



UNIVERSIDAD DE CHILE  
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS  
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

**DESARROLLO DE ALGORITMOS DE APRENDIZAJE DE MÁQUINAS  
PARA LA CLASIFICACIÓN AUTOMÁTICA DE COBERTURAS DE SUELO,  
CON EL USO DE IMÁGENES SATELITALES**

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL ELÉCTRICO

PEDRO IGNACIO BRAVO COLLADO

PROFESOR GUÍA:  
JAIME ORTEGA PALMA

MIEMBROS DE LA COMISIÓN:  
LUIS REYES ROJAS  
PABLO ESTÉVEZ VALENCIA

SANTIAGO DE CHILE  
2024

## DESARROLLO DE ALGORITMOS DE APRENDIZAJE DE MÁQUINAS PARA LA CLASIFICACIÓN AUTOMÁTICA DE COBERTURAS DE SUELO, CON EL USO DE IMÁGENES SATELITALES

Existe un creciente interés por contar con mapas que contengan variadas clases de cobertura y uso de suelo (bosques, urbano, agrícola, etc.), debido a que la utilización de estos mapas permitiría mejorar la planificación territorial, evaluar cambios en el medioambiente y realizar estudios sobre el cambio climático, entre otros.

Una herramienta de gran importancia para ello son las imágenes satelitales, ya que permiten diferenciar entre coberturas de suelo cubriendo grandes extensiones de terreno y de manera dinámica al contar con imágenes periódicas y por largos periodos de tiempo.

El objetivo de esta memoria es crear una metodología para el entrenamiento de algoritmos de aprendizaje de máquinas escalable y replicable para la clasificación automática de cobertura de suelos a partir de imágenes multiespectrales de acceso libre del programa Copernicus.

Se utiliza información publicada por CONAF en conjunto a imágenes satelitales del programa Copernicus en el entrenamiento de redes neuronales convolucionales, las cuales a su vez son aplicadas a las mismas imágenes satelitales, haciendo un cruce entre las predicciones resultantes y la información publicada por CONAF. Mediante este método se logra la limpieza automática de datos, generando un conjunto de píxeles altamente representativo para cada una de los tipos de cobertura y uso de suelo a clasificar, permitiendo la caracterización espectral de cada una de las clases de interés.

Utilizando este grupo de respuestas espectrales altamente representativas, se logra entrenar diversos algoritmos de aprendizaje de máquinas (Classification and Regression Trees, Random Forest, Gradient Boosted Decision Trees y Perceptrón Multicapas), con los cuales se logra la creación automática de mapas de cobertura y uso de suelo, distinguiendo entre las clases de interés y aquellas zonas no clasificables dentro de ninguna de ellas.

Finalmente, a partir de métricas tales como *accuracy*, *precision* y *recall* se logra cuantificar la mejoría en desempeño de este tipo de algoritmos asociada a la limpieza automática previamente realizada, a la vez, se utilizan dichas métricas para concluir la mayor efectividad se logra al utilizar un modelo *Boosted Gradient Trees* entrenado con datos generados mediante la limpieza automática propuesta, obteniéndose un *accuracy* del 91 %, un *macro precision* del 82 % y un *macro recall* del 89 %.

# Tabla de Contenido

<b>1. Introducción</b>	<b>1</b>
1.1. Mapas de cobertura de suelo . . . . .	1
1.1.1. Uso de mapas de cobertura de suelo . . . . .	1
1.2. Sensores multiespectrales . . . . .	2
1.3. El programa Copernicus . . . . .	2
1.4. Objetivos del trabajo de título . . . . .	2
1.4.1. General . . . . .	2
1.4.2. Específicos . . . . .	2
<b>2. Marco Teórico y Estado del Arte</b>	<b>3</b>
2.1. Marco Teórico . . . . .	3
2.1.1. Firmas Espectrales (F.E.) . . . . .	3
2.1.2. Índices Espectrales . . . . .	4
2.1.3. Centro Regional Copernicus . . . . .	4
2.1.4. Características del satélite (Sentinel-2) . . . . .	5
2.1.5. Redes Neuronales . . . . .	6
2.1.5.1. Perceptrón Multicapa . . . . .	6
2.1.5.2. Red Neuronal Convolutiva . . . . .	6
2.1.6. Árboles de clasificación . . . . .	7
2.1.6.1. Árbol de clasificación . . . . .	7
2.1.6.2. Random Forest . . . . .	7
2.1.6.3. Gradient Boosted Trees . . . . .	7
2.1.7. Métricas de evaluación . . . . .	8
2.1.7.1. Matriz de confusión: . . . . .	8
2.1.7.2. <i>Accuracy</i> . . . . .	8
2.1.7.3. <i>Precision</i> . . . . .	9
2.1.7.4. <i>Recall</i> . . . . .	9
2.1.7.5. <i>F1</i> . . . . .	9
2.2. Estado del Arte . . . . .	10
2.2.1. Uso del satélite Sentinel-2 . . . . .	10
2.2.2. Uso de imágenes hiperespectrales . . . . .	10
2.2.3. Mapas de cobertura de suelo en Chile . . . . .	11
2.2.3.1. CONAF . . . . .	11
2.2.3.2. Copernicus . . . . .	13
<b>3. Metodología</b>	<b>14</b>
3.1. Propuesta metodológica . . . . .	14

3.1.1.	Datos a utilizar . . . . .	14
3.1.1.1.	Datos auxiliares para evaluación de los algoritmos . . . . .	14
3.1.2.	Zona geográfica del estudio . . . . .	14
3.1.3.	Técnicas a comparar . . . . .	15
3.1.4.	Clases a identificar . . . . .	15
3.1.5.	Tratamiento de datos . . . . .	15
3.1.5.1.	Corrección geométrica . . . . .	15
3.1.5.2.	Eliminación de nubes . . . . .	16
3.1.6.	Uso de información complementaria . . . . .	16
3.1.6.1.	Uso de índices espectrales . . . . .	16
3.1.7.	Generación de datos para entrenar y evaluar los distintos algoritmos. . . . .	16
3.1.7.1.	Limpieza automática de datos . . . . .	17
<b>4.</b>	<b>Contribución y Resultados</b>	<b>18</b>
4.1.	Contribución . . . . .	18
4.1.1.	Entregables esperados . . . . .	18
4.2.	Resultados . . . . .	19
4.2.1.	Adquisición y visualización de datos . . . . .	19
4.2.1.1.	Catastro de cobertura de suelos (CONAF) . . . . .	19
4.2.1.2.	Transformación de las clases de uso/cobertura de suelo . . . . .	19
4.2.1.3.	Subdivisión del área de interés en bloques . . . . .	20
4.2.1.4.	Adquisición de las imágenes espectrales para bloques de interés . . . . .	20
4.2.1.5.	Creación de ventanas representativas para cada clase . . . . .	20
4.2.2.	Creación, entrenamiento y evaluación de algoritmos . . . . .	21
4.2.2.1.	Red convolucional base (CNN) . . . . .	21
4.2.2.2.	Red paralela rotacional (CNN) . . . . .	23
4.2.2.3.	Incorporación de índices espectrales (CNN) . . . . .	25
4.2.2.4.	Utilización de EuroSat sobre arquitecturas anteriores . . . . .	27
4.2.2.5.	Random Forest . . . . .	28
4.2.2.6.	Creación de píxeles representativos utilizando red convolucional . . . . .	29
4.2.2.7.	Caracterización de respuestas espectrales por clase utilizando píxeles representativos . . . . .	31
4.2.2.8.	Random Forest entrenado utilizando píxeles representativos . . . . .	35
4.2.2.9.	Otros tipos de algoritmos utilizando píxeles representativos . . . . .	45
<b>5.</b>	<b>Análisis de los resultados</b>	<b>49</b>
5.1.	Redes Convolucionales . . . . .	49
5.1.1.	Desempeño de redes convolucionales . . . . .	49
5.1.2.	Limpieza automática . . . . .	50
5.2.	Algoritmos para clasificación automática de cobertura/uso de suelos . . . . .	51
<b>6.</b>	<b>Conclusión</b>	<b>52</b>
	<b>Bibliografía</b>	<b>53</b>
	<b>Anexos</b>	<b>56</b>
	<b>Anexo A. Mapas generados (RF y GBDT)</b>	<b>56</b>

# Índice de Tablas

2.1.	Descripción de bandas del satélite Sentinel-2. . . . .	5
2.2.	Matriz de confusión normalizada por columnas. . . . .	9
2.3.	Matriz de confusión normalizada por filas. . . . .	9
2.4.	Comparación de <i>accuracy</i> logrado para diferentes aplicaciones de machine learning sobre datos Sentinel-2 <sup>[20]</sup> . . . . .	10
2.5.	Clasificaciones de cobertura de suelo CONAF. . . . .	11
2.6.	Años de catastros CONAF. . . . .	11
2.7.	Clasificaciones de cobertura de suelo Copernicus. . . . .	13
4.1.	Relación entre clases utilizadas por CONAF y clases propuestas. . . . .	19
4.2.	Resultados para modelo convolucional base. . . . .	21
4.3.	Resultados para modelo convolucional rotacional. . . . .	23
4.4.	Resultados para modelo convolucional inclusivo de índices espectrales. . . . .	25
4.5.	Resultados para arquitecturas convolucionales propuestas aplicadas en dataset EuroSat. . . . .	27
4.6.	Resultados para modelo Random Forest entrenado con datos de CONAF. . . . .	28
4.7.	Resultados para Random Forest entrenado con datos sin filtrar sobre datos filtrados. . . . .	35
4.8.	Tabla comparativa resultados para modelos Random Forest. . . . .	37
4.9.	Resultados para Random Forest entrenado con datos sin filtrar sobre datos filtrados, con un umbral de certeza igual al 60%. . . . .	37
4.10.	Métricas resultantes para modelos entrenados con distintos algoritmos de limpieza, con exigencia de 60 % de certeza. . . . .	39
4.11.	Resultados modelos RF, primera zona geográfica. . . . .	40
4.12.	Resultados modelos RF, segunda zona geográfica. . . . .	41
4.13.	Resultados modelos RF, tercera zona geográfica. . . . .	42
4.14.	Resultados modelos RF, cuarta zona geográfica. . . . .	43
4.15.	Resultados modelos RF, quinta zona geográfica. . . . .	44
5.1.	Métricas resultantes para modelos convolucionales entrenados usando dataset generado a partir de datos CONAF. . . . .	49
5.2.	Métricas resultantes para modelos convolucionales entrenados usando dataset EuroSat. . . . .	49
5.3.	Métricas resultantes para modelos clasificadores de píxeles entrenados usando dataset limpiado automáticamente. . . . .	51
5.4.	Métricas resultantes para modelos clasificadores de píxeles entrenados usando dataset limpiado automáticamente, umbral de certeza igual al 60%. . . . .	51

# Índice de Ilustraciones

2.1.	Representación gráfica de una F.E. cualquiera <sup>[9]</sup> . . . . .	3
2.2.	Variación de F.E. de acuerdo a tipo de vegetación <sup>[9]</sup> . . . . .	3
2.3.	Red Convolutiva Simple [15]. . . . .	6
2.4.	Clasificación de tipo de iris utilizando árbol de clasificación [16]. . . . .	7
2.5.	Ejemplo de matriz de confusión no normalizada. [19] . . . . .	8
2.6.	Parque Oncol <sup>[9]</sup> . . . . .	10
2.7.	Cobertura de suelo en región del Biobío, actualización CONAF. . . . .	12
3.1.	Metodología para limpieza automática . . . . .	17
4.1.	Cambios al modificar clases (CONAF a C.R. Copernicus). . . . .	19
4.2.	División de grilla sobre la región del Biobío. . . . .	20
4.3.	Modelo Convolutiva Base. . . . .	21
4.4.	Matriz de confusión para modelo base con máximo <i>accuracy</i> . . . . .	22
4.5.	Matriz de confusión para modelo base con máximo <i>macro recall</i> . . . . .	22
4.6.	Ejemplo de red aplicada a detección de supernovas [33]. . . . .	23
4.7.	Modelo Convolutiva Rotacional. . . . .	23
4.8.	Matriz de confusión para modelo rotacional con máximo <i>accuracy</i> . . . . .	24
4.9.	Matriz de confusión para modelo rotacional con máximo <i>macro recall</i> . . . . .	24
4.10.	Modelo Convolutiva con inclusión de índices espectrales. . . . .	25
4.11.	Matriz de confusión para modelo inclusivo de índices espectrales con máximo <i>accuracy</i> . . . . .	26
4.12.	Matriz de confusión para modelo inclusivo de índices espectrales con máximo <i>macro recall</i> . . . . .	26
4.13.	Matriz de confusión para modelo rotacional utilizando dataset EuroSat. . . . .	27
4.14.	Matriz de confusión para modelo Random Forest entrenado con datos de CONAF. . . . .	28
4.15.	Imagen satelital junto a mapa de cobertura / uso de suelo CONAF. . . . .	29
4.16.	Diagrama de deslizamiento de ventana de 3x3, ilustrando la cantidad de predicciones por pixel. . . . .	29
4.17.	Demostración limpieza automática aplicada a clase <i>Artificial Forest</i> . . . . .	30
4.18.	Demostración limpieza automática aplicada a clase <i>Urban / Built Up</i> . . . . .	30
4.19.	Respuesta espectral de la clase <i>Agriculture</i> . . . . .	31
4.20.	Respuesta espectral de la clase <i>Artificial Forest</i> . . . . .	31
4.21.	Respuesta espectral de la clase <i>Herbaceous Vegetation</i> . . . . .	32
4.22.	Respuesta espectral de la clase <i>Natural Forest</i> . . . . .	32
4.23.	Respuesta espectral de la clase <i>Permanent Water Bodies</i> . . . . .	33
4.24.	Respuesta espectral de la clase <i>Urban / Built Up</i> . . . . .	33
4.25.	Comparación de las respuestas espectrales de la clase <i>Agriculture</i> y <i>Herbaceous Vegetation</i> . . . . .	34

4.26.	Comparación de las respuestas espectrales de la clase <i>Artificial Forest</i> y <i>Natural Forest</i> . . . . .	34
4.27.	Matriz de confusión para modelo entrenado con datos sin filtrar sobre datos filtrados. . . . .	35
4.28.	Matriz de confusión para modelo entrenado con datos filtrados mediante el método 1. . . . .	36
4.29.	Matriz de confusión para modelo entrenado con datos filtrados mediante el método 2. . . . .	36
4.30.	Matriz de confusión para modelo entrenado con datos sin filtrar sobre datos filtrados, con un umbral de certeza igual al 60 %. . . . .	37
4.31.	Matriz de confusión para modelo entrenado con datos filtrados mediante el método 1, con un umbral de certeza igual al 60 %. . . . .	38
4.32.	Matriz de confusión para modelo entrenado con datos filtrados mediante el método 2, con un umbral de certeza igual al 60 %. . . . .	38
4.33.	Relación entre cobertura, <i>accuracy</i> y certeza. . . . .	39
4.34.	Primer ejemplo aplicado a generación de mapas, modelo entrenado directamente con datos CONAF. . . . .	40
4.35.	Primer ejemplo aplicado a generación de mapas, modelo entrenado con datos limpios (método 2). . . . .	40
4.36.	Segundo ejemplo aplicado a generación de mapas, modelo entrenado directamente con datos CONAF . . . . .	41
4.37.	Segundo ejemplo aplicado a generación de mapas, modelo entrenado con datos limpios (método 2). . . . .	41
4.38.	Tercer ejemplo aplicado a generación de mapas, modelo entrenado directamente con datos CONAF. . . . .	42
4.39.	Tercer ejemplo aplicado a generación de mapas, modelo entrenado con datos limpios (método 2). . . . .	42
4.40.	Cuarto ejemplo aplicado a generación de mapas, modelo entrenado directamente con datos CONAF. . . . .	43
4.41.	Cuarto ejemplo aplicado a generación de mapas, modelo entrenado con datos limpios (método 2). . . . .	43
4.42.	Quinto ejemplo aplicado a generación de mapas, modelo entrenado directamente con datos CONAF. . . . .	44
4.43.	Quinto ejemplo aplicado a generación de mapas, modelo entrenado con datos limpios (método 2). . . . .	44
4.44.	Matriz de confusión para modelo CART. . . . .	45
4.45.	Matriz de confusión para modelo CART, con un umbral de certeza igual a 60 %. . . . .	45
4.46.	Matriz de confusión para modelo Gradient Boosted Trees. . . . .	46
4.47.	Matriz de confusión para modelo Gradient Boosted Trees, con un umbral de certeza igual a 60 %. . . . .	46
4.48.	Matriz de confusión para modelo Random Forest con índices espectrales. . . . .	47
4.49.	Matriz de confusión para modelo Random Forest con índices espectrales, utilizando un umbral de certeza igual a 60 %. . . . .	47
4.50.	Matriz de confusión para modelo perceptrón multicapas. . . . .	48
4.51.	Matriz de confusión para modelo perceptrón multicapas, con un umbral de certeza igual a 60 %. . . . .	48
5.1.	Porcentaje de píxeles clasificados correctamente para cada nivel de certeza. . . . .	50

A.1.	Comparación predicciones random forest común y gradient boosted, primer ejemplo. . . . .	56
A.2.	Comparación predicciones random forest común y gradient boosted, segundo ejemplo. . . . .	57
A.3.	Comparación predicciones random forest común y gradient boosted, tercer ejemplo.	57
A.4.	Comparación predicciones random forest común y gradient boosted, cuarto ejemplo. . . . .	58
A.5.	Comparación predicciones random forest común y gradient boosted, quinto ejemplo. . . . .	58



# 1 | Introducción

## 1.1. Mapas de cobertura de suelo

Durante las últimas décadas se han utilizado imágenes aéreas para monitorear la evolución de las poblaciones vegetativas en diversos ambientes, esto permite cuantificar el efecto de fenómenos como incendios, urbanización y el cambio climático en la superficie forestada. La efectividad de esta tarea ha sido facilitada por el análisis multiespectral, el cual permite diferenciar entre especies distintas de árboles e incluso detectar diferencias en madurez y nivel de estrés <sup>[1]</sup>.

### 1.1.1. Uso de mapas de cobertura de suelo

Dentro de los usos de mapas de cobertura de suelo se encuentra (pero no se limita a):

- **Planificación y gestión del territorio** Se ha utilizado data disponibilizada por el programa Copernicus para desarrollar un modelo evaluador de la disipación calórica a partir del grado de cobertura arbórea en contextos urbanos<sup>[2]</sup>, lo cual puede ayudar a planificar la cantidad adecuada de áreas verdes al expandir zonas urbanas.
- **Evaluación ambiental y conservación** La calidad de los mapas de cobertura de suelo está ligada a su capacidad ser utilizadas en evaluar el grado de recarga de aguas subterráneas<sup>[3]</sup>
- **Estudios de cambio climático y modelado** Los mapas de cobertura y uso de suelo juegan un rol central en evaluar el cambio climático<sup>[4]</sup>
- **Evaluación de riesgos naturales** El nivel de cambio entre mapas de cobertura de suelo realizados a lo largo de un periodo extenso de tiempo puede correlacionarse con el riesgo de inundación<sup>[5]</sup>

## 1.2. Sensores multiespectrales

Utilizando imágenes multiespectrales, se han creado mapas de cobertura de suelo logrando diferenciar entre tipos de bosque (hoja ancha, coníferos, mixtos, etc.) con un *accuracy* de más de 90 % utilizando tan solo cuatro bandas (Rojo, Verde, Azul y *NIR*)<sup>[6]</sup>.

## 1.3. El programa Copernicus

El programa Copernicus, de la Unión Europea, es un sistema complejo que incluye diversos satélites y sensores monitoreando la tierra cuyos datos son de libre acceso<sup>[7]</sup>. Para esta memoria se pretende utilizar las imágenes obtenidas a partir del programa Copernicus, específicamente las obtenidas con el satélite **Sentinel-2** de 13 bandas<sup>[8]</sup>.

## 1.4. Objetivos del trabajo de título

### 1.4.1. General

Desarrollar métodos escalables y replicables para la creación de algoritmos de aprendizaje de máquinas para la clasificación automática de cobertura de suelos a partir de imágenes de acceso libre del programa Copernicus.

Este trabajo de título involucra dos partes principales, diseño de un método para la adquisición de respuestas multiespectrales altamente representativas para cada clase y en segundo lugar la utilización de estos datos para el entrenamiento y evaluación de modelos basados en árboles de decisión.

### 1.4.2. Específicos

- Lograr la distinción entre al menos las siguientes clases:
  - Agricultura (Agriculture).
  - Bosque Artificial (Artificial Forest).
  - Vegetación herbácea (Herbaceous Vegetation).
  - Bosque Nativo (Natural Forest).
  - Cuerpos de Agua (Permanent Water Bodies).
  - Zona Urbana (Urban/ Built Up).
- Evaluar el impacto de la utilización de índices espectrales en las métricas resultantes de los modelos desarrollados.
- Adquirir respuestas espectrales características para las clases utilizadas.
- Evaluar el impacto de limpieza de datos automática en el desempeño de modelos basados en árboles de decisión.
- Realizar comparaciones en el desempeño de distintos modelos del grupo de árboles de decisión.

# 2 | Marco Teórico y Estado del Arte

## 2.1. Marco Teórico

### 2.1.1. Firmas Espectrales (F.E.)

Se conoce como firma espectral (ver figura 2.1) a la respuesta distintiva entre materiales al reflejar de una manera propia y específica los estímulos lumínicos <sup>[9]</sup>

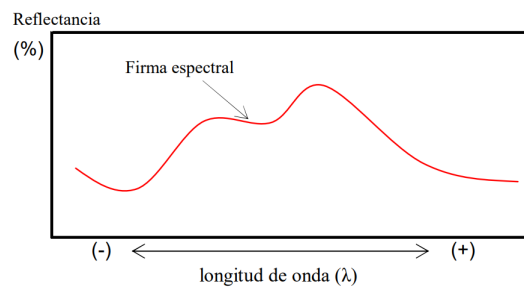


Figura 2.1: Representación gráfica de una F.E. cualquiera<sup>[9]</sup>.

Utilizando firmas espectrales, se puede distinguir entre distintos tipos de vegetación (ver figura 2.2). En el caso de Sentinel-2A no se habla de firma, sino respuesta espectral ya que no se tiene la caracterización continua a lo largo del espectro lumínico, aún así dicha respuesta podría ser utilizada para distinguir entre distintos tipos de cobertura de suelo.

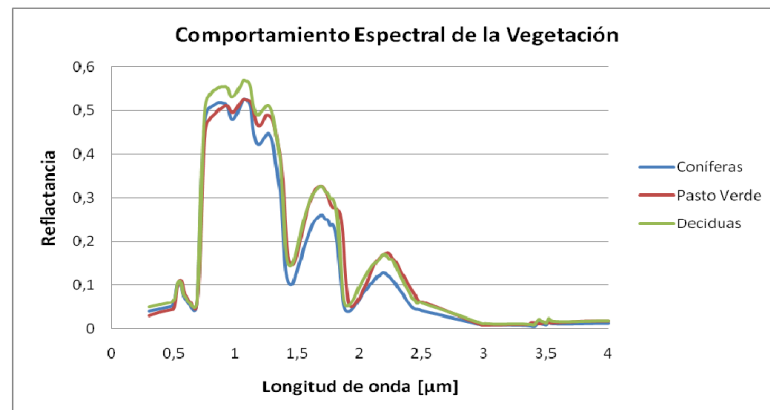


Figura 2.2: Variación de F.E. de acuerdo a tipo de vegetación<sup>[9]</sup>.

### 2.1.2. Índices Espectrales

Se conoce como índice espectral a la representación numérica de características propias de un objeto observado mediante sensores multispectrales, su utilización es variada, como algunos ejemplos tenemos:

- **Normalized Difference Vegetation Index (INDV):** Corresponde a un índice caracterizador de la salud y densidad de la vegetación. <sup>[10]</sup>
- **Normalized Burn Ratio (NBR):** Es un índice utilizado en la estimación de intensidad de incendios en áreas quemadas <sup>[11]</sup>
- **Change Vector (CV):** Captura el valor de cambio absoluto espectral entre dos fechas. <sup>[12]</sup>
- **Relative Change Vector Maximum (RCVMAX):** Captura el valor de cambio relativo espectral entre dos fechas. <sup>[12]</sup>

### 2.1.3. Centro Regional Copernicus

El centro regional Copernicus (CRC), tiene como objetivo difundir los beneficios del programa Copernicus dentro de la zona, promoviendo el uso de los datos satelitales provistos por dicho programa tanto como para uso público y privado. Además de esto, se aporta a la creación de un ecosistema de observación de la Tierra en Chile, buscando que la comunidad de usuarios comparta información, estando formado por los siguientes miembros: <sup>[13]</sup>

- ACTI.
- Biblioteca del Congreso.
- Centro de Innovación TI para Aplicaciones Sociales, U. de Santiago.
- Centro de Modelamiento Matemático (CMM), U. de Chile.
- Centro de Resiliencia del Clima (U. de Chile, U. de Concepción, U. Austral).
- ChileTec.
- Instituto Milenio de Oceanografía (U. de Concepción, PUCV, PUC).
- Laboratorio Nacional de Computación de Alto Rendimiento.
- Ministerio de Obras Públicas.
- Observatorio de Desastres Socio-Naturales (U. de Chile, SENAPRED, SHOA, DGAC, Sernageomin, CONAF).
- SENAPRED.
- Programa de Riesgo Sísmico, U. de Chile.

El desarrollo de la presente memoria haciendo uso de los datos disponibilizados por el programa Copernicus se alinea con el interés del Centro Regional de promover el uso de los datos y difundir los beneficios del programa, en consecuencia se ha recibido un apoyo continuo por parte de integrantes del CRC, quienes disponibilizaron el uso de sus instalaciones en el Centro de Modelamiento Matemático para el desarrollo de la memoria.

### 2.1.4. Características del satélite (Sentinel-2)

Sentinel-2 es un satélite de alta tasa de revisita (5 días en operación conjunta entre Sentinel-2A y Sentinel-2B), alta resolución espacial (10 a 60m dependiendo de la banda espectral), y 13 bandas espectrales desde 433nm a 2280nm, tanto la longitud de onda central como el ancho de banda asociado a cada banda espectral se pueden encontrar en la tabla 2.1.

Tabla 2.1: Descripción de bandas del satélite Sentinel-2.

<b>Banda</b>	<b>Longitud de onda central [nm]</b>	<b>Ancho de banda [nm]</b>
1	433	20
2	490	65
3	560	35
4	665	30
5	705	15
6	740	15
7	783	20
8	842	115
8a	865	20
9	945	20
10	1380	30
11	1610	90
12	2190	180

## 2.1.5. Redes Neuronales

### 2.1.5.1. Perceptrón Multicapa

El perceptrón es la unidad básica de las redes neuronales, también es conocido como *neurona* o *máquina lineal*. Se le denomina perceptrón multicapa a un algoritmo de aprendizaje de máquinas formado por múltiples perceptrones, en este tipo de estructura existen capas de neuronas, donde la segunda capa recibe como entrada el resultante de los perceptrones que forman la primera capa.

Cada perceptrón  $i$  cuenta con 4 características principales, el tamaño de entrada  $n_i$ , el vector de pesos  $w_i$ , formado por  $n$  valores  $w_{ij}$ , el valor de sesgo  $\theta_i$  y su función de activación  $f_a(x)$ , la cual debe ser diferenciable y creciente. Su valor resultante ( $o_i$ ) para un vector de entradas  $x_i$ , formado por  $n$  valores  $x_{ij}$  puede ser descrito utilizando la ecuación 2.1 [14].

$$o_i = f\left(\theta_i + \sum_{j=1}^n (w_{ij} \cdot x_{ij})\right) \quad (2.1)$$

Los valores de  $w_i$  y de  $\theta_i$  pueden ser actualizados mediante entrenamiento, proceso en el cual se logra ajustar dichos parámetros minimizando el error (entre predicciones y resultados deseados) para realizar tareas tales como regresión o clasificación multiclase [14].

### 2.1.5.2. Red Neuronal Convolutiva

Las redes neuronales convolucionales se diferencian del perceptrón multicapas en estar concebidas para su aplicación en imágenes, por lo cual en vez de recibir un vector como entrada, se recibe un tensor con un alto, un ancho y una profundidad (3 para RGB).

Este tipo de redes cuentan con tres tipos de capas: las capas convolucionales (actúan sobre un *kernel*, generalmente de 3x3 píxeles), las capas de *pooling* (disminuyen la dimensionalidad de la entrada para la capa siguiente) y las capas *fully-connected*, equivalentes al perceptrón multicapas. Generalmente se utiliza una estructura intercalada entre capas convolucionales y de *pooling*, terminando con capas *fully-connected*, mientras más complejas sean las formas a distinguir se recomienda aumentar la cantidad de capas convolucionales entre capas de *pooling* [15].

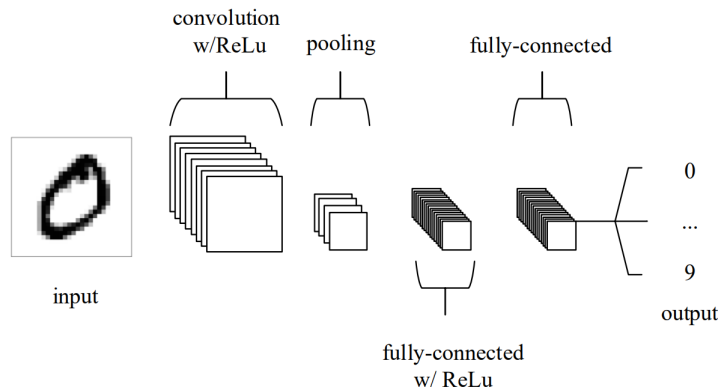


Figura 2.3: Red Convolutiva Simple [15].

## 2.1.6. Árboles de clasificación

### 2.1.6.1. Árbol de clasificación

Un árbol de clasificación corresponde a un algoritmo utilizado para distinguir entre distintos tipos de vectores de datos a través de una serie de *nodos* en los cuales se discrimina de acuerdo a un umbral específico para uno de los valores de dicho vector. Un ejemplo ilustrativo de este tipo de algoritmo se puede apreciar a continuación, en el cual un árbol de clasificación se utiliza para distinguir entre tipos de iris (flor) a partir del ancho y largo de cada pétalo [16]:

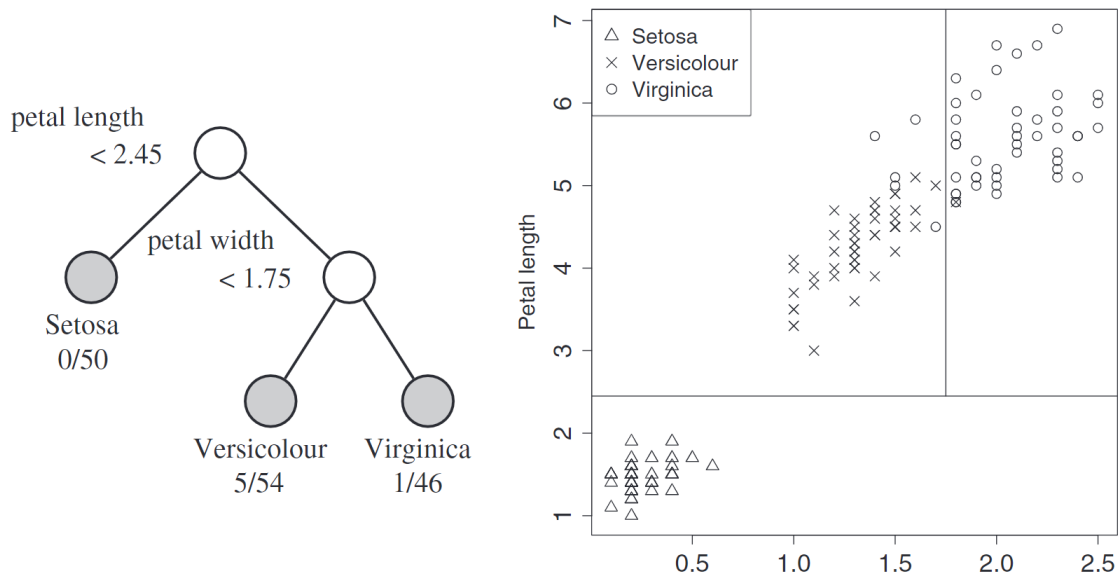


Figura 2.4: Clasificación de tipo de iris utilizando árbol de clasificación [16].

Existen distintos tipos de árboles de clasificación, los cuales son creados utilizando diferentes parámetros para determinar la cantidad de nodos, las variables y umbrales utilizadas, entre otros. Algunos tipos de árboles son CART, CHAID, CRUISE, GUIDE entre otros [16].

### 2.1.6.2. Random Forest

Es posible mejorar el desempeño de un algoritmo de clasificación al utilizar una cantidad de árboles de clasificación distintos entre sí, tomando como clasificación final a la clasificación más popular entre todos los árboles utilizados. La creación de árboles distintos puede ser implementada con diferentes métodos, tal como seleccionar como la partición para cada nodo una variable aleatoria entre las  $k$  mejores variables disponibles o crear cada árbol utilizando distintas particiones aleatorias del conjunto de entrenamiento. [17]

### 2.1.6.3. Gradient Boosted Trees

Éste tipo de modelo utiliza árboles de decisión como base, con la particularidad que cada nuevo árbol de decisión es creado y entrenado con la tarea clasificar aquellos datos que no fueron clasificados correctamente por el árbol anterior [18].

## 2.1.7. Métricas de evaluación

Dentro de un problema de clasificación, múltiples métricas de evaluación pueden ser descritas utilizando tan solo cuatro parámetros:

- **True Positives (TP):** corresponde a la cantidad de ejemplos clasificados correctamente dentro de la clase de interés.
- **True Negatives (TN):** corresponde a la cantidad de ejemplos clasificados correctamente fuera de la clase de interés.
- **False Positives (FP):** corresponde a aquellos ejemplos que incorrectamente fueron clasificados dentro de la clase de interés.
- **False Negatives (FN):** corresponde a aquellos ejemplos que incorrectamente fueron clasificados fuera de la clase de interés.

### 2.1.7.1. Matriz de confusión:

Corresponde a una tabla con dos ejes, uno correspondiente a la clase real u observada y otro a las predicciones, tomando como eje  $x$  las predicciones y al eje  $y$  las clases reales, en el bloque  $b_{ij}$  se representa la cantidad de ejemplos correspondientes a la clase  $y_j$  clasificados en la clase  $x_i$ , tal como se puede ver en la siguiente figura:

		PREDICTED classification				Total
		Classes	a	b	c	
ACTUAL classification	a	6	0	1	2	9
	b	3	9	1	1	14
	c	1	0	10	2	13
	d	1	2	1	12	16
Total		11	11	13	17	52

Figura 2.5: Ejemplo de matriz de confusión no normalizada. [19]

### 2.1.7.2. Accuracy

Puede ser expresada por la fórmula  $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ , e interpretada simplemente como la proporción entre ejemplos correctamente clasificados y la cantidad de ejemplos totales [19]. Para la matriz de la figura 2.5 y la clase  $a$  se detalla el cálculo en la ecuación 2.2.

$$acc_a = \frac{6 + (9 + 1 + 1 + 0 + 10 + 2 + 2 + 1 + 12)}{6 + (9 + 1 + 1 + 0 + 10 + 2 + 2 + 1 + 12) + (3 + 1 + 1) + (0 + 1 + 2)} = 0.846 \quad (2.2)$$

Es fácil demostrar que debido a propiedades de la asociatividad de la suma ( $a + [b + c] = [a + b] + c$ ) el  $accuracy$  de cada clase es igual entre si, por lo que generalmente se habla de simplemente  $accuracy$ , y no  $accuracy_{clase}$  [19].



### 2.1.7.3. Precision

Al tener el mismo resultado para todas las clases, la *accuracy* no es una métrica óptima para evaluar el desempeño de problemas con clases no-balanceadas [19]. Otra métrica de posible interés es la *precision* [19], definida como  $precision = \frac{TP}{TP+FP}$ , nuevamente podemos utilizar de ejemplo la clase *a* (ver ecuación 2.3).

$$precision_a = \frac{6}{6 + (3 + 1 + 1)} = 0.545 \quad (2.3)$$

Si la matriz de confusión es normalizada por columnas, los valores de la diagonal corresponden a la *precision* por clase, como se puede apreciar en la tabla 2.2.

Tabla 2.2: Matriz de confusión normalizada por columnas.

		Predicted			
		a	b	c	d
Actual	Classes				
	a	0.55	0	0.08	0.12
	b	0.27	0.81	0.09	0.06
	c	0.09	0	0.77	0.12
d	0.09	2	0.08	0.71	

Se puede obtener una métrica de *macro precision* al sumar los valores de la diagonal y dividiendo por clase (*macro precision*= 0.71, métrica que le asigna la misma importancia a todas las clases, o ponderando por la cantidad de ejemplos reales por clase (*weighted precision*= 0.73), lo cual asigna mayor importancia a las clases más populares. Esta métrica indica qué tan confiable es un modelo para clasificaciones positivas [19].

### 2.1.7.4. Recall

El *recall* indica la capacidad de un modelo predictivo para hallar cada clase de interés [19], se puede calcular como  $\frac{TP}{TP+FN}$ , para la clase *a* puede ser calculado como  $\frac{6}{6+(0+1+2)} = 0.67$ , de manera similar al *precision*, si se normaliza una matriz de confusión por filas, la diagonal corresponde al *recall* por clase (ver tabla 2.3), también existiendo una versión *macro* y *weighted* respectivamente.

Tabla 2.3: Matriz de confusión normalizada por filas.

		Predicted			
		a	b	c	d
Actual	Classes				
	a	0.67	0	0.11	0.22
	b	0.21	0.64	0.7	0.7
	c	0.08	0	0.77	0.15
d	0.06	0.13	0.06	0.75	

En este caso, el *macro recall*= 0.71.

### 2.1.7.5. F1

El *F1* es una métrica resumen del *precision* y el *recall* calculada como  $\frac{2 \cdot precision \cdot recall}{precision + recall}$  [19].

## 2.2. Estado del Arte

### 2.2.1. Uso del satélite Sentinel-2

Se ha utilizado el satélite Sentinel-2 para monitoreo del desarrollo urbano, la agricultura y bosques, enfocándonos en estos últimos, se compila una pequeña lista (ver tabla 2.4) de estudios recientes que utilizan la misma métrica de desempeño<sup>[20]</sup>.

Tabla 2.4: Comparación de *accuracy* logrado para diferentes aplicaciones de machine learning sobre datos Sentinel-2<sup>[20]</sup>.

Aplicación Específica	País	Métodos	Accuracy
Extensión de bosque.	Polonia, China, Burkina Faso, Sudáfrica, Madagascar, Zimbabwe y Bulgaria.	ML, cloud computing.	80-90 %
Tipos de bosque.	Italia, Ghana, Sudáfrica y Togo.	Linear discrimination analysis, índices espectrales y ML.	88-90 %
Identificación de especies.	Alemania e Italia.	OBIA-RF, Regresión paso a paso.	65-76 %
Productividad.	Alemania, Sudáfrica y el sur de África.	ML (Random Forest), Invertible Forest Reflectance Model.	90-92 %
Mapeo humedales.	China, Canadá, Sudáfrica, Senegal y Ghana.	ML, Google Earth Engine, OBIA.	83-90 %
Incendios forestales.	Isla Madeira, Bulgaria, RD del Congo y África en general.	Algoritmo propio, Active fire products with SAR data.	80-89 %
Mapeo zonas áridas.	Alemania y el sur de África.	Clasificación sub-píxel, simulaciones BiomeBGC.	82 %
Renovación de bosques.	Brasil y Polonia.	SVM, RF, OBIA.	90-97 %
Salud forestal.	Polonia.	ML.	75-78 %

### 2.2.2. Uso de imágenes hiperespectrales

Por otro lado, utilizando técnicas como Support Vector Machine (SVM), RF y redes neuronales (NN) sobre imágenes espectrales con una mayor resolución espectral (188 bandas) combinadas con información adquirida mediante un sensor LIDAR se ha logrado distinguir entre 21 especies arbóreas dominantes del parque Oncol, en la región de la Araucanía con un 88 % de precisión.<sup>[21]</sup>



Figura 2.6: Parque Oncol<sup>[9]</sup>.

## 2.2.3. Mapas de cobertura de suelo en Chile

### 2.2.3.1. CONAF

CONAF pone a disposición del público general *catastros vegetacionales*, los cuales son mapas de cobertura de suelo altamente especializados en la diferenciación entre tipos de bosques.<sup>[22]</sup> Las clases utilizadas son propias de la institución y se hallan ordenadas en niveles (ver tabla 2.5).

Tabla 2.5: Clasificaciones de cobertura de suelo CONAF.

Nivel 1	Nivel 2	Nivel 3
Áreas Urbanas e Industriales.	Ciudades, pueblos y zonas industriales. Minería industrial.	
Tierrenos agrícolas	Terrenos de uso agrícola. Rotación cultivo-pradera.	
Praderas y matorrales.	Praderas Matorral-pradera Matorral Matorral arborescente Plantación de arbustos	
Bosques.	Plantaciones Bosque nativo Bosque mixto	Pino, eucalipto, etc. Araucaria, ciprés lenga, etc.
Humedales.	Marismas herbáceas Vegas	
	Otros terrenos húmedos	
Áreas desprovistas de vegetación	Playas y dunas	
	Afloramientos rocosos	
	Áreas sobre límite vegetación	
	Corridas de lava y escoriales	
	Derrumbes sin vegetación	
	Otros terrenos sin vegetación	
	Cajas de ríos	
Nieves eternas y glaciares	Nieves	
Cuerpos de agua	Ríos	
	Lagos, lagunas, embalses y tranques.	

Dichos catastros fueron validados experimentalmente, la validación asociada al método dificulta la posibilidad de realizar los catastros manualmente resultando en diferencia en el año de publicación de los catastros de cada región, tal como se muestra en la tabla 2.6.

Tabla 2.6: Años de catastros CONAF.

Región:	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIV	XV	XVI	RM
Abreviatura Nacional	TA	AN	AT	CO	VA	LI	ML	BI	AR	LL	AI	MA	LR	AP	NB	RM
Año Catastro CONAF	2016	2018*	2018*	2014	2013	2013	2016	2015	2014	2013	2011	2017*	2014	2015	2015	2013

Los años marcados con un \* indican que no se halla disponible el catastro original, pero se pueden utilizar mapas obtenidos a partir de detección de cambios

Además de los catastros, CONAF realiza un sistema de monitoreo basado en la detección de cambios, el cual puede ser utilizado para obtener mapas de cobertura de suelos para años posteriores a la publicación del catastro original, estas actualizaciones se realizaron hasta el año 2021<sup>[23]</sup>. El sistema de actualización está basado en un algoritmo de detección de cambios a partir de índices espectrales<sup>[24]</sup>, utilizando los índices INDV, NBR, CV y RCVMAX<sup>[12]</sup>.

En la figura 2.7 se muestra el mapa de cobertura de suelo para la región del Biobío obtenido a través de la actualización del catastro original (en particular se escogió el 2019 de manera arbitraria).

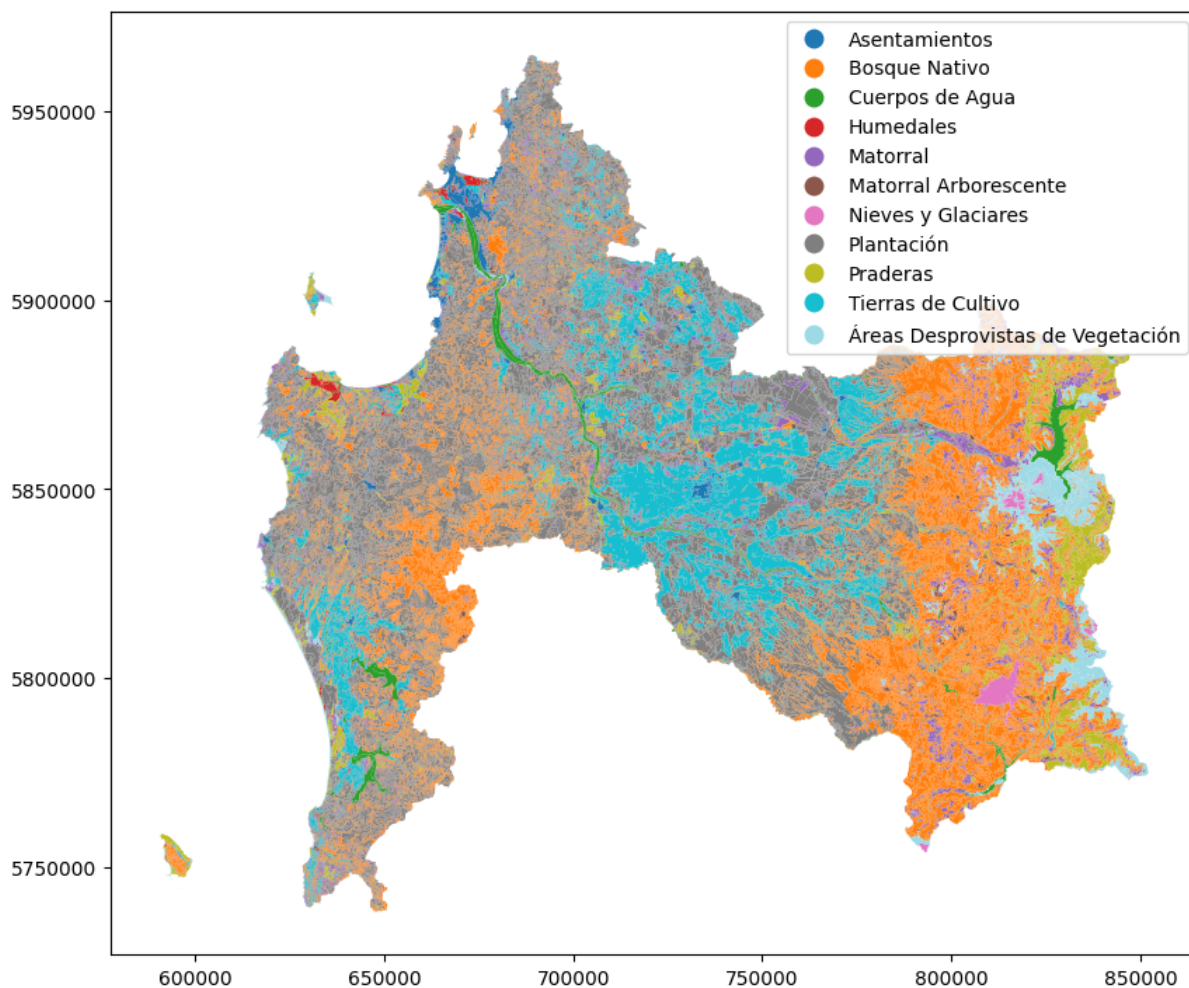


Figura 2.7: Cobertura de suelo en región del Biobío, actualización CONAF.

### 2.2.3.2. Copernicus

El programa copernicus pone a disposición general productos globales de cobertura y uso de suelo con resolución de 100m<sup>2</sup>, utilizando sus propias clases<sup>[25]</sup>:

Tabla 2.7: Clasificaciones de cobertura de suelo Copernicus.

Nivel 1	Nivel 2	Nivel 3
Bosques	Hoja de aguja siempreverde	Abierto Cerrado
	Hoja ancha siempreverde	Abierto Cerrado
	Hoja de aguja caducifolia	Abierto Cerrado
	Hoja ancha caducifolia	Abierto Cerrado
	Bosque mixto	Abierto Cerrado
	Bosque intacto	Abierto Cerrado
	Bosque secundario	
	Plantación	
Arbustos	Siempreverde	
	Caducifolios	
Vegetación herbácea	Pastizales	
	Pastizales naturales	
	Pastizal africano	
	Sabana	
Vegetación controlada/agricultura	Cultivos irrigados	
	Cultivos llovidos	
	Campos	
	Cultivos permanentes	
	Cultivos en hileras	
	Cultivos de ciclo largo-corto	
	Cultivos con multiples ciclos	
Zona urbana/construida		
Vegetación escasa		
Hielos y nieves		
Aguas abiertas		
Humedales		

Dichas clases se encuentran detalladas en la tabla 2.7, en la cual se puede apreciar un foco en productividad al existir un grado de especificidad mayor en agricultura.

# 3 | Metodología

## 3.1. Propuesta metodológica

El uso del satélite Sentinel-2 en conjunto con técnicas de aprendizaje de máquinas para generar mapas de cobertura de suelos entrega resultados satisfactorios (*overall accuracy* > 80 %), siendo utilizado por ventajas tales como alta resolución espacial, temporal y acceso a bandas NIR, lo cual tiene múltiples aplicaciones <sup>[26]</sup>.

### 3.1.1. Datos a utilizar

En primer lugar, se accede a los mapas de cobertura de suelo generados por CONAF, en particular a los catastros correspondientes a años en los cuales Sentinel-2 ya haya estado en órbita. Estos mapas son de libre descarga y pueden ser adquiridos en formato vectorial shapefile (*.shp*), que almacena la geometría y atributos de características geográficas, generalmente con forma de puntos, líneas y polígonos. <sup>[22]</sup> <sup>[23]</sup>. En segundo lugar, se accede a los registros históricos de las imágenes tomadas con el satélite Sentinel-2, los cuales pueden ser descargados en formato *.jp2* y se tiene acceso a múltiples fotografías por año<sup>[27]</sup>.

Para demostrar la capacidad de aprendizaje de la red y evaluar su capacidad de distinción entre las clases propuestas no es necesaria la utilización de datos correspondientes a todos los años disponibles ni a todas las estaciones, así como tampoco el balanceo de datos utilizando áreas correspondientes a otras regiones geográficas.

#### 3.1.1.1. Datos auxiliares para evaluación de los algoritmos

Se considera prudente evaluar cada arquitectura sobre dos bases de datos, una creada del modo mencionado anteriormente y otra a partir de una base de datos conocida y utilizada a nivel global. Se utiliza la base de datos EuroSat luego de hacer una disminución de resolución para obtener píxeles del mismo tamaño.

### 3.1.2. Zona geográfica del estudio

Se propone utilizar una zona geográfica que cuente con cantidad significativa de bosque nativo y artificial, dependiendo principalmente de qué tan reciente sea la actualización al catastro de uso de suelo para dicha región geográfica por parte de CONAF. Se tiene acceso a mapas de cambio de cobertura de suelo para la región del Biobío en los años 2015, 2017, 2019 Y 2021 disponibilizados por CONAF <sup>[23]</sup> y también se tiene acceso a imágenes multiespectrales tomadas por Sentinel-2A desde diciembre del 2016 hasta la fecha<sup>[27]</sup>, al existir suficiente representación de bosques nativos y artificiales<sup>[28]</sup> se considera que la región del Biobío es idónea para la realización de la presente memoria.

### 3.1.3. Técnicas a comparar

Debido a su desempeño en el estado del arte[2.4], se entrena una red Random Forest para clasificación pixel a pixel y una red neuronal convolucional. Además se desea incorporar dos técnicas extra por cada red:

- **Incorporación de índices espectrales.**
- **Utilización de redes en paralelo aplicadas sobre rotaciones del input.**

### 3.1.4. Clases a identificar

Se desea clasificar entre las siguientes clases (clases propuestas recientemente por el centro regional de Copernicus en Chile), sujeto a disponibilidad y grado de representación de dicho tipo de cobertura/uso de suelo sobre el área de estudio:

- Natural Forest.
- Artificial Forest.
- Desert.
- Shrubs.
- Herbaceous vegetation.
- Bare / sparse vegetation.
- Agriculture (cropland).
- Urban / built up.
- Snow and Ice.
- Permanent water bodies.
- Open sea.
- Unclassified/uncertain.

### 3.1.5. Tratamiento de datos

Pese a los múltiples beneficios que ofrece trabajar con las imágenes tomadas con Sentinel-2, se deben tratar dichas imágenes de modo de contrarrestar algunas de sus características no deseables<sup>[26]</sup>

#### 3.1.5.1. Corrección geométrica

Se deben tratar las imágenes multispectrales y el catastro de cobertura de suelo de modo de que ambos muestren a la Región del Biobío con la misma estructura geométrica y proyección geográfica, esto es de suma importancia ya que se requiere la superposición de las imágenes multispectrales con los catastros de cobertura de suelo para poder generar los conjuntos de entrenamiento, validación y prueba. Para esto, se propone utilizar solamente el área de la región ubicada al oeste del meridiano 72, correspondiente al sistema de referencia EPSG:32718, ya que éste es en el cual se presentan los datos de CONAF. De este modo, no es necesario aplicar ningún tipo de reproyección.

### 3.1.5.2. Eliminación de nubes

La aparición de nubes en una imagen satelital impide la correcta caracterización de la cobertura de suelo bajo dicha nube, por lo tanto es deseable procesar los datos minimizando la cantidad de nubes en cada imagen. Para ello, una alternativa efectiva consiste en la adquisición de múltiples imágenes con baja nubosidad y realizar el cálculo de la mediana para cada pixel por separado.

### 3.1.6. Uso de información complementaria

Anteriormente se mencionó la posibilidad de incorporar índices espectrales, modificando la arquitectura de la red.

#### 3.1.6.1. Uso de índices espectrales

Se hará uso de diversos índices espectrales relacionados a las clases a identificar, en particular se utilizará:

- **NDVI** Corresponde a un índice caracterizador de la salud y densidad de la vegetación.<sup>[10]</sup>, un alto nivel de este índice podría ayudar a diferenciar zonas donde hay vegetación de aquellas en las que no.
- **NDBI** Corresponde a un índice correlacionado con el grado en el que un sector se halla edificado<sup>[29]</sup>, un alto nivel de este índice permitiría identificar con mayor facilidad el área urbana.
- **NDBAI** Corresponde a un índice empleado para identificar tan vacía se halla un área<sup>[30]</sup>, esto permitiría asegurar que un sector no debería ser etiquetado dentro de ninguna de las clases de interés
- **MuWI** Corresponde a un índice diseñado para mapeo de cuerpos de agua a partir de las imágenes del Sentinel-2<sup>[31]</sup>, naturalmente, este índice podría ser útil en identificar cuerpos de agua.

### 3.1.7. Generación de datos para entrenar y evaluar los distintos algoritmos.

En primer lugar se debe realizar una reasignación de clases, modificando las etiquetas de cada polígono del archivo *shapefile* del modo detallado en la tabla 4.1.

Posterior a la reasignación de clases, se descargan 456 imágenes multiespectrales de 3870m x 4360m (total de 769417.92ha), comprendiendo información entre el 21 de diciembre, 2018 y 20 de marzo, 2019 dentro de la Región del Biobío.

La resolución espacial de dichas imágenes es tal que cada pixel es equivalente a 20m x 20m, a partir del catastro de uso de suelos dispuestos por CONAF para las mismas zonas geográficas se obtienen ventanas de 32x32 pixeles representando cada clase, las cuales se utilizan para entrenar modelos de redes convolucionales, y además con la totalidad de los pixeles disponibles (sin generación de ventanas) se entrena un algoritmo del tipo Random Forest.



### 3.1.7.1. Limpieza automática de datos

Se obtienen las predicciones de la red convolucional mediante un método de ventanas deslizantes al ejecutarse sobre las 456 imágenes mencionadas previamente. Aquellos píxeles donde existe un acuerdo entre la predicción y el catastro de CONAF se denominan como píxeles *altamente representativos*, tal como ilustra el diagrama mostrado en la figura 3.1.

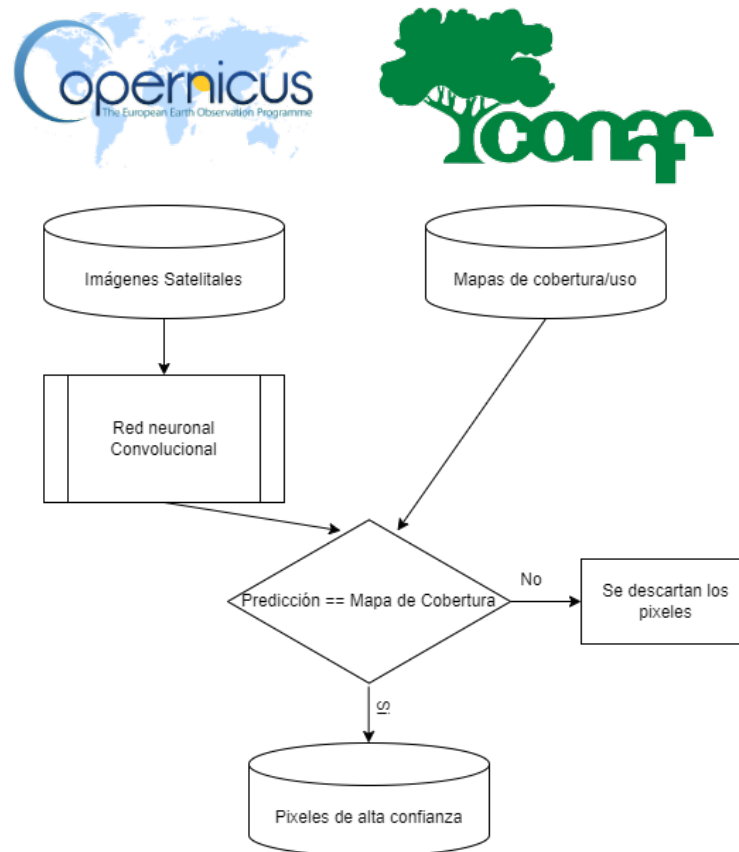


Figura 3.1: Metodología para limpieza automática

Dichos píxeles altamente representativos se utilizan para entrenar un segundo modelo del tipo RF, el cual permite cuantificar la efectividad de la limpieza automática en aumentar el desempeño de algoritmos de clasificación, además, se caracteriza cada una de las clases de acuerdo a sus respuestas espectrales y se finaliza con el entrenamiento y evaluación de otros tipos de algoritmos de aprendizaje de máquinas (CART, Gradient Boosted Decision Trees, y perceptrón multicapas).

# 4 | Contribución y Resultados

## 4.1. Contribución

CONAF ha puesto a disposición catastros de cobertura de suelo, pero como se puede ver en la tabla 2.6, no se tiene certeza de que vaya a haber información reciente, siendo los últimos mapas actualizaciones de los catastros originales y solo generados hasta el año 2021<sup>[23]</sup>, legalmente, CONAF solo se halla obligado a realizar nuevos catastros cada 10 años (artículo 4, ley número 20.283<sup>[32]</sup>).

Por otro lado, Copernicus disponibilizó catastros de cobertura de suelo entre los años 2015-2019, con resolución de 100m<sup>2</sup> y sin la distinción deseable entre bosque nativo y artificial

Entendiendo lo anterior, se tiene que la posibilidad generar automáticamente mapas de cobertura de suelo todos los años con una resolución de 20 m<sup>2</sup> correspondería a una mejora significativa respecto a la disponibilidad actual de mapas de cobertura de suelo en Chile.

### 4.1.1. Entregables esperados

Como entregables se encuentran las matrices de confusión resultantes al probar el desempeño de los algoritmos desarrollados, así como tablas en las cuales se resume sus métricas principales (*accuracy*, *recall*, *precision* y *F1-Score*), gráficos representando la respuesta espectral de cada clase trabajada y 5 mapas demostrativos del desempeño de cada modelo..

Debido a limitaciones de la librería empleada para la creación de algoritmos de, no se emplea validación cruzada en el entrenamiento de ninguno de los modelos basados en árboles de decisión, además, se observó experimentalmente una muy baja variación en desempeño de los modelos a la hora de variar la *semilla aleatoria* asociada a cada momento (varianza < % 1 en *macro-recall* y *macro-precision* en modelo Random Forest), por lo que solo se muestran los resultados para un modelo por arquitectura generada utilizando la librería *tensorflow\_decision\_forests*.

Además de esto, se entrega cada modelo entrenado (arquitectura y pesos), instrucciones precisas de cómo utilizarlos y el código mediante el cual se llevó a cabo el entrenamiento y la generación de los datos.

## 4.2. Resultados

### 4.2.1. Adquisición y visualización de datos

#### 4.2.1.1. Catastro de cobertura de suelos (CONAF)

Se adquiere tanto el catastro de cobertura de suelos y sus actualizaciones para la zona de interés (región del Biobío) para los años 2017, 2019 y 2021, tal como se puede apreciar en la figura 2.7.

#### 4.2.1.2. Transformación de las clases de uso/cobertura de suelo

En primer lugar se establece una relación entre las sub-clases utilizadas por CONAF y las clases propuestas por el centro regional de Copernicus (ver tabla 4.1).

Tabla 4.1: Relación entre clases utilizadas por CONAF y clases propuestas.

CONAF	Propuesta Copernicus
Asentamientos	Urban / Built Up
Bosque Nativo	Natural Forest
Cuerpos de Agua	Permanent Water Bodies
Humedales	Unclassified*
Matorral	Shrubs
Matorral Arborecente	Shrubs
Nieves y Glaciares	Snow and Ice
Plantación	Artificial Forest
Praderas	Herbaceous Vegetation
Tierras de Cultivo	Agriculture
Áreas Desprovistas de Vegetación	Bare / Sparse Vegetation

La reasignación de clases vuelve necesaria la agrupación de las clases ‘Matorral’ y ‘Matorral Arborecente’ en una sola (‘Shrubs’), así como también la obviación de la clase ‘Humedales’. A partir de esta relación entre clases, se modifican los archivos proporcionados por CONAF, obteniendo una variación apreciable en la figura 4.1.

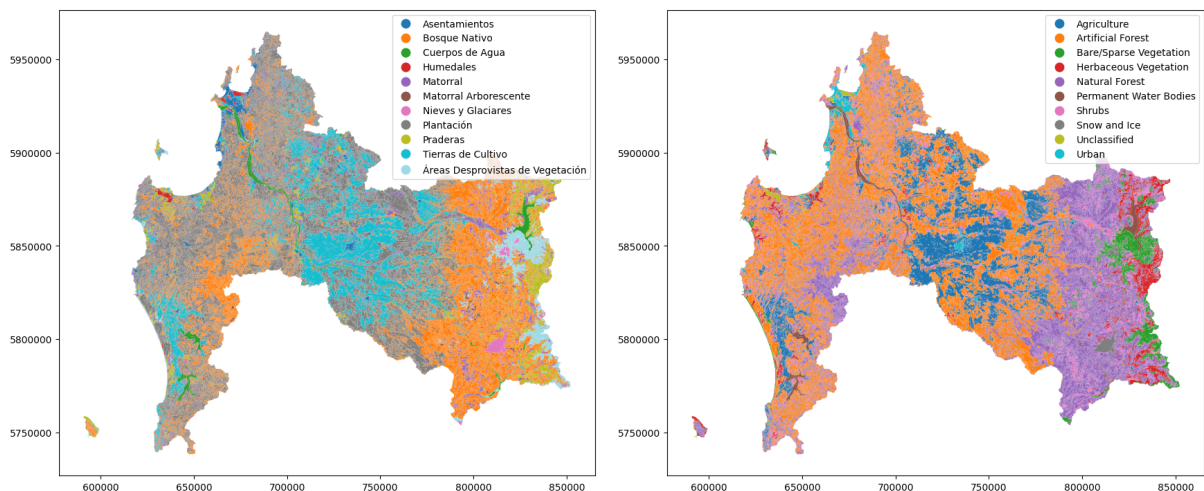


Figura 4.1: Cambios al modificar clases (CONAF a C.R. Copernicus).

### 4.2.1.3. Subdivisión del área de interés en bloques

Para facilitar la adquisición de datos y la compatibilidad entre las imágenes satelitales con el catastro de uso de suelos, se subdivide la región en múltiples bloques de interés, siendo utilizados aquellos que no sobrepasan los límites de la región ni se hallen al este del meridiano 72, evitando reproyecciones. En la figura 4.2 se muestra la división correspondiente, con una línea roja marcando el meridiano 72, límite divisorio entre el sistema de proyección EPSG:32718 a la izquierda y el sistema EPSG:32719 a la derecha.

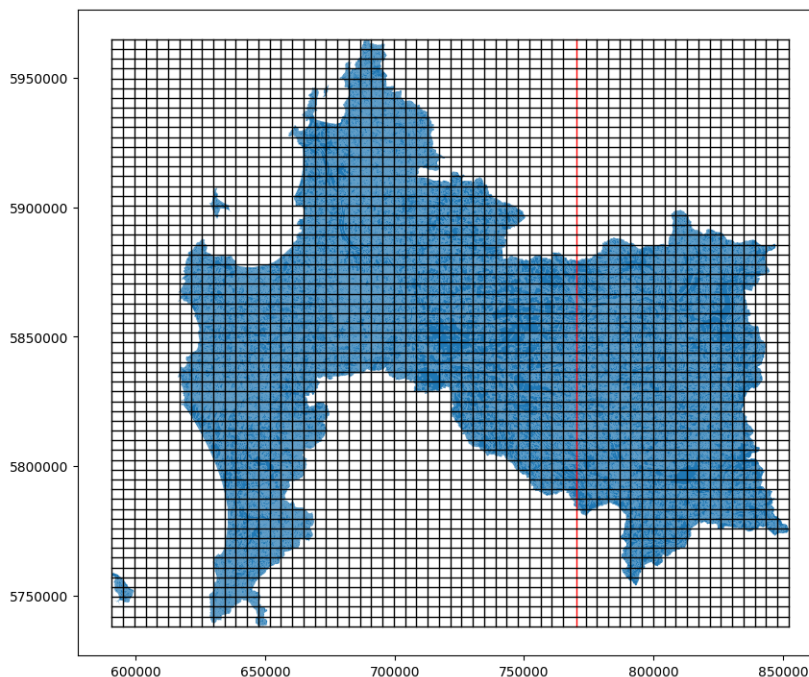


Figura 4.2: División de grilla sobre la región del Biobío.

### 4.2.1.4. Adquisición de las imágenes espectrales para bloques de interés

Utilizando la API *planetary computer* para descargar datos de Copernicus, se adquieren 456 imágenes correspondientes a bloques de interés, además se rasterizan los mapas asociados a las 456 secciones correspondientes en la actualización de cobertura y uso de suelos proporcionada por CONAF, estableciéndose como etiquetas de la cobertura de suelo para cada pixel disponible.

### 4.2.1.5. Creación de ventanas representativas para cada clase

Se rasterizan los mapas de cobertura y uso de suelos puestos a disposición por CONAF, luego se genera una grilla similar a la de la figura 2.7 para cada bloque, siendo el tamaño de cada ventana igual a 32x32 píxeles. Aquellas ventanas con una cantidad mayor al 70% de píxeles correspondientes a la misma clase (> 716 píxeles) se las guarda en una carpeta asociada a su clase en Google Drive. Paralelamente, se utilizan los mapas previamente rasterizados para entrenar y evaluar un modelo del tipo Random Forest, sin necesidad de generar ventanas.

## 4.2.2. Creación, entrenamiento y evaluación de algoritmos

Para cada arquitectura de red presentada, se entrena el modelo 30 veces y se seleccionan como modelos de interés aquellos con máximo *accuracy* o *macro-recall* en el conjunto de prueba. Se utiliza el optimizador *Adam* con un learning rate de 0.001, minimizando la pérdida *Categorical Cross-Entropy* en el conjunto de validación.

### 4.2.2.1. Red convolucional base (CNN)

Se utiliza el siguiente diseño de red neuronal como base para las otras dos arquitecturas probadas, tomando como entrada ventanas de 32x32 píxeles con 12 canales, cuenta con 3 capas convolucionales seguidas de capas *maxpool* y luego de esto se continúa hacia un *flatten*, iniciando la sección *fully connected*, conformada por dos capas *Dense* con un *Dropout* entre ellas. Se detalla la arquitectura de la red en la figura 4.3.

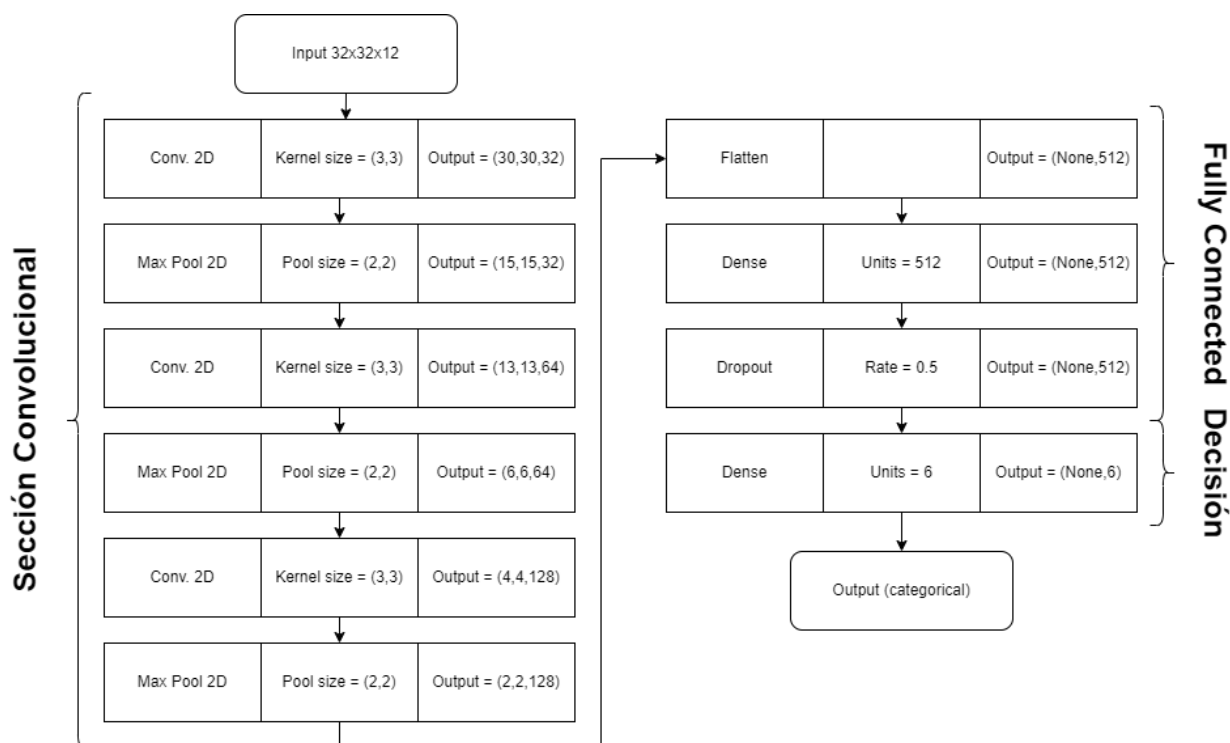


Figura 4.3: Modelo Convolucional Base.

Para cada una de las redes convolucionales probadas, se muestra la matriz de confusión para el modelo con máximo *accuracy* a un solo nivel y *macro-recall*, así como también se presenta en una tabla resumen los *accuracies* a primer y segundo nivel, y las métricas *macro* de *precision*, *recall* y *F1*.

Tabla 4.2: Resultados para modelo convolucional base.

Modelo	Acc. a un nivel [%]	Acc. a dos niveles [%]	<i>Macro Precision</i> [%]	<i>Macro Recall</i> [%]	<i>Macro F1</i> [%]
Base, máxima <i>accuracy</i>	62	80	63	60	58
Base, máximo <i>recall</i>	59	83	65	61	60

Se pueden encontrar las matrices de confusión para el modelo base con el máximo *accuracy* y *macro recall* en las figuras 4.4 y 4.5 respectivamente.

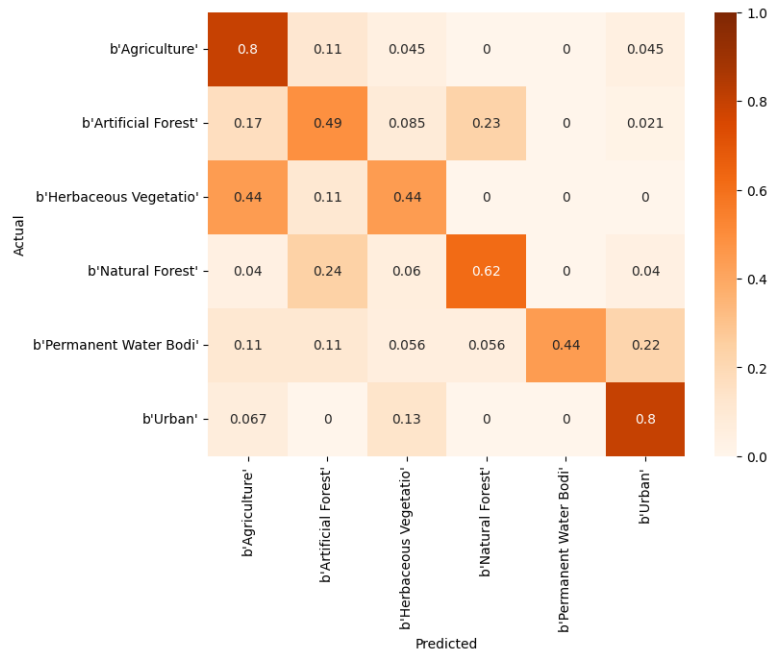


Figura 4.4: Matriz de confusión para modelo base con máximo *accuracy*.



Figura 4.5: Matriz de confusión para modelo base con máximo *macro recall*.

Cabe destacar que en una matriz de confusión normalizada por filas, los valores de la diagonal corresponden al *recall* clase a clase, por lo que es lógico que un modelo con *recall* alto en una clase altamente representada en la región, tal como *Agriculture* pueda presentar un mejor *accuracy*. Por otro lado, se puede apreciar que el modelo con máximo recall obtiene mejores resultados para clases menos representadas, como *Herbaceous Vegetation* y *Urban*.

### 4.2.2.2. Red paralela rotacional (CNN)

Es posible modificar la arquitectura secuencial básica de una red neuronal convolucional, estableciendo cuatro ramas paralelas operando sobre rotaciones sobre imágenes con simetría cíclica, tal como imágenes asociadas a fenómenos astronómicos[33] (ver figura 4.6), es posible que esta arquitectura también pueda aprovechar la simetría cíclica presente en imágenes satelitales.

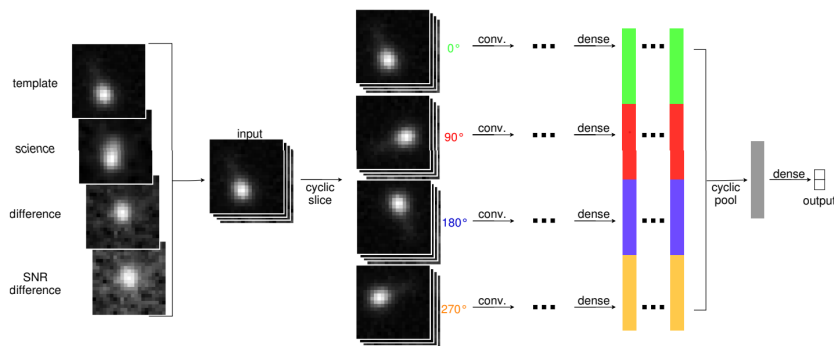


Figura 4.6: Ejemplo de red aplicada a detección de supernovas [33].

Con esto en mente, se utiliza la arquitectura de red detallada en la figura 4.7.

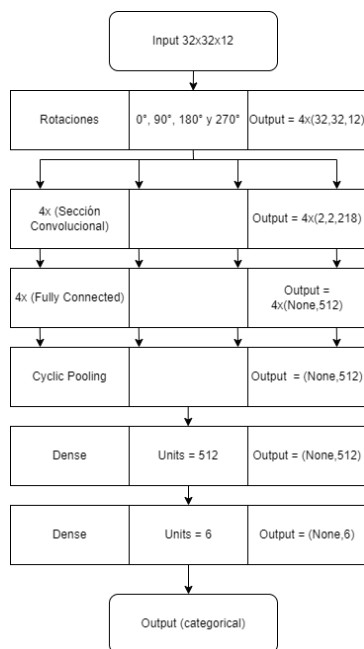


Figura 4.7: Modelo Convolutivo Rotacional.

Las métricas resultantes para esta arquitectura se encuentran en la tabla 4.3, obteniéndose un aumento en *macro-recall* en comparación al modelo base.

Tabla 4.3: Resultados para modelo convolutivo rotacional.

Modelo	Acc. a un nivel [%]	Acc. a dos niveles [%]	Macro Precision [%]	Macro Recall [%]	Macro F1 [%]
Rotacional, máxima <i>accuracy</i>	62	80	57	60	57
Rotacional, máximo <i>recall</i>	60	84	60	65	60

En las figuras 4.8 y 4.12 se encuentran las matrices con mejor *accuracy* y *macro recall* respectivamente.

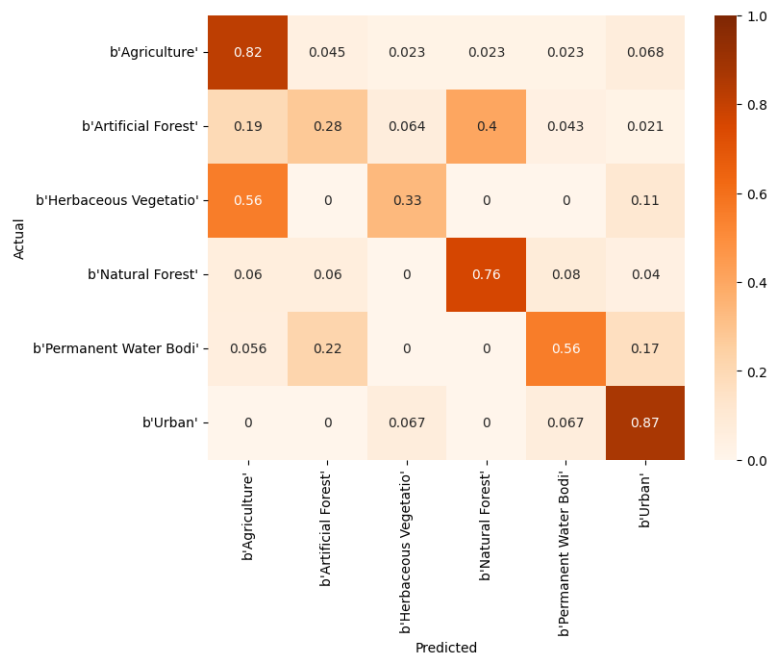


Figura 4.8: Matriz de confusión para modelo rotacional con máximo *accuracy*.

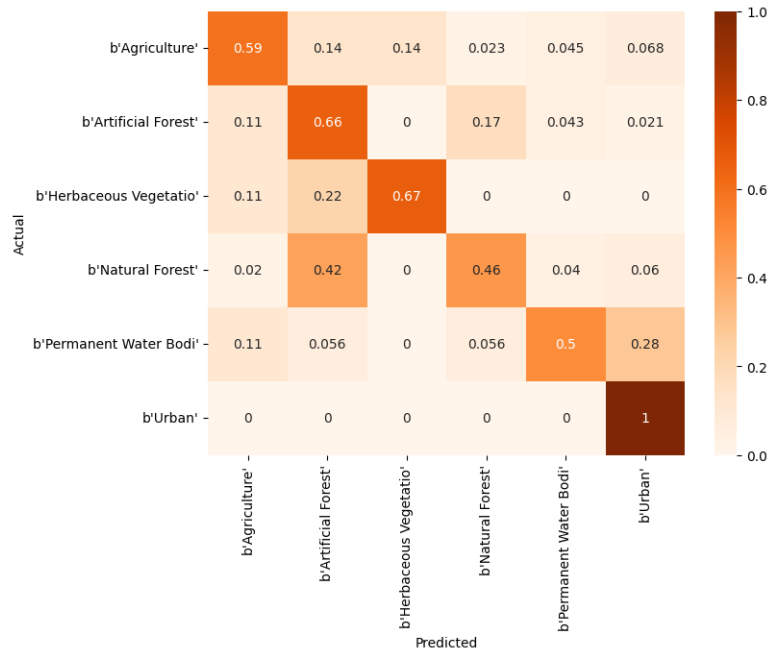


Figura 4.9: Matriz de confusión para modelo rotacional con máximo *macro recall*.

Nuevamente se observa el fenómeno en el cual el modelo con mayor *accuracy* es aquel con mayor *recall* para la clase *agriculture*.



### 4.2.2.3. Incorporación de índices espectrales (CNN)

Aumentando el tamaño del input a 32x32x17, se incluyen los índices:

$$NDVI = \frac{B12 - B8}{B12 + B8}$$

$$NDBI = \frac{B11 - B8}{B11 + B8}$$

$$AFRI = \frac{B8A - 0.66 \cdot B11}{B8A + 0.66 \cdot B11}$$

$$NdBaI = \frac{B6 - B11}{B6 + B11}$$

$$WBI = \frac{B2 - B4}{B2 + B4}$$

De este modo la estructura de esta red se puede apreciar en la figura 4.10.

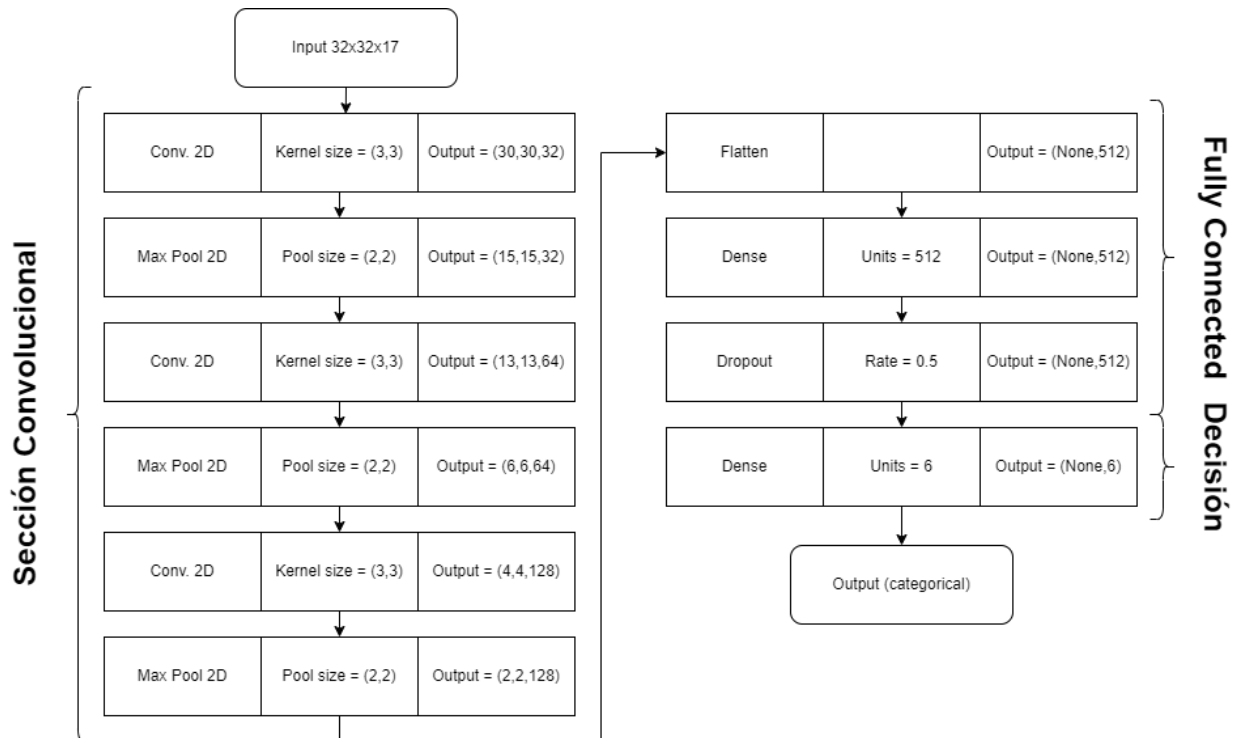


Figura 4.10: Modelo Convolucional con inclusión de índices espectrales.

Tabla 4.4: Resultados para modelo convolucional inclusivo de índices espectrales.

Modelo	Acc. a un nivel [%]	Acc. a dos niveles [%]	Macro Precision [%]	Macro Recall [%]	Macro F1 [%]
Rotacional, máxima <i>accuracy</i>	60	82	55	57	54
Rotacional, máximo <i>recall</i>	58	82	56	62	57

Las matrices de confusión respectivas se encuentran en las figuras 4.11 y 4.12.

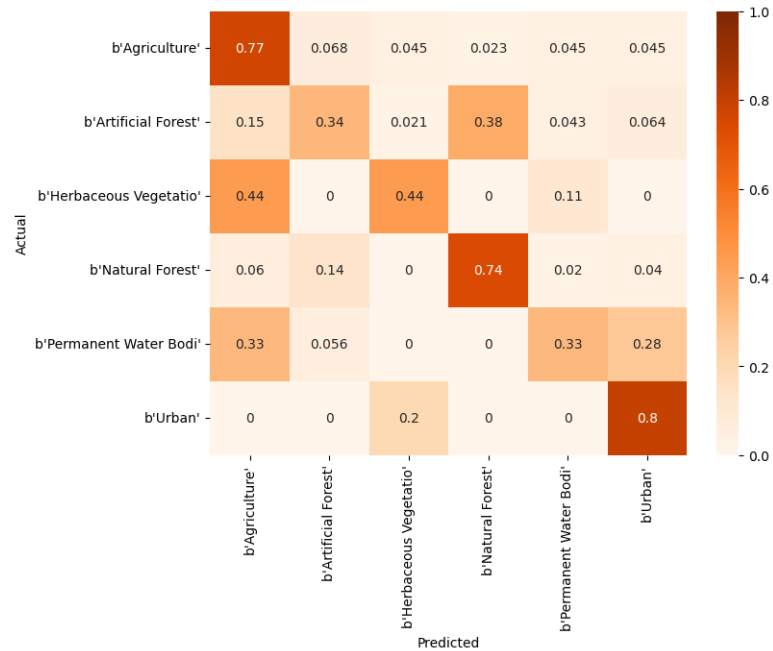


Figura 4.11: Matriz de confusión para modelo inclusivo de índices espectrales con máximo *accuracy*.



Figura 4.12: Matriz de confusión para modelo inclusivo de índices espectrales con máximo *macro recall*.

Nuevamente, se puede apreciar que el uso del *accuracy* favorece de sobremanera aquellos modelos con sesgo de clasificación hacia la clase *agriculture*, siendo el *macro recall* una métrica más oportuna para identificar modelos capaces de discriminar transversalmente entre clases.

#### 4.2.2.4. Utilización de EuroSat sobre arquitecturas anteriores

Con la finalidad de verificar si los resultados anteriores son cercanos al desempeño máximo posible para las arquitecturas propuestas, se utilizan modelos con la misma estructura sobre un dataset más extenso y validado manualmente. A continuación se muestra la matriz de confusión con máximo *accuracy* entre todas las iteraciones para las 3 arquitecturas.

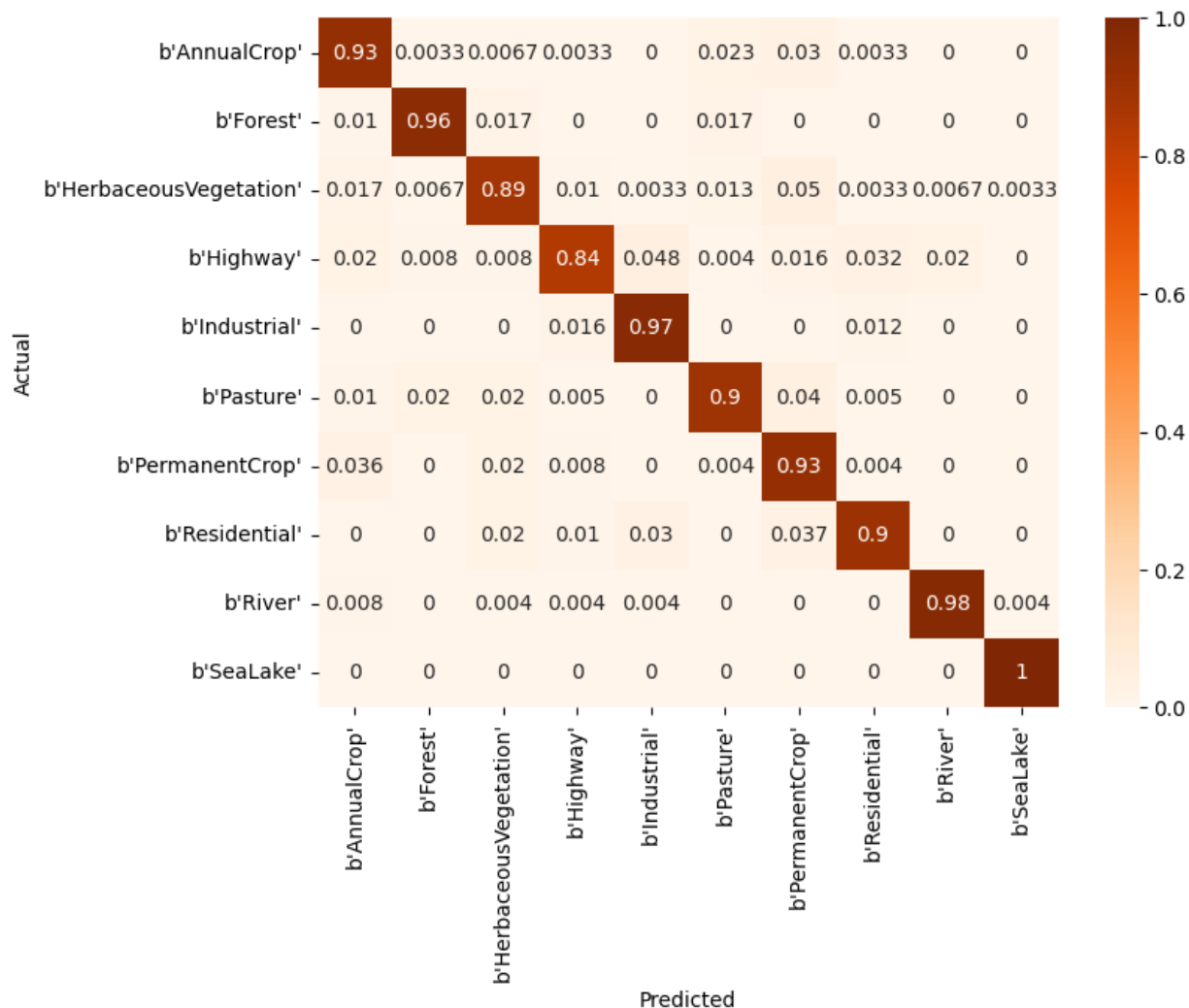


Figura 4.13: Matriz de confusión para modelo rotacional utilizando dataset EuroSat.

Tabla 4.5: Resultados para arquitecturas convolucionales propuestas aplicadas en dataset EuroSat.

Modelo	Acc. a un nivel [%]	Acc. a dos niveles [%]	Macro Precision [%]	Macro Recall [%]	Macro F1 [%]
Base	92	98	92	92	92
Rotacional	93	98	93	93	93
Con índices	92	98	92	92	92

#### 4.2.2.5. Random Forest

Se entrena un modelo Random Forest utilizando la librería *tensorflow\_decision\_forests*, [34], por defecto todos los modelos Random Forest tienen un total de 300 árboles *CART* con profundidad máxima de 16 nodos, configuración que se mantiene hasta confirmar mejoras de desempeño asociadas a la limpieza automática propuesta.

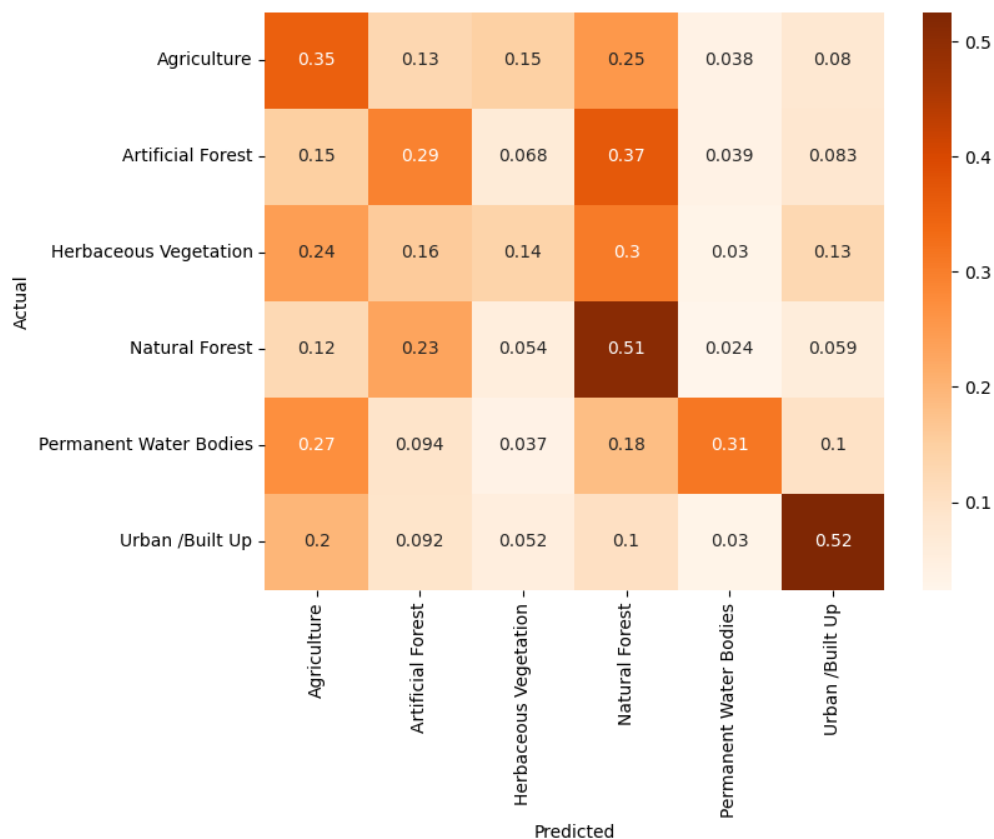


Figura 4.14: Matriz de confusión para modelo Random Forest entrenado con datos de CONAF.

Se obtienen una matriz de confusión (figura 4.14) y una tabla de métricas (4.6).

Tabla 4.6: Resultados para modelo Random Forest entrenado con datos de CONAF.

Métrica	<i>Precision</i> [%]	<i>Recall</i> [%]	<i>F1</i> [%]
<b>Promedio <i>macro</i></b>	28	36	28
<b>Promedio <i>weighted</i></b>	51	34	34

Se puede apreciar tanto en la figura 4.14 como en la tabla 4.6 un desempeño inadecuado para la tarea de realizar mapas de uso/cobertura de suelo. Tanto los bajos valores en las métricas obtenidas para este modelo como la diferencia en el desempeño entre las redes convolucionales entrenadas con los datos de CONAF en contraste al dataset EuroSat podrían indicar una posible baja confiabilidad de los datos de CONAF para crear datasets de entrenamiento de algoritmos de aprendizaje de máquinas.

### 4.2.2.6. Creación de píxeles representativos utilizando red convolucional

En la figura 4.15 se muestra una imagen satelital junto a la clasificación de uso y cobertura de suelos asociada por parte de CONAF.

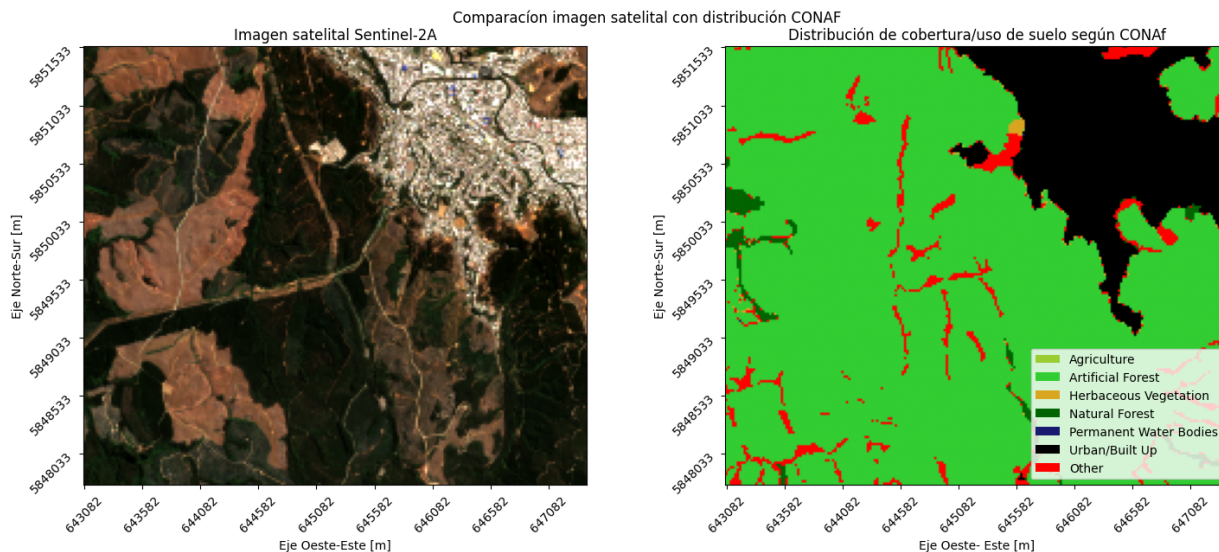


Figura 4.15: Imagen satelital junto a mapa de cobertura / uso de suelo CONAF.

Se puede apreciar que existe una cantidad considerable de suelo desnudo incorrectamente clasificado como bosque artificial por parte de CONAF, además, hay un río cruzando la zona urbana mostrada en la figura que no está clasificado como tal.

Con el objetivo de disminuir la cantidad de píxeles incorrectos utilizados para entrenar modelos basados en árboles de decisión, se establece un método en el cual se realizan todas las posibles predicciones de la red neuronal convolucional entrenada para cada píxel, luego de esto se establecen como *píxeles de alta confianza* aquellos en los que la mayoría de las predicciones por píxel sean iguales a la clasificación hecha por CONAF para el mismo.

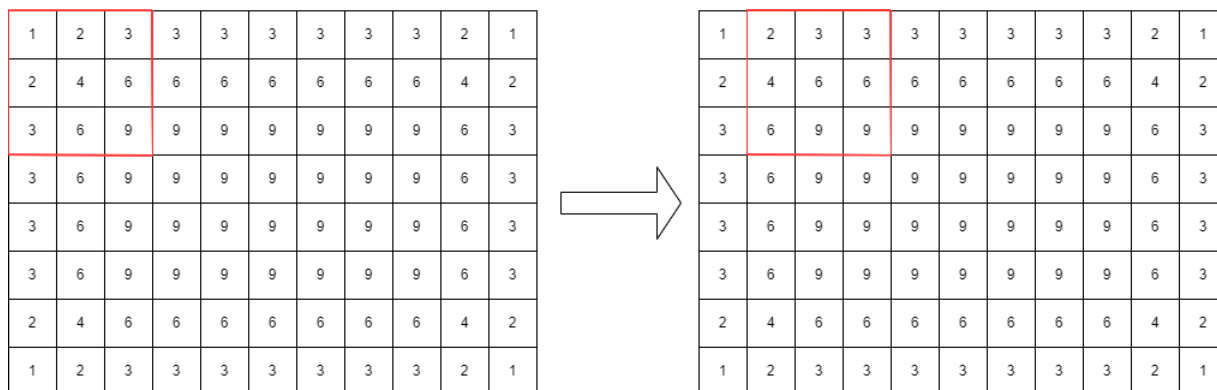


Figura 4.16: Diagrama de deslizamiento de ventana de 3x3, ilustrando la cantidad de predicciones por píxel.

Una vez realizado el calce entre predicciones y mapas de cobertura, se puede efectuar un segundo filtrado al eliminar *outliers* mediante un umbral definido como el doble de la desviación estandar del total de pixeles hallados en común para cada clase. A continuación se muestra el efecto de la limpieza para las dos clases dominantes en el área:

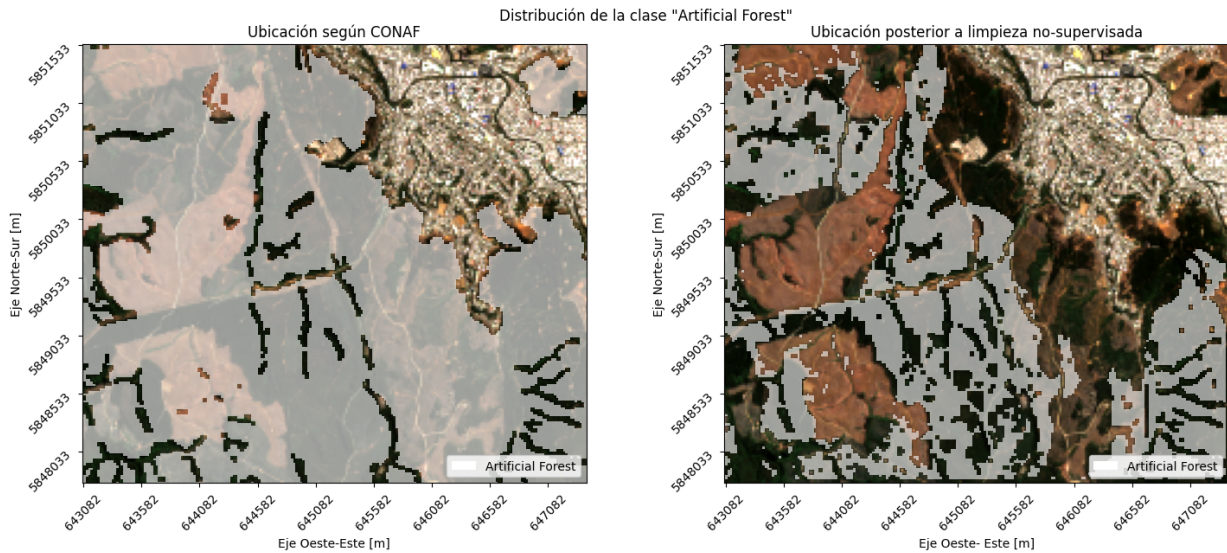


Figura 4.17: Demostración limpieza automática aplicada a clase *Artificial Forest*.

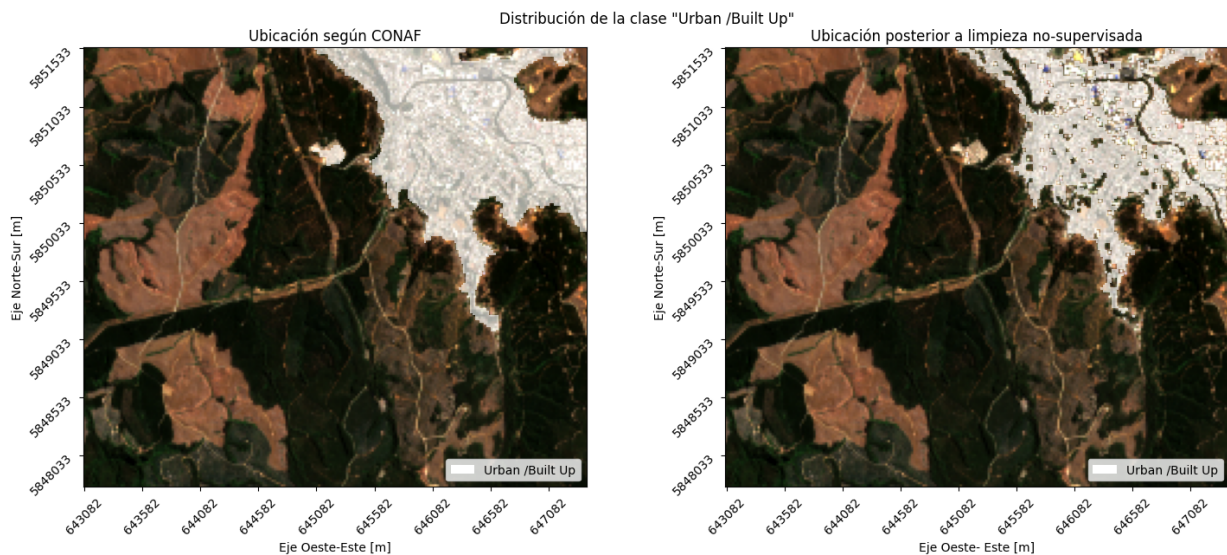


Figura 4.18: Demostración limpieza automática aplicada a clase *Urban / Built Up*.

Se puede apreciar la disminución casi total del suelo desnudo para el primer caso, y una eliminación parcial del río presente dentro del área asociada a la zona urbana, así como también algunas partes correspondientes a áreas verdes. Al grupo de pixeles obtenidos con el método descrito anteriormente se les denomina *pixeles representativos* o *pixeles de alta confianza*, ya que existe una minimización automatizada de la cantidad falsos positivos para cada clase.

### 4.2.2.7. Caracterización de respuestas espectrales por clase utilizando pixeles representativos

Considerando el rango máximo en el cual se pueden codificar las imágenes multispectrales de Sentinel-2A (0 a 4095 por pixel) el ratio de reflectancia para una banda específica de un pixel corresponde simplemente al valor de dicho pixel dividido en 4095

A continuación se muestran las respuestas obtenidas, en orden alfabético comenzando por la clase *Agriculture* (figura 4.19).

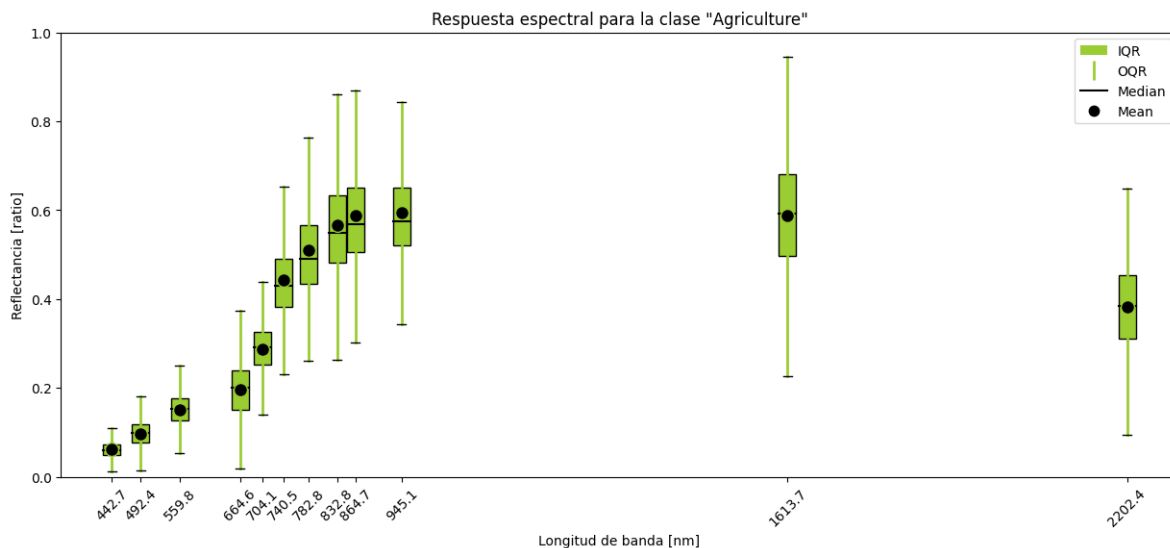


Figura 4.19: Respuesta espectral de la clase *Agriculture*.

Continuamos con la clase *Artificial Forest* (figura 4.20), ya es posible ver que se puede distinguir fácilmente entre estas dos respuestas al comparar la banda 11 (1613.7 [nm]), teniendo el bosque artificial una reflectancia mucho menor en esta longitud de banda

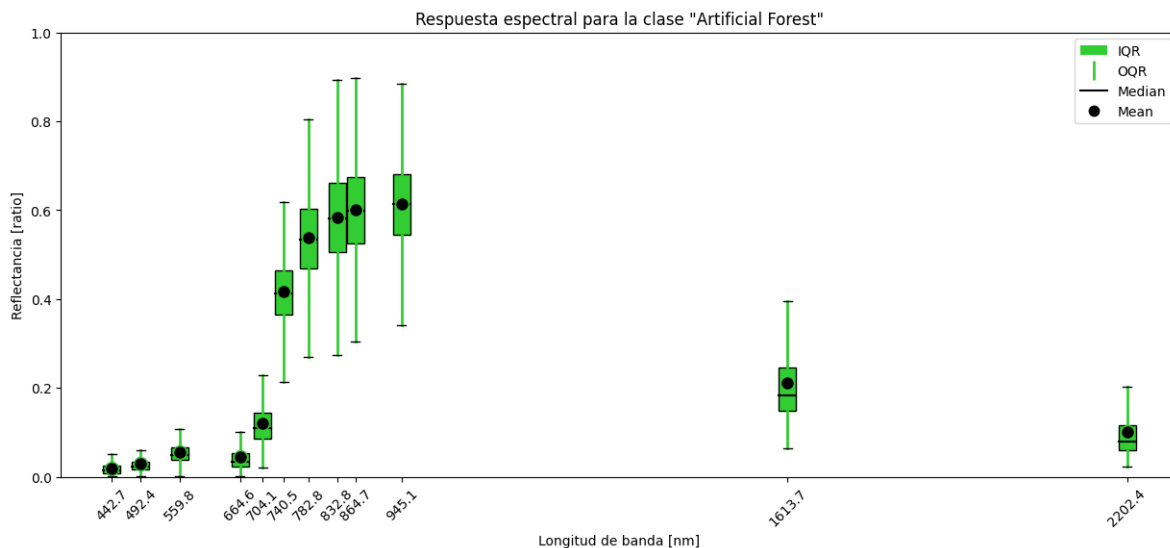


Figura 4.20: Respuesta espectral de la clase *Artificial Forest*.

Por otro lado se puede apreciar la respuesta para la clase *Herbaceous Vegetation* (figura 4.21), la cual es similar a la de la clase *Agriculture*, por lo que luego de mostrar el resto de las respuestas se mostrará un gráfico comparativo de las respuestas para ambas clases.

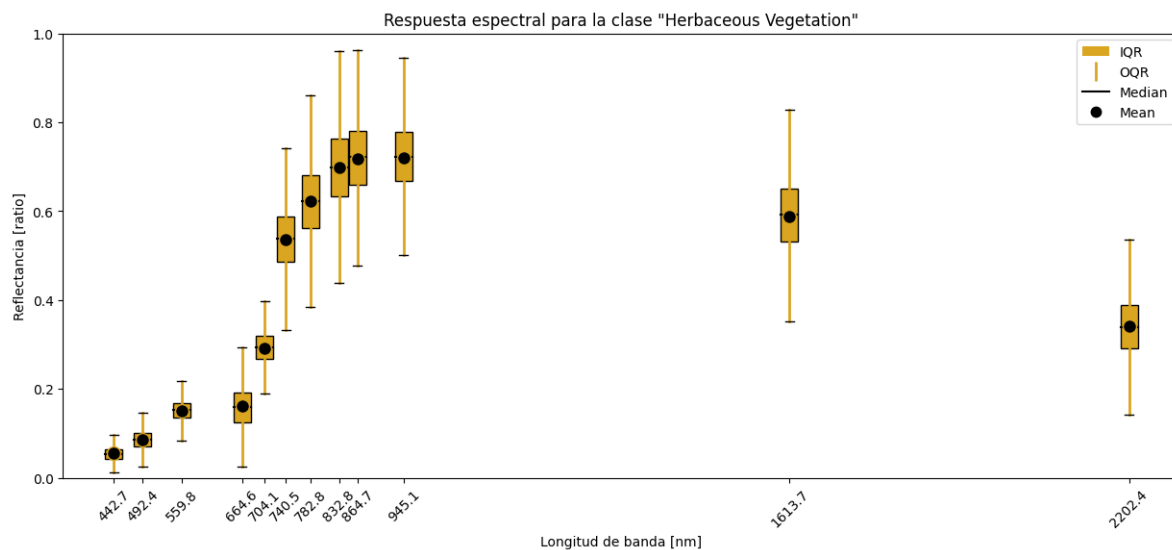


Figura 4.21: Respuesta espectral de la clase *Herbaceous Vegetation*.

Siguendo, se muestra la respuesta para la clase *Natural Forest* (4.22), la cual tiene un parecido a a la clase *Artificial Forest*, al igual que para el caso anterior, se mostrará un gráfico comparativo al final de esta sección.

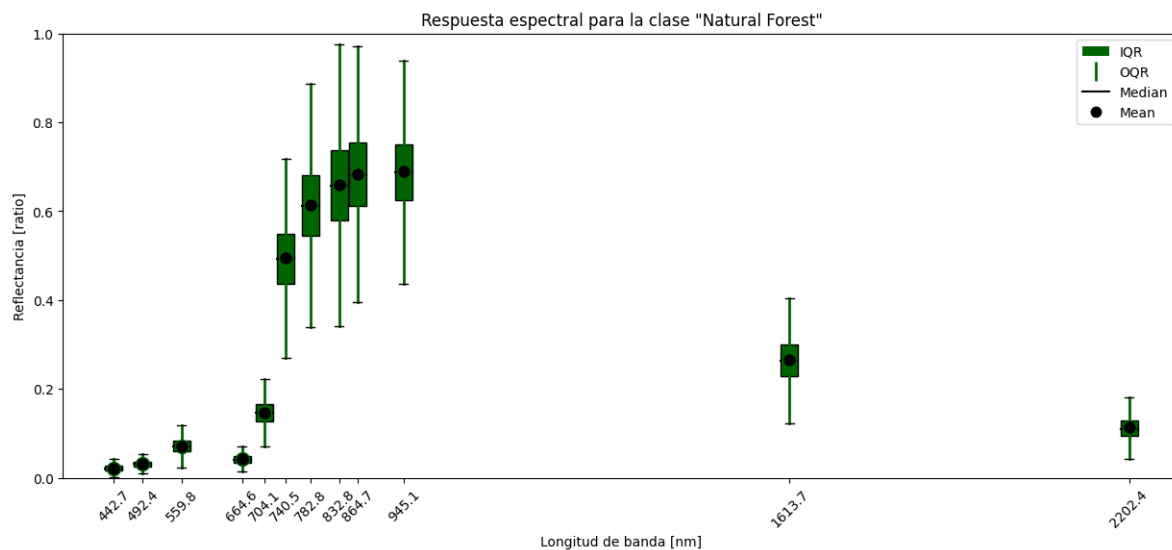


Figura 4.22: Respuesta espectral de la clase *Natural Forest*.



En la figura 4.23 se muestra la respuesta para la clase *Permanent Water Bodies*, fácilmente caracterizable por una baja reflectancia para todas las bandas disponibles.

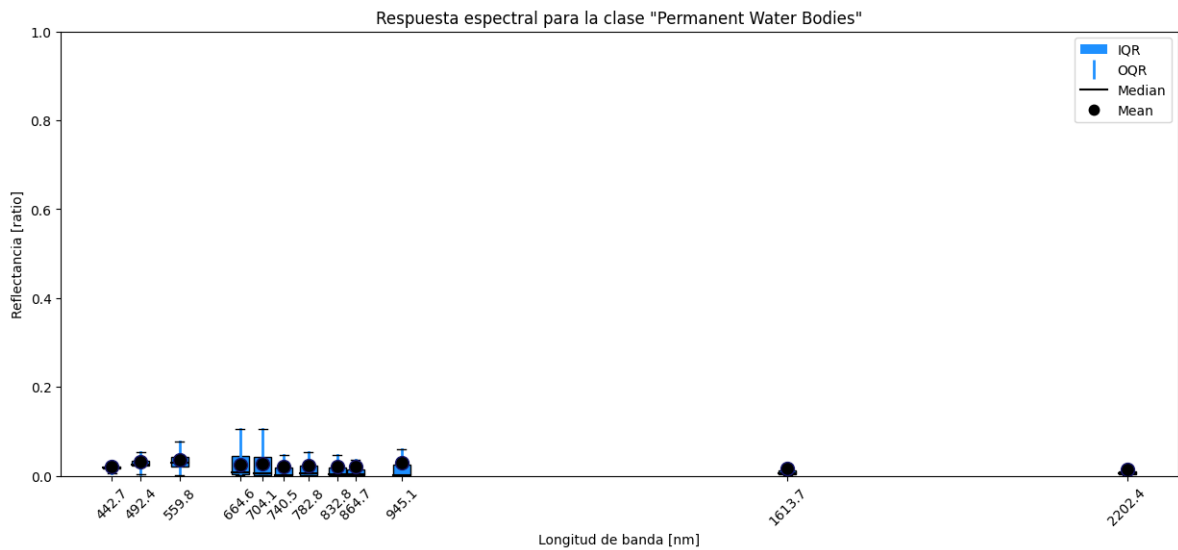


Figura 4.23: Respuesta espectral de la clase *Permanent Water Bodies*.

En último lugar podemos ver la respuesta espectral asociada a la clase *Urban / Built Up* (figura 4.24), con una respuesta característicamente plana pero de reflectancia promedio muy superior a la clase *Permanent Water Bodies*

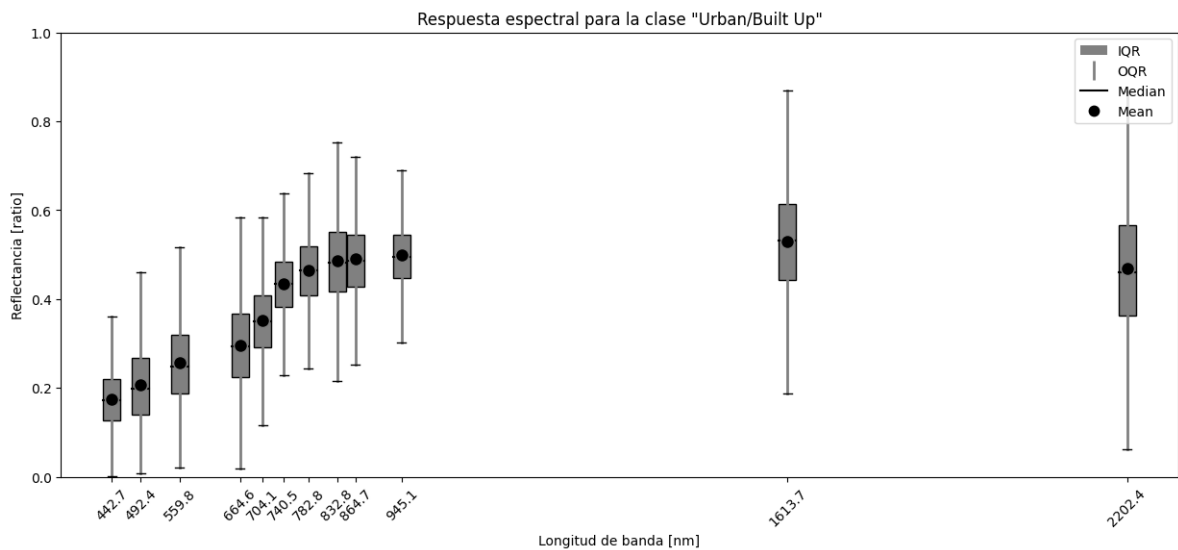


Figura 4.24: Respuesta espectral de la clase *Urban / Built Up*.

Tal como había sido mencionado anteriormente, se obtiene un gráfico comparativo para las respuestas de las clases *Agriculture* y *Herbaceous Vegetation* (figura 4.25).

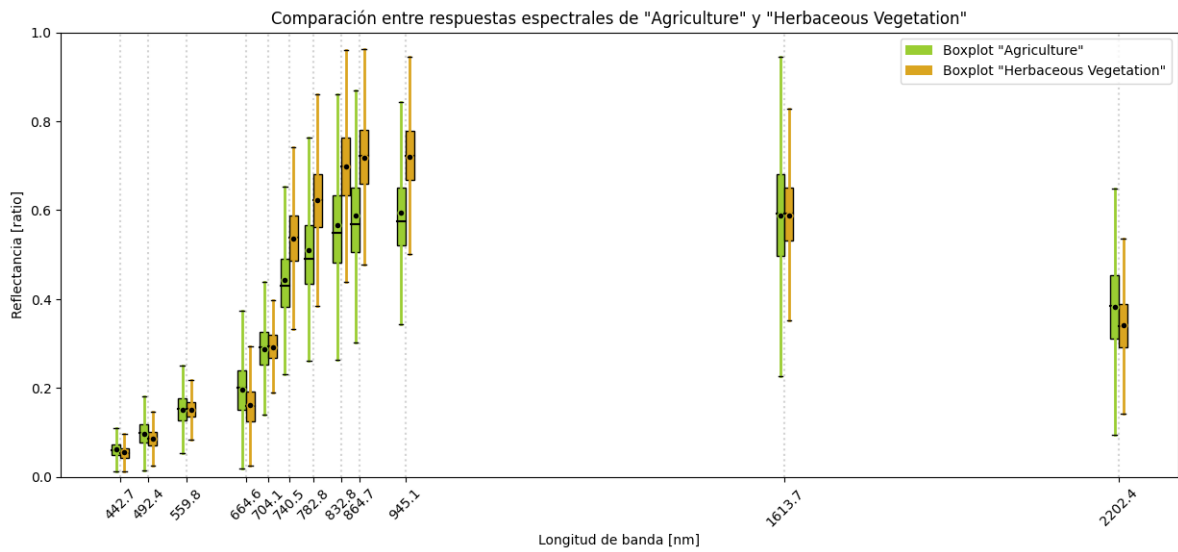


Figura 4.25: Comparación de las respuestas espectrales de la clase *Agriculture* y *Herbaceous Vegetation*.

Finalmente, en la figura 4.26 se muestra el gráfico comparativo obtenido para las respuestas comparativas de ambos tipos de bosque.

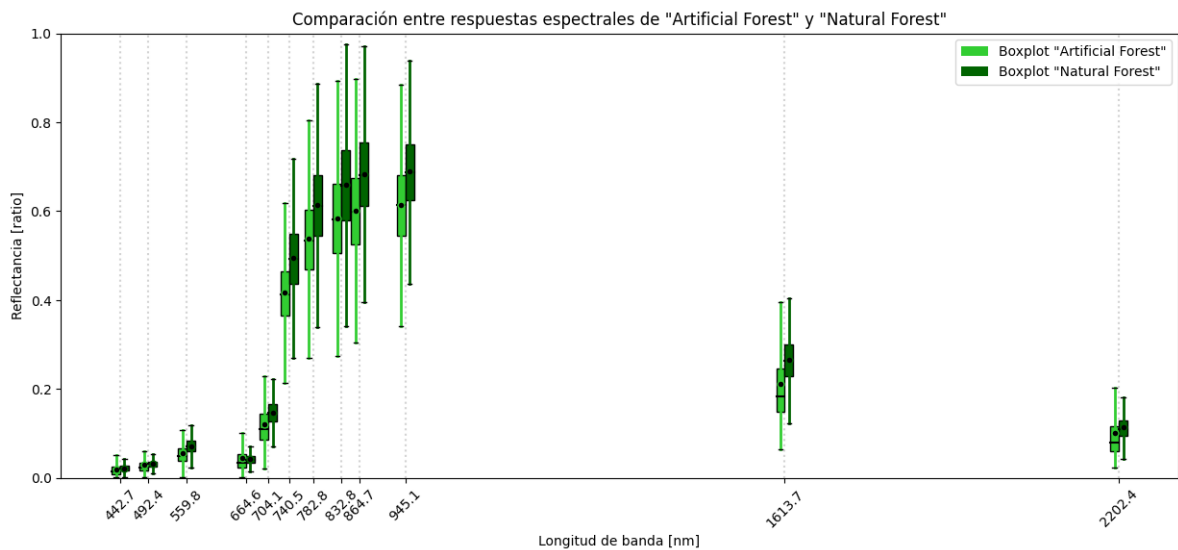


Figura 4.26: Comparación de las respuestas espectrales de la clase *Artificial Forest* y *Natural Forest*.

Se puede apreciar un grado de distinción entre todas las clases, sin embargo, existe un solapamiento considerable para las clases incluidas en la figura 4.25 y en la figura 4.26.

#### 4.2.2.8. Random Forest entrenado utilizando pixeles representativos

Se establecen dos métodos de limpieza automática de datos:

- **Método 1:** Se determinan los pixeles altamente representativos mediante la intersección entre el grupo de pixeles clasificados para cada clase utilizando la red convolucional base con máximo *macro recall* y el mapa de actualización de cobertura y uso de suelo publicado por CONAF.
- **Método 2:** Se calcula el promedio y desviación estandar de reflectancia para los pixeles hallados utilizando el método 1, luego se efectúa un segundo filtrado al eliminar *outliers* con una distancia mayor a dos desviaciones estandar del promedio.

Se seleccionan manualmente 5 áreas de interés, las cuales son unidas a un conjunto de prueba obtenido por la selección aleatoria del 10% de los datos disponibles. Se utiliza como conjunto de prueba transversal el conjunto de prueba asociado a los pixeles del **método 2**, y se vuelven a realizar pruebas sobre el modelo Random Forest entrenado con datos sin filtrar.

La figura 4.30 muestra la matriz de confusión resultante para el modelo entrenado sobre datos CONAF sin filtrar.

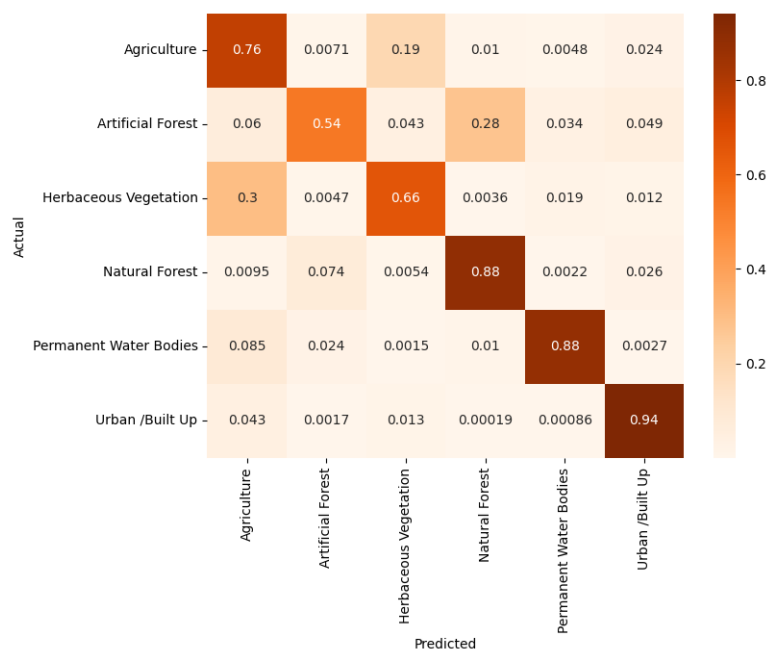


Figura 4.27: Matriz de confusión para modelo entrenado con datos sin filtrar sobre datos filtrados.

Tabla 4.7: Resultados para Random Forest entrenado con datos sin filtrar sobre datos filtrados.

	<i>Accuracy</i> [%]	<i>Macro Precision</i> [%]	<i>Macro Recall</i> [%]	<i>Macro F1</i> [%]
Métrica	65	57	78	61

Se observa que el desempeño del modelo sobre datos limpios es notoriamente superior al que se pudo ver en la matriz de la figura 4.14 y la tabla 4.6.

Por su parte, se entrenan dos nuevos modelos del tipo Random Forest, uno utilizando datos obtenidos mediante el método 1 y otro utilizando datos obtenidos mediante el método 2, obteniéndose las matrices de confusión ilustradas en las figuras 4.28 y 4.29 respectivamente.

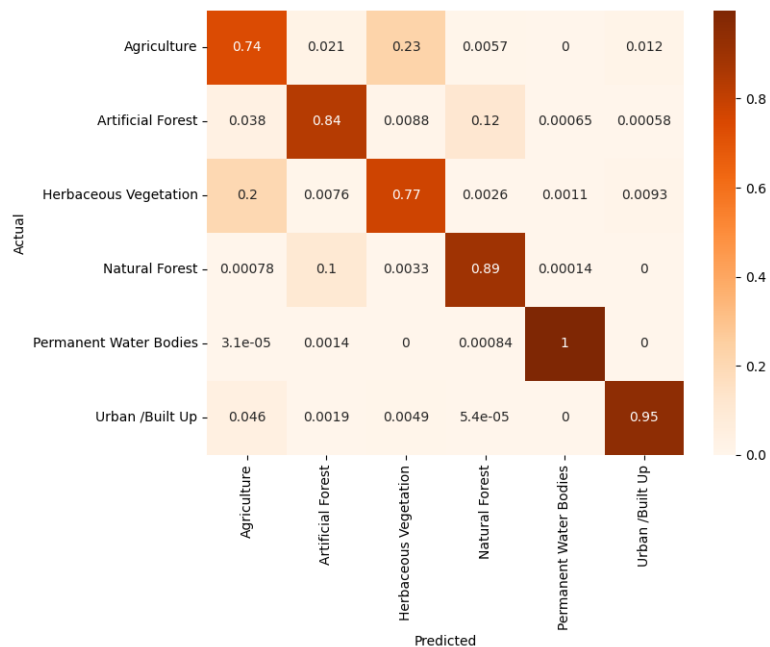


Figura 4.28: Matriz de confusión para modelo entrenado con datos filtrados mediante el método 1.

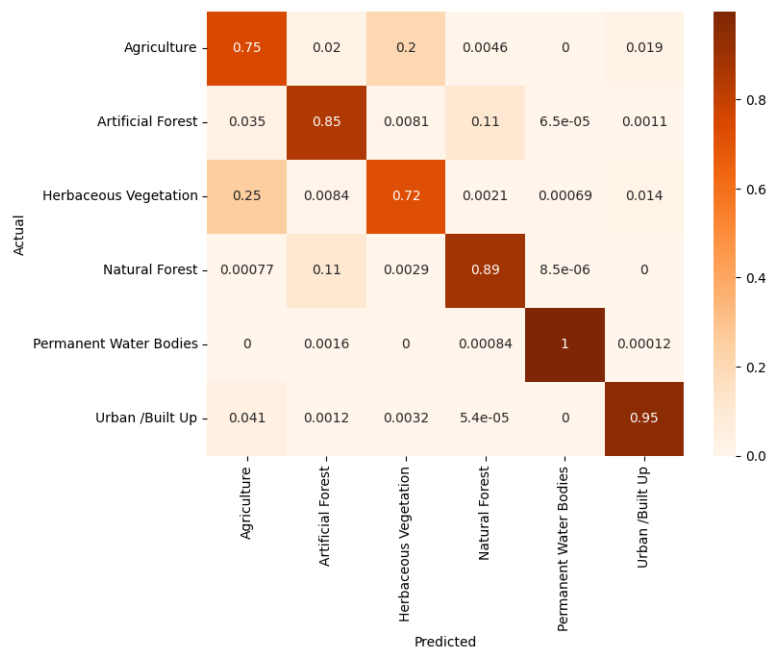


Figura 4.29: Matriz de confusión para modelo entrenado con datos filtrados mediante el método 2.

La tabla 4.8 contrasta el desempeño entre estos tres modelos.

Tabla 4.8: Tabla comparativa resultados para modelos Random Forest.

	<i>Accuracy</i> [%]	<i>Macro Precision</i> [%]	<i>Macro Recall</i> [%]	<i>Macro F1</i> [%]
<b>Sin Filtrar</b>	65	57	78	61
<b>Método 1</b>	85	76	86	79
<b>Método 2</b>	86	76	86	79

Con el fin de aumentar la confiabilidad de los modelos, es posible clasificar únicamente aquellos pixeles en los cuales la predicción sea realizada sobre un umbral arbitrario de certeza, se implementa una función simple capaz de realizar esta tarea, con la cual se obtienen nuevas métricas para los tres modelos RF con un umbral de certeza igual al 60 %, la figura 4.30 permite apreciar la matriz de confusión para el modelo entrenado con datos CONAF sin filtrar:

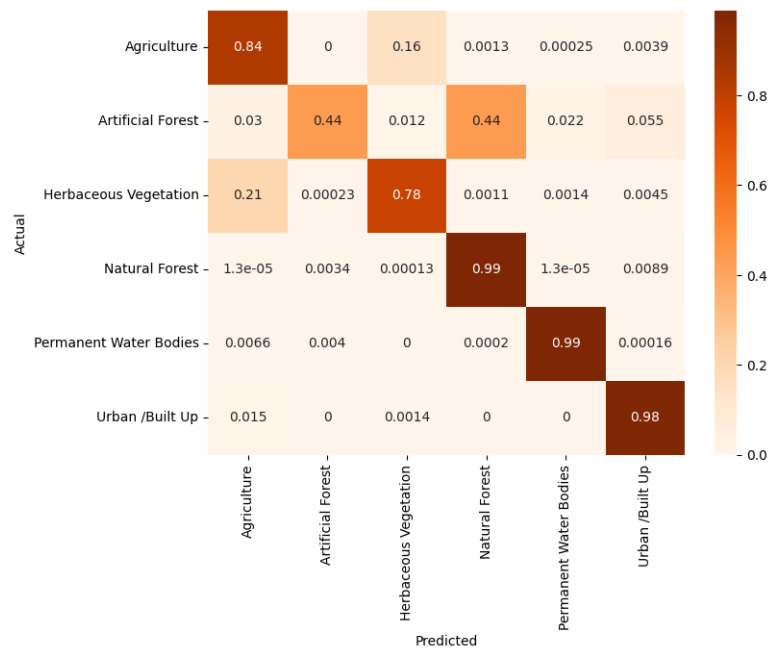


Figura 4.30: Matriz de confusión para modelo entrenado con datos sin filtrar sobre datos filtrados, con un umbral de certeza igual al 60 %.

Tabla 4.9: Resultados para Random Forest entrenado con datos sin filtrar sobre datos filtrados, con un umbral de certeza igual al 60 %.

	<i>Accuracy</i> [%]	<i>Macro Precision</i> [%]	<i>Macro Recall</i> [%]	<i>Macro F1</i> [%]
<b>Sin Filtrar</b>	82	80	84	79

Se puede apreciar un aumento significativo en todas las métricas, al comparar con la tabla 4.7 y la matriz 4.30, sin embargo, es importante recordar que cada vez que se establece un umbral de certeza disminuye la cantidad de predicciones, lo cual afecta directamente a la calidad de posibles mapas generados mediante este tipo de algoritmos.

Análogamente, se obtienen las matrices de confusión para los dos modelos entrenados con datos filtrados al utilizar un umbral de 60% de certeza, mostrándose en las figuras 4.31 y 4.32 respectivamente.

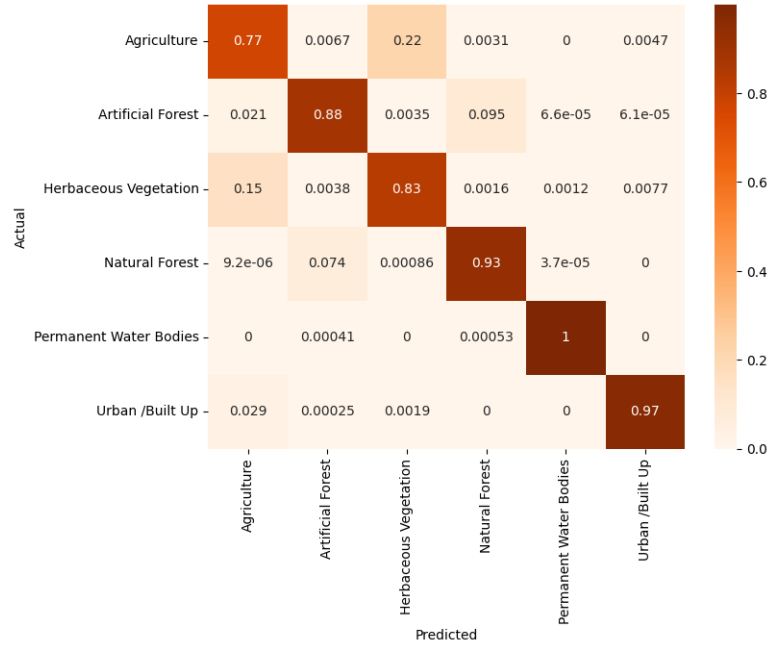


Figura 4.31: Matriz de confusión para modelo entrenado con datos filtrados mediante el método 1, con un umbral de certeza igual al 60%.

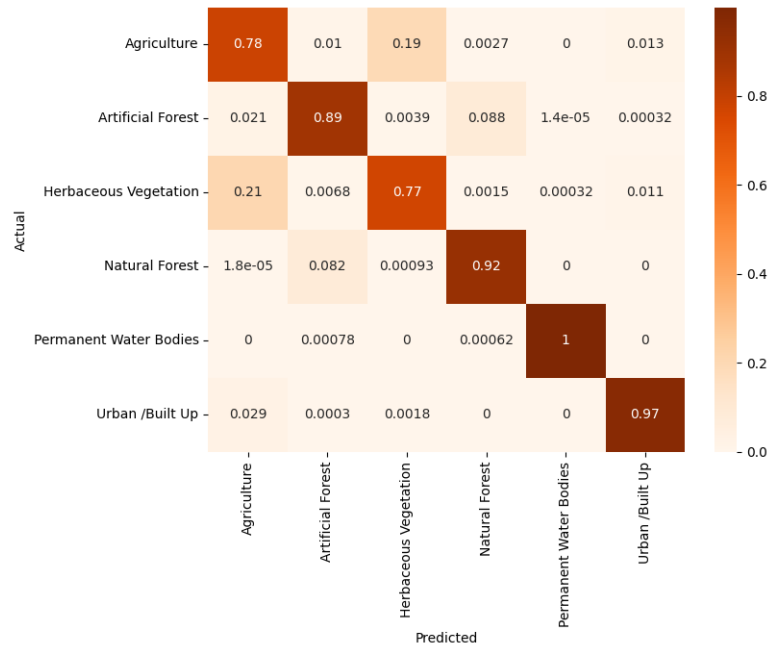


Figura 4.32: Matriz de confusión para modelo entrenado con datos filtrados mediante el método 2, con un umbral de certeza igual al 60%.

La tabla 4.10 agrupa las métricas resultantes de cada modelo al establecer el umbral fijado anteriormente.

Tabla 4.10: Métricas resultantes para modelos entrenados con distintos algoritmos de limpieza, con exigencia de 60 % de certeza.

	<i>Accuracy</i> [%]	<i>Macro Precision</i> [%]	<i>Macro Recall</i> [%]	<i>Macro F1</i> [%]
<b>Sin Filtrar</b>	82	80	84	79
<b>Método 1</b>	89	80	90	83
<b>Método 2</b>	89	80	89	83

Si bien es útil comparar entre estos modelos utilizando estas métricas, se debe hacer una evaluación más completa al comparar el grado de cobertura (en porcentaje de pixeles clasificados por sobre un umbral arbitrario) para cada nivel de certeza, así como el *accuracy* promedio en función de la misma variable. A continuación se puede observar dicha variación para cada modelo:

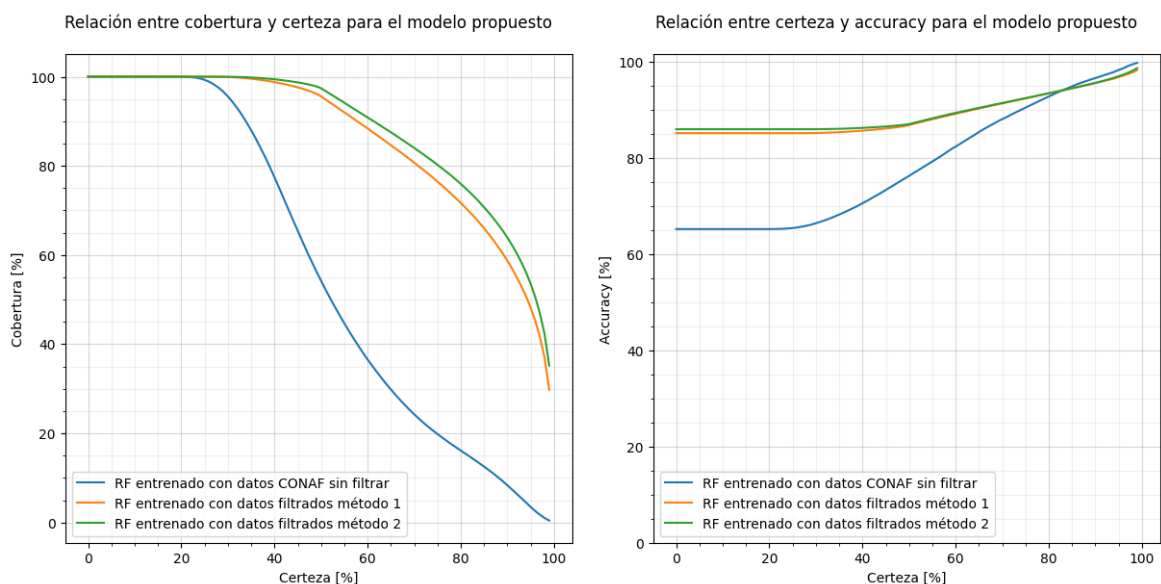


Figura 4.33: Relación entre cobertura, *accuracy* y certeza.

Se puede apreciar un grado de cobertura mayor para el modelo entrenado con datos obtenidos mediante el método 2 de manera transversal, así como un mayor *accuracy* hasta una certeza igual al 80 %, certeza para la cual el *accuracy* es mayor para el modelo entrenado con datos sin filtrar, sin embargo, la cobertura para dicho modelo es inferior al 20 % para dicho umbral, comparativamente mucho inferior.

Luego de estos resultados, se muestra la aplicabilidad de estos modelos al hacer una comparativa entre el efecto de ejecutar el modelo entrenado con datos sin filtrar y aquel filtrado mediante el segundo método de filtrado.

Veamos los resultados para la primera zona de prueba (figura ??), comenzando por el mapa generado por el modelo entrenado directamente con datos CONAF, con una cobertura del 16 %:

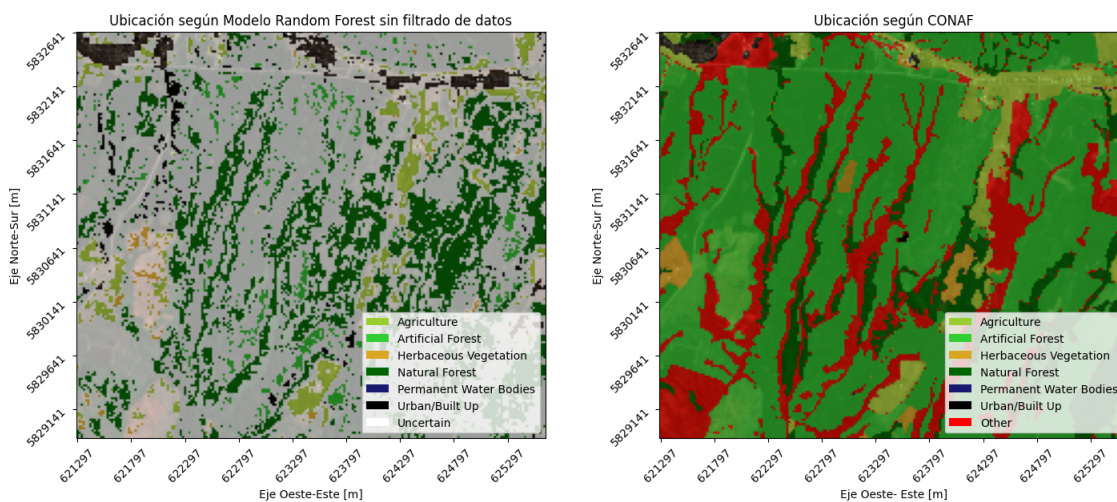


Figura 4.34: Primer ejemplo aplicado a generación de mapas, modelo entrenado directamente con datos CONAF.

Por otro lado se muestra su análogo para el modelo entrenado con los datos limpiados utilizando el método 2 (figura 4.35), con una cobertura del 90 %:

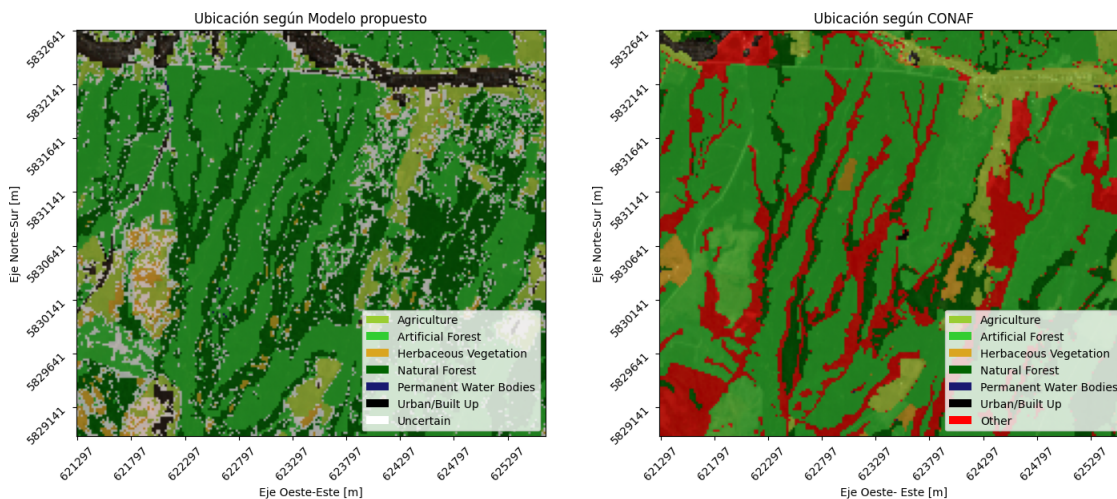


Figura 4.35: Primer ejemplo aplicado a generación de mapas, modelo entrenado con datos limpios (método 2).

Las métricas resultantes para esta zona geográfica en los pixeles altamente representativos se encuentran en la tabla 4.11

Tabla 4.11: Resultados modelos RF, primera zona geográfica.

Modelo	Weighted Precision [%]	Weighted Recall [%]	Weighted F1 [%]
Entrenado con datos CONAF sin filtrar	80	47	40
Entrenado con datos filtrados (M2)	94	89	91



Veamos los resultados para la segunda zona de prueba (figura ??), la cual tiene como punto de interés mayor un río que la atraviesa, para este caso se logra una cobertura del 10 %:

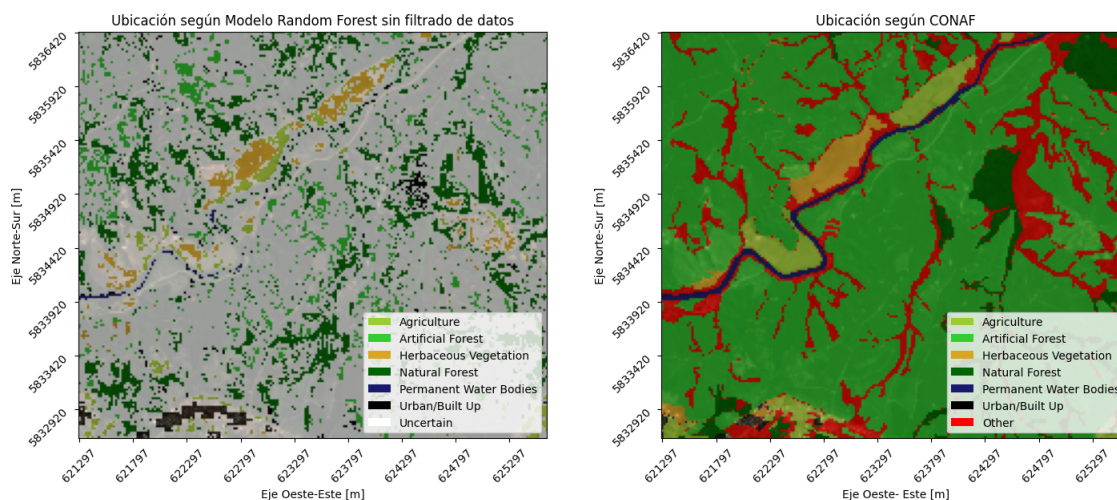


Figura 4.36: Segundo ejemplo aplicado a generación de mapas, modelo entrenado directamente con datos CONAF

Tal como para el caso anterior, se muestra su análogo para el modelo entrenado con datos obtenidos con el método 2 (figura 4.35), lográndose una cobertura del 94 %.

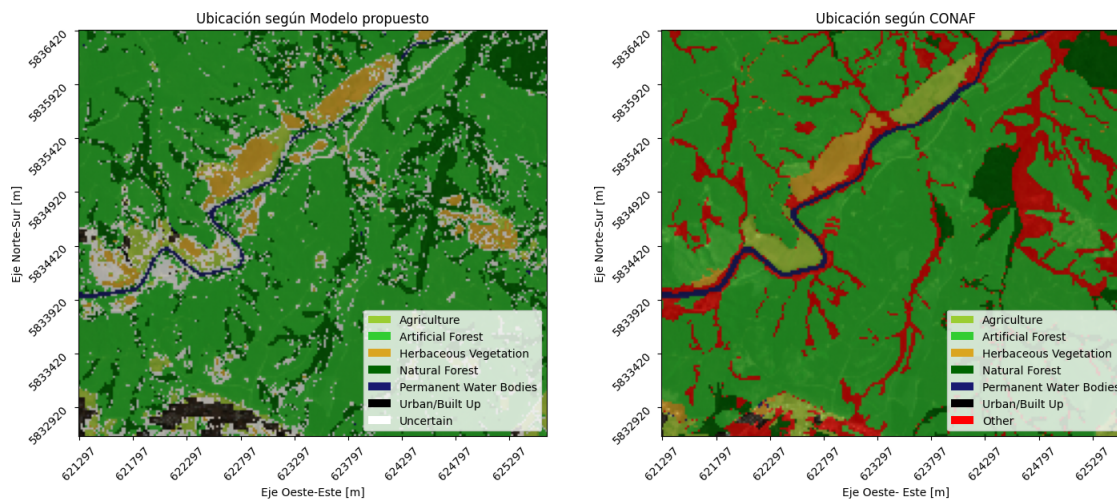


Figura 4.37: Segundo ejemplo aplicado a generación de mapas, modelo entrenado con datos limpios (método 2).

Las métricas resultantes para esta zona geográfica en los pixeles altamente representativos se hallan en la tabla 4.12.

Tabla 4.12: Resultados modelos RF, segunda zona geográfica.

Modelo	<i>Weighted Precision</i> [%]	<i>Weighted Recall</i> [%]	<i>Weighted F1</i> [%]
Entrenado con datos CONAF sin filtrar	87	46	53
Entrenado con datos filtrados (M2)	96	93	94

Por otro lado tenemos la tercera área de interés, con un gran cuerpo de agua dentro de ella, se logra una cobertura del 10 % con el modelo entrenado utilizando datos de CONAF sin filtrar (ver figura 4.38).

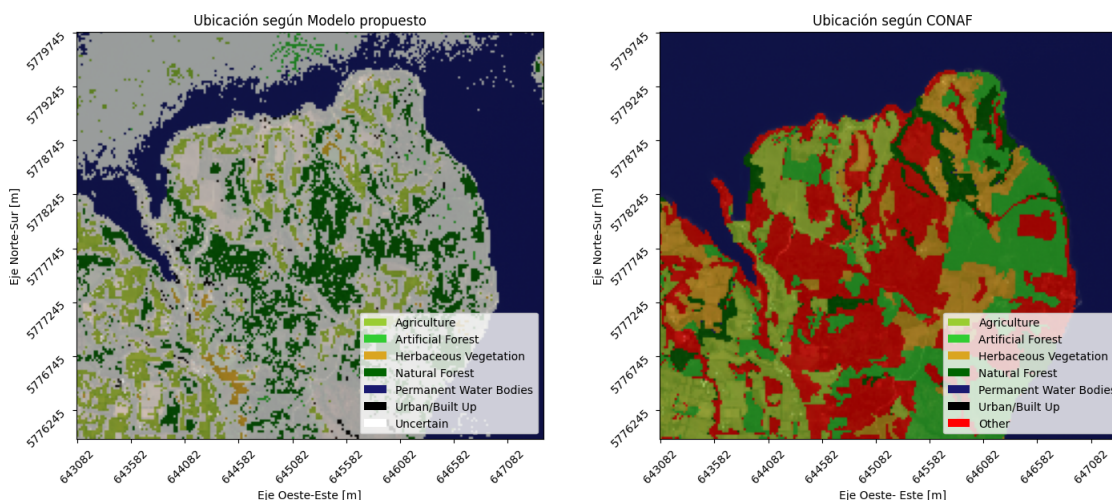


Figura 4.38: Tercer ejemplo aplicado a generación de mapas, modelo entrenado directamente con datos CONAF.

Por otro lado, para el análogo utilizando el modelo entrenado con datos limpiados utilizando el método 2 (figura 4.39), se logra una cobertura del 96 %:

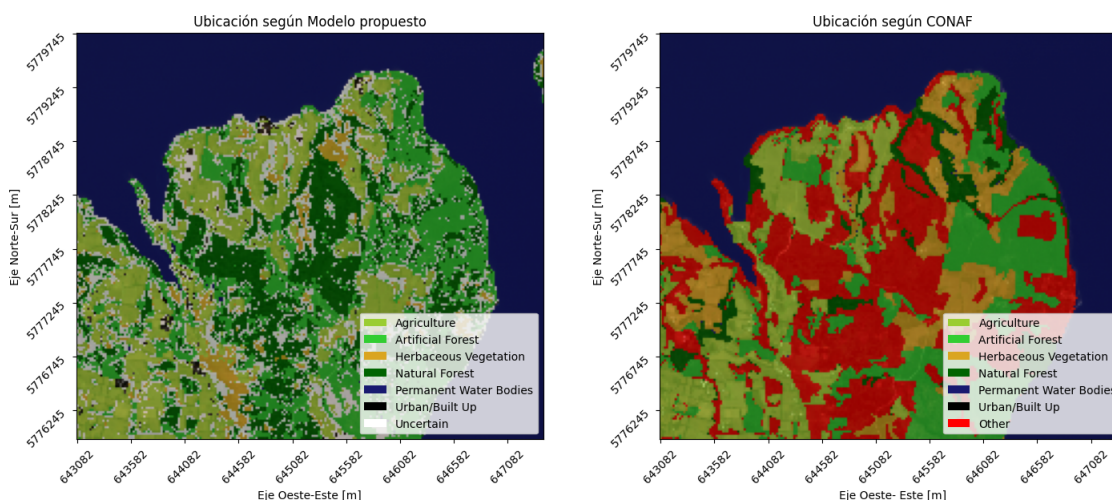


Figura 4.39: Tercer ejemplo aplicado a generación de mapas, modelo entrenado con datos limpios (método 2).

En la tabla 4.13 se encuentran as métricas resultantes para esta zona geográfica en los pixeles altamente representativos.

Tabla 4.13: Resultados modelos RF, tercera zona geográfica.

Modelo	<i>Weighted Precision</i> [%]	<i>Weighted Recall</i> [%]	<i>Weighted F1</i> [%]
Entrenado con datos CONAF sin filtrar	87	46	53
Entrenado con datos filtrados (M2)	99	98	98

La cuarta área de interés (ver figura ??) cuenta con una mezcla entre bosques, agricultura y zona urbana, para el modelo entrenado con datos sin filtrar se logra una cobertura del 32 %.

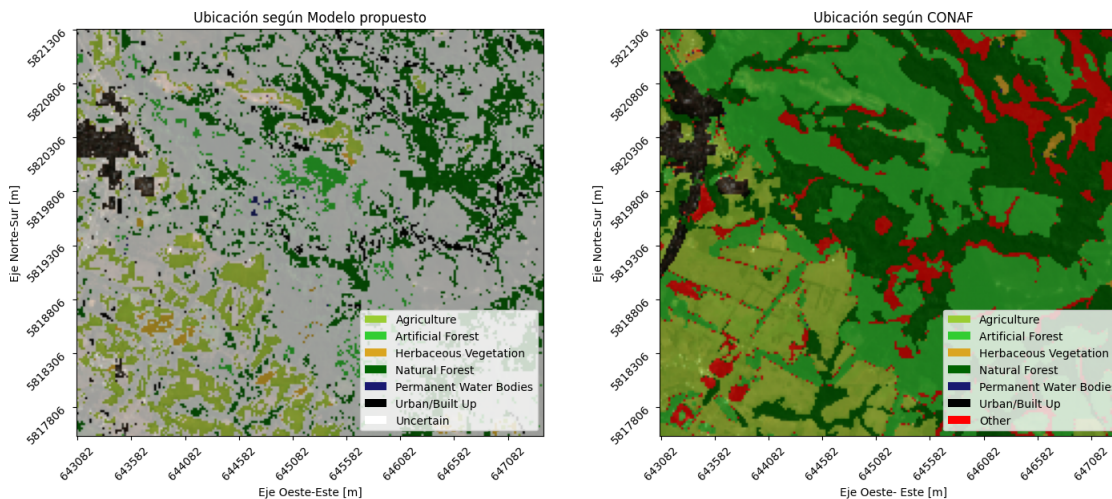


Figura 4.40: Cuarto ejemplo aplicado a generación de mapas, modelo entrenado directamente con datos CONAF.

Por otro lado, para el análogo utilizando el modelo entrenado con datos adquiridos usando el método 2 (figura 4.41), se logra una cobertura del 89 %:

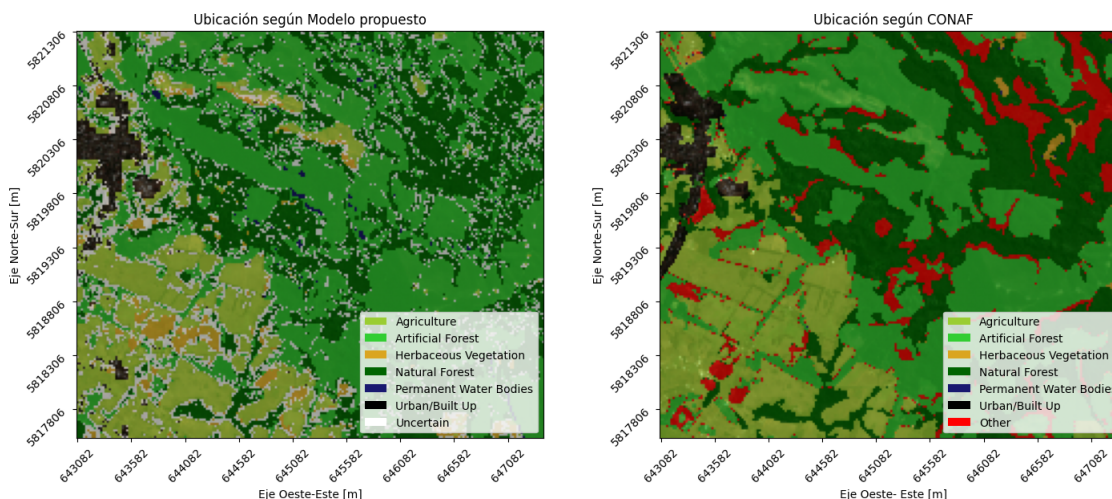


Figura 4.41: Cuarto ejemplo aplicado a generación de mapas, modelo entrenado con datos limpios (método 2).

Las métricas resultantes para esta zona geográfica en los pixeles altamente representativos se hallan en la tabla 4.14

Tabla 4.14: Resultados modelos RF, cuarta zona geográfica.

Modelo	Weighted Precision [%]	Weighted Recall [%]	Weighted F1 [%]
Entrenado con datos CONAF sin filtrar	89	85	85
Entrenado con datos filtrados (M2)	89	87	88

Finalmente, la quinta área de interés cuenta con un principalmente dos clases, zona urbana

y bosque artificial, con una zona importante clasificada incorrectamente como bosque artificial por conaf (no incluida en el conjunto de prueba, gracias a la limpieza), para el caso del modelo entrenado con datos sin ningún tipo de limpieza se logra una cobertura del 33 %.

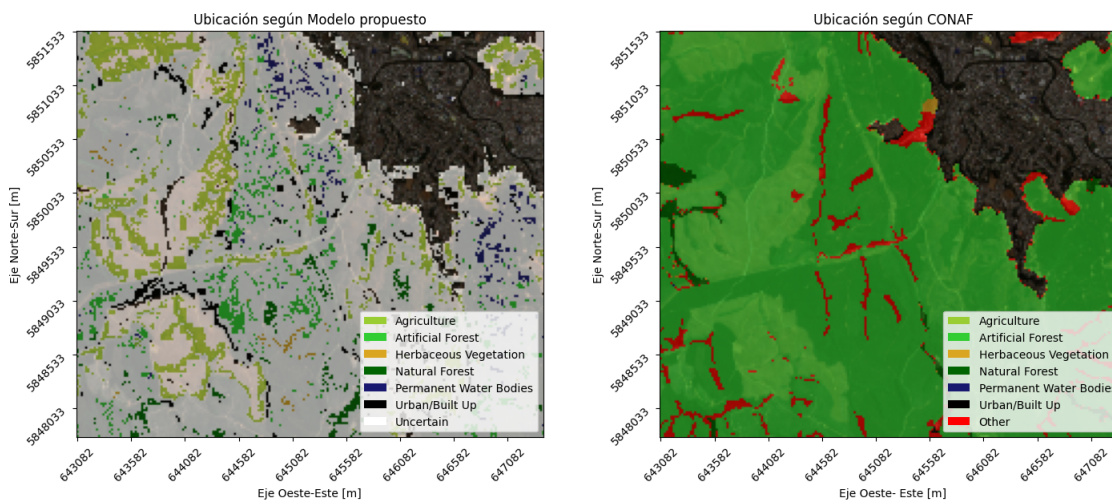


Figura 4.42: Quinto ejemplo aplicado a generación de mapas, modelo entrenado directamente con datos CONAF.

Por otro lado, para el análogo utilizando el modelo entrenado con datos limpios (método 2), se logra una cobertura del 90 %, existiendo gran contraste entre las figuras 4.42 y 4.43

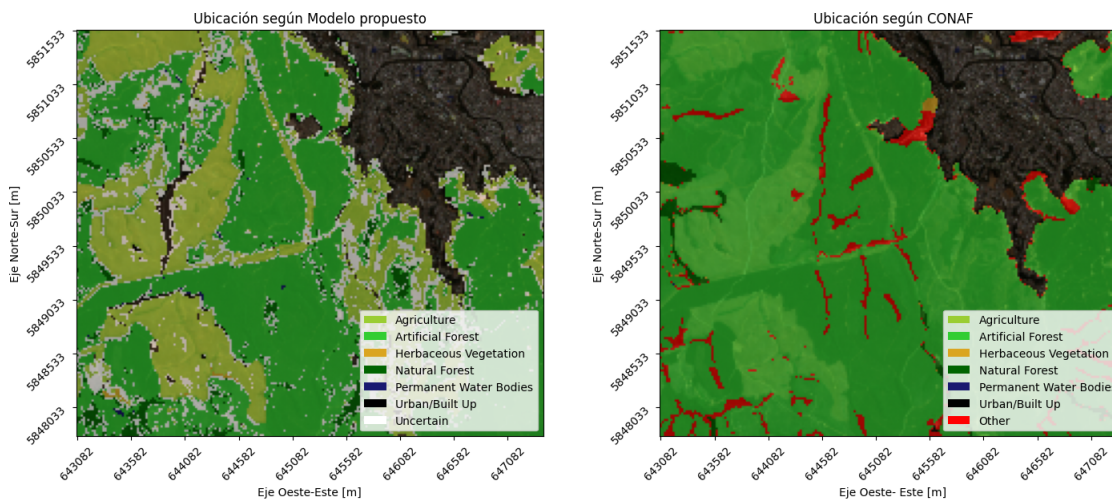


Figura 4.43: Quinto ejemplo aplicado a generación de mapas, modelo entrenado con datos limpios (método 2).

Las métricas resultantes para esta zona geográfica en los pixeles altamente representativos se encuentran detalladas en la tabla 4.15

Tabla 4.15: Resultados modelos RF, quinta zona geográfica.

Modelo	<i>Weighted Precision</i> [%]	<i>Weighted Recall</i> [%]	<i>Weighted F1</i> [%]
Entrenado con datos CONAF sin filtrar	96	77	81
Entrenado con datos filtrados (M2)	100	87	93

#### 4.2.2.9. Otros tipos de algoritmos utilizando pixeles representativos

Un modelo del tipo Random Forest está formado por múltiples árboles de decisión, sin embargo es posible entrenar uno sólo de estos árboles (CART), lo cual conlleva un desempeño menor en cuanto a clasificación:

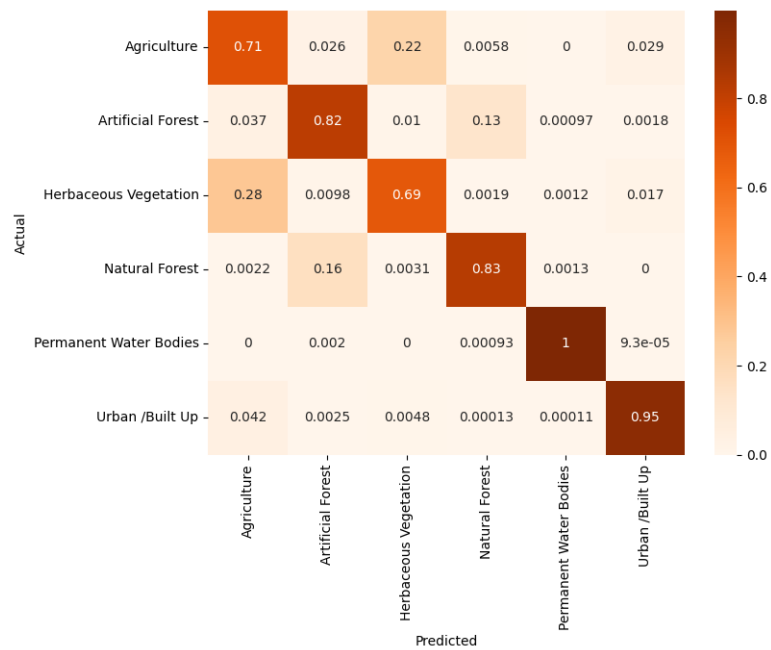


Figura 4.44: Matriz de confusión para modelo CART.

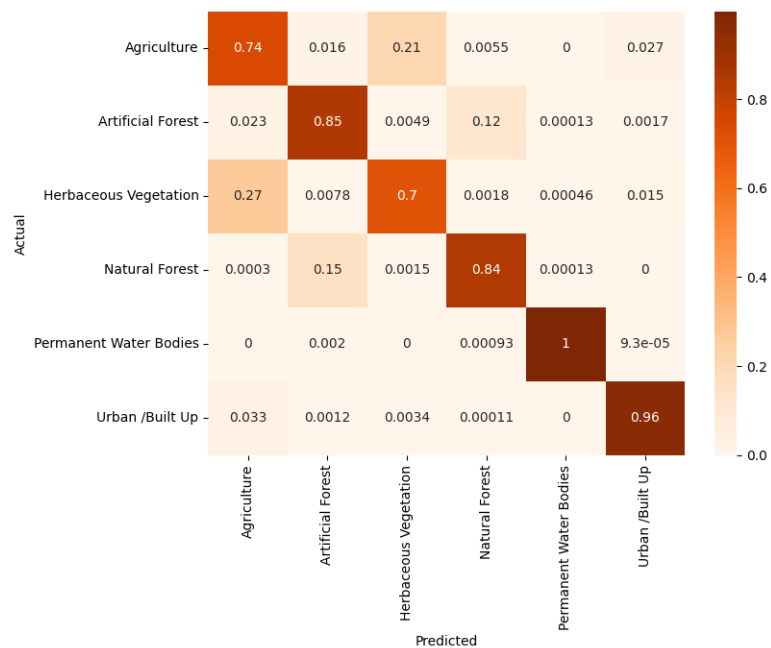


Figura 4.45: Matriz de confusión para modelo CART, con un umbral de certeza igual a 60 %.

Se prueba también con un modelo del tipo *gradient boosted trees*(ver marco teórico).

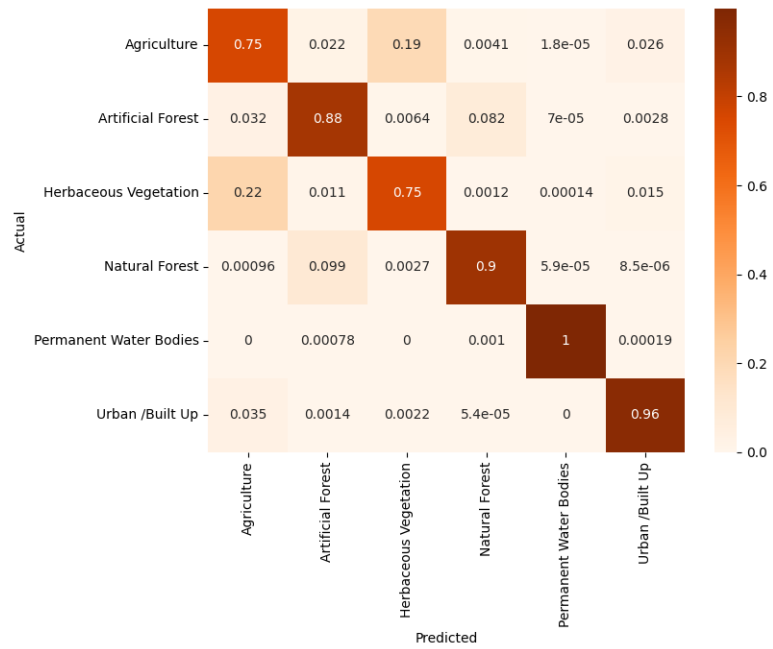


Figura 4.46: Matriz de confusión para modelo Gradient Boosted Trees.

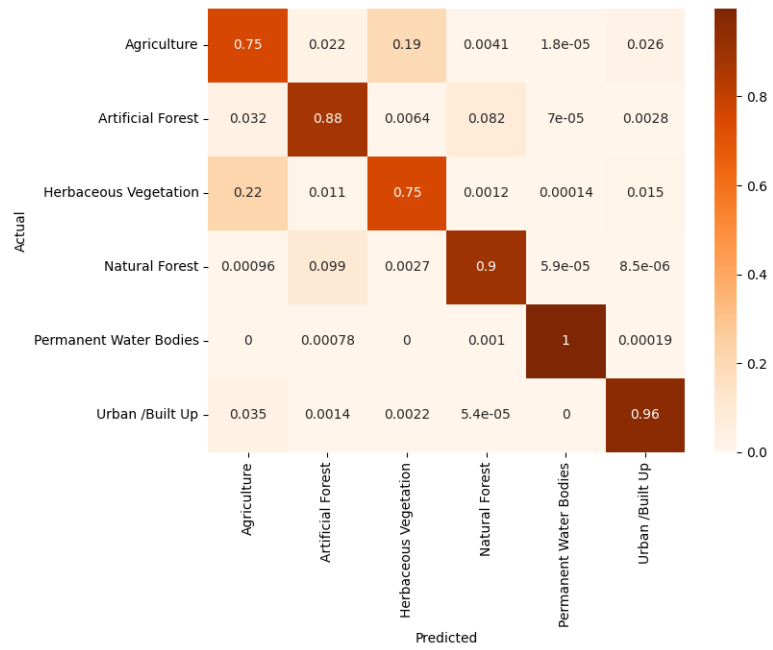


Figura 4.47: Matriz de confusión para modelo Gradient Boosted Trees, con un umbral de certeza igual a 60%.

Al igual que para las redes neuronales convolucionales, se prueba con incluir índices espectrales (los mismos que en CNN), para un modelo del tipo Random Forest:

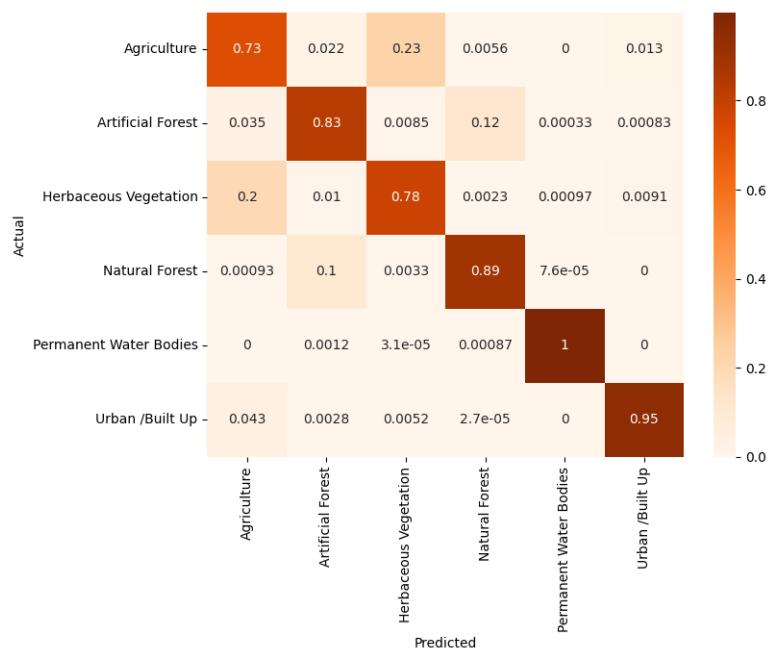


Figura 4.48: Matriz de confusión para modelo Random Forest con índices espectrales.

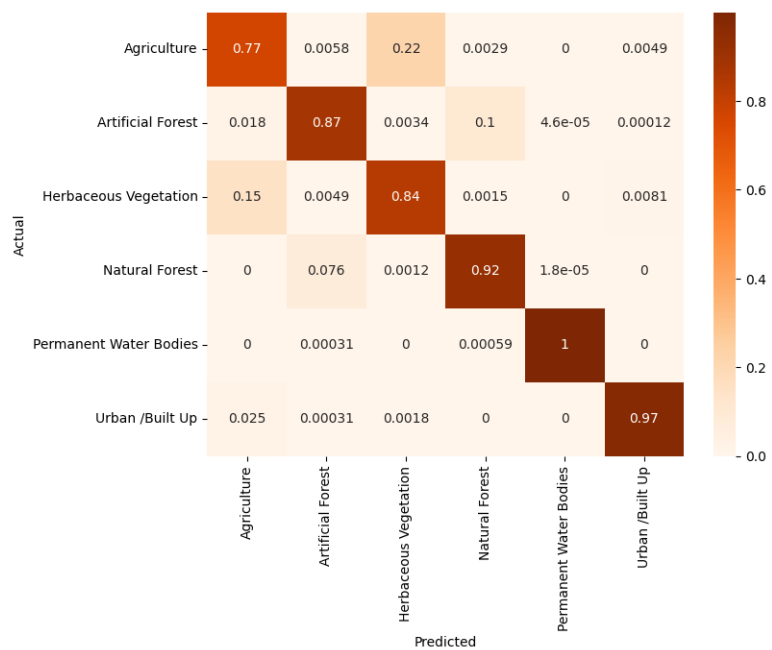


Figura 4.49: Matriz de confusión para modelo Random Forest con índices espectrales, utilizando un umbral de certeza igual a 60%.

En último lugar, se prueba un modelo perceptrón multicapa con dos capas ocultas, una de 300 y otra de 100 neuronas, obteniéndose las matrices ilustradas a continuación:

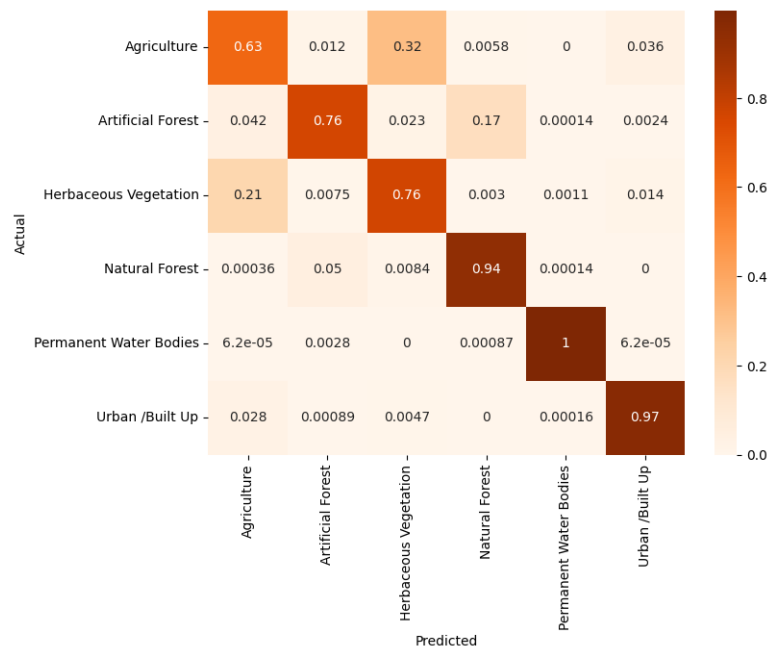


Figura 4.50: Matriz de confusión para modelo perceptrón multicapas.



Figura 4.51: Matriz de confusión para modelo perceptrón multicapas, con un umbral de certeza igual a 60 %.



# 5 | Análisis de los resultados

## 5.1. Redes Convolucionales

### 5.1.1. Desempeño de redes convolucionales

Tal como se puede apreciar en las tablas 5.1 y 5.2, Se logra una desempeño notablemente inferior en todas las métricas para cada arquitectura probada a la hora de comparar los modelos entrenados con datos CONAF con aquellos entrenados con datos EuroSat.

Utilizando sólo esta información no es evidente si esto se debe a una menor cantidad de datos o a una peor calidad de los mismos, sin embargo hay un aumento considerable en desempeño en modelos del tipo Random Forest al emplear datos limpiados automáticamente (ver figura 4.33 y tabla 4.8).

Dicha mejoría permite inferir que si bien podría ser posible un aumento en desempeño al incrementar la cantidad de datos utilizados, el mal etiquetado de zonas por parte de CONAF es una fuente considerable de error, explicando la drástica diferencia de métricas de desempeño entre las redes convolucionales entrenadas con datos CONAF en contraste a aquellas entrenadas con el dataset EuroSat.

Tabla 5.1: Métricas resultantes para modelos convolucionales entrenados usando dataset generado a partir de datos CONAF.

Modelo	Acc. a un nivel [%]	Acc. a dos niveles [%]	Macro Precision [%]	Macro Recall [%]	Macro F1 [%]
Base, máxima <i>accuracy</i>	62	80	63	60	58
Base, máximo <i>recall</i>	59	83	65	61	60
Rotacional, máxima <i>accuracy</i>	62	80	57	60	57
Rotacional, máximo <i>recall</i>	60	84	60	65	60
Con índices, máxima <i>accuracy</i>	60	82	55	57	54
Con índices, máximo <i>recall</i>	58	82	56	62	57

Tabla 5.2: Métricas resultantes para modelos convolucionales entrenados usando dataset EuroSat.

Modelo	Acc. a un nivel [%]	Acc. a dos niveles [%]	Macro Precision [%]	Macro Recall [%]	Macro F1 [%]
Base, máxima <i>accuracy</i>	92	98	92	92	92
Base, máximo <i>recall</i>	92	98	92	92	92
Rotacional, máximo <i>accuracy</i>	92	98	93	93	93
Rotacional, máximo <i>recall</i>	92	98	92	92	92
Con índices, máxima <i>accuracy</i>	92	98	92	92	92
Con índices, máximo <i>recall</i>	92	98	92	92	92

## 5.1.2. Limpieza automática

Tal como fue mencionado anteriormente, es posible apreciar un aumento considerable en desempeño en modelos del tipo Random Forest al emplear datos limpiados automáticamente (ver figura 4.33 y tabla 4.8).

Es posible a la vez ponderar ambos gráficos presentes en la figura 4.33, de modo de obtener el porcentaje de aciertos para cada nivel de certeza, tal como se puede apreciar en la figura 5.1.

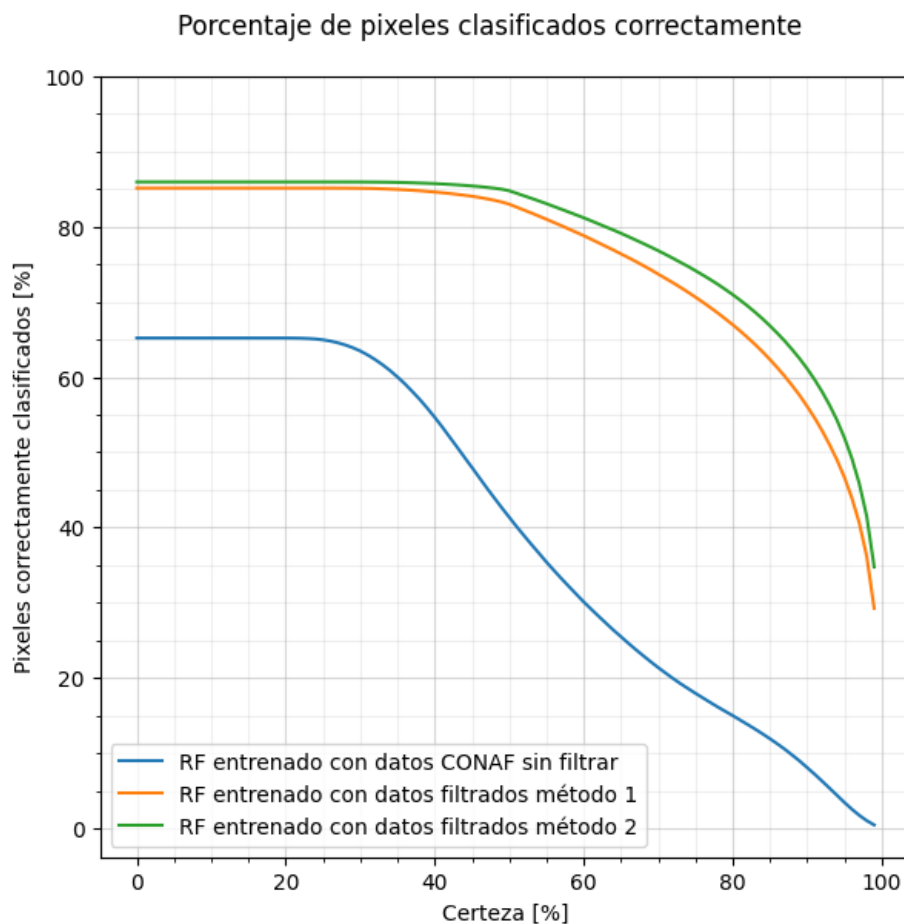


Figura 5.1: Porcentaje de pixeles clasificados correctamente para cada nivel de certeza.

Se desprende del gráfico anterior una relación proporcional entre la calidad de los datos y el poder de clasificación de los modelos, considerándose como beneficioso, al aumentar en un 20 % la cantidad de pixeles clasificados correctamente para baja certeza (65 % a 86 %) y en 50 % para pixeles clasificados con alta certeza (30 % a 80 % en 60 % de certeza).

Es de suma importancia recordar que el porcentaje de pixeles correctamente clasificados no es equivalente al *accuracy*, ya que a medida que aumenta la certeza disminuye la cantidad de pixeles clasificados pero el *accuracy*, al contrario que la cantidad de pixeles correctamente clasificados, aumenta (ver figura 4.33).

## 5.2. Algoritmos para clasificación automática de cobertura/uso de suelos

Se entrenan y comparan cinco modelos utilizando los mismos datos de entrenamiento y de prueba, lográndose un desempeño similar entre los 3 modelos formados por múltiples árboles de decisión (Random Forest, Boosted Gradient Trees y Random Forest con índices espectrales).

Tal como se puede apreciar en la tabla , el modelo para el cual se obtuvo mejor desempeño (en cada una de las cuatro métricas empleadas) es el modelo *Gradient Boosted Trees*, modelo diseñado de modo que cada nuevo árbol generado complemente las clasificaciones incorrectas realizadas por el árbol anterior (a diferencia de modelos Random Forest, en los cuales cada árbol es generado aleatoriamente) [18].

Tabla 5.3: Métricas resultantes para modelos clasificadores de píxeles entrenados usando dataset limpiado automáticamente.

Modelo	Acc. a un nivel [%]	Macro Precision [%]	Macro Recall [%]	Macro F1 [%]
Random Forest	86	76	86	79
CART	83	74	83	76
Boosted Gradient Trees	88	78	87	81
Random Forest con índices	85	76	86	79
Perceptrón Multicapa	81	72	84	74

Al comparar los resultados presentados en la tabla 5.3 con aquellos presentados en la tabla 5.4 se puede observar un aumento en desempeño en todas las métricas, para todos los modelos al incrementar el requerimiento de certeza. Esta diferencia justifica la utilización de la clase *uncertain* a la hora de clasificar cobertura/uso de suelos.

Tabla 5.4: Métricas resultantes para modelos clasificadores de píxeles entrenados usando dataset limpiado automáticamente, umbral de certeza igual al 60 %.

Modelo	Acc. a un nivel [%]	Macro Precision [%]	Macro Recall [%]	Macro F1 [%]
Random Forest	89	80	89	83
CART	85	76	85	79
Boosted Gradient Trees	91	82	90	84
Random Forest con índices	89	80	90	83
Perceptrón Multicapa	86	78	86	79

Se tienen resultados básicamente idénticos para el modelo Random Forest común y aquel en el cual se hace inclusión de índices espectrales, probablemente debido a un grado de redundancia en la información discriminatoria disponible para estos modelos, naturalmente el modelo CART entrega resultados inferiores a los demás modelos basados en árboles de decisión, al ser equivalente a un modelo de Random Forest con un solo árbol. El modelo con mejor desempeño en cada una de las métricas empleadas corresponde al modelo Boosted Gradient Trees, con un *accuracy* de 91 %.

## 6 | Conclusión

Se logra utilizar redes neuronales convolucionales entrenadas sobre datos subóptimos para realizar una limpieza automática de dichos datos (véanse figuras 4.17 y 4.18), adquiriéndose múltiples respuestas espectrales representativas para las clases utilizadas. A partir de esta limpieza, se logra caracterizar la distribución de respuestas clase a clase (figuras 4.19, 4.20, 4.21, 4.22, 4.23 y 4.24), así como entrenar algoritmos de clasificación automática de cobertura y uso de suelos, permitiendo estudiar en la inclusión de los mismos parámetros tales como índices espectrales (tablas 5.3 y 5.4).

El impacto de la limpieza automática implementada constituye en un incremento significativo en *accuracy*, *precision*, *recall* y *F1* (ver tabla 4.8), logrando un *accuracy* comparable al del estado del arte (tabla 2.4) al utilizar las clases propuestas, así como también una mayor cobertura y cantidad de píxeles clasificados correctamente para niveles arbitrarios de certeza (figuras 4.33 y 5.1). Hasta el momento, los mejores resultados se obtuvieron luego del entrenamiento de una red convolucional, su utilización en la limpieza de datos automática (ver figura 3.1) mediante el denominado método 2 (ver figuras 4.33 y 5.1) y el uso de dichos datos en el entrenamiento de un algoritmo del tipo Boosted Gradient Trees (ver tablas 5.3 y 5.4).

La metodología utilizada, en particular el trabajarse únicamente con datos de acceso público (En cuanto a CONAF, tanto el catastro de cobertura/uso de suelo como sus actualizaciones, y en cuanto a imágenes satelitales aquellas disponibilizadas por Copernicus), permite que estos resultados sean replicables y aplicables a la creación de distintos tipos de algoritmos de clasificación automática de cobertura y uso de suelo (tablas 5.3 y 5.4). La posibilidad de utilizar datos asociados a una mayor extensión territorial y temporal permite asegurar la escalabilidad de la metodología diseñada, cumpliéndose en su totalidad el objetivo general de la presente memoria, así también como cada uno de los objetivos específicos.

Si bien se cuenta con resultados satisfactorios, la utilización de los datos obtenidos a través de la metodología propuesta en el presente informe podría permitir la obtención de modelos con mayor capacidad de discriminación al ajustar los parámetros disponibles para cada uno de los algoritmos utilizados (por ejemplo profundidad de árboles de decisión, cantidad de árboles, etc.), esta búsqueda de parámetros óptimos puede ser realizada con ayuda de algoritmos *tuner*.

Es posible concluir satisfactoriamente sobre la limpieza automática de datos y el desempeño de los algoritmos entrenados utilizando los datos resultantes, existiendo múltiples aplicaciones para estos últimos, tal como monitoreo del área vegetativa, disminución de cuerpos de agua debido a sequía o crecimiento de zonas urbanas, entre otras.

# Bibliografía

- [1] R.M. Hoffer and C.J. Johannsen, “[Ecological Potential in Spectral Signature Analysis](#),” Febrero de 1969.
- [2] R. Pace, F. Chiocchini, M. Sarti, T. A. Endreny, C. Calfapietra, and M. Ciolfi, “[Integrating Copernicus land cover data into the i-Tree Cool Air model to evaluate and map urban heat mitigation by tree cover](#),” *European Journal of Remote Sensing*, vol. 0, no. 0, pp. 1–18, 2022.
- [3] Z. Zomlot, B. Verbeiren, M. Huysmans, and O. Batelaan, “[Trajectory analysis of land use and land cover maps to improve spatial–temporal patterns, and impact assessment on groundwater recharge](#),” *Journal of Hydrology*, vol. 554, pp. 558–569, 2017.
- [4] P. H. Verburg, K. Neumann, and L. Nol, “[Challenges in using land use and land cover data for global change studies](#),” *Global Change Biology*, vol. 17, no. 2, pp. 974–989, 2011.
- [5] P. E. Zope, T. I. Eldho, and V. Jothiprakash, “[Hydrological impacts of land use–land cover change and detention basins on urban flood hazard: a case study of Poisar River basin, Mumbai, India](#),” *Natural Hazards*, vol. 87, no. 3, pp. 1267–1283, Jul 2017.
- [6] Martin Machala and Lucie Zejdová, “[Forest Mapping Through Object-based Image Analysis of Multispectral and LiDAR Aerial Data](#),” *European Journal of Remote Sensing*, vol. 47, pp. 117–131, 17 Feb 2017.
- [7] K. Bereta, H. Caumont, U. Daniels, E. Goor, M. Koubarakis, D.-A. Pantazi, G. Stamoulis, S. Ubels, V. Venus, and F. Wahyudi, “[The Copernicus App Lab project: Easy access to Copernicus data.](#),” in *EDBT*, pp. 501–511, 2019.
- [8] S. Jutz and M.P. Milagro-Pérez, “[Copernicus: the European Earth Observation programme](#),” *Revista de Teledetección*, vol. 56, 27 Nov 2020.
- [9] J. H. P. y Daniel Montaner F, “[Patrones de Respuesta Espectral](#),” Lab. Geomática y Ecología del Paisaje (GEP) – Facultad de Ciencias Forestales Universidad de Chile.
- [10] S. Huang, L. Tang, J. P. Hupy, Y. Wang, and G. Shao, “[A commentary review on the use of normalized difference vegetation index \(NDVI\) in the era of popular remote sensing](#),” *Journal of Forestry Research*, vol. 32, pp. 1–6, Feb 2021.
- [11] D. Roy, L. Boschetti, and S. Trigg, “[Remote sensing of fire severity: assessing the performance of the normalized burn ratio](#),” *IEEE Geoscience and Remote Sensing Letters*, vol. 3, no. 1, pp. 112–116, 2006.
- [12] S. Jin, L. Yang, G. Xian, P. Danielson, and C. Homer, “[A Multi-Index Integrated Change Detection Method for Updating the National Land Cover Database](#),” *AGU Fall Meeting Abstracts*, 12 2010.

- [13] C. Relays, “[Acerca de Copernicus Chile](#),” 2018.
- [14] F. Murtagh, “[Multilayer perceptrons for classification and regression](#),” *Neurocomputing*, vol. 2, no. 5, pp. 183–197, 1991.
- [15] K. O’Shea and R. Nash, “[An Introduction to Convolutional Neural Networks](#),” 2015.
- [16] W.-Y. Loh, “[Fifty Years of Classification and Regression Trees](#),” *International Statistical Review*, vol. 82, no. 3, pp. 329–348, 2014.
- [17] L. Breiman, “[Random Forests](#),” *Machine Learning*, vol. 45, pp. 5–32, Oct 2001.
- [18] Z. Zhang and C. Jung, “[GBDT-MO: Gradient-Boosted Decision Trees for Multiple Outputs](#),” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, pp. 3156–3167, 2021.
- [19] M. Grandini, E. Bagli, and G. Visani, “Metrics for multi-class classification: an overview,” 2020.
- [20] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage, “[Sentinel-2 data for land cover/use mapping: A review](#),” *Remote Sensing*, vol. 12, no. 14, p. 2291, 2020.
- [21] R. Saavedra, “[Detección Espectral de Especies Nativas a partir de datos Hiperespectrales HySpex y LiDAR en Bosques Templados al Sur de Chile](#),” 2022.
- [22] CONAF, “[Fichas Regionales Catastro Vegetacional](#),” 2015\*.
- [23] CONAF, “[Monitoreos de cambio de uso de la tierra](#),” 2017.
- [24] CONAF, “[Metodología Multi-índices para la detección de cambios de usos de la tierra](#),” 2017.
- [25] Copernicus, “[Copernicus Global Land Service](#),” 2015-2019.
- [26] D. Phiri, M. Simwanda, S. Salekin, V. R. Nyirenda, Y. Murayama, and M. Ranagalage, “[Sentinel-2 Data for Land Cover/Use Mapping: A Review](#),” *Remote Sensing*, vol. 12, no. 14, 2020.
- [27] E. Copernicus, “[Acerca de Copernicus Chile](#),” 2016.
- [28] CONAF, “[Monitoreo de cambios, Corrección Cartográfica y Actualización del Catastro de los Recursos Vegetacionales Nativos de la región de Biobío](#),” 2017.
- [29] A. Varshney, “[Improved NDBI differencing algorithm for built-up regions change detection from remote-sensing data: an automated approach](#),” *Remote Sensing Letters*, vol. 4, no. 5, pp. 504–512, 2013.
- [30] T. Hung, “[Urban Bare Land Classification Using NDBaI Index Based on Combination of Sentinel 2 MSI and Landsat 8 Multiresolution Images](#),” *VNU Journal of Science: Earth and Environmental Sciences*, vol. 36, no. 2, 2020.
- [31] Z. Wang, J. Liu, J. Li, and D. D. Zhang, “[Multi-Spectral Water Index \(MuWI\): A Native 10-m Multi-Spectral Water Index for Accurate Water Mapping on Sentinel-2](#),” *Remote Sensing*, vol. 10, no. 10, 2018.
- [32] C. Nacional, “[Ley sobre recuperación del bosque nativo y fomento forestal](#),” 2017.
- [33] E. Reyes, P. A. Estévez, I. Reyes, G. Cabrera-Vives, P. Huijse, R. Carrasco, and F. Förster, “[Enhanced Rotational Invariant Convolutional Neural Network for Supernovae De-](#)

- tection,” in *2018 International Joint Conference on Neural Networks (IJCNN)*, IEEE, July 2018.
- [34] M. Guillame-Bert, S. Bruch, R. Stotz, and J. Pfeifer, “**Yggdrasil Decision Forests: A Fast and Extensible Decision Forests Library**,” in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pp. 4068–4077, 2023.

# Anexos

## Anexo A. Mapas generados (RF y GBDT)

A continuación se muestra una tabla resumen de las métricas resultantes para el modelo Random Forest y el modelo Gradient Boosted Trees al realizar predicciones sobre las mismas zonas geográficas:

Modelo	Imagen	<i>Accuracy</i> [%]	<i>Weighted Precision</i> [%]	<i>Weighted Recall</i> [%]	<i>Weighted F1</i> [%]
RF	1	89	94	89	91
	2	93	96	93	94
	3	98	99	98	98
	4	87	89	87	88
	5	87	100	87	87
GBDT	1	93	95	93	93
	2	96	97	96	96
	3	98	99	98	99
	4	88	90	88	89
	5	90	100	90	94

Se puede apreciar un desempeño ligeramente superior de manera transversal para el modelo GBDT, con resultados similares gráficamente:

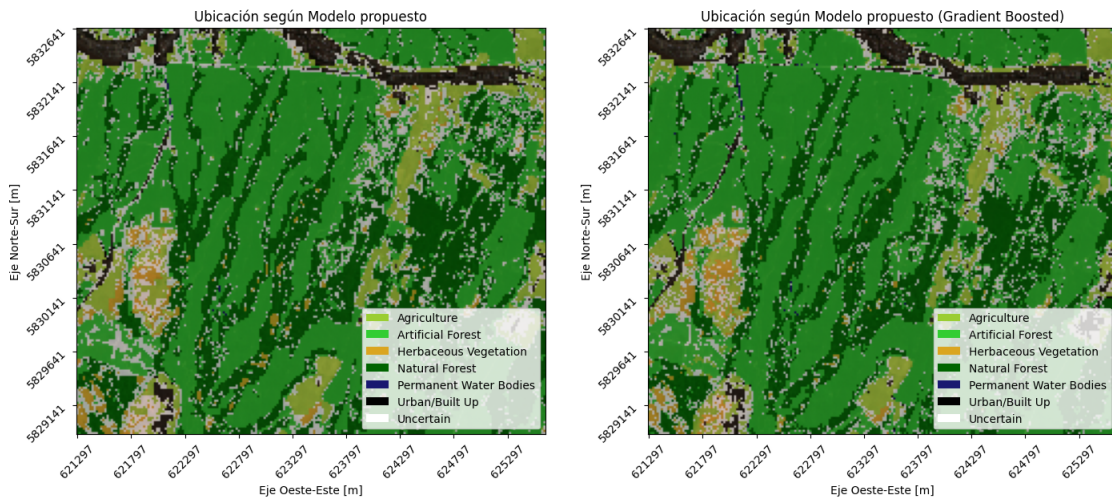


Figura A.1: Comparación predicciones random forest común y gradient boosted, primer ejemplo.



Para la segunda zona de interés se puede apreciar un grosor ligeramente mayor para el río que la atraviesa, nuevamente se puede apreciar una no-categorización de caminos de tierra.

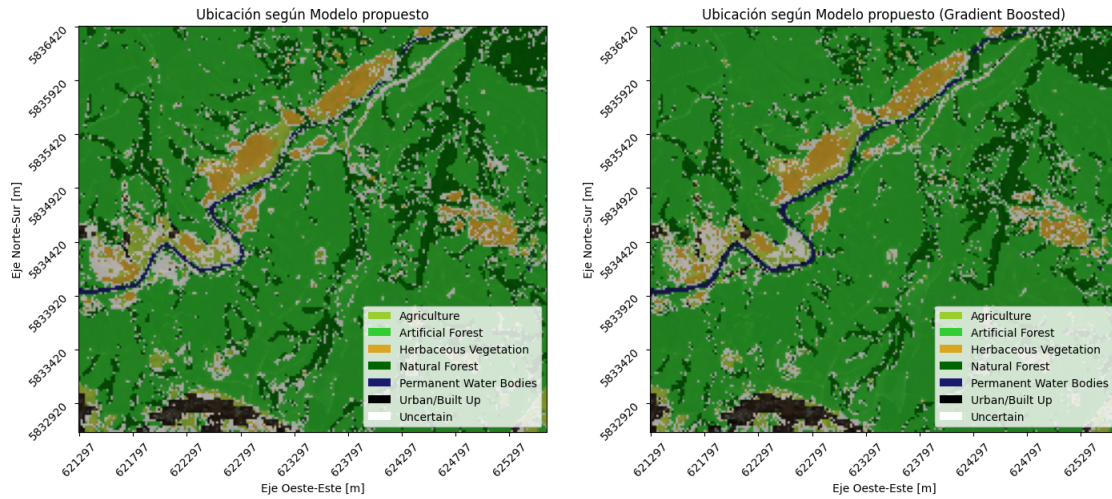


Figura A.2: Comparación predicciones random forest común y gradient boosted, segundo ejemplo.

Por otro lado, en la tercera zona de interés se observan predicciones muy cercanas entre sí, salvo la tendencia del modelo GBDT en clasificar una mayor cantidad de píxeles como zona urbana.

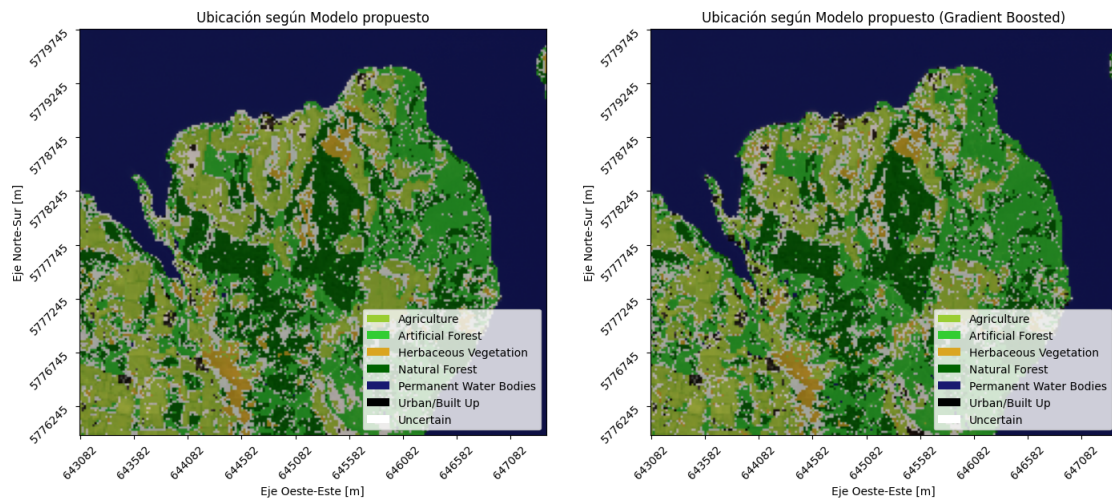


Figura A.3: Comparación predicciones random forest común y gradient boosted, tercer ejemplo.

Finalmente, para tanto el penúltimo como último ejemplo, no se logra detectar una diferencia significativa.

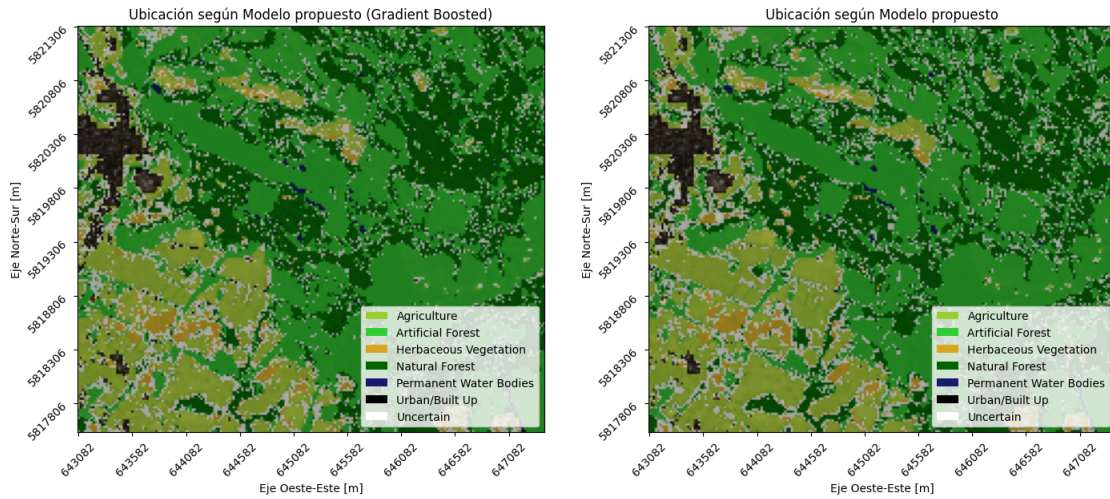


Figura A.4: Comparación predicciones random forest común y gradient boosted, cuarto ejemplo.

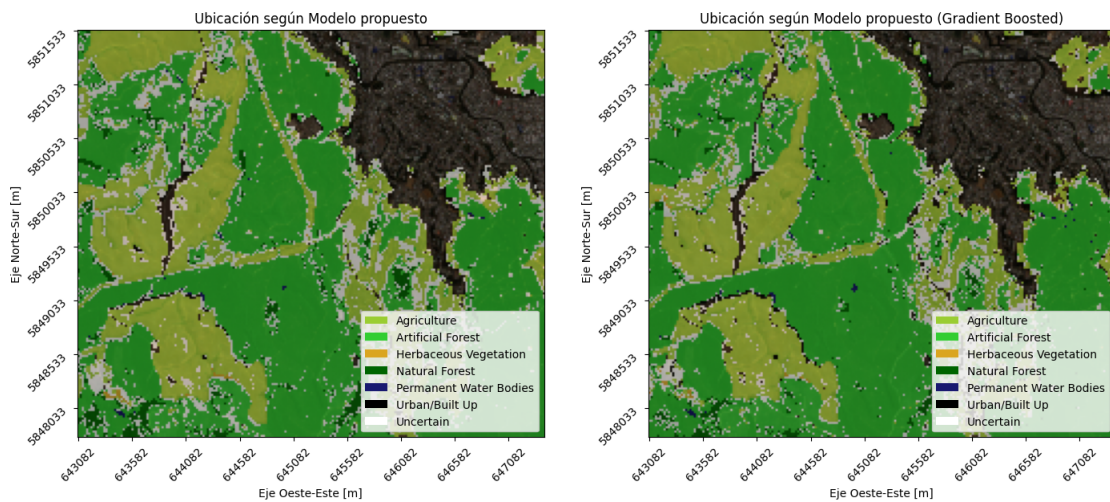


Figura A.5: Comparación predicciones random forest común y gradient boosted, quinto ejemplo.