



TESIS DOCTORAL

Método de identificación de cultivos mediante el análisis masivo multitemporal de imágenes multiespectrales de satélite

MARCOS FERNÁNDEZ SELLERS

PROGRAMA DE DOCTORADO EN TECNOLOGÍAS INFORMÁTICAS (TIN)

Con la conformidad del director Dr. Adolfo J. Lozano Tello

Esta tesis cuenta con la autorización del director de la misma y de la Comisión Académica del programa. Dichas autorizaciones constan en el Servicio de la Escuela Internacional de Doctorado de la Universidad de Extremadura.

2023

A mi madre, un ejemplo de resistencia y valentía.

A mi padre, un referente personal y profesional.

A Clara, por su incansable apoyo.

Agradecimientos

Una vez más, se cumple el tópico de no haber imaginado hace años la llegada de este momento, en el que finalizo la redacción de mi Tesis Doctoral. En estos casos suele haber un punto de inflexión. El mío fue un correo electrónico, allá por septiembre de 2016, en el que consultaba la posibilidad de realizar mi Trabajo Fin de Grado con mi entonces tutor, Adolfo Lozano. La (buena) relación académica se convirtió también en relación laboral, llegando a la dirección y tutorización de este trabajo. Mi primer agradecimiento debe ser para él, por apostar por mí, darme su confianza y haberme dado la oportunidad de aprender y trabajar con él.

Mi otro agradecimiento hacia Adolfo va por mi integración en el grupo Quercus de Ingeniería del Software de la Universidad de Extremadura, donde he tenido el privilegio de compartir tiempo y conocimientos con grandes profesionales, y que de una u otra manera me han aportado tanto profesional como personalmente durante estos años. Aunque no pueda citarlos a todos, sí que me gustaría agradecer de manera especial a Sergio Laso, por su desinteresada colaboración, y a aquellos cuyo compañerismo ya ha trascendido la frontera de la amistad: Enrique Moguel, Julio Acedo y Jorge Perriñez.

La elaboración de este trabajo tampoco hubiera sido posible sin mis compañeros del proyecto InteliPAC: Andrés Caballero, Guillermo Siesto y Miguel Ros, con sus largas horas de trabajo en nuestro querido *zulo*; y la inestimable colaboración de Elia Quirós y Laura Fragoso.

Ya en el plano más personal, agradecer a mis compañeros y amigos que me han acompañado durante todos estos años, siempre con mis dificultades de agenda, con una mención especial a Francisco Pajuelo, por todo su apoyo y sus acertados consejos.

Y siguiendo con los tópicos, se dice que los últimos serán los primeros. La vida nos ha dado muchas vueltas durante el tiempo que ha durado este trabajo, especialmente en los últimos meses, en los que me he sentido muy arropado por toda mi familia, pero especialmente por tres personas. No sé si tendré reencarnaciones suficientes para compensar todo el apoyo brindado y todo el tiempo que no os he podido dedicar durante todo este periodo. Este también es el fruto de vuestro esfuerzo y trabajo, incansable durante tantos años. A Carmen, que no pudo conocer a su niño siendo doctor, a César y Clara, y a la memoria de aquellos que ya no nos acompañan, pero que estarán igual de orgullosos desde ahí arriba.

A todos, de corazón, gracias.

Marcos.

Resumen

La detección de cultivos mediante imágenes de satélite ha experimentado un gran auge en los últimos años, debido principalmente a dos factores: por un lado, las nuevas misiones satelitales, que ofrecen gratuitamente imágenes de satélite de manera periódica y con una buena resolución de imagen; y, por otro lado, la popularización de sistemas de aprendizaje máquina de gran precisión aplicados a la monitorización de cultivos. En muchos trabajos relacionados se pueden apreciar numerosas propuestas referidas a la detección de cultivos, pero que se centran en cultivos, regiones o técnicas concretas. La literatura carece de una metodología completa que permita implementar un sistema de identificación de cultivos desde cero, con la descripción secuencial de fases y técnicas que se deben seguir, y que pueda adaptarse de manera concreta a las necesidades de cada proyecto.

Esta tesis doctoral propone un método completo de identificación de cultivos, haciendo uso de imágenes de satélite multiespectrales (que disponen de varias bandas de información), y realizando el análisis de las imágenes a lo largo del tiempo. El método comprende una serie de fases y recomendaciones, que pueden adaptarse a los requisitos del proyecto que se desea implementar. Partiendo de la identificación de requisitos del proyecto y del diseño del sistema de información, el método propone cómo seleccionar y estandarizar la información de los recintos agrícolas que se desea analizar. Esa información es utilizada para la descarga de imágenes de satélite y su tratamiento, incluyendo el recorte y normalización de las imágenes, así como el etiquetado de los datos descargados. En las siguientes fases, se proponen técnicas para la selección y filtrado de datos, detección de anomalías e interferencias de los datos. Desde estos datos, el método

propone la generación de imágenes sintéticas multitemporales y multiespectrales para el aprendizaje, y todas las cuestiones relacionadas con el diseño del sistema. Finalmente, la propuesta incluye un método específico para determinar el periodo de detección adecuado para cada cultivo mediante técnicas heurísticas.

El resultado de aplicar el método consiste en un sistema software completo para la identificación de cultivos. Los procesos de análisis se centran a nivel de píxel, de manera masiva y automática, y generando modelos de aprendizaje máquina para la identificación de cultivos en función de las necesidades del proyecto. Este trabajo incluye también un caso de uso para probar su aplicabilidad a un proyecto real.

Abstract

Crop detection using satellite images has experienced a great increase in recent years, due to two main factors: on the one hand, the new satellite missions, which offer free good resolution satellite images regularly; and on the other hand, the popularization of high-precision machine learning systems applied to crop monitoring. Several proposals related to crop detection can be appreciated in many related works, but they focus on specific crops, regions, or techniques. The literature lacks a complete methodology that allows the implementation of a crop identification system from scratch, sequentially describing the phases and techniques to follow, and that can be specifically adapted to the needs of each project.

This doctoral thesis proposes a complete method for crop identification, using multispectral satellite images, which have several bands of information, and analysing the images over time. The method embodies different phases and recommendations, adapting each of them to the requirements of the project that needs to be implemented. Starting from the identification of project requirements and the design of the information system, the method proposes how to select and standardize the information of the agricultural plots. This information is used for the download of satellite images and their treatment, including image cropping and normalization, as well as the labelling of the downloaded data. In the following phases, techniques for data selection, filtering and anomaly detection are proposed. From these data, the method proposes the generation of multitemporal and multispectral synthetic images for the learning, and all considerations related to the system design. Finally, the proposal includes a specific method to determine the suitable sensing time period for each crop identification

using heuristic techniques.

The result of applying the method consists of a complete software system for crop identification. The analysis processes are focused on a pixel level, in a massive and automatic way, and generate machine learning models for crop identification based on the needs of the project. This work also includes a use case to test its applicability to a real project.

Glosario de términos

A

área de trabajo:

Región geográfica donde se desea aplicar el método de detección de cultivos. Puede ser, por ejemplo, una comarca, una comunidad autónoma, o un país.

B

banda de información:

Información proveniente de un sensor de satélite que comprende un rango definido del espectro electromagnético. También conocida como banda espectral.

C

coordenada relativa de píxel:

Posición que ocupa un píxel en una imagen de zona, en formato fila y columna, considerando la imagen de zona como una matriz de píxeles.

H

hiperparámetros:

Valores de configuración utilizados durante el proceso de entrenamiento de un sistema de aprendizaje máquina. Generalmente indicado por el científico de datos.

I

imagen de zona:

Imagen satelital recortada con las mismas dimensiones que una zona geométrica.

Imagen Sintética Multitemporal y Multiespectral (ISMM):

Matriz de dos dimensiones, cuyos ejes se corresponden con los periodos temporales y con la información de las bandas de información, respectivamente, que incluye datos extraídos de las imágenes de satélite sobre un píxel concreto a lo largo del periodo.

M

modelo de aprendizaje automático:

Archivo entrenado mediante un algoritmo de aprendizaje máquina, empleando un conjunto de datos, y utilizado para reconocer determinados patrones de un cultivo.

multiespectral:

Que abarca información de varios espectros o bandas de información de las imágenes de satélite.

N

normalización:

Proceso mediante el cual se unifica la resolución de las distintas bandas de información de una imagen de satélite.

P

periodo temporal:

Rango de fechas que abarca una o varias imágenes satelitales. Será utilizado para agrupar datos de las imágenes de satélite y elaborar las ISMM.

píxel:

Superficie homogénea más pequeña de las que componen una imagen de satélite, cuyos valores se obtienen de las mediciones de los sensores del satélite en un punto geográfico concreto.

R**recinto geométrico:**

Conjunto de coordenadas que representan geoméricamente el polígono de un recinto agrícola. Es decir, una parcela de cultivo delimitada geográficamente.

resolución:

Tamaño de un píxel de la imagen satelital, medido en metros sobre el terreno.

S**serie temporal:**

Sucesión de datos medidos en determinados momentos en un lugar concreto y ordenados cronológicamente.

U**umbral de acierto:**

Porcentaje mínimo de acierto requerido para un modelo de aprendizaje máquina.

Z**zona geométrica:**

Polígono geométrico de coordenadas con un tamaño determinado. Se utiliza para dividir el área de trabajo en regiones con un tamaño menor para obtener las imágenes de zona.

Índice general

Resumen	v
Abstract	vii
Glosario	ix
Índice de figuras	xvii
1. Introducción	1
1.1. Contexto	1
1.2. Estructura de la Tesis	7
1.3. Publicaciones relacionadas	8
2. Estado del arte	9
2.1. Síntesis de la evidencia disponible	9
2.2. Conclusiones	18
3. Objetivos y contribuciones	21
3.1. Objetivos	21
3.2. Contribuciones al estado del arte	23

3.3. Premisas	24
3.4. Hipótesis	25
3.5. Restricciones	25
4. Método de identificación de cultivos	27
4.1. Análisis de requisitos y diseño del sistema de información (M1)	29
4.1.1. Identificación de requisitos y limitaciones (M1.1) . .	29
4.1.2. Diseño del sistema de información (M1.2)	31
4.2. Organización de los datos fuente de recintos (M2)	32
4.2.1. Selección y estandarización de recintos identificables (M2.1)	32
4.2.2. Creación de zonas geométricas (M2.2)	33
4.2.3. Emparejamiento de recintos y zonas (M2.3)	35
4.3. Descarga masiva y tratamiento preliminar de datos (M3) . .	37
4.3.1. Descarga masiva de datos (M3.1)	37
4.3.2. Recorte y normalización de las imágenes descargadas (M3.2)	38
4.3.3. Emparejamiento y etiquetado de los datos (M3.3) .	39
4.4. Selección, agrupación y filtrado de datos (M4)	41
4.4.1. Selección de datos (M4.1)	41
4.4.2. Eliminación de datos con ruido (M4.2)	42
4.4.3. Agrupación de los datos (M4.3)	43
4.4.4. Interpolación de los datos (M4.4)	45
4.4.5. Reducción de anomalías e interferencias (M4.5) . . .	46
4.5. Configuración de los ejemplos de entrada para el aprendizaje (M5)	47

4.5.1.	Generación de imágenes sintéticas (M5.1)	47
4.5.2.	Especificación de los modelos para el entrenamiento (M5.2)	49
4.6.	Diseño del sistema de aprendizaje (M6)	50
4.6.1.	Selección y configuración del sistema de aprendizaje (M6.1)	50
4.6.2.	Diseño del entrenamiento de los modelos (M6.2)	51
4.7.	Búsqueda del periodo de detección de los cultivos (M7)	53
4.7.1.	Establecer función heurística y de parada (M7.1)	54
4.7.2.	Definición del estado inicial (M7.2)	54
4.7.3.	Ejecución del algoritmo de detección del periodo (M7.3)	55
4.8.	Identificación de los cultivos (M8)	57
5.	Evaluación	61
5.1.	Análisis de requisitos y diseño del sistema de información	62
5.1.1.	Identificación de requisitos y limitaciones	62
5.1.2.	Diseño del sistema de información	63
5.2.	Organización de los datos fuente de recintos	63
5.2.1.	Selección y estandarización de recintos identificables	63
5.2.2.	Creación de zonas geométricas	66
5.2.3.	Emparejamiento de recintos y zonas	68
5.3.	Descarga masiva y tratamiento preliminar de datos	69
5.3.1.	Descarga masiva de datos	69
5.3.2.	Recorte y normalización de las imágenes descargadas	70
5.3.3.	Emparejamiento y etiquetado de los datos	71
5.4.	Selección, agrupación y filtrado de datos	71

5.4.1.	Selección de datos	71
5.4.2.	Eliminación de datos con ruido	72
5.4.3.	Agrupación de los datos	72
5.4.4.	Interpolación de los datos	73
5.4.5.	Reducción de anomalías e interferencias	73
5.5.	Configuración de los ejemplos de entrada para el aprendizaje	75
5.5.1.	Generación de imágenes sintéticas	75
5.5.2.	Especificación de los modelos para el entrenamiento	76
5.6.	Diseño del sistema de aprendizaje	77
5.6.1.	Selección y configuración del sistema de aprendizaje	77
5.6.2.	Diseño del entrenamiento de los modelos	77
5.7.	Búsqueda del periodo de detección de los cultivos	78
5.7.1.	Establecer función heurística y de parada	78
5.7.2.	Definición del estado inicial	79
5.7.3.	Ejecución del algoritmo de detección del periodo	79
5.8.	Identificación de los cultivos	81
5.9.	Conclusiones de la evaluación	82
6.	Conclusiones y trabajos futuros	83
6.1.	Conclusiones	83
6.2.	Trabajos futuros	85
	Referencias	87

Índice de figuras

4.1. Esquema general del método propuesto.	28
4.2. Ejemplo de un conjunto de zonas geométricas para una región determinada (Extremadura).	35
4.3. Ejemplo de emparejamiento entre recintos y zonas geométricas.	36
4.4. Ejemplo de emparejamiento entre recintos y píxeles.	40
4.5. Representación de la extracción de la información por píxel de las distintas bandas y periodos para generar una ISMM.	48
4.6. Ejemplo de periodos adyacentes para calcular vecindades en el algoritmo de HC.	56
5.1. Conjunto de recintos declarados en el área de trabajo en el año 2022.	65
5.2. Ejemplo de coordenadas de un recinto con alguna malformación.	66
5.3. Malla de barrido de los satélites de la misión Sentinel-2 a su paso por la península Ibérica.	67
5.4. Relación de zonas geométricas en la comunidad autónoma de Extremadura, resaltando aquellas correspondientes al área de trabajo.	68
5.5. Ejemplo de funcionamiento del filtro de Hampel sobre una serie temporal.	74

5.6. Ejemplo de ISMM generada.	76
5.7. Traza de ejecución del algoritmo HC para obtener el periodo adecuado de detección del tabaco.	80

Capítulo 1

Introducción

1.1. Contexto

La aplicación de técnicas de aprendizaje automático en el campo de la identificación de cultivos, utilizando imágenes de satélite, se está generalizando debido a los avances en la teledetección y en el área de la Inteligencia Artificial. Esta universalización está generando que tanto organismos como investigadores realicen novedosos estudios y propuestas, centrados mayoritariamente en regiones o cultivos concretos, sin profundizar en la reutilización y estandarización de estos métodos de teledetección aplicados a la agricultura de precisión.

El Instituto Geográfico Nacional (IGN) define la teledetección como *la técnica de adquisición de datos de la superficie terrestre desde sensores instalados en plataformas espaciales* [1]. Así, en la teledetección distinguimos tres elementos esenciales [2]: la plataforma para sostener el instrumento de observación, como aviones, drones, o, especialmente, en lo que comúnmente conocemos como satélites artificiales; el objeto que se va a observar, que en el caso más común será la superficie terrestre; y el instrumento o sensor para observar dicho objeto. No obstante, se puede deducir que existe un cuarto elemento, derivado de los anteriores pero el que adquiere mayor importancia: la información obtenida de la observación.

Recién celebradas las bodas de platino del lanzamiento del Sputnik 1 en 1957 por parte de la Unión Soviética, el primer satélite artificial de la historia [3], pocos se aventurarían a pronosticar el futuro de los satélites en conjunción con una gran variedad de aplicaciones [4], como son: la geología, facilitando la exploración de zonas de difícil acceso; la meteorología, no solo para la predicción del tiempo, sino para otras actividades como la monitorización de la actividad volcánica, con un caso muy reciente en nuestro país en la isla de La Palma; las ciencias marinas, monitorizando distintos aspectos del medio como las olas o las corrientes marinas; la protección del medio ambiente, en situaciones con incendios forestales, sequías o inundaciones [5]; y, especialmente, un campo de estudio no tan adelantado en dicha fecha, como es la agricultura. Tal es la magnitud de las posibilidades que ofrece la teledetección que España cuenta con un Plan Nacional de Teledetección [6], con el objetivo de ofrecer cobertura a nivel nacional de un amplio conjunto de imágenes de satélite.

Centrando el objetivo en las aplicaciones de la teledetección en la agricultura, se puede establecer que ambos avanzan de la mano a pasos agigantados en los últimos años. Su relación es ya bastante longeva, especialmente desde el lanzamiento del programa norteamericano Landsat, en 1972, el primer satélite de observación de la Tierra que se lanzó con la intención expresa de estudiar y monitorizar la superficie terrestre [7]. Gracias a él, se encuentran los primeros estudios referidos a la aplicación de la teledetección a la monitorización y clasificación de cultivos [8]. No obstante, la identificación de cultivos no es la única aplicación que se puede conseguir con este tipo de técnicas, sino que también está muy extendida para otros usos como la predicción del rendimiento y la producción de cultivos; el seguimiento de catástrofes sobre cosechas [9], como inundaciones, diluvios, sequías, heladas o los distintos tipos de plagas; o el seguimiento del pastoreo, que si bien estaría más enfocado al ámbito de la ganadería, cumple un papel fundamental en el mantenimiento de zonas agrícolas como son los terrenos de pastos.

Es importante aclarar que, en teledetección, se disponen de distintos tipos de sensores en función de su interacción con la superficie de la tierra: sensores que registran la energía solar reflejada por la superficie, conocida como teledetección pasiva (recoge tanto la radiación visible como aque-

lla fuera del rango de visión humano, como la infrarroja); por otro lado, sensores que captan la energía reflejada por un haz de energía artificial, denominada como teledetección activa [4] (por ejemplo, los sensores con tecnología radar, a través de ondas, o lidar, a través de haces de luz); y por último, la radiación emitida por la superficie, como la radiación térmica o las microondas [10]. En el caso concreto de la teledetección en el campo de la agricultura, esta información recopilada por los sensores varía notablemente en función de la superficie, donde la vegetación ofrece unos resultados característicos, aunque también presenta suficientes variaciones para discernir entre sus distintos tipos.

Si se desea establecer un punto de inflexión en la monitorización de cultivos mediante imágenes de satélites en los últimos años, sería el despliegue de la misión Sentinel-2 en el año 2015 [11], en este caso por parte de la Agencia Espacial Europea (*ESA*), dentro del programa Copernicus [12]. La influencia de esta misión en el campo de la teledetección radica en dos componentes principales: por un lado, la periodicidad de las imágenes, donde actualmente se puede disponer de información de la misma región geográfica con una frecuencia de 5 días, lo que aporta un gran potencial para realizar análisis multitemporales de una misma región. En segundo lugar, destaca por la buena resolución de sus imágenes, que si bien no llegan a considerarse de alta resolución, sí que proporciona información con un detalle de hasta 100 metros cuadrados por píxel en algunas de sus bandas de información.

No obstante, y ya avanzando temporalmente hasta la actualidad, se dispone de una amplia variedad de opciones de satélites para obtener imágenes para la detección de cultivos, siendo algunas de las más frecuentes los satélites de la citada misión Sentinel-2 [13], Landsat [14] –mayoritariamente Landsat-8–, RapidEye [15], GeoEye-1 [16], MODIS [17], HJ-1 [18], PROBA-V [19], entre otros, aunque es posible encontrar múltiples combinaciones entre ellos [20] [21] [19]. También son frecuentes los casos en los que se combinan datos ópticos de las fuentes citadas anteriormente con datos de radar, como puede ser la misión Sentinel-1 [22] [23]. Existe también algún caso en el que las propias imágenes se combinan con datos meteorológicos [24]. Disponer de tal cantidad de fuentes de datos es otra ventaja añadida.

Otra de las claves del éxito actual de la teledetección en el campo de la agricultura ha sido el avance de las técnicas relacionadas con la inteligencia artificial, y más concretamente dentro del subconjunto del aprendizaje máquina (*machine learning*) [25]. Este avance radica, en primer lugar, en las mejoras relacionadas con la propia generación de modelos de aprendizaje máquina, a través del estudio de patrones, y que permite no solo que dichos modelos sean más acertados a la hora de realizar predicciones, sino que se puedan emplear grandes cantidades de datos como ejemplos para el entrenamiento, permitiendo consecuentemente obtener unos resultados mejores. Esto facilita, por tanto, la existencia de trabajos que combinan el aprendizaje máquina con imágenes de satélite, que generalmente incluyen una gran cantidad de información.

Por otra parte, es necesario resaltar el desarrollo y mejora de las técnicas de aprendizaje máquina, partiendo de los algoritmos básicos utilizados en este campo durante décadas, como el análisis de componentes principales (en inglés, *PCA*) o el algoritmo K-medias, pasando por otros métodos más sofisticados como las redes neuronales artificiales, y llegando en la actualidad a los métodos de aprendizaje profundo (*deep learning*) y las redes convolucionales, que ya han demostrado obtener resultados exitosos [26]. Asimismo, los distintos métodos de aprendizaje máquina han demostrado que pueden ser aplicados a los distintos casos de uso de la agricultura de precisión [27], como el análisis de las propiedades del suelo, la predicción del tiempo, la predicción de rendimiento de cultivos, la detección de enfermedades y malezas, o la gestión del riego por goteo, además de la propia detección e identificación de los distintos tipos de cultivos.

Debido al éxito de estos avances, desde el año 2018, la Comisión Europea (CE) viene promoviendo la implantación de la monitorización de cultivos mediante las técnicas citadas anteriormente, en lugar de los controles de campo empleados tradicionalmente [28]. Esto viene motivado por la gran apuesta que realiza la Unión Europea (UE) a través de la Política Agraria Común (PAC), operativa desde 1962, y que en el año 2021 suponía el 33,1 % del presupuesto de la UE, llegando hasta los 55.710 millones de euros [29]. La relevancia de la PAC se manifiesta ante la importancia de mantener la producción alimentaria para los estados miembros, favorecer el desarrollo de las comunidades rurales y apostar por una agricultura ambientalmente

sostenible [30]. El objetivo a medio-largo plazo es incluir los procesos de monitorización automáticos dentro del sistema de gestión de pagos proporcionados por la PAC, de manera que se puedan extender los controles a un mayor número de recintos. De hecho, la CE habla textualmente de utilizar las técnicas más recientes de aprendizaje máquina para tal fin, aplicadas concretamente a los recursos proporcionados por el programa Sentinel.

Con este enfoque, la CE tiene como objetivo prioritario la monitorización automática de la PAC y su sistema de gestión de ayudas para toda la UE, tanto a nivel estratégico como económico. Además, se insta a los organismos gestores de las ayudas públicas a la agricultura a implantar un sistema de monitorización de cultivos. Los grandes avances en el campo de la teledetección, atendiendo no solo a la cantidad de fuentes de datos disponibles sino a la calidad de la información ofrecida, así como la alta periodicidad de los mismos, y las nuevas y avanzadas técnicas de aprendizaje máquina, son ejes prioritarios para la UE, que ya han demostrado sus buenos resultados aplicados en el campo de la teledetección agrícola.

Los satélites de la misión Sentinel-2 son los que se utilizan en Europa para la monitorización de la PAC (entre otros usos), y cuenta con dos satélites desplegados en órbita: el Sentinel-2A, lanzado en junio de 2015, y el Sentinel-2B, en marzo de 2017. Ambos satélites proporcionan la misma información, con un desfase temporal debido a sus distintas órbitas, consistente en 13 bandas espectrales [31]: cuatro bandas con una resolución espacial de 10 metros, seis bandas con 20 metros y tres bandas con 60 metros. Lo relevante es el alto nivel de periodicidad de las imágenes ofrecidas: 10 días en el ecuador con un satélite y 5 días con 2 satélites; y, en latitudes medias, 2 o 3 días [32].

Gracias a las distintas bandas de información, también es posible calcular una serie de índices espectrales, que permiten analizar aspectos como la vegetación, la nieve, el suelo, los recursos hídricos, etc. [33]. Algunos de los índices más conocidos son: el índice de Vegetación de la Diferencia Normalizada (NDVI), que tiene a su vez variaciones como el Índice de Vegetación de la Diferencia Normalizada Verde (GNDVI); el índice de vegetación mejorado (EVI); el Índice Diferencial de Agua Normalizado (NDWI); el Índice Diferencial Normalizado de Nieve (NDSI); entre otros [34]. Cabe destacar que estos índices son universales, y no dependen de una única fuente de

datos, sino que se pueden calcular, por ejemplo, partiendo tanto de las imágenes de Sentinel-2 como de Landsat-8.

Otro aspecto a favor de la misión Sentinel es la Plataforma de Aplicaciones de Sentinel (*Sentinel Application Platform*, SNAP), ofrecida de forma gratuita por la ESA, que incluye un amplio conjunto de herramientas para trabajar con las imágenes ofrecidas por la misión, aunque también permiten trabajar con recursos externos [35]. Entre sus herramientas, se encuentran la visualización de imágenes o la posibilidad de combinar las distintas bandas para generar los índices espectrales citados anteriormente.

No obstante, es necesario aclarar que no todo son bondades en el uso de este tipo de datos. El principal inconveniente que encontramos, sobre todo en áreas con una climatología adversa, es la presencia de nubosidad en las imágenes, lo que distorsiona la realidad del terreno, tanto por la propia presencia de nubes como por su sombra, ya que ambos fenómenos meteorológicos afectan a los datos recabados por los distintos sensores. Otro de los inconvenientes presentes en este campo es la gran cantidad de almacenamiento y capacidad de cómputo necesario para procesar las imágenes de satélite, especialmente cuando el área de estudio presenta una extensión amplia. Por ello, hay que tener muy en cuenta la eficiencia de todos los procesos relacionados con las distintas operaciones a realizar.

En base a lo anterior, son muchas las propuestas que estudian técnicas específicas para la identificación y monitorización de cultivos, con distintas alternativas para abordar el problema. Es frecuente encontrar soluciones que trabajan con una única imagen, en lugar de aprovechar las ventajas de realizar análisis multitemporales. Además, algunas de las propuestas que contemplan la utilización de varias imágenes presentan una serie de dificultades para replicarlas a gran escala, como la ausencia de una metodología definida, o no tener en cuenta las distintas restricciones y requisitos que conlleva un procesamiento masivo para aplicar el sistema a gran escala. También existen otras alternativas que utilizan herramientas software específicas o de pago, enfocadas al caso de uso concreto, que dificultan su utilización de manera universal. Asimismo, es frecuente encontrar propuestas que se centran en cultivos o regiones concretas [36], y no ofrecen una solución global. En resumen, este campo adolece de una propuesta genérica para poder extrapolar el sistema a distintos territorios a gran escala,

y, sobre todo, que permita realizar todos los procesos de manera masiva y automática.

Por las razones mencionadas, sería útil disponer de un método integral para conseguir, de manera guiada, un sistema de identificación de cultivos, haciendo uso de las imágenes multispectrales obtenidas de los satélites, durante periodos de tiempo significativos para cada cultivo, desde el primer proceso de recopilación de las propias imágenes hasta la obtención final de los resultados del sistema de aprendizaje.

1.2. Estructura de la Tesis

A continuación, se enumeran las distintas secciones de esta tesis doctoral, resumiendo sus contenidos principales.

- El **Capítulo 2** proporciona una visión general de los trabajos relacionados con la monitorización y clasificación de cultivos mediante imágenes de satélite, centrándose en todas las fases del proceso: desde la adquisición de las imágenes hasta el desarrollo de los sistemas de aprendizaje.
- El **Capítulo 3** define los objetivos, contribuciones, premisas, hipótesis y restricciones del trabajo.
- En el **Capítulo 4** se establece la propuesta de un método genérico para elaborar, paso a paso, un sistema completo de detección de cultivos, que realice esta tarea de manera masiva y automática.
- El **Capítulo 5** presenta la evaluación del trabajo, con la aplicación de un desarrollo completo del método propuesto.
- En el **Capítulo 6** se extraen las principales conclusiones y se presentan algunas ideas para trabajos futuros.

1.3. Publicaciones relacionadas

En esta sección se relaciona producción científica relacionada con esta tesis doctoral. Concretamente, se compone de tres publicaciones con índice de impacto JCR y una publicación en congreso:

- A. Lozano-Tello, **M. Fernández-Sellers**, E. Quirós, L. Fragoso-Campón, A. García-Martín, J. A. Gutiérrez Gallego, C. Mateos, R. Trenado, and P. Muñoz, “Crop identification by massive processing of multiannual satellite imagery for EU Common Agriculture Policy Subsidy Control”, *European Journal of Remote Sensing*, vol. 54, no. 1, pp. 1–12, 2020. [JCR(2020)=3.647], Q2 (16/32).
- **M. Fernández-Sellers**, G. Siesto, A. Lozano-Tello, and P. J. Clemente, “Finding a suitable sensing time period for crop identification using heuristic techniques with multi-temporal satellite images”, *International Journal of Remote Sensing*, vol. 43, no. 15-16, pp. 6038–6055, 2021. [JCR(2021)=3,531], Q3 (18/34).
- G. Siesto, **M. Fernández-Sellers**, and A. Lozano-Tello, “Crop classification of satellite imagery using synthetic multitemporal and multispectral images in Convolutional Neural Networks”, *Remote Sensing*, vol. 13, no. 17, pp. 3378, 2021. [JCR(2021)=5,349], Q2 (11/34).
- E. Quirós, A. Lozano-Tello, **M. Fernández-Sellers**, L. Fragoso-Campón, “Aplicación de redes neuronales para la monitorización de cultivos de la PAC en Extremadura”, XIX Congreso de la Asociación Española de Teledetección. Pamplona, 29 junio - 1 julio 2022.

Capítulo 2

Estado del arte

En este capítulo se recogen los trabajos existentes relacionados con los distintos procesos presentes en la identificación de cultivos, tales como la selección de imágenes de satélite y su pretratamiento, la determinación de los periodos de detección, la agrupación de datos para el aprendizaje máquina o el diseño del propio sistema de aprendizaje, entre otros.

2.1. Síntesis de la evidencia disponible

El primer paso para realizar la identificación de cultivos mediante imágenes de satélite es obtener las imágenes fuente, partiendo de los distintos satélites con la información del terreno disponible. En algunos casos, las imágenes se ofrecen para ser descargadas por parte de usuarios externos desde distintos repositorios, entre las que se pueden citar como las más usadas: el Centro de Datos Terrestres Theia (de origen francés), como proponen en [37]; el Centro de Observación y Ciencia de los Recursos Terrestres (EROS, por sus siglas en inglés, de origen estadounidense), empleado por [38]; o la plataforma de acceso gratuito a los recursos del programa Copernicus, a nivel europeo, y utilizado en trabajos como [39]. Es posible encontrar otros trabajos como [22] y [40] en los que se utiliza otra de las fuentes de datos más importantes, como es Google Earth Engine.

Una vez descargadas estas imágenes, comúnmente se les aplican una serie de operaciones para evitar anomalías y poder trabajar con ellas correctamente. En primer lugar, se deben recopilar las imágenes con las que se desea trabajar, atendiendo comúnmente al criterio de seleccionar las imágenes con un adecuado porcentaje de nubosidad. Aunque es un factor importante, en algunos trabajos no se especifica este umbral [41] [39]. Otros, como en [37], hacen referencia a imágenes sin nubes o casi sin nubes, de manera literal sin mencionar el criterio. Otros autores como [32] sitúan un umbral de nubosidad fijo, en este caso un 20 %, a partir del cual se descartaría la imagen completa. En estos procesos de filtrado, en algunos casos se opta por seleccionar un número equitativo de imágenes por periodo para no descompensar los meses con mayor y menor presencia de nubosidad, como por ejemplo, seleccionar una imagen nítida por mes, como proponen en [23].

Una vez que las imágenes se han seleccionado, es frecuente que se necesite aplicar un conjunto de procedimientos de normalización para que puedan ser posteriormente analizadas. El caso más extendido se encuentra al utilizar fuentes de datos cuyos espectros tengan resoluciones diferentes (como es el caso de Sentinel-2 o Landsat-8, por ejemplo), donde es necesario realizar una normalización de las distintas resoluciones de las bandas de información para procesarlas de manera homogénea (por ejemplo, si una banda X dispone de píxeles con resolución de 20m, y queremos equipararla a una banda Y con resolución de 10m, por cada píxel de la banda X acabaremos teniendo dos píxeles idénticos con una resolución de 10m, de manera que todos los píxeles resultantes tengan la misma resolución). Esto es aplicado por diversos autores como [38], [39] o [21]. Relacionado con lo anterior, también se encuentran casos en los que se emplean algoritmos concretos para esta tarea. Es el caso de [42], que utiliza el algoritmo de k vecinos más cercanos (en inglés, *k-nearest neighbors*, abreviado k -nn), que utiliza la proximidad para hacer clasificaciones sobre la agrupación de un punto de datos individual. La normalización de las resoluciones también es importante en el caso de combinar fuentes de datos distintas, como Landsat o Sentinel, con diversas opciones, como explican Li et al. [43]. Por otra parte, estudios como el de [44] llegan incluso a comparar el uso y la efectividad de las distintas resoluciones para la clasificación de cultivos.

A pesar de que en ocasiones se criben las imágenes por su porcentaje de nubosidad, es común disponer de imágenes en las que nos encontramos datos inexistentes o inválidos, debidos al mencionado problema de la nubosidad, a errores puntuales en los sensores, u otras causas. En este caso debemos dividir el problema en dos fases: en primer lugar, detectar esos datos inválidos, como puede ser la presencia de nubes o las propias sombras de las nubes; y, una vez detectados y eliminados, completar la falta de información para cada una de las bandas. Para conseguir localizar estos datos que se deben descartar se puede utilizar la propia capa de datos adicional que ofrecen algunas fuentes de datos, como Sentinel, en la que se ofrece clasificación de la escena [45] –una capa de imagen en la que nos indica la clasificación de cada píxel, para comprobar si hay nubosidad, nieve, agua, etc.–. No obstante, esta máscara de nubes no estuvo disponible en la plataforma de descarga (Sentinel Hub) hasta mayo de 2020 [46], por lo que hasta esa fecha también se exploraron otras opciones. El trabajo de [47] propone utilizar Google Earth Engine para este fin. También se pueden encontrar trabajos específicos sobre este tema, ya que tanto [48] como [49] generan sus propios algoritmos para detectar nubes y las sombras que provocan, basándose en combinaciones de bandas o índices. El trabajo de [50] propone utilizar elementos más complejos como redes neuronales. Otros como Hollstein et al. [51] han generado métodos más completos para clasificar los píxeles incluyendo, además de nubes y sombras, otras anomalías como nieve o agua, haciendo uso de árboles de decisión. También es posible aplicar filtros estadísticos para localizar datos anómalos o extremos (debidos a distintas causas, como interferencias con los sensores de los satélites), como pueden ser el filtro de Hampel, descrito en [52].

Tras detectar y suprimir los datos con nubes o con alguna anomalía, es necesario completar los huecos de esos datos para que las series temporales sean homogéneas en cuanto al número de datos. Una de las soluciones que se plantea es combinar datos ópticos con datos de radar, que no se ven afectados por estos problemas de nubosidad. Tanto Arslan et al. [53] como Tuvdendorj et al. [54] proponen combinar, en este caso, datos de Sentinel-1 (radar) con Sentinel-2 (óptico), aunque también es posible que estos datos sean de diferentes fuentes, como proponen Kordi y Yousefi [55], integrando datos de Sentinel-1 con datos de Landsat-8. No obstante, incluir diferentes fuentes de datos puede afectar a la capacidad de almacenamiento

y cómputo, así como al tiempo de procesamiento.

Centrándonos, por tanto, en soluciones para el citado problema que no impliquen utilizar fuentes de datos adicionales, lo más sencillo sería aplicar una interpolación lineal clásica, más rápida y menos costosa a nivel computacional en comparación con otras soluciones, aunque encontramos en la literatura diversas opciones alternativas. En [56] se propone utilizar las similitudes a nivel espacial, espectral y temporal para calcular los valores desconocidos. Li et al. [38] proponen como alternativa el uso de mapas autoorganizados de Kohonen. En el trabajo de [57] podemos encontrar como propuesta de solución la creación de un modelo de mezcla gaussiana a nivel de recinto, basado en modelos probabilísticos. Otra opción es la combinación de imágenes sin presencia de nubosidad en fechas cercanas a las fechas de los datos descartados, como indican en [58].

Una vez que se dispone de las imágenes almacenadas y tratadas, es posible que la cantidad de información pendiente de ser procesada sea demasiado elevada por tener que analizar muchos recintos, especialmente si atendemos al tiempo de procesamiento que ello implica. Para reducir el tamaño de las imágenes, trabajos como [39] y [17] aplican máscaras sobre las mismas para almacenar únicamente las regiones de interés, es decir, aquellas que abarquen los recintos que se desean estudiar. El trabajo de Gardiraju et al. [59] crea agrupaciones de píxeles que pertenezcan a la misma clase o cultivo, de manera que se reduce también la cantidad de información almacenada. No obstante, también existen casos en los que se analizan y clasifican las imágenes completas, como en [60], ya que la aplicación de alguna de las técnicas anteriormente mencionadas también influye en el tiempo total de preprocesamiento.

Finalmente, a las imágenes se les debe añadir una capa adicional conocida como *verdad terreno*, que implica conocer el tipo de cultivo que se presupone en cada sección de la imagen (píxel), con el objetivo de disponer de ejemplos fiables para el posterior entrenamiento del sistema de aprendizaje. Para la obtención de la información relativa a los diferentes cultivos presentes en las distintas áreas, Wang et al. [61] utilizan la capa de datos de tierras de cultivo, una de las fuentes de datos más comunes, proporcionada por el departamento de agricultura de Estados Unidos [62]. Otros trabajos ponen más énfasis en este apartado: Rauf et al. [39] proponen aplicar es-

tudios de campo (visitas, encuestas a los dueños de los recintos, etc.), algo que puede resultar demasiado laborioso e impreciso; en [32] se llevan a cabo etiquetados manuales de las distintas áreas; aunque también se puede realizar una clasificación supervisada de imágenes de alta resolución combinada con una interpretación visual por parte de expertos, como se describe en el trabajo de Hu et al. [17]. Como alternativa, [36] propone la obtención de muestras históricas de los píxeles con la frecuencia del cultivo por año en dicho píxel –por ejemplo, en un muestreo de 8 años, si durante 5 años se catalogó como un determinado un cultivo, se establece que es un posible píxel de ese cultivo–. En ocasiones, el etiquetado también puede ajustarse a otras variantes, ya que se puede elaborar de manera jerárquica con varios niveles, tal y como se explica en [63] –por ejemplo, podemos tener como primer nivel la etiqueta “cultivos extensivos”, un segundo nivel con “cereales de grano grande”, donde encontraríamos un tercer nivel que incluiría cultivos como el maíz o el sorgo–. [13] agrupa los cultivos en grupos por similitud, con el objetivo de evitar cultivos con pocos ejemplos y crear clases mejor representadas.

Todo lo tratado anteriormente podría ser válido tanto para trabajos que realicen su investigación empleando una única imagen, como para aquellos que opten por analizar un conjunto de ellas a lo largo del tiempo. Sin embargo, también se aprecia como necesario establecer qué imagen o imágenes se van a utilizar para el estudio, ya que es fundamental seleccionar las imágenes apropiadas según el periodo fenológico del cultivo que se va a analizar. Centrando el estudio en aquellos trabajos que realizan análisis multitemporales (varias imágenes), es importante determinar cuáles van a ser las imágenes que se necesitan descargar, es decir, cuál será el periodo de detección. En el trabajo de [64] realizan diversas pruebas intentando determinar el número de imágenes necesario para obtener un resultado aceptable, sin llegar a proponer un método concreto para llevarlo a cabo. Masse et al. [65] presentan un algoritmo genético para seleccionar un conjunto de imágenes, maximizando distintos índices de acierto y minimizando el número de imágenes. El trabajo de Van Niel y McVicar [66] establece distintas metodologías para determinar las imágenes que se van a utilizar, destacando la que denomina como clasificación multi-fecha, en la que se combinan un conjunto reducido de imágenes. Otros estudios como [67] evalúan cuál es la ganancia obtenida aumentando tanto el número de

imágenes (eje temporal) como el número de bandas de información, entendiendo la ganancia como la mejora en los resultados de clasificación de los cultivos.

El enfoque más común a la hora de determinar qué imágenes utilizar es tener en cuenta el periodo fenológico promedio de los cultivos, algo que se realiza en diversos trabajos como en [68], [69] o [70]. Otros autores como [71] y [72] abordan el problema analizando la temporada de crecimiento de los cultivos. Ambos enfoques se basan en experiencias previas u opiniones de expertos en cuanto a los periodos de cada cultivo, sin aplicar ninguna tecnología específica como solución. Por otro lado, el estudio de [32] demuestra la importancia de seleccionar unas fechas frente a otras a la hora de identificar los cultivos, pero sin proponer una metodología para establecer el periodo completo de detección de manera automática.

Dentro de los propios recintos de estudio también es posible aplicar un nuevo filtrado, de manera similar al trabajo realizado con las imágenes, en este caso prescindiendo de algunos píxeles poco significativos en base a distintos criterios. El trabajo de [73] propone el descarte de algunos píxeles, bien por encontrarse en el borde del recinto o por pertenecer a parcelas demasiado estrechas. Otra alternativa es la encontrada en [74], donde se descartan directamente todos aquellos píxeles pertenecientes a recintos que no superen una superficie mínima.

Para generar los modelos de detección basados en el aprendizaje máquina, una vez que se dispone de la información preparada para el entrenamiento, es necesario decidir qué datos y con qué valores se va realizar el proceso. Existen investigaciones a nivel de píxel, aunque es destacable señalar que también existen en la literatura otros que obtienen buenos resultados generalizando el área de aplicación. En [75] se aplica el indicador NDVI a nivel de parcela, mientras que otros trabajos como [16] combinan índices y bandas también a nivel de parcela, generando como entrada del sistema de aprendizaje el conjunto de series temporales de cuatro bandas espectrales y tres índices de vegetación.

Existen trabajos que analizan cómo agrupar las bandas de los satélites, o si se van a generar nuevos datos sintéticos a partir de los datos fuente. Es posible encontrar una considerable cantidad de variantes (si se trabaja

a nivel de píxel): [41] propone generar la media de distintos indicadores (EVI, NDVI, NDWI, etc.) para cada píxel, siendo este conjunto de medias la entrada para el aprendizaje. No obstante, esta aproximación no permite comparar la evolución a lo largo del tiempo de los cultivos. Otros trabajos como [13], [39] o [68] combinan las propias bandas espectrales con los citados índices. En [37] y [32] simplemente se generan las series temporales de los distintos índices a nivel de píxel, destacando como caso más extendido aquellos que únicamente emplean el NDVI, reflejado en [76] o en [77]. Como alternativa, tanto [78] como [14] filtran las propias bandas utilizadas, sin emplear todas las disponibles.

Respecto al ajuste de la distribución de los datos para mejorar el entrenamiento de los modelos, [41] propone equiparar el número de píxeles por cultivo para que los algoritmos de aprendizaje máquina no se sobreentrenen. Otra opción puede ser utilizar conjuntos de datos auxiliares que ayuden a caracterizar otros parámetros como el tipo de vegetación o las propiedades del suelo, como es el caso de [79].

Independientemente de los datos utilizados para el aprendizaje y su agrupación, nos encontramos con una variedad muy amplia de métodos para generar los modelos de detección de los cultivos. Generalmente, la literatura clasifica los algoritmos de aprendizaje máquina en dos grandes grupos, tal y como hacen Saxena et al. [80]: no supervisados y supervisados –aunque también existen los denominados como semisupervisados, citados en trabajos como el de Jing y Chao [81]–. El conjunto de algoritmos no supervisados es el que menos tipos presenta, destacando métodos como ISODATA (*Iterative Self-Organizing Data Analysis Techniques*, método iterativo que utiliza la distancia euclídea como medida de similitud para agrupar datos en diferentes clases), o el método de agrupamiento K-medias (*K-means*, algoritmo que agrupa objetos en k grupos basándose en sus características). También podemos encontrar otras posibles soluciones dentro de los métodos no supervisados: en [82] se propone una versión distribuida de ISODATA; [83] aplica el agrupamiento jerárquico, donde se van agrupando los datos por pares en función de su similitud, generando un árbol de agrupamientos desde la base hasta la cima. Otra alternativa es la propuesta por [72], que aplica el agrupamiento isométrico de los componentes principales (PCIB de las siglas en inglés), basado en el PCA, que

extrae información de los datos para clasificarlos en subgrupos.

Según el trabajo anterior de [80] y el de otros autores como Wang et al. [84], uno de los principales métodos de clasificación supervisados serían los árboles de decisión. Por ejemplo, en el trabajo de [16], se aplican a mapas con los posibles resultados de una serie de decisiones, de manera que se puede anticipar la mejor opción de manera matemática; en [36] se decantan por los algoritmos k-*nn*, que utilizan la proximidad para realizar clasificaciones o predicciones; también son muy utilizados los bosques aleatorios (*random forests*), que crean múltiples árboles de decisión y los combinan para obtener una predicción más precisa y estable. Además, tal y como se describe en [85], los bosques aleatorios se pueden aplicar utilizando series temporales. Otro de los métodos de clasificación supervisados más extendido son las máquinas de vector soporte (del inglés *support-vector machines*, SVM), que se aplican satisfactoriamente en trabajos como el de [18] o el de [86].

En el trabajo de Siachalou et al. [20] se apuesta por otra solución menos convencional, ya que aplican la técnica de clasificación de modelos ocultos de Markov (*Hidden Markov Models*, HMM), un modelo estadístico para el reconocimiento de patrones y su clasificación. El trabajo de [87] elige la técnica estadística de potenciación del gradiente (*Gradient boosting*), que forma un modelo predictivo conjunto en base a modelos de predicción más simples, como los árboles de decisión. Otras soluciones se pueden encontrar en [88], realizando predicciones mediante el algoritmo de máxima verosimilitud, o en [89], que aplica la Regresión Logística Multinomial (*Multinomial Logistic Regression*, MLR).

Además de estas técnicas, en los últimos años predomina una variante denominada aprendizaje profundo, en el que destacan diversos tipos de redes neuronales. Uno de los tipos más destacados y efectivos son las redes neuronales convolucionales (CNN), que abarcan distintas variantes. Uno de los trabajos más completos del ámbito lo encontramos en Wang et al. [90], que realiza una revisión de los distintos métodos de *deep learning* para teledetección en agricultura y sus distintas aplicaciones. En el apartado de agricultura de precisión aplicada a la clasificación del terreno, destacan las tres variantes principales de CNN: 1D-CNN, 2D-CNN y 3D-CNN. Todas ellas presentan buenos resultados, e incluso la literatura muestra distintas

comparativas entre estas variantes, como el trabajo de Yang et al. [91], que compara una arquitectura concreta de 1D-CNN con otra basada en 3D-CNN. Por contra, estudios como [92] defienden el uso de las 2D-CNN. Existen otros tipos de CNN, como el de Haut et al. [93], que proponen un novedoso sistema de CNNs denominado TResNets, enfocado al consumo eficiente de recursos; o el trabajo de Zhu et al. [94], donde presentan alternativas como las redes VGG, también muy extendidas.

Además de las CNN en la identificación de cultivos, también encontramos otros tipos de redes neuronales aplicadas a la detección e identificación de cultivos. Jia et al. [95] estudian las redes neuronales recurrentes (RNN), aunque destacan un subtipo concreto dentro de estas redes, las redes LSTM (*long short-term memory*), que se caracterizan por tener “memoria” a corto y largo plazo durante el procesamiento dentro de la red. En [96] se muestra otra alternativa, las redes neuronales de retroprogragación (BPNN, por sus siglas en inglés). Por otra parte, [97] compara otro tipo especial de red neuronal, como son los codificadores automáticos (*autoencoder*) con las CNN.

Por último, también es interesante comentar que se encuentran trabajos que combinan distintos métodos de aprendizaje en monitorización para obtener los resultados. Por ejemplo, el trabajo de [42] propone combinar bosques aleatorios y redes neuronales. Algo parecido se propone en [98], en este caso combinando los bosques aleatorios con redes neuronales convolucionales. Como solución más compleja que las anteriores tenemos la propuesta por Mazzia et al. [99], que combina redes neuronales recurrentes con redes neuronales convolucionales. Una combinación similar se puede encontrar en [100], sustituyendo las redes neuronales recurrentes por las redes LSTM, en una combinación que denomina ConvLSTM.

2.2. Conclusiones

Tras estudiar y analizar el estado del arte sobre los distintos subprocesos para la identificación de cultivos mediante imágenes de satélite, se pueden extraer una serie de conclusiones que motivan la realización de esta tesis.

En general, no se especifican ni se detallan los procesos de obtención de las imágenes desde los servidores fuente. Es común que los estudios únicamente se limiten a citar la fuente o la web desde donde han descargado las propias imágenes. Además, tampoco aclaran si las imágenes se obtienen mediante algún proceso automático o se descargan manualmente, lo que conllevaría una ingente cantidad de tiempo cuando se requieren cientos de imágenes si se va a realizar un procesamiento masivo.

Con las imágenes descargadas, se ha observado que algunos autores deciden descartar las imágenes en función de algún porcentaje de nubosidad presente en la imagen. No obstante, esto no siempre puede ser adecuado, ya que si se desea trabajar con cultivos minoritarios (pocos recintos asociados, o recintos pequeños con pocos píxeles), cualquier píxel que se pueda rescatar de estas imágenes es probable que sea muy valioso de cara al entrenamiento o a la identificación. Esto también es aplicable a aquellas propuestas que descartan los recintos menores a un cierto tamaño. Sin embargo, al disponer de una mayor cantidad de información almacenada y procesada, parece que trabajar con las imágenes completas puede suponer un gran esfuerzo en términos de almacenamiento, capacidad de cómputo y tiempo de procesamiento, por lo que puede resultar interesante reducir las imágenes para trabajar sólo con aquellos píxeles que sean de interés, es decir, los que pertenezcan a los recintos de estudio.

Sean o no descartadas las imágenes, parece lógico que es necesario aplicar algún algoritmo o técnica para completar las series temporales de las imágenes en caso de presencia de nubosidad o de datos inválidos o anómalos, y solo algunos trabajos abordan este problema.

Con respecto a establecer el periodo de detección de los cultivos, no se ha encontrado mucha literatura al respecto, puesto que la mayoría de trabajos utilizan los periodos convencionales para cada cultivo, y suelen ser útiles solo en zonas geográficas específicas. Además, las propuestas en-

contradas no proponen una metodología automática para establecer dicho periodo, por lo que se aprecia una necesidad en este aspecto.

Existen, así mismo, muchas variantes para combinar los datos obtenidos de cara a generar los distintos modelos de detección, pero no se establece una comparación clara entre las distintas alternativas. Algo similar ocurre con los algoritmos de aprendizaje máquina, donde la cantidad de alternativas para este tipo de estudios resulta elevada, destacando los métodos de aprendizaje profundo.

Como se ha visto en esta sección, existen muchas técnicas y alternativas para abordar el proceso de identificación de cultivos, e incluso distintas variantes dentro de cada uno de los subprocesos estudiados, pero no se ha encontrado una metodología que proponga un conjunto completo de procesos para la creación de modelos de cultivos y la identificación masiva. Para resolver esta carencia, en el presente trabajo se propone un método que permite a los usuarios crear modelos adecuados para la identificación de cultivos desde imágenes de satélites de manera guiada y justificada.

Capítulo 3

Objetivos y contribuciones

Este capítulo presenta los objetivos de este trabajo y el problema que se pretende abordar (Sección 3.1), junto con sus principales contribuciones (Sección 3.2), las premisas, hipótesis y restricciones (Secciones 3.3, 3.4 y 3.5, respectivamente).

3.1. Objetivos

Atendiendo a la gran cantidad de situaciones que pueden beneficiarse de un sistema como el aquí propuesto, en este trabajo se intenta responder a la siguiente pregunta: *¿es posible monitorizar de manera masiva y automática un conjunto de recintos de cultivos desde imágenes de satélite?*. Para abordar el problema planteado, se identifican los siguientes **objetivos**:

- O1. Modelar la recopilación masiva y tratamiento de datos, obtenidos a partir de imágenes multiespectrales de satélite, para que sean adecuados para el análisis de cultivos.
- O2. Diseñar una metodología para determinar de manera automática el periodo idóneo de detección de cultivos.

- O3. Establecer un sistema completo basado en aprendizaje máquina para la detección de cultivos que seleccione los datos de entrada y configure el sistema de manera óptima.

Para conseguir el primer objetivo, consideramos las siguientes cuestiones:

- Algunas soluciones de tratamiento de imágenes de satélite seleccionan y descargan manualmente las imágenes recogidas por los satélites, y son analizadas y tratadas de manera individual. Esto reduce el número de imágenes descargadas, y consecuentemente limita las posibilidades que ofrecen las propias imágenes. Para solucionar este problema, se pretende establecer un procedimiento automatizado para seleccionar las imágenes a descargar, partiendo de la información cartográfica de los distintos recintos de cultivos que se quieren analizar. Esta operación debe incluir la selección de aquella información contenida dentro de los propios recintos, de manera que los datos posteriormente analizados se correspondan con aquellos que sean de interés. Además, se deberán tratar los datos descargados, incluyendo filtrados de valores anómalos por causas ajenas al sistema, como pueden ser fallos en los propios sensores del satélite o la presencia de nubosidad en la zona que altere los valores reales.

Para cumplir con el segundo objetivo, se debe tener en cuenta:

- En función del caso de uso del sistema y sus correspondientes requisitos, será necesario establecer una temporalidad de observación de los cultivos. Generalmente se establece el periodo de estudio de un cultivo de manera manual, basándose en experiencias previas u otros trabajos relacionados. No obstante, no se desea dejar toda la responsabilidad en manos de los expertos, sino que el objetivo plantea la posibilidad de establecer un sistema eficiente que, de manera automática, identifique el periodo sobre el que trabajar para cada cultivo y caso de uso.

Para alcanzar el tercer objetivo, debemos resolver la siguiente cuestión:

- El uso eficiente de sistemas de identificación de cultivos mediante el aprendizaje máquina con imágenes de satélite va a depender de los objetivos concretos del proyecto que se va a desarrollar. La preparación y combinación de los datos, y la generación de los ejemplos de entrada para el entrenamiento, podrán tener diferentes configuraciones, y deberían establecerse pautas justificadas para realizar estos procesos. Para cumplir este objetivo, se propone definir un método integral de detección de cultivos utilizando imágenes de satélite, que contenga una serie de pasos bien definidos que sirvan de guía en el proceso de desarrollo del sistema.

3.2. Contribuciones al estado del arte

El trabajo realizado pretende dar solución a los objetivos y problemas descritos en la Sección 3.1. Las **contribuciones** para conseguir los mencionados objetivos son:

C1. Método automático de recopilación de datos.

Con respecto al estado del arte actual, se proporcionará un nuevo método que ayude a los desarrolladores a recopilar y procesar la información ofrecida por los satélites de manera masiva y automática, atendiendo a los recintos o regiones que se quieren analizar y el periodo de tiempo requerido. Esta contribución está estrechamente relacionada con el O1.

C2. Sistema para determinar el periodo de detección adecuado.

Partiendo del O2, se proporcionará un sistema automático para determinar, dado un cultivo concreto, su periodo temporal de detección adecuado, combinando los diferentes parámetros involucrados, como son: el umbral de acierto requerido, la capacidad máxima de almacenamiento o el tiempo necesario para realizar los distintos procesos de descarga y procesamiento de la información.

C3. Método completo para el desarrollo de sistemas de análisis de cultivos.

En relación al O3, se establecerá un método integral conteniendo una secuencia de pasos bien definidos que permita el análisis de cultivos según las condiciones del proyecto. Estará basado en el aprendizaje automático, utilizando imágenes sintéticas a nivel de píxel que engloben la información disponible de los distintos recintos de cultivos disponibles.

3.3. Premisas

Se asumen una serie de **premisas** para la consecución de los objetivos planteados para este trabajo:

- P1. Se trata de identificar un cultivo determinado a través de su evolución fenológica en el tiempo, no de detección de objetos que identifiquen al cultivo.
- P2. Se dispone de la información cartográfica de los recintos de cultivos tanto para el entrenamiento como para su monitorización.
- P3. Es posible acceder a los datos de las imágenes de satélite para su descarga.
- P4. La cantidad de datos disponibles será suficiente para poder crear modelos de aprendizaje adecuados.
- P5. El acceso a los datos de las imágenes de satélite permite automatizar su descarga.
- P6. Se dispone de capacidad de almacenamiento para recopilar las imágenes que requiera el sistema.
- P7. La capacidad de cómputo disponible será suficiente para procesar los datos y elaborar los procesos de aprendizaje.

3.4. Hipótesis

Una vez identificadas las premisas, se describen una serie de **hipótesis** sobre el trabajo:

- H1. Es posible automatizar la descarga de imágenes de satélite dados un periodo de tiempo determinado y un área de trabajo concreto.
- H2. Es posible procesar automáticamente las imágenes descargadas, seleccionar los datos de interés y generar una entrada determinada para desarrollar un proceso de aprendizaje máquina.
- H3. Es posible crear un método integral que genere modelos de aprendizaje máquina de manera automática para monitorizar los resultados y obtener conclusiones sobre los datos obtenidos.

3.5. Restricciones

Finalmente, las siguientes **restricciones** definen los límites de la contribución de este trabajo, delimitan el problema investigado y permiten la mejora incremental de la investigación a través de los trabajos futuros propuestos:

- R1. La presencia de nubosidad puede afectar significativamente a los resultados obtenidos, teniendo en cuenta que existen regiones donde la presencia de nubes es persistente durante varias semanas o meses consecutivos.
- R2. El número de datos debe ser suficientemente relevante y diverso para establecer unos modelos de aprendizaje máquina genéricos, cuyos resultados puedan extrapolarse a un amplio número de regiones.
- R3. El periodo de tiempo de estudio de las imágenes (entendido como el tiempo entre la imagen más antigua y la más reciente) debe ser lo suficientemente amplio para generar un modelo de aprendizaje máquina eficaz.

Capítulo 4

Método de identificación de cultivos

El método propuesto en esta tesis doctoral, esquematizado en la Figura 4.1, establece los pasos necesarios para disponer de un sistema de monitorización e identificación de cultivos mediante imágenes de satélite, que permita el análisis masivo y automático de un conjunto de recintos agrícolas. El resultado del método será conseguir modelos de aprendizaje máquina para la identificación de cultivos, adecuados a las necesidades específicas del proyecto que se vaya a desarrollar, por lo que se generará un modelo por cada cultivo que se desee identificar.

El método propone identificar los distintos requisitos y limitaciones del sistema (M1); cómo recopilar los datos fuente –principalmente, la información de los recintos que se desea monitorizar y los que serán utilizados para el aprendizaje– (M2); la forma de implementar el sistema de descarga masiva de datos de imágenes de satélite (M3); los posteriores procesamientos y tratamientos de los datos descargados (M4); la manera de generar los datos de entrada para el sistema de aprendizaje (M5); cómo diseñar e implementar el sistema de aprendizaje (M6); cómo implementar un método para establecer el periodo óptimo de análisis empleando modelos heurísticos (M7); y finalmente cómo realizar las predicciones sobre los recintos con los modelos generados (M8).

M1. Análisis de requisitos y diseño del sistema de información

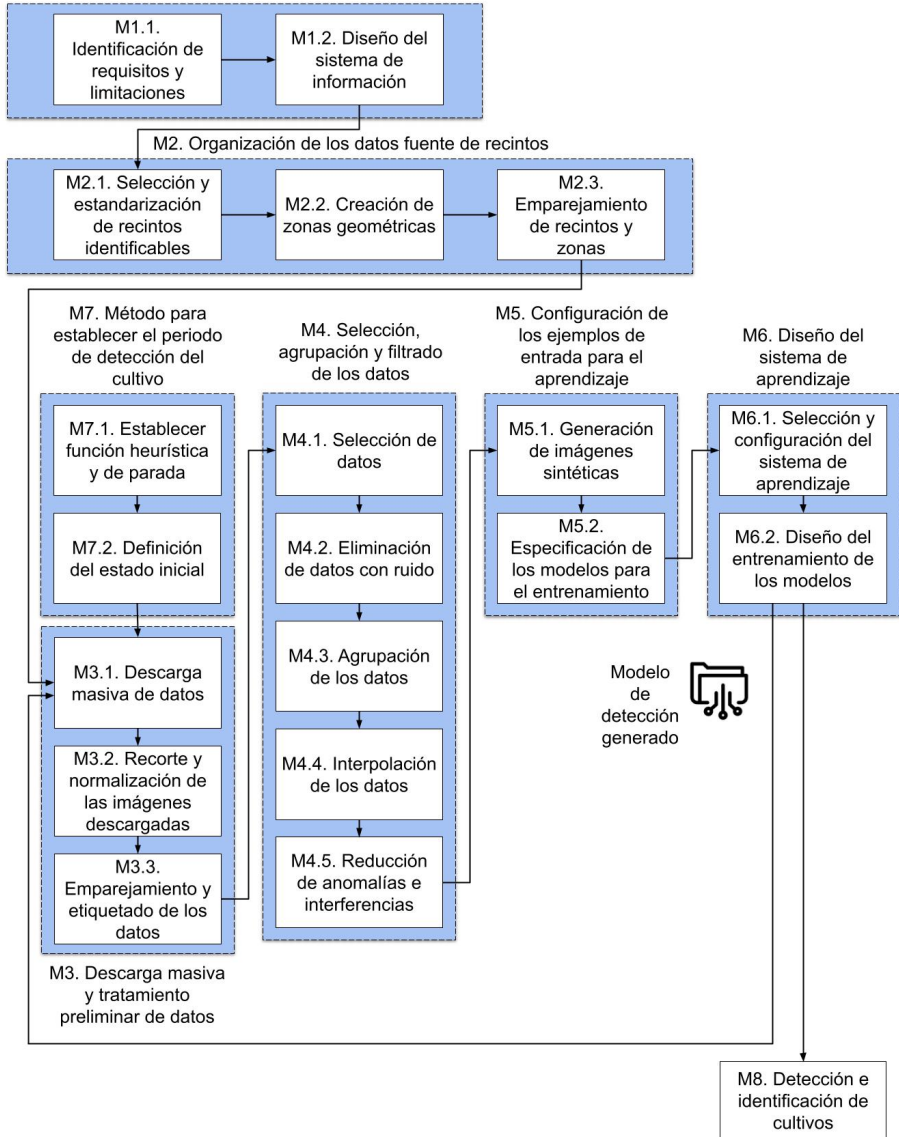


Figura 4.1: Esquema general del método propuesto.

El elemento básico de representación que será utilizado en el método es la evolución de cada píxel de un recinto a lo largo del tiempo, combinando las distintas bandas de información que ofrecen las imágenes de satélite. De esta manera, cada entrada del sistema de aprendizaje será un elemento multitemporal y multiespectral a nivel de píxel, que habrá que procesar para los recintos estudiados. Por lo tanto, será necesario descargar un conjunto de imágenes a lo largo del tiempo y extraer toda la información de las bandas disponibles agrupadas a nivel de píxel, tal y como se explicará en las siguientes secciones.

4.1. Análisis de requisitos y diseño del sistema de información (M1)

En esta primera fase del método se recopila información previa sobre los requisitos y limitaciones del proyecto concreto que se quiere desarrollar. Estas pueden afectar a la implementación del método, y son relativas especialmente al límite de tiempo o a la capacidad de almacenamiento disponible. Los requisitos y limitaciones los establece el gestor del proyecto, es decir, la persona que conoce todas las peculiaridades del proyecto, y que adquiere la responsabilidad en la toma de decisiones. Estas decisiones, además de establecer los requisitos y limitaciones, también incluyen otras cuestiones referentes al presente método, como establecer el umbral mínimo de acierto, necesario en la Sección 4.7.1.

También es necesario realizar en esta fase el diseño general del sistema de información, acorde a dichas limitaciones, para poder almacenar y tratar los datos resultantes de la ejecución del método. Esta tarea será realizada por el propio desarrollador del proyecto.

4.1.1. Identificación de requisitos y limitaciones (M1.1)

El primer paso para poder aplicar el método de manera correcta es disponer de todos los requisitos y limitaciones que presente el proyecto que se va a abordar, y las características del sistema software de la organización,

4.1. ANÁLISIS DE REQUISITOS Y DISEÑO DEL SISTEMA DE INFORMACIÓN (M1)

por lo que el objetivo de esta primera fase del método es establecer estos requisitos y limitaciones. En este caso, se destacan, en primer lugar, tres aspectos importantes: el tiempo disponible, la capacidad de almacenamiento y el porcentaje de acierto mínimo.

El tiempo total de ejecución del sistema puede ser una de las mayores restricciones que tenga la aplicación del método, sobre todo si se está abarcando un gran número de recintos para el análisis. En algunos casos, por ejemplo debido a procedimientos administrativos de los organismos públicos de gestión, el análisis completo de los cultivos debe realizarse antes de una fecha administrativa concreta –por ejemplo, cuando se trata de gestionar subvenciones de la PAC, o alegaciones a expedientes administrativos–. Esta fecha límite de entrega de resultados deberá tener en cuenta el tiempo de análisis y procesado de las imágenes, y la generación de los propios resultados. Por tanto, la última fecha en la que se puedan obtener y analizar imágenes de satélite estará condicionada por estos tiempos de cálculo.

La capacidad de almacenamiento es la otra gran limitación que se puede encontrar en un proyecto de este tipo. Es común que las administraciones públicas no permitan exportar los datos a otras fuentes externas, como pueden ser sistemas de almacenamiento en la nube, y restrinjan el análisis a un uso local de los datos por eficiencia o por seguridad. Por esta razón, la infraestructura donde se aplique el método debe estar preparada para ello. En el caso de encontrarse con limitaciones en la infraestructura, se deberá tener en cuenta a la hora de establecer el número de imágenes que se necesitan descargar.

Otro de los requisitos que se debe establecer de manera previa es el valor mínimo de confianza del modelo, que se denominará como umbral de acierto. Este valor será el porcentaje de acierto mínimo de cada modelo basado en los resultados de su predicción. Por tanto, se debe determinar el umbral de acierto mínimo deseado generado por cada modelo según las exigencias del proyecto. Este punto también es clave en el caso de servir de soporte para los sistemas de justificación de expedientes administrativos si dependen de los resultados de estos procesos. El sistema puede diseñarse para garantizar un umbral de acierto determinado, pero ha de tenerse en cuenta que cuanto más grande sea el umbral de acierto requerido, mayor pueden ser los requisitos de tiempo y capacidad de almacenamiento

mencionados anteriormente, por lo que habrá que evaluar los requisitos del proyecto y ponderar estos factores para lograr los objetivos establecidos.

4.1.2. Diseño del sistema de información (M1.2)

El siguiente paso antes de implementar el sistema de información es realizar su diseño, que incluirá, al menos, una máquina o servidor para realizar los cálculos, así como una base de datos donde almacenar la información que se va generando como parte del propio método.

Las prestaciones de la máquina de ejecución que realizará la generación de modelos y el análisis de recintos debe tener unas características adecuadas, ya que cuanto más potente sea, menos tiempo será necesario invertir en todos los procesos, teniendo en cuenta las restricciones de tiempo de la sección anterior. Se recomienda que la máquina utilizada disponga de tarjeta gráfica dedicada, puesto que los procesos de entrenamiento del sistema de aprendizaje se ejecutarán mucho más rápido en este tipo de componente. También puede llegar a ser interesante disponer de varias máquinas para paralelizar y agilizar los procesos, especialmente cuando las restricciones de tiempo sean estrictas.

Un aspecto importante a tener en cuenta es la capacidad de almacenamiento del sistema, ya sea interna o externa. Los proyectos a los que va orientado el método propuesto se supone que tratan con varios miles de recintos, y sus respectivas imágenes satelitales a lo largo del tiempo, por lo que el número de terabytes que se debe almacenar puede ser elevado.

En cuanto a la propia base de datos, será suficiente con disponer de una base de datos relacional clásica, ya que las imágenes de satélite descargadas se almacenarán directamente en los dispositivos de almacenamiento (como se indica en la Sección 4.3.1). El motivo principal de no integrar las propias imágenes en la base de datos es reducir la complejidad del sistema de información, de tal forma que en la base de datos relacional se almacenará la ruta o directorio donde se almacena cada imagen. De esta forma, cuando se desee trabajar con la imagen, será suficiente obtener el directorio donde se encuentra almacenada y el nombre de la imagen para poder acceder a ella. Esta base de datos almacenará distintos tipos de información, que se detallarán en cada una de las secciones del método.

4.2. Organización de los datos fuente de recintos (M2)

Una vez establecidos los requisitos y limitaciones, se debe recopilar la información de los recintos que se desean analizar. En esta fase se seleccionarán, estandarizarán y estructurarán los recintos geométricos de las parcelas de cultivo que se quieren analizar. Los procesos que se mencionan en esta fase se aplicarán tanto a los recintos que se utilizarán para generar los modelos, como para los recintos que van a ser monitorizados e identificados sus cultivos. La selección de los recintos vendrá determinada por la elección del área del trabajo, que será definida por el gestor del proyecto en función del objetivo del propio proyecto, mientras que el resto de procesos requerirán del software correspondiente implementado por el desarrollador.

4.2.1. Selección y estandarización de recintos identificables (M2.1)

Inicialmente, es necesario disponer de la información geográfica de las parcelas o recintos que se desean analizar –entendiendo por información geográfica al polígono formado por una secuencia ordenada de coordenadas que delimitan el recinto–. También será necesario conocer la información de los cultivos declarados en los recintos en años anteriores, para poder entrenar el sistema de aprendizaje. Esto es, el cultivo o tipo de cultivo del recinto. Por tanto, en esta fase se recopilarán los recintos geométricos que se desean analizar y se estandarizará su información.

La información de los recintos que se van a analizar y los que se utilizarán como ejemplos de entrenamiento puede proceder de distintas fuentes o presentarse en distintos formatos, por lo que es necesario unificarla y estandarizarla en un formato determinado. Algunos de los formatos de representación geográfica más extendidos son Shapefile, KML o GeoJSON, entre otros [101], por lo que habrá que seleccionar uno adecuado para que el software que se implemente pueda manejar estos datos.

Una vez unificado el formato, se deben estandarizar los distintos campos de información de los recintos, como son:

CAPÍTULO 4. MÉTODO DE IDENTIFICACIÓN DE CULTIVOS

- El sistema de coordenadas. Si los datos provienen de distintas fuentes, es posible que se encuentren representados en distintos sistemas de coordenadas geográficas, y deben unificarse en el sistema de coordenadas que vaya a ser utilizado en el proyecto.
- Los errores de formación de los polígonos. Con respecto a las coordenadas, también es posible que los polígonos de los recintos incluyan, en algunos casos, errores geométricos que impidan su procesamiento (por ejemplo que no estén cerrados, que tengan vértices anómalos, etc).
- La información del cultivo declarado. Debe unificarse la nomenclatura de estos atributos para poder tipificar los cultivos. También debe tenerse en cuenta que puede requerirse, según las necesidades de análisis, la segregación por variedades de cultivos o el agrupamiento por cultivos afines.

Para que este proceso sea eficiente, lógicamente deberán implementarse módulos software o *scripts* que, de manera masiva y automática, analicen los recintos, estandaricen y curen los datos, y la almacenen en el sistema de información.

4.2.2. Creación de zonas geométricas (M2.2)

El conjunto de recintos identificados conformará el área de trabajo para la aplicación del resto de fases del método. El área de trabajo se corresponde con una región terrestre determinada que abarca todos los recintos seleccionados en la sección anterior, y tendrá que proyectarse en las correspondientes imágenes de satélite que contengan esta área. El método está enfocado a realizar procesamientos de manera masiva, por lo que la extensión del área puede ser tan grande como se necesite, teniendo en cuenta los requisitos y limitaciones expuestos en la Sección 4.1.

Para aplicar las distintas operaciones que se explicarán en la Sección 4.3, se puede trabajar directamente con las imágenes descargadas de la fuente de datos seleccionada. No obstante, para mejorar la eficiencia de ejecución

4.2. ORGANIZACIÓN DE LOS DATOS FUENTE DE RECINTOS (M2)

del procesamiento para el emparejamiento con los recintos, en esta fase se propone dividir el área de trabajo en una serie de zonas geométricas –polígonos geográficos cuadrados– con un tamaño fijo.

El objetivo principal de utilizar estas zonas geométricas reside en la posibilidad de trabajar con imágenes de satélite con un tamaño más reducido, lo que favorecerá la eficiencia de cómputo necesaria para aplicar el método, y en general reducirá también el tiempo de procesamiento, incluso teniendo en cuenta que el uso de este sistema por zonas conlleva un procesamiento previo.

Para generar estas zonas geométricas, se propone partir de la malla de barrido del satélite seleccionado, y subdividir las cuadrículas del barrido en otras más pequeñas, con unas dimensiones determinadas y comunes para todas las zonas. El tamaño de las zonas geométricas debe venir determinado por el tamaño de las imágenes satelitales de la fuente de datos seleccionada y por la efectividad para el uso de estas zonas por el sistema de información. El tamaño de las zonas geométricas no debe ser excesivamente pequeño, ya que la eficiencia ganada al trabajar con imágenes más reducidas se perdería si se usa un gran número de imágenes. Así mismo, si el tamaño es cercano al tamaño original de las imágenes, la ganancia en cuanto a la eficiencia tampoco será interesante, puesto que el coste computacional seguirá siendo elevado al procesar imágenes de gran tamaño.

Estas zonas generadas deben emparejarse con el área real de trabajo, de manera que se almacenen únicamente aquellas zonas geométricas que sean de interés. Un ejemplo para la región de Extremadura se puede apreciar en la Figura 4.2, donde cada cuadrícula del mapa se corresponde con una zona geométrica.

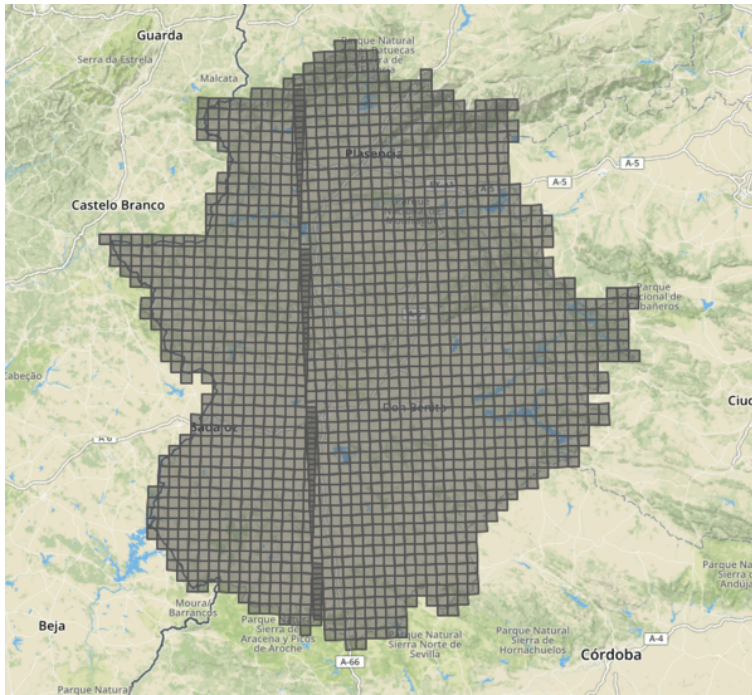


Figura 4.2: Ejemplo de un conjunto de zonas geométricas para una región determinada (Extremadura).

4.2.3. Emparejamiento de recintos y zonas (M2.3)

Al utilizar el sistema de división por zonas geométricas, es necesario realizar el proceso de emparejar cada uno de los recintos con las correspondientes zonas geométricas donde se encuentran ubicados geográficamente. Esta fase consiste en realizar este emparejamiento iterando por todos los recintos e identificando con qué zonas geométricas existe intersección, con el consiguiente almacenamiento en la base de datos.

En la mayoría de los casos, cada uno de los recintos de estudio estarán incluidos en una única zona geométrica, pero puede darse el caso de que, al encontrarse en el límite de una zona geométrica, las coordenadas del recinto

4.2. ORGANIZACIÓN DE LOS DATOS FUENTE DE RECINTOS (M2)

estén contenidas en dos o más zonas (el número de zonas también puede ser elevado en función de la extensión del recinto), como puede apreciarse en el ejemplo de la Figura 4.3.

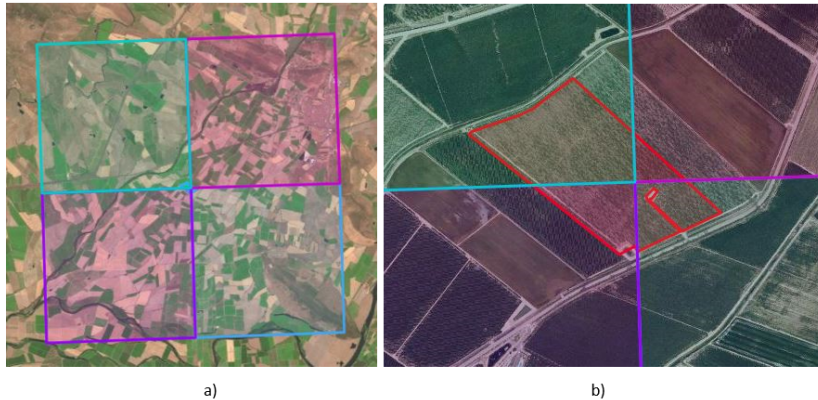


Figura 4.3: Ejemplo de emparejamiento entre recintos y zonas geométricas.

a) Ejemplo de cuatro zonas geométricas adyacentes. b) Ejemplo de coordenadas de un recinto (en rojo) emparejado con cuatro zonas debido a su ubicación geográfica.

Existe, además, otra peculiaridad a la hora de realizar el emparejamiento. Debido a la presencia de distintos husos de coordenadas, en las regiones limítrofes entre husos, las zonas geométricas pueden solaparse, como se aprecia en la Figura 4.2 presentada anteriormente. Por este motivo, a la hora de realizar el emparejamiento, puede que un recinto esté encuadrado en dos zonas con husos distintos, lo que provocaría que se duplicase la información. En este caso, debe programarse la selección de sólo una de ellas. Así mismo, será necesario automatizar esta fase mediante los correspondientes módulos software y almacenar todos los resultados en la base de datos.

Por tanto, al finalizar la fase M2 del método, se dispondrá de un conjunto de recintos geométricos estandarizados, un conjunto de zonas geométri-

cas homogéneas que abarquen el área de trabajo, y el emparejamiento entre ambos conjuntos. Esta información se utilizará para la descarga y extracción de datos.

4.3. Descarga masiva y tratamiento preliminar de datos (M3)

En esta sección se detallan los procesos encaminados a la descarga masiva de las imágenes desde los servidores de los satélites y su procesamiento previo.

4.3.1. Descarga masiva de datos (M3.1)

En esta fase se describe el proceso de descarga de las imágenes de satélite, comenzando por la selección de la fuente de datos (es decir, el satélite o satélites escogidos). En general, la propuesta se describe para utilizar una única fuente de datos, pero se puede adaptar al uso combinado de varias, como pueden ser diferentes satélites, LIDAR, ortofotos, drones, etc.

Una vez que se dispone de la información geográfica normalizada de los recintos emparejados a las zonas geométricas correspondientes, obtenidas en las fases anteriores, se deben seleccionar aquellas imágenes de satélite que cubren el área de trabajo, es decir, que cubren todos los recintos. La restricción en este caso se basa en la necesidad de disponer de alguna herramienta software que permita la descarga automática de las imágenes de la fuente seleccionada. Este software accederá al servidor de descarga de imágenes del sistema de satélites, que habitualmente suele disponer de una API o servicio similar para poder acceder a los datos correspondientes.

Para la descarga de las imágenes necesitamos, por tanto, establecer dos parámetros: la fecha o rango de fechas del que se desean obtener las imágenes, y el área geográfica que se quiere analizar. Siguiendo con la propuesta de utilizar las zonas geométricas, se descargarán las imágenes correspondientes para todas las zonas geométricas que hayamos generado. El hecho de que la descarga se denomine como “masiva” implica que se

4.3. DESCARGA MASIVA Y TRATAMIENTO PRELIMINAR DE DATOS (M3)

descargará una cantidad de imágenes proporcional a la extensión del área de trabajo, y multiplicada por la amplitud del rango de fechas que se vaya a estudiar.

Tras este proceso se irán descargando imágenes que abarcarán los recintos seleccionados y en las fechas establecidas. En cada paso, se comprobará si la imagen seleccionada ya está actualmente almacenada de manera local pues, para un conjunto de recintos cercanos, es posible que solo se requiera de una única imagen por fecha. En el caso de no disponer de ninguna imagen descargada para esta zona geométrica, se deberá seleccionar una de ellas (recalcando que todo este proceso se debe realizar de manera automática). Una vez la imagen es seleccionada, se descarga y se almacena localmente para su uso posterior.

4.3.2. Recorte y normalización de las imágenes descargadas (M3.2)

Con el objetivo de reducir el tiempo de procesamiento, atendiendo a que este se realiza de manera masiva, en esta fase se propone recortar las imágenes originales descargadas en las zonas geométricas explicadas en la Sección 4.2.2. Se entiende como recorte de la imagen almacenar únicamente el subconjunto de píxeles correspondiente a las zonas geométricas determinadas. Con ello se consigue que el procesamiento de estas imágenes sea menos costoso en términos tanto de computación como de tiempo. Se debe tener en cuenta que esta fase también debe realizarse de manera masiva y automática. Todo este proceso requerirá de la implementación software correspondiente para realizarse de manera automática. Las imágenes recortadas por cada zona geométrica, a las que denominaremos imágenes de zona, se almacenarán debidamente etiquetadas para su uso. Por tanto, una zona geométrica consiste en un polígono rectangular que se ha fijado para manejar más eficiente los datos, y que se relacionará con una imagen de zona, que es un fragmento de la imagen del satélite original que contiene todos los valores de las bandas espectrales correspondientes de cada uno de los píxeles.

En el caso común donde las imágenes dispongan de varias bandas de información, si se desea trabajar con varias bandas, se requiere de un pro-

cedimiento adicional, aplicado durante el proceso de recorte, comúnmente denominado como normalización de bandas. La normalización consiste en equiparar la resolución de las distintas bandas de información de una imagen de satélite. Cada una de las bandas de información puede tener su propia resolución (tamaño o dimensiones por cada píxel), por lo que se deben normalizar las distintas resoluciones de la imagen, indispensable para poder trabajar a nivel de píxel, y que cada píxel disponga de la misma resolución para todas las bandas espectrales.

4.3.3. Emparejamiento y etiquetado de los datos (M3.3)

Disponiendo tanto de la información geográfica de los recintos seleccionados para ser analizados, así como de las imágenes de zonas, el siguiente paso, descrito en esta fase, es identificar los píxeles concretos que se encuentran dentro de los recintos, de manera que cada píxel del recinto quede referenciado a las imágenes de zona correspondientes.

Para cada uno de los recintos, y considerando las zonas que tiene emparejadas, acorde a lo explicado en la Sección 4.2.3, se obtendrán las coordenadas geográficas de las imágenes de zona. Para cada conjunto de coordenadas de cada píxel (un píxel se delimita con cuatro pares de coordenadas geográficas, relativos a sus cuatro esquinas), se debe comprobar, de manera automática, que el píxel se encuentra íntegramente dentro del recinto, es decir, que existe intersección completa entre ambos conjuntos de coordenadas. Es importante aclarar que se descartarán aquellos píxeles que presenten una intersección parcial con el recinto, ya que la muestra de datos puede estar contaminada debido al posible solapamiento de recintos de otros cultivos, tal y como se puede visualizar en el ejemplo de la Figura 4.4.

El resultado de esta fase es disponer del listado de píxeles de las imágenes de zona asociados a cada recinto, con una identificación de posición. Si entendemos la imagen como una matriz bidimensional, y etiquetamos sus píxeles como un listado de coordenadas relativas ($[0,0]$, $[0,1]$, $[0,2]$...), considerando el primer campo como la posición de la columna y el segundo como la posición de la fila, para cada recinto almacenaremos el listado de píxeles del mismo. A esta identificación la denominamos coordenadas relativas de

4.3. DESCARGA MASIVA Y TRATAMIENTO PRELIMINAR DE DATOS (M3)

píxeles. En el caso del ejemplo de la Figura 4.4, el listado de coordenadas relativas de píxeles sería: $([1,8], [1,9], [2,7], [2,8], [2,9]...)$, como se puede ver en la imagen d) de la Figura 4.4. Este listado deberá almacenarse en el sistema de información, descrito en la Sección 4.1.2.

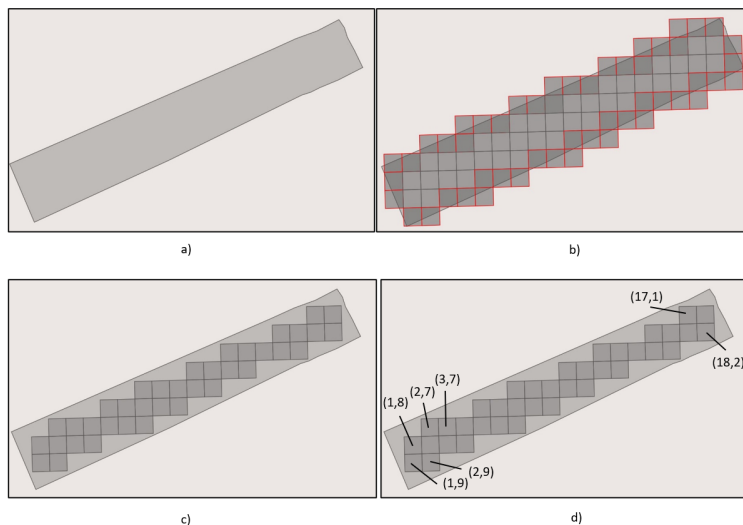


Figura 4.4: Ejemplo de emparejamiento entre recintos y píxeles.

- Ejemplo de polígono de un recinto, representado por una secuencia ordenada de coordenadas geométricas.
- Conjunto de píxeles analizados para el recinto en cuestión, marcados en rojo aquellos píxeles que no tienen intersección completa con el recinto.
- Píxeles seleccionados y emparejados al recinto –siendo solamente aquellos que se encuentran íntegramente dentro del propio recinto–.
- Etiquetado de las coordenadas relativas de píxeles.

Todo este proceso de emparejamiento servirá para, posteriormente, obtener la información concreta de cada píxel en la fecha de la imagen descargada, y, de manera global, agrupar toda la información relativa a los distintos cultivos.

4.4. Selección, agrupación y filtrado de datos (M4)

Disponiendo de la información en bruto descargada, el objetivo de esta fase es establecer qué datos se van a emplear para los procesos de aprendizaje e identificación de cultivos (Sección 4.4.1), filtrar y eliminar aquellos que puedan ocasionar distorsión en el entrenamiento (Sección 4.4.2), decidir cómo se agrupan estos datos (Sección 4.4.3), generar datos en aquellos huecos donde haya carencias (Sección 4.4.4), y finalmente suavizar los datos considerados como anómalos (Sección 4.4.5). Obviamente, estos procesos deberán realizarse mediante la implementación de sus respectivos módulos software.

4.4.1. Selección de datos (M4.1)

En esta fase se deben identificar las bandas espectrales e indicadores derivados que se va a utilizar en el proceso de entrenamiento. A la hora de establecer qué datos se van a utilizar, la opción principal sería utilizar todas las bandas espectrales de información disponibles, frente a otras alternativas como el uso de indicadores de vegetación concretos (como NDVI, NDWI, etc), que conllevarían un tiempo adicional para generar y calcular los citados indicadores, y además conlleva una pérdida de información que podría mejorar los modelos de aprendizaje. Se debe tener en cuenta que, al tratarse de un procesamiento masivo, se podría estar hablando de cientos de imágenes con cientos o incluso miles de gigabytes de información.

Aunque los modelos de aprendizaje generados con todas las bandas espectrales requieren de un mayor tiempo de entrenamiento (al tener que procesar mayor cantidad de datos), esta sería la elección recomendada si se quisiera maximizar el porcentaje de acierto del modelo. La opción del uso de indicadores de vegetación se plantearía si las limitaciones de tiempo lo requiriesen o se fuera a emplear para otros usos, como generación de informes u otros análisis. Adicionalmente, aunque la opción del uso de indicadores reduciría ligeramente las necesidades en cuanto a capacidad de almacenamiento se refiere, no habría diferencias significativas entre ambos enfoques en este aspecto. La opción elegida para este método sería utilizar

4.4. SELECCIÓN, AGRUPACIÓN Y FILTRADO DE DATOS (M4)

todas las bandas disponibles, aunque se podría adaptar a otras opciones en función de las necesidades del proyecto o los recursos disponibles.

Por otra parte, se propone el uso del píxel como unidad mínima de trabajo y predicción para el método [102]. Trabajar a nivel de píxel ofrece varios beneficios con respecto a trabajar con los recintos completos como una única unidad. Principalmente, la elección de este sistema se basa en la posibilidad que aporta poder detectar la superficie concreta del recinto asociada a un cultivo determinado. Es decir, permite detectar si dentro de un mismo recinto coexisten distintos cultivos, de manera que permite detectar una posible plantación ilegal o invasión de delimitaciones con vecinos. También es muy interesante a la hora de aplicar el método para otros fines, como la detección de plagas o enfermedades, análisis de la fertilidad del suelo, o la identificación de problemas de riego en parte del recinto, entre otros.

Además, trabajar a nivel de píxel permite ponderar de manera correcta y equitativa los ejemplos seleccionados para el sistema de aprendizaje: si se trabajase a nivel de recinto como elemento de entrada para el sistema de aprendizaje, deberíamos tener en cuenta que no debería tener el mismo peso un recinto con el doble de superficie que otro. Trabajando a nivel de píxel, todas las entradas al sistema tendrán la misma ponderación. No hay que olvidar, además, que utilizar el píxel en vez del valor medio por cada recinto ofrece una cantidad significativamente mayor de ejemplos de entrenamiento, por lo que implicará, en la mayoría de los casos, mejores modelos de detección.

Por tanto, los datos utilizados en el método serán las bandas espectrales (o índices seleccionados si es el caso) de las imágenes del satélite en un periodo de tiempo, agrupados a nivel de píxel.

4.4.2. Eliminación de datos con ruido (M4.2)

En este paso, el objetivo es eliminar aquellos datos que puedan producir ruido durante el entrenamiento, entendiendo el ruido como un valor erróneo en ciertos datos de las imágenes del satélite. Como se ha tratado en las investigaciones previas, descritas en la Sección 2.1, el origen principal de

la aparición de ruidos e interferencias en las imágenes de satélite es la presencia de nubosidad, que a su vez proyecta su propia sombra sobre el terreno, alterando los valores de la imagen.

Se deben detectar aquellos valores afectados por ciertas condiciones climáticas que pueden desvirtuar la calidad de los datos que se van a analizar. También se debe tener en cuenta que estas condiciones climáticas engloban diversas opciones, como nubosidad, sombras de las propias nubes, o nieve, entre otras, por lo que habrá que seleccionar un sistema de detección de estos fenómenos con un nivel de acierto adecuado, y todo ello a nivel de píxel.

Para conseguirlo, se deben conocer los valores medios de cada banda espectral de información y sus desviaciones, de manera que, si una o varias de las bandas se desvía del rango de datos habitual, por encima o por debajo de un determinado umbral, se debería descartar la información de ese píxel. Estos umbrales se establecen en base a las características de los sensores del satélite o fuente de datos seleccionada y, consecuentemente, atendiendo a sus distintas bandas de información, por lo que no se puede establecer unos umbrales de manera universal.

Para la identificación de píxeles con ruido, se suelen emplear técnicas basadas en aprendizaje con árboles de decisión, que combinen las comprobaciones de los valores de las bandas, tal y como explican Hollstein et al. [51]. No obstante, es usual que, en algunas fuentes de datos satelitales, esta información venga ya incluida en la propia imagen, catalogando los píxeles en estas categorías de ruidos.

4.4.3. Agrupación de los datos (M4.3)

En esta fase, se determina cómo se agrupan los datos para posteriormente generar las entradas del sistema de aprendizaje. Los datos seleccionados para el entrenamiento deben agruparse, de manera que conformen una estructura de datos que relacione el conjunto de bandas o índices a lo largo del tiempo. Es decir, la unidad de aprendizaje y de análisis estará formada por el conjunto de bandas e índices seleccionados, provenientes de las imágenes de satélite, ordenadas cronológicamente.

4.4. SELECCIÓN, AGRUPACIÓN Y FILTRADO DE DATOS (M4)

A la hora de generar las series temporales de datos para el entrenamiento de los modelos, es necesario que las series sean similares en cuanto al número de datos que incluyen, ya que los modelos de aprendizaje requieren ejemplos de entrada homogéneos. Para conseguirlo, se deben tener en cuenta las cuestiones presentadas en esta sección.

Cuando se procesan imágenes de distintos años, es probable que sus fechas relativas de día y mes no coincidan. Esto depende de la frecuencia de disponibilidad de imágenes, y también depende del número de días que tenga el año, considerando los años bisiestos. Por ejemplo, en una zona determinada podríamos tener una imagen el 2 de marzo de 2019, pero el año siguiente la fecha de la misma imagen sería el 1 de marzo de 2020, debido a la fecha de paso del satélite y del posible desfase por el año bisiesto.

Otra de las cuestiones tiene que ver con el solapamiento de las órbitas satelitales es la existencia de varios satélites de una misma constelación. En el caso de misiones que tengan varios satélites, pueden existir algunas regiones en las que se solapen sus órbitas, por lo que se dispondrá de imágenes con mayor periodicidad. En estos casos, las series temporales de los píxeles ubicados en estas regiones tendrían más datos que el resto de regiones.

Para solucionar estas dos situaciones, se propone crear un conjunto de periodos temporales fijos. En estos periodos se realizará una media de los distintos valores que incluyan las imágenes, de forma que en cada periodo se asigne un valor único. El rango de fechas de los periodos podrían ser semanas, quincenas, meses, etc. Por ejemplo, si se decide utilizar quincenas, el posible conjunto de periodos temporales podría ser: del 1 al 15 de abril, del 16 al 30 de abril, del 1 al 15 de mayo, y así sucesivamente. No obstante, cuanto menor sea el periodo, mayor será el número de datos de la serie temporal, y presumiblemente mejores serán los resultados que ofrezca el modelo. En cambio, utilizar un periodo temporal corto tiene como desventaja la posibilidad de disponer de varios periodos sin ninguna imagen válida, debido a las posibles interferencias comentadas en la Sección 4.4.2.

Por tanto, se propone aplicar el método trabajando con unos periodos de duración mínima, donde se asegure disponer de al menos una imagen por periodo en función de la frecuencia de disponibilidad de imágenes del satélite seleccionado. Cuando en un periodo se disponga de varias imáge-

nes, una manera habitual de unificar la información es aplicar la media aritmética para cada banda a nivel de píxel, resultando en un único valor por cada banda, periodo y píxel.

4.4.4. Interpolación de los datos (M4.4)

La interpolación de datos dentro de una serie temporal de datos de satélite consiste en el cálculo de los valores desconocidos de la serie a partir de otros valores cuyo valor sí es conocido (este proceso también se denomina como generación de datos sintéticos). Estos valores desconocidos pueden aparecer ya que, una vez retirados los píxeles con valores no válidos y agrupados por periodos, es posible que en alguno de los periodos no exista ningún valor, por lo que es necesario completar las series temporales del mismo píxel a lo largo del tiempo (es decir, a través de varias imágenes), de manera que todas las series temporales (una por píxel) contengan un número equivalente de datos.

Sería posible igualar el valor eliminado a un valor estático predeterminado, por ejemplo a 0. No obstante, esto podría dar lugar a la creación de patrones irreales, que produzcan ruido en el proceso de entrenamiento. Por ejemplo, en un año en el que haya una fecha donde la región de estudio al completo esté nublada, todos los valores de esa fecha se establecerían a 0, y el sistema de aprendizaje podría interpretar erróneamente como patrón que ese valor esté a 0 para hallar la concordancia con algunos de los cultivos estudiados.

Para evitar esta anomalía en el aprendizaje, se propone la aplicación de una interpolación que genere los datos incompletos, de manera que todos los píxeles dispongan de la misma cantidad de datos y se establezcan series temporales homogéneas en cuanto al número de valores que incluyen. Esta interpolación se puede aplicar mediante diferentes técnicas [38] [56] [57]: realizando una interpolación lineal con valores vecinos, utilizando redes neuronales para simular los datos incompletos, o mediante distintos modelos estadísticos.

4.4.5. Reducción de anomalías e interferencias (M4.5)

En esta fase se van a suavizar aquellos valores considerados anómalos y que puedan suponer un perjuicio en el entrenamiento. Este caso es diferente al mencionado en la Sección 4.4.2, donde se detectaban y eliminaban datos erróneos. Aquí se detectan valores desviados de la tendencia de sus vecinos, pero que aportan información a la evolución del píxel. Estos valores anómalos pueden ser debidos a que han sufrido algún problema con los sensores de los satélites, o bien porque los propios algoritmos de detección de condiciones climatológicas adversas no hayan funcionado correctamente. Por ello, se propone aplicar un filtrado adicional que compruebe si existe algún valor en las bandas excesivamente desviado de la tendencia, para suavizarlo y que no perturbe los resultados finales de los modelos. Es importante recalcar que el objetivo en este punto es suavizar dichos valores anómalos, y no eliminarlos completamente, ya que entonces deberíamos volver a aplicar una interpolación según lo expuesto en la sección anterior, además de perder información.

Este filtro debe comprobar cada valor de la serie temporal, y compararlo con sus valores vecinos, de manera que se detecte si el valor sufre una desviación excesiva. Aunque en este punto también existen redes neuronales específicas para detectar los valores anómalos, una de las técnicas más extendidas es la implementación de un filtro de Hampel [52], para detectar estos valores y reemplazarlos por otros más cercanos a la tendencia de la serie temporal. Para saber si una muestra es atípica o anómala, durante la aplicación del filtro de Hampel se calculan la media y la desviación estándar de sus K vecinos, y se comprueba si el valor de la misma sobrepasa una cierta cantidad de desvíos, declarando en tal caso el valor como atípico, y reemplazándolo por su valor medio.

Por tanto, el resultado de la fase M4 será un conjunto de series temporales de datos, habiendo filtrado sus posibles anomalías y agrupados por periodos temporales, recordando que se genera una serie temporal por cada banda de información de cada píxel.

4.5. Configuración de los ejemplos de entrada para el aprendizaje (M5)

Con los datos ya seleccionados, agrupados y filtrados, es necesario estructurarlos con el objetivo de adecuarlos como ejemplos de entrada para el sistema de aprendizaje. Hay que recordar que los distintos elementos utilizados para el aprendizaje contienen varias bandas de información, y cada una de estas bandas abarca un rango de fechas determinado, por lo que se dispone de información multidimensional, que debe ser formateada para su uso. Esta fase desarrolla esta configuración de los datos que conformarán las entradas del sistema de aprendizaje.

4.5.1. Generación de imágenes sintéticas (M5.1)

En este proceso se van a generar los ejemplos de entrada para la generación de modelos. La información recopilada debe relacionarse de manera que pueda servir como entrada válida para un sistema de aprendizaje máquina. En este caso, al disponer de datos multiespectrales (varias bandas de información o índices, que además pueden proceder de distintas fuentes de datos) y multitemporales (varios periodos), se propone crear una matriz para cada píxel y año, cuyas dimensiones se corresponden con las citadas bandas y periodos. Esta matriz será el formato de entrada para el sistema de aprendizaje, que denominaremos como Imagen Sintética Multitemporal y Multiespectral (ISMM). Por tanto, la ISMM será una matriz de dos dimensiones, cuyas variables se corresponden con los periodos temporales y con la información de las bandas e indicadores, respectivamente.

En la Figura 4.5 se puede apreciar un ejemplo del proceso de composición de la ISMM, aplicado automáticamente a cada píxel. Teniendo en cuenta un único píxel –representado en verde en cada una de las bandas de información–, se deben considerar el conjunto de imágenes que incluyan a este píxel durante un período de tiempo. Como cada una de las imágenes contiene información de las bandas seleccionadas, estos datos se extraen para cada una de las fechas (de las imágenes correspondientes) para generar la entrada al sistema de aprendizaje, como se puede ver en la gráfica

4.5. CONFIGURACIÓN DE LOS EJEMPLOS DE ENTRADA PARA EL APRENDIZAJE (M5)

situada a la derecha de la figura. En ella se muestra, en el eje de ordenadas, el valor del píxel (normalizado a valores entre 0 y 1) en cada una de las bandas para un periodo concreto, y en el eje de abscisas, la evolución de los valores a lo largo del tiempo (en este ejemplo, en datos de 20 periodos temporales).

La elección de este formato de entrada se basa en la posibilidad de ser válida para distintos sistemas de aprendizaje, principalmente para las Redes Neuronales Convolucionales, que son muy utilizadas en el análisis de imágenes, y más concretamente de imágenes satelitales, como se comentó en la Sección 2.1.

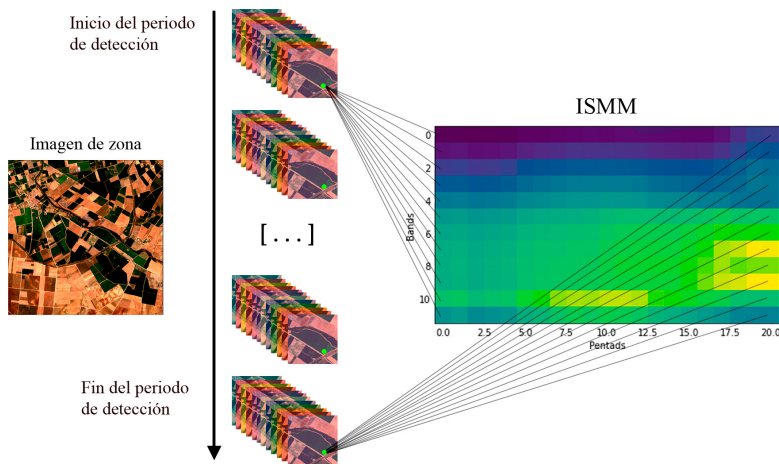


Figura 4.5: Representación de la extracción de la información por píxel de las distintas bandas y periodos para generar una ISMM.

Aunque los modelos de aprendizaje pueden generarse con los datos de cultivos de un único año, el método propone utilizar datos plurianuales, de manera que se disponga de información de diversos años para el entrenamiento. Cuando se refiere a un año, debe entenderse que son los meses significativos de un cultivo que abarca un periodo temporal en un año de siembra.

Usar varios años en el proceso de creación de modelos puede ser clave para obtener unos resultados con mayor índice de acierto al tener más datos de entrada pero, sobre todo, ayuda a evitar que el modelo aprenda patrones sobre la fenología de un año en concreto, y sea capaz de interpretar los patrones de manera más genérica para cada cultivo. Por ejemplo, cuando se entrena con datos de años de sequía, para que el modelo sea capaz de realizar la predicción correctamente sería recomendable combinar los datos de entrada con datos de otros años y diluir la tendencia de ese año de sequía. Un caso similar se podría encontrar en años con periodos de nubosidad extensos, donde se hayan podido analizar pocos datos representativos. Así se logra disminuir la dependencia de los modelos de los años concretos del entrenamiento.

Por lo tanto, el elemento básico de entrada al modelo de entrenamiento será cada una de las ISMM generadas, que se corresponden con la información de las bandas de un píxel a lo largo de un periodo dentro de un año, y abarcando, a ser posible, diferentes años.

4.5.2. Especificación de los modelos para el entrenamiento (M5.2)

En esta fase se deben establecer qué modelos de aprendizaje máquina se van a necesitar para cumplir los objetivos marcados por el proyecto concreto. De esto dependerá el diseño del sistema de aprendizaje, donde se seleccionarán los recintos que se emplearán como ejemplos en el proceso antes de comenzar con el entrenamiento de los modelos.

Tal y como se explicó en la Sección 4.2.1, se puede disponer de la información tanto del cultivo declarado como de su variedad para cada uno de los recintos. Por tanto, algunas opciones para el posterior entrenamiento podrían ser generar un modelo por cada cultivo, por cada variedad de un cultivo, agrupando varios cultivos en una clase común, o incluso por características concretas de los recintos (por ejemplo, por coeficiente de regadío, situación de registro de actividad, etc).

La decisión de generar unos modelos u otros depende de los requisitos y objetivos del proyecto, pero es fundamental que los objetivos estén perfec-

tamente definidos antes de proceder al diseño del sistema de aprendizaje. A modo de ejemplo, es posible encontrar situaciones en las que el proyecto fije la necesidad de identificar un cultivo con una variedad concreta, por lo que los modelos deberán ser muy específicos. En otros casos, también a modo de ejemplo, es posible que se desee distinguir entre cultivos de tierra arable y recintos de árboles frutales, para lo cual se podrán agrupar los cultivos en grandes grupos con características similares.

4.6. Diseño del sistema de aprendizaje (M6)

Habiendo definido los modelos de aprendizaje máquina que se van a necesitar en el proyecto, se debe seleccionar el tipo de sistema de aprendizaje concreto, que sea más conveniente para los objetivos marcados, y diseñar su estructura y parámetros internos. Una vez diseñado el sistema de aprendizaje, se debe diseñar el propio entrenamiento de los modelos, estableciendo principalmente qué datos se utilizan para entrenar cada modelo, y teniendo especial atención a la distribución de los ejemplos positivos y negativos.

4.6.1. Selección y configuración del sistema de aprendizaje (M6.1)

En esta fase se debe decidir el sistema de aprendizaje para generar los modelos. A pesar de disponer de una gran cantidad de alternativas, mencionadas en la Sección 2.1, una de las opciones más recomendables es trabajar con Redes Neuronales Convolucionales (CNN), debido a su eficacia a la hora de detectar e identificar cultivos partiendo de imágenes de satélite, como se argumenta en Wang et al. [90], y ser especialmente útiles en la detección de patrones dentro de imágenes. Las imágenes sintéticas (ISMM), como se explica en la Sección 4.5.1, son idóneas para aplicarlas en las CNNs [103], por lo que el método propone utilizar este tipo de redes.

La primera decisión de diseño para modelizar la CNN es establecer su topología. Esto incluye la capa de entrada, cuyas dimensiones se corresponden con las dimensiones de las ISMM; las distintas capas internas; y la capa de salida, que en este caso tendrá un único nodo, que representará el

porcentaje de similitud del píxel analizado con el cultivo del modelo de la CNN.

En cuanto a la elección del resto de hiperparámetros, comúnmente se aplican distintas técnicas [104], donde la más habitual es la búsqueda manual a base de prueba y error, realizando pruebas con un modelo concreto y variando los parámetros hasta encontrar la mejor solución. Sin embargo, esta técnica presenta un coste computacional y temporal elevado, por lo que se han de tener en cuenta otras alternativas como el uso de algoritmos genéticos [105], la búsqueda aleatoria [106], o la optimización bayesiana [107], entre otras. A pesar de que todos estos métodos suelen ser costosos, también se pueden encontrar algunas propuestas para acelerar el ajuste de hiperparámetros [108]. La elección de una de estas técnicas va a depender de las limitaciones de tiempo, cómputo y la precisión que requiera el sistema.

4.6.2. Diseño del entrenamiento de los modelos (M6.2)

Habiendo seleccionado la división de los datos para cada modelo, según la Sección 4.5.2, y una vez está definido el sistema de aprendizaje, el objetivo de esta fase es definir la distribución de los datos en dos grandes bloques: ejemplos positivos –aquellos datos del cultivo o cultivos que se desea predecir en cada modelo– y ejemplos negativos –el resto de datos, es decir, otros cultivos o categorías de los recintos–. Una práctica habitual es que los datos positivos y negativos estén balanceados (que cada grupo contenga el mismo número de ejemplos de entrada). No obstante, esta proporción se debe adaptar a las necesidades del proyecto, ya que es posible que en algunos casos interese aumentar el número de ejemplos negativos introducidos para el entrenamiento, como puede ser el caso de grupos de cultivos minoritarios (con pocos datos), o también en el caso de querer aumentar todavía más la precisión en un cultivo determinado. Es importante recordar que cuanto mayor sea el número de ejemplos utilizado para generar un modelo, mayor será el tiempo de entrenamiento, aunque generalmente también será mayor la precisión. A modo de ejemplo, si se desea generar un modelo de detección del cultivo del tabaco y maximizar su precisión, como ejemplos positivos para el entrenamiento se utilizarán todos

los datos de los cultivos de tabaco que estén disponibles, y en el máximo número de años; como ejemplos negativos se incluirán datos de todos los cultivos restantes, para que el sistema aprenda a distinguirlos con mayor precisión. En este caso, es interesante considerar cultivos con una evolución fenológica similar, correspondiente a los cultivos de verano, como pueden ser el tomate, el arroz, el maíz, etc. Esto no quiere decir que solo se incluyan estos datos como ejemplos negativos, pero sí que se deberá priorizar su uso por delante de otros como los datos de árboles frutales, por ejemplo. Como regla general, se debe intentar buscar ejemplos negativos que difieran poco con los ejemplos positivos para que los modelos logren distinguirlos.

Disponiendo de ambos bloques, también se deben distribuir cada uno en otros tres grupos: el conjunto de datos para el entrenamiento, el conjunto para la validación, y el conjunto para las pruebas o tests [109]. La distribución en cada uno de estos tres conjuntos deberá decidirse en función de la cantidad de datos disponibles para el entrenamiento, por lo que deberá ser ajustada en cada caso. Habiendo seleccionado las proporciones, sí es importante aplicar estas divisiones de manera aleatoria, y teniendo en cuenta que este proceso se realiza a nivel de píxel. Las proporciones más habituales para distribuir los datos son entre un 60-80 % para el conjunto de entrenamiento, y entre un 10-20 % tanto para la validación como para los tests.

Con las entradas a nivel de píxel generadas, los datos agrupados por cultivo o grupos de cultivos, la red neuronal diseñada y creada, y los datos agrupados en entrenamiento, validación y test, en este punto se estaría en disposición de comenzar el entrenamiento de los modelos de aprendizaje con un periodo de detección determinado. Para concluir, hay que resaltar que esta fase de selección de ejemplos es crítica para que los modelos se adecúen a la identificación de los cultivos que se quieren detectar, y así se pueda cumplir con los objetivos concretos especificados en el proyecto.

4.7. Búsqueda del periodo de detección de los cultivos (M7)

Esta sección propone un método automático para establecer el periodo adecuado en el año para la detección de los cultivos que se desean analizar. Para ello, se debe combinar el método propuesto con todas las fases desarrolladas desde la Sección 4.3 en adelante. Cabe destacar que todo lo descrito en esta sección puede ser aplicado de manera individual por cada cultivo o grupo de cultivos que se desea analizar. Establecer un periodo de detección adecuado es muy importante, ya que puede influir considerablemente en el porcentaje de acierto obtenido con los modelos de aprendizaje. Incluir periodos o meses poco significativos (por ejemplo, los meses de invierno para los cultivos de verano) ocupa una cantidad de memoria adicional que no aporta mejoras al sistema de aprendizaje, además de incrementar el tiempo de cálculo y generar ruido en el entrenamiento, por lo que la elección de un periodo de detección adecuado gana mucha importancia. Se debe señalar que este proceso es habitual que se ejecute en las fases iniciales del proyecto, cuando no se tengan definidos estos periodos de detección. Cuando el proyecto se va a aplicar en campañas posteriores, se pueden utilizar los periodos calculados inicialmente, ya que no suele haber grandes variaciones de un año hacia otro.

La técnica propuesta [110] se basa en el algoritmo de ascenso de colinas (más conocido por su nomenclatura en inglés, *hill climbing*, HC), y más concretamente la variación denominada como máxima pendiente (*steepest ascent hill climbing*) [111], que consiste en explorar todas las alternativas desde el estado actual para conseguir un resultado adecuado sin analizar todas las combinaciones de todas las alternativas.

El algoritmo HC, de manera general, parte de un estado inicial, al que se le asigna una puntuación de acierto, y busca alternativas que se irán puntuando de manera iterativa. Escogiendo la mejor alternativa, y siempre que sea mejor que el estado actual, se repite iterativamente el mismo proceso hasta que no haya mejora. En el caso de esta propuesta, los estados se corresponden con los valores de predicción en los distintos periodos temporales y las alternativas serán periodos contiguos al periodo actual.

4.7. BÚSQUEDA DEL PERIODO DE DETECCIÓN DE LOS CULTIVOS (M7)

4.7.1. Establecer función heurística y de parada (M7.1)

En primer lugar, en esta fase se va a realizar la definición de las condiciones y requisitos del algoritmo para la detección de periodos, centrándose en establecer la métrica por la que se evaluará cada estado o modelo durante la ejecución del algoritmo.

Existen diversas métricas extendidas para valorar los modelos de aprendizaje, como pueden ser la precisión o acierto (*accuracy*), el valor de pérdida (*loss value*), o algunas más elaboradas como la métrica F-Score [112]. Se puede seleccionar una métrica u otra en función de las necesidades, y la métrica elegida como unidad comparativa servirá para determinar si un modelo es mejor que otro. De cualquier forma, va a estar condicionada por los parámetros de acierto que se estén buscando en el proyecto. Como idea básica, la función heurística del algoritmo que se propone es la medida F-Score de los distintos modelos, ya que funciona mejor en los casos donde la cantidad de ejemplos esté desbalanceada (por ejemplo, que se encuentren más datos de unos cultivos que de otros, como puede ser el caso).

Otro paso importante es establecer las condiciones de parada del algoritmo. El comportamiento general de HC es parar cuando se alcanza un modelo cuya puntuación no mejora a los anteriores. No obstante, por motivos de rapidez, es interesante plantear que en este punto se puede establecer un umbral de acierto para los resultados de los modelos (que puede venir determinado por los objetivos del proyecto), y que una vez alcanzado dicho umbral, el algoritmo pare. Además, este umbral puede ser modificado en función de los distintos tipos de cultivos o de las distintas necesidades administrativas para las que se aplique el método de detección del cultivo. Es decir, este umbral de acierto puede definirse dependiendo de la precisión y confianza en los modelos que se requieran en los diferentes procesos administrativos. Por tanto, será responsabilidad del gestor del proyecto definir el umbral de acuerdo mínimo de los modelos.

4.7.2. Definición del estado inicial (M7.2)

El algoritmo siempre necesita un estado inicial para comenzar el procesamiento, por lo que es necesario determinar el periodo de detección de

dicho estado. En este punto existen dos recomendaciones: por un lado, en caso de disponer de experiencia previa en el análisis de cultivos, el periodo inicial se puede basar en dicha experiencia, habiendo entrenado modelos de los mismos cultivos en años previos. No obstante, si se carece de experiencia o los resultados de años anteriores no fueron suficientemente satisfactorios, la recomendación es buscar consejo de expertos, como agrónomos y especialistas en la materia. Este estado inicial dependerá principalmente de las condiciones fenológicas que presente el cultivo o grupo de cultivos (por ejemplo, será diferente para un cultivo de verano que para un cultivo de invierno).

Es importante recalcar que la selección de un estado inicial adecuado reduce el tiempo de procesamiento del algoritmo, pero no es un aspecto crítico, ya que la técnica empleada por el algoritmo es capaz de alcanzar resultados similares con un periodo inicial inadecuado, a costa, eso sí, de un mayor número de iteraciones y una mayor cantidad de tiempo.

4.7.3. Ejecución del algoritmo de detección del periodo (M7.3)

Una vez seleccionado el periodo inicial, es necesario descargar las imágenes de satélite contenidas en el correspondiente periodo de detección. Por cada paso del algoritmo, se entrenará un nuevo modelo de detección y se evaluará su resultado para asignarle una puntuación. El resultado final de la ejecución del algoritmo será un modelo de detección entrenado con imágenes de un determinado periodo. Cabe recalcar que realizar los procesos que se describen en este apartado deberían realizarse de manera automática, implementados en el correspondiente módulo software.

Podría llegar a ocurrir que la situación inicial cumpliera con el umbral de resultados preestablecido –es decir, con el porcentaje de acierto mínimo– por lo que el algoritmo pararía y devolvería los resultados. No obstante, la situación común es que el algoritmo comience a iterar en busca de una solución mejor. Aquí es donde el algoritmo ofrece distintas variaciones. La propuesta incluye la variante de ascenso más empinado, citada anteriormente. Consiste en explorar todos los periodos vecinos del estado actual, y comenzar a iterar por el mejor de ellos, siempre y cuando sea mejor que el actual.

4.7. BÚSQUEDA DEL PERIODO DE DETECCIÓN DE LOS CULTIVOS (M7)

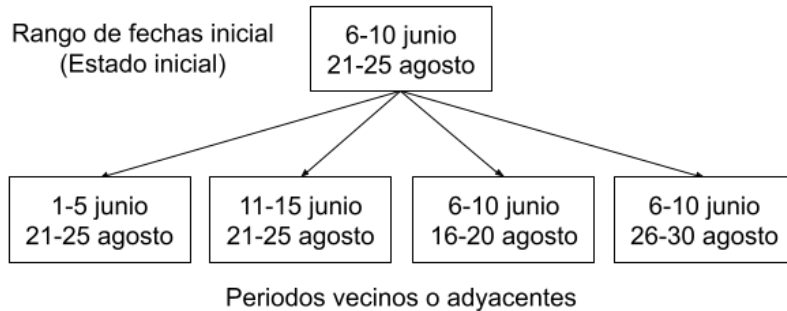


Figura 4.6: Ejemplo de periodos adyacentes para calcular vecindades en el algoritmo de HC.

Para entender correctamente el concepto de periodo adyacente o periodo vecino, se describe el siguiente ejemplo (representado en la Figura 4.6): consideramos periodos temporales de 5 días, donde el rango de fechas inicial abarca desde el periodo 6-10 de junio hasta el periodo 21-25 de agosto. Las combinaciones para poder explorar nuevas soluciones consisten en alargar o reducir el rango de fechas por cada uno de los extremos del periodo. Por tanto, los cuatro vecinos que resultarían serían los rangos de fechas entre el 1-5 de junio y el 21-25 de agosto, entre el 11-15 de junio y el 21-25 de agosto, entre el 6-10 de junio y el 16-20 de agosto, y entre el 6-10 de junio y el 26-31 de agosto.

Es importante aclarar que mientras se exploran los posibles periodos adyacentes, es necesario comprobar que dichos periodos se encuentran dentro de las restricciones establecidas por el sistema, concretamente la denominada como fecha límite real, además de asegurarnos de que el periodo vecino no ha sido explorado con anterioridad.

Una vez calculados los periodos adyacentes, es necesario comenzar con la sucesión de pasos del método, descritas a partir de la Sección 4.3, e incluyendo la descarga de imágenes, su procesamiento, la generación de los ejemplos de entrada, y a continuación calcular los modelos de detección que incluyan la información de dichos periodos, continuando de esta manera con

la búsqueda, y generando el árbol de búsqueda del algoritmo.

Cuando finalice la ejecución del algoritmo, dispondremos de un modelo de detección del cultivo estudiado que cumplirá con los distintos requisitos y criterios de aceptación establecidos previamente, según lo descrito en la Sección 4.7.1.

Es conocido que los algoritmos de HC tienen la posibilidad de quedarse atascados en máximos locales o mesetas, evitando de esta manera encontrar una solución mejor. Hay algunas alternativas para aquellos que no tengan restricciones de tiempo o de espacio de almacenamiento, como los reinicios aleatorios (empezando la ejecución del algoritmo desde un estado inicial diferente para explorar otras regiones del espectro de posibilidades), o aplicar otras variaciones, como la relajación de mejora de alternativas hijas, o como el método de escalada estocástico, presente también en la recopilación de algoritmos de [111].

4.8. Identificación de los cultivos (M8)

Llegados a este punto, ya se disponen de los distintos modelos de detección (al menos uno por cultivo o grupo de cultivos), y se estaría en disposición de realizar las distintas predicciones de manera masiva para todos los recintos de estudio. Cabe recordar que lo ideal en este tipo de métodos es entrenar con datos de años anteriores y predecir los resultados con los datos del año en curso, sin incluir estos últimos en el entrenamiento, ya que estaríamos condicionándolo. Obviamente, esta fase de identificación masiva de cultivos debe automatizarse con los adecuados módulos software, que deberían estar optimizados para manejar el considerable volumen de datos.

En cuanto a la descarga de las imágenes del año en curso para realizar la predicción, por simplicidad se puede realizar de manera simultánea a la descarga del resto de imágenes utilizadas para el entrenamiento, aplicando las respectivas operaciones de preprocesamiento descritas en las Secciones 4.3, 4.4 y 4.5. No obstante, una opción alternativa para ahorrar tiempo y espacio sería descargar las imágenes para la predicción una vez esté ya generado el modelo y establecido el periodo de detección descrito en la sección

anterior, de manera que se descarguen exclusivamente aquellas imágenes que vayan a ser utilizadas.

Para que el software sea efectivo, se propone que la implementación se realice de la siguiente manera: con el listado de zonas geométricas del área de trabajo, se recorre cada una de ellas, y a su vez se itera por los recintos que se encuentran emparejados con la zona. Para cada recinto, se generan las ISMM, teniendo en cuenta el periodo de detección establecido en la sección anterior. A cada una de las ISMM se le aplicará el modelo de aprendizaje ya entrenado, obteniendo un porcentaje de similitud o coincidencia con el cultivo correspondiente.

Es interesante volver a aclarar que todo el proceso de detección se realiza a nivel de píxel, por lo que las distintas predicciones realizadas nos pueden dar información parcial sobre los distintos cultivos dentro de un mismo recinto. Por tanto, para todos los recintos de estudio con necesidad de ser analizados, se les aplicará, de manera masiva y automática, la predicción basada en el modelo generado para su propio cultivo, de manera que obtendremos información del porcentaje de similitud de cada píxel del recinto con el cultivo declarado.

También es posible aplicar modelos de detección de otros cultivos sobre cada píxel, de manera que obtengamos también los porcentajes de similitud con otros cultivos diferentes al declarado. Esto es interesante especialmente en aquellos casos en los que el porcentaje de similitud con el cultivo declarado sea bajo, para identificar cuál sería el cultivo posible que se detecta.

Además de obtener la predicción del cultivo por cada píxel, puede ser muy valioso agrupar esta información para su interpretación. Por ejemplo, realizando la media de las predicciones de todos los píxeles de un recinto, de manera que obtengamos el porcentaje de similitud del recinto completo con respecto al cultivo estudiado. Esto suele ser habitual, por ejemplo, cuando se quiere examinar un recinto en conjunto para determinar algún tipo de ayuda. También puede ser interesante disponer de la desviación típica de las similitudes de los píxeles, para saber si los datos se encuentran agrupados en valores cercanos a la media o, por contra, si existen valores muy dispares dentro del recinto. La necesidad de obtener estos datos debería ser determinada por los usuarios del sistema que quieran obtener informes que

CAPÍTULO 4. MÉTODO DE IDENTIFICACIÓN DE CULTIVOS

cumplan los requerimientos de sus procesos administrativos.

Para facilitar la interpretación de los resultados, es muy habitual en los procesos administrativos que se aplique un “semáforo de umbrales de acierto” [28], con el objetivo de catalogar los distintos recintos. Este semáforo, aplicable tanto a nivel de píxel como a nivel de recinto, debería disponer de dos umbrales: uno que limite la diferencia entre los valores en verde (similitud alta con el cultivo) y en amarillo (similitud media); y otro entre los valores en amarillo y en rojo (similitud baja con el cultivo). Estos umbrales pueden ser ajustados acordes a las necesidades de catalogación del proyecto, proponiendo que el umbral del verde se encuentre, por ejemplo, en torno al 80-90 % de similitud, y el umbral del amarillo entorno al 60 %. Esto facilitará la interpretación de los resultados de manera visual y directa.

Disponiendo de la información a nivel de píxel, sus datos de evolución por la imágenes del periodo temporal, y los niveles medios y desviaciones, es muy útil generar documentación basada en las decisiones tomadas durante la aplicación del método. Así, conforme se vayan aplicando las distintas fases del método, todas las decisiones tomadas podrán servir para justificar los resultados de manera documentada.

Con todo lo anterior, finalmente dispondremos de un sistema de identificación de cultivos a nivel de píxel, genérico en cuanto a la fuente de datos utilizada, personalizable en función de las necesidades del proyecto, y asegurando un porcentaje de acierto, que nos permitirá obtener distintas predicciones sobre los recintos estudiados, y cuya interpretación de los resultados permitirá facilitar la toma de decisiones con respecto a los correspondientes recintos, especialmente en el ámbito administrativo.

4.8. IDENTIFICACIÓN DE LOS CULTIVOS (M8)

Capítulo 5

Evaluación

En este capítulo se expone un desarrollo completo del método propuesto en el Capítulo 4. Con este caso se valida la aplicabilidad del método partiendo de un caso de estudio real, con sus correspondientes requisitos, limitaciones y objetivos. La finalidad es verificar, paso a paso, que la aplicación del método da lugar a un sistema completo operativo para la gestión de trámites administrativos relacionados con la identificación de cultivos de forma masiva a través de imágenes de satélite.

El caso de estudio consiste en un proyecto que resuelve una tarea específica dentro del sistema de información InteliPAC [102], desarrollado de forma conjunta por la Junta de Extremadura y el grupo de investigación Quercus de la Universidad de Extremadura. Su objetivo principal es la identificación de cultivos orientada a la PAC y a otros servicios relacionados, que lleva usándose por diferentes servicios de la Consejería de Agricultura desde el año 2018. Actualmente, el sistema dispone de datos de recintos desde 2017 de más de 300 cultivos distintos, y en el último año 2022 se tienen almacenados 2,3 millones de recintos, que abarcan más de 30.000 km². El sistema InteliPAC, cuyos subproyectos se han desarrollado siguiendo el método propuesto en esta tesis, utiliza las imágenes de la misión Sentinel-2, recomendado por la Comisión Europea [28], debido, entre otras razones, a ser una misión de la Agencia Espacial Europea, que ofrece imágenes de manera periódica con una resolución adecuada para la

5.1. ANÁLISIS DE REQUISITOS Y DISEÑO DEL SISTEMA DE INFORMACIÓN

monitorización de cultivos.

El objetivo concreto del proyecto que se va a describir en este capítulo es disponer de un modelo de detección del cultivo de tabaco, lo más preciso posible, para detectar plantaciones de tabaco ilegal en la comunidad autónoma de Extremadura, en las comarcas de La Vera y parte del Campo Arañuelo, ocupando una región de unos 2.000 km². Las especificaciones de cada fase y los correspondientes desarrollos han sido realizados por un equipo de 4 personas del grupo Quercus, entre las que se encuentra el autor de este trabajo, completando todas las fases descritas en el método, y tomando las decisiones correspondientes para adaptarlo al proyecto.

A continuación, se va a proceder a describir en cada apartado las decisiones de implementación y organización de la información que se han tomado para resolver los objetivos del proyecto siguiendo la metodología propuesta.

5.1. Análisis de requisitos y diseño del sistema de información

5.1.1. Identificación de requisitos y limitaciones

El proyecto en el que se basa este desarrollo tiene una serie de restricciones marcadas por la Junta de Extremadura. En primer lugar, la fecha límite para disponer tanto de los modelos generados como de los propios resultados de la predicción es el 30 de septiembre de 2022, fecha en la que deben gestionar los distintos expedientes administrativos relacionados con las ayudas y sanciones dentro del marco de la PAC. Por tanto, esa fecha condicionará el periodo de análisis de los cultivos.

En cuanto a la capacidad de almacenamiento, no existen restricciones como tal en cuanto al máximo disponible, aunque sí se precisa que se desea minimizar este aspecto para poder reducir costes en un futuro en cuanto a la adquisición de dispositivos de almacenamiento.

Con respecto al umbral de acierto, la Junta de Extremadura hace hincapié en maximizar el porcentaje de acierto del modelo de detección de

tabaco, al ser un cultivo especialmente sensible en la zona por su alta concentración, y la aparición de plantaciones ilegales del mismo. Las parcelas identificadas seguirán un proceso de seguimiento en campo y sancionador. Por tanto, se centran los esfuerzos en conseguir una precisión alta en este modelo para mejorar la eficiencia en la gestión.

5.1.2. Diseño del sistema de información

Para desarrollar las fases del método, incluyendo la descarga de imágenes, su procesamiento y el entrenamiento de modelos, se ha utilizado una única máquina, con las siguientes características:

- Procesador: i-8700K CPU 3,7GHz
- Memoria RAM: 32GB DDR4
- Tarjeta gráfica: NVIDIA GeForce GTX 1080 Ti
- Almacenamiento interno: 512GB SSD
- Almacenamiento externo: 4TB SSD

En el equipo anterior se ha instalado, bajo un sistema operativo Ubuntu, un sistema gestor MariaDB / MySQL para almacenar toda la información del sistema, y los correspondientes directorios donde se almacenarán las imágenes del área de trabajo del proyecto. Esta configuración se ha considerado suficiente para poder almacenar y ejecutar los módulos software en los tiempos requeridos.

5.2. Organización de los datos fuente de recintos

5.2.1. Selección y estandarización de recintos identificables

El proceso de selección de los recintos se ha circunscrito a los recintos del área donde se va a realizar el análisis. Para ello, la información proporcionada por parte de la Junta de Extremadura consiste, por una parte, en

5.2. ORGANIZACIÓN DE LOS DATOS FUENTE DE RECINTOS

el conjunto de recintos que han solicitado ayudas de la PAC en los años 2020 y 2021 en el área de trabajo, y que han sido validados por diferentes procesos (inspecciones de campo, inspecciones visuales vía satélite, etc.). Estos recintos servirán para realizar el entrenamiento del modelo de detección de tabaco, ya que disponen de la catalogación de sus respectivos cultivos. De los recintos se tiene, además del cultivo asignado, el polígono que delimita su forma para poder extraer información de sus píxeles.

Por otra parte, se dispone del conjunto de recintos agrícolas que se desea analizar para el año 2022, también en el mismo área de trabajo. De esta manera, se dispone de información para entrenar el modelo de detección de tabaco (con datos de los años 2020 y 2021), e información que se desea cotejar con el modelo generado (datos de 2022), es decir, de los recintos de ese área que van a ser analizados con el modelo de tabaco.

Respecto a la estandarización de los datos, en este caso se encontró disparidad en los formatos de recepción de la información. Por distintas cuestiones, algunos recintos se han recibido en formato ShapeFile (SHP), mientras que otros se recibieron en formato CSV, por lo que es necesario estandarizarlos en un único formato. El formato elegido para trabajar con esta información es el estándar GeoJSON [113], debido a su sencillez, a la gran cantidad de librerías disponibles para trabajar con él, y a su fácil integración en los distintos sistemas de representación gráfica de mapas. En cuanto al sistema de coordenadas, la propuesta consiste en unificar todos los datos geográficos basándonos en el sistema de referencia de coordenadas geográficas, con valores absolutos de latitud y longitud en grados sexagesimales, ya que, además de ser un formato estandarizado, es compatible con el estándar GeoJSON, y permite su representación gráfica directa sobre una gran cantidad de aplicaciones. Para aplicar las operaciones anteriores se han utilizado las librerías de Python *geojson* [114] y *shapely* [115], respectivamente.

En la Figura 5.1 se muestra la distribución geográfica de los distintos recintos geométricos, proporcionados por la Junta de Extremadura, declarados en la zona de interés en el año 2022 para situarlos geográficamente. Para ese año se disponen de 4.342 recintos de tabaco y 36.583 de otros cultivos. Con respecto a los años utilizados para generar datos de entrenamiento, en 2020 se registraron 4.561 recintos de tabaco, así como 21.062

recintos de otros cultivos –110 tipos de cultivos diferentes–, como maíz, tomate, pimiento, olivar, etc., que servirán como ejemplos negativos. Por otra parte, en 2021 se registraron 4.221 recintos de tabaco y 37.166 de otros cultivos. El aumento significativo de recintos entre 2020 y 2021 se debió a la situación sanitaria vivida a partir de marzo de 2020 a causa de la pandemia mundial del Covid-19.

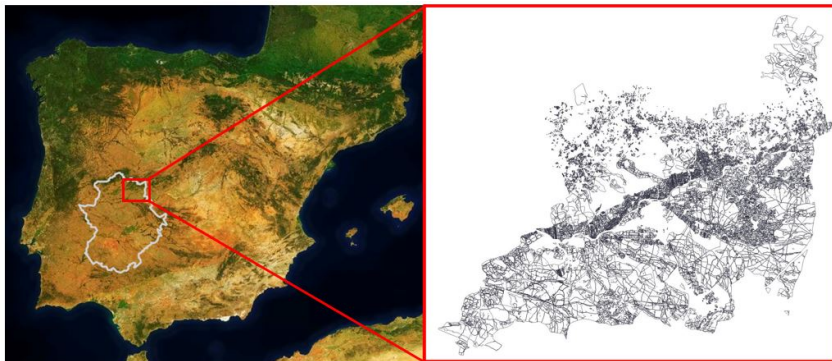


Figura 5.1: Conjunto de recintos declarados en el área de trabajo en el año 2022.

Algunos de los datos de recintos que se proporcionaron por parte de la Junta de Extremadura presentaban errores de formación de polígonos, como por ejemplo el que se aprecia en la Figura 5.2. Estos problemas en ocasiones aparecen debido a que las coordenadas del recinto se dibujan manualmente. Para solucionarlo, se ha hecho uso también de la citada librería *shapely*, de manera que se han obtenido las coordenadas simplificadas, como podemos ver en la imagen (b) de la figura.

Como indica el método, debe unificarse la nomenclatura de los cultivos declarados. En este caso, desde la Junta se pedía detección de recintos con tabaco sin especificar la variedad, por lo que se unificó con el valor “tabaco” a los recintos que habían solicitado la PAC, independientemente de su variedad. El objetivo, como ya se ha comentado, es crear un único modelo de tabaco para detectar su cultivo ilegal en la región delimitada.

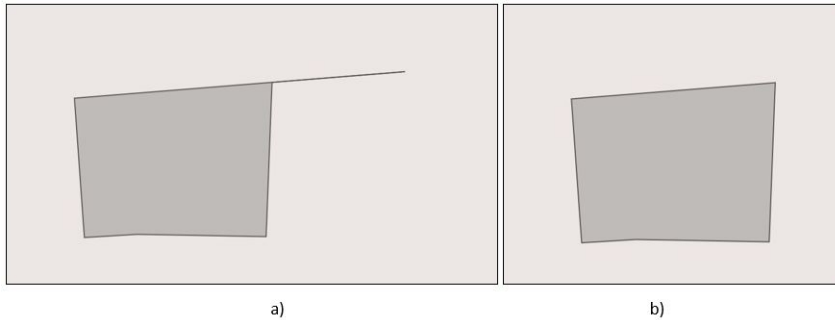


Figura 5.2: Ejemplo de coordenadas de un recinto con alguna malformación.

Hay que recordar que toda la información de los recintos por cada año, incluyendo sus coordenadas geográficas y la información del cultivo y variedad que tienen asociados, se almacenan en la base de datos. En este caso, todas las interacciones del sistema con la base de datos se realizan haciendo uso de la librería *pymysql* [116]. Asimismo, es importante aclarar que el tiempo de ejecución de esta fase no resulta excesivamente elevado, siendo necesario para este caso concreto un tiempo de ejecución aproximado de media hora.

5.2.2. Creación de zonas geométricas

Con respecto a la creación de zonas geométricas, el método indica que se debe partir de la malla de barrido del satélite seleccionado, siendo en este caso los satélites de la misión Sentinel-2. La malla original consiste en un conjunto de cuadrículas de $100 \times 100 \text{ km}^2$ y proyección UTM/WGS84, representada en la Figura 5.3.

Con el objetivo de que el software se ejecute de forma eficiente, en el método se recomienda que el tamaño de las zonas geométricas tenga una proporción adecuada respecto al tamaño original. Al tratarse de unas dimensiones de $100 \times 100 \text{ km}^2$, se decide utilizar alguno de los divisores de 100 como medida de las zonas geométricas. Esta decisión se ha tomado para cubrir completamente la malla original con un número entero de zonas. En

5.2. ORGANIZACIÓN DE LOS DATOS FUENTE DE RECINTOS

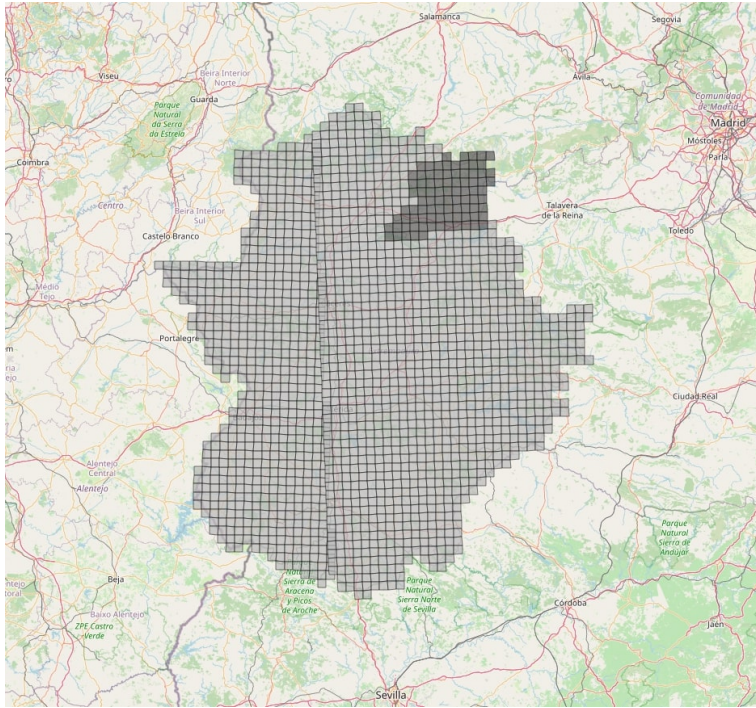


Figura 5.4: Relación de zonas geométricas en la comunidad autónoma de Extremadura, resaltando aquellas correspondientes al área de trabajo.

5.2.3. Emparejamiento de recintos y zonas

Para cada uno de los recintos, se aplica el proceso de emparejamiento con sus respectivas zonas geométricas, utilizando, de nuevo, la librería *shapely*. Como indica el método, para cada recinto se han procesado sus coordenadas, y se ha comprobado con qué zonas existe una intersección, todo ello de forma automática, dando lugar al listado de zonas geométricas asociadas a cada recinto. A modo de ejemplo, para los más de 25.000 recintos de 2020, alrededor de 20.000 recintos están asociados a una única zona, mientras que algo más de 5.000 se encuentran asociados a dos zonas, debido a que se encuentran en el límite de las zonas geométricas. Menos del 1%

de los recintos están asociados a 3 o más zonas. La ejecución automática de esta fase ha supuesto un tiempo de ejecución aproximado de una hora y media.

A pesar de que en el método se establece la peculiaridad de los recintos presentes en zonas geométricas en distintos husos de coordenadas, el área de trabajo de este proyecto se encuentra situado en una región con un solo huso de coordenadas, por lo que no será necesario aplicar ninguna selección adicional.

5.3. Descarga masiva y tratamiento preliminar de datos

5.3.1. Descarga masiva de datos

Habiendo seleccionado la fuente de datos, en este caso Sentinel-2, el método establece la necesidad de disponer de una herramienta que permita la descarga masiva y automática de sus imágenes. Concretamente, el programa Copernicus, que gestiona los datos de las distintas misiones Sentinel, ofrece una plataforma de acceso abierto a los datos [117], donde podemos encontrar una API para interactuar con la plataforma. No obstante, existen también librerías en algunos lenguajes de programación que facilitan dicha interacción, como *SentinelSat* [118], disponible para Python, que permite la recolección masiva y automática de las imágenes a partir de los datos geográficos facilitados, y que ha sido elegida para elaborar el software necesario para realizar la descarga, ya que es una de las librerías más utilizadas en este área.

A la hora de descargar las imágenes, se opta por descargar aquellas imágenes que ya tienen aplicada la corrección atmosférica, que en el caso de Sentinel se corresponden con las imágenes de nivel L2A. El uso de este formato de imágenes proporciona, además, una capa adicional de información que clasifica el terreno en función de diversos factores, como se explica en la Sección 5.4.2. Sin embargo, si fuera necesario o si se quisiera optar por las imágenes de nivel L1C –sin corrección atmosférica aplicada–, se podría

emplear algún procesador de imágenes como el SEN2COR [119], ofrecido por la ESA. Por tanto, se descargan las distintas imágenes necesarias para las fechas establecidas, comprobando en cada momento si la imagen seleccionada está almacenada de manera local para no repetir el proceso.

5.3.2. Recorte y normalización de las imágenes descargadas

Para el recorte de las imágenes originales y la generación de las imágenes de zona, así como para otras operaciones que se comentarán posteriormente, es habitual el uso de programas de escritorio, como puede ser SNAP (Sentinel Application Platform) [35], puesto a disposición también por la ESA. No obstante, al realizar el procesamiento de manera automática, se descarta utilizar directamente este tipo de aplicaciones, por lo que ha sido necesario buscar un componente que permita automatizar el proceso. En este caso se ha utilizado *Snappy* [120], la librería para Python seleccionada, ya que ofrece acceso a todas las funcionalidades y herramientas disponibles en SNAP, pero que se pueden aplicar de manera masiva mediante su implementación en programas o *scripts*.

Es necesario recordar que durante el proceso de recorte de imágenes para generar las imágenes de zona, el método requiere aplicar también la normalización de las distintas resoluciones de las bandas de información, realizado de manera automática. En este caso, se decide optar por la resolución mínima que ofrece Sentinel-2, que es de 10x10 metros para cada píxel en alguna de sus bandas, por lo que se normalizan todas las bandas a esta resolución. El motivo de esta decisión es disponer de la mayor cantidad de información posible: cuanto más reducido es el tamaño del píxel, mayor cantidad de píxeles se pueden representar en un mismo área, y consecuentemente se dispone de un mayor número de ejemplos para el entrenamiento del sistema de aprendizaje. Por tanto, tras ejecutar los procesos anteriores, se dispone de un conjunto de imágenes de zona, cuyas resoluciones de imagen se encuentran ya normalizadas a 10x10 metros por píxel.

El proceso de descarga y recorte de imágenes es uno de los más costosos a nivel temporal. Para el área de trabajo establecido, teniendo en cuenta las 88 zonas geométricas, el tiempo estimado para descargar información es de unas doce horas por cada mes completo de información. Finalmente,

como se describirá en la Sección 5.7, se delimitó el periodo a 6 meses (2 meses por cada uno de los tres años), por lo que el tiempo total empleado en esta fase fue de 72 horas.

5.3.3. Emparejamiento y etiquetado de los datos

Al igual que sucedía en la Sección 5.2.3, es necesario volver a recorrer todos los recintos, pero en esta ocasión con la finalidad de emparejar cada uno de ellos con sus píxeles significativos dentro de las imágenes de zona, obteniendo, para cada uno de los recintos, el listado de píxeles asociados al mismo, que se encuentran íntegramente dentro de sus coordenadas geográficas, tal como indica el método.

Todo este proceso se ha ejecutado de manera automática, empleando el software desarrollado específicamente para este fin, y con una duración aproximada de 24 horas. El resultado de esta fase ha sido doble: para cada recinto se almacena, por una parte, el listado de coordenadas relativas de píxeles; y por otra, las coordenadas geométricas de los píxeles asociados a cada recinto, para poder ser representados de manera gráfica en caso de ser necesario. Todo ello ha sido debidamente clasificado e insertado en la base de datos del sistema de información.

Atendiendo a los datos de recintos identificables expuestos en la Sección 5.2.1, el resultado de aplicar esta fase fue de 3,9 millones de píxeles asociados a los recintos de 2020; 8,79 millones de píxeles en 2021; y 10,83 millones de píxeles en 2022.

5.4. Selección, agrupación y filtrado de datos

5.4.1. Selección de datos

Como establece el método, se va a utilizar la información a nivel de píxel, y en este caso a la resolución normalizada descrita en la fase anterior. Debido a que para este proyecto se necesita el mayor porcentaje de acierto posible para la predicción de tabaco ilegal, se ha optado por utilizar todas

las bandas de información disponibles. En este caso son las 12 bandas ofrecidas por Sentinel-2, de manera que para cada píxel se disponen de 12 valores de información, que serán agrupados por el conjunto de imágenes que comprendan el periodo de análisis.

5.4.2. Eliminación de datos con ruido

Para detectar y eliminar aquellos píxeles con información no válida que puedan producir ruido durante el entrenamiento, el método establece la necesidad de establecer unos umbrales en cada una de las bandas que establezcan si el valor es anómalo o no. Para ello, se ha utilizado la clasificación de escenas facilitada en las imágenes de nivel L2A de Sentinel-2 [45]. Este algoritmo proporciona una capa adicional de datos en las imágenes, que indica, a nivel de píxel, si la información representada está incluida en alguna de las siguientes categorías: nubosidad, con hasta cuatro clases distintas; sombras y sombras de nubes; vegetación; suelo sin vegetación; agua; y nieve.

El algoritmo de clasificación de escenas se aplica sobre la imagen original para determinar la categoría de cada uno de los píxeles. Este algoritmo utiliza un árbol de decisión con diferentes fórmulas, y asigna una categoría concreta a cada píxel. Por tanto, una vez catalogados todos los píxeles, se descartan aquellos que estén catalogados en una categoría distinta a “vegetación”.

5.4.3. Agrupación de los datos

El método propone la creación de un conjunto de periodos temporales fijos, con una duración mínima, que asegure disponer de al menos una imagen por periodo. Las imágenes ofrecidas por la misión Sentinel-2 ofrecen una periodicidad mínima de imágenes cada 5 días, que en algunas regiones puede ser de entre 2 y 3 días por el solapamiento de las órbitas de sus dos satélites. Por tanto, al aplicar el método a las imágenes de Sentinel-2, se decide crear periodos temporales cada 5 días, que es el tiempo mínimo en el que los satélites disponen de una imagen nueva.

A modo de ejemplo, podríamos tener los periodos 1-5 de enero (ambas fechas inclusive), 6-10 de enero, 11-15 de enero, y así sucesivamente, adaptando el número de días de cada mes en el último periodo: en los meses con 30 días iría desde el 26 al 30, y en los meses con 31 días se le incluiría esa fecha adicional. Algo similar habría que aplicar en el mes de febrero, cuyo último periodo contaría desde el día 26 hasta el 28 o 29, dependiendo del año.

Así pues, para este proyecto se han ejecutado *scripts*, desarrollados de manera específica, que calculan los valores medios de cada píxel para cada uno de los periodos establecidos, que serán los valores que se almacenan en la base de datos.

5.4.4. Interpolación de los datos

Con el objetivo de que todos los píxeles dispongan de la misma cantidad de datos, tal y como indica el método, se deben interpolar aquellos datos desconocidos para completar las series temporales de datos. En este proceso es necesario hacer la interpolación en los casos en los que se eliminan datos con ruido (acorde a la Sección 5.4.2), para que todos los píxeles dispongan de una serie temporal con la misma cantidad de datos.

Para hacer esta tarea, se hace uso de la librería para Python *numpy* [121], y más concretamente, se aplica una interpolación lineal, mediante el método “*interp*”. Esta función, aplicada sobre la serie temporal de cada píxel, ofrece como resultado una serie temporal completa, sin valores desconocidos, de manera que, para todos los píxeles, se disponen de series temporales homogéneas en cuanto al número de valores.

5.4.5. Reducción de anomalías e interferencias

Como se ha comentado en el método, es posible que algunos valores anómalos no se hayan descartado correctamente en la Sección 5.4.2, debido a problemas con los sensores o errores en los algoritmos de detección de datos con ruido, por lo que es interesante aplicar un nuevo filtro para suavizar posibles valores atípicos. Por experiencia con el uso de datos de imágenes

de Sentinel-2, este caso es relativamente habitual. Por ello se decide aplicar en este proyecto un filtro de Hampel [52] para detectar estos valores y suavizarlos.

Para saber si una muestra es atípica o anómala (en este caso se entiende como muestra a un valor determinado dentro de la serie temporal de un píxel), durante la aplicación del filtro de Hampel se calculan la media y la desviación estándar de sus K vecinos (se aplica considerando hasta 7 vecinos), y se comprueba si el valor de la misma sobrepasa una cierta cantidad de desvíos, declarando en tal caso el valor como atípico, y reemplazándolo por su valor medio. Por tanto, no se modifica la cantidad de elementos de cada serie temporal, sino que únicamente se suavizan aquellos que se consideren como atípicos, tal y como se ve en la Figura 5.5. Todo este proceso se ha aplicado de manera masiva y automática para todas las series temporales, ejecutando distintos *scripts* implementados en Python, apoyados en la librería pandas [122].

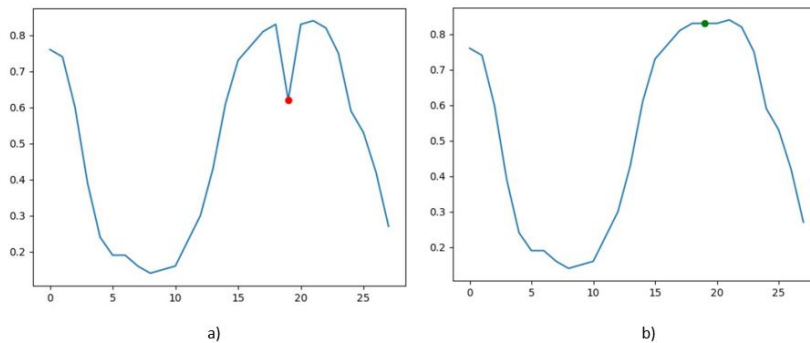


Figura 5.5: Ejemplo de funcionamiento del filtro de Hampel sobre una serie temporal.

Es importante aclarar que, a nivel software, todas las operaciones descritas en la Sección 5.4 se realizan de manera conjunta. Para el volumen de datos con el que se ha trabajado, el tiempo de ejecución del proceso explicado en esta sección es de unas 12 horas por cada mes de información.

Como el periodo de estudio fue de 6 meses (2 meses por cada uno de los tres años), ha llevado un total de 72 horas.

5.5. Configuración de los ejemplos de entrada para el aprendizaje

5.5.1. Generación de imágenes sintéticas

En este punto ya se dispone de la información de cada píxel agrupada por periodos, habiendo filtrado sus posibles valores anómalos, conformando una serie temporal por cada píxel y banda de información, y cuyos valores son los valores medios dentro de cada periodo y banda.

Como propone el método, para cada uno de los píxeles, ha de generarse el ejemplo de entrada para el sistema de información. Se han combinado las series temporales de sus bandas de información, conformando una única matriz por píxel, tal y como se explica en la Sección 4.5.1, dando lugar a las ISMM, y recordando que se dispone de una ISMM por cada píxel.

A modo de ejemplo, en la Figura 5.6 se puede visualizar una ISMM generada a partir de los datos de un píxel concreto. El eje de abscisas representa los valores de todas las bandas de información del píxel en un periodo temporal concreto, y el eje de ordenadas representa un conjunto de periodos temporales, con el correspondiente rango de valores asignado a cada color. Como recordatorio, considerando la ISMM como una matriz de valores, el valor de la posición $[0,0]$ se corresponde con el valor de la primera banda (en este caso B1) en el primer periodo.

5.5. CONFIGURACIÓN DE LOS EJEMPLOS DE ENTRADA PARA EL APRENDIZAJE

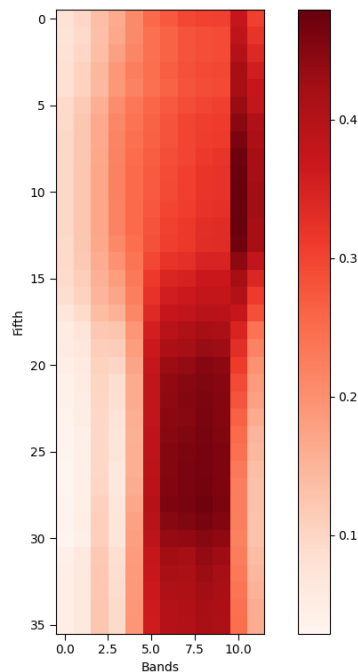


Figura 5.6: Ejemplo de ISMM generada.

El proceso completo de generación de todas las ISMM conlleva menos de una hora de ejecución, ya que todos los datos están ya procesados y agrupados.

5.5.2. Especificación de los modelos para el entrenamiento

Esta ejecución del método se centra en crear un único modelo de detección de tabaco lo más preciso posible. Por tanto, para generar el citado modelo, no será necesario ningún tipo de agrupamiento concreto, ya que será suficiente con incluir como ejemplos positivos los datos de todos los recintos de tabaco disponibles, independientemente de la variedad.

5.6. Diseño del sistema de aprendizaje

5.6.1. Selección y configuración del sistema de aprendizaje

Como propone el método que se presenta en esta tesis, se ha decidido utilizar redes neuronales convolucionales para generar los modelos de aprendizaje. En este caso concreto, el diseño de la CNN parte de la entrada bidimensional de la ISMM, a la que se le aplican dos convoluciones para extraer características relevantes de la imagen, una incrementando los filtros y otra disminuyéndolos, creando distintos tamaños en las capas internas, para así encauzar el entrenamiento para discriminar mejor las características adecuadas.

En cuanto a los parámetros, y siguiendo las recomendaciones de [123], entre cada convolución se aplican las técnicas de *Batch-Normalization* y *Dropout* para conseguir repartir los datos dentro de la red y realizar un entrenamiento más homogéneo. Finalmente, se aplica una capa de aplanamiento, seguida de la capa de salida con un único nodo, que será el que indique el resultado de la predicción. En cuanto a la activación de las capas, se aplica una función de activación de unidad lineal rectificadora (ReLU) para las capas convolucionales, y la función *Softmax* en la capa densa.

Durante todo el proceso se utiliza la técnica de parada anticipada (*early-stopping*), que monitorea el progreso de los valores de pérdida para evitar el sobreentrenamiento de la red, deteniendo el entrenamiento cuando la red detecta que el progreso del entrenamiento no presenta mejoras durante varias iteraciones, e incluso cuando detecta que el progreso va a peor.

5.6.2. Diseño del entrenamiento de los modelos

Tal y como establece el método, si se desea maximizar la precisión de un modelo concreto, se recomienda utilizar el máximo número de datos posible. En este caso, para maximizar el porcentaje de acierto del modelo de tabaco, se incluyen como ejemplos negativos datos de todos los cultivos que no sean tabaco, recordando que se entrena con datos de 2020 y 2021 en la región de interés. En total, se dispone de casi 1,2M de píxeles asociados a cultivos de tabaco. Para balancear el entrenamiento y que no quede

5.7. BÚSQUEDA DEL PERIODO DE DETECCIÓN DE LOS CULTIVOS

descompensado, se decide utilizar el mismo número de ejemplos negativos que positivos, resultando también en 1,2M de píxeles de otros cultivos. Es necesario recordar que, para cada uno de los píxeles, tanto asociados al tabaco como a otros cultivos, se ha generado una ISMM, que sirven como ejemplos de entrada para el entrenamiento, disponiendo en total de 2,38M de ISMM.

Estos datos se dividen, a su vez, en los grupos de entrenamiento (70 %), validación (15 %) y test (15 %), de manera que obtenemos la siguiente distribución de datos de entrada al sistema de aprendizaje: 850K píxeles de ejemplos positivos para el entrenamiento, 170K píxeles para la validación, y 170K para el test. Consecuentemente, los ejemplos negativos tendrán la misma distribución.

5.7. Búsqueda del periodo de detección de los cultivos

A la hora de establecer el periodo de detección adecuado, se realizó la ejecución completa para el modelo de tabaco utilizando los datos de 2020 y 2021. Los resultados con el periodo obtenido fueron significativamente muy buenos, y fue validado por ingenieros agrícolas. Una vez establecido el periodo, ya se dispone de manera simultánea del modelo de aprendizaje entrenado

5.7.1. Establecer función heurística y de parada

Se ha seleccionado la métrica F-Score propuesta por el método para comparar los distintos modelos y establecer el periodo de detección adecuado, debido a las ventajas comentadas en la Sección 4.7.1. En este punto también se debe establecer un umbral de acierto. No obstante, ya que el objetivo del proyecto es maximizar el porcentaje de acierto del modelo de tabaco, no se impone ningún umbral de acierto mínimo, sino que se deja finalizar el algoritmo completo para obtener la mejor solución posible.

5.7.2. Definición del estado inicial

Basado en las opiniones de expertos agrónomos involucrados en el proyecto, se decide utilizar como estado inicial aquellos periodos localizados entre el 16 de julio y el 31 de julio, rango de fechas en el cual se puede observar un patrón característico de crecimiento de las plantaciones de tabaco.

5.7.3. Ejecución del algoritmo de detección del periodo

La traza de ejecución del algoritmo se presenta en la Figura 5.7, mostrando el proceso de exploración y sus correspondientes puntuaciones. Comenzando desde el rango de periodos inicial (entre el 16 y el 31 de julio), la primera puntuación calculada resulta en 0,94765. Después de valorar el estado inicial, el algoritmo de HC explora sus vecinos, tal y como se explica en la Sección 4.7.3. A modo de recordatorio, explorar los vecinos consiste en generar los periodos adyacentes, descargar las imágenes relacionadas a dichos periodos, y calcular las puntuaciones para los nuevos rangos de periodos. Por tanto, tras la primera iteración, se muestran los cuatro nuevos rangos de fechas con sus correspondientes puntuaciones, donde es necesario comprobar cuál es la más alta y verificar si es mejor que la puntuación del nodo anterior o no.

Las iteraciones y la exploración de vecinos continúa hasta que se alcanza un estado donde ninguno de sus vecinos obtiene mejor puntuación. Es importante aclarar que no todas las exploraciones resultan en cuatro vecinos, debido a que dichos rangos pueden haber sido explorados previamente.

El algoritmo finalizó después de 10 iteraciones, cuando la puntuación no mejoraba la mejor encontrada previamente, resultando el periodo final el comprendido entre el 1 de julio y el 31 de agosto, con una puntuación F-Score de 0,97028, que ofrece una mejora de casi 0,025 puntos (2,4 %) frente a la puntuación original. Como se ha mencionado anteriormente, el algoritmo HC no explora todas las combinaciones, teniendo en cuenta que el método intenta explorar las menos posibles. En este caso, han sido necesarias 10 iteraciones, con 29 combinaciones de periodos exploradas. Para realizar la

5.7. BÚSQUEDA DEL PERIODO DE DETECCIÓN DE LOS CULTIVOS

ejecución completa del algoritmo se empleó alrededor de 25 horas, debido a la necesidad de entrenar distintos modelos de aprendizaje.

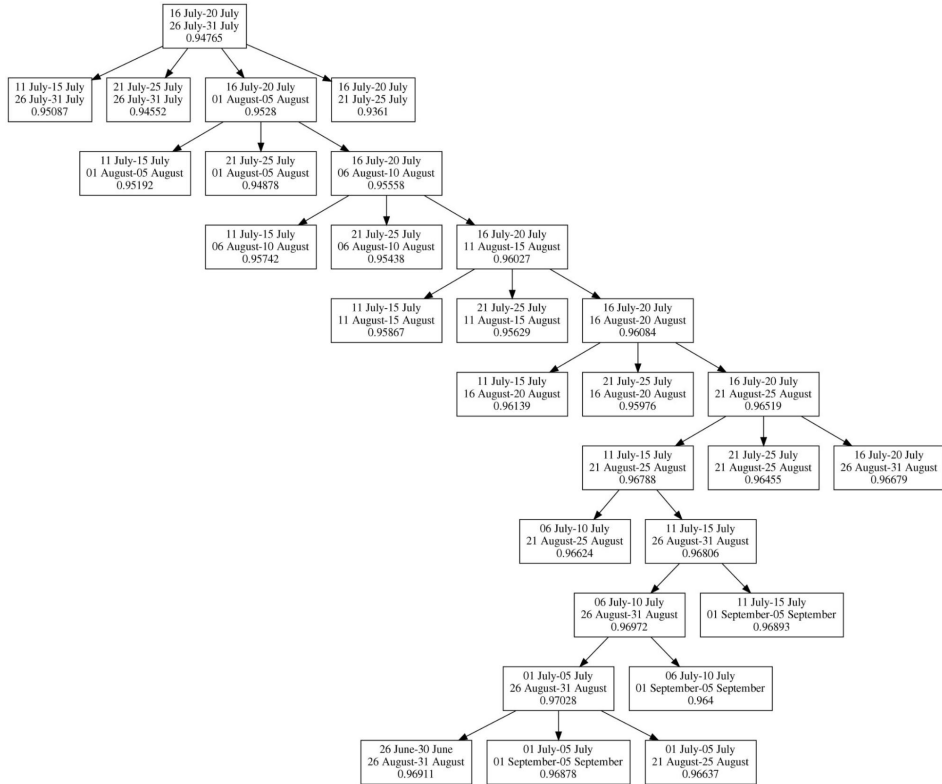


Figura 5.7: Traza de ejecución del algoritmo HC para obtener el periodo adecuado de detección del tabaco.

Por tanto, el modelo entrenado para el periodo seleccionado se estableció como el modelo definitivo de detección de tabaco con los datos de 2020 y 2021, ofreciendo unos resultados muy precisos, con un 0,103 de valor de pérdida (*loss*), un 97,097% de acierto (*accuracy*), y un F-Score de 0,971. El tiempo de entrenamiento del modelo fue de aproximadamente 40 minutos.

5.8. Identificación de los cultivos

Como se indica en el método, en este punto ya se dispone del modelo de detección de tabaco entrenado y preparado para realizar la predicción sobre los datos de 2022. En este caso, se cumplen las indicaciones del método de haber entrenado con datos de años anteriores (2020 y 2021), y predecir con datos de un año posterior (2022).

Con las imágenes de 2022 descargadas y los datos de los recintos procesados y preparados para realizar la predicción, se ha aplicado el modelo sobre todos los píxeles de los recintos de 2022, y más concretamente sobre las ISMM de cada uno de ellos, con el objetivo de detectar posible tabaco plantado de manera ilegal en cualquier recinto. El tiempo total de la predicción sobre cada una de las ISMM fue de unas ocho horas.

En este caso, aunque la predicción se realiza a nivel de píxel y se obtiene un porcentaje de similitud con el modelo de tabaco por cada píxel, se ha realizado la media de todos los píxeles de cada recinto, obteniendo el porcentaje de similitud medio del recinto, así como la desviación típica obtenida al realizar la media. Para interpretar estos resultados, se ha establecido un “semáforo de umbrales de acierto”, recomendado por el método, donde se ha establecido el umbral del verde en un 90 %, de manera que aquellos recintos cuya predicción media de tabaco supere el 90 % serán considerados como posibles plantaciones de tabaco ilegal.

Analizando todos los recintos que no declararon tabaco en 2022, se detectaron 97 de ellos con una similitud media con el modelo de tabaco por encima del 90 %, que podrían llegar a ser consideradas como plantaciones de tabaco ilegal. Con los datos recopilados en el sistema, se generaron de manera semiautomática los 97 informes sobre cada uno de los recintos, basados en las distintas decisiones tomadas durante la aplicación del método, y que justifican las predicciones obtenidas.

5.9. Conclusiones de la evaluación

En este capítulo se ha expuesto el desarrollo de un sistema completo de detección del cultivo de tabaco, basado en el método propuesto en esta tesis doctoral. Para ello, se han seguido cada una de las fases del método, tomando las decisiones oportunas para adaptar el método a los objetivos de este proyecto concreto.

Tal y como se puede observar en la sección anterior, aplicar el método a un caso específico conlleva establecer las pautas de actuación en cada fase y ofrece una guía para la toma de decisiones en el diseño del proyecto. Para este caso, en todas las fases se han podido seguir las recomendaciones del método, cumpliendo tanto con las restricciones del proyecto como con sus objetivos. El tiempo total de ejecución del método ha sido de unos nueve días completos, sin tener en cuenta el tiempo necesario para el desarrollo del software y la toma de decisiones.

Así mismo, la aplicación del método le ha servido al equipo de trabajo para generar documentación basada en las distintas decisiones tomadas en su desarrollo, y se ha entregado a la Junta de Extremadura para justificar las decisiones tomadas en cada fase.

Capítulo 6

Conclusiones y trabajos futuros

Esta tesis propone una metodología completa para desarrollar e implementar un sistema de detección de cultivos, a través de una serie de indicaciones y recomendaciones, útil para la identificación de cultivos de manera masiva y automática. Este capítulo resume las distintas aportaciones al estado del arte y muestra las conclusiones extraídas tras la finalización de este trabajo. Así mismo, también se exponen las líneas de trabajo futuras que sería interesante abordar en próximas investigaciones.

6.1. Conclusiones

La principal contribución de este trabajo es un **método que describe los procedimientos y técnicas para desarrollar un sistema de análisis de cultivos usando imágenes satelitales**. El ámbito de aplicación es óptimo para grandes extensiones de terreno, y con un número elevado de recintos, por lo que se denomina de tratamiento masivo. Además, se basa en el análisis múltiple de imágenes de cada recinto para considerar su evolución en el tiempo, por lo que también se le denomina multitemporal. Propone, partiendo de las delimitaciones de un conjunto de recintos de cultivos, descargar de manera masiva y automática las imágenes de satélite

que los cubren, y procesar esta información para su posterior uso en un sistema de aprendizaje máquina. Este procesamiento incluye la generación de imágenes con un tamaño más reducido (imágenes de zona), los procesos de extracción de información de las imágenes, y la selección, agrupación y filtrado de los datos. En cada fase, el método incluye recomendaciones para adaptar las decisiones de implementación a distintas situaciones en función de los requisitos del proyecto, las fuentes de datos disponibles, la precisión en la identificación y los tiempos necesarios de respuesta. La idea general de esta contribución fue publicada en el *European Journal of Remote Sensing* [102].

La segunda aportación principal del trabajo es el uso de un nuevo enfoque en la construcción de los ejemplos de aprendizaje mediante la **generación de imágenes sintéticas para detección de cultivos**. Se ha propuesto un sistema que, partiendo de la información de un píxel de un recinto de cultivo a lo largo del tiempo, genera imágenes sintéticas multi-temporales y multispectrales (ISMM), que serán empleadas como ejemplos de entrenamiento en un sistema de aprendizaje, con el objetivo de extraer sus características más significativas. Esta aportación fue publicada en *Remote Sensing* [103].

Otra aportación principal es la propuesta de un **método específico para determinar el periodo de detección adecuado de un cultivo**. Seleccionar el periodo en el que se analizan las imágenes de satélite en teledetección de cultivos resulta clave para obtener unos buenos resultados. Escoger meses con información poco diferenciadora, o dejar fuera semanas que caracterizan a un cultivo puede conllevar a una pérdida significativa en el nivel de acierto. Con la metodología propuesta, es posible determinar un periodo de detección de un cultivo determinado cuyo modelo de aprendizaje cumpla con un umbral de acierto mínimo preestablecido, de manera que se minimice el número de imágenes necesarias maximizando el acierto. Esta propuesta fue publicada en el *International Journal of Remote Sensing* [110].

El método de esta tesis ha sido aplicado en la implementación del proyecto InteliPAC, un sistema completo de monitorización de cultivos implantado en la Junta de Extremadura en un entorno real, que lleva usándose de manera eficiente desde 2018 por diferentes Servicios en tareas relaciona-

das con la detección de cultivos, como la verificación de usos SIGPAC, la validación de cultivos para aprobar las ayudas ofrecidas por la PAC, o la detección de plantaciones de tabaco ilegal.

De acuerdo con lo anterior, se concluye que se han cumplido los objetivos propuestos al comienzo de este trabajo, proponiendo una solución integral para la identificación masiva de cultivos mediante múltiples imágenes de satélite, y que ya dispone de distintas aplicaciones reales que se encuentran en uso actualmente por parte de la administración pública.

6.2. Trabajos futuros

Una vez presentadas las conclusiones obtenidas tras la realización de esta tesis doctoral, se muestran algunos problemas que no han sido completamente resueltos, y que se proponen como posibles líneas de trabajo futuras:

- **Incluir en la metodología el procedimiento completo para integrar varias fuentes heterogéneas de datos.** A pesar de que en el método se ha indicado la posibilidad de combinar distintas fuentes de datos de imágenes de satélite (desde diferentes misiones de satélites, y con diferentes tipos de sensores), no se ha podido profundizar en una propuesta que establezca de manera detallada cómo debe realizarse este proceso. Existen en la literatura diversos trabajos que ya usan distintas fuentes de datos, tal y como se mostró en la Sección 2.1. por lo que sería interesante analizar con detalle cómo realizan la integración de las distintas fuentes y establecer, de manera genérica, una fase específica detallada y funcional para llevarlo a cabo, que se incorporaría como una ampliación en la fase M3 del método propuesto en este trabajo.

- **Establecer una estructura combinada de sistemas de aprendizaje.** La metodología propuesta incluye el uso de CNNs, debido a su probada eficiencia en el tratamiento de imágenes en la detección de cultivos, especialmente con imágenes sintéticas como las planteadas en el trabajo. No obstante, sería interesante estudiar cómo otros sistemas de aprendizaje podrían obtener resultados según las características de los proyectos concretos.

Referencias

- [1] IGN, “Teledetección.” <https://www.ign.es/web/resources/docs/IGNCnig/OBS-Teledeteccion.pdf>.
- [2] ESA, “¿qué es la teledetección?.” https://www.esa.int/SPECIALS/Eduspace_ES/SEM01U3FEXF_0.html.
- [3] NASA History Division, “Sputnik and the dawn of the space age.” <https://history.nasa.gov/sputnik.html>.
- [4] E. D. A. Inc., “Teledetección satelital: Tipos, usos y aplicaciones.” <https://eos.com/es/blog/teledeteccion/>, 2021.
- [5] IGN, “Plan nacional de teledetección. aplicaciones.” <https://pnt.ign.es/aplicaciones>.
- [6] IGN, “Plan nacional de teledetección.” <https://pnt.ign.es>.
- [7] NASA, “Landsat 1.” <https://landsat.gsfc.nasa.gov/satellites/landsat-1/>.
- [8] J. L. Engvall, J. D. Tubbs, and Q. A. Holmes, “Pattern recognition of landsat data based upon temporal trend analysis,” *Remote Sensing of Environment*, vol. 6, no. 4, pp. 303–314, 1977.
- [9] D. Alonso, “50 aplicaciones de la teledetección.” <https://mappinggis.com/2018/10/50-aplicaciones-de-la-teledeteccion/>.

-
- [10] Instituto Andaluz de Investigación y Formación agraria, pesquera, alimentaria y de la producción ecológica, “Teledetección para la agricultura.” <https://www.juntadeandalucia.es/agriculturaypesca/ifapa/servifapa/registro-servifapa/bdac6055-21a9-4dd7-8cba-06a3e9d67873/download>.
- [11] Copernicus, “Sentinel-2.” <https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2>.
- [12] Copernicus, “Copernicus web page.” <https://www.copernicus.eu/es>.
- [13] M. Račić, K. Oštir, D. Peressutti, A. Zupanc, and L. C. Zajc, “Application of temporal convolutional neural network for the classification of crops on sentinel-2 time series,” vol. 43, pp. 1337–1342, International Society for Photogrammetry and Remote Sensing, 2020.
- [14] Y. Zhai, N. Wang, L. Zhang, L. Hao, and C. Hao, “Automatic crop classification in northeastern china by improved nonlinear dimensionality reduction for satellite image time series,” *Remote Sensing*, vol. 12, no. 17, p. 2726, 2020.
- [15] S. Siachalou, G. Mallinis, and M. Tsakiri-Strati, “A hidden markov models approach for crop classification: Linking crop phenology to time series of multi-sensor remote sensing data,” *Remote Sensing*, vol. 7, no. 4, pp. 3633–3650, 2015.
- [16] L. García-Torres, J. J. Caballero-Novella, D. Gómez-Candón, and J. M. Peña, “Census parcels cropping system classification from multitemporal remote imagery: A proposed universal methodology,” *PLoS ONE*, vol. 10, no. 2, 2015.
- [17] Q. Hu, D. Sulla-Menashe, B. Xu, H. Yin, H. Tang, P. Yang, and W. Wu, “A phenology-based spectral and temporal feature selection method for crop mapping from satellite time series,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 80, pp. 218–229, 2019.

- [18] K. Jia, B. Wu, and Q. Li, “Crop classification using hj satellite multispectral data in the north china plain,” *Journal of Applied Remote Sensing*, vol. 7, no. 1, 2013.
- [19] X. Zhang, M. Zhang, Y. Zheng, and B. Wu, “Crop mapping using proba-v time series data at the yucheng and hongxing farm in china,” *Remote Sensing*, vol. 8, no. 11, p. 915, 2016.
- [20] S. Siachalou, G. Mallinis, and M. Tsakiri-Strati, “Analysis of time-series spectral index data to enhance crop identification over a mediterranean rural landscape,” *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 9, pp. 1508–1512, 2017.
- [21] Z. Li, G. Chen, and T. Zhang, “Temporal attention networks for multitemporal multisensor crop classification,” *IEEE Access*, vol. 7, pp. 134677–134690, 2019.
- [22] G. Ghazaryan, O. Dubovyk, F. Löw, M. Lavreniuk, A. Kolotii, J. Schellberg, and N. Kussul, “A rule-based approach for crop identification using multi-temporal and multi-sensor phenological metrics,” *European Journal of Remote Sensing*, vol. 51, no. 1, pp. 511–524, 2018.
- [23] H. Snevajs, K. Charvat, V. Onckelet, J. Kvapil, F. Zadrazil, H. Kubickova, J. Seidlova, and I. Bartlova, “Crop detection using time series of sentinel-2 and sentinel-1 and existing land parcel information systems,” *Remote Sensing*, vol. 14, no. 5, p. 1095, 2022.
- [24] N. Addimando, M. Engel, F. Schwarz, and M. Batič, “A deep learning approach for crop type mapping based on combined time series of satellite and weather data,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 43, no. B3, pp. 1301–1308, 2022.
- [25] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, “Machine learning in geosciences and remote sensing,” *Geoscience Frontiers*, vol. 7, no. 1, p. 3–10, 2016.

-
- [26] K. Schulz, R. Hänsch, and U. Sörgel, “Machine learning methods for remote sensing applications: an overview,” *Earth Resources and Environmental Remote Sensing / GIS Applications IX*, 10 2018.
- [27] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, “Machine learning applications for precision agriculture: A comprehensive review,” *IEEE Access*, vol. 9, pp. 4843–4873, 2021.
- [28] W. Devos, G. Lemoine, P. Milenov, D. Fasbender, P. Loudjani, C. Wirnhardt, A. Sima, and P. Griffiths, “Second discussion document on the introduction of monitoring to substitute otsc: rules for processing application in 2018-2019,” *Publications Office of the European Union, Luxembourg*, 2018.
- [29] F. Nègre, “La financiación de la pac.” <https://www.europarl.europa.eu/factsheets/es/sheet/106/la-financiacion-de-la-pac>, 2022.
- [30] European Comission, “The common agricultural policy at a glance.” https://agriculture.ec.europa.eu/common-agricultural-policy/cap-overview/cap-glance_en.
- [31] Q. Zheng, W. Huang, X. Cui, Y. Shi, and L. Liu, “New spectral index for detecting wheat yellow rust using sentinel-2 multispectral imagery,” *Sensors*, vol. 18, no. 3, p. 868, 2018.
- [32] H. Zhang, J. Kang, X. Xu, and L. Zhang, “Accessing the temporal and spectral features in crop type mapping using multi-temporal sentinel-2 imagery: A case study of yi’an county, heilongjiang province, china,” *Computers and Electronics in Agriculture*, vol. 176, p. 105618, 2020.
- [33] F. Pucha, “Lista de índices espectrales en sentinel 2 y landsat.” <https://acolita.com/lista-de-indices-espectrales-en-sentinel-2-y-landsat/>, 2019.
- [34] J. Xue and B. Su, “Significant remote sensing vegetation indices: A review of developments and applications,” *Journal of Sensors*, vol. 2017, pp. 1–17, 2017.

- [35] ESA, “Snap.” <https://earth.esa.int/eogateway/tools/snap>.
- [36] P. Hao, L. Wang, Y. Zhan, C. Wang, Z. Niu, and M. Wu, “Crop classification using crop knowledge of the previous-year: Case study in southwest kansas, usa,” *European Journal of Remote Sensing*, vol. 49, pp. 1061–1077, 2016.
- [37] A. Htitiou, A. Boudhar, Y. Lebrini, R. Hadria, H. Lionboui, and T. Benabdelouahab, “A comparative analysis of different phenological information retrieved from sentinel-2 time series images to improve crop classification: a machine learning approach,” *Geocarto International*, vol. 37, no. 5, pp. 1426–1449, 2020.
- [38] Z. Li, G. Chen, and T. Zhang, “A cnn-transformer hybrid approach for crop classification using multitemporal multisensor images,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 847–858, 2020.
- [39] U. Rauf, W. S. Qureshi, H. Jabbar, A. Zeb, A. Mirza, E. Alanazi, U. S. Khan, and N. Rashid, “A new method for pixel classification for rice variety identification using spectral and time series data from sentinel-2 satellite imagery,” *Computers and Electronics in Agriculture*, vol. 193, p. 106731, 2022.
- [40] L. Yin, N. You, G. Zhang, J. Huang, and J. Dong, “Optimizing feature selection of individual crop types for improved crop mapping,” *Remote Sensing*, vol. 12, no. 1, p. 162, 2020.
- [41] L. Viskovic, I. N. Kosovic, and T. Mastelic, “Crop classification using multi-spectral and multitemporal satellite imagery with machine learning,” Institute of Electrical and Electronics Engineers Inc., 2019.
- [42] J. Yao, J. Wu, C. Xiao, Z. Zhang, and J. Li, “The classification method study of crops remote sensing with deep learning, machine learning, and google earth engine,” *Remote Sensing*, vol. 14, no. 12, p. 2758, 2022.

-
- [43] Y. Li, Q. Shi, L. He, R. Cai, L. Meng, J. Li, and A. Plaza, “Fusing sentinel-2 and landsat-8 surface reflectance data via pixel-wise local normalization,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 7359–7374, 2022.
- [44] J. Fan, X. Zhang, C. Zhao, Z. Qin, M. D. Vroey, and P. Defourny, “Evaluation of crop type classification with different high resolution satellite data sources,” *Remote Sensing*, vol. 13, no. 5, pp. 1–17, 2021.
- [45] Copernicus, “Level-2a algorithm overview.” <https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-2-msi/level-2a/algorithm>.
- [46] EO Research, “Cloud masks at your service.” <https://medium.com/sentinel-hub/cloud-masks-at-your-service-6e5b2cb2ce8a>, 2020.
- [47] N. You and J. Dong, “Examining earliest identifiable timing of crops using all available sentinel 1/2 imagery and google earth engine,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 161, pp. 109–123, 2020.
- [48] S.-H. Kim and J. Eun, “Development of cloud and shadow detection algorithm for periodic composite of sentinel-2a/b satellite images,” *Korean Journal of Remote Sensing*, vol. 37, no. 5, pp. 989–998, 2021.
- [49] J. D. Braaten, W. B. Cohen, and Z. Yang, “Automated cloud and cloud shadow identification in landsat mss imagery for temperate ecosystems,” *Remote Sensing of Environment*, vol. 169, pp. 128–138, 2015.
- [50] M. Czerkawski, P. Upadhyay, C. Davison, A. Werkmeister, J. Cardona, R. Atkinson, C. Michie, I. Andonovic, M. Macdonald, and C. Tachtatzis, “Deep internal learning for inpainting of cloud-affected regions in satellite imagery,” *Remote Sensing*, vol. 14, no. 6, p. 1342, 2022.
- [51] A. Hollstein, K. Segl, L. Guanter, M. Brell, and M. Enesco, “Ready-to-use methods for the detection of clouds, cirrus, snow, shadow,

- water and clear sky pixels in sentinel-2 msi images,” *Remote Sensing*, vol. 8, no. 8, p. 666, 2016.
- [52] R. K. Pearson, Y. Neuvo, J. Astola, and M. Gabbouj, “Generalized hampel filters,” *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 87, 2016.
- [53] I. Arslan, M. Topakci, and N. Demir, “Monitoring maize growth and calculating plant heights with synthetic aperture radar (sar) and optical satellite images,” *Agriculture*, vol. 12, no. 6, p. 800, 2022.
- [54] B. Tuvdendorj, H. Zeng, B. Wu, A. Elnashar, M. Zhang, F. Tian, M. Nabil, L. Nanzad, A. Bulkhbai, and N. Natsagdorj, “Performance and the optimal integration of sentinel-1/2 time-series features for crop classification in northern mongolia,” *Remote Sensing*, vol. 14, no. 8, p. 1830, 2022.
- [55] F. Kordi and H. Yousefi, “Crop classification based on phenology information by using time series of optical and synthetic-aperture radar images,” *Remote Sensing Applications: Society and Environment*, vol. 27, p. 100812, 2022.
- [56] S. Valero, C. Pelletier, and M. Bertolino, “Patch-based reconstruction of high resolution satellite image time series with missing values using spatial, spectral and temporal similarities,” *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2016.
- [57] F. Mouret, M. Albughdadi, S. Duthoit, D. Kouamé, G. Rieu, and J. Y. Tourneret, “Reconstruction of sentinel-2 derived time series using robust gaussian mixture models — application to the detection of anomalous crop development,” *Computers and Electronics in Agriculture*, vol. 198, p. 106983, 7 2022.
- [58] L. E. Christovam, M. H. Shimabukuro, M. De, L. B. T. Galo, and E. Honkavaara, “Pix2pix conditional generative adversarial network with mlp loss function for cloud removal in a cropland time series,” *Remote Sensing*, vol. 14, no. 1, p. 144, 2021.

-
- [59] K. K. Gadiraju, B. Ramachandra, Z. Chen, R. R. Vatsavai, and A. C. M. S. A. C. M. SIGMOD, “Multimodal deep learning based crop classification using multispectral and multitemporal satellite imagery,” pp. 3234–3242, Association for Computing Machinery, 2020.
- [60] Z. Li, G. Zhou, and T. Zhang, “Interleaved group convolutions for multitemporal multisensor crop classification,” *Infrared Physics and Technology*, vol. 102, 2019.
- [61] Y. Wang, Z. Zhang, L. Feng, Y. Ma, and Q. Du, “A new attention-based cnn approach for crop mapping using time series sentinel-2 images,” *Computers and Electronics in Agriculture*, vol. 184, p. 106090, 2021.
- [62] C. Boryan, Z. Yang, R. Mueller, and M. Craig, “Monitoring us agriculture: the us department of agriculture, national agricultural statistics service, cropland data layer program,” *Geocarto International*, vol. 26, pp. 341–358, 8 2011.
- [63] M. O. Turkoglu, S. D’Aronco, G. Perich, F. Liebisch, C. Streit, K. Schindler, and J. D. Wegner, “Crop mapping from image time series: Deep learning with multi-scale label hierarchies,” *Remote Sensing of Environment*, vol. 264, 2021.
- [64] A. Moumni, B. Sebbar, V. Simonneaux, J. Ezzahar, A. Lahrouni, and T. S. of Photo-Optical Instrumentation Engineers (SPIE), “Sample period dependent classification approach for the cartography of crops in the haouz plain, morocco,” vol. 11149, SPIE, 2019.
- [65] A. Masse, D. Ducrot, and P. Marthon, “Tools for multitemporal analysis and classification of multisource satellite imagery,” *2011 6th International Workshop on the Analysis of Multi-Temporal Remote Sensing Images*, pp. 209–212, 2011.
- [66] T. G. V. Niel and T. R. McVicar, “Determining temporal windows for crop discrimination with remote sensing: A case study in south-eastern australia,” *Computers and Electronics in Agriculture*, vol. 45, no. 1-3, pp. 91–108, 2004.

- [67] H. Carrão, P. Gonçalves, and M. Caetano, “Contribution of multispectral and multitemporal information from modis images to land cover classification,” *Remote Sensing of Environment*, vol. 112, no. 3, pp. 986–997, 2007.
- [68] D. Zhang, S. Fang, B. She, H. Zhang, N. Jin, H. Xia, Y. Yang, and Y. Ding, “Winter wheat mapping based on sentinel-2 data in heterogeneous planting conditions,” *Remote Sensing*, vol. 11, no. 22, p. 2647, 2019.
- [69] P. Fang, X. Zhang, P. Wei, Y. Wang, H. Zhang, F. Liu, and J. Zhao, “The classification performance and mechanism of machine learning algorithms in winter wheat mapping using sentinel-2 10 m resolution imagery,” *Applied Sciences*, vol. 10, no. 15, p. 5075, 2020.
- [70] C. Sun, Y. Bian, T. Zhou, and J. Pan, “Using of multi-source and multi-temporal remote sensing data improves crop-type mapping in the subtropical agriculture region,” *Sensors*, vol. 19, no. 10, p. 2401, 2019.
- [71] B. Chen, H. Zheng, L. Wang, O. Hellwich, C. Chen, L. Yang, T. Liu, G. Luo, A. Bao, and X. Chen, “A joint learning im-bilstm model for incomplete time-series sentinel-2a data imputation and crop classification,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 108, 4 2022.
- [72] Z. Ma, Z. Liu, Y. Zhao, L. Zhang, D. Liu, T. Ren, X. Zhang, and S. Li, “An unsupervised crop classification method based on principal components isometric binning,” *ISPRS International Journal of Geo-Information*, vol. 9, no. 11, p. 648, 2020.
- [73] M. A. Peña, R. Liao, and A. Brenning, “Using spectrottemporal indices to improve the fruit-tree crop classification accuracy,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 128, pp. 158–169, 2017.
- [74] S. Mangiarotti, A. K. Sharma, S. Corgne, L. Hubert-Moy, L. Ruiz, M. Sekhar, and Y. Kerr, “Can the global modeling technique be

- used for crop classification?,” *Chaos, Solitons and Fractals*, vol. 106, pp. 363–378, 2018.
- [75] L. A. Ruiz, J. Almonacid-Caballer, P. Crespo-Peremarch, J. A. Recio, J. E. Pardo-Pascual, and E. Sánchez-García, “Automated classification of crop types and condition in a mediterranean area using a fine-tuned convolutional neural network,” vol. 43, pp. 1061–1068, International Society for Photogrammetry and Remote Sensing, 2020.
- [76] A. Moumni, M. Oujaoura, J. Ezzahar, and A. Lahrouni, “A new synergistic approach for crop discrimination in a semi-arid region using sentinel-2 time series and the multiple combination of machine learning classifiers,” vol. 1743, p. 012026, IOP Publishing Ltd, 2021.
- [77] U. Kanjir, N. Đurić, and T. Veljanovski, “Sentinel-2 based temporal detection of agricultural land use anomalies in support of common agricultural policy monitoring,” *ISPRS International Journal of Geo-Information*, vol. 7, no. 10, p. 405, 2018.
- [78] M. A. Peña and A. Brenning, “Assessing fruit-tree crop classification from landsat-8 time series for the maipo valley, chile,” *Remote Sensing of Environment*, vol. 171, pp. 234–244, 2015.
- [79] P. Benevides, H. Costa, F. D. Moreira, D. Moraes, M. Caetano, T. I. of Electrical, E. E. Geoscience, and R. S. S. (GRSS), “Annual crop classification experiments in portugal using sentinel-2,” pp. 5838–5841, Institute of Electrical and Electronics Engineers Inc., 2021.
- [80] V. Saxena, R. K. Dwivedi, A. Kumar, and iNurture Education Solutions Private Limited; Teerthanker Mahaveer University, “Analysis of machine learning algorithms for crop mapping on satellite image data,” pp. 23–28, Institute of Electrical and Electronics Engineers Inc., 2021.
- [81] S. Jing and T. Chao, “Time series land cover classification based on semi-supervised convolutional long short-term memory neural networks,” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Science*, vol. 43, no. B2, pp. 1521–1528, 2020.

- [82] M. K. Dhodhi, J. A. Saghri, I. Ahmad, and R. Ul-Mustafa, “Disodata: A distributed algorithm for unsupervised classification of remotely sensed data on network of workstations,” *Journal of Parallel and Distributed Computing*, vol. 59, pp. 280–301, 11 1999.
- [83] A. J. Rivera, M. D. Pérez-Godoy, D. Elizondo, L. Deka, and M. J. del Jesus, “A preliminary study on crop classification with unsupervised algorithms for time series on images with olive trees and cereal crops,” *Advances in Intelligent Systems and Computing*, vol. 1268, pp. 276–285, 2021.
- [84] X. Wang, J. Zhang, L. Xun, J. Wang, Z. Wu, M. HENCHIRI, S. Zhang, S. Zhang, Y. Bai, S. Yang, S. Li, and X. Yu, “Evaluating the effectiveness of machine learning and deep learning models combined time-series satellite data for multiple crop types classification over a large-scale region,” *Remote Sensing*, vol. 14, no. 10, p. 2341, 2022.
- [85] K. Tatsumi, Y. Yamashiki, M. A. C. Torres, and C. L. R. Taipe, “Crop classification of upland fields using random forest of time-series landsat 7 etm+ data,” *Computers and Electronics in Agriculture*, vol. 115, pp. 171–179, 2015.
- [86] A. Khobragade, P. Athawale, and M. Raguwanshi, *Optimization of statistical learning algorithm for crop discrimination using remote sensing data*. 2015.
- [87] L. Yang, L. R. Mansaray, J. Huang, and L. Wang, “Optimal segmentation scale parameter, feature subset and classification algorithm for geographic object-based crop recognition using multisource satellite imagery,” *Remote Sensing*, vol. 11, no. 5, p. 514, 2019.
- [88] J. Pluto-Kossakowska, “Review on multitemporal classification methods of satellite images for crop and arable land recognition,” *Agriculture*, vol. 11, no. 10, 2021.
- [89] A. Li, S. Liang, A. Wang, and J. Qin, “Estimating crop yield from multi-temporal satellite data using multivariate regression and neural network techniques,” *Photogrammetric Engineering and Remote Sensing*, vol. 73, no. 10, pp. 1149–1157, 2007.

-
- [90] D. Wang, W. Cao, F. Zhang, Z. Li, S. Xu, and X. Wu, “A review of deep learning in multiscale agricultural sensing,” *Remote Sensing*, vol. 60, pp. 1–16, 2022.
- [91] S. Yang, L. Gu, X. Li, F. Gao, and T. Jiang, “Fully automated classification method for crops based on spatiotemporal deep-learning fusion technology,” *IEEE Transactions On Geoscience And Remote Sensing*, vol. 60, pp. 1–16, 2022.
- [92] M. Y. Moreno-Revelo, L. Guachi-Guachi, J. B. Gómez-Mendoza, J. Revelo-Fuelagán, and D. H. Peluffo-Ordóñez, “Enhanced convolutional-neural-network architecture for crop classification,” *Applied Sciences*, vol. 11, no. 9, p. 4192, 2021.
- [93] J. M. Haut, A. Alcolea, M. E. Paoletti, J. Plaza, J. Resano, and A. Plaza, “Gpu-friendly neural networks for remote sensing scene classification,” *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.
- [94] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, “Deep learning in remote sensing: A comprehensive review and list of resources,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.
- [95] X. Jia, A. Khandelwal, D. J. Mulla, P. G. Pardey, and V. Kumar, “Bringing automated, remote-sensed, machine learning methods to monitoring crop landscapes at scale,” *Agricultural Economics*, vol. 50, no. S1, pp. 41–50, 2019.
- [96] B. She, Y. Yang, Z. Zhao, L. Huang, D. Liang, and D. Zhang, “Identification and mapping of soybean and maize crops based on sentinel-2 data,” *International Journal of Agricultural and Biological Engineering*, vol. 13, no. 6, pp. 171–182, 2020.
- [97] J. B. Castro, R. Q. Feitosa, L. C. L. Rosa, P. A. Diaz, and I. Sanchez, “A comparative analysis of deep learning techniques for sub-tropical crop types recognition from multitemporal optical/sar image sequences; a comparative analysis of deep learning techniques for sub-tropical crop types recognition from multitemporal optical/sar

- image sequences,” *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2017.
- [98] S. Yang, L. Gu, X. Li, T. Jiang, and R. Ren, “Crop classification method based on optimal feature selection and hybrid cnn-rf networks for multi-temporal remote sensing imagery,” *Remote Sensing*, vol. 12, no. 19, p. 3119, 2020.
- [99] V. Mazzia, A. Khaliq, and M. Chiaberge, “Improvement in land cover and crop classification based on temporal features learning from sentinel-2 data using recurrent-convolutional neural network (r-cnn),” *Applied Sciences*, vol. 10, no. 1, p. 238, 2020.
- [100] G. S. Phartiyal and D. Singh, “Comparative study on deep neural network models for crop classification using time series polsar and optical data,” vol. 42, pp. 675–681, International Society for Photogrammetry and Remote Sensing, 2018.
- [101] A. Morales, “Los 10 formatos gis vectoriales más populares.” <https://mappinggis.com/2013/11/los-formatos-gis-vectoriales-mas-populares>.
- [102] A. Lozano-Tello, M. Fernández-Sellers, E. Quirós, L. Fragoso-Campón, A. García-Martín, J. A. G. Gallego, C. Mateos, R. Trenado, and P. Muñoz, “Crop identification by massive processing of multiannual satellite imagery for eu common agriculture policy subsidy control,” *European Journal of Remote Sensing*, vol. 54, no. 1, pp. 1–12, 2021.
- [103] G. Siesto, M. Fernández-Sellers, A. Lozano-Tello, D. S. Junior, and L. S. Shiratsuchi, “Crop classification of satellite imagery using synthetic multitemporal and multispectral images in convolutional neural networks,” *Remote Sensing*, vol. 13, p. 3378, 8 2021.
- [104] M. Claesen and B. D. Moor, “Hyperparameter search in machine learning.” <https://arxiv.org/pdf/1502.02127.pdf>, 2015.
- [105] J. T. Tsai, J. H. Chou, and T. K. Liu, “Tuning the structure and parameters of a neural network by using hybrid taguchi-genetic algo-

-
- rithm,” *IEEE Transactions on Neural Networks*, vol. 17, pp. 69–80, 1 2006.
- [106] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization yoshua bengio,” *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.
- [107] D. R. Jones, M. Schonlau, and W. J. Welch, “Efficient global optimization of expensive black-box functions,” *Journal of Global Optimization*, vol. 13, no. 4, pp. 455–492, 1998.
- [108] M. Aach, R. Sedona, A. Lintermann, G. Cavallaro, H. Neukirchen, and M. Riedel, “Accelerating hyperparameter tuning of a deep learning model for remote sensing image classification,” *International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 2022, pp. 263–266, 2022.
- [109] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. New York: Springer, 2013.
- [110] M. Fernández-Sellers, G. Siesto, A. Lozano-Tello, and P. J. Clemente, “Finding a suitable sensing time period for crop identification using heuristic techniques with multi-temporal satellite images,” *International Journal of Remote Sensing*, vol. 43, no. 15-16, pp. 6038–6055, 2021.
- [111] S. Russell and P. Norving, *Artificial Intelligence: A Modern Approach*. Harlow: Pearson, 2022.
- [112] Y. Sasaki, “The truth of the f-measure.” <https://www.cs.odu.edu/~mukka/cs795sum09dm/Lecturenotes/Day3/F-measure-YS-26Oct07.pdf>, 2007.
- [113] GeoJSON, “Geojson web page.” <https://geojson.org>.
- [114] S. Gillies, M. Russell, C. Farwell, and B. Grotewold, “Python-geojson library.” <https://python-geojson.readthedocs.io>.

REFERENCIAS

- [115] S. Gillies, “Shapely library.” <https://shapely.readthedocs.io>.
- [116] Y. Matsubara, “Pymysql.” <https://pymysql.readthedocs.io>.
- [117] ESA, “Copernicus open access hub.” <https://scihub.copernicus.eu>.
- [118] M. Wille and K. Clauss, “Sentinelat.” <https://sentinelat.readthedocs.io>.
- [119] ESA, “Sen2cor.” <https://step.esa.int/main/snap-supported-plugins/sen2cor>.
- [120] A. Schwankner, “Esa snap with python.” <https://github.com/schwankner/esa-snap-with-python>.
- [121] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.
- [122] Wes McKinney, “Data Structures for Statistical Computing in Python,” in *Proceedings of the 9th Python in Science Conference* (Stéfan van der Walt and Jarrod Millman, eds.), pp. 56–61, 2010.
- [123] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” <https://arxiv.org/pdf/1502.03167.pdf>, 2015.