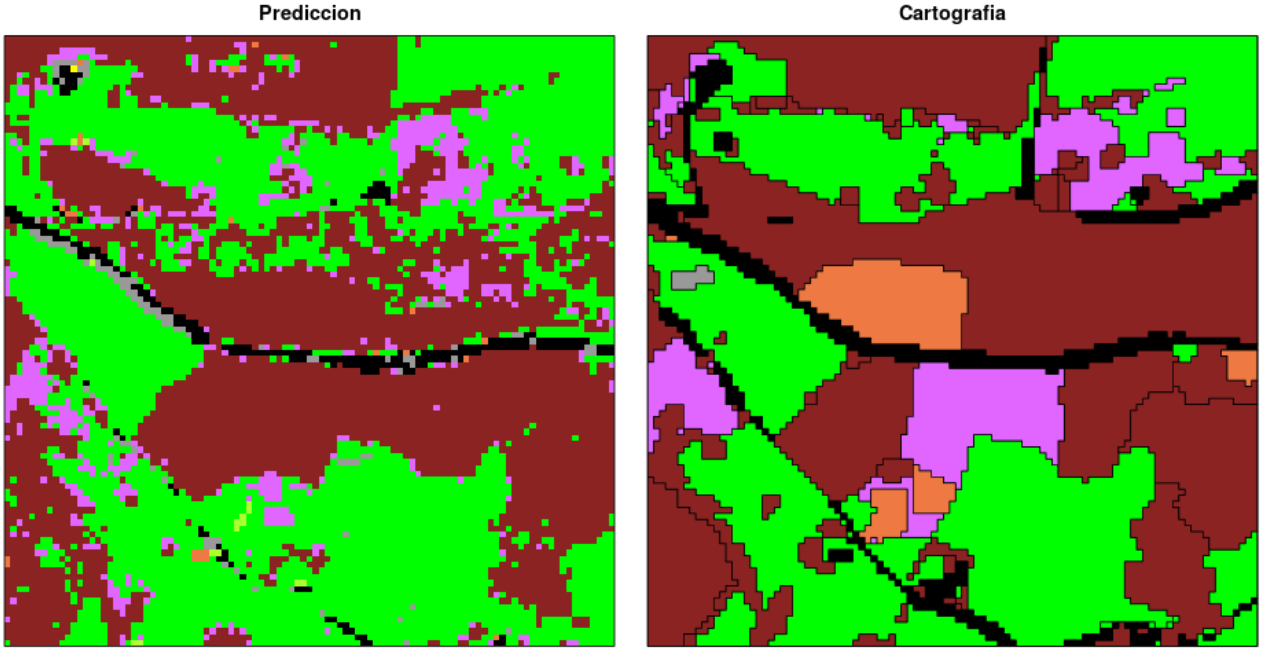

Ilustración 4.8 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

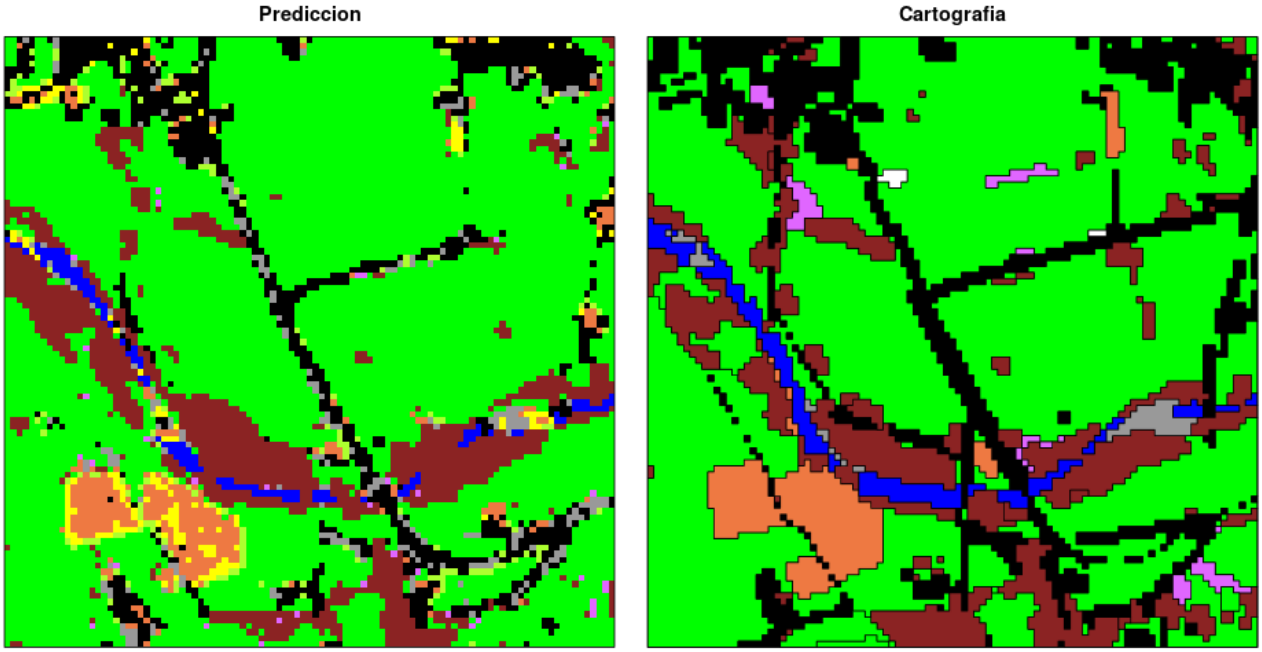

Ilustración 4.9 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

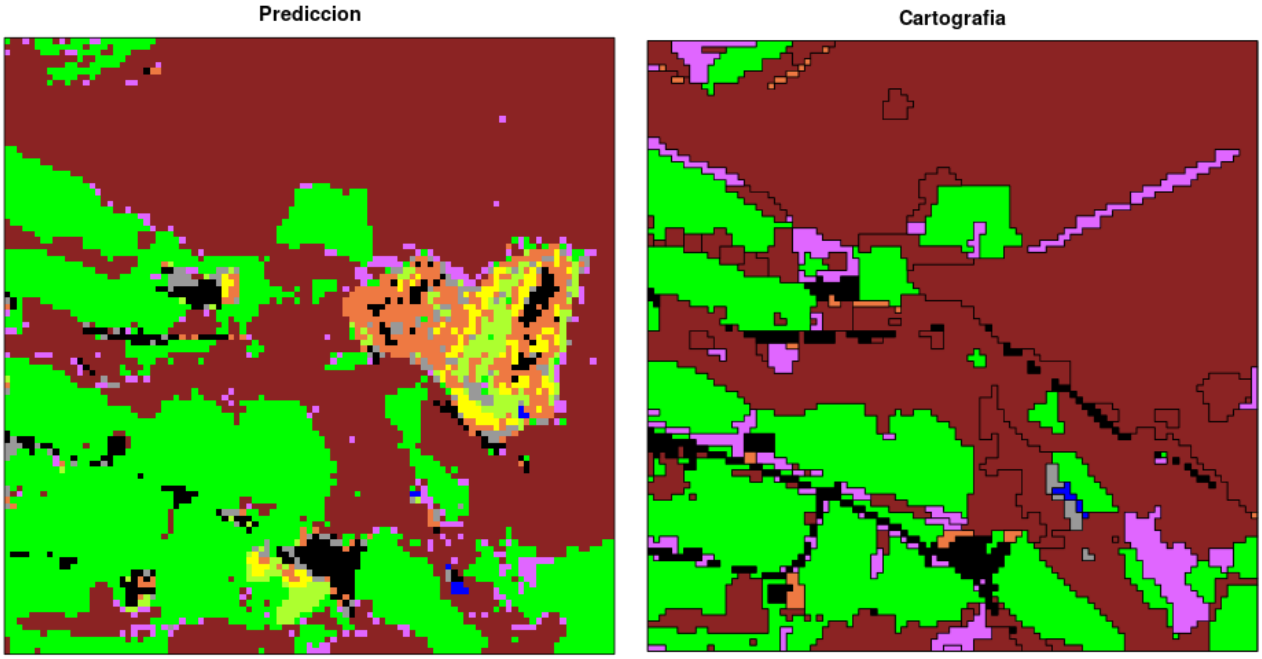

Ilustración 4.10 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

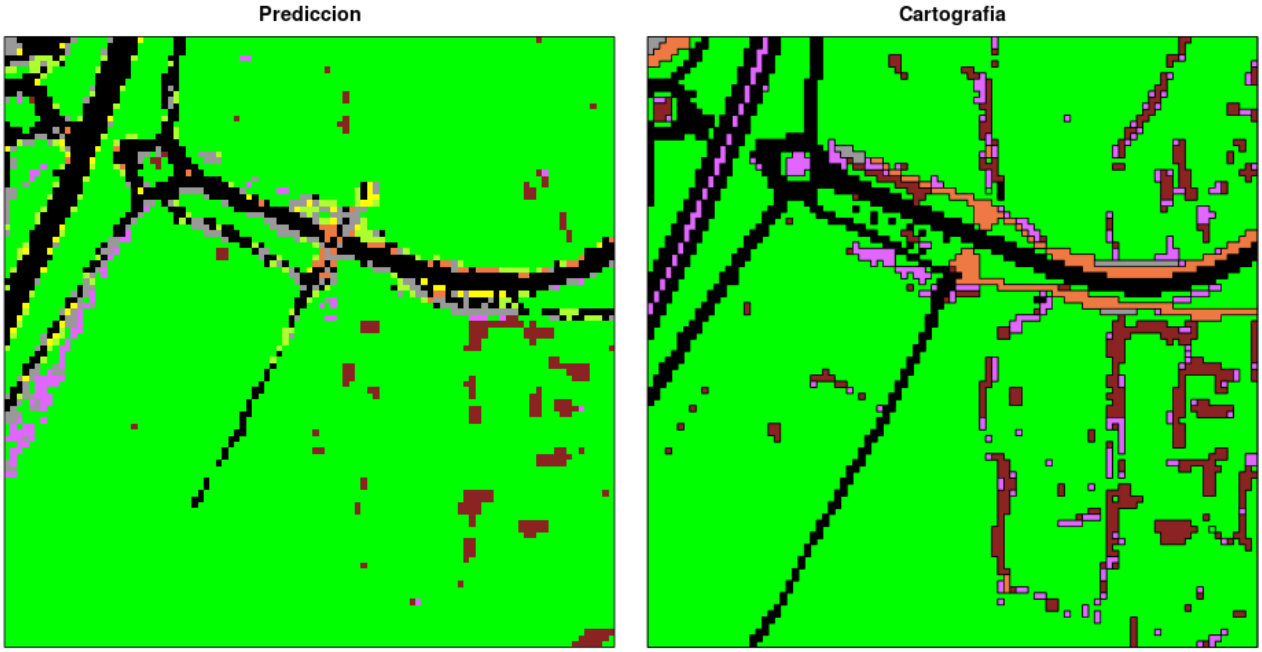

Ilustración 4.11 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

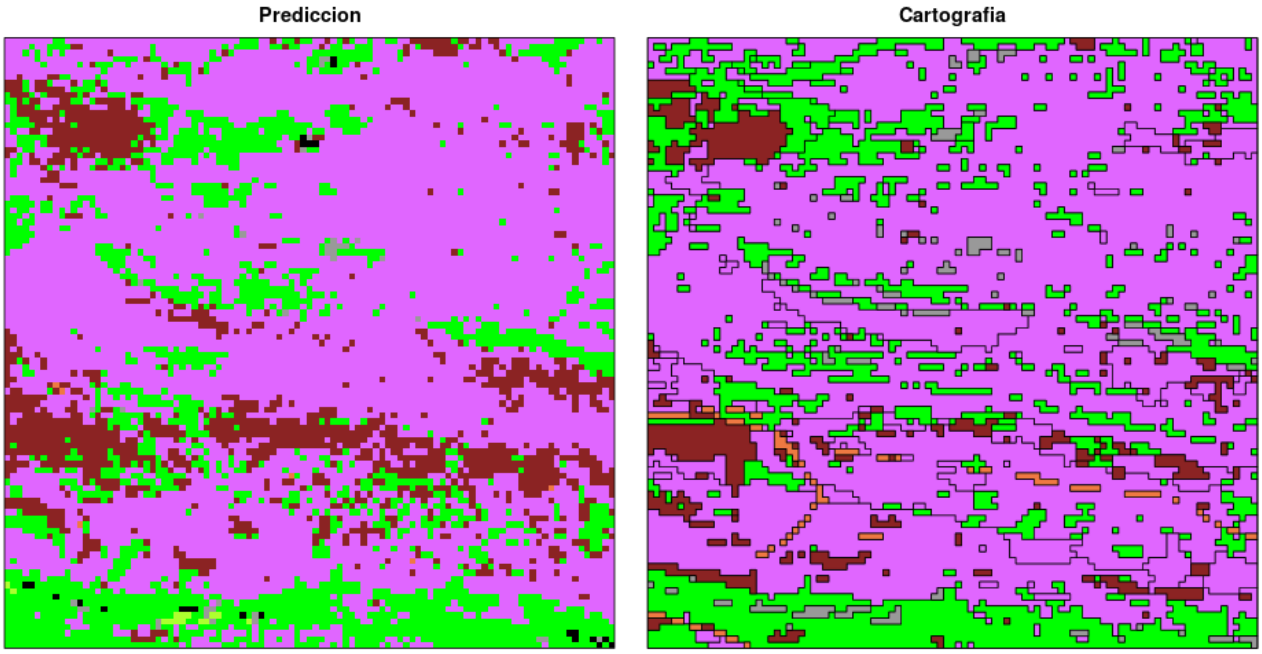

Ilustración 4.12 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

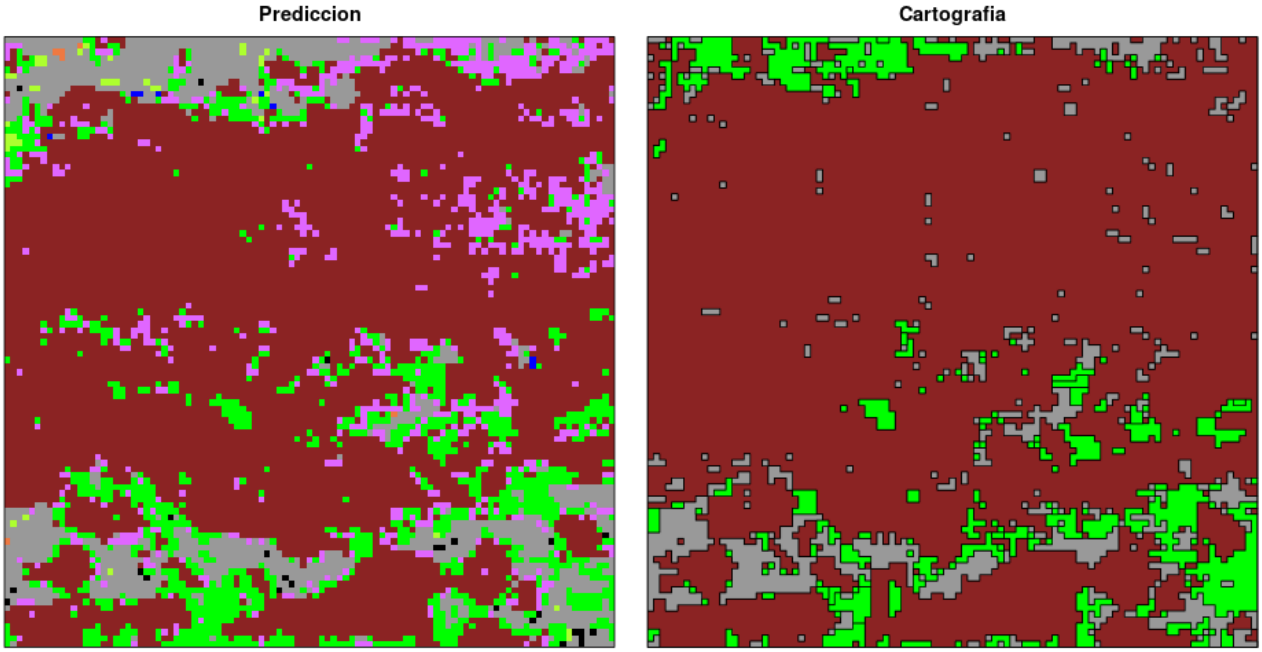

Ilustración 4.13 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

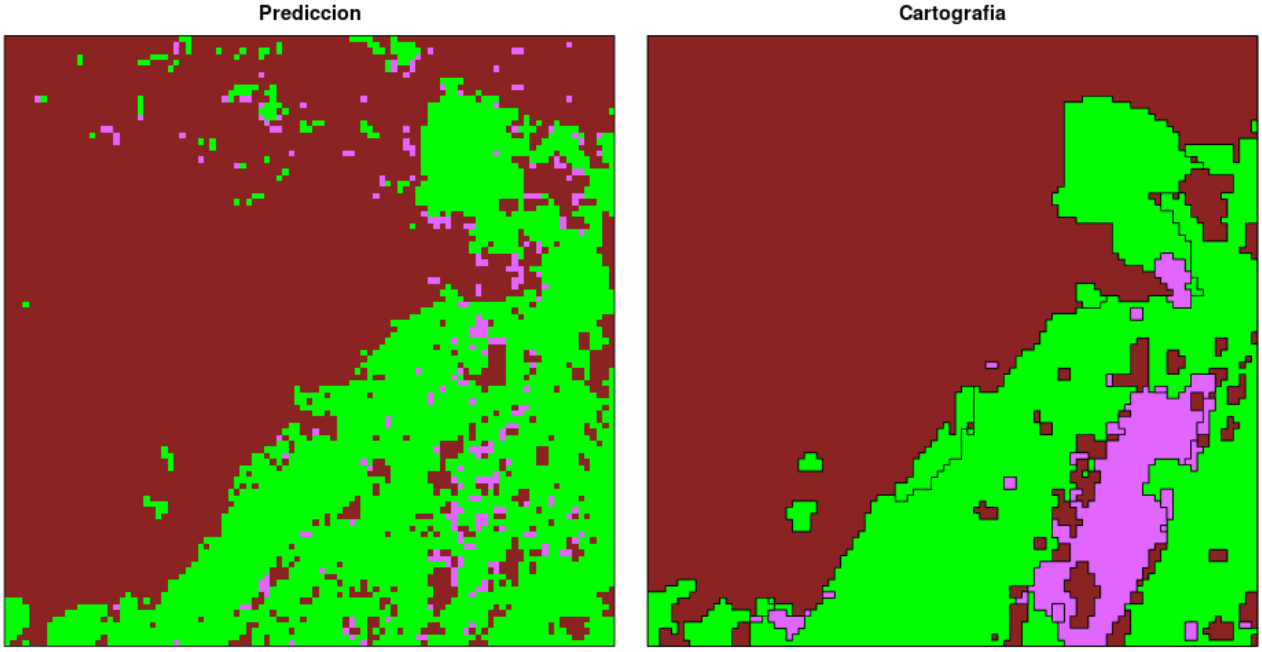

Ilustración 4.14 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

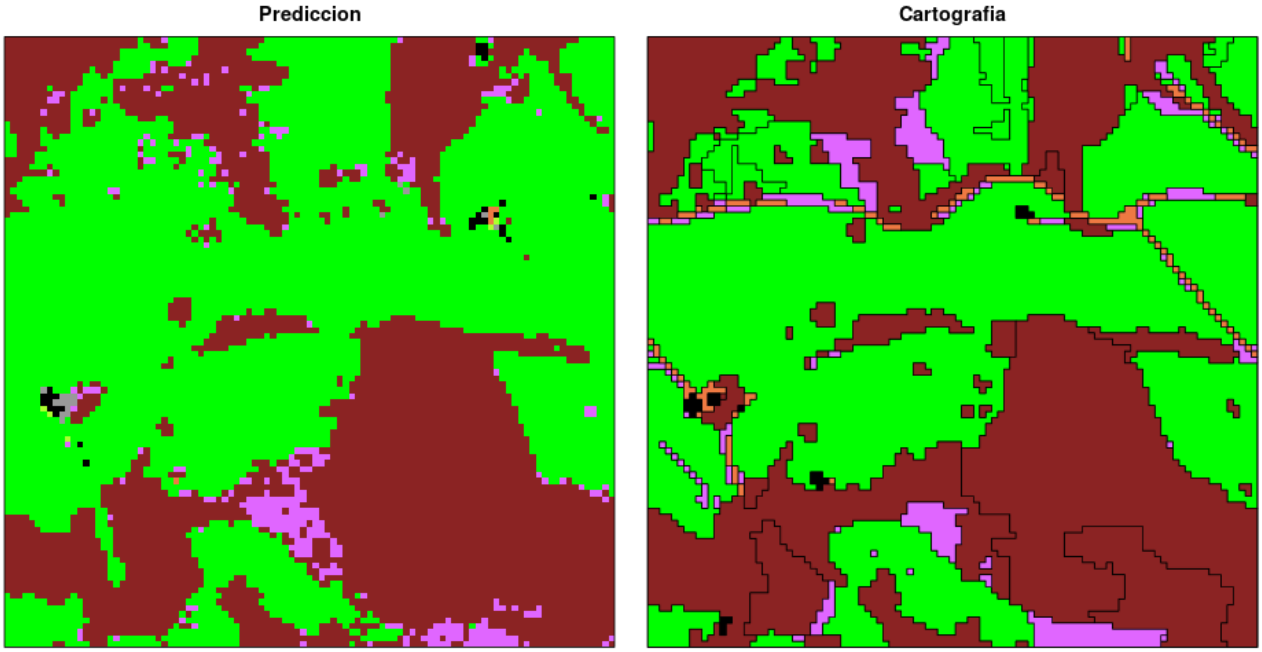

Ilustración 4.15 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

A la vista de estas imágenes, los resultados parecen satisfactorios salvo por la excepción de las clases arbustivas. A continuación, se muestra una imagen de la predicción realizada de una zona de Alto Campoo donde no se disponía de cartografía y la compararemos con la ortofoto de esa misma zona. Al sur de Cantabria el *NDVI* de las clases de herbáceas es, por lo general, más bajo que en la zona norte, por tanto, se puede considerar que los puntos amarillos de la imagen de la predicción son herbáceas, aunque de peor calidad. La leyenda es la misma que la empleada en las anteriores ilustraciones.

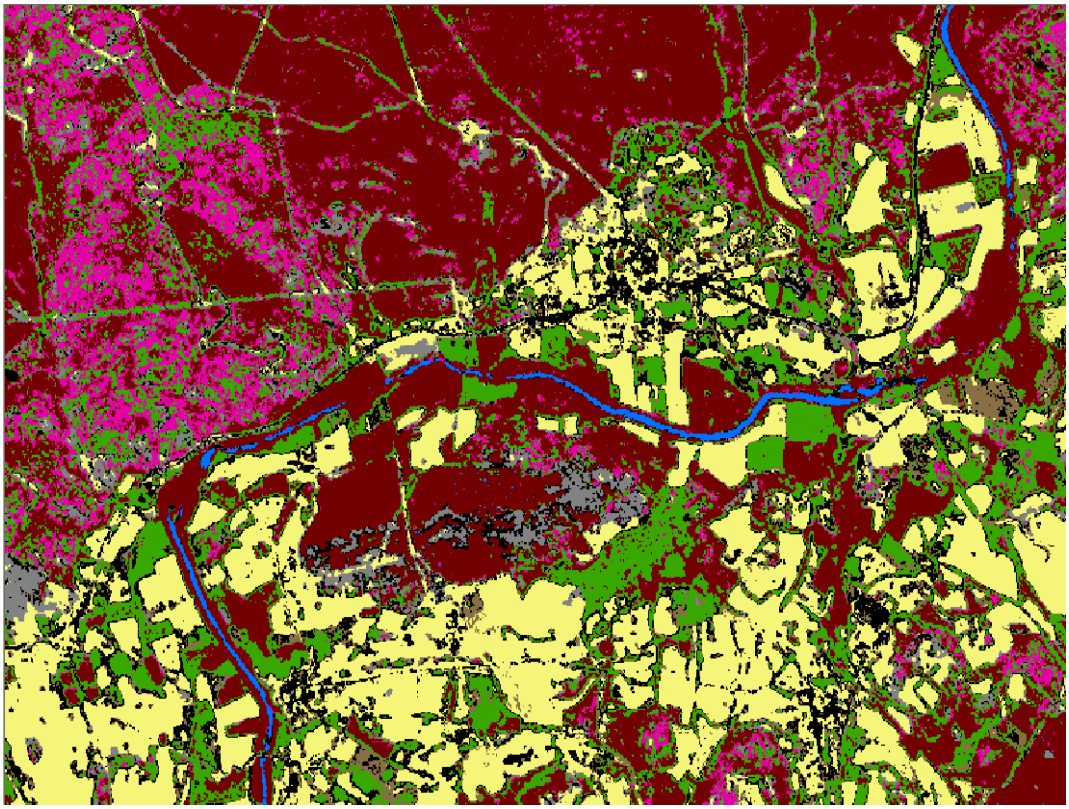

Ilustración 4.16 En esta ilustración se muestra la clasificación realizada en una zona no cartografiada.

Ilustración 4.17 En esta ilustración se muestra la ortofoto de la misma zona que la Ilustración 4.16

Por último y para cerrar este capítulo, mostraremos una imagen de la predicción realizada en la zona *B*. La leyenda de esta imagen es la misma que la que se ha empleado a lo largo del capítulo.

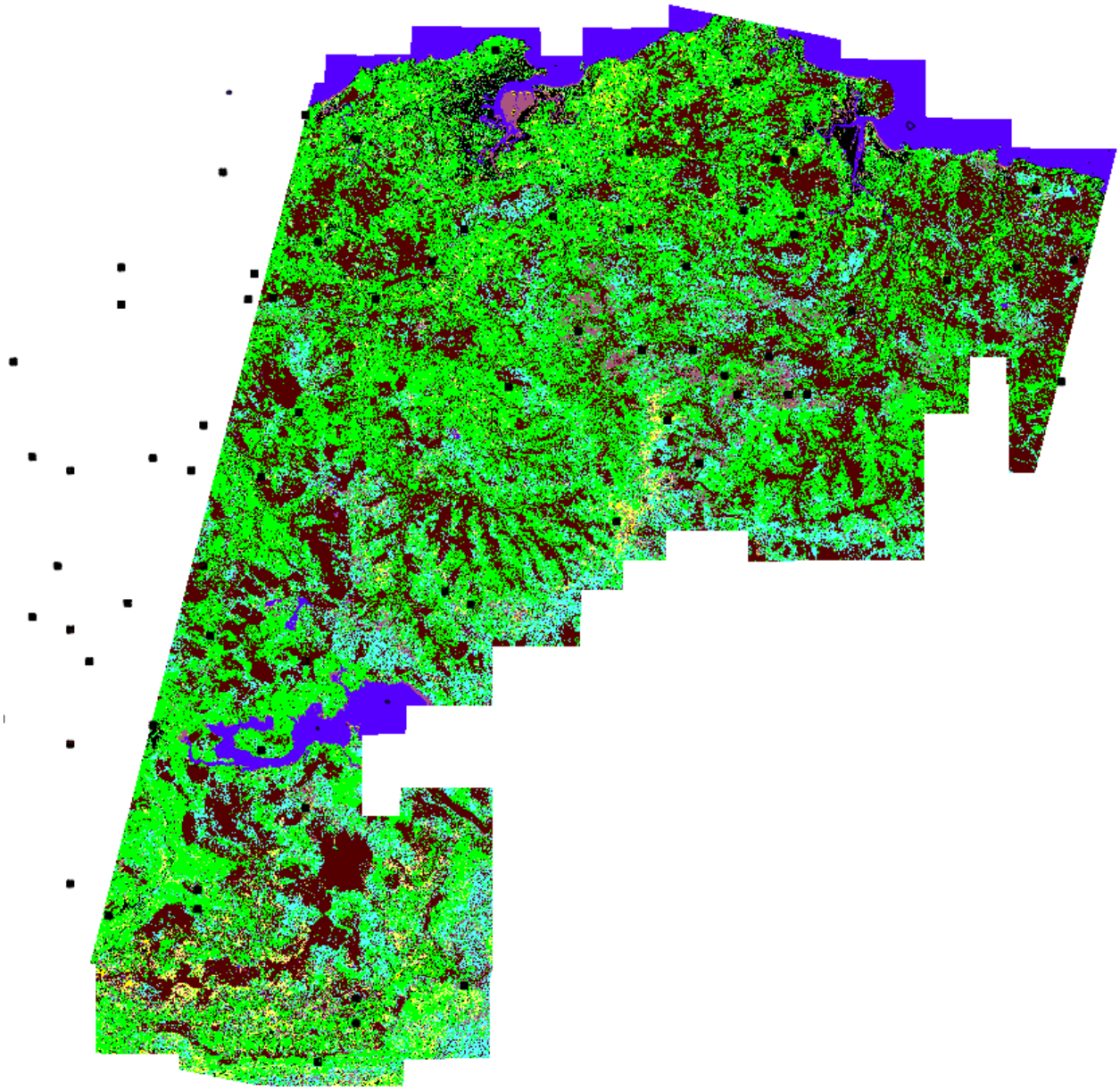

Ilustración 4.18 Imagen de la predicción realizada en la zona *B*. Los cuadrados que aparecen en color negro representan los cuadrados donde se disponía de cartografía en la zona.

5 Conclusiones y trabajo futuro

5.1 Conclusiones

La metodología desarrollada en este trabajo ha sido aplicada a las 8 zonas en las que se ha dividido Cantabria obteniendo una precisión que varía del 67% al 75% en los conjuntos de validación. No sólo son unos resultados satisfactorios sino que, además, el hecho de aplicar esta metodología en varias zonas de Cantabria diferentes y el que no haya sobreajuste a los datos en ninguna de estas zonas, nos lleva a pensar que la metodología desarrollada puede ser generalizada a otras regiones con climas y clases de vegetación distintas.

La metodología consiste en entrenar un clasificador binario probabilista *(MaxEnt)* siguiendo la estrategia *ONE vs ALL* para cada una de las clases de vegetación y, posteriormente, entrenar un árbol de decisión que utilice como entrada las salidas de *MaxEnt* y devuelva la clasificación definitiva de los puntos. Se ha comprobado que los árboles de decisión hacen un mayor uso de aquellos modelos cuya fiabilidad es mayor, es decir, que no sólo sirven para juntar toda la información procedente de *MaxEnt* sino que, además, realizan una selección de los mejores modelos.

5.2 Trabajo futuro

En los siguientes puntos se expone una serie de posibles trabajos futuros a seguir:

- Trabajar con nuevos datos como pueden ser datos que estimen la altitud de los píxeles como los datos *LIDAR*[6] o con más bandas del satélite y observar cómo la metodología descrita conlleva a una mejora de los resultados.
- Encontrar, al menos, otro método de clasificación tan competitivo para este problema como *MaxEnt* lo cual nos permitiría juntar la información procedente de ambos logrando unos resultados aún más satisfactorios.
- Ajustar la *Q* del método *MaxEnt*, que recordemos que proporciona la probabilidad de encontrar la clase en los distintos puntos. Se podría ajustar del siguiente modo: cuando esté presente la clase de vegetación como clase primaria y como clase secundaria la *Q* tomará un 1 en ese punto, cuando no esté ni como primaria ni como secundaria será un 0, cuando esté únicamente presente como primaria sea *p* y cuando sólo esté como secundaria sea un *1-p*. Evidentemente, se trataría de optimizar el *p*.
- En la construcción del árbol de decisión se podría penalizar aquellas confusiones que sean indeseables como por ejemplo confundir las clases de improductivo (rocas, urbano, etc.) con las clases de plantas (arbustos, árboles, herbáceas, etc.).

6 https://es.wikipedia.org/wiki/LIDAR

- Se podría tratar de disminuir la constante de regularización β en el término de penalización de complejidad de la función de ganancia de *MaxEnt* dado que no hay sobreajuste a los datos, lo cual permitiría a *MaxEnt* trabajar con más transformaciones de variables y, posiblemente, se obtendrían unos resultados más satisfactorios. Sin embargo, habría que asegurarse de que con esa nueva constante no se produzca sobreajuste.
- Se podría usar una medida de evaluación más compleja para evaluar la salida del árbol de decisión otorgándonos una mayor información sobre cómo de buena es la clasificación que se ha realizado.

6 Referencias

Bibliografía

1: Max Tuni. Análisis de técnicas de aprendizaje automático en el campo de la teledetección. 2011.
2: Instituto Geográfico español. (s.f.). Recuperado el 18 de junio de 2015, de http://www.ign.es/ign/layoutIn/teledeteccionQueEs.do. .
3: C.C.Petit y E.F. Lambin. Integration of multi-source remote sensing data for land cover change detection. 2001.
4: Agencia Estatal Europea. (s.f.). Recuperado el 18 de junio de 2015, de http://www.esa.int/esl/ESA_in_your_country/Spain/Cientificos_de_todo_el_mundo_repasan_los_avances_en_teledeteccion_en_Valencia/%28print%29. .
5: Miguel Ernesto Alva Huayaney y Juan Felipe Meléndez de la Cruz. Aplicación de la teledetección para el análisis multitemporal de la regresión glaciar en la Cordillera Blanca. 2009.
6: F. Gonzáles Alonso. Aplicaciones de la teledetección espacial al estudio de los incendios forestales. Detección de incendios en Galicia. 1993.
7: J. Delegido, L. Alonso, M. P. Cendrero, A. Forner y J. Moreno. Aplicación de la teledetección hiperspectral a la estimación del contenido en clorofila de las plantas. 2009.
8: S. Arenas, J. F. Haeger y D. Jordano. Aplicación de técnicas de teledetección y GIS sobre imágenes Quickbird para identificar y mapear individuos de peral silvestre (Pyrus bourgeana) en bosque esclerófilo mediterráneo. 1998.
9: J. Cabello, J.M. Paruelo. La teledetección en estudios ecológicos. 2008.
10: Gerard Moré, Xavier Pons, José Ángel Burriel, Rafael Castells, Joan Josep Ibáñez, Xavier Roijals. Generación de cartografía detallada de vegetación mediante procesamiento digital de imágenes de Landsat, variables orográficas y climáticas. 2005.
11: Peter Flach. Machine Learning. The Art and Science of Algorithms that Make Sense of Data. 2012.
12: Ian H. Witten, Eibe Frank, Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. 2001.
13: Tom Mitchell. Machine Learning. 1997.
14: Dengsheng Lu, Hongli Ge, Shizhen He, Aijun Xu, Guomo Zhou, Huaqiang Du. Pixel-based Minnaert Correction Method for Reducing Topographic Effects on a Landsat 7 ETM+ Image. 2008.
15: Johnson, L.F. y Trout, T.J.. Satellite NDVI assisted monitoring of vegetable crop evapotranspiration in California’s San Joaquin Valley. 2012.
16: Ludmila I. Kuncheva. Combining Patter Classifiers. Methods and Algorithms. 2004.
17: Brian M. Steele, David A. Patterson. Land cover Mapping Using Combination and Ensemble Classifiers. 2002.
18: Phillips, S. J.. A maximum entropy approach to species distribution modeling.. 2004.
19: Trevor Hastie, Will Fithian. Inference from presence-only data; the ongoing controversy. 2012.
20: Cory Merw, Matthew J. Smith y John A. Silander, Jr.. A practical guide to MaxEnt for modelling species' distributions: what it does, and why inputs and settings matter. 2013.
21: Jane Elith, Steven J. Phillips, Trevor Hastie, Miroslav Dudík, Yung En Chee, Colin J. Yates. A statistical explanation of MaxEnt for ecologists. 2011.
22: Steven J. Phillips y Miroslav Dudík. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. 2008.
23: Rokach, Lior y Maimon, O.. Data mining with decision trees: theory and applications. 2008.
24: Tom Fawcett. ROC Graphs: Notes and Practical Considerations for Researchers. 2003.
25: Payam Refaeilzadeh, Lei Tang y Huan Lui. K-fold cross-validation, Arizona State University. 2008.

7 Apéndice

En este apéndice se mostrarán ilustraciones y tablas que, debido a la magnitud de la experimentación llevada a cabo, no se han podido mostrar con anterioridad.

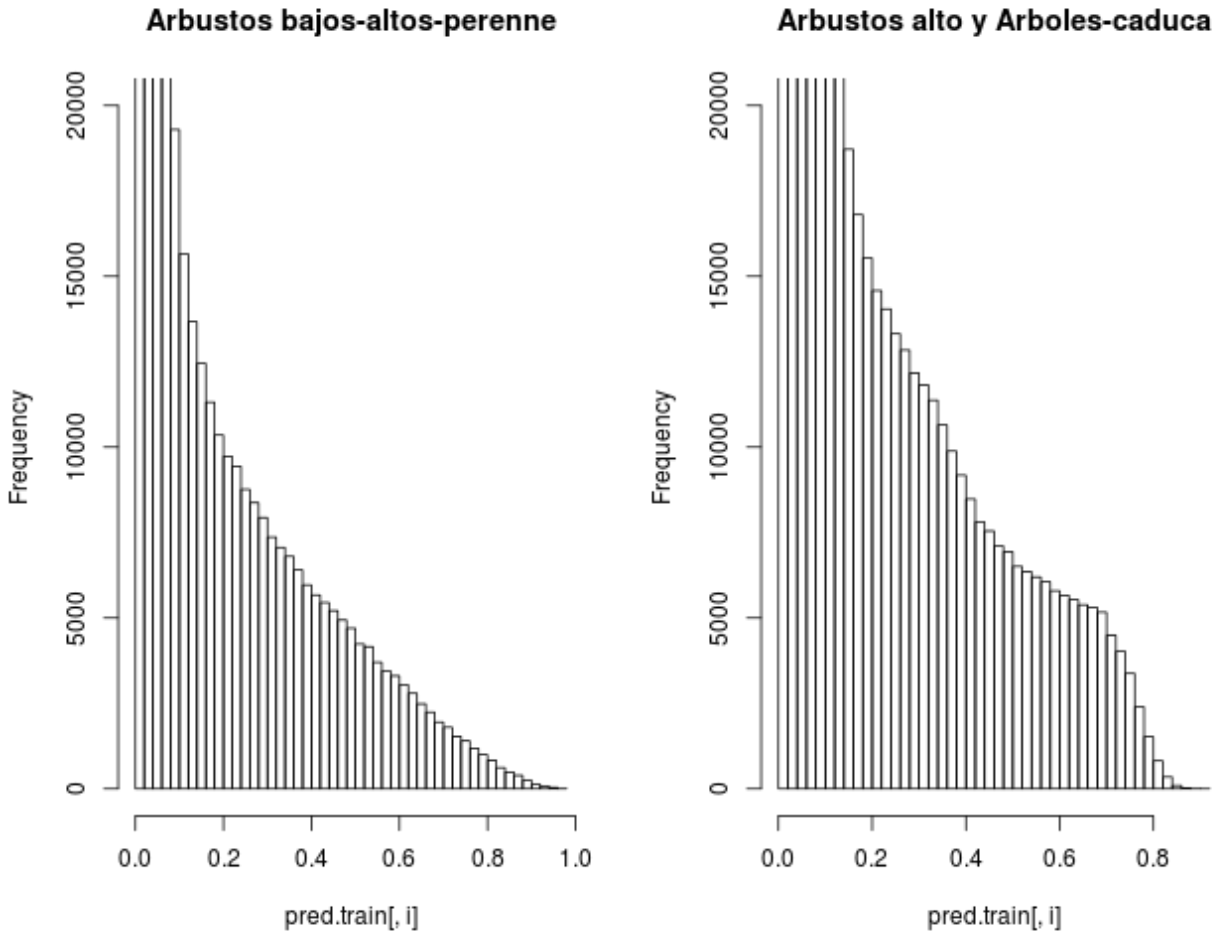

Ilustración 7.1 En esta ilustración se muestra en el eje Y el número de píxeles y en el eje X los valores de la predicción dada por el modelo de la clase arbustos bajos-altos-perenne, a la izquierda, y por la clase arbustos altos y árboles-caduca, a la derecha.

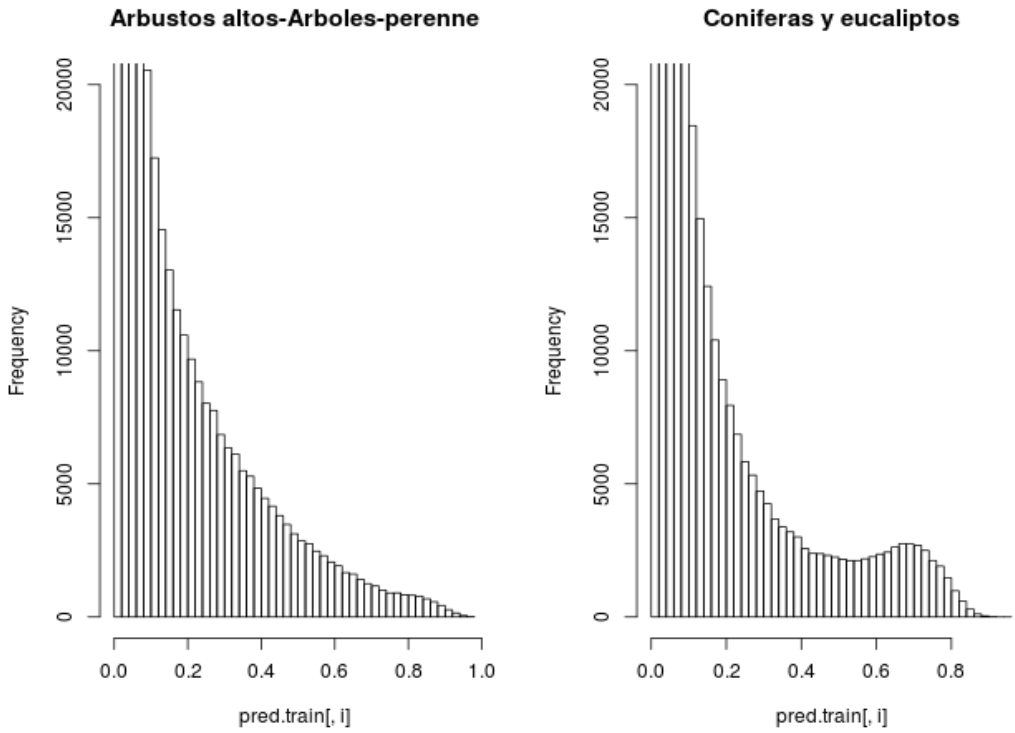

Ilustración 7.2 En esta ilustración se muestra en el eje Y el número de píxeles y en el eje X los valores de la predicción dada por el modelo de la clase arbustos altos-árboles-perenne, a la izquierda, y por la clase coníferas y eucaliptos, a la derecha.

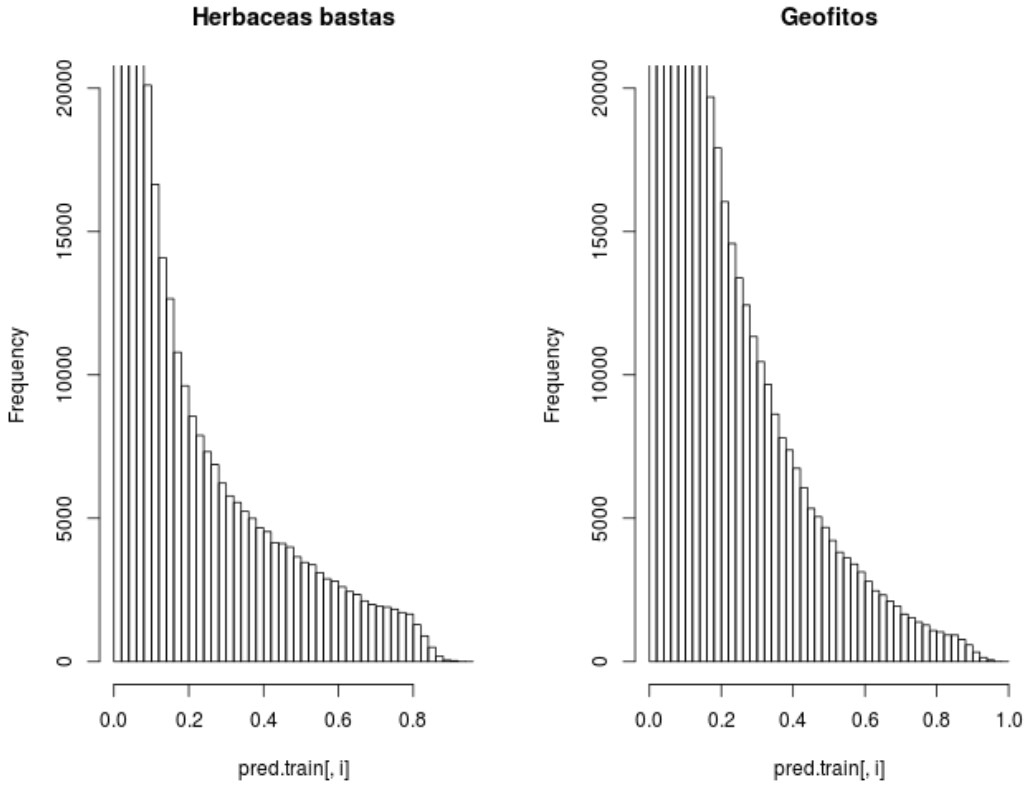

Ilustración 7.3 En esta ilustración se muestra en el eje Y el número de píxeles y en el eje X los valores de la predicción dada por el modelo de la clase herbáceas bastas, a la izquierda, y por la clase geófitos, a la derecha.

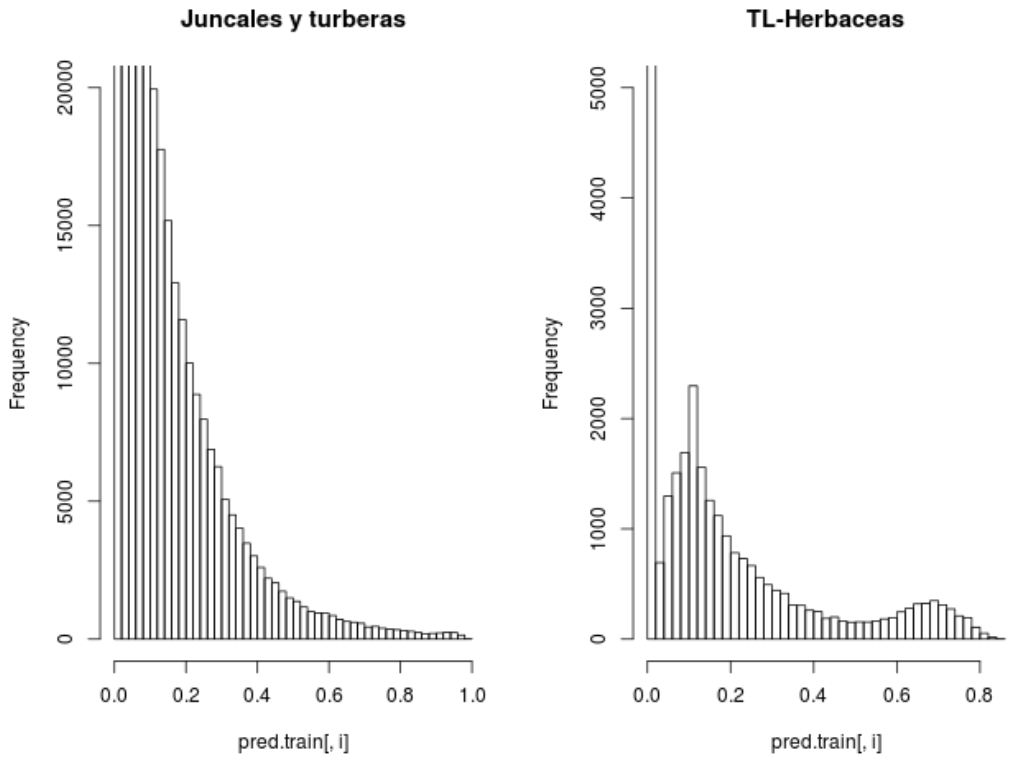

Ilustración 7.4 En esta ilustración se muestra en el eje Y el número de píxeles y en el eje X los valores de la predicción dada por el modelo de la clase juncales y turberas, a la izquierda, y por la clase TL-herbáceas, a la derecha.

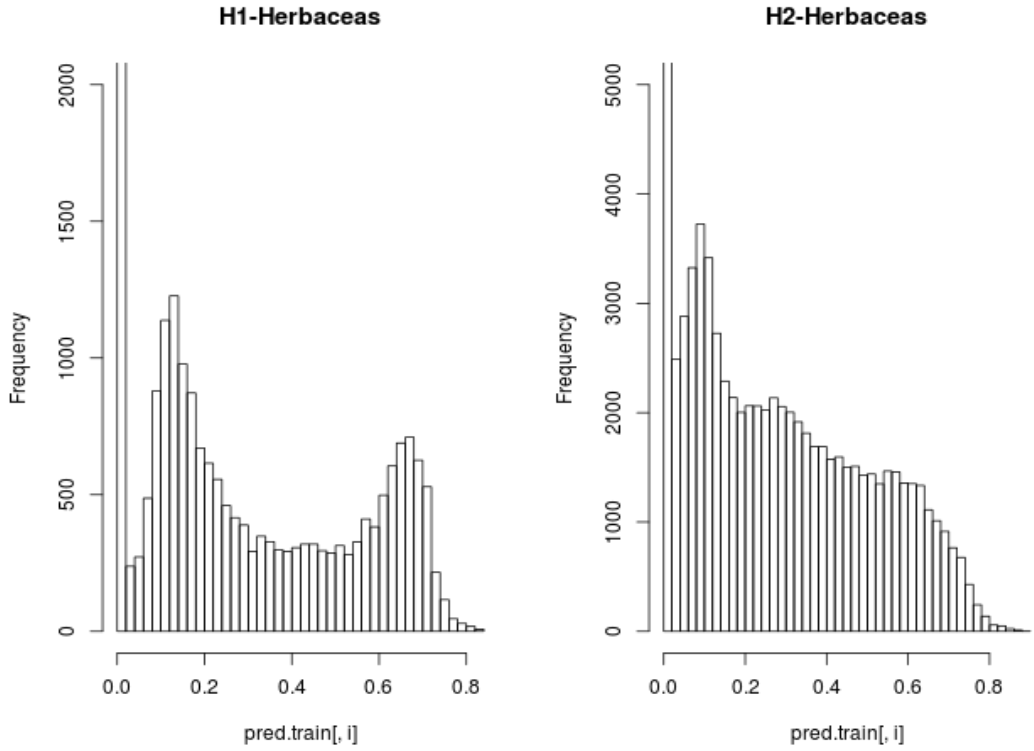

Ilustración 7.5 En esta ilustración se muestra en el eje Y el número de píxeles y en el eje X los valores de la predicción dada por el modelo de la clase H1-herbáceas, a la izquierda, y por la H2-herbáceas, a la derecha.

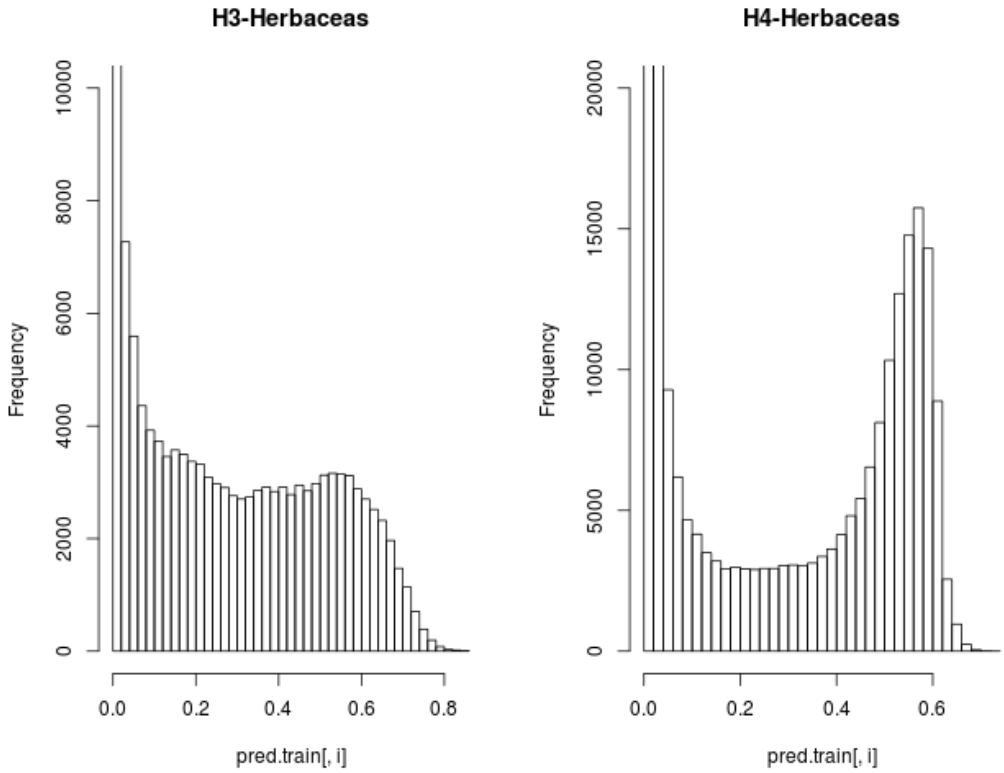

Ilustración 7.6 En esta ilustración se muestra en el eje Y el número de píxeles y en el eje X los valores de la predicción dada por el modelo de la clase H3-herbáceas, a la izquierda, y por la clase H4-herbáceas, a la derecha.

Clase de vegetación	Sensibilidad	Especificidad
Urbano	0,6287878788	0,9895013123
Agua	0,9945711183	0,9990872581
Roca-Piedras	0,2727272727	0,9936948298
Suelo desnudo	0,2086720867	0,9844911833
Arbustos bajos-medios	0,5818181818	0,9825256417
Arbustos medios-sin hojas	0	0,9983302896
Arbustos bajos-medios-sin hojas	0,5344036697	0,9659567332
Arbustos medios-altos-caduca	0,1255924171	0,9803462051
Arbustos bajos-altos-perenne	0,7142857143	0,999791482
Arbustos alto y Arboles-caduca	0,6207978485	0,9501355652
Arbustos altos-Arboles-perenne	0,4003795066	0,983074451
Coníferas y eucaliptos	0,7070193286	0,9665718763
Herbáceas bastas	0,4746450304	0,9861526946
Geófitos	0,4829268293	0,9708342505
Juncales y turberas	0,1780821918	0,996862581
TL-Herbáceas	0,7655502392	0,9974194228
H1-Herbáceas	0,6259259259	0,9946637079
H2-Herbáceas	0,8402266289	0,9838228545
H3-Herbáceas	0,8682042834	0,9861878453
H4-Herbáceas	0,945625511	0,9920601755

Tabla 7.1 En esta tabla se muestra la sensibilidad y la especificidad de cada clase de vegetación del segundo conjunto de validación.

Clase de vegetación	Sensibilidad	Especificidad
Urbano	0,6618426417	0,990116764
Agua	0,9959747916	0,9993318531
Roca-Piedras	0,3007676204	0,9941517396
Suelo desnudo	0,2394047234	0,986144099
Arbustos bajos-medios	0,6361297699	0,9846412779
Arbustos medios-sin hojas	0,072	0,9986551972
Arbustos bajos-medios-sin hojas	0,5708467674	0,9685799478
Arbustos medios-altos-caduca	0,1625264271	0,9812465233
Arbustos bajos-altos-perenne	0,7	0,9997566995
Arbustos alto y Arboles-caduca	0,6501564906	0,9538643972
Arbustos altos-Arboles-perenne	0,4363279546	0,9846295636
Coníferas y eucaliptos	0,7268065268	0,969872627
Herbáceas bastas	0,4935064935	0,9853675967
Geófitos	0,5228681342	0,9726757564
Juncales y turberas	0,2027972028	0,9966870288
TL-Herbáceas	0,7946859903	0,9980129741
H1-Herbáceas	0,6532999165	0,9951283075
H2-Herbáceas	0,8653541387	0,9858723459
H3-Herbáceas	0,900749406	0,989581268
H4-Herbáceas	0,9499005245	0,9926452041

Tabla 7.2 En esta tabla se muestra la sensibilidad y la especificidad de cada clase de vegetación del segundo conjunto de entrenamiento.

Clase de vegetación	Sensibilidad	Especificidad
Urbano	0,5843920145	0,9877185455
Agua	0,9952082565	0,9992113565
Roca-Piedras	0,1569767442	0,9923784494
Suelo desnudo	0,2072368421	0,9872439528
Arbustos bajos-medios	0,588761175	0,9825133051
Arbustos medios-sin hojas	0	0,9982257475
Arbustos bajos-medios-sin hojas	0,5873749038	0,9700525198
Arbustos medios-altos-caduca	0,1094420601	0,9778442155
Arbustos bajos-altos-perenne	0,5294117647	0,9995828989
Arbustos alto y Arboles-caduca	0,6375227687	0,9531792248
Arbustos altos-Arboles-perenne	0,3606237817	0,9824448726
Coníferas y eucaliptos	0,6827268068	0,9641076192
Herbáceas bastas	0,435546875	0,9845330479
Geófitos	0,4629981025	0,9688033952
Juncales y turberas	0,1888888889	0,9961794107
TL-Herbáceas	0,7889447236	0,997789241
H1-Herbáceas	0,5735294118	0,9938705416
H2-Herbáceas	0,8614084507	0,9858799219
H3-Herbáceas	0,8930155211	0,9889035819
H4-Herbáceas	0,9562600321	0,9934750075

Tabla 7.3 En esta tabla se muestra la sensibilidad y la especificidad de cada clase de vegetación del tercer conjunto de validación.

Clase de vegetación	Sensibilidad	Especificidad
Urbano	0,6190866271	0,9889208556
Agua	0,9939947251	0,9990008169
Roca-Piedras	0,2124824684	0,9934457803
Suelo desnudo	0,219581749	0,9854784506
Arbustos bajos-medios	0,6258591065	0,9842383371
Arbustos medios-sin hojas	0,0201612903	0,9985914513
Arbustos bajos-medios-sin hojas	0,6109368376	0,9714917594
Arbustos medios-altos-caduca	0,1360962567	0,9808845976
Arbustos bajos-altos-perenne	0,4671532847	0,9995771278
Arbustos alto y Arboles-caduca	0,6711961912	0,9565536494
Arbustos altos-Arboles-perenne	0,4239130435	0,9842417611
Coníferas y eucaliptos	0,7077147016	0,967735923
Herbáceas bastas	0,4819950331	0,9850953351
Geófitos	0,5191338836	0,9725521149
Juncales y turberas	0,1919770774	0,9967222261
TL-Herbáceas	0,8157262905	0,9982057276
H1-Herbáceas	0,5982441472	0,994359468
H2-Herbáceas	0,873993657	0,986787747
H3-Herbáceas	0,8963481436	0,9891067956
H4-Herbáceas	0,9529678296	0,9931119605

Tabla 7.4 En esta tabla se muestra la sensibilidad y la especificidad de cada clase de vegetación del tercer conjunto de entrenamiento.

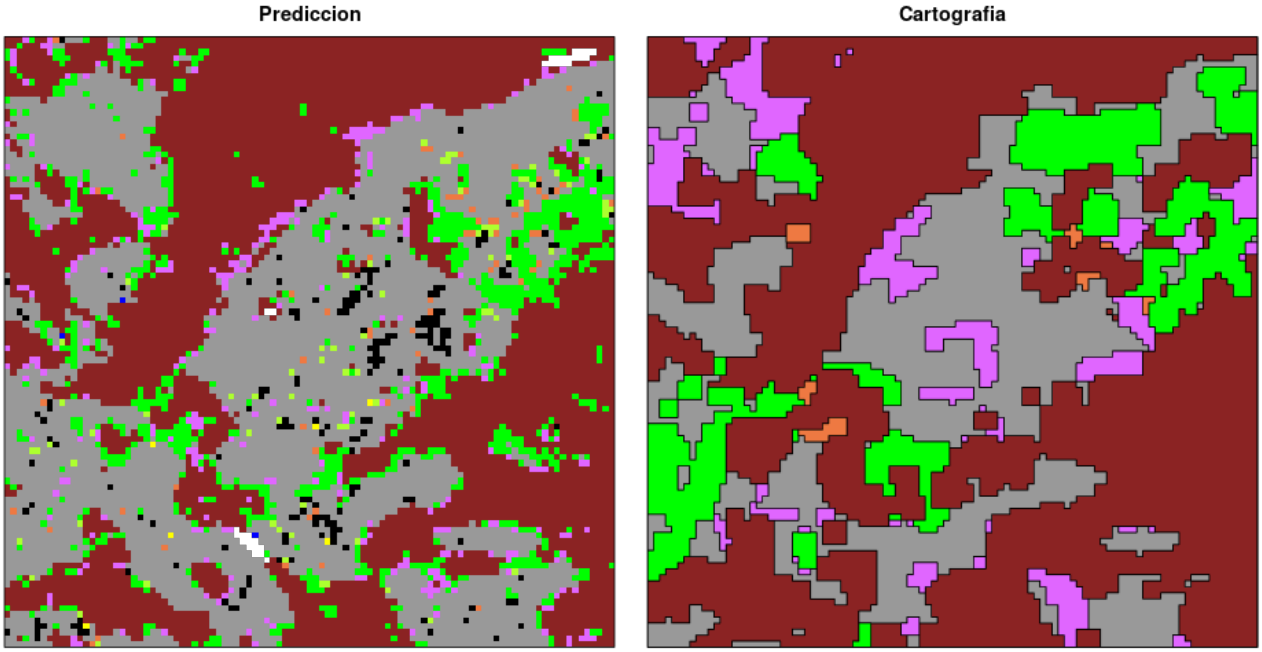

Ilustración 7.7 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

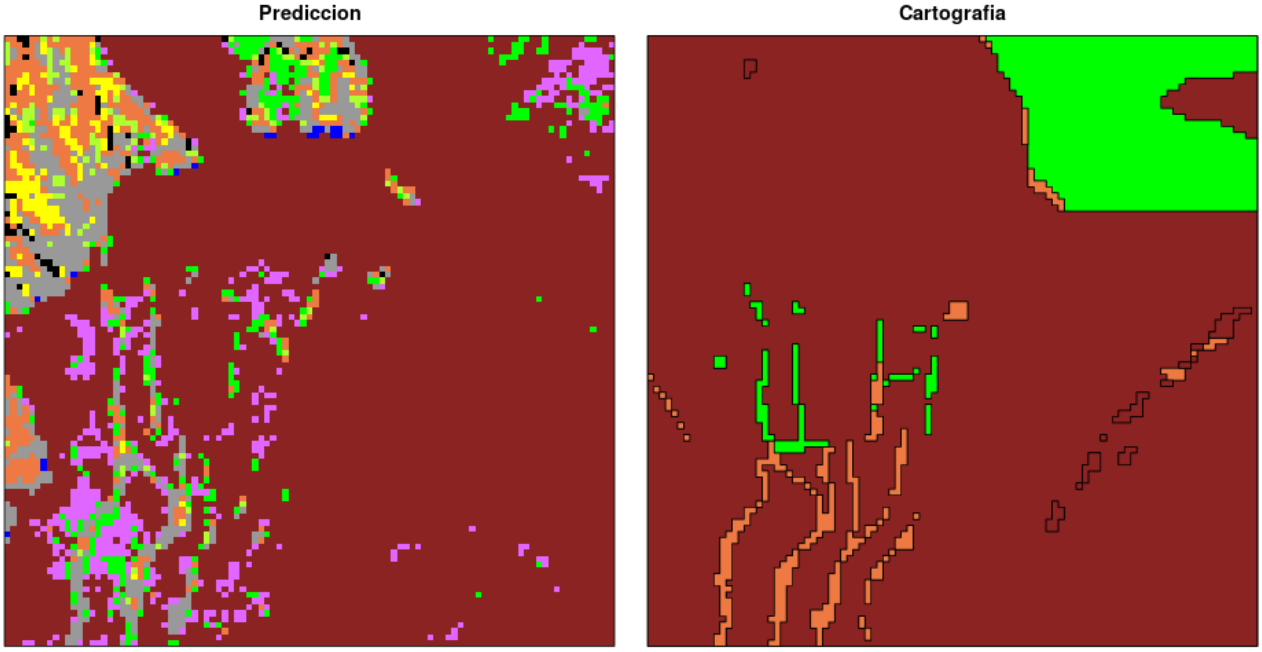

Ilustración 7.8 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

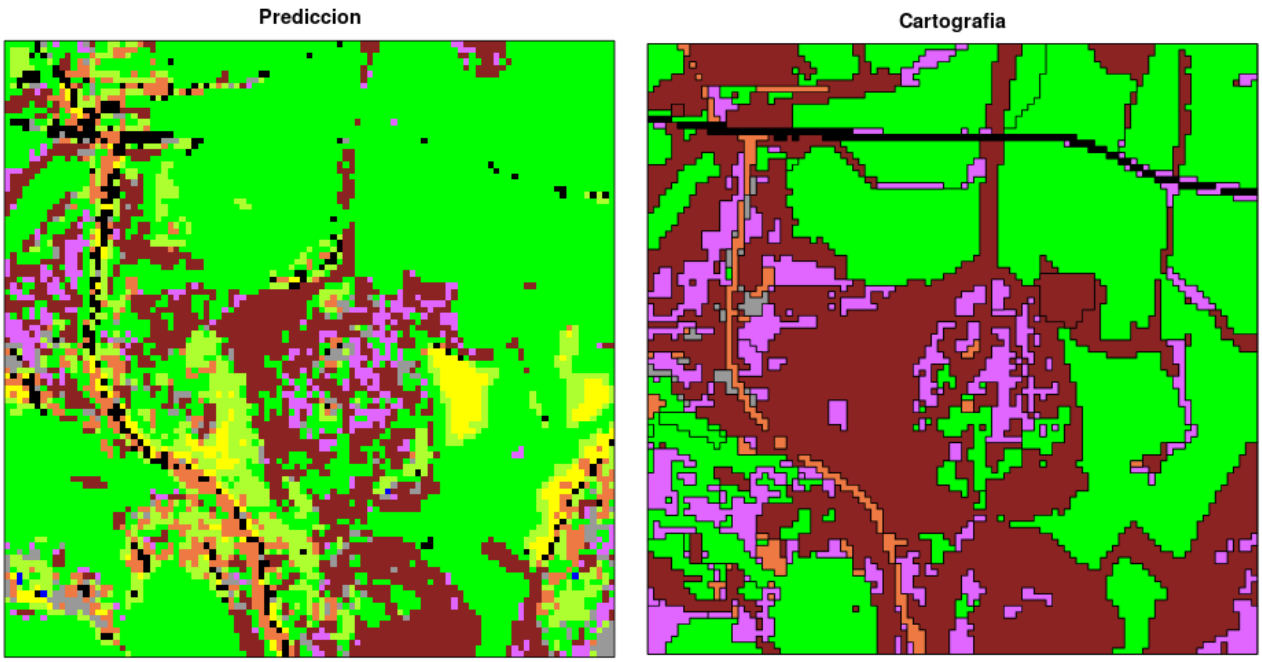

Ilustración 7.9 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

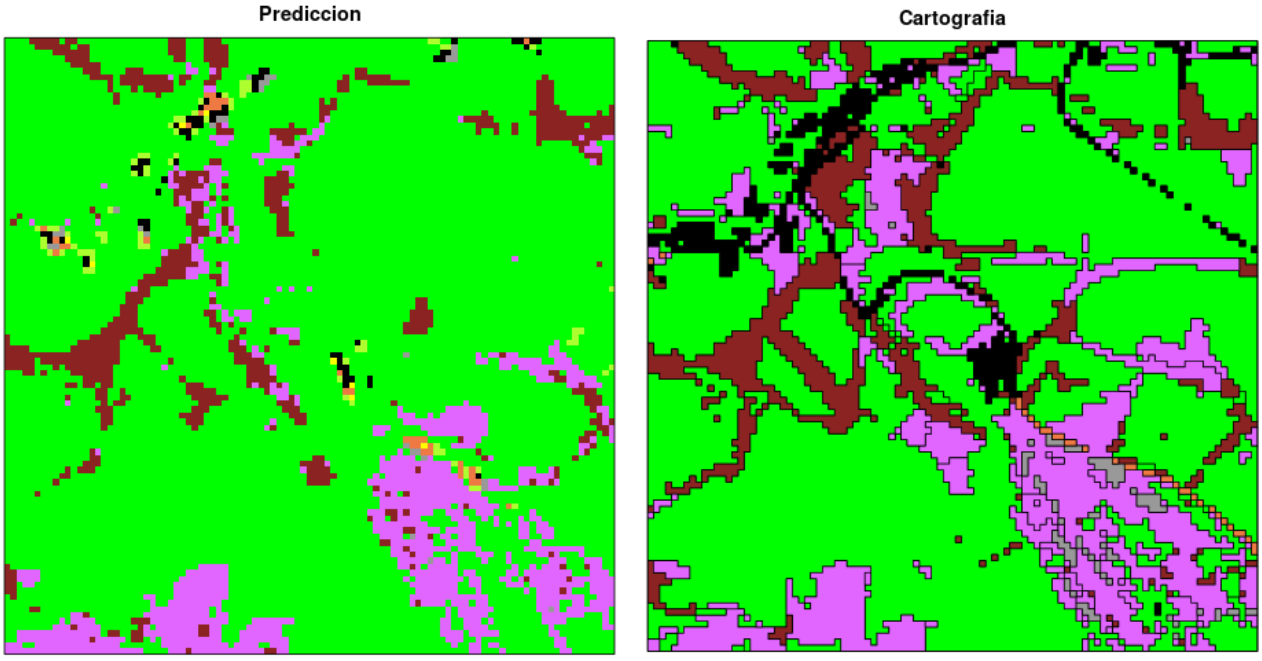

Ilustración 7.10 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

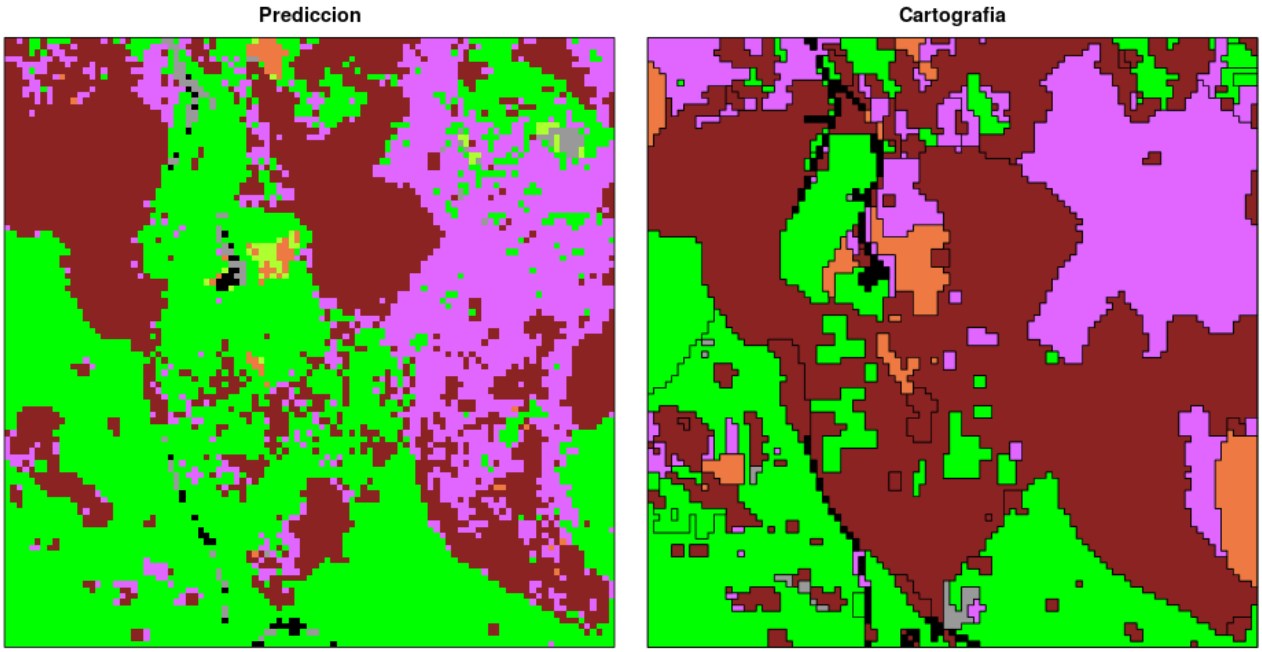

Ilustración 7.11 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

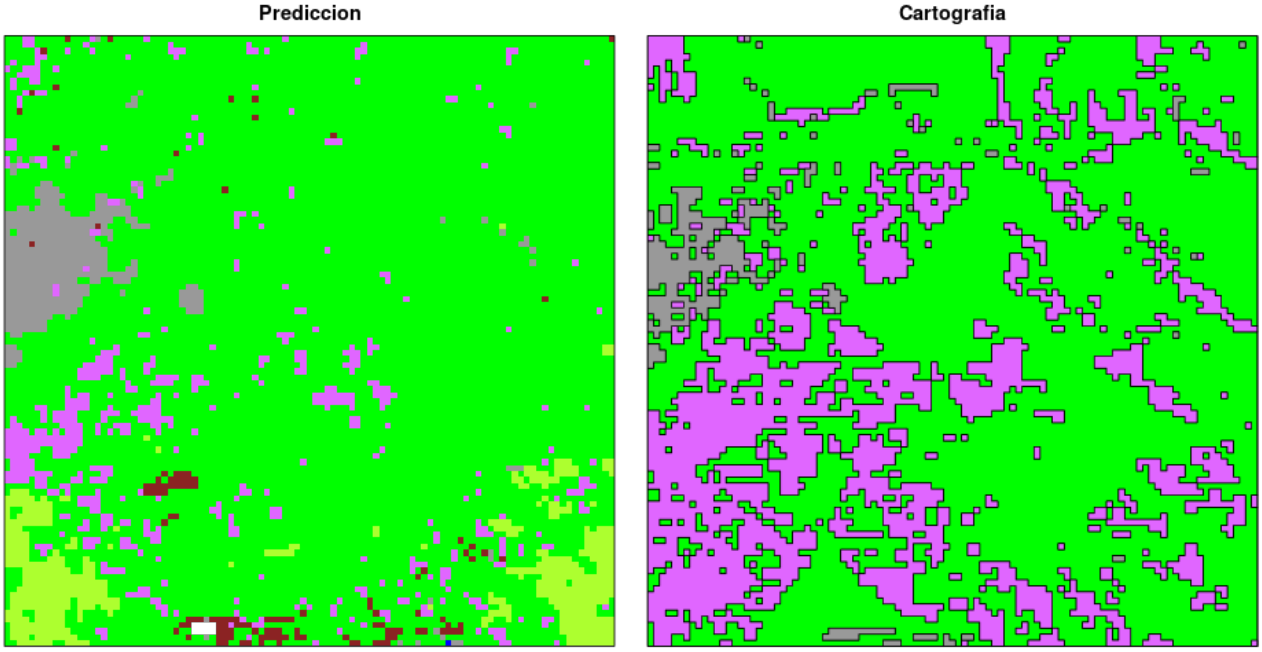

Ilustración 7.12 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

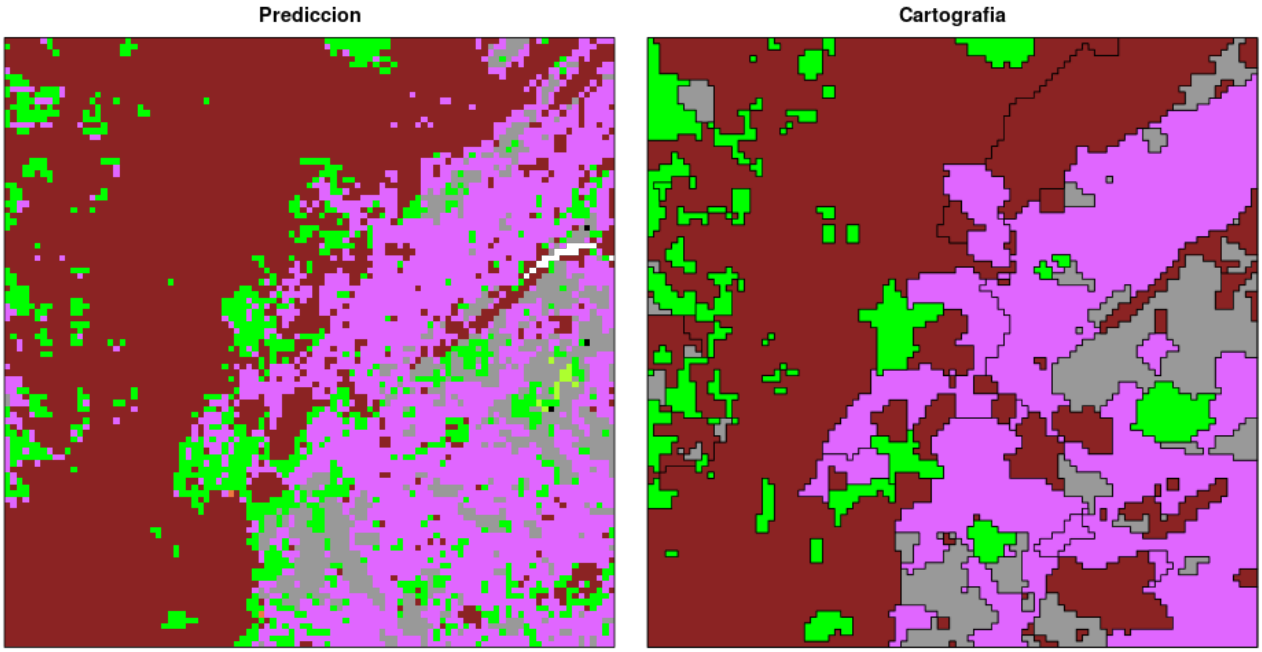

Ilustración 7.13 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

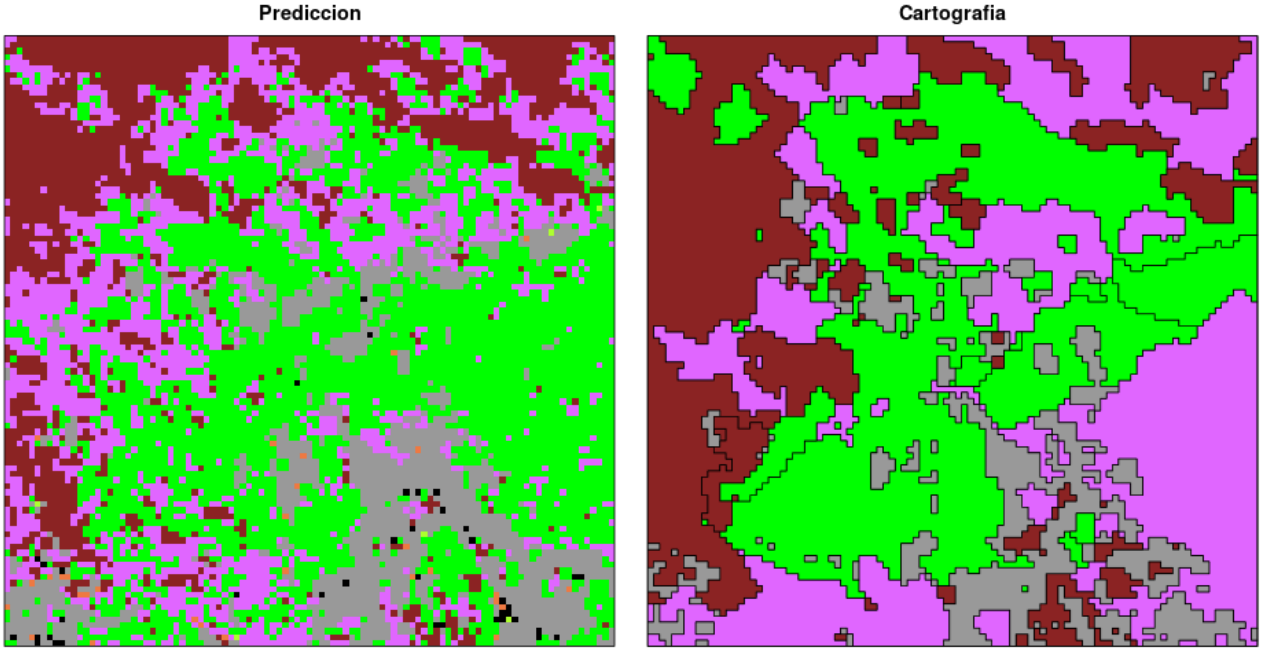

Ilustración 7.14 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

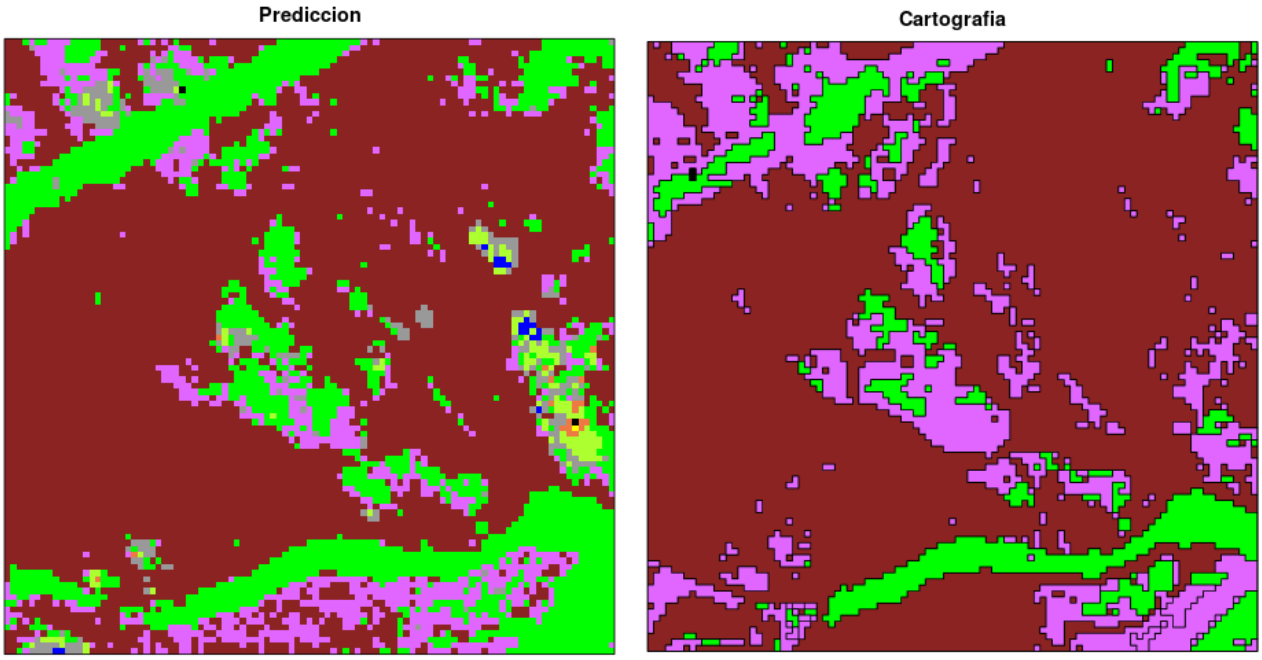

Ilustración 7.15 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

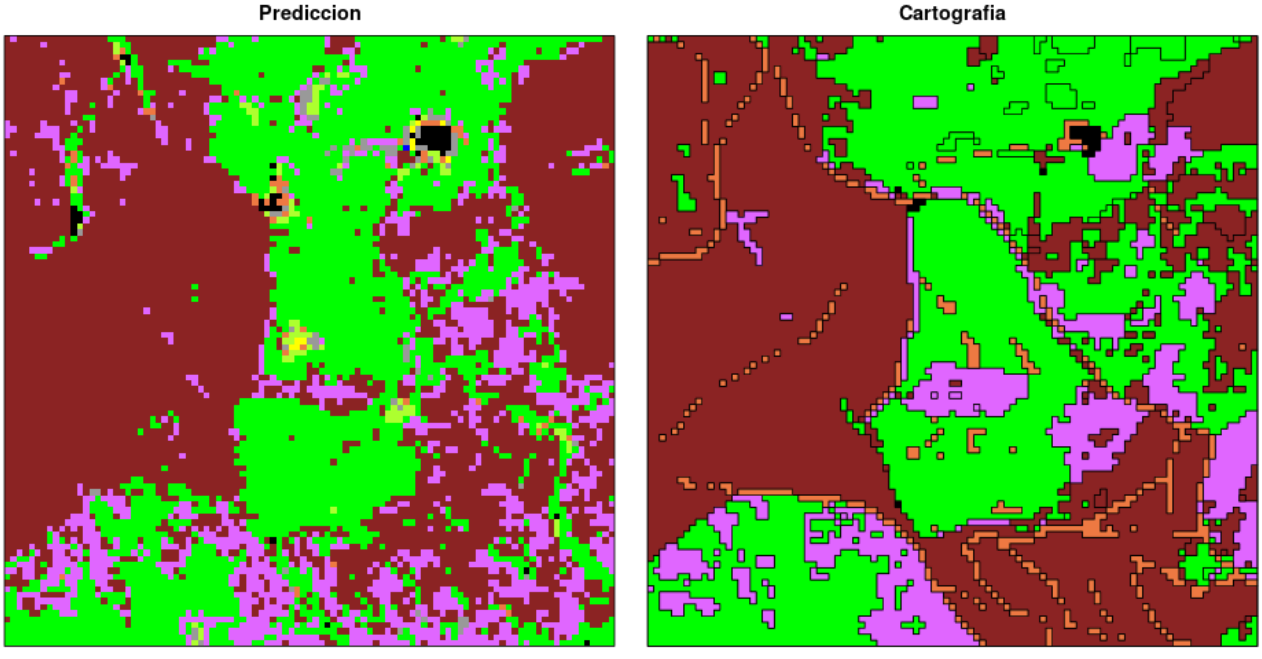

Ilustración 7.16 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

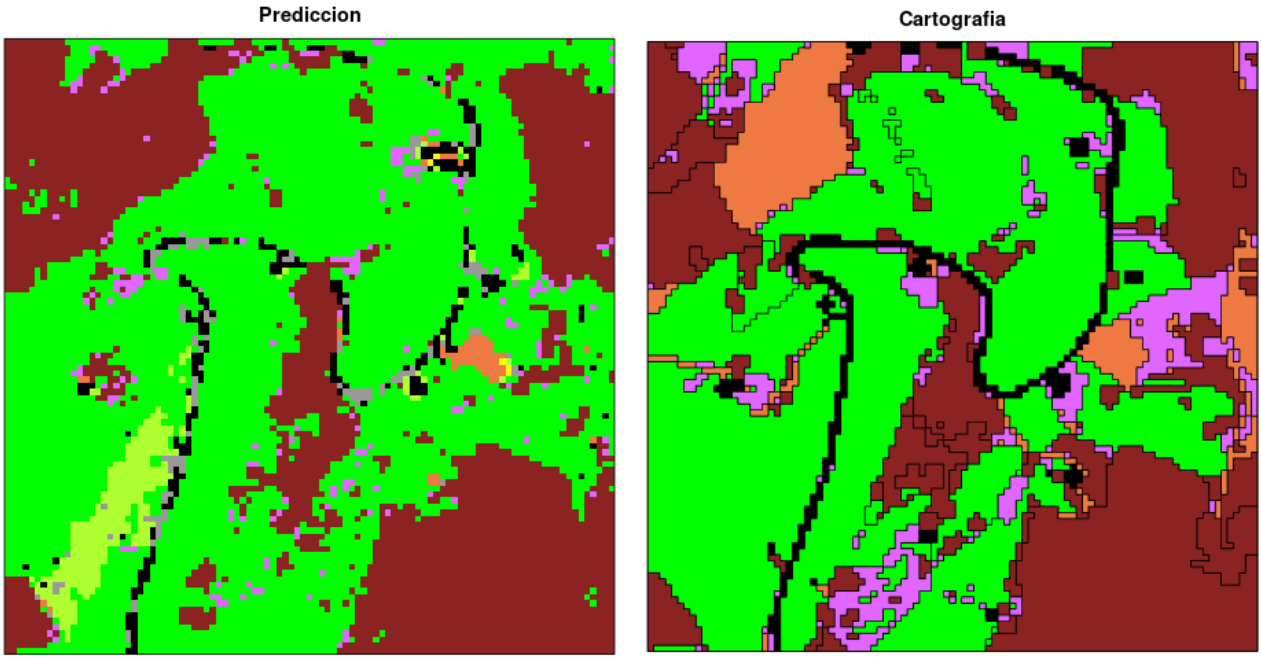

Ilustración 7.17 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

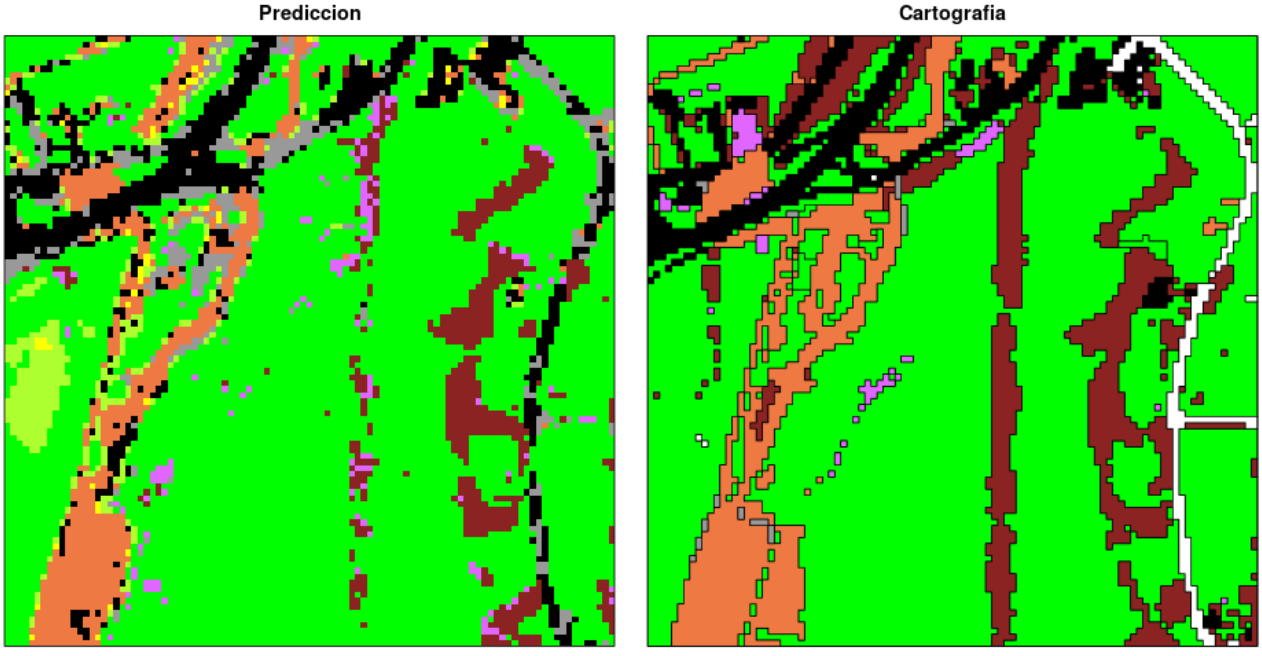

Ilustración 7.18 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

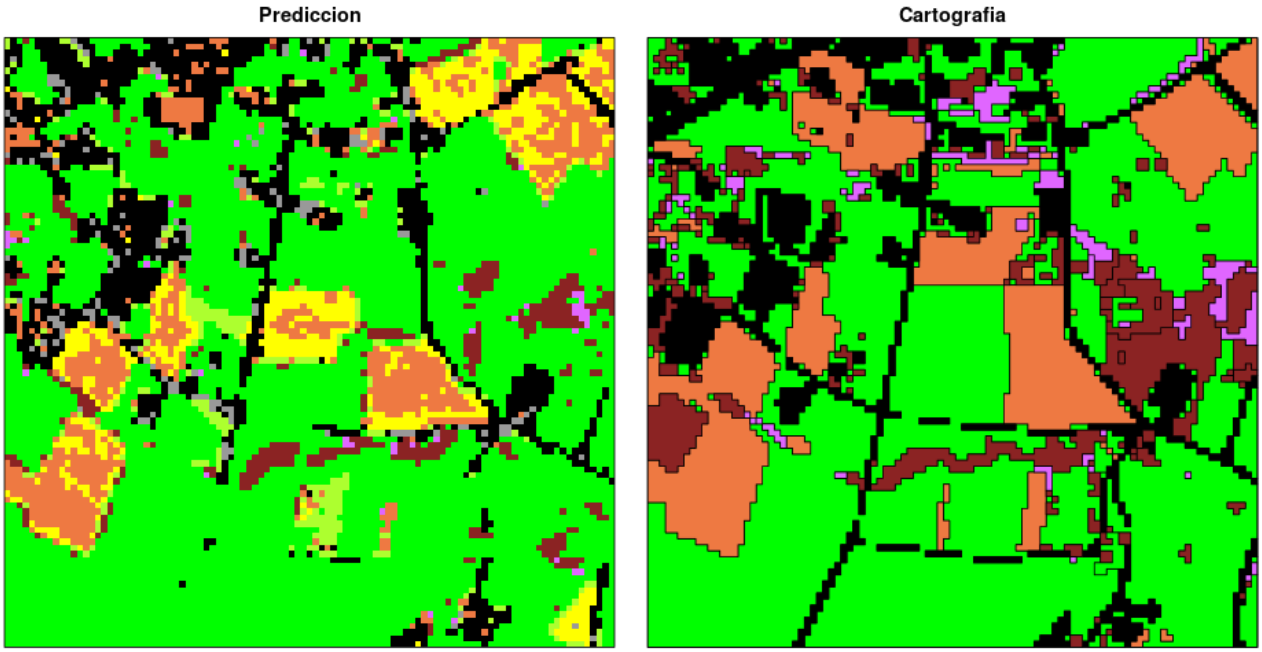

Ilustración 7.19 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

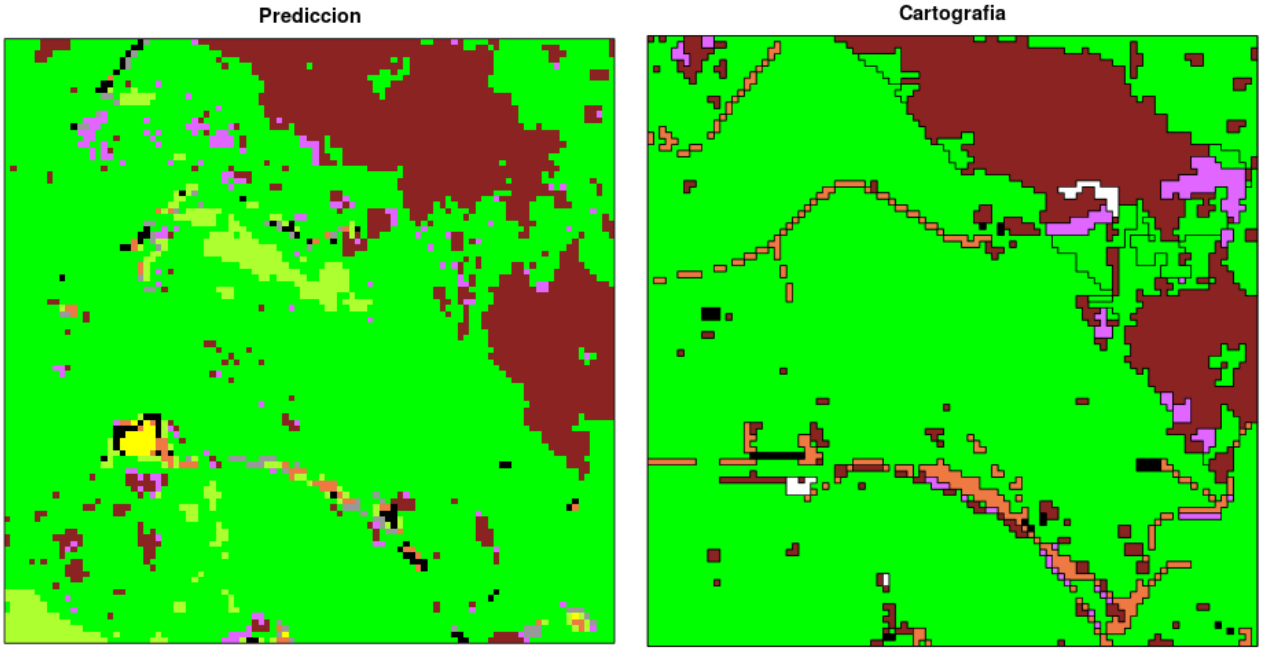

Ilustración 7.20 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

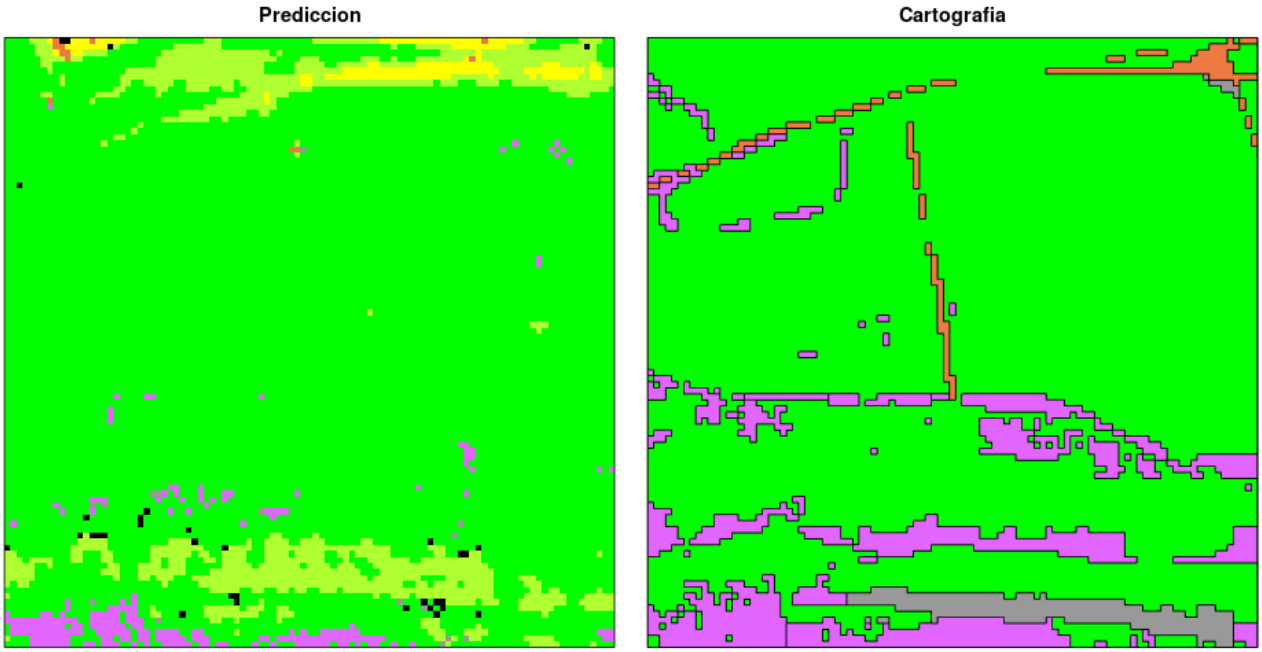

Ilustración 7.21 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

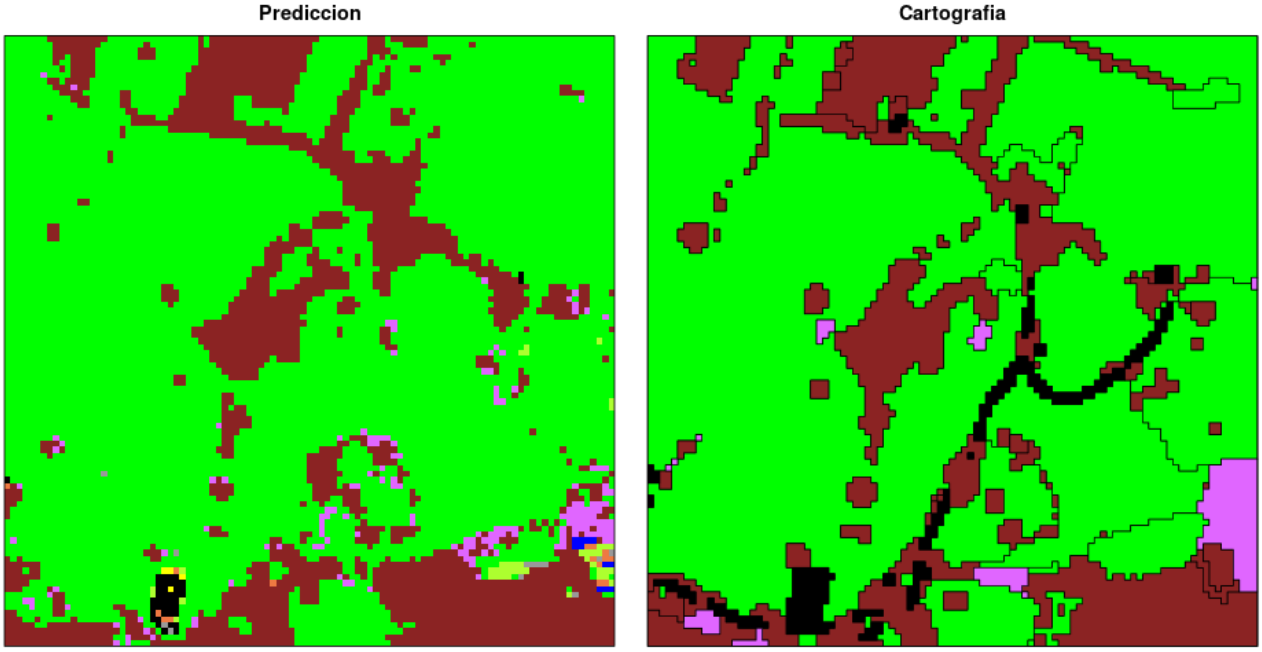

Ilustración 7.22 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

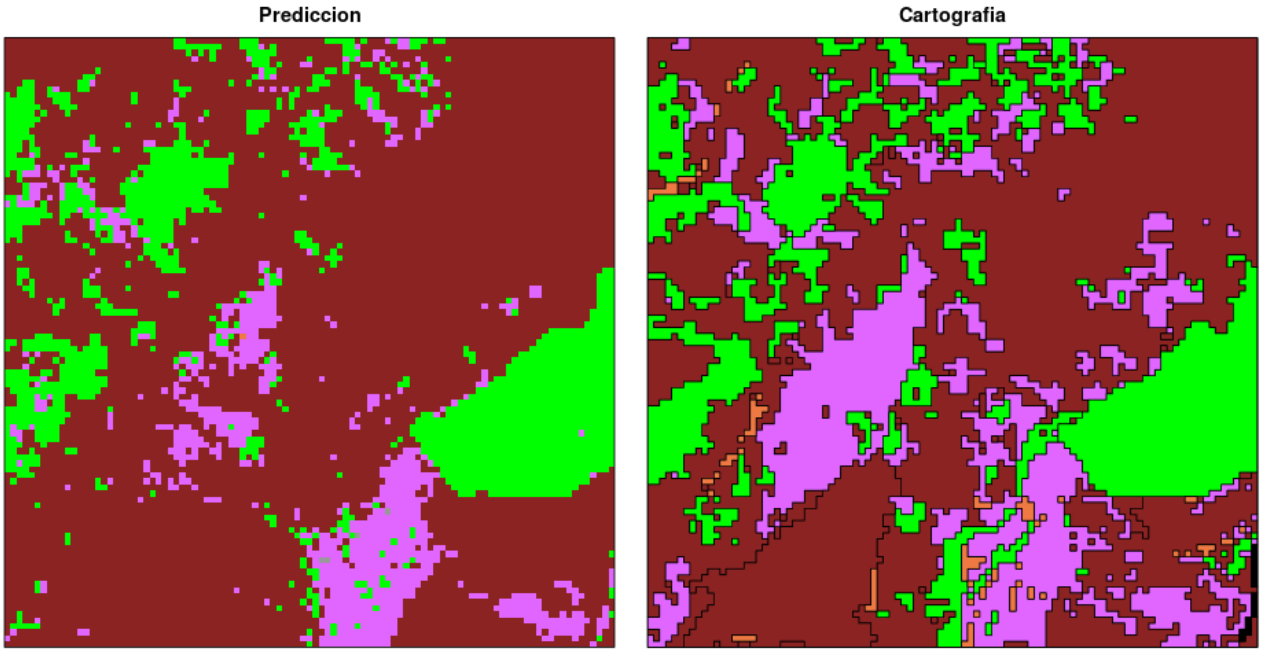

Ilustración 7.23 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

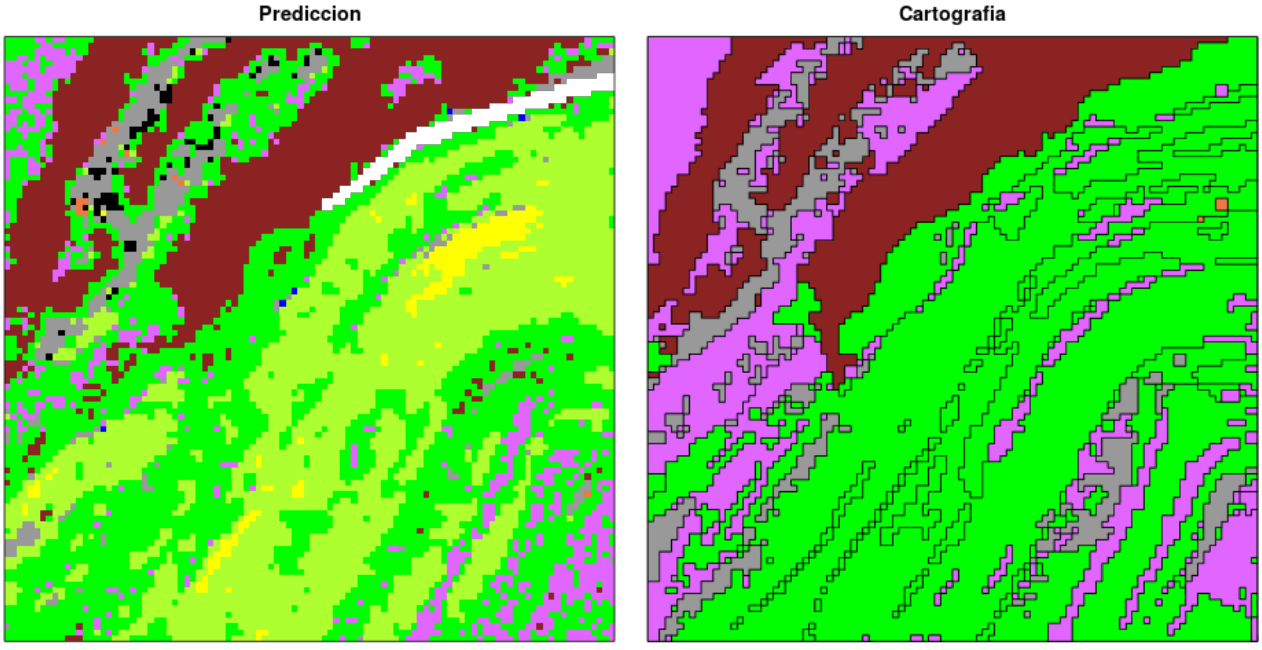

Ilustración 7.24 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

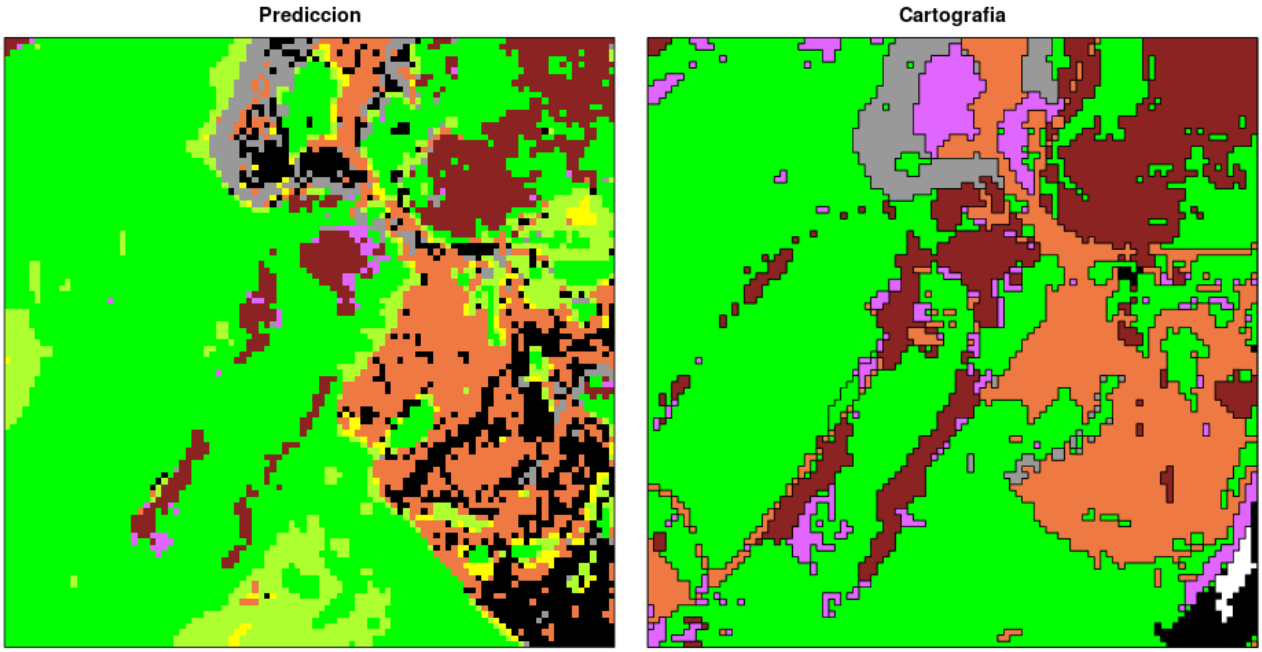

Ilustración 7.25 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

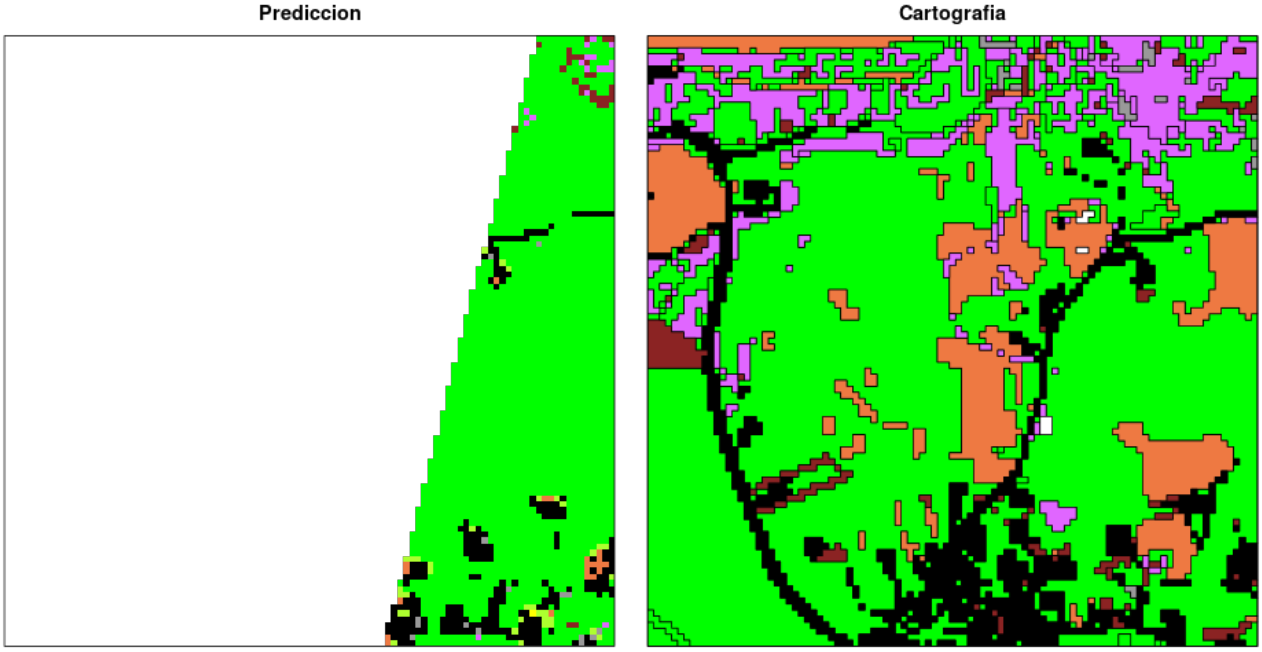

Ilustración 7.26 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

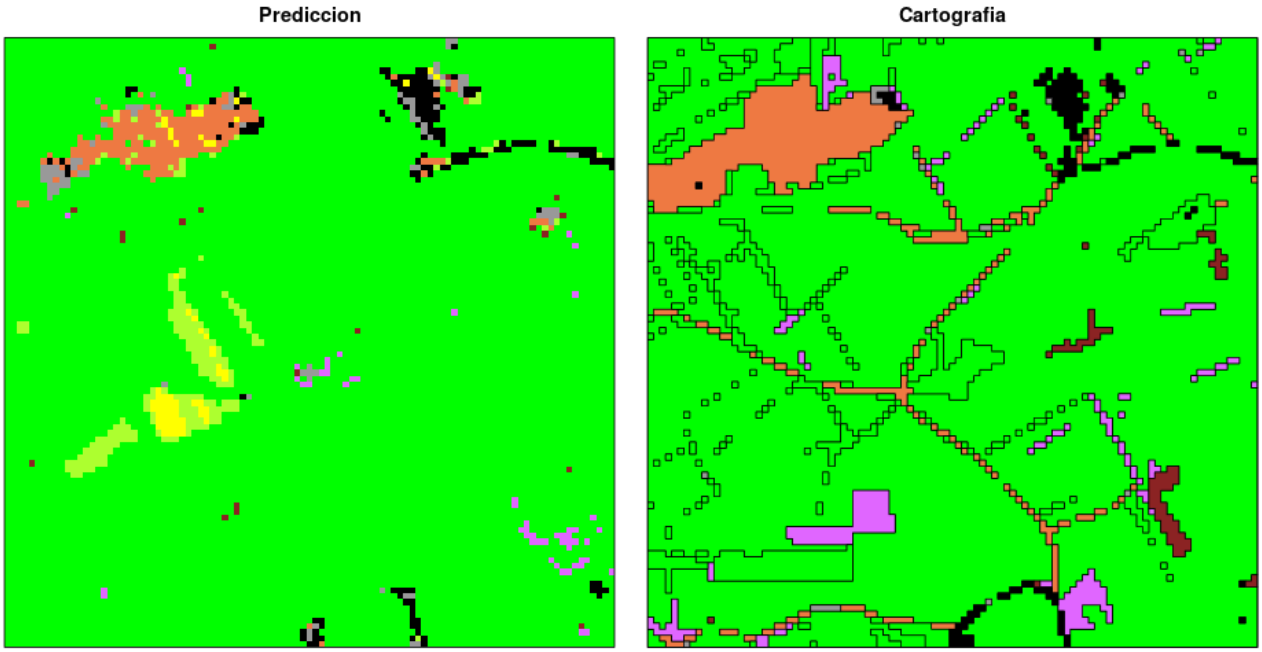

Ilustración 7.27 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

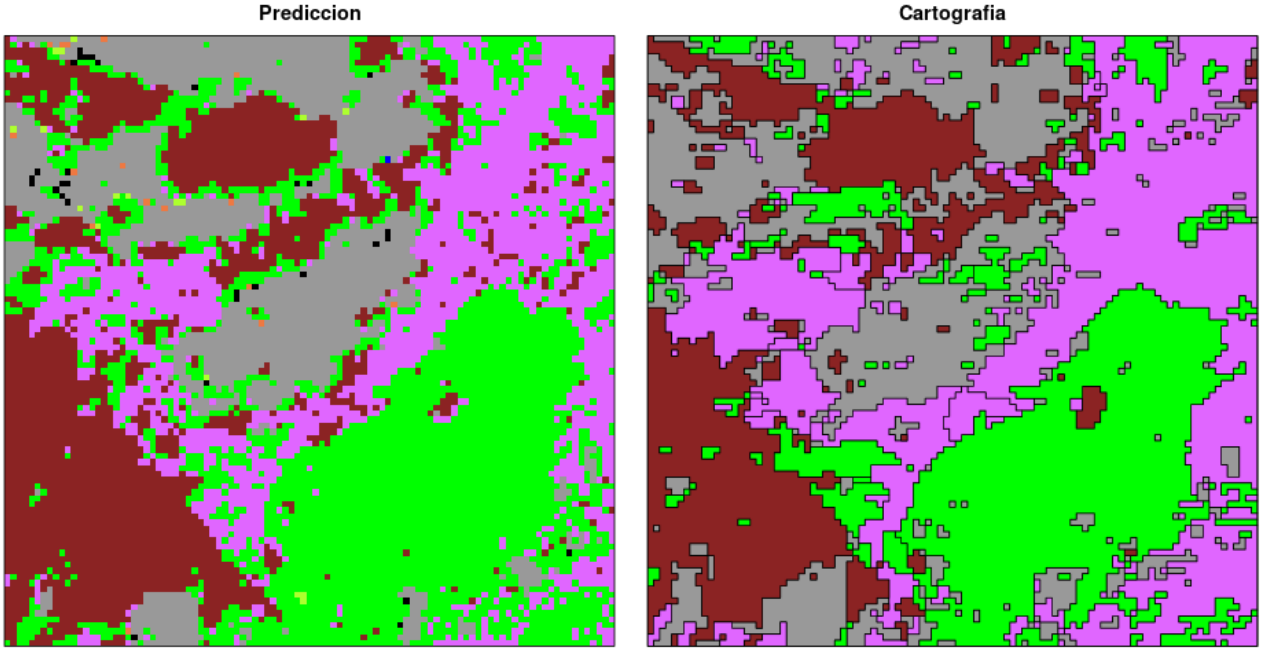

Ilustración 7.28 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

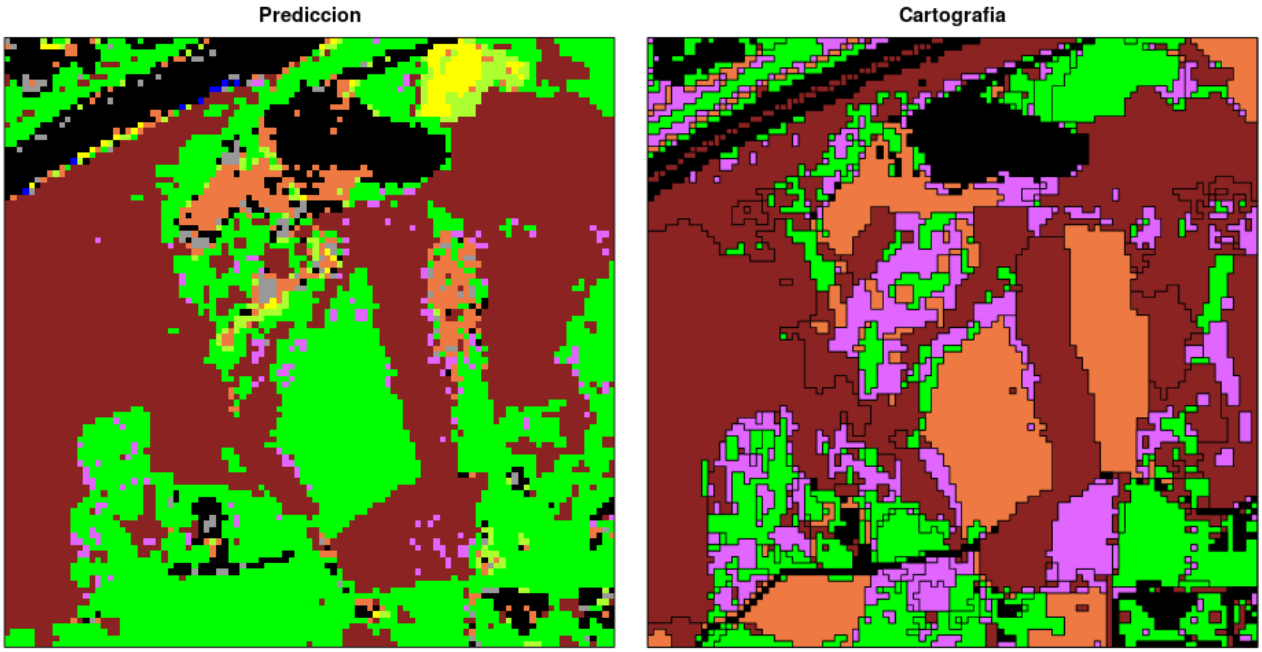

Ilustración 7.29 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

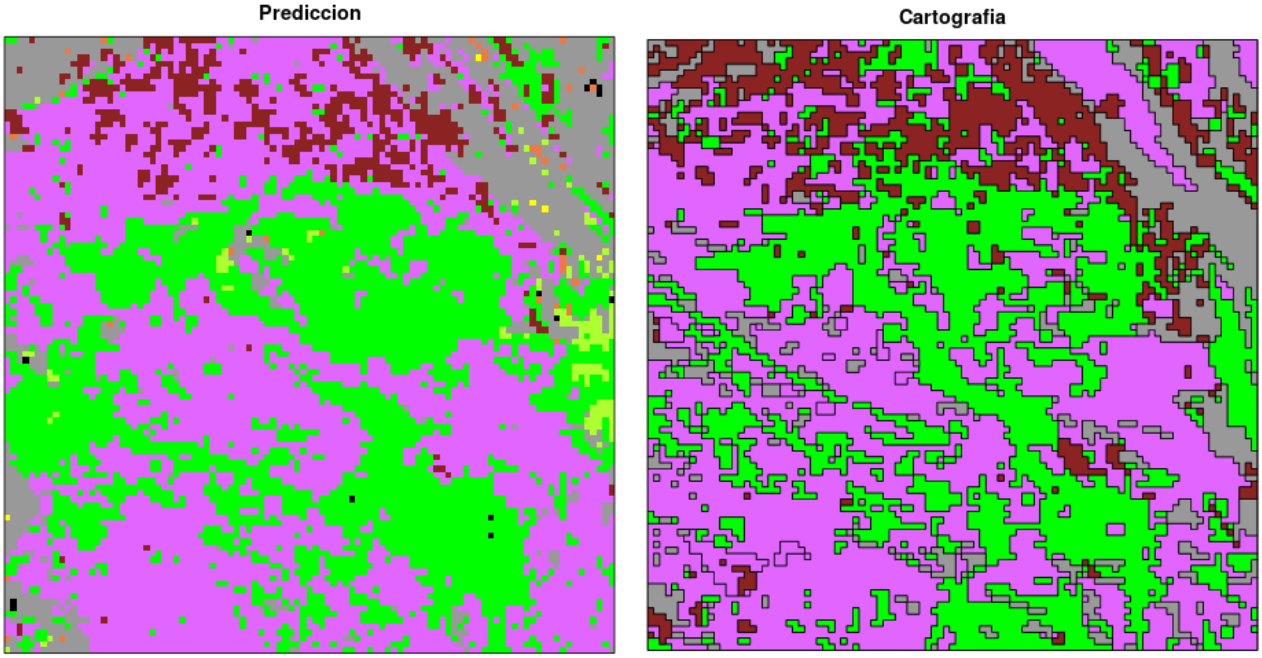

Ilustración 7.30 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

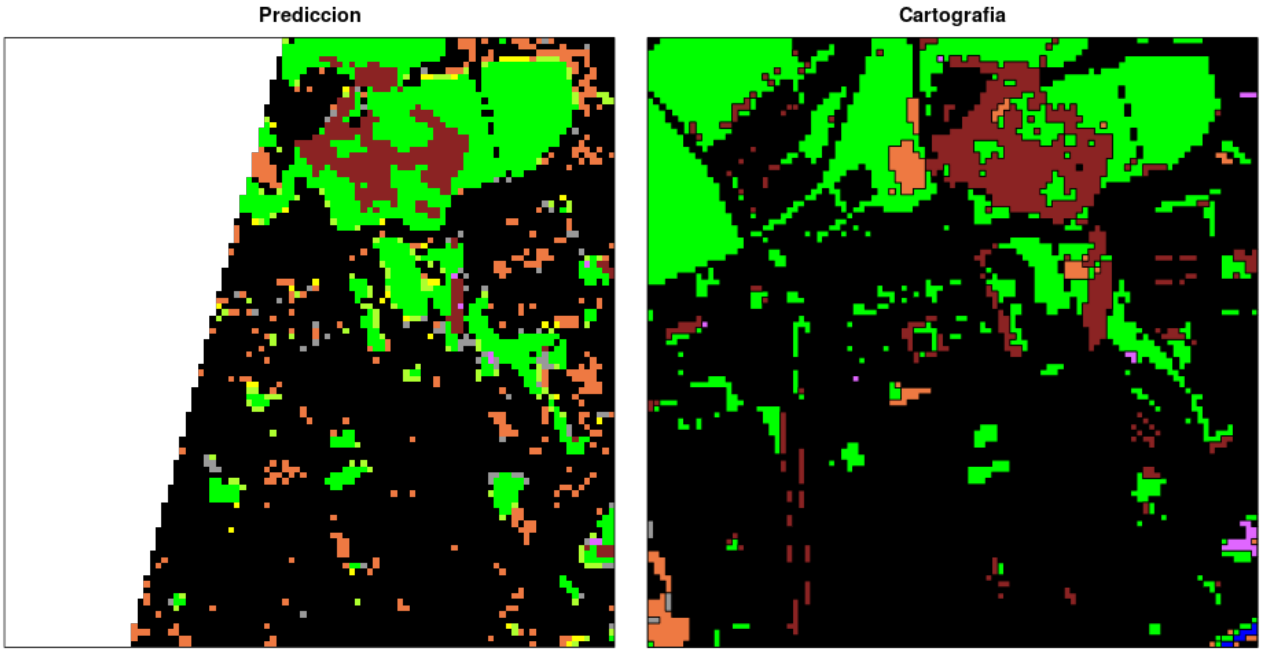

Ilustración 7.31 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

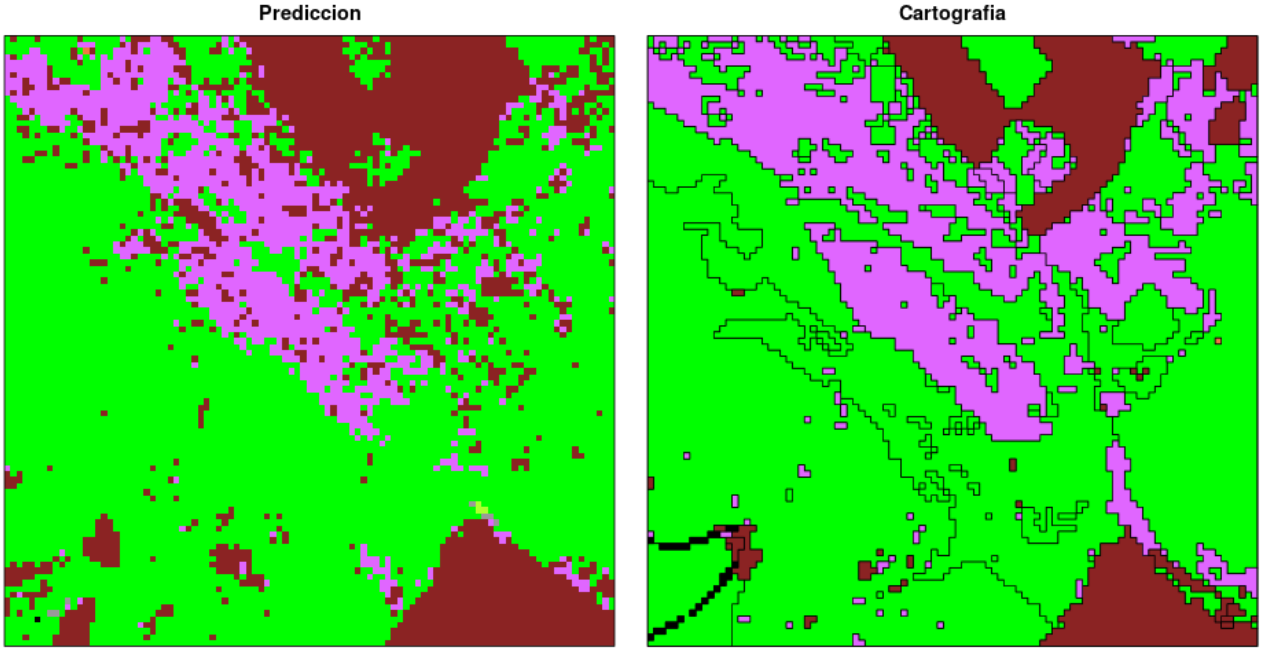

Ilustración 7.32 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

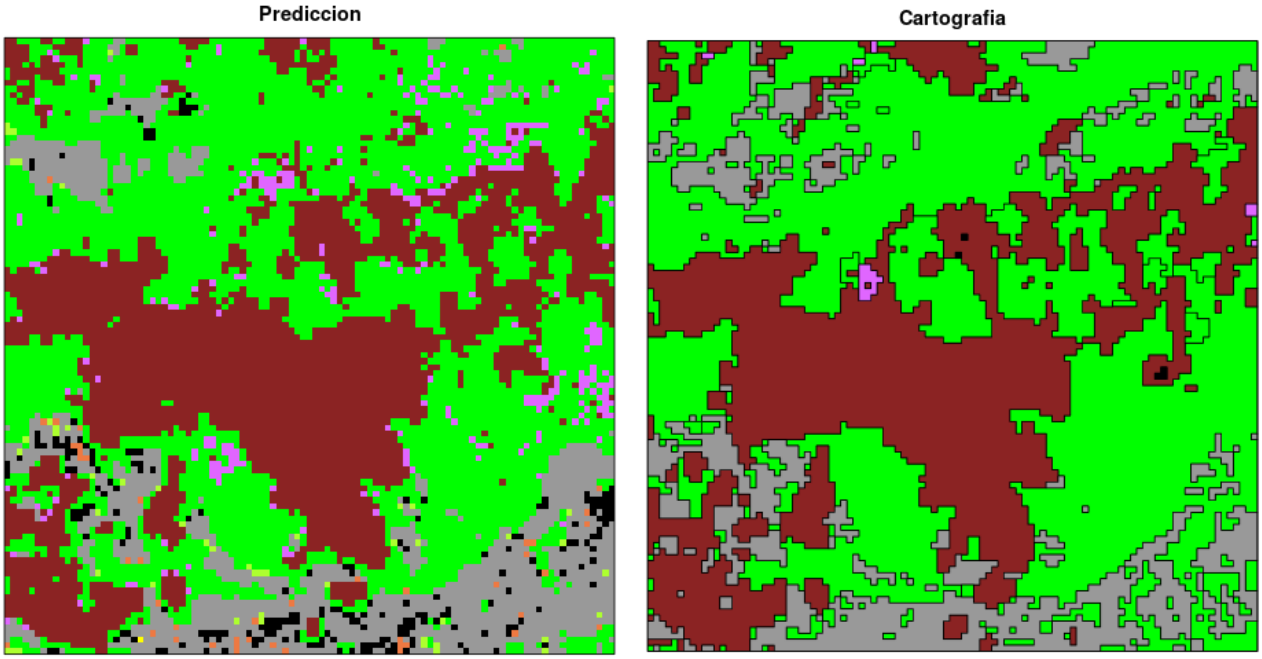

Ilustración 7.33 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

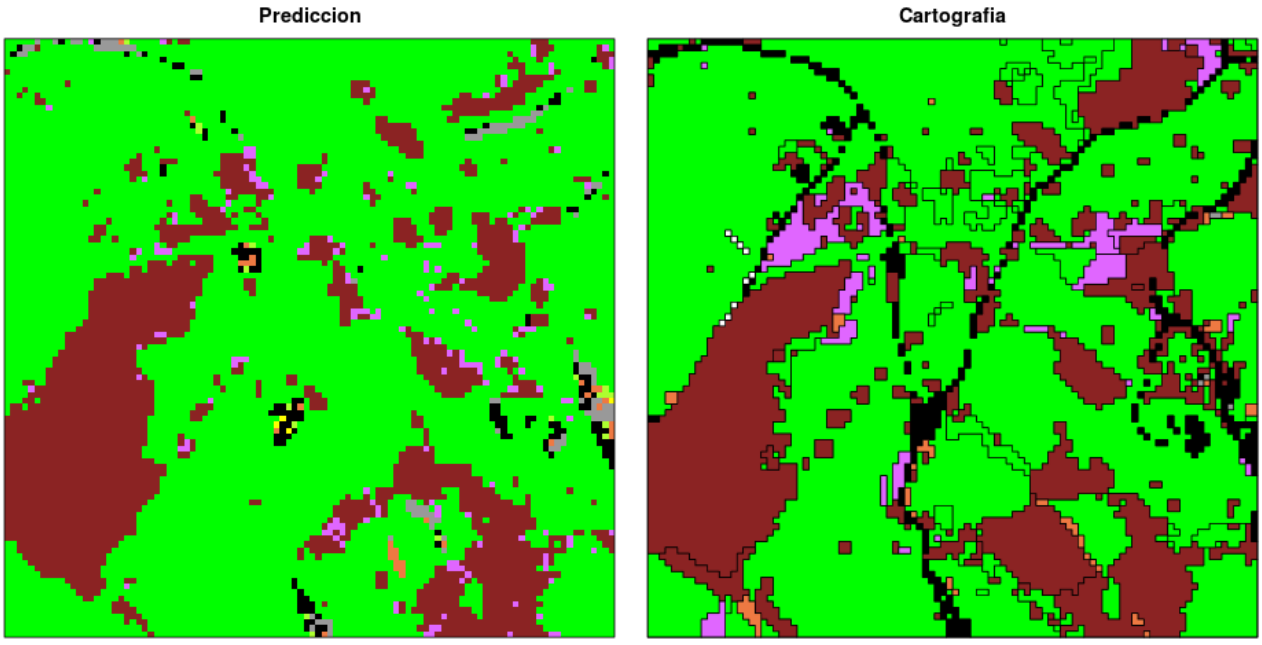

Ilustración 7.34 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

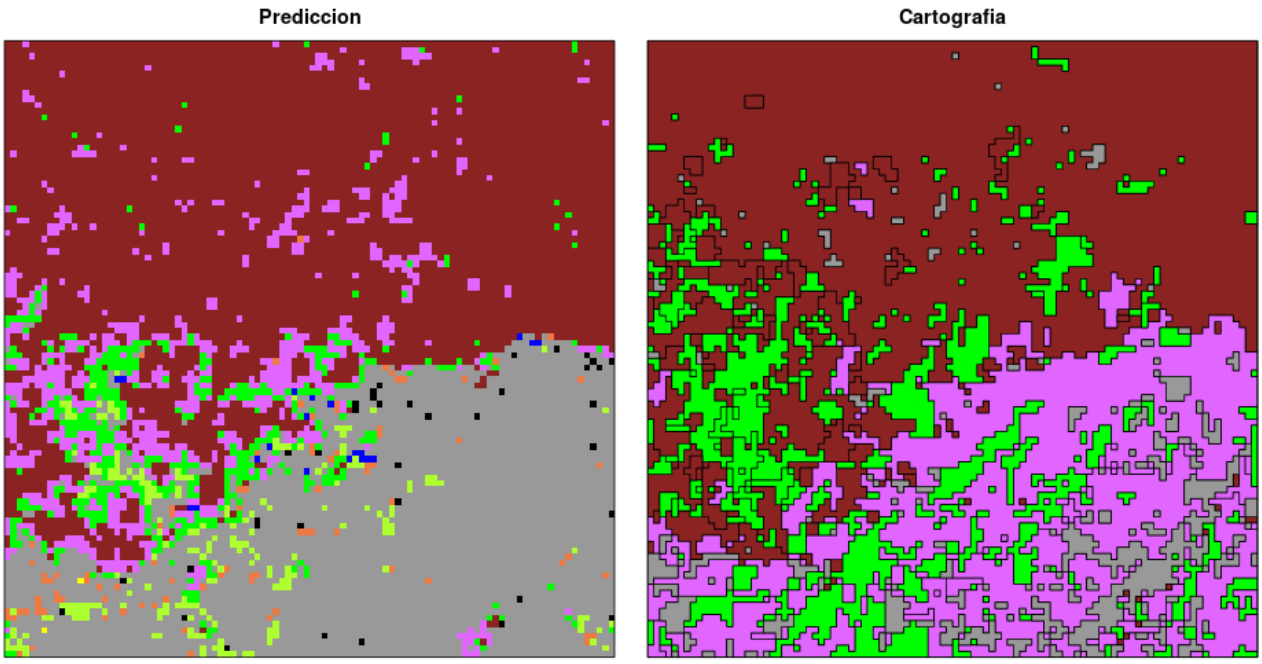

Ilustración 7.35 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

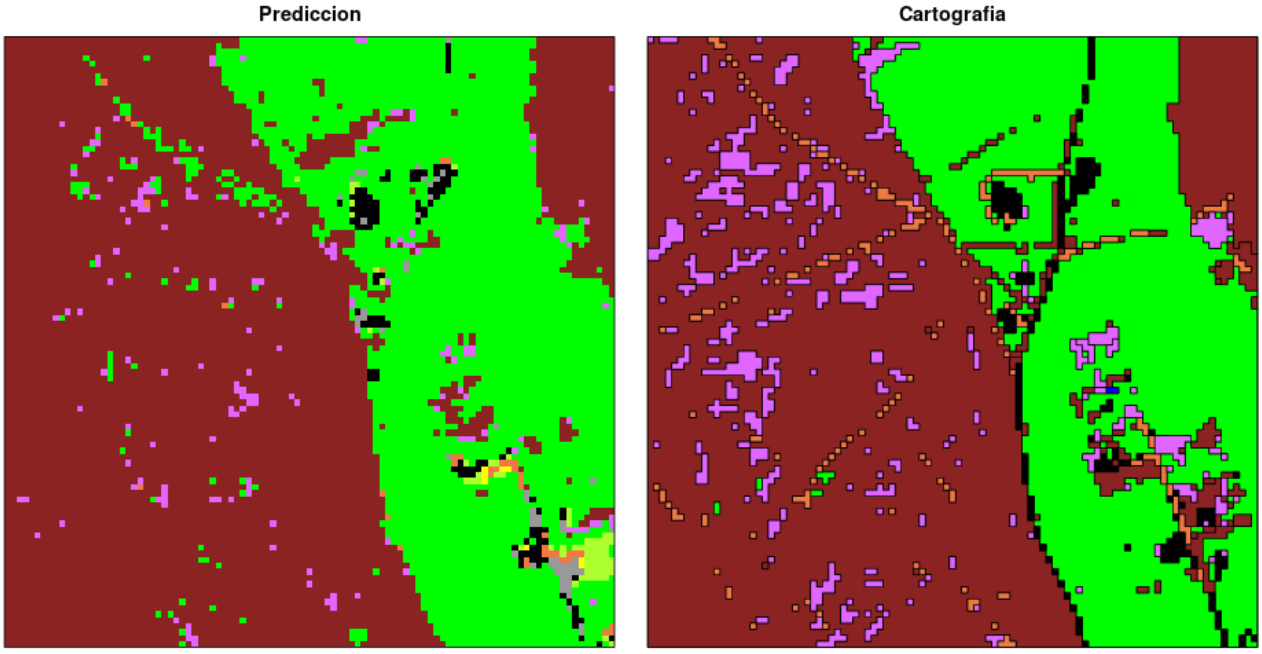

Ilustración 7.36 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

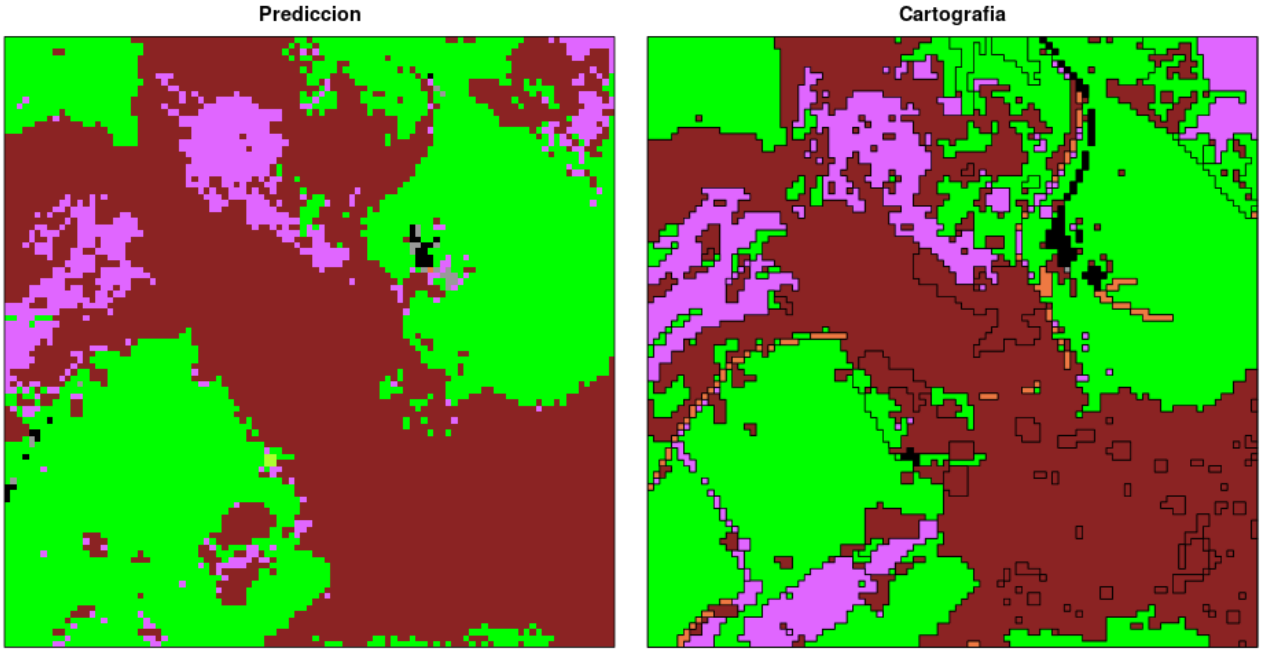

Ilustración 7.37 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

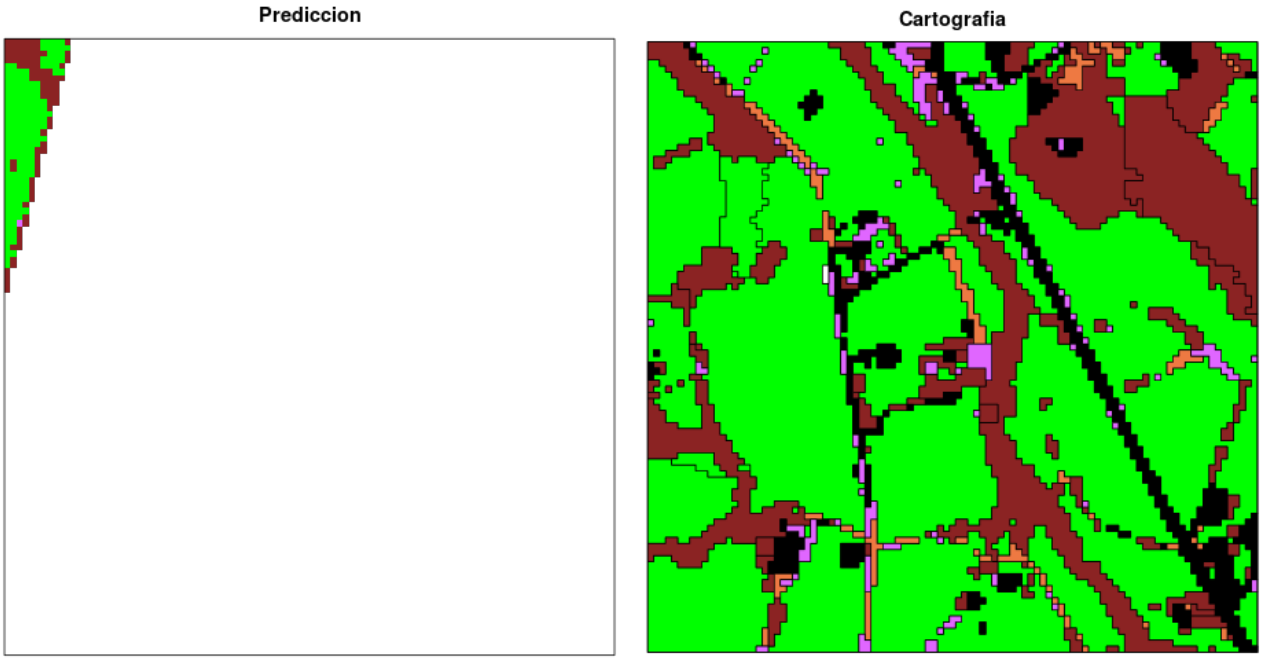

Ilustración 7.38 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

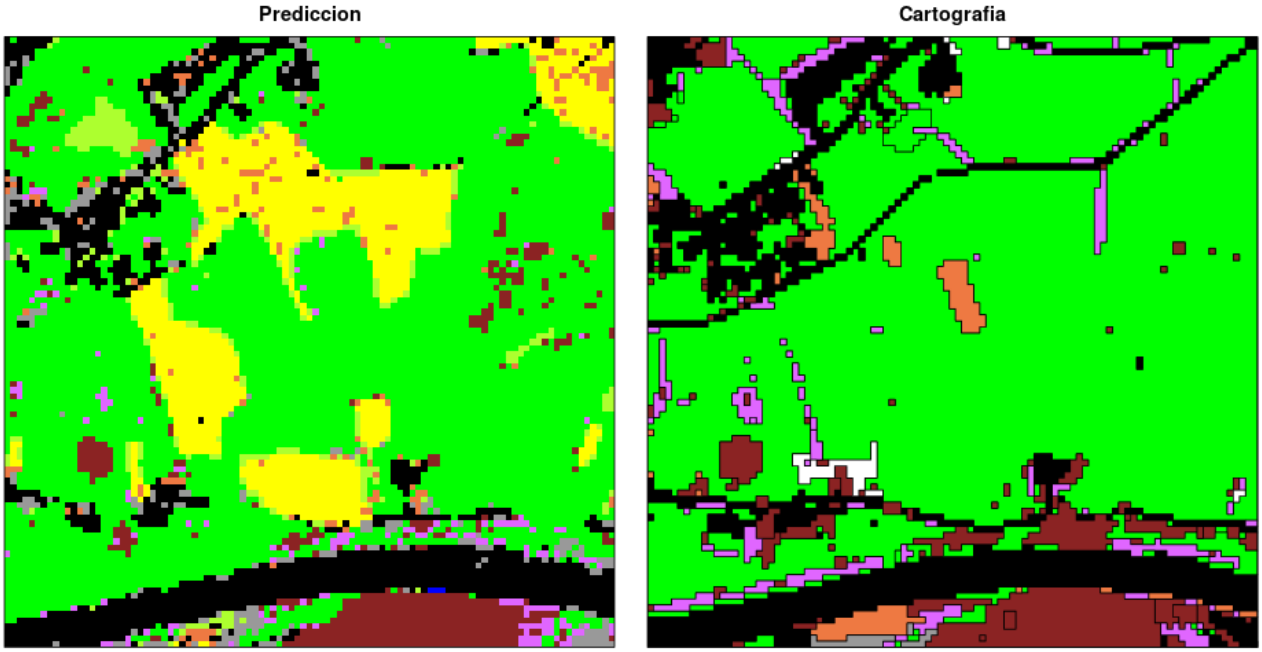

Ilustración 7.39 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

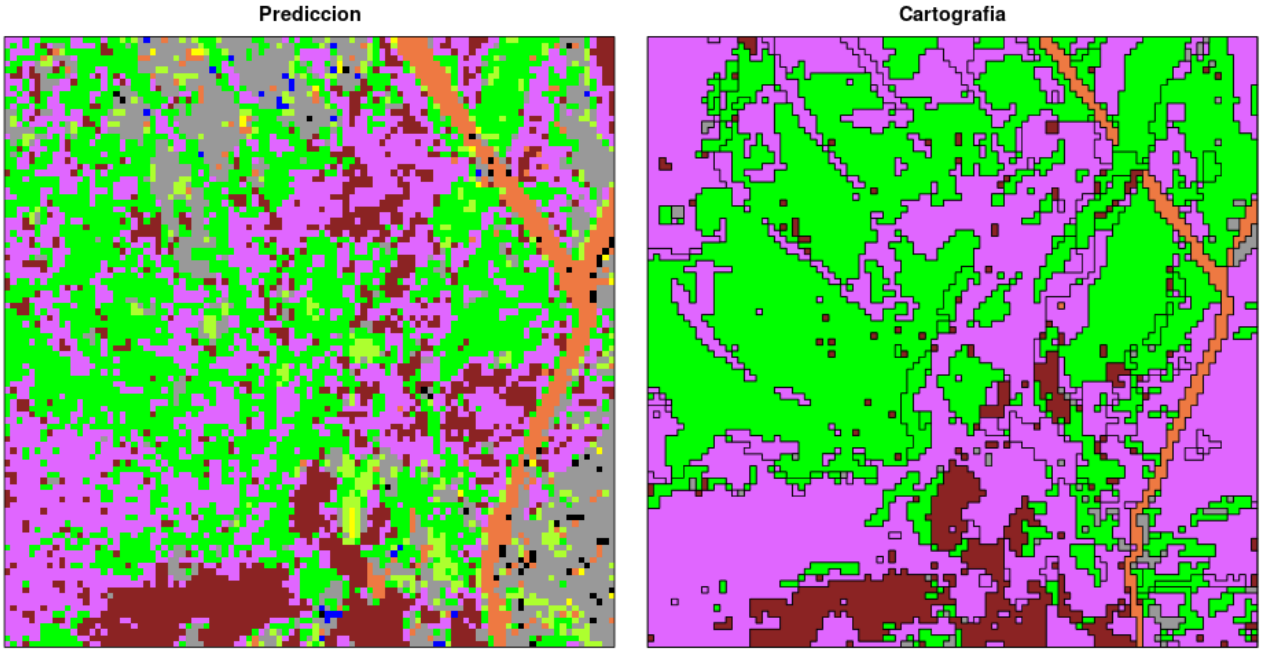

Ilustración 7.40 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

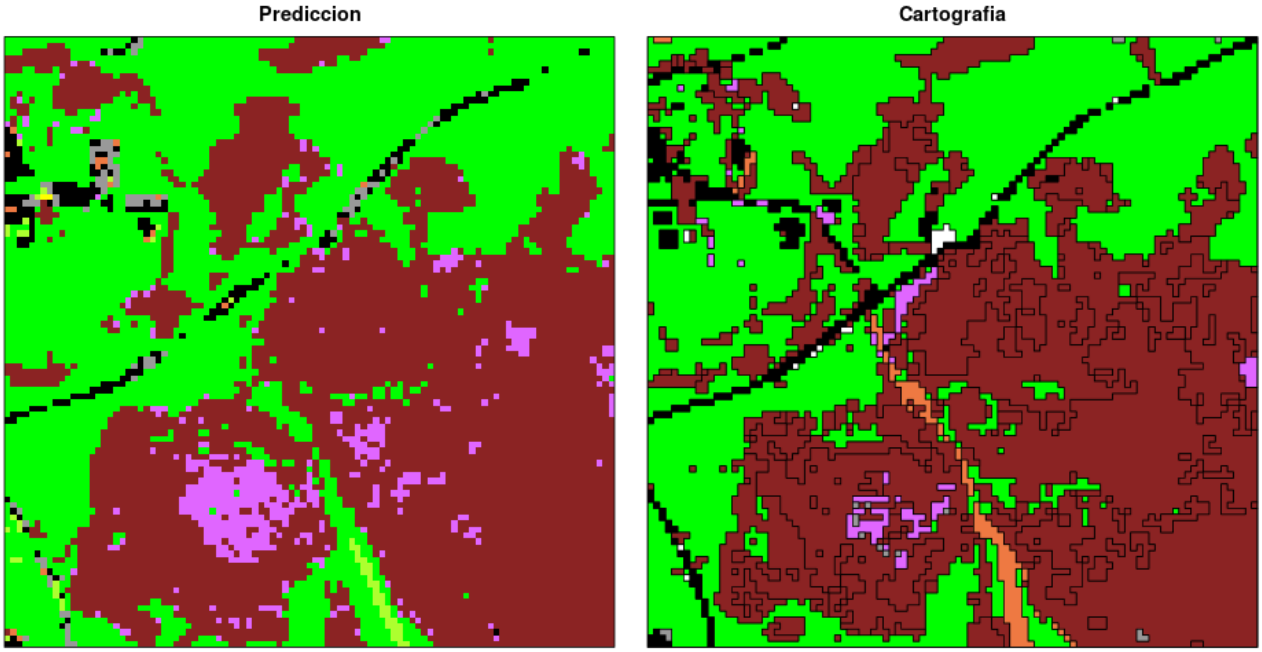

Ilustración 7.41 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

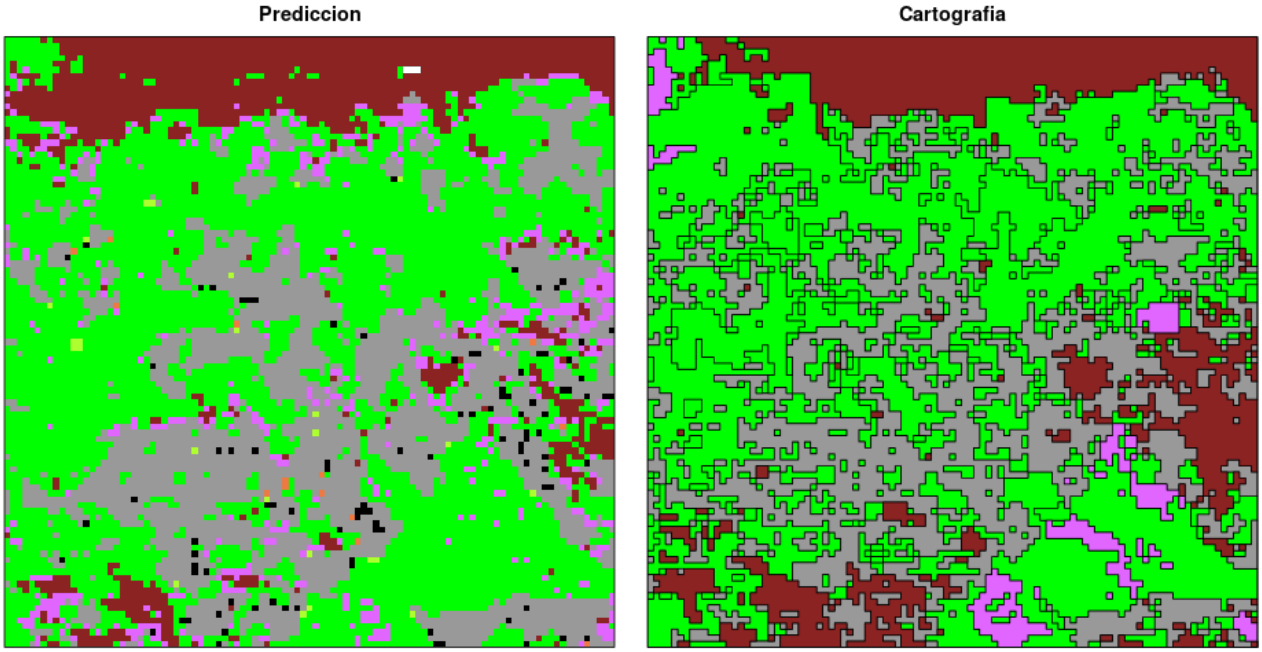

Ilustración 7.42 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

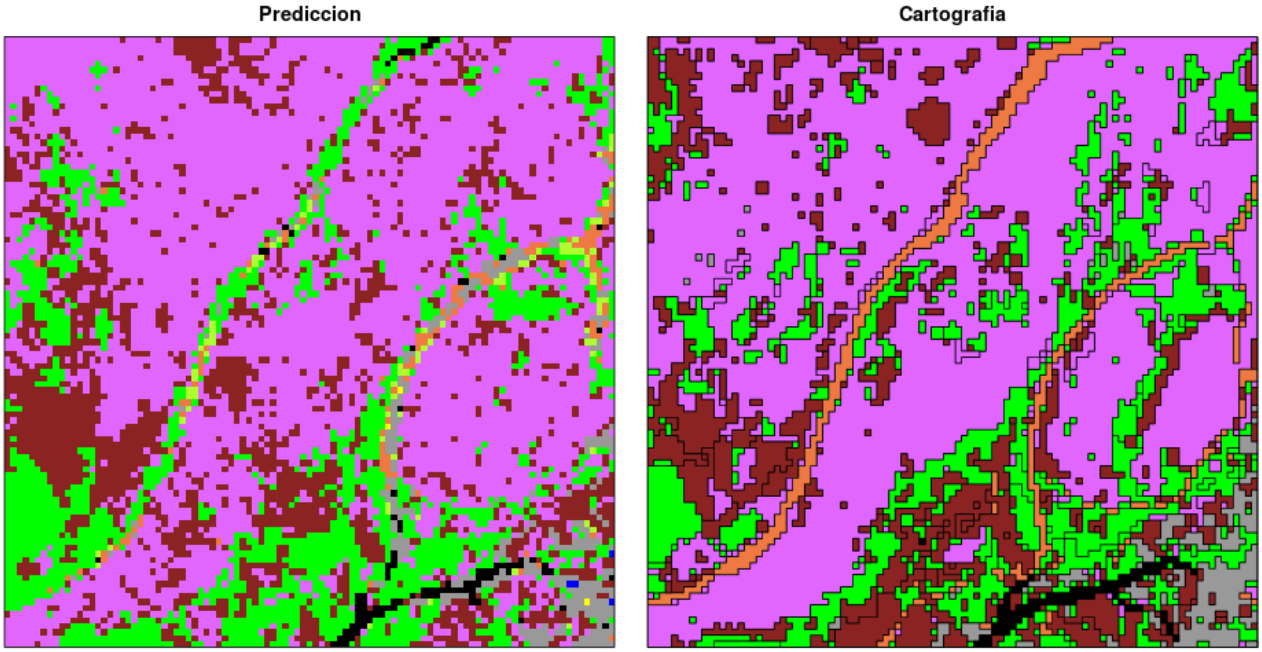

Ilustración 7.43 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

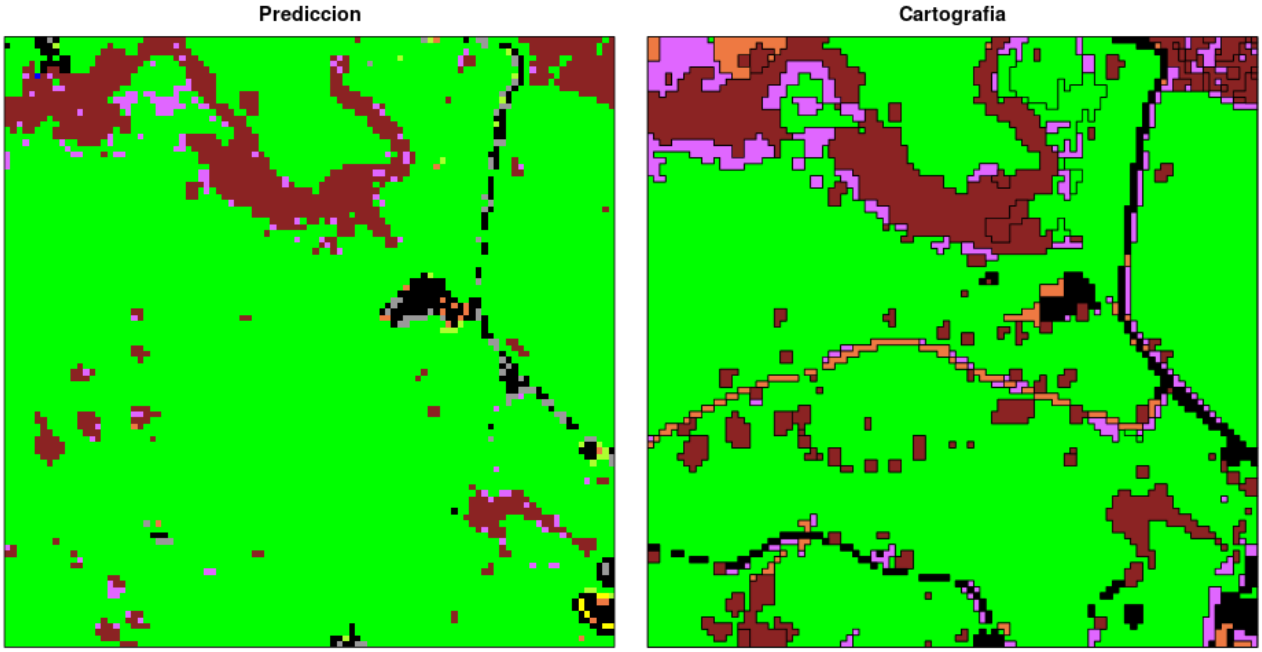

Ilustración 7.44 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

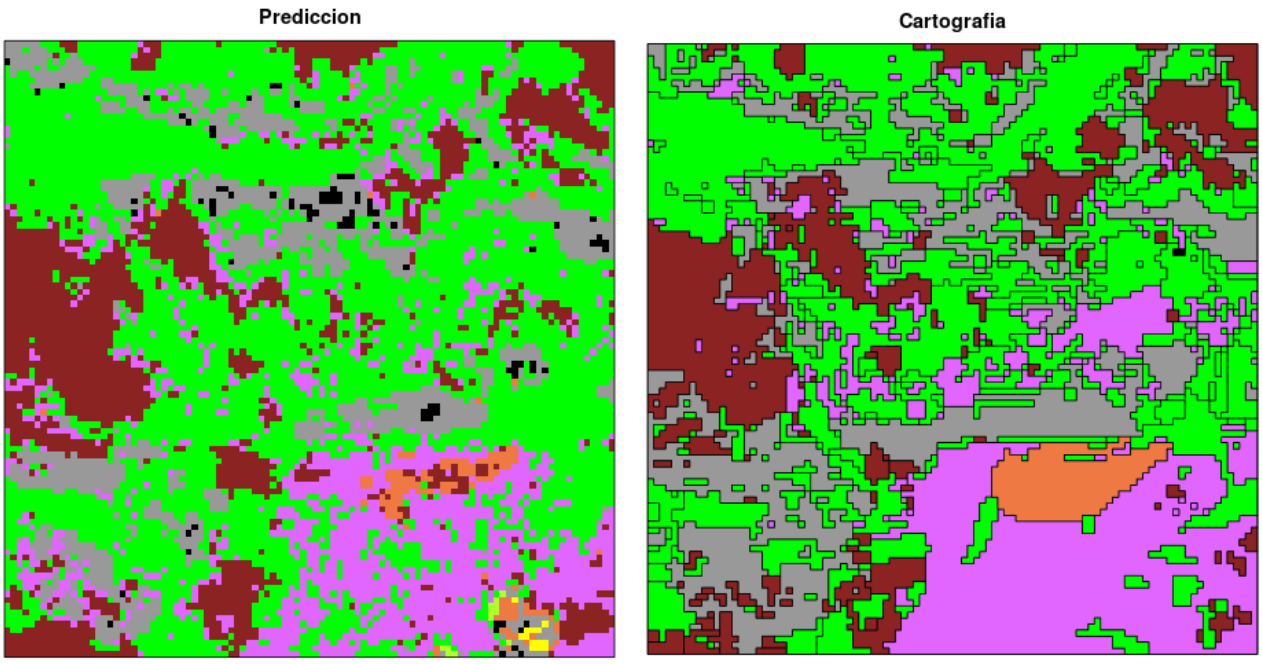

Ilustración 7.45 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

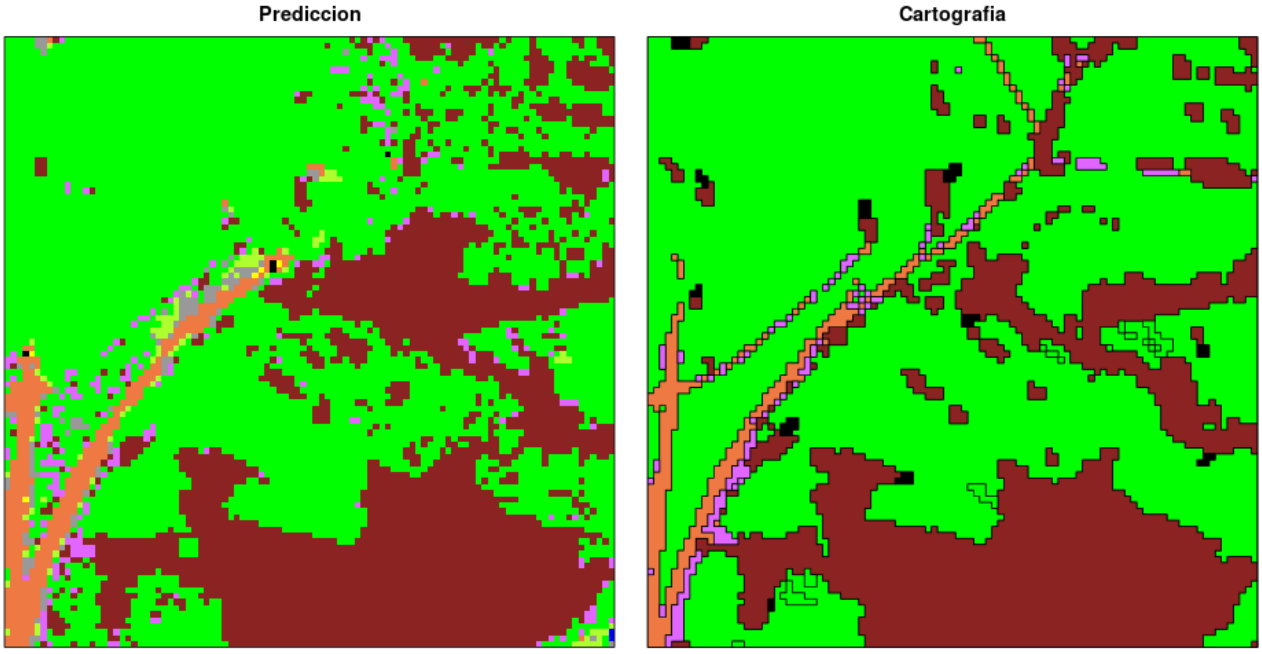

Ilustración 7.46 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

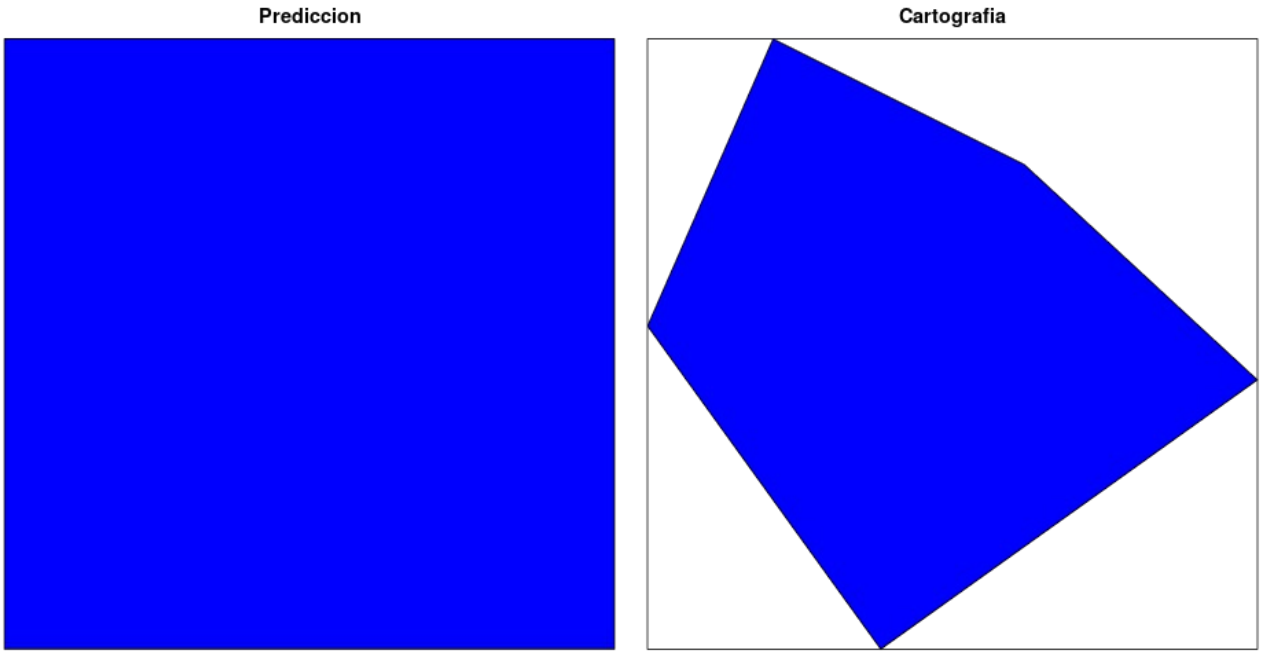

Ilustración 7.47 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

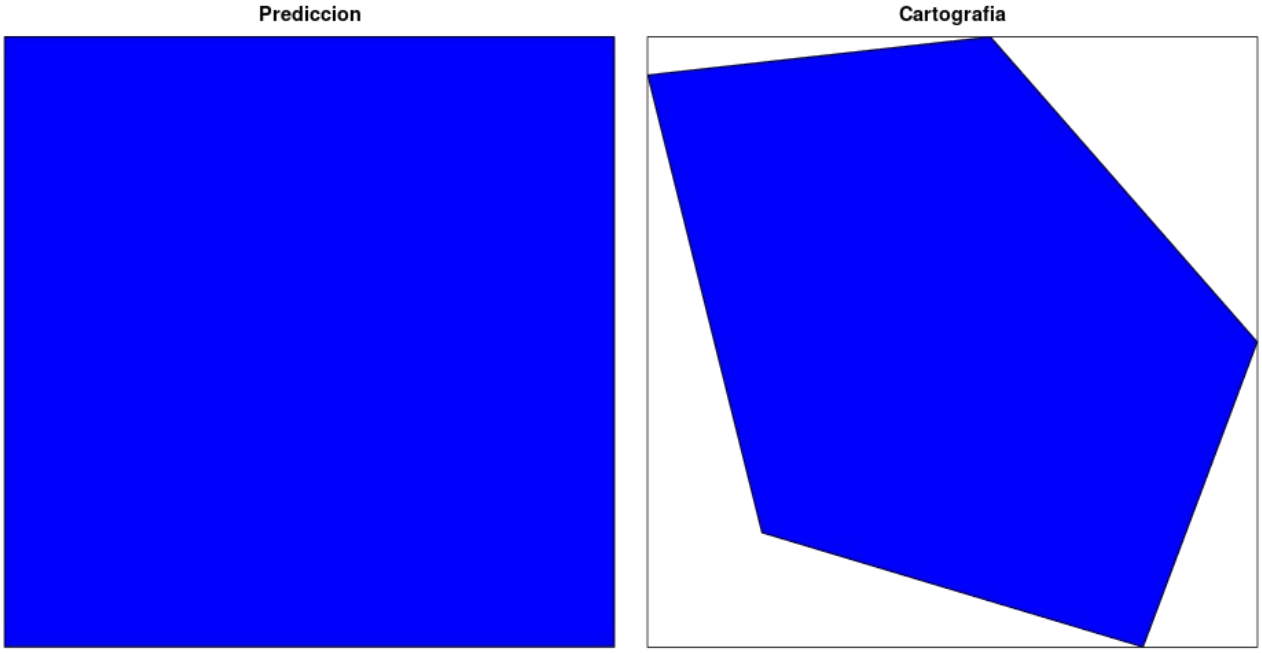

Ilustración 7.48 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.

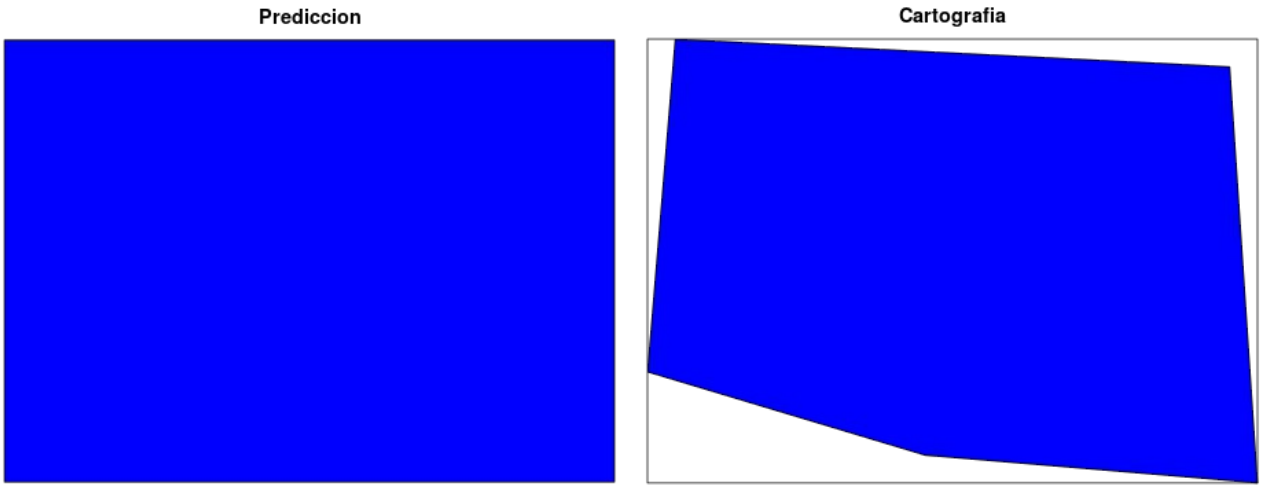

Ilustración 7.49 En esta ilustración se muestran las clases de vegetación de una zona cartografiada. A la izquierda se muestra la predicción realizada y a la derecha la cartografía de vegetación.