



Escuela
Politécnica
Superior

Técnicas de aprendizaje automático aplicadas a la estimación del estado de cultivos mediante series temporales



Grado en Ingeniería en Sonido e
Imagen en Telecomunicación

Trabajo Fin de Grado

Autor:

Anaida Fernández García

Tutor/es:

Juan Manuel López Sánchez

Tomás Martínez Marín

Julio 2020



Universitat d'Alacant
Universidad de Alicante

Técnicas de aprendizaje automático aplicadas a la estimación del estado de cultivos mediante series temporales

Autor

Anaida Fernández García

Tutor/es

Juan Manuel López Sánchez

Dpto. de Física, Ing. Sistemas y Teoría de la Señal

Tomás Martínez Marín

Dpto. de Física, Ing. Sistemas y Teoría de la Señal



Grado en Ingeniería en Sonido e Imagen en Telecomunicación



Escuela
Politécnica
Superior



Universitat d'Alacant
Universidad de Alicante

ALICANTE, Julio 2020

Justificación y Objetivos

“Las razones que me han llevado a realizar este Trabajo Fin de Grado (TFG) son colaborar en una investigación en la que el Departamento de Física, Ingeniería de Sistemas y Teoría de la Señal de la Escuela Politécnica Superior lleva trabajando más de 6 años con amplias expectativas de futuro, profundizar mi conocimiento sobre área de las Telecomunicaciones que engloba los sistemas radar y las comunicaciones vía satélite y, por último, adentrarme en las tecnologías emergentes que tanta repercusión van a tener en nuestra vida como son las técnicas de aprendizaje automático o Machine Learning (ML).”

Agradecimientos

El hecho de que la realización de este trabajo fin de grado haya sido posible no recae solo en mí, muchas personas han contribuido a lo largo de mi vida para que pueda estar en este momento finalizando mi carrera, por ello quiero agradecer:

A mis padres, Ana García y Diego Fernández, por darme las oportunidades y abrirme las puertas siempre para hacer todo lo que a mí me moviera.

A mi hermana Lorena, por ser modelo a seguir y por inspirarme con tu constancia, esfuerzo y motivación.

A todas mis amigas, de Murcia y de Alicante, que me han animado desde que tengo recuerdo y han estado ahí incluso cuando solo se podía estar en casa, qué bien que vosotras seáis casa.

A Aina, Enrique, Robert y Alberto por haber sido apoyo y comprensión, compañeros, amigos e incluso profesores durante estos 4 años.

A todos los profesores y personal académico, desde el instituto hasta la universidad, que me han enseñado tanto técnica y personalmente, cómo quiero ser en un futuro, y han sido inspiración mostrándome las infinitas posibilidades de este mundo. Y en especial a Plens, por poner su conocimiento a disposición de todos.

A los tutores de este proyecto, Juan Manuel López y Tomás Marín, que pensaron en mí para trabajar en las áreas que a mí me apasionan con ellos.

A Finlandia, y a todas las personas que hicieron de esa experiencia una de las mejores de mi vida.

Y a Javi, porque lo haces todo posible.

La distancia, que es el impedimento principal del progreso de la humanidad, será completamente superada, en palabra y acción.

*La humanidad estará unida, las guerras serán imposibles,
y la paz reinará en todo el planeta.*

Nikola Tesla.

Índice general

1	Introducción	1
1.1	Contexto	1
1.2	Objetivos	2
1.3	Estructura de la memoria	2
2	Marco Teórico	5
2.1	Teledetección	5
2.1.1	Tecnología radar	5
2.1.2	Satélites en teledetección	8
2.2	Técnicas de regresión y Machine Learning	10
2.2.1	Clasificación de técnicas de Machine Learning	11
2.2.2	Modelos de Machine Learning	12
2.3	Estimación de parámetros físicos de cultivos mediante teledetección	14
2.3.1	Metodología general basada en espacio de estados	15
2.3.2	Regresión aplicada a la estimación	17
3	Metodología	19
3.1	Datos para análisis y modelado	19
3.2	Procesamiento y división de datos	21
3.3	Modelo de regresión y optimización	21
3.4	Evaluación de resultados	22
4	Resultados	25
4.1	Método por parcelas	25
4.1.1	Optimización	25
4.1.2	Salidas del modelo	26
4.1.2.1	Salidas de función de densidad de probabilidad	26
4.1.2.2	Salidas de valor único	28
4.1.3	Evaluación de resultados	31
4.2	Método por píxeles	36
4.2.1	Optimización	36
4.2.2	Salidas del modelo	37
4.2.2.1	Salidas de función de densidad de probabilidad	37
4.2.2.2	Salidas de valor único	40
4.2.3	Evaluación de resultados	43
4.3	Comparativa de métodos	45
5	Conclusiones	47

Bibliografía	49
Lista de Acrónimos y Abreviaturas	51

Índice de figuras

2.1	Imagen Synthetic Aperture Radar (SAR) de Sierra Nevada tomada por el satélite de Sentinel-1 con polarización VV (dB). [1]	6
2.2	TOPS-SAR Sentinel-1 [2]	8
2.3	Swath a día 1 de 6 en Sentinel-1 [3]	9
2.4	Esquema general del funcionamiento de las técnicas de ML.	10
3.1	Mapa de la zona de estudio con las parcelas utilizadas destacadas. [1]	20
4.1	Optimización del número de árboles para Random Forest Regressor (RFR) en el modelo de salida Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie (BBCH)	25
4.2	Ejemplo de salidas Probability Density Function (PDF) normalizadas del modelo para estimación de BBCH y la altura.	27
4.3	Ejemplo de salidas PDF normalizadas del modelo para estimación de BBCH y la altura.	28
4.4	Comparación de la salida predicha y la verdad de tierra de los modelos de salida individual para la estimación de BBCH y la altura.	29
4.5	Comparación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de BBCH y la altura.	30
4.6	Relación de la salida predicha y la verdad de tierra de los modelos de variables independientes.	33
4.7	Relación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de BBCH y la altura.	34
4.8	Relación de las variables de datos de entrada al modelo con respecto a la BBCH	36
4.9	Optimización del número de árboles para RFR en el modelo de salida BBCH	37
4.10	Ejemplo de salida PDF normalizada los modelos para estimación de las variables independientes.	38
4.11	Ejemplo de salidas PDF normalizadas del modelo para estimación de BBCH y la altura.	39
4.12	Comparación de la salida predicha y la verdad de tierra de los modelos de salidas individuales.	41
4.13	Comparación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de BBCH y la altura.	42
4.15	Relación de la salida predicha y la verdad de tierra de los modelos de salidas independientes.	43
4.16	Relación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de BBCH y la altura.	44

Índice de tablas

4.1	Parámetros de optimización de entrada y modelo	26
4.2	Influencia de los parámetros de entrada en el modelo de estimación.	35
4.3	Parámetros de optimización de entrada y modelo	37
4.4	Influencia de los parámetros de entrada en el modelo de estimación.	45
4.5	Índices estadísticos de las predicciones con datos de entrada a nivel de parcela.	46
4.6	Índices estadísticos de las predicciones con datos de entrada a nivel de pixel . .	46

1 Introducción

La telecomunicación se puede definir como toda transmisión y/o emisión y recepción de señales que representan signos, escritura, imágenes y sonidos o información de cualquier naturaleza por hilo, radioelectricidad, medios ópticos u otros sistemas electromagnéticos [4]. Esto permite compartir información útil a distancia y engloba un amplio conjunto de sistemas y tecnologías.

En este apartado nos vamos a centrar en situarnos dentro de los distintos sistemas de telecomunicación, y más detenidamente en los relevantes para este proyecto. A continuación, se expondrán los objetivos concretos que se quieren alcanzar. Y, por último, cómo se va a organizar la memoria del proyecto.

1.1 Contexto

Las telecomunicaciones forman parte de nuestro día a día y tienen cometidos de lo más variados: desde mandar un simple mensaje hasta comunicarse con una estación espacial, pero todos ellos engloban el manejo o el hecho de compartir información a distancia.

Dentro de los sistemas de telecomunicación encontramos el sistema de la teledetección, definido como la adquisición de información de un objeto, área o fenómeno, con instrumentos que no están en contacto directo con el objeto, según la Real Academia de Ingeniería (RAI) [4]. Estos instrumentos van a medir la radiación electromagnética que emiten o reflejan los objetos observados. Algunos instrumentos pueden ser, por ejemplo, las cámaras fotográficas o los sistemas de radar (RADio Detection And Ranging) o sonar.

Las imágenes obtenidas desde satélite por sistemas radar son una gran fuente de información para aplicaciones de teledetección, como, por ejemplo, las predicciones meteorológicas, la realización de mapas topográficos o la monitorización de cultivos. Esta última, en la que se va a centrar este proyecto, requiere disponer de suficiente información periódica durante el tiempo que engloba el desarrollo completo del cultivo. En la monitorización de cultivos se encuentra la estimación del estado de los mismos, así como de variables descriptoras de este estado (biomasa, altura, etc.), que se obtienen a partir de imágenes que pueden ser analizadas de forma independiente o utilizando técnicas que aprovechen las series temporales de datos para la estimación. Estas técnicas son muy útiles en estos casos en los que la estimación va estrechamente ligada al transcurso del tiempo.

En la Universidad de Alicante (UA), el grupo de investigación Señales, Sistemas y Telecomunicación (SST) ha diseñado un marco de trabajo sobre este tema basado en espacio de estados, que permite combinar de forma óptima los modelos de evolución esperable de los

cultivos con los datos de otras fuentes como las imágenes SAR de satélite o la temperatura acumulada medida por una estación meteorológica.

Hasta la fecha, el modelo de observación utilizado que relacionaba las observaciones proporcionadas por las imágenes SAR con el estado fenológico de los cultivos era bastante simplificado y sus resultados no eran óptimos. Aquí entra el propósito de este TFG, contribuir a su optimización mediante la generación de modelos de observación más complejos para las imágenes SAR. Estos modelos están basados en regresión con técnicas de machine learning que introduzcan la información de estas imágenes en el modelo de espacio de estados mencionado previamente.

1.2 Objetivos

El objetivo general de este TFG es estimar el estado de cultivos de arroz mediante el análisis series temporales con técnicas de aprendizaje automático y su unión a la línea de procesamiento original.

Los objetivos concretos serían:

- Analizar las posibles técnicas de regresión de aprendizaje autónomo (por ejemplo, regresión con Random Forest (RF)) para estimar directamente el estado de los cultivos a partir de series temporales de datos.
- Analizar las posibles técnicas de aprendizaje autónomo para ser combinados con algoritmos ya disponibles de dinámica de sistemas en la estimación del estado de cultivos.
- Incorporar dichas técnicas en la cadena de procesado disponible.

1.3 Estructura de la memoria

La estructura de la memoria se divide en 4 capítulos principales, omitiendo este de introducción, los cuales son: marco teórico, metodología, resultados y conclusiones.

En el marco teórico se expone toda la teoría necesaria para la compresión de este proyecto en términos técnicos y dentro de un contexto y una investigación previa que este continúa. Se ven en él 3 secciones:

- Teledetección, incluyendo cómo funcionan los sistemas radar, en concreto los SAR, qué información obtenemos de ellos y qué papel desempeñan los satélites y las técnicas de detección en el proceso de obtener e interpretar esta información.
- Técnicas de regresión y Machine Learning, donde se encuentra la clasificación de las distintas técnicas de ML y algunos modelos, más concretamente extendido el análisis mediante regresión.
- Estimación de parámetros físicos de cultivos mediante teledetección, donde se presentan la metodología general basada en el espacio de estados presentada en el marco de trabajo previo y la estimación de los parámetros físicos de los cultivos mediante regresión.

El siguiente capítulo, metodología, expone cómo se ha realizado este trabajo y está dividido en 3 secciones principales:

- Datos para análisis y modelado, que incluye la obtención de datos que se disponen para este estudio y la descripción detallada de todos ellos y su utilidad.
- Procesamiento y división de datos, donde se explica el procesamiento de datos realizado para su preparación para la creación del modelo e incluyendo la división de datos según su funcionalidad dentro del modelo.
- Modelo de regresión y optimización, aquí se comparte qué modelo se va a utilizar, cómo funciona, las necesidades que tiene y la manera de optimizarlo para solucionar un problema concreto.
- Evaluación de resultados, expone cómo se van a presentar los resultados y qué métodos se van a utilizar para contrastar su eficiencia y fiabilidad.

Seguidamente, el apartado de resultados tiene como objetivo exponer los resultados obtenidos con los distintos métodos y casos estudiados. Este apartado consta de las siguientes secciones:

- Método por parcelas, donde se muestra la optimización realizada para este método, las salidas del modelo implementado y su evaluación.
- Método por píxeles, donde se muestra, al igual que en el apartado anterior, la optimización realizada para este método, las salidas del modelo implementado y su evaluación.
- Comparativa de métodos, se evalúan los métodos utilizados para todos los casos y se justifican los resultados obtenidos.

Por último, el capítulo de conclusiones expone la evaluación general sobre el trabajo realizado y los resultados obtenidos, las posibles mejoras que se podrían implementar y las líneas futuras del proyecto.

2 Marco Teórico

A continuación se expone la teoría necesaria para la comprensión de este TFG, ampliando la información presentada en el capítulo 1.

2.1 Teledetección

Conociendo de qué trata este sector de las telecomunicaciones de una manera general y en el contexto en el que se sitúa la contribución de este TFG, en este apartado se van a explicar conceptos más concretos de cómo funciona y cómo va a ser utilizada esta tecnología en el proyecto.

Como ya se menciona en el capítulo 1, los instrumentos de teledetección se caracterizan por medir la radiación electromagnética emitida o reflejada por un objeto o superficie que se encuentra a distancia del mismo. Estos instrumentos se pueden clasificar en dos tipos: pasivos, miden la radiación natural emitida o reflejada por el objeto observado, o activos, emiten energía que posteriormente será reflejada y detectada.

Algunos de estos instrumentos son cámaras fotográficas, láseres o radares. Todos ellos trabajan en un determinado rango en el espectro electromagnético, dependiendo de lo que quieran captar, por ejemplo los sensores ópticos necesitan trabajar en frecuencias del espectro visible, mientras que los radares pueden trabajar a frecuencias de microondas.

2.1.1 Tecnología radar

Para este proyecto nos vamos a centrar en el instrumento llamado radar. Se trata de un instrumento activo, que trabaja en el espectro de las microondas, concretamente entre 1-40 GHz. Por ello, son sensores sensibles a objetos del tamaño de sus longitudes de onda, es decir entre 30-0.75 cm, mucho mayores que los sensores ópticos. Además, otras ventajas frente a estos es su penetración en nubosidades e incluso parcialmente en la superficie terrestre o la vegetación, por lo que, por ejemplo, el tiempo atmosférico, no supone un impedimento para la captación de información en la mayoría de los casos, como sucede en los sensores ópticos, y esto también se debe a la mayor longitud de onda.

Al tratarse de un instrumento activo, este también ilumina con pulsos electromagnéticos el área a observar, emitidos por una antena propia, por lo que no depende de fuentes externas de radiación, esto es que, incluso objetos que no emitan radiación ni reflejen de otras fuentes, se podrán detectar gracias a la reflexión de esta incidencia.

La información que se obtiene en una antena receptora (o la misma emisora con doble función) de esta reflexión es otra onda electromagnética de la cual se puede medir su potencia,

fase y polarización para obtener información útil del área observada. Para todas ellas deben tenerse en cuenta parámetros adicionales que influyen durante el trayecto del pulso, como es el retardo de fase o las pérdidas de potencia. El producto típico que esto genera se conoce como imagen SAR, cuyo ejemplo se puede ver en la figura 2.1, que es una representación gráfica cuyo eje horizontal corresponde al azimuth y el eje vertical al rango o distancia, y se presentan, normalmente en escala de grises, las contribuciones para cada “pixel”, siendo el blanco la máxima y el negro la mínima. También se puede representar de igual manera la fase, en escala de grises o color, repartidos entre los 0-360°.

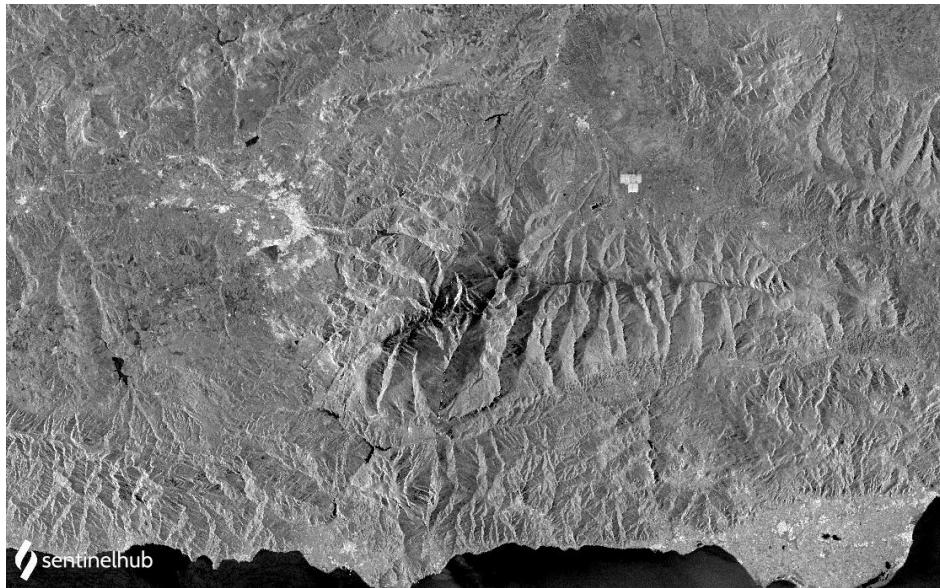


Figura 2.1: Imagen SAR de Sierra Nevada tomada por el satélite de Sentinel-1 con polarización VV (dB). [1]

La potencia recibida (P_R) se puede considerar como la contribución de los siguientes parámetros: potencia transmitida (P_T), longitud de onda (λ), ganancia de la antena (G), pérdidas en sistema (L_s) y en atmósfera (L_a), la longitud del trayecto (R), la superficie del área observada (S) y, por último, el coeficiente de backscattering (σ_0), teniendo en cuenta ambos trayectos de ida y vuelta y la propagación esférica de la onda, como se muestra en la fórmula 2.1.

$$P_R = \frac{P_T \lambda^2 G^2 \sigma_0 S}{(4\pi)^3 L_s L_a R^4} \quad (2.1)$$

Como podemos observar, el parámetro más interesante que nos va a dar información sobre el área observada es el coeficiente de backscattering. Este parámetro es un valor adimensional (dB) que representa la relación entre la proporción de área equivalente si el objeto observado fuera un blanco isótropo (reflexión total) (m^2) y su área o superficie observada real (m^2). Este parámetro va a depender de la frecuencia utilizada por el radar, la polarización de la

onda, el ángulo de incidencia del pulso y del material y geometría de la superficie observada. Teniendo en cuenta solamente el coeficiente de backscattering y la longitud de onda empleada (λ), ya existen ciertos rangos que suelen representar distintos tipos de superficies observadas:

- $\sigma_0 > 0$ dB: típicamente objeto artificial liso que está encarado al ángulo de incidencia del radar y actúa como un espejo.
- $-10 \text{ dB} < \sigma_0 < 0 \text{ dB}$: superficies muy rugosas como pueden ser vegetaciones densas donde hay mucha probabilidad de reflexión.
- $-20 \text{ dB} < \sigma_0 < -10 \text{ dB}$: superficies rugosas como vegetaciones menos densas entre las que se incluirían los cultivos.
- $\sigma_0 < -20 \text{ dB}$: superficies lisas que no encaran el haz de incidencia del radar por lo que reflejan casi todo a otra dirección, esto se da en masas de agua en calma, carreteras o suelos muy secos.

Por otra parte, existen técnicas de detección más complejas que consideran también la información proveniente de la fase y la polarización para el desarrollo de modelos. Los principales son la interferometría, la interferometría diferencial y la polarimetría.

Con el objetivo de maximizar la resolución espacial del área observada por el radar, esto es, la distancia mínima distingible, se necesita mejorar la resolución en rango y en azimuth. La resolución en rango depende del tamaño del pulso, ya que los objetos podrán ser diferenciados si están a una distancia mayor que un pulso, por lo que cuánto más pequeño sea este, mayor resolución de rango se obtendrá, aunque se debe mantener la duración del pulso para el rango de frecuencias asignado. Por otra parte, para mejorar la resolución en azimuth, se debe incrementar la resolución angular. Esta resolución es inversamente proporcional al ángulo de observación, ya que cuanto mayor es este, más objetos o áreas considera a la misma distancia y no es posible diferenciarlos. Para obtener ángulos pequeños se necesita un swath, o haz de área iluminada, muy directivo, y ello se consigue con una apertura o longitud de radar grande. Es aquí donde entran los sistemas radar de los que se va a extraer la información para este proyecto, los SAR. Estos sistemas consiguen aumentar la resolución en azimuth, con una longitud no muy grande, realizando un barrido en azimuth y posterior procesamiento para ampliar el área observada sin compensar empeorando la resolución en azimuth. Un ejemplo de este barrido se puede ver en la figura 2.2.

Además de la resolución espacial, también existen otros parámetros que determinan la calidad de la información adquirida por los sistemas radar. Uno de ellos es el ruido que aparece como puntos de máximas y mínimas contribuciones debido a fenómenos puntuales que puede dificultar la interpretación de la información. Este ruido se denomina speckle, y la técnica utilizada para atenuar este efecto es la reducción por multi-look. Como su nombre indica, esta técnica se basa en la toma de varias imágenes de información de la misma área y el posterior promediado todas ellas, obteniendo una información más plana, viéndose reducidos los píxeles de información aleatorios máximos y mínimos.

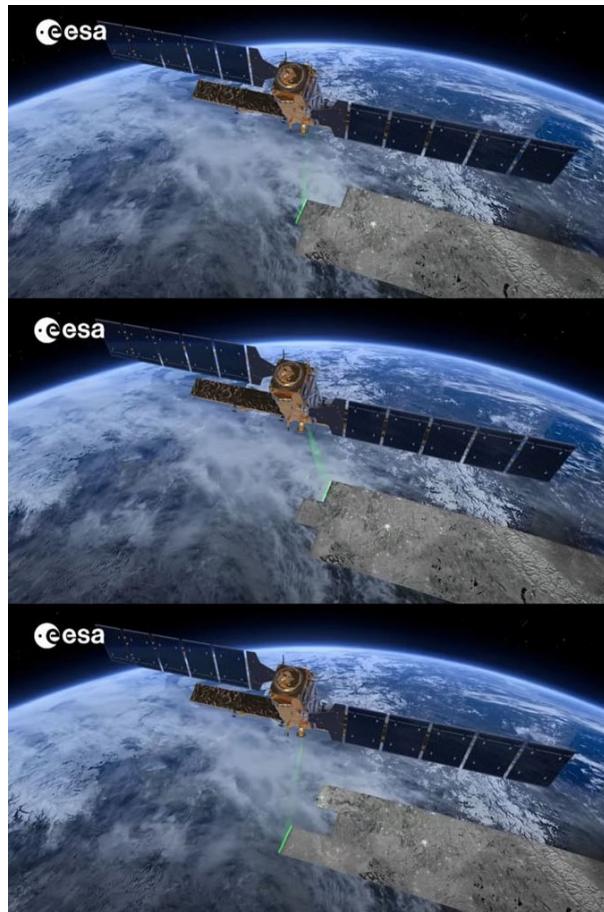


Figura 2.2: TOPS-SAR Sentinel-1 [2]

2.1.2 Satélites en teledetección

Los sistemas SAR van a ser utilizados en este proyecto para observar parcelas cultivadas de la Tierra, por lo que los sistemas en los que se van a emplazar estos son los satélites. La misión satelital de la que se va a obtener información es Sentinel-1, del Programa Copérnico de la European Space Agency (ESA). Es una misión de órbita polar que engloba 2 satélites, A y B, cuyo objetivo es la observación de la superficie de la Tierra tanto terrestre como oceánica. Sentinel-1 concluyó sus lanzamientos de satélites en abril de 2016. El rango de frecuencias de trabajo de Sentinel-1 proporciona información de la banda C, esto es entre 4-8 GHz, aunque, del mismo programa, la misión Sentinel-2 utiliza tecnología multiespectral, esto significa que trabaja en 13 bandas distintas, las cuales engloban la luz visible, el infrarrojo cercano y el infrarrojo de onda corta. Esto proporciona información más precisa y adecuada para cada fenómeno a observar [5], lo cual puede ser utilizado para futuros avances de esta investigación.

La órbita se traza en el eje polar de la Tierra con una pequeña inclinación y sincrónica al Sol, un periodo de revista global de 6 y una altitud de 693 km para Sentinel-1 ambos satélites, A y B. En la figura 2.3 se puede observar la órbita trazada por Sentinel-1 en un

periodo de un día. Además, el sistema SAR no trabaja desde una posición perpendicular a la superficie terrestre a medir, ya que algunas superficies serían consideradas a la misma distancia por simetría en el swath, por lo que la visión del radar es lateral derecha. Esto deberá ser considerado para el procesamiento de extracción de información.

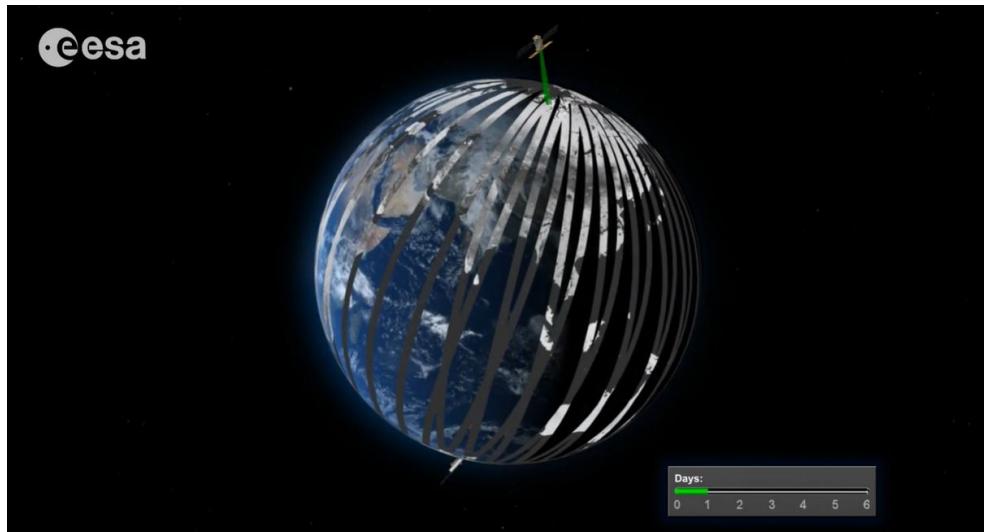


Figura 2.3: Swath a día 1 de 6 en Sentinel-1 [3]

Sentinel-1 tiene 4 modos de adquisición principales según el área que se pretende observar cuyos swath y resolución espacial varían. El primer modo, llamado Stripmap Mode, presenta un swath de 80 km y una resolución de 5x5 m. Este modo se utiliza para monitorización de islas pequeñas y emergencias puntuales. El segundo modo es Interferometric Wide Swath, de 250 km de swath y 5x20 m de resolución, es utilizado principalmente para todas las áreas de superficie terrestre, tanto áreas habitadas, como zonas montañosas o llanuras (donde se incluyen los cultivos). El tercer modo se conoce como Extra Wide Swath Mode, consta de un swath de 400 km y una resolución de 20x40 m, es utilizado para zonas marítimas, polares o cubiertas de hielo, donde se buscan grandes coberturas y un tiempo de revista corto, ya que, por el eje elegido para su órbita, las zonas polares se cubren en menor tiempo. Por último, cabe destacar el Wave Mode, cuyo swath se caracteriza por considerarse de superficie cuadrada de 20x20 km, y con una resolución de 20x5 m. Este es utilizado para la observación de los océanos [6].

Para que la utilización de estos modos sea posible, se necesita una tecnología SAR acorde con estas necesidades. El radar tiene unas dimensiones en Sentinel-1 de antena de 12.3 m x 0.821 m una vez desplegado. El rango del ángulo de incidencia con respecto a la Tierra es de 20"-46". Los modos de adquisición también pueden trabajar con distintas polarizaciones. Las ofrecidas por los satélites de Sentinel-1 para la emisión son Horizontal (H) y Vertical (V). Para la recepción se pueden elegir la misma polarización utilizada en emisión, lo que sería HH o VV, o recibir ambas polarizaciones independientemente de cuál haya sido enviada, HH+HV o VV+VH [6]. Una emisión con polarización doble entorpecería el procesamiento ya que no

se podría reconocer en la recepción qué parte de la señal correspondía a cada una.

2.2 Técnicas de regresión y Machine Learning

Las técnicas de regresión proporcionan una estimación de una variable dependiente de otras la cual es útil para realizar predicciones, por lo que están relacionados con el aprendizaje automático o ML. El ML es un tipo de inteligencia artificial, que se caracteriza por la generación de un modelo estimado de manera automática por un computador. Esta estimación se realiza con un entrenamiento previo aplicado a un algoritmo de aprendizaje específico a una serie de datos de entrenamiento. Con este aprendizaje se elabora un modelo que es capaz de devolver una salida o solución a partir de unos parámetros de entrada que deben ser del mismo tipo que los utilizados en la fase de aprendizaje, el esquema general se puede ver en la imagen 2.4.

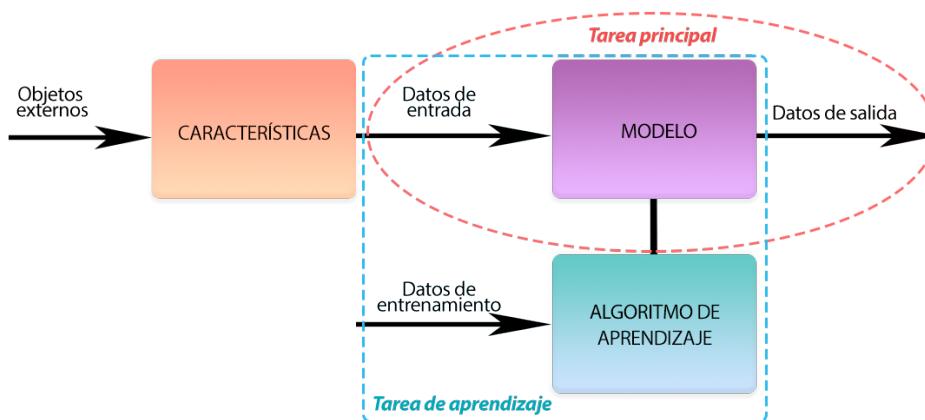


Figura 2.4: Esquema general del funcionamiento de las técnicas de ML.

Finalmente, el objetivo de las técnicas de ML puede ser clasificar una información o realizar una previsión acorde con un modelo estimado. Como se puede ver, el objetivo de este y las técnicas de regresión pueden coincidir para el segundo caso, y esto lleva a un desarrollo conjunto de ambas técnicas.

2.2.1 Clasificación de técnicas de Machine Learning

Los modelos empleados en ML son numerosos, y su clasificación se puede realizar dependiendo de su algoritmo de aprendizaje y del tipo de razonamiento en el que se basa. Comenzando por la clasificación según su algoritmo de aprendizaje, que principalmente se dividen según el feedback del que aprenden, los modelos se pueden clasificar de la siguiente manera [7]:

- Aprendizaje no supervisado: este aprendizaje se basa en la clasificación o agrupación de los objetos de entrada según patrones que cumplen las distintas entradas de estos. Estos métodos no devuelven un nombre específico para cada grupo o cluster ya que no se le han proporcionado referencias o etiquetas en la etapa de entrenamiento. Necesita numerosas entradas en el entrenamiento para detectar patrones suficientemente estables.
- Aprendizaje por refuerzo: el aprendizaje se realiza por refuerzo positivo, que sería una recompensa, o negativo, penalización. Este algoritmo buscaría la estimación del modelo para obtener el máximo refuerzo positivo posible. Así, tras suficiente entrenamiento construye un modelo muy preciso para nuevas entradas.
- Aprendizaje supervisado: tanto las entradas como las salidas están previamente definidas en la etapa de aprendizaje. Se realiza un entrenamiento en el que se utilizan las entradas con sus correspondientes salidas para elaborar el modelo. Una vez suficientemente entrenado, este puede obtener salidas previamente desconocidas a partir de entradas similares a las del entrenamiento. Los resultados están limitados a los proporcionados en la etapa de entrenamiento, pero estos suelen ser más estables con menos cantidad de datos que otros modelos.
- Aprendizaje semi-supervisado: este aprendizaje recibe algunas de sus entradas correctamente etiquetadas y el resto de ellas, la mayoría, sin etiquetar, así tiene algunas referencias para la clasificación fiables pero no toma las etiquetas como una referencia totalmente cierta para toda la clasificación como ocurre en el aprendizaje supervisado. Así se evitan malos aprendizajes por ruido o etiquetas erróneas en los datos de entrada. Bastante común en grandes masas de datos para aprendizaje.

Por otra parte, teniendo en cuenta la base de los razonamientos internos que los algoritmos realizan para obtener las salidas correspondientes, aunque no considerando esta división estricta, las técnicas se pueden clasificar de la siguiente manera [8]:

- Geométricos: los modelos geométricos son aquellos cuyos objetos pueden ser representados en un espacio de instancias (X) en el que cada instancia corresponde a un posible objeto, esto es, habrá tantas instancias como objetos con distintas combinaciones de entradas posibles. Por otra parte, las etiquetas también se representan como un espacio de etiquetas (Y) con un número finito de posibilidades [9]. Utilizando estos conceptos, el algoritmo se desarrolla con otros conceptos geométricos como son líneas, planos y distancias. Estos métodos suelen ser aplicados cuando X e Y están formados por valores numéricos, que son fácilmente representables en ejes de coordenadas.
 - Probabilísticos: los modelos probabilísticos parten de la base de que las entradas de los objetos están basadas en un proceso aleatorio que hacen referencia a una distribución
-

de probabilidad desconocida. Se busca definir esa distribución $P(Y|X)$, siendo X el conjunto de objetos posibles e Y las etiquetas correspondientes. Aquí el modelo tendría como salidas probabilidades para cada una de las opciones posibles.

- Lógicos: los modelos lógicos son los más cercanos al razonamiento humano y los más comprensibles también como algoritmos. Se basan en decisiones lógicas, estructuradas típicamente en forma de árbol, esto es llamado árbol de decisiones y según las características de los parámetros de entrada nos vamos desplazando hacia la base del árbol, obteniendo al final una única salida para cada objeto de entrada. *Introducir esquema*
- Agrupaciones y gradiente: estos modelos se incluyen en los anteriores, ya que es una clasificación paralela según el tratamiento del espacio de instancias (X): agrupaciones seccionando estos espacios en un número de segmentos definido, fácilmente representables y con una única solución, en cambio; en los gradientes, no existe una segmentación previamente definida, por lo que el modelo trata todo el espacio como uno solo.

2.2.2 Modelos de Machine Learning

Una vez presentadas todas las posibles clasificaciones de técnicas de ML, podemos adentrarnos en los modelos más comunes, a qué tipo de los anteriores pertenecen y algunas aplicaciones. Algunos de los más conocidos son los siguientes:

- Redes neuronales artificiales: este modelo es de tipo geométrico y de aprendizaje supervisado, ya que el entrenamiento consta de entradas etiquetadas con su correspondiente salida. Este modelo se caracteriza por estar inspirado por las redes neuronales naturales del cerebro animal, obteniendo resultados sin unas reglas preestablecidas de análisis. Estas redes están compuestas por capas de neuronas, las cuales representan un peso y una función de activación por la que una parte de la información de entrada se va a procesar. Estas funciones y pesos se van ajustando mediante el entrenamiento hasta tener una red óptima para su funcionamiento. En cuanto a aplicaciones, la más común es el reconocimiento en imágenes de objetos o caracteres.
- Máquinas de vectores de soporte: es un modelo geométrico y supervisado. Este modelo utiliza el espacio de instancias para representar los objetos de entrenamiento como puntos y las salidas como líneas o hiperplanos, dependiendo del número de dimensiones. Una vez ajustado este modelo, las nuevas entradas se clasificarán según al espacio al que pertenezcan. Este modelo está muy relacionado con la clasificación/agrupación y la regresión. Algunas de sus aplicaciones son la clasificación de textos [10] o, el más interesante para este proyecto, la clasificación de información procedente de un SAR [11].
- Redes bayesianas: es un modelo probabilístico y gráfico, a la vez que lógico, de aprendizaje supervisado. Se basa en un modelo gráfico de nodos que corresponden a variables conocidas o desconocidas y el tratamiento probabilístico simplificado con la regla de la cadena. Son muy utilizadas en aplicaciones relacionadas con las ciencias de la salud para modelar comportamientos biológicos.

- Árboles de decisión: es un modelo lógico y de aprendizaje supervisado. Es lo de los modelos lógicos más ilustrativos porque se basa en árboles que siguen las reglas de decisión, yendo desde el primer nodo donde se sitúa la entrada resolviendo condiciones de estas hasta llegar a una única salida, alcanzable por un camino único, que es la salida del modelo. En el aprendizaje, este modelo va ajustando sus condiciones y elaborando el árbol más coherente para llegar a las soluciones necesarias. Este método es bastante sencillo de implementar y comprender. Relacionado con ese modelo también encontramos el conocido como RF. Este modelo se caracteriza por generar numerosos árboles de decisión provenientes de un factor aleatorio con la misma distribución. Al obtener los resultados de cada uno de los árboles, se realiza un promediado, tomando la respuesta más repetida como la más probable pero teniendo la posibilidad de considerar el resto de salidas en su correspondiente porcentaje. De esta manera, un modelo que era limitado a una respuesta única, se abre devolviendo una respuesta probabilística. Este tipo de respuesta es especialmente interesante para la inclusión de estas estimaciones con datos de otros modelos predictores, ya que no excluye el resto de estimaciones.
- Las técnicas de regresión: son todas aquellas técnicas, típicamente de aprendizaje supervisado, que buscan la relación de una variable dependiente con una o más variables independientes mediante la estimación de su función de regresión. Para ello se consideran y ponderan todos los valores de la variable dependiente para unos valores fijos de las variables independientes. Además, en estos análisis, también se puede tener en cuenta la varianza de la variable dependiente para estos mismos valores, pudiendo ser estudiada también mediante su distribución de probabilidad. Esta varianza indica la fiabilidad de nuestras estimaciones o el “ruido” en las medidas de la variable dependiente. En general, estas técnicas ajustan una función de regresión, por lo que se aplican cuando las variables tienen sentido numérico. Además, estas pueden ser combinadas con otros modelos, como los conjuntos de árboles de decisiones de regresión, conocido como RFR.

Ya que las técnicas de regresión pretenden modelar una función para una variable, a partir de otras conocidas, este va a ser el método más conveniente para realizar las predicciones del estado fenológico de los cultivos. El caso más sencillo de regresión es en el que solo tenemos una variable dependiente y otra independiente, este caso se conoce como regresión lineal simple, ya que la función de regresión estimada se corresponde a una ecuación lineal de una recta. Los datos que obtenemos para la variable dependiente que vamos a relacionar tienen, aparte de las componentes lineales, una componente aleatoria de ruido que puede deberse a distintos fenómenos como la precisión mínima del instrumento de medida, el ruido que existe generalmente en la medida o contribuciones de fuentes externas, consideradas como ruido también. Esta función de regresión es frecuentemente estimada mediante el Método de Mínimos Cuadrados (MMC). También existe la regresión lineal múltiple, que funciona de la misma manera pero con mayor número de variables independientes, por lo que en lugar de una recta, la función de regresión representa un plano en el que coinciden N dimensiones, siendo N el número de variables independientes total. Su expresión analítica se presenta en la ecuación , donde Y corresponde a la variable dependiente, X a las variables independientes, β parámetro de influencia de cada variable independiente, y ε el término aleatorio.

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_N X_N + \varepsilon \quad (2.2)$$

Cuando la función de regresión no es una función lineal, la regresión es no lineal, ya que la respuesta de la variable dependiente puede ser exponencial, logarítmica o polinomial, entre otras, por lo que la función de regresión presentará mayor complejidad. Aquí también es común utilizar el MMC o la regresión segmentada, que ajusta como regresión lineal segmentos de la original no lineal.

Cualquier variable independiente que tenga relación con la dependiente es útil en mayor o menor medida pero siempre proporciona información aunque su varianza sea muy grande o su contribución relativamente pequeña. Cualquier tipo de información extra proporciona un ajuste a la estimación final positivo si esta se ha modelado correctamente.

A parte de las regresiones lineales y no lineales mencionados, también encontramos otros métodos de regresión como son los mínimos errores absoluto (bastante similar al MMC), la regresión no paramétrica o la regresión lineal bayesiana.

2.3 Estimación de parámetros físicos de cultivos mediante teledetección

La fenología es la ciencia que estudia la relación entre los factores climáticos y los ciclos de los seres vivos [12]. Esto es, el estudio del desarrollo de plantas y animales en relación con parámetros ambientales. Este proyecto se centra en el estudio fenológico de plantas, en concreto de cultivos de arroz. Por lo que, a continuación, se van a presentar los conceptos físicos que definen el estado de desarrollo de estos cultivos, para comprender cuáles son los parámetros clave y utilizados en el marco de investigación. Una vez introducidos, se expone el concepto de la técnica utilizada hasta el momento, los modelos de evolución y observación. Por último, se verá la aplicación de las técnicas de regresión a los datos obtenidos para elaborar un modelo representativo del estado fenológico de un cultivo.

Algunos de los parámetros descriptores clave utilizados en el marco creado de este proyecto son los siguientes:

- Escala BBCH, que recibe su nombre por los participantes en su estudio y desarrollo, es una escala numérica de intervalo 0-9 que describe la fenología [13]. Cada valor numérico corresponde a un estado de desarrollo, desde la germinación o primeros brotes, correspondientes al estado 0, hasta la senectud, estado 9. Cada estado puede estar dividido hasta en 10 sub-etapas. El rango de que cada etapa abarca, concretamente para los cultivos de arroz, se puede observar en la tabla *insertar tabla* [14]. Para que uno de estos estados sea considerado el nivel general de una parcela, no solo tiene que ser este estado el mayoritario, sino que debe abarcar más del 50% del cultivo.
- Índice de Vegetación de Diferencia Normalizada (NDVI), enunciado anteriormente, es un observable proporcionado por los sensores ópticos que suele ser usado como índice de vegetación de diferencia normalizada, para estimar la cantidad, calidad y desarrollo de

la vegetación con base a la medición de la intensidad de la radiación de ciertas bandas del espectro electromagnético en ella. Estas bandas son concretamente las bandas del rojo y del infrarrojo cercano, con rangos de reflexión entre 0 y 1 cada una de ellas. El coeficiente glsndvi se obtiene según la fórmula 2.3, conformando un rango entre -1 y 1, y representa el desarrollo de la vegetación, ya que la contribución de la banda infrarroja cercana está ligada a la reflexión de la celulosa, por tanto a las áreas verdes y frondosas, mientras que la banda roja es mucho menos sensible a estas contribuciones y más a la absorción de clorofila. En resumen, un buen desarrollo vegetal tiene valores de NDVI más cercanos a la unidad positiva [15].

$$NDVI = \frac{IRCercano - ROJO}{IRCercano + ROJO} \quad (2.3)$$

- Temperatura del aire, o el calor acumulado durante todo el proceso de desarrollo de un cultivo, es una fuente de observaciones para el que existen modelos de observación que lo relacionan con el estado fenológico. Concretamente en los cultivos de arroz tiene un impacto notable, por lo que se considera otro de los parámetros a tener en cuenta en su monitorización y en la elaboración de modelos de predicción [16].

2.3.1 Metodología general basada en espacio de estados

En este área hay estudios previos donde se crea la metodología basada en espacio de estados utilizada. El primer artículo [17] de este marco de trabajo, de 2014, trata de estimar el estado fenológico de cultivos en tiempo real empleando espacio de estados y técnicas de sistemas dinámicos utilizando información del pasado y actualizaciones. Esto se ve representado por dos modelos diferenciables, los cuales son el modelo de evolución: modelo que predice el estado fenológico de un cultivo según el desarrollo cronológico típico del mismo, y el modelo de observación: modelo que trata de predecir el estado de fenológico a partir de variables físicas observables. Estos modelos son combinados para considerar tanto el desarrollo habitual de un cultivo como posibles variaciones temporales (adelantos o retrasos) en el mismo que pueden ser debidos a múltiples factores.

Dentro del espacio de estados se define que cada etapa de la evolución corresponde a un único estado, el cual está contenido en un sistema dinámico o proceso, ya que tiene una evolución temporal. Este sistema se define según las siguientes ecuaciones, donde la fórmula 2.4 es el proceso recursivo, correspondiente al modelo de evolución, y la fórmula 2.5 es la ecuación de medida, que relaciona una nueva observación con el sistema, por lo que corresponde al modelo de observación.

$$\dot{x}(t) = \frac{dx(t)}{dt} = f(x(t), t, v(t)) \quad (2.4)$$

$$z(t) = h(x(t), t, w(t)) \quad (2.5)$$

El vector de estados $x(t)$ es el vector de n variables de estado que describe el sistema en un momento determinado t , por lo que el espacio de estados dispone de n dimensiones. Las funciones que aportan el ratio de cambio para cada estado está representado por $f()$ y el

ruido en la evolución es $v(t)$. $z(t)$ representa el vector de salida, donde $w(t)$ es el ruido, y $h(t)$ la relación entre el vector de estados y la observación $z(t)$.

Las dos etapas en las que se divide este algoritmo son predicción y actualización. La actualización se produce cuando existe información observable nueva, esta se introduce al sistema mediante una Extended Kalman Filter (EKF) y genera un nuevo estado y una matriz de transiciones, la cual representa la probabilidad de los siguientes posibles estados.

Para la creación del modelo de evolución se utilizan los mismo observables polarimétricos (información del radar polarimétrico, vertical (V) y horizontal (H), del satélite Radsat-2) que para las actualizaciones. Los valores de este modelo no son exactamente la escala BBCH ya que no hay un registro continuo para ello pero están discretizados y agrupados en rangos equivalentes a valores fenológicos. Estas agrupaciones o clusters van ajustando su centro con las actualizaciones de los observables. La estimación, entonces, se realiza combinando la información de este modelo, una vez que está generado y conociendo la información temporal del cultivo a estimar, con la información observable por medio del EKF.

En 2016, el marco de trabajo publica otro artículo [16] en el que se trata de estimar el NDVI, siendo este un descriptor del estado fenológico también, de cultivos de arroz a partir de dos fuentes de información observables: imágenes SAR del satélite TerraSAR-X y temperatura del aire registrada, combinadas en tiempo real mediante filtros de partículas. Se mantiene el método de trabajo de espacio de estados con predicción y actualización.

Además de las mejoras que presenta por la consideración de la temperatura como fuente de información, este artículo introduce cómo trabajar con las imágenes SAR. Las imágenes SAR dan una información de el coeficiente de reflexión de la superficie observada, según una polarización de onda emitida y recibida, siendo este representado en ejes de azimuth y rango, como se ha explicado anteriormente. Estas imágenes, de libre acceso por la ESA, cuentan con canales de polarización VV y HH para TerraSAR-X y VV y VH para Sentinel-1, esto último es, emisión vertical y recepción tanto vertical como horizontal. La segunda configuración es más sensible al crecimiento de los cultivos de arroz por la verticalidad de sus tallos. Por otra parte, las imágenes de polarización horizontal recibida proporciona información extra que puede ser útil y que es bastante semejante a la obtenida en HV, es decir, emisión horizontal y recepción vertical, por lo que se abarca la mayor parte de información útil posible. A parte de las distintas polarizaciones, las imágenes SAR obtenidas por el satélite TerraSAR-X son equivalentes a las obtenidas por Sentinel-1, por lo que su marco de trabajo es perfectamente compatible.

Las imágenes SAR pueden ser adquiridas con distintos formatos. El que se va a utilizar aquí es el formato Ground Range Detected (GRD) 14m x 3m por conveniencia para este estudio. Este formato consiste en imágenes SAR multi-look (reducción de speckle) y proyectadas al rango de la tierra utilizando un modelo de elipsoide de la Tierra. La información de la fase es suprimida, lo cual no es un problema, ya que no era uno de los parámetros clave utilizados, y los píxeles que presenta la imagen son aproximadamente cuadrados [18].

El primer problema reconocible en la obtención de estas imágenes es su falta de correspondencia con las coordenadas geográficas comúnmente utilizadas, además de las dimensiones y orientación de estas, ya que las imágenes abarcan áreas de mucho mayor tamaño a las áreas de cultivos aquí estudiados, y la orientación no es totalmente paralela al eje polar de la Tierra, sino que presenta cierta inclinación. Para solucionar todo ello, se realiza un pre-procesado desarrollado por el Departamento de Física, Ingeniería de Sistemas y Teoría de la Señal de la Escuela Politécnica Superior, utilizando el software libre SNAP cedido por la ESA, que se divide en los siguientes pasos:

1. Lectura de las imágenes.
2. Actualización de la información orbital en las imágenes cargadas.
3. Cancelación de ruido térmico.
4. Recorte del área de interés.
5. Calibración radiométrica para obtener de salida de ambos canales (VH y VV) el formato σ_0 .
6. Filtrado de speckle.
7. Conversión de escala lineal a dB.
8. Geo-referenciación: genera un mapa en una rejilla uniforme de coordenadas cartográficas con tamaño de píxel elegido de aproximadamente 10 m para cada coordenada (latitud y longitud).
9. Escritura del producto en formato propio de SNAP: DEAM-DIMAP

Esto se realiza tantas veces como número de imágenes de distintas fechas se hayan obtenido. Para finalizar el procesamiento, se ajustan unas imágenes con otras para que todas estén referenciadas a las mismas coordenadas cartográficas en los mismos píxeles. Finalmente, estos píxeles tienen una resolución de 10m x 10m, debido a la degradación por el procesamiento. Una vez finalizado este proceso, las imágenes están preparadas para tratar su información de manera más sencilla.

2.3.2 Regresión aplicada a la estimación

Dentro del marco de trabajo previo, este TFG se va a centrar en elaborar un modelo de observación a partir de las imágenes SAR de satélite para cultivos de arroz, haciendo uso de la regresión de series temporales y técnicas de aprendizaje automático. Modelo que posteriormente se combina con el de evolución para la elaboración de estimaciones más completas y fiables. La regresión busca generar un modelo de predicción del fenómeno que se está estudiando a partir de una información de la que el fenómeno depende. En este caso, ese fenómeno es el estado fenológico de los cultivos de arroz y la información a partir de la cuál se va a generar este modelo son los coeficientes de backscattering obtenidos para las distintas polarizaciones (VV y VH), su ratio y la desviación estándar de cada uno, todo ello a nivel de media de parcela y día.

A la hora de crear un modelo a partir de regresión, se debe considerar la evolución que se quiere estudiar para elegir la información que se va a utilizar. Para el estudio de la fenología se debe escoger información que complete un periodo de desarrollo desde el primer estado de la siembra del cultivo, hasta la madurez y recogida del producto, información que debe presentar una resolución temporal suficiente para que sean distinguibles las distintas etapas dentro del proceso. A esto se suma que para la generación de un modelo fiable este tiene que ajustarse con un número determinado de ciclos enteros de información, en mayor o menor número dependiendo de la fiabilidad y constancia de los mismos.

Habiendo generado un modelo utilizando técnicas de regresión, este debe ser comparado con información contrastada de los cultivos que indiquen si el modelo se aproxima lo suficiente a la realidad del terreno y, por tanto, es un modelo fiable. Esto significaría que la información utilizada, a priori, tiene relación con el fenómeno estudiado, y que la cantidad de información ha sido suficiente para ajustar el modelo, por lo que puede ser utilizado para predicción, en este caso, del estado fenológico del nuevo cultivo observado.

En este modelo no se incluye la parte dinámica del sistema, ya que esta está representada en el modelo de evolución y, por tanto, en la unión de ambos modelos. Pero este modelo se genera únicamente a partir de los observables anteriormente mencionados. Las salidas de este modelo son funciones de densidad de probabilidad (PDF) del estado fenológico del cultivo combinables con las salidas del modelo de evolución.

3 Metodología

El desarrollo de este TFG va a consistir en la creación de un script de Python que tenga como entradas los datos SAR de Sentinel-1, y como salida la función de densidad de probabilidad para la BBCH y/o altura de la planta, utilizando RFR para realizar esta estimación.

3.1 Datos para análisis y modelado

Los datos con los que contamos para el estudio constan de 7 parcelas de cultivo de arroz llamadas: Calogne, Ermita, Puntal, Mínima, Puebla, Reboso y Vega, situadas en Sevilla, provincia del suroeste de España, mapa en 3.1. De ellas se disponen de los datos tanto de satélite como de mediciones en campo para los años 2017 y 2018.

Los datos de entrada de satélite de los que disponemos para estas 7 parcelas son los coeficientes de backscattering para polarizaciones VV y VH en unidades de dB, obtenidos de las imágenes SAR post-procesadas de los satélites Sentinel-1. De ellos se emplean estos dos datos en dB, su ratio y la desviación estándar de todos los anteriores. Los datos tienen un periodo de revista de Sentinel-1 es de 6 días, lo que aporta una resolución temporal de cada ciclo de desarrollo aceptable. Esta información presenta unas limitaciones temporales debido al reciente desarrollo del programa Copérnico. Aún así, parece válida para una aproximación suficientemente fiable, en comparación con los periodos de 12 días de los datos anteriores a 2016.

Para realizar el modelo de observación necesitamos contrastar con datos reales tomados en tierra de los mismos cultivos, por lo que se recogen también de los siguientes datos de las 7 parcelas: su posición geográfica, área, días de siembra y de cosecha, producción, BBCH total, mínima y máxima, la altura media del cultivo y los días del año para los cuales se han tomado estos datos. De todos ellos, los más relevantes para este estudio son la BBCH por parcela, la altura del cultivo, por su relación con el desarrollo del cultivo para este caso, y los días de siembra, cosecha y de toma de datos, los cuáles no tienen porqué coincidir con el periodo de revista de los satélites de Sentinel-1. Los periodos de cultivo del arroz de aproximadamente 4 meses, disponemos de una cosecha por año y parcela, por lo que por cada época de cultivo se dispone de hasta 7 ciclos de desarrollo del cultivo, valor no muy amplio. En el caso de no disponer de suficiente información, esta se podría generar de manera estadística, partiendo de la original.

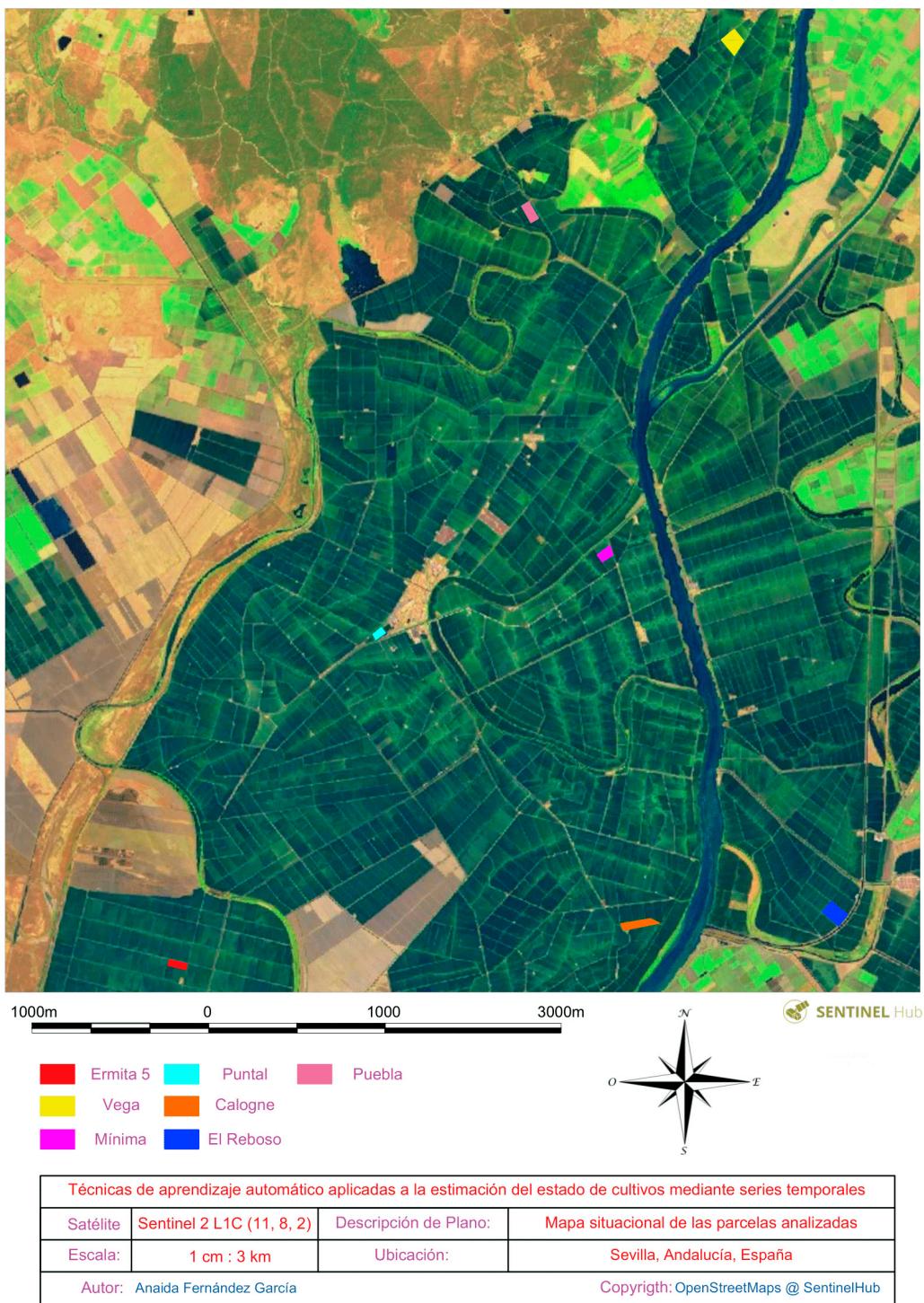


Figura 3.1: Mapa de la zona de estudio con las parcelas utilizadas destacadas. [1]

3.2 Procesamiento y división de datos

El procesamiento que los datos de satélite necesitan para la elaboración del modelo es distinto dependiendo del método utilizado. En este proyecto se realizan pruebas para la estimación de 3 casos de salidas distintas: BBCH, altura (cm) y BBCH junto a altura (cm). Además, cada uno de estos casos se evalúa utilizando los datos de entrada a nivel de parcela (mediante la media) y a nivel de pixel, desconociendo a priori qué método obtiene mejores resultados. Para ambos métodos, el procesamiento de los datos comienza restringiendo la información al periodo que nos interesa: desde el día de siembra hasta el de cosecha. A continuación, se ajustan los días de los que se tiene información realizando una interpolación para los datos de BBCH y/o altura, haciéndolos coincidir con las fechas de toma de datos del satélite, cuyos datos no deben interpolarse ya que su evolución no es tan creciente lineal como la altura o la fenología. El siguiente procesamiento se da únicamente para el método a nivel de parcela: se realiza la media y la desviación estándar de los datos que vamos a utilizar como entradas del sistema (VV en dB, VH en dB, y el ratio entre ambos, VH-VV, también en dB). Si se está trabajando con el dato de altura, además, se realiza una limpieza de datos corruptos, ya que para algunas fechas concretas faltan datos de tierra con los que contrastar.

Para finalizar la preparación de los datos, se dividen estos en sets de entrenamiento y de test. La división se realiza por parcelas completas, para que sea más sencillo y completo su entrenamiento y posterior visualización. Para todos los casos se reservan 6 parcelas de entrenamiento y 1 de test de resultados, seleccionando este número y las parcelas concretas que mejor entrena el sistema y optimizan los resultados. Los datos reales con los que se va a entrenar y examinar el modelo se preparan con la interpolación mencionada anteriormente y la división de parcelas que sigue los mismos requisitos que los datos de satélite. La única adaptación extra que tienen estos datos se da en el caso de nivel de pixel: como no contamos con la información medida en tierra a ese nivel de BBCH ni de altura, los valores generales para cada parcela interpolados deben ser asignados a cada uno de los píxeles correspondientes de esa parcela, es decir, todos los píxeles tendrán el mismo valor de salida para una misma parcela y día. Una interpretación gráfica aproximada de este procesamiento se puede ver en la figura *INSERTAR CROQUIS*.

Una vez procesados todos los datos y creados los sets de entrenamiento y test, estos datos pueden ser guardados y cargados para no repetir el procesamiento en futuras ocasiones.

3.3 Modelo de regresión y optimización

A continuación, la implementación del regresor, su entrenamiento y evaluación completa la creación del modelo de observación y este está listo para realizar predicciones. La técnica de regresión que se va a emplear es RFR, una técnica de aprendizaje automático supervisado basado en árboles de decisión como se ha visto anteriormente y que ha sido probada su eficacia en generaciones de modelos similares para imágenes SAR en 2019 [19]. En este caso, las salidas son los posibles valores de BBCH y/o la altura del cultivo, a los que se llega mediante decisiones sobre los rangos de valores que presentan los parámetros de entrada. Como todas las técnicas de aprendizaje automático, cuenta con una etapa de aprendizaje para la creación

del modelo, y una etapa de test para comprobar su correcto funcionamiento. Normalmente, para cada vector de entrada, este modelo devuelve una salida única que es el estado fenológico o la altura más probable, pero por motivos de integración con el modelo de evolución, va a presentar la probabilidad para cada estado.

En la creación del modelo existen distintos parámetros modificables según la técnica de RFR para optimizarlo acorde con las características del problema que se intenta resolver. El número de estimadores o número de árboles es uno de los principales parámetros, por defecto 100, el cual va representar el número de árboles de decisión de los que se va a componer el regresor. Mayor número de árboles implica una mayor complejidad y la posibilidad de generar soluciones más profundas y precisas, con el riesgo de hacer un sistema excesivamente complejo que tenga sobreajuste u overfitting en su entrenamiento y que tenga un costo computacional muy elevado sin realmente aportar mejoras significativas. Otros parámetros variables en la creación del regresor, ya dentro de los árboles de decisión, son la profundidad máxima (número máximo de nodos y niveles de cada árbol, por defecto nulo), el mínimo número de muestras para dividir un nodo interno (por defecto 2), el número mínimo de muestras para un nodo final (por defecto 1) o estado aleatorio inicial en la creación de los árboles de decisión (por defecto nulo), entre otros. Debido a las características de nuestro sistema, se mantienen los valores por defecto de todos los parámetros excepto el número de estimadores y el estado de aleatoriedad, parámetros que se destinan a la optimización del regresor ya que son los más influyentes en los resultados finales.

La optimización del número de estimadores se realiza ejecutando una prueba para todas las posibilidades entre los valores 1 y 1000 y escogiendo el valor mínimo de estimadores que presente un máximo local no puntual del error cuadrático, es decir, al que tiende de manera progresiva, y con una mejora considerable. La optimización del estado de aleatoriedad se realiza de la misma manera. Este último parámetro también garantiza una generación aleatoria similar para los mismos parámetros independientemente de las veces que se realicen las pruebas, esto es, la “semilla” de la que parte esta generación es la misma, por lo que se puede realizar una evaluación fiable de los resultados, ya que no van a depender de cómo se hayan generado inicialmente los árboles de decisión.

3.4 Evaluación de resultados

Los resultados generales obtenidos para cada uno de los casos mencionados anteriormente, una vez optimizados los datos utilizados y los parámetros del regresor, se deben contrastar y evaluar para determinar la fiabilidad del modelo. Las evaluaciones de estos resultados se realizan con indicadores de error como la media del error absoluto (Mean Absolute Error (MAE)), correspondiente a la fórmula 3.1, la raíz del error cuadrático medio (Root-Mean-Square Error (RMSE)), fórmula 3.2 o el coeficiente de determinación (R^2), fórmula 3.3 para las predicciones del set de test, siendo las salidas del modelo una predicción de valor único.

$$MAE = \frac{\sum_{i=1}^n |y_i - x_i|}{n} \quad (3.1)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - x_i)^2}{n}} \quad (3.2)$$

$$R^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \sigma_y^2} \quad (3.3)$$

Siendo y_i los valores de salida predichos por el sistema y x_i la verdad de tierra medida correspondiente para las fórmulas de MAE y RMSE. El coeficiente de determinación se calcularía según la fórmula 3.3 siendo σ_{xy} la covarianza de x e y , valores predichos y contrastados; respectivamente, σ_x^2 la varianza de x y σ_y^2 la varianza de y .

Además, se evalúa también la influencia que ha tenido cada parámetro de entrada en las salidas, esto es, saber cuáles son los que presentan una mayor correlación con la variable de salida del sistema y por tanto son más útiles para su predicción. También se realizan representaciones gráficas comparativas entre las predicciones y los datos de tierra, de donde se puede observar para qué intervalos de la variable de salida el modelo funciona mejor.

4 Resultados

4.1 Método por parcelas

4.1.1 Optimización

La optimización de los datos de entrada tiene su base en la elección de las variables de entrada y de los sets de parcelas de entrenamiento y test. Los mejores resultados para todos los casos el uso de un conjunto de 6 parcelas para entrenamiento y 1 para la evaluación. Con respecto a datos de entrada al sistema de los mencionados anteriormente: media por día y parcela de VV, VH y ratio VH/VV y la desviación estándar de cada uno de ellos, se utilizan todos, exceptuando el caso de predicción de la altura, donde solo se emplean los 3 primeros, ya que los demás empeoran los resultados. En cuanto a la optimización del regresor, se basa en la determinación del número de árboles que lo componen. La relación entre el número de árboles para un modelo y el coeficiente de determinación es un buen descriptor para la elección de este parámetro, como se presenta en la imagen de ejemplo 4.1 con el caso de salida BBCH. En ella, se puede ver que una vez alcanzado cierto nivel de coeficiente, la mejora de este en relación al aumento del número de árboles no es significativa con respecto al costo computacional y a la complejidad del sistema que se crea.

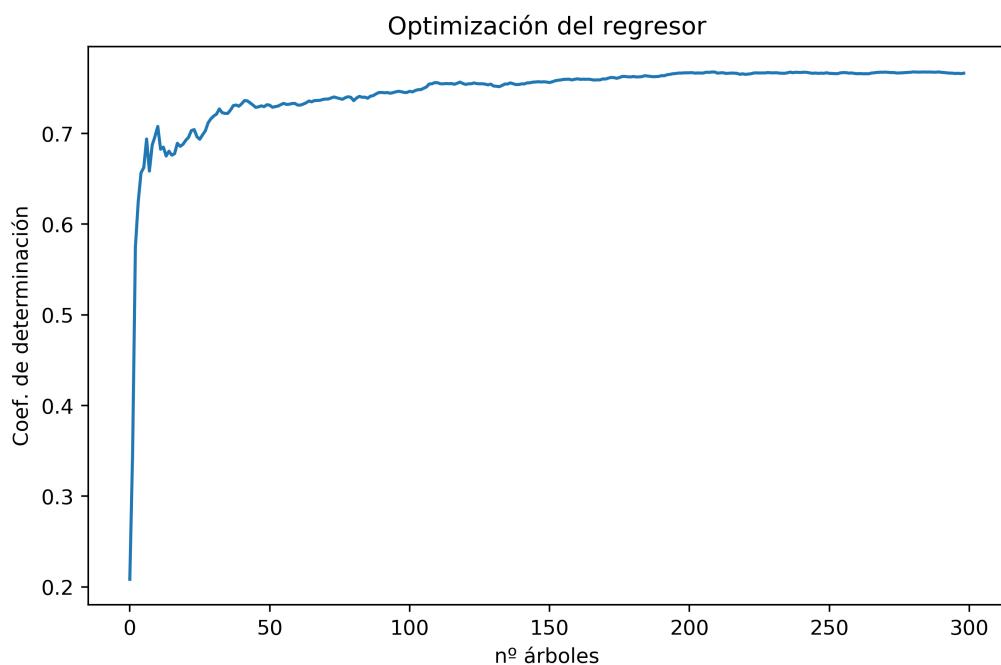


Figura 4.1: Optimización del número de árboles para RFR en el modelo de salida BBCH

Finalmente, los parámetros utilizados en este método para cada caso se presentan en la tabla 4.1, donde se pueden ver: la parcela utilizada para el periodo de test del modelo; siendo el resto de parcelas utilizadas en el entrenamiento, y el número de árboles óptimo para RFR, de acuerdo con la evolución del coeficiente de determinación para cada caso.

	BBCH	Altura	BBCH&Altura
Parcela de test	'Mínima'	'Puntal'	'Mínima'
Número de árboles	42	77	54

Tabla 4.1: Parámetros de optimización de entrada y modelo

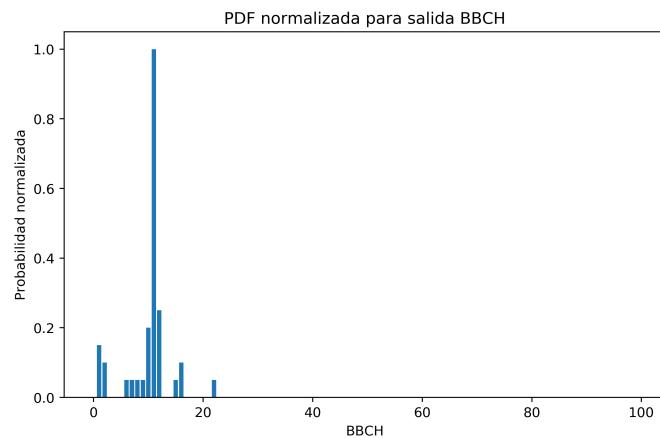
Como se puede observar, los 3 casos de este método constan de un número óptimo de árboles de, al menos, el mismo orden. Cabe destacar que el aumento de complejidad en el modelo y diferente parcela de test en el caso de la altura como salida del sistema. Probablemente se debe a la diferente adaptación a los datos de entrada disponibles, que son distintos a los otros dos casos.

4.1.2 Salidas del modelo

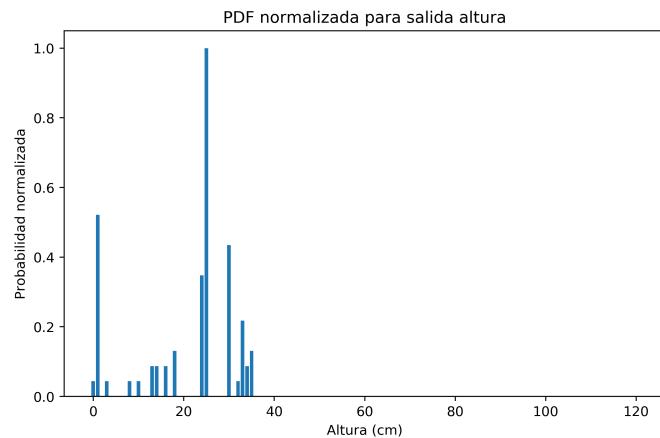
4.1.2.1 Salidas de función de densidad de probabilidad

Como ya se ha comentado, las salidas del modelo son por defecto predicciones únicas, pero el marco de trabajo anterior demanda salidas en formato PDF para ser integradas con el resto del modelo. En las figuras 4.2a (modelo de salida BBCH), 4.2b (modelo de salida altura) y 4.3 (modelo de ambas salidas) se puede apreciar cómo son algunas de estas salidas, siendo ejemplos extraídos a partir de los mismos datos para los 3 casos estudiados.

Las PDFs se presentan normalizadas con respecto a la salida con mayor probabilidad. En estos ejemplos se puede ver como, aunque hay una salida que predomina con respecto al resto, no se excluyen las demás soluciones posibles. Esto facilita la integración con el modelo de predicción temporal ya que se pueden combinar ambas salidas PDF para ver dónde coinciden y con qué probabilidad. En general, los resultados de la predicción temporal son bastante certeros, con oportunidad de fallo para ajustes finos en cultivos que hayan podido sufrir retrasos o adelantos en su desarrollo típico. Es ahí donde la contribución de este modelo es importante, aportando información extra en tiempo real y creando una predicción de cuáles serían los posibles estados de desarrollo en los que se encuentra el cultivo según la información actual de satélite.



(a) Ejemplo del modelo de doble salida: PDF estimación de BBCH

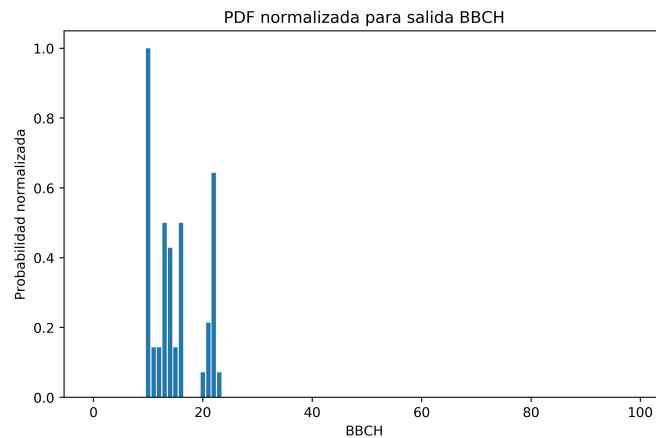


(b) Ejemplo del modelo de doble salida: PDF estimación de altura

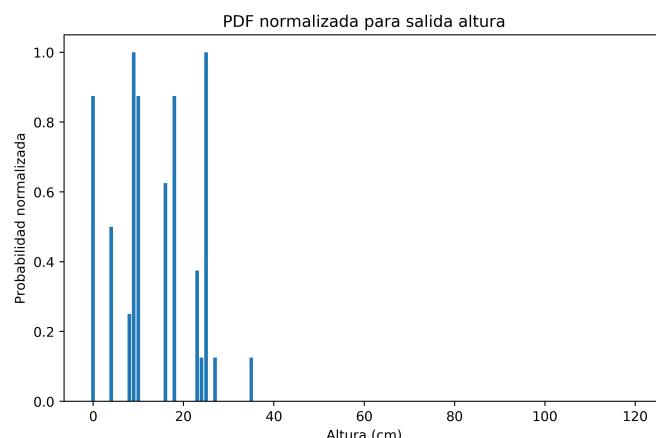
Figura 4.2: Ejemplo de salidas PDF normalizadas del modelo para estimación de BBCH y la altura.

Las 3 figuras 4.2a, 4.2b y 4.3 han sido generadas con los mismos datos de entrada, es decir, los mismos datos de satélite en la misma parcela y fecha, por lo que las salidas para los modelos generados de estimación de BBCH (Figura 4.2a) y altura (Figura 4.2b) como modelos independientes se pueden comparar con las salidas del modelo de estimación de ambas (Figura 4.3). La principal diferencia que encontramos entre los mismos tipos de datos de salida para cada modelo es que, en general, las salidas individuales presentan una estimación principal con una diferencia de probabilidad mucho mayor con respecto al resto que las estimaciones del modelo de doble salida. El rango para cada salida se mantiene bastante similar en ambos modelos, además del valor con la probabilidad más alta, aunque la disminución de diferencias con las demás estimaciones, sobre todo en el caso de la altura (4.3b), indica que ese sistema es menos preciso y estable. El hecho de que sea la salida de la altura la que se vea más perjudicada en este modelo de dos salidas puede deberse a que la optimización de este, en

cuanto a número de árboles del regresor, es más similar, y por lo tanto más beneficiosa, con respecto al caso individual de estimación de BBCH.



(a) Ejemplo del modelo de doble salida: PDF estimación de BBCH

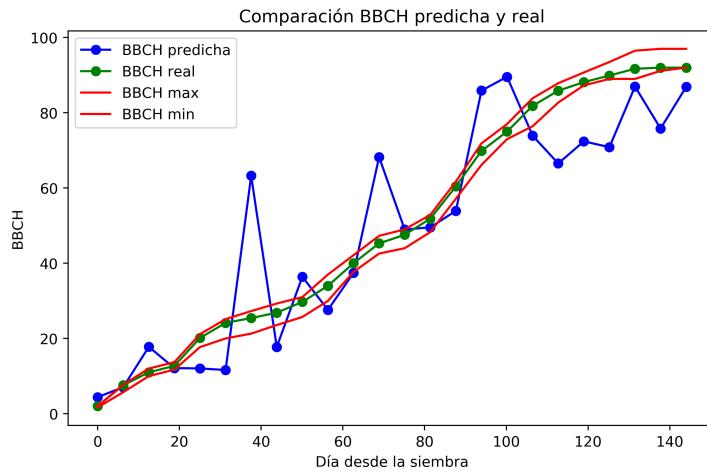


(b) Ejemplo del modelo de doble salida: PDF estimación de altura

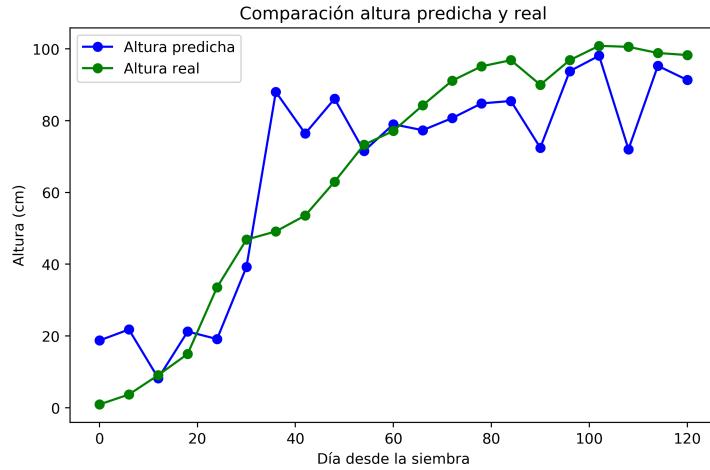
Figura 4.3: Ejemplo de salidas PDF normalizadas del modelo para estimación de BBCH y la altura.

4.1.2.2 Salidas de valor único

Para continuar con la presentación de los resultados obtenidos, se ilustran en las figuras 4.4a, 4.4b y 4.5 las comparaciones de las soluciones únicas obtenidas (salida con máxima probabilidad) y los datos de tierra tomados correspondientes para la parcela de test en 2018: BBCH general, máxima y mínima en la parcela, casos individual (4.4a) y de doble salida (4.5a), o la altura general del cultivo, también para ambos casos (figuras 4.4b y 4.5b, respectivamente). Estas salidas, aunque carecen de la información completa, son las más sencillas de comparar y evaluar con los resultados reales, ya que estos también son un único valor.



(a) Comparación de la salida predicha y la verdad de tierra del modelo para estimación de BBCH.

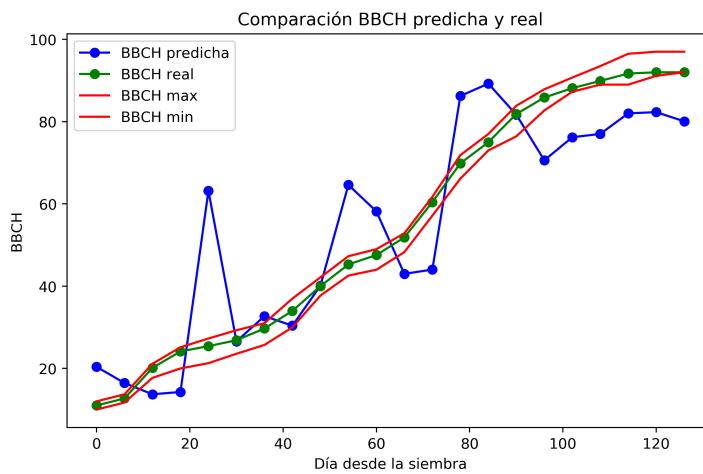


(b) Comparación de la salida predicha y la verdad de tierra del modelo para estimación de la altura.

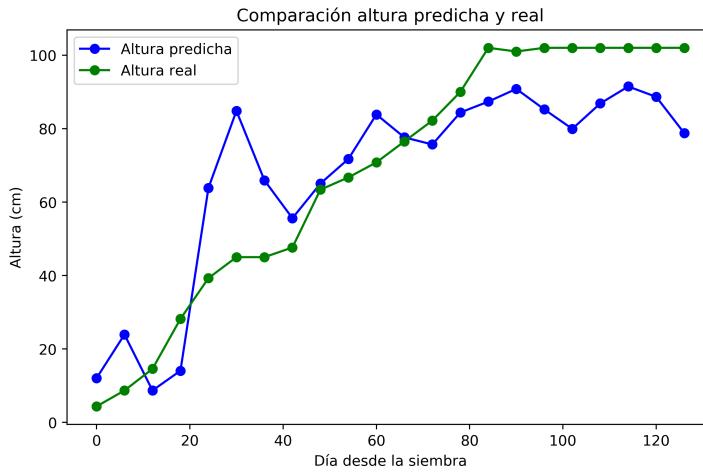
Figura 4.4: Comparación de la salida predicha y la verdad de tierra de los modelos de salida individual para la estimación de BBCH y la altura.

Comenzando por los modelos de predicción de BBCH representados en las figuras 4.4a y 4.5a, se pueden observar estimaciones y errores muy parecidos en ambos modelos. Estos son fácilmente comparables ya que la parcela de test coincide para los dos casos. Ambos modelos, presentan estimaciones que, en su mayoría, están fuera de los límites marcados por las líneas de BBCH máximas y mínimas medidas en campo. En general, se ve una tendencia creciente a lo largo de la variable estimada entorno a los valores reales, aunque con 2 o 3 zonas con picos de error mayor en las que el sistema predice un estado fenológico mayor al real. De hecho, el primer pico en torno a los días 30-40 después de la siembra está presente tanto en la estimación de BBCH como de la altura en las figuras 4.4b y 4.5b. Teniendo en cuenta que los valores de entrada son iguales para todos los casos, se puede atribuir este

fenómeno a anomalías presentes en los datos de satélite, los cuales, en una etapa concreta del desarrollo del cultivo, pueden presentar valores muy similares a los obtenidos para etapas finales, y por ello cometer un error de predicción bastante grande. Este error sería corregible bien detectando la anomalía y eliminando o modificando esta parte de los datos de entrada, tanto en entrenamiento como en test o futuros usos, o bien comprobando que el modelo está generando predicciones correctas con menor densidad de probabilidad, estando representadas en las salidas de PDF, y pudiendo ser consideradas sin un procesamiento extra de los datos, simplemente al contrastarlos con los resultados del modelo temporal, ya que es una diferencia de etapas muy grande y en este modelo no se daría un error de etapa de tal dimensión.



(a) Comparación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de BBCH.



(b) Comparación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de la altura

Figura 4.5: Comparación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de BBCH y la altura.

Visualizando las salidas para la altura, se puede también comparar, aunque no directamente, el funcionamiento de ambos modelos. En este caso las parcelas de test no son las mismas debido a la optimización para cada modelo, como tampoco los datos de entrada, por lo que la comparación simplemente visual entre las dos representaciones es menos intuitiva. Aún así, se aprecian coincidencias como el error en la etapa cercana a 40 días, como se ha mencionado antes, o la estimación de mayor altura para los días de 0 a 10 después de la siembra. En general, ambas representaciones tienen también una tendencia creciente bastante similar a la real, con aparentemente mejor estimación para las etapas finales en el modelo de salida única de altura, ya que en la figura 4.4b se observa una estimación menor constante para todas las etapas desde el día 80 tras la siembra hasta la cosecha (día 120 aproximadamente).

4.1.3 Evaluación de resultados

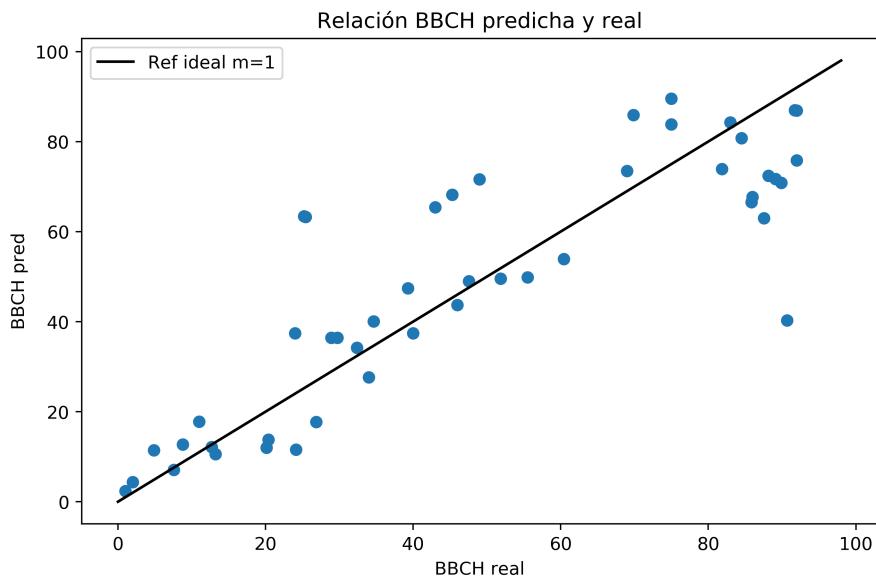
Para la evaluación de los resultados obtenidos para el método por parcelas se van a utilizar, como se ha mencionado anteriormente, las salidas de valor único por su facilidad para ser representadas y comparadas con los valores únicos medidos.

La primera evaluación realizada se basa en la relación directa entre las salidas estimadas y las medidas reales, lo cuál muestra la fiabilidad del modelo de una forma más clara y así puede verse en las figuras 4.6a, para el modelo de única salida de BBCH, 4.6b, para el modelo de única salida de altura y la figura 4.7 para el modelo de doble salida. En estas representaciones se añade, además, una recta de pendiente (m) unidad, la cual sirve de referencia ya que esa relación representaría un modelo ideal. Se han utilizado para todas las representaciones los valores contrastados de la parcela de test correspondiente a cada caso, teniendo en cuenta ambos años de datos, 2017 y 2018.

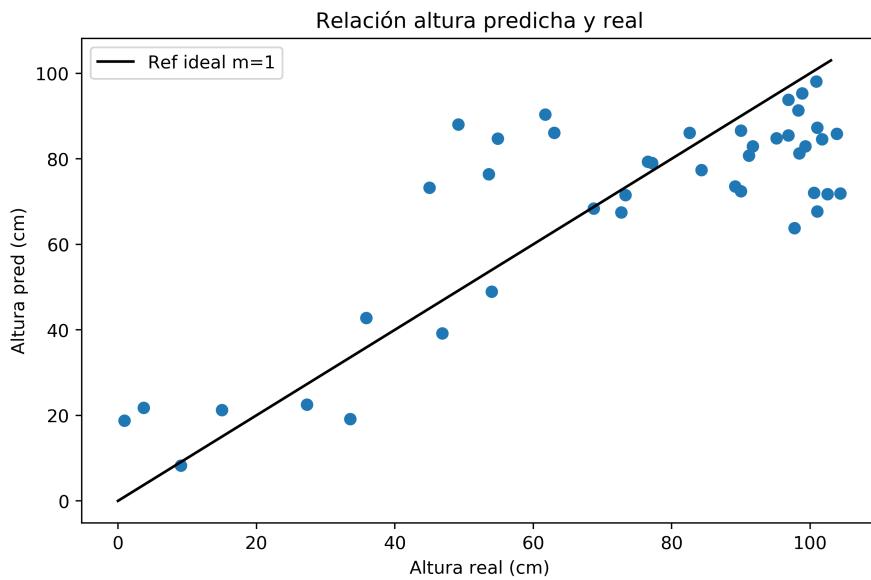
En la figura 4.6a se aprecia una distribución uniforme, lo cuál indica que se disponen de datos suficientes para cubrir las distintas etapas del desarrollo del cultivo. Además, las nubes de puntos se mantienen, en general, cercanos a la recta de referencia, por lo que, excepto para puntos concretos como el pico de las etapas tempranas y las últimas etapas del cultivo, se obtienen resultados bastante buenos. Comparando esta información con la obtenida para el modelo de dos salidas, representado en la figura 4.7a, se puede asegurar a simple vista qué modelo da mejores resultados. En esta última figura las nubes de puntos se presentan más dispersas que en el anterior, aunque siguiendo la tendencia de la recta de referencia. Vemos, contrariamente, agrupaciones de puntos y huecos, por lo que algunas etapas no están siendo estimadas correctamente. Esto se puede deber al hecho de que el procesamiento de este modelo incluye una etapa de limpieza de datos corruptos que se encuentran en los datos de altura, por lo que en algunas etapas, tanto para la BBCH como para la altura, faltan datos que aporten la información completa al sistema.

Comparando a continuación las figuras de los modelos con salida de altura del cultivo, 4.7 para salida única y 4.7b para salida doble, se observa en ambos una mayor inestabilidad en general debida a la concentración de puntos en algunas etapas. La altura de los cultivos no tiene un crecimiento tan lineal progresivo como la fenología de un cultivo, por lo que, a parte de por la corrección de datos, se encuentran concentraciones de datos debido a la estabilidad de la altura durante distintas etapas. Es por ello que, sobre todo en las etapas finales, hay una mayor concentración de puntos, cuando el arroz se mantiene en altura. Aún así, en la

figura 4.7 se aprecia que todas las estimaciones de altura tienden a la baja para casi todos los datos de altura mayor a 80cm. Una posible explicación sería una tendencia generalizada más baja en las parcelas utilizadas para entrenar el modelo o la poca variación en los parámetros de entrada para alturas desde 50cm hasta 110cm, intervalo en el que la altura estimada se mantiene en torno a la misma franja de valores.

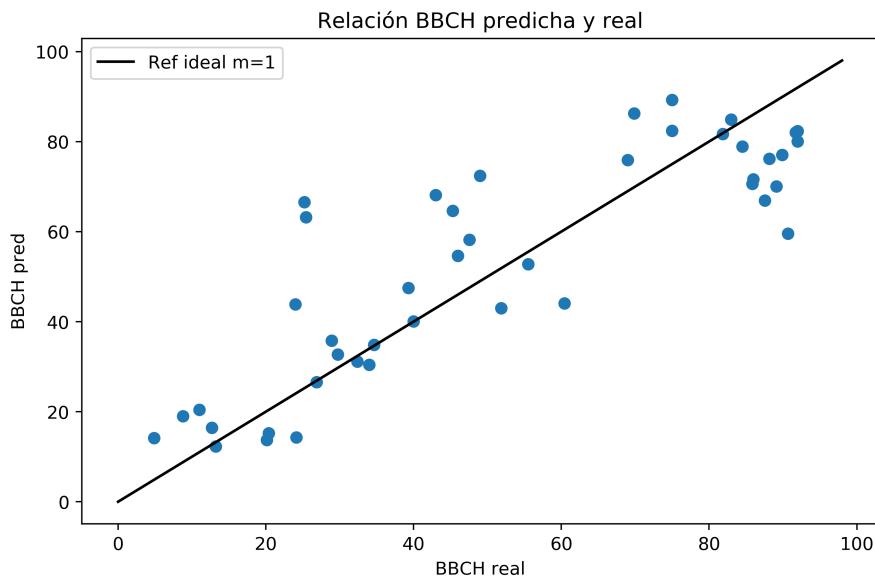


(a) Relación de la salida predicha y la verdad de tierra del modelo para estimación de BBCH.

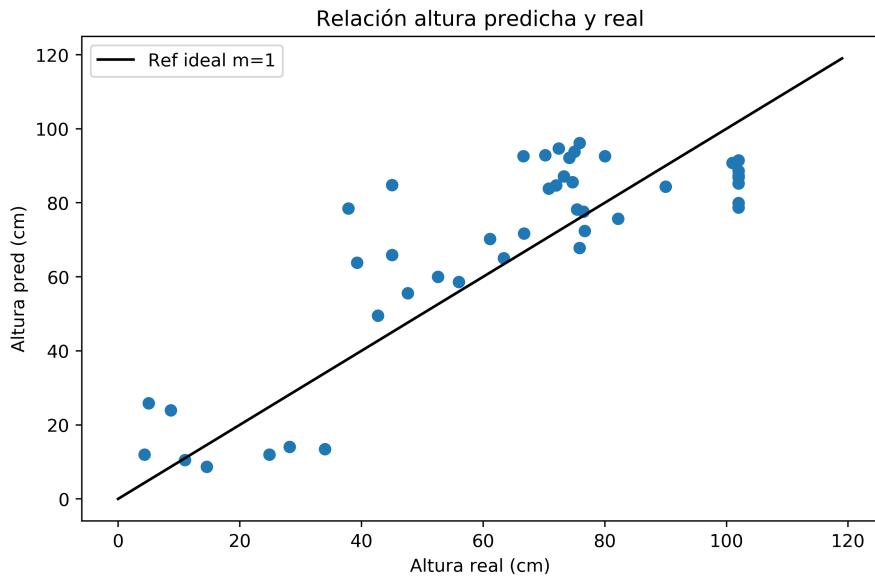


(b) Relación de la salida predicha y la verdad de tierra del modelo para estimación de la altura.

Figura 4.6: Relación de la salida predicha y la verdad de tierra de los modelos de variables independientes.



(a) Relación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de BBCH.



(b) Relación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de la altura

Figura 4.7: Relación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de BBCH y la altura.

A parte de la evaluación con descriptores estadísticos que se realizan a la hora de comparar los dos métodos, se realiza también una evaluación de los datos de entrada utilizados. La tabla 4.2 representa el peso que tiene cada una de las variables de entrada en el modelo generado para cada uno de los 3 casos tratados.

	BBCH	Altura	BBCH&Altura
VV	0.12	0.17	0.12
VH	0.52	0.69	0.53
Ratio VH/VV	0.11	0.14	0.10
Dev est. VV	0.09	-	0.10
Dev est. VH	0.07	-	0.09
Dev est. ratio	0.07	-	0.06

Tabla 4.2: Influencia de los parámetros de entrada en el modelo de estimación.

En la tabla 4.2 se presentan proporciones altamente similares para los 2 casos con las mismas entradas, y proporcionalmente parecidos para el caso de predicción de la altura. El parámetro de entrada que destaca principalmente sobre el resto es el coeficiente de backscattering para la polarización VH, con más del 50% de la predicción elaborada en base a él. Los siguientes parámetros más influyentes son el coeficiente de backscattering para la polarización VV y el ratio de ambos, y, por último, las desviaciones estándar, que suponen menos del 10% de la predicción final. Para comprender mejor porqué existen estas diferencias, se representan en la figura 4.8 la relación entre cada una de las variables con BBCH.

Lo primero que se observa en la figura 4.8 es que todas las subfiguras, en general, presentan una tendencia creciente conforme BBCH va creciendo, que a los primeros estados corresponden valores más bajos y que en todas aparece el pico de detección entorno a un valor de BBCH de 20-30, que se asemeja a los valores obtenidos por los mismos parámetros para BBCH cercanas a la cosecha. Esto confirma que las predicciones erróneas cercanas a estas etapas vienen dadas por esta anomalía a la que se le podría buscar corrección.

Comparando los valores que se obtienen para los distintos parámetros a partir de una BBCH de 40, se aprecia en la subfigura 4.8b que su evolución es la más estable y además creciente. Esto quiere decir que, al contrario que en las subfiguras 4.8a y 4.8c, no hay oscilaciones en los valores de entrada y, por tanto, la predicción a partir de estos valores es más sencilla que si para un mismo valor de entrada pueden corresponder distintas BBCH o alturas. Además, estas etapas finales en la subfigura 4.8b se presentan estables pero no constantes, que sería el peor caso que se puede hallar, ya que no aportaría ninguna distinción para intervalos de BBCH muy amplios.

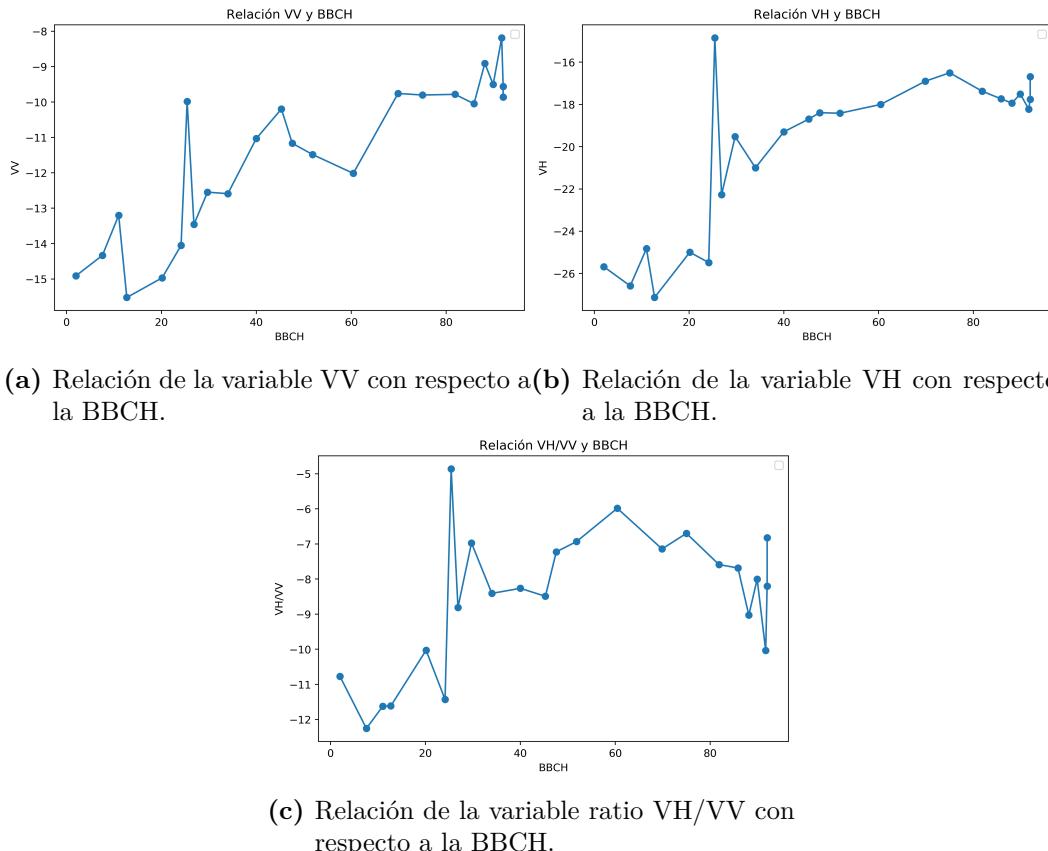


Figura 4.8: Relación de las variables de datos de entrada al modelo con respecto a la BBCH

4.2 Método por píxeles

4.2.1 Optimización

La optimización de los datos de entrada para este método se realiza de igual manera que el anterior, coincidiendo los mejores resultados para todos los casos en el uso de un conjunto de 6 parcelas para entrenamiento y 1 para la evaluación y los datos de entrada, se reducen a los tres principales (VV, VH y ratio), ya que la desviación estándar no aporta una gran mejora para este método. En cuanto a la optimización del regresor, se utiliza también la relación entre el número de árboles para un modelo y el coeficiente de determinación, presentado en la imagen de ejemplo 4.9 con el caso de salida BBCH. En ella, se puede ver que una vez alcanzado cierto nivel de coeficiente, la mejora de este en relación al aumento del número de árboles no es significativa con respecto al costo computacional y a la complejidad del sistema que se crea.

Finalmente, los parámetros utilizados en este método para cada caso se presentan en la tabla 4.3, donde se pueden ver: la parcela utilizada para el periodo de test del modelo; siendo el resto de parcelas utilizadas en el entrenamiento, y el número de árboles óptimo para RFR, de acuerdo con la evolución del coeficiente de determinación para cada caso.

Como se puede observar, los 3 casos coinciden en el set de parcelas de entrenamiento y

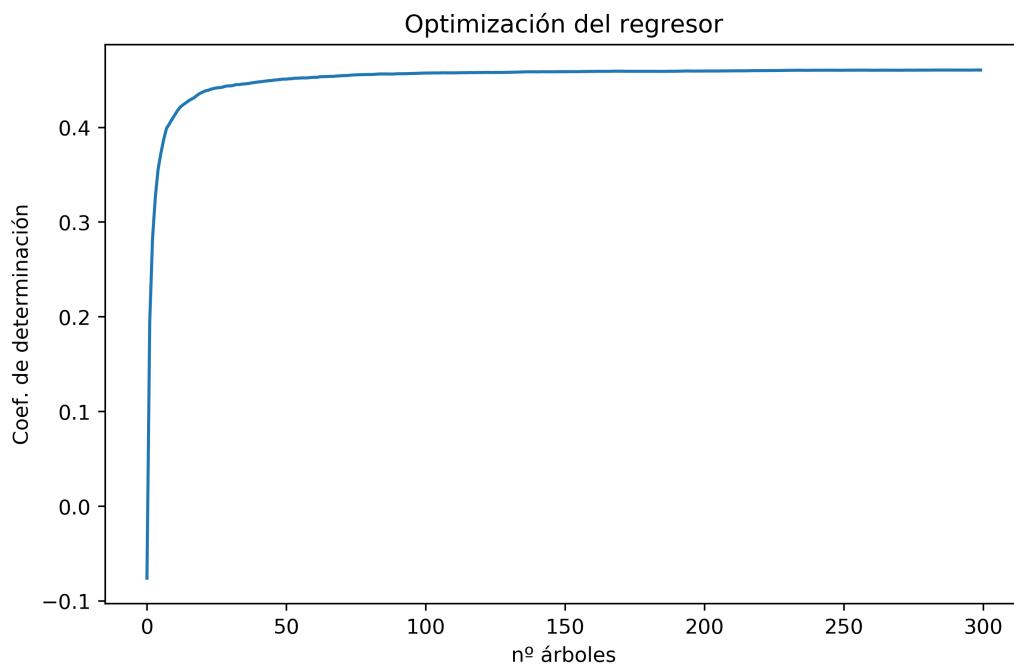


Figura 4.9: Optimización del número de árboles para RFR en el modelo de salida BBCH

	BBCH	Altura	BBCH&Altura
Parcela de test	‘Mínima’	‘Mínima’	‘Mínima’
Número de árboles	46	48	47

Tabla 4.3: Parámetros de optimización de entrada y modelo

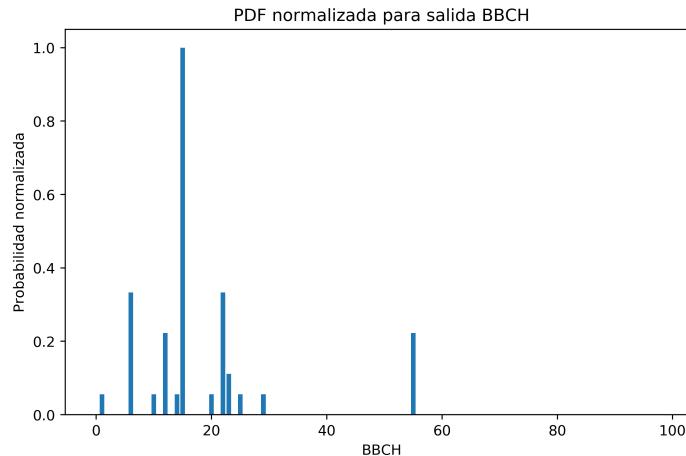
test con el que se obtienen mejores resultados. Probablemente esto se debe a anomalías en el desarrollo de algunas parcelas, las cuales, si se toman como set de test, el modelo no las habría podido tener en cuenta en el aprendizaje y el error en la predicción sería mayor. Los 3 casos de este método constan de un número óptimo de árboles muy similares, con una diferencia de 1 y 2 con respecto al caso menor. Cabe destacar que el aumento de complejidad, aunque escaso, en el modelo de salida de la altura, como ocurre para el método anterior. Además, tanto para la metodología por píxeles como por parcelas, se obtiene un número óptimo de árboles para el caso de 2 salidas intermedio a los óptimos para las mismas salidas en modelos independientes. Esto se debe a una compensación en la estimación de ambas variables, para que una variable sea óptima sin que su mejora sea a costa de la otra se llega a un valor intermedio en el que ninguna de las salidas está totalmente optimizada pero tienen el mejor resultado del sistema compartido completo.

4.2.2 Salidas del modelo

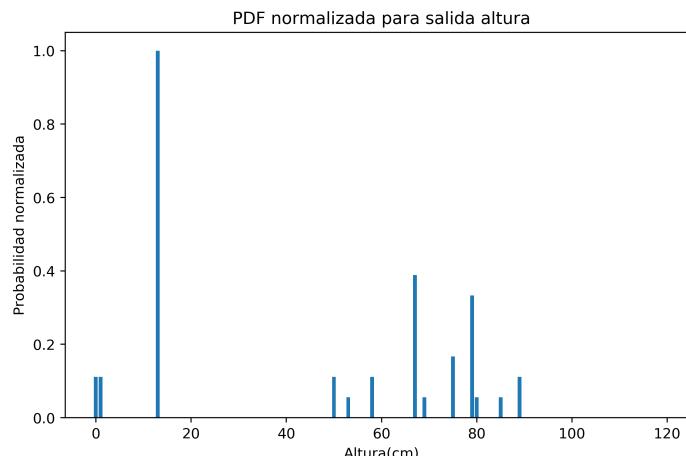
4.2.2.1 Salidas de función de densidad de probabilidad

Como se ha mencionado en la metodología anterior, las salidas útiles de este trabajo son las PDF. En las figuras 4.10a (modelo de salida BBCH), 4.10b (modelo de salida altura)

y 4.11 (modelo de ambas salidas) se puede apreciar cómo son algunas de estas salidas en esta metodología, siendo ejemplos extraídos a partir de los mismos datos para los 3 casos estudiados.



(a) Ejemplo de salida PDF normalizada del modelo para estimación de BBCH.



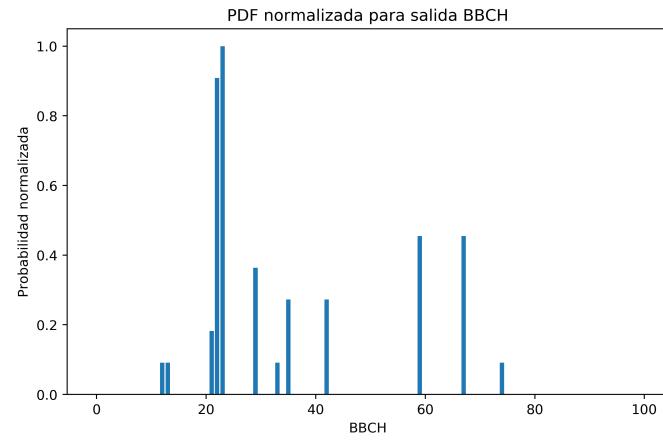
(b) Ejemplo de salida PDF normalizada del modelo para estimación de la altura.

Figura 4.10: Ejemplo de salida PDF normalizada los modelos para estimación de las variables independientes.

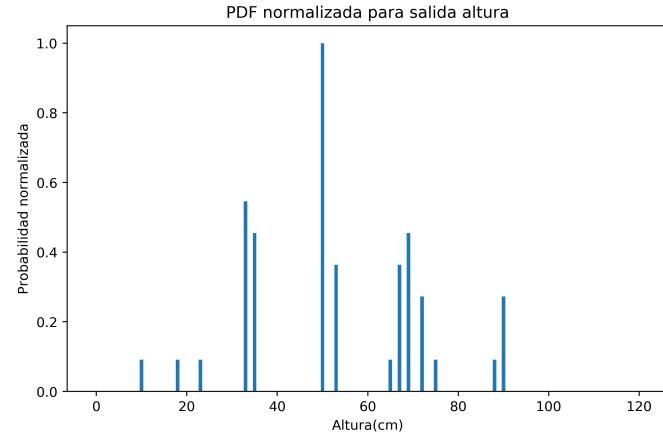
Estas representaciones tienen las mismas características mencionadas anteriormente: están normalizadas con respecto a la salida con mayor probabilidad, destaca una salida predominante pero no se excluyen el resto a la hora de integrarse con el modelo de predicción temporal.

Las figuras 4.10 y 4.11 han sido generadas con los mismos datos de entrada, es decir, los mismos datos de satélite en la misma parcela y fecha, por lo que las salidas para los

modelos generados de estimación de BBCH (Figura 4.10a) y altura (Figura 4.10b) como modelos independientes se pueden comparar con las salidas del modelo de estimación de ambas (Figura 4.11). La principal diferencia que se vuelve a encontrar entre los mismos tipos de datos de salida para cada modelo es que, las salidas individuales presentan una estimación principal con una diferencia de probabilidad mayor con respecto al resto que las estimaciones del modelo de doble salida. El rango para cada salida es bastante más amplio y disperso para el modelo de doble salida, esto junto a la disminución de diferencias con las demás estimaciones, indica que ese sistema es menos preciso y estable. En este modelo de dos salidas no se ve una clara optimización para una de ellas con respecto a la otra, ya que además ambas tiene valores muy cercanos al sistema individual.



(a) Ejemplo del modelo de doble salida: PDF estimación de BBCH



(b) Ejemplo del modelo de doble salida: PDF estimación de altura

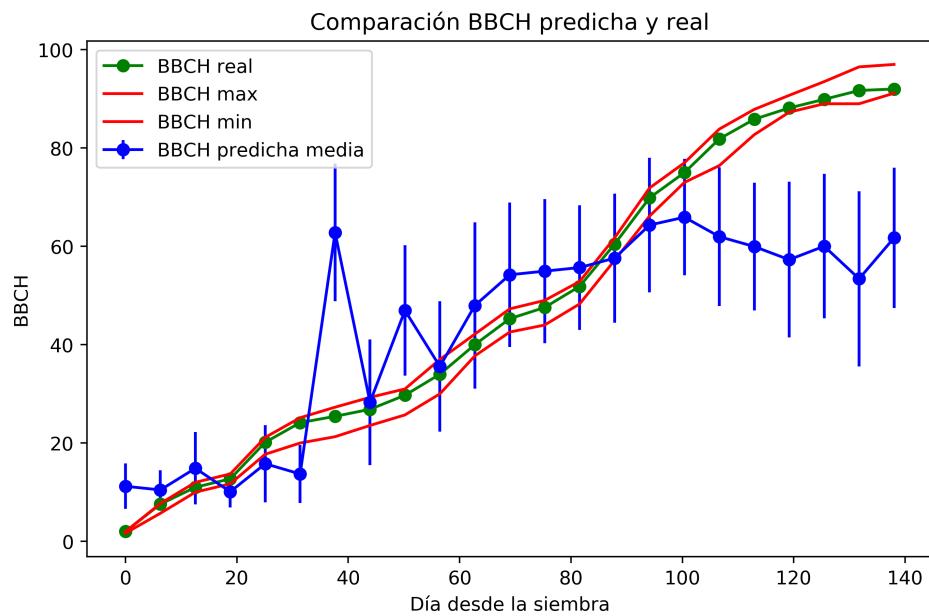
Figura 4.11: Ejemplo de salidas PDF normalizadas del modelo para estimación de BBCH y la altura.

4.2.2.2 Salidas de valor único

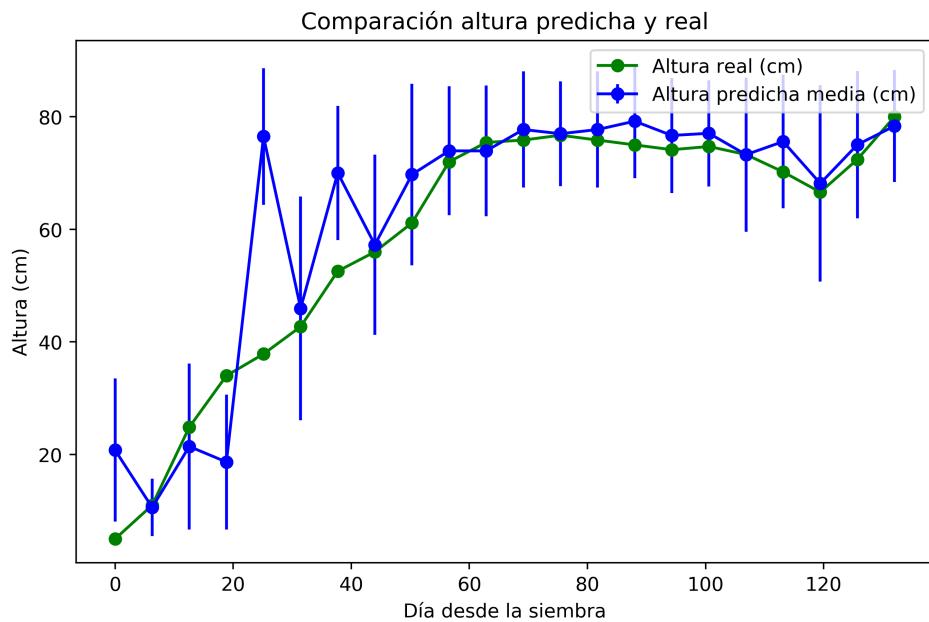
Para continuar con la presentación de los resultados obtenidos, se ilustran en las figuras 4.12a, 4.12b y 4.13 las comparaciones de las soluciones únicas obtenidas (salida con máxima probabilidad) y los datos de tierra tomados correspondientes para la parcela de test en 2018: BBCH general, máxima y mínima en la parcela, casos individual (4.4a) y de doble salida (4.13a), o la altura general del cultivo, también para ambos casos (figuras 4.12b y 4.13b, respectivamente). Se ha realizado una adaptación en los datos para poder representarlos de esta manera, ya que aquí la estimación de valor único se realiza a nivel de pixel y no de parcela, como los datos para contrastar. Es por ello que, para una representación y comprensión más sencillas, se representa la media y la desviación estándar para los píxeles a nivel de parcela y día.

En este método ambos modelos son comparables para las dos salidas posibles ya que todos ellos comparten la misma parcela de test, por lo que se compara la eficiencia para los mismos datos. Comenzando por los modelos de predicción de BBCH representados en las figuras 4.12a y 4.13a, se pueden observar estimaciones y errores de nuevo muy parecidos. Ambos modelos presentan una tendencia creciente, que estima a la baja la fenología para las últimas etapas del desarrollo y los mismos picos de predicción errónea sobre los días desde la siembra de 30-40 ya comentados anteriormente. Exceptuando las últimas etapas, se aprecian medias bastante cercanas a los valores reales, aunque con unas desviaciones estándar muy grandes a partir de los 40 días desde la siembra. Estas desviaciones se deben tanto a diferencias reales en la parcela, áreas con diferente nivel de desarrollo, como a un mal ajuste del modelo ya que en el entrenamiento existían estas diferencias en la misma parcela y se ha considerado el mismo valor de BBCH por la disponibilidad de datos de verdad de tierra. En esta metodología no aparecen diferencias tan claras entre los dos modelos implementados, ya que eran bastante similares.

Considerando las salidas para la altura, también se puede apreciar la gran similitud que presentan las representaciones comparadas: figuras 4.12b y 4.13b, difícilmente diferenciables a simple vista. Las coincidencias más obvias son las mencionadas con la estimación de la BBCH: la tendencia creciente, los picos de predicciones erróneas para las mismas etapas y grandes desviaciones estándar, en este caso, durante todo el proceso. Hay que puntualizar, que las medias para cada parcela tienen valores muy próximos a los reales, sobretodo en las etapas intermedias y finales, cuando la mayoría de los modelos presentan mayor inexactitud. Aunque los resultados globales evaluados para la metodología a nivel de pixel puedan ser numéricamente peores debido a la evaluación independiente de cada pixel, este es el método que mejores resultados ha obtenido para la estimación de la altura.

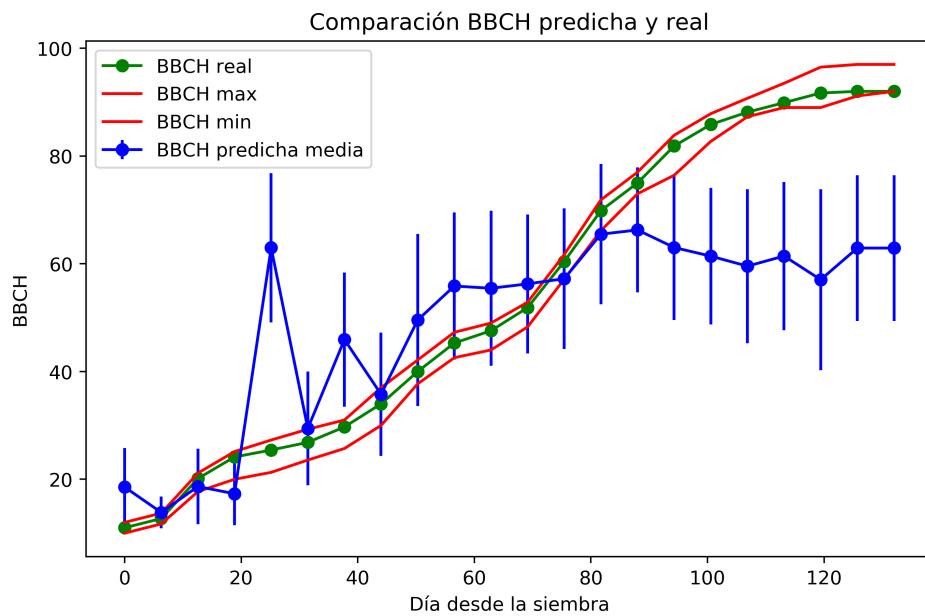


(a) Comparación de la salida predicha y la verdad de tierra del modelo para estimación de la BBCH.

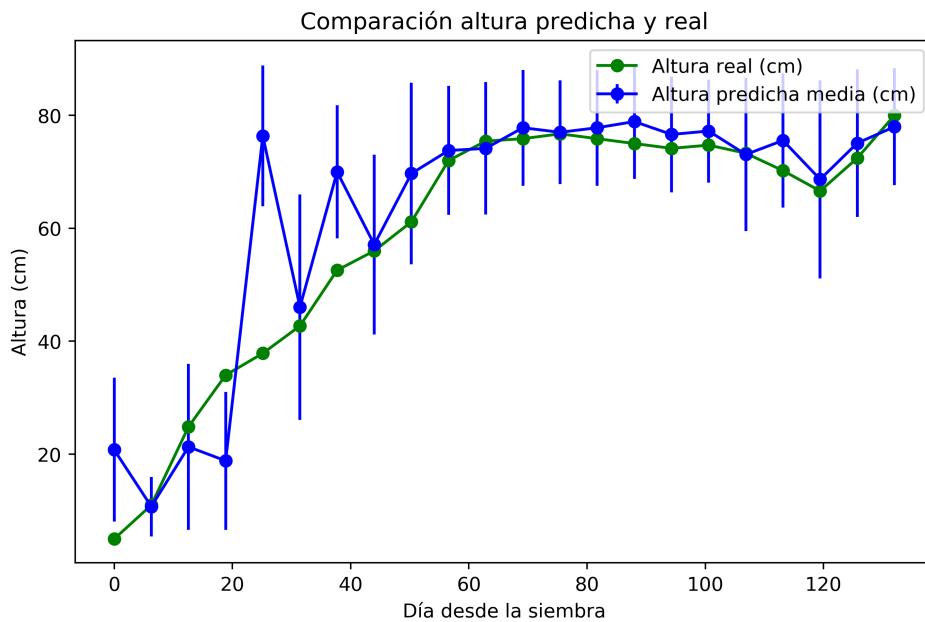


(b) Comparación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de la altura

Figura 4.12: Comparación de la salida predicha y la verdad de tierra de los modelos de salidas individuales.



(a) Comparación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de BBCH.

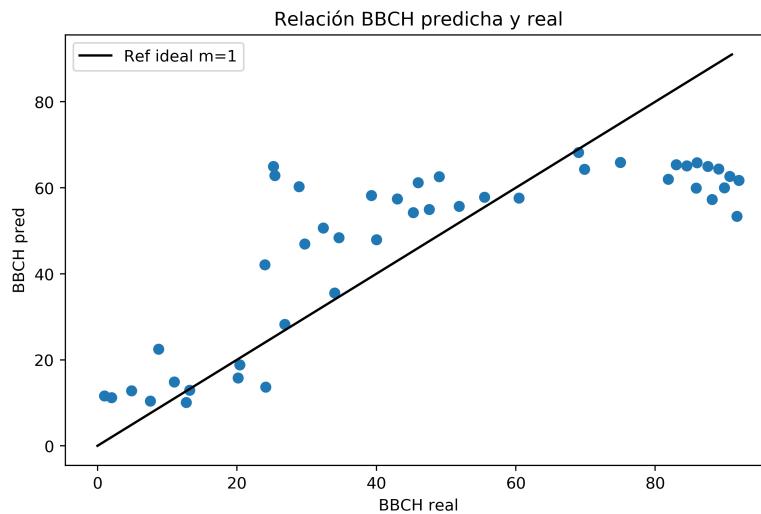


(b) Comparación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de la altura

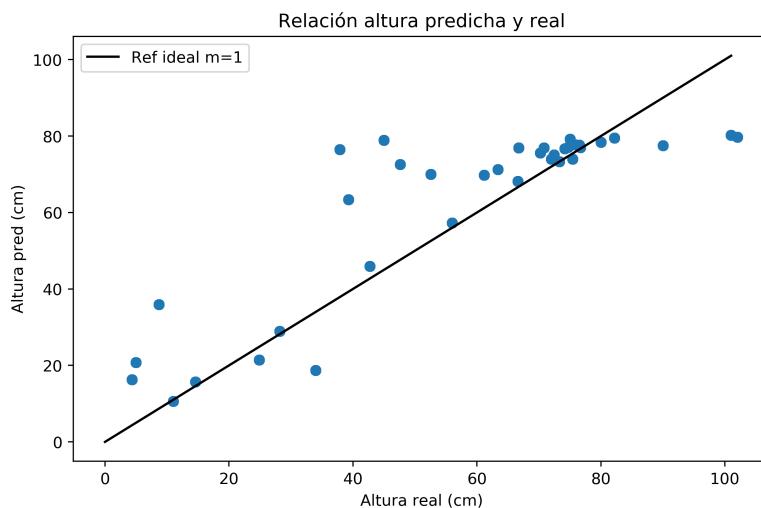
Figura 4.13: Comparación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de BBCH y la altura.

4.2.3 Evaluación de resultados

Para la evaluación de los resultados obtenidos para el método por píxeles se van a utilizar, como se ha mencionado anteriormente, las salidas de valor único por su facilidad para ser representadas y comparadas con los valores únicos medidos.



(a) Relación de la salida predicha y la verdad de tierra del modelo para estimación de BBCH.

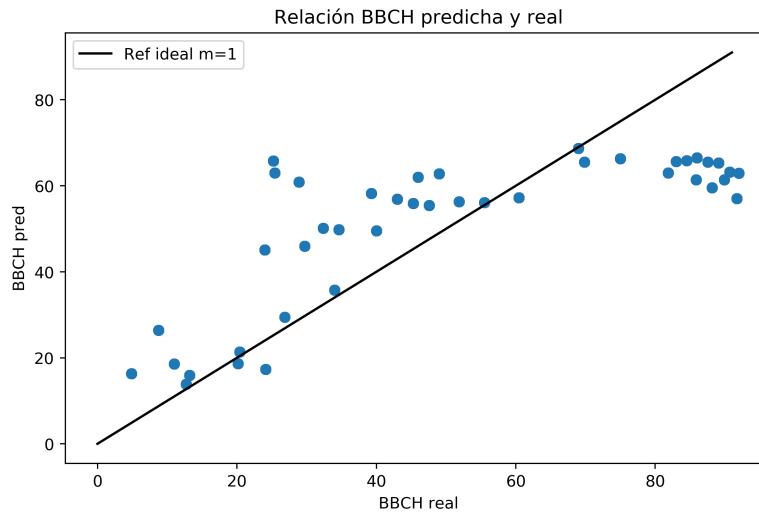


(b) Relación de la salida predicha y la verdad de tierra del modelo para estimación de la altura.

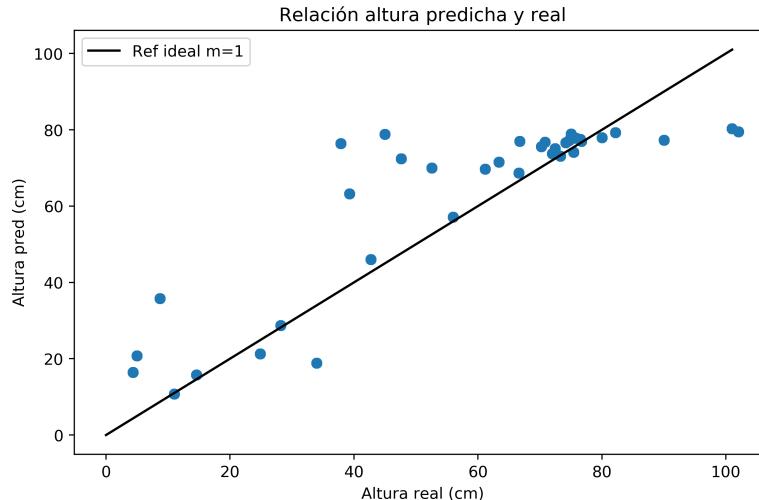
Figura 4.15: Relación de la salida predicha y la verdad de tierra de los modelos de salidas independientes.

Se realiza la primera evaluación, la relación directa entre las salidas estimadas medias por día y las medidas reales, en las figuras 4.14a, para el modelo de única salida de BBCH, 4.14b,

para el modelo de única salida de altura y la figura 4.16 para el modelo de doble salida. Se incluye en ellas, igualmente, la recta de referencia de pendiente (m) unidad, como modelo ideal. Se han utilizado para todas las representaciones los valores contrastados de la parcela de test correspondiente a cada caso, teniendo en cuenta ambos años de datos, 2017 y 2018.



(a) Relación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de BBCH.



(b) Relación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de la altura

Figura 4.16: Relación de la salida predicha y la verdad de tierra del modelo de doble salida para la estimación de BBCH y la altura.

En todas las figuras se aprecia una distribución uniforme, lo cuál indica que se disponen de datos suficientes para cubrir las distintas etapas del desarrollo del cultivo. Solo se encuentran

aglomeraciones de puntos en las figuras 4.14b y 4.16b, las correspondientes a la variable altura, donde en las últimas etapas se mantiene una altura entorno a 70cm por más tiempo, por lo que se tienen más muestras. Además, las nubes de puntos se mantienen, en general, cercanos a la recta de referencia, exceptuando el pico de las etapas tempranas, presente en todos los casos, y las últimas etapas del cultivo, para las que la estimación es de nuevo inferior a la real, viéndose este error de predicción más severo para las figuras de BBCH, 4.14a y 4.16a. Se sigue apreciando también en estas representaciones que la estimación de la altura para las etapas finales obtiene los mejores resultados.

Las diferencias entre los resultados expuestos en la figura 4.15 y 4.16, casos de salida individual y doble respectivamente, son difícilmente apreciables por la similitud de los modelos, y por tanto, de sus salidas. Se puede concluir entonces, que ambos modelos son eficientes y útiles en las estimaciones de BBCH y alturas, ya sea de manera individual o conjunta.

A parte de la evaluación con descriptores estadísticos que se realizan a la hora de comparar los dos métodos, se realiza una evaluación de los datos de entrada utilizados. La tabla 4.4 representa el peso que tiene cada una de las variables de entrada en el modelo generado para cada uno de los 3 casos tratados.

	BBCH	Altura	BBCH&Altura
VV	0.19	0.16	0.19
VH	0.61	0.61	0.59
Ratio VH/VV	0.19	0.23	0.22

Tabla 4.4: Influencia de los parámetros de entrada en el modelo de estimación.

Aunque aquí se encuentran algunas diferencias, más que en la metodología anterior, todas son muy similares. El parámetro con más peso siempre es el coeficiente de backscattering VH, seguido por las otras dos variables aproximadamente igualadas. En esta metodología el parámetro principal tiene más peso, entorno a un 60%, debido bien a la ausencia de las variables de entrada de desviación estándar, o bien al análisis a nivel de pixel, que obtenga mejores resultados gracias a esta variable. La razón por la que este parámetro funciona mejor como estimador se ha visto y comentado con las representaciones de la figura 4.8, totalmente válidas para este caso ya que las relaciones entre los datos de entrada y, por ejemplo, la BBCH son iguales.

4.3 Comparativa de métodos

La evaluación general de resultados realizada es la comparación entre los dos métodos de procesamiento de datos de entrada mencionados: a nivel de parcela o de pixel. En las siguientes tablas se pueden ver la evaluación de los resultados, divididos en entradas a nivel de parcela 4.5 y a nivel de pixel 4.6, según los índices estadísticos de MAE, RMSE y el coeficiente de determinación (R^2). Ambos conjuntos corresponden al mejor caso de cada método, esto es, habiendo seleccionado sus parámetros de entrada óptimos en cuanto a número de variables y set de parcelas de entrenamiento, y habiendo optimizado también los parámetros del regresor.

Datos de entrada a nivel de parcela			
	BBCH	Altura	BBCH&Altura
R^2	0.74	0.65	0.70
RMSE	15.39	17.87	15.69
MAE	11.13	14.21	12.63

Tabla 4.5: Índices estadísticos de las predicciones con datos de entrada a nivel de parcela.

Datos de entrada a nivel de pixel			
	BBCH	Altura	BBCH&Altura
R^2	0.45	0.52	0.45
RMSE	22.23	19.89	21.17
MAE	17.26	15.64	16.58

Tabla 4.6: Índices estadísticos de las predicciones con datos de entrada a nivel de pixel.

Como se puede observar, para las 3 posibles salidas para las que se han diseñado los modelos, el método que emplea datos de entrada a nivel de parcela consigue notablemente mejores resultados. Se obtienen valores de coeficiente de determinación más cercanos a 1, valor ideal, apreciándose en el modelo de predicción de BBCH una mejora para el caso de parcelas de un 64.4% con respecto al de pixel. Lo mismo ocurre en los índices de error, cuyo valor óptimo es 0: se obtienen valores inferiores para el método a nivel de parcela en todos los casos estudiados.

Los mejores resultados obtenidos para el procesamiento a nivel de parcelas pueden recaer en que todos los datos recogidos de verdad de tierra se presentan por parcelas, por lo que no existen unos datos para contrastar cada pixel extraído de la información de satélite con el terreno. Dentro de cada parcela el cultivo no tiene porqué desarrollarse de manera homogénea, pero la información recibida de este consta del valor de BBCH (con el requisito de ser el correcto para al menos el 50% del cultivo) y la altura generalizados por parcela, además de los máximos y mínimos de cada uno, por lo que las variaciones que se puede hallar en distintas zonas del cultivo no se tienen en cuenta. Por ello, el método de pixel implementado asigna un solo valor de salida para conjuntos de píxeles de la misma parcela, los cuales pueden estar representando unos niveles de BBCH o altura distintos, y esto conlleva un ajuste del modelo erróneo y, por tanto, peores índices estadísticos de evaluación. Aún así, este método no ha sido descartado, ya que el análisis a nivel de pixel es de gran interés para detectar esas variaciones de desarrollo de un cultivo dentro de una misma parcela.

5 Conclusiones

A rasgos generales, este TFG ha cumplido con los objetivos propuestos: se ha desarrollado un modelo de observación para la estimación del desarrollo de los cultivos utilizando imágenes SAR de satélite y un modelo regresor de aprendizaje automático, RFR. Además de cumplimentar el objetivo general de obtener un modelo, se han obtenido resultados considerablemente útiles para la integración con el modelo de predicción temporal anterior perteneciente al marco de trabajo marcado. Se han llegado a estos resultados gracias a la diversidad de pruebas realizadas tanto en metodología como en variables a estimar, lo cual ha aportado riqueza en las evaluaciones de resultados. Aunque, en general, los resultados de las estimaciones a nivel de media hayan obtenido mejor evaluación, la metodología de pixel aporta gran información sobre los rangos de la parcela y, en un futuro estudio, puede ser muy interesante para diferenciar zonas dentro de una misma parcela con distinto nivel de desarrollo, y poder detectar anomalías y actuar para corregirlas.

Dentro de este TFG se pueden ampliar algunos campos para profundizar en una línea de trabajo futura o, simplemente, para obtener mejores resultados en la investigación actual. Un ejemplo de esto sería la optimización de generación del modelo, ya que aquí se han tenido en cuenta 2 parámetros principales de RFR, pero python permite la variación de muchos otros más complejos y menos intuitivos que se han mantenido por defecto y pueden llevar a un estudio completo. Otra implementación que se podría llevar a cabo para la mejora del modelo sería la obtención de más datos de entrada, ya fueran datos reales obtenido por el paso de más tiempo o datos estimados a partir de los ya disponibles. Esto haría la etapa de aprendizaje más completa y generalizaría el modelo para distintos comportamientos de desarrollo. Siguiendo en esta línea, un caso ideal de implementación sería la generación de modelos de observación y predicción temporal enfocados a cada parcela individualmente. La generación de modelos en este caso sería mucho más personalizada, ajustándose a las características propias de cada parcela, las cuales serían más estable que los modelos generados para parcelas distintas. Se ha mencionado que sería un caso ideal ya que para su implementación útil se necesitarían una cantidad de datos sobre una sola parcela mucho mayor de la que disponemos actualmente, aunque también su modelado sería menos compleja, debido a la semejanza que habría en todos los casos.

Dejando a un lado el enfoque de mejora de los modelos creados, otras líneas futuras de investigación de este proyecto podría ser la implementación y comparación de distintos métodos de aprendizaje automático o análisis de regresión, realizando una evaluación y justificación de qué métodos funcionan mejor con este tipo de datos y por qué. Es una prueba que en principio estaba contemplada, aunque no a gran escala, en este proyecto pero no se ha podido llevar a cabo por tiempo, habiendo optado por centrarse el proyecto en una obtención de un modelo útil en vez de al estudio de los distintos métodos de realización. Otra línea de

trabajo futura obvia que se realizará será la integración de las salidas de estos modelos con las PDF's del marco de trabajo, objetivo último de este TFG que no se ha llevado a cabo. Aunque esta integración a priori resulte fácil e intuitiva, es digna de estudio, ya que habrá que buscar una compensación entre ambas salidas e integrarlas con un filtro de partículas, según estaba previsto por su buen funcionamiento en el marco de trabajo previo.

Para finalizar, se concluye que se ha implementado una herramienta muy útil, fácilmente manejable y necesaria en la investigación a la que está enfocada para completar el uso de la información disponible de una manera óptima, además de estar abierta a líneas futuras tanto de investigaciones relacionadas como la mejora de esta misma. Se ha contrastado la eficiencia de RFR para datos de este tipo y se corrobora la versatilidad de aplicaciones que puede tener esta herramienta dentro de la regresión y estimación de series temporales.

Bibliografía

- [1] SINERGISE LABORATORY FOR GEOGRAPHICAL INFORMATION SYSTEMS, L. *Sentinel hub playground*. URL <https://apps.sentinel-hub.com/sentinel-playground/?source=S2&lat=40.4&lng=-3.73000000000018&zoom=12&preset=1-NATURAL-COLOR&layers=B01,B02,B03&maxcc=20&gain=1.0&gamma=1.0&time=2019-12-01%7C2020-06-29&atmFilter=&showDates=false>.
- [2] ESA. *Sentinel-1: Radar mission*, 2014. URL <https://www.youtube.com/watch?v=FJWzLxdSMyA>.
- [3] ESA. *Sentinel-1 constellation*, 2014. URL https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-1/Satellite_constellation.
- [4] DE INGENIERÍA, R.A. *Diccionario español de ingeniería*, 2014. URL <http://diccionario.raing.es/es>.
- [5] ESA. *Copernicus overview*, 2016. URL https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Overview4.
- [6] EARTH OBSERVING SYSTEM (EOS). *Sentinel 1*;, 2014. URL <https://eos.com/sentinel-1/>.
- [7] RUSSELL, S.J. and NORVIG, P. *Artificial Intelligence A Modern Approach*, 2010.
- [8] FLACH, P. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, 2012.
- [9] FÜRNKRANZ, J. and HÜLLERMEIER, E. *Preference Learning: An Introduction*, 2011.
- [10] JOACHIMS, T. *Text categorization with support vector machines: Learning with many relevant features*. *Machine Learning: ECML-98. Lecture Notes in Computer Science*, 1998.
- [11] MAITY, A. *Supervised classification of radarsat-2 polarimetric data for different land features*. *Corr*, 2016. URL <https://dblp.org/rec/journals/corr/Mait16.bib>.
- [12] CASTILLO, F.E. and SENTÍS, F.C. *Agrometeorología*, 2001.
- [13] MEIER, U. *Growth stages of mono-and dicotyledonous plants*. *Federal Biological Research Centre for Agriculture and Forestry*, 2001. URL <https://web.archive.org/web/20180427154542/https://ojs.openagrar.de/index.php/BBCH/article/download/515/464>.

- [14] LANCASHIRE, P.D., BLEIHOLDER, H., BOOM, T.V.D., LANGELUDDEKE, P., STAUSS, R., WEBER, E., and WITZENBERGER, A. *A uniform decimal code for growth stages of crops and weeds. Annals of Applied Biology*, 1991.
- [15] VERDIN, J., PEDREROS, D., and EILERTS, G. *Índice diferencial de vegetación normalizado (ndvi). FEWS - Red de Alerta Temprana Contra la Inseguridad Alimentaria, Centroamérica, USGS/EROS Data Center*, 2003.
- [16] BERNARDIS, C.D., VICENTE-GUIJALBA, F., MARTINEZ-MARIN, T., and LOPEZ-SANCHEZ, J.M. *Contribution to real-time estimation of crop phenological states in a dynamical framework based on ndvi time series: Data fusion with sar and temperature*, 2016.
- [17] VICENTE-GUIJALBA, F., MARTINEZ-MARIN, T., and LOPEZ-SANCHEZ, J.M. *Dynamical approach for real-time monitoring of agricultural crops*, 2014.
- [18] ESA. *Missions: Sentinel -1. data products*, 2020. URL <https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1/data-products>.
- [19] WANGA, H., MAGAGIA, R., GOÏTAA, K., TRUDELA, M., McNAIRNB, H., and POWERS, J. *Crop phenology retrieval via polarimetric sar decomposition and random forest algorithm*. Elsevier, 2019.
- [20] DECOSTE, D. and SCHOLKOPF, B. *Training invariant support vector machines*. Kluwer Academic Publishers, 2002.
- [21] BARGHOUT, L. *Spatial-taxon information granules as used in iterative fuzzy-decision-making for image segmentation. Granular Computing and Decision-Making*, 2015.

Lista de Acrónimos y Abreviaturas

AEMA	Agencia Europea de Medio Ambiente.
BBCH	Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie.
EKF	Extended Kalman Filter.
EOS	Earth Observing System.
ESA	European Space Agency.
GRD	Ground Range Detected.
MAE	Mean Absolute Error.
ML	Machine Learning.
MMC	Método de Mínimos Cuadrados.
NDVI	Índice de Vegetación de Diferencia Normalizada.
PDF	Probability Density Function.
RAI	Real Academia de Ingeniería.
RF	Random Forest.
RFR	Random Forest Regressor.
RMSE	Root-Mean-Square Error.
SAR	Synthetic Aperture Radar.
SST	Señales, Sistemas y Telecomunicación.
TFG	Trabajo Fin de Grado.
UA	Universidad de Alicante.