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

Junio 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, Junio 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

Es a ellos a quien dedico este trabajo.

dedicatoria

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

Índice de figuras

Índice de tablas

Índice de Códigos

Lista de Acrónimos y Abreviaturas

AEMA	Agencia Europea de Medio Ambiente.
BBCH	Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie.
EOS	Earth Observing System.
ESA	European Space Agency.
GRD	Ground Range Detected.
ML	Machine Learning.
MMC	Método de Mínimos Cuadrados.
NDVI	Índice de Vegetación de Diferencia Normalizada.
RAI	Real Academia de Ingeniería.
RF	Random Forest.
SAR	Synthetic Aperture Radar.
SST	Señales, Sistemas y Telecomunicación.
TFG	Trabajo Fin de Grado.
UA	Universidad de Alicante.

1 Capítulo 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 [3]. 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) [3]. 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.

En este área ya hay estudios previos que, a partir de datos similares que comparten estos programas, se obtiene un estado de la fenología aproximado de los cultivos observados. Algunos estudios previos precedentes y que sirven de base para este TFG son:

- [4], artículo de 2014 que 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 y, finalmente una extensión del filtro de Kalman. La información que utiliza proviene de un radar polarimétrico del satélite Radsat-2 y los cultivos son 3 tipos de cereales.

- [5], artículo de 2016 que trata, de estimar el Índice de Vegetación de Diferencia Normalizada (NDVI), el cual representa el estado de la fenología, en tiempo real empleando filtros de partículas para integrar las dos fuentes de información utilizadas: imágenes SAR y temperatura del aire registrada. El satélite del que se obtiene la información es el TerraSAR-X y los cultivos observados son arrozales, como va a ser nuestro caso. Este obtienen resultados algo mejores que en el anterior artículo y se utiliza la misma tecnología que encontramos en este proyecto: SAR.

- [6], artículo de 2019 todavía más similar al objetivo de este proyecto, en él se estima el estado fenológico de distintos tipos de cultivos utilizando imágenes SAR proporcionadas por el satélite RADARSAT-2 y el método Random Forest (RF) para series temporales, que es uno de los elegidos también para este proyecto.

En resumen, para este proyecto en particular, el cultivo observado son arrozales, los datos empleados son imágenes SAR de los satélites Sentinel-1A y Sentinel-1B con ciclos periódicos de 6 días teniendo en cuenta ambos a partir de 2016, y las técnicas de estimación se basarán en las regresiones de series temporales y técnicas de aprendizaje automático.

1.2 Objetivos

Contribuyendo a la línea de investigación de los artículos [4] y [5], cuyos autores Juan Manuel López Sanchéz y Tomás Martínez Marín son el tutor y co-tutor de este TFG, respectivamente, el objetivo general 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 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 va a dividir en 3 capítulos principales las cuales son: marco teórico, metodología y resultados. Además de unas conclusiones finales valorando los resultados obtenidos.

En el marco teórico se expondrá 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. Veremos en él 3 secciones:

- Teledetección, incluyendo cómo funcionan los sistemas radar, en concreto los SAR, qué información obtenemos de ellos en ciertos programas de satélites y las técnicas de detección que determinan cómo interpretar esta información.
- Técnicas de regresión y Machine Learning, donde se encuentra la clasificación de las distintas técnicas de ML, algunos modelos y sus aplicaciones existentes, y, por último, el análisis mediante regresión.
- Estimación de parámetros físicos de cultivos mediante teledetección, donde se presentan la metodología general basada en el espacio de estados, los parámetros físicos de los cultivos con los que se puede trabajar y la estimación de estos mediante regresión.

En cuanto a la metodología, se incluirán tanto las técnicas y métodos concretos que se van a utilizar, por qué motivos y qué esperamos obtener de ellos, como el software, el lenguaje de programación que vamos a emplear, las herramientas utilizadas y las bases de datos con las que vamos a trabajar, incluyendo su procedencia y procesamiento previo.

Por último, el apartado de resultados expondrá los resultados obtenidos con las diferentes técnicas de regresión y aprendizaje automático para los mismos datos. Estos resultados podrán ser fácilmente evaluados ya que se contrastarán, además, con los datos reales tomados en tierra de los mismos cultivos que se presentan en el dataset.

2 Capítulo 2. Marco Teórico

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

2.1 Teledetección

Conociendo de qué trata este sector de las telecomunicaciones de una manera general, en este apartado se van a explicar conceptos más concretos de cómo funciona esta tecnología, qué tipos existen y qué técnicas se emplean en teledetección.

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 terreste 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, 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 0-360º.

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.

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

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.

Con el objetivo de maximizar la resolución espacial del área observada por el radar, esto es, la distancia mínima distinguible, 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 2.1.

Además de la resolución espacial, también existen otros parámetros que determinan la calidad de la información adquirida por los sistemas radar. Uno de ellos que aparece, similar al

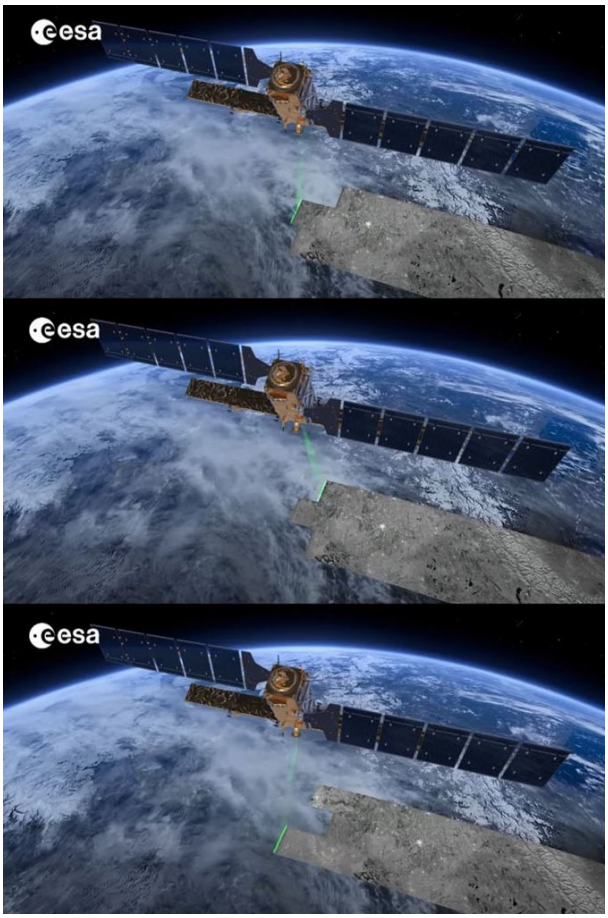

Figura 2.1: TOPS-SAR Sentinel-1 [1]

ruido blanco, como puntos de máximas y mínimas contribuciones debido a distintos fenómenos 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.

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. Las misiones satelitales de las que se va a obtener información para este proyecto son Sentinel-1 y Sentinel-2, del Programa Copérnico de la European Space Agency (ESA). Son dos misiones de órbita polar que engloban cada uno de ellos 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 y Sentinel-2 lo hizo en marzo de 2017. La principal diferencia entre estos dos es el rango de frecuencias de trabajo de cada uno, mientras que

Sentinel-1 proporciona información de la banda C, esto es entre 4-8 GHz, Sentinel-2 utiliza tecnología multiespectral, por lo 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 [7].

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 5 días y una altitud de 693 km y 786 km para Sentinel-1 y Sentinel-2, respectivamente y considerando ambos satélites, A y B, en ambos casos. En la figura 2.2 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.

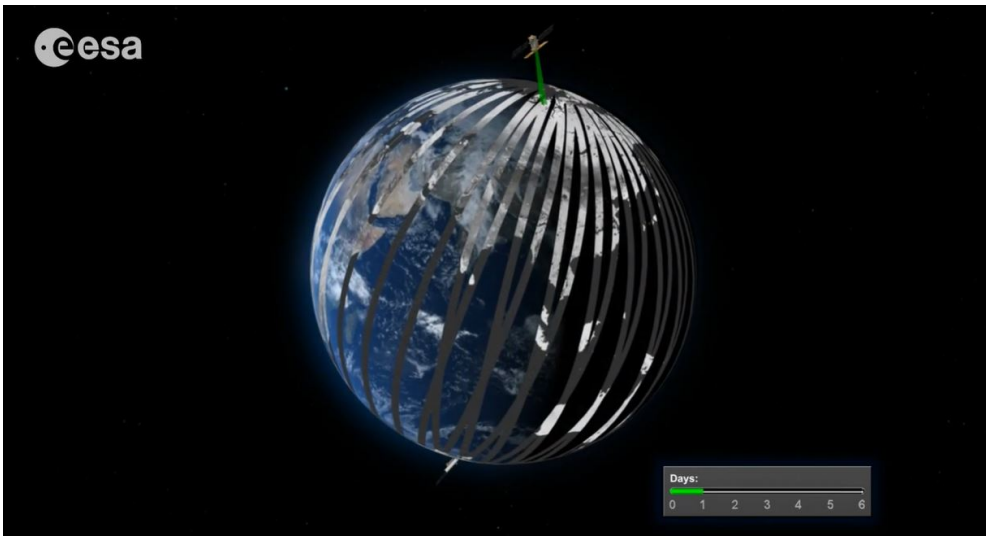

Figura 2.2: Swath a día 1 de 6 en Sentinel-1 [2]

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 [8].

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 [8]. 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.

2.1.3 Técnicas de detección

Con el objetivo de que la información captada por el radar sea comprensible y refleje una información coherente, existen distintos tipos de técnicas de detección según el tipo de información que se quiera extraer. Teniendo en cuenta solamente el coeficiente de backscattering en los sistemas SAR 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 dB $< \sigma_0 <$ 0 dB: superficies muy rugosas como pueden ser vegetaciones densas donde hay mucha probabilidad de reflexión.
- -20 dB $< \sigma_0 <$ -10 dB: superficies rugosas como vegetaciones menos densas entre las que se incluirían los cultivos.
- $\sigma_0 <$ -20 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:

- Interferometría: técnica que toma en consideración la fase obtenida en cada medición como la contribución del trayecto de la señal, el desfase introducido en el procesado y el cambio de fase debido a la reflectividad de la superficie observada. Crea modelos topográficos a partir de las diferencias de fase para una misma área con posiciones distintas de radar (pequeños desplazamientos o variaciones de posición en cada ciclo). Las fases podrán ser útiles si existe una correlación entre ellas, es decir, si son coherentes.
- Interferometría diferencial: tiene como objetivo recrear un modelo topográfico temporal que represente las deformaciones del terreno con el paso del tiempo. Se basa en el mismo principio que la técnica anterior, ya que utiliza la diferencia de fase, primero entre dos imágenes para crear el modelo topográfico, y, a continuación, con una tercera para detectar las zonas de la superficie en las que ha habido un desplazamiento. También se

puede realizar utilizando un modelo digital del terreno que aporta la topografía, y la diferencia de fase de dos imágenes para los desplazamientos.

- Polarimetría: se basa en la polarización de la onda recibida como fuente de información. Analiza su polarización estudiando su orientación y elipticidad media, con ellas se obtienen las matrices o vectores de scattering que definen una superficie por el cambio que esta realiza a la polarización de una onda incidente polarizada conocida. Esto da información sobre el tipo de superficie que se está observando.

2.2 Técnicas de regresión y Machine Learning

Las técnicas de regresión proporcionan una estimación 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.

* Introducir aquí esquema 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 y esto lleva a que tienen parte de su desarrollo en común.

2.2.1 Clasificación de técnicas de Machine Leaning

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 [9]:

- 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.

- 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 [10]:

- 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 [11]. 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 coordinadas.

- 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 y aplicaciones

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 cuáles son sus aplicaciones más usuales y en las que son más efectivos. 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 de este modelo es el reconocimiento en imágenes de objetos o caracteres. Cuando analizamos imágenes cada unidad de información a la entrada para un objeto está formada por un pixel, y cada capa de neuronas irá reconociendo formas, colores, etc. hasta devolver la clasificación que corresponde a ese objeto de entrada.

- 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, por lo que las aplicaciones son diversas. Relacionado con ese modelo también encontramos el conocido como RF, anteriormente mencionado. 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 y teniendo en cuenta el resto en su correspondiente porcentaje. De esta manera, un modelo que era limitado a una respuesta única, se abre devolviendo una respuesta probabilística.

- Máquinas de vectores de soporte: es un modelo geométrico y supervisado. Este modelo geométrico utiliza el espacio de instancias para representar los objetos de entrenamiento como puntos y las clases o salidas como líneas o hyperplanos, dependiendo del número de dimensiones, lo más separados posibles unos de otros. 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 el reconocimiento de caracteres escritos a mano [12] y clasificación de textos [13], la clasificación de imágenes por segmentación [14] o, el más interesante para este proyecto, la clasificación de información procedente de un SAR [15]

- 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.

- Algoritmos genéticos: son algoritmos basados en la genética biológica. Estos comienzan con unas muestras aleatorias de posibles salidas, se evalúan y se realizan una serie de transformaciones que incluyen la selección de los mejores resultados, su recombinación

y alteración de algunos para volver a empezar con la siguiente entrada, así hasta tener un modelo estable, fiable y calibrado. Este modelo se utiliza bastante para análisis y predicciones.

- También se consideran, entre otros métodos, los análisis de regresión. Técnicas que se pueden aplicar, como se puede ver en el siguiente apartado, al ML, ya que comparten un mismo objetivo.

2.2.3 Análisis de regresión

Las técnicas de regresión son todas aquellas técnicas 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 tiene 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.

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 este mismo general 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.

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_N X_N + \varepsilon \tag{2.2}$$

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.

Introducir aquí parte analítica regresión no lineal

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

A continuación, se van a presentar los conceptos físicos que definen el estado de desarrollo de los cultivos, para comprender cuáles son los parámetros clave y objetivos de este proyecto. Posteriormete, se expone la información útil extraíble de las imágenes SAR, a partir de la cuál se va a trabajar para el objetivo anterior. 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.

2.3.1 Metodología general basada en espacio de estados

2.3.2 Parámetros físicos de la fenología

La fenología es la ciencia que estudia la relación entre los factores climáticos y los ciclos de los seres vivos [16]. Esto es, el estudio del desarrollo de plantas y animales en relación con parámetros ambientales. Este proyecto se centrará en el estudio fenológico de plantas, en concreto de cultivos de arroz, por lo que se va a profundizar en las características fenológicas que los describen. Algunos de los parámetros clave son los siguientes:

- Escala Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie (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 [17]. 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* [18]. 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.

- 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 [19].

$$NDVI = \frac{IRCercano - ROJO}{IRCercano + ROJO} \tag{2.3}$$

- 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 [5].

2.3.3 Extracción de información de 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. Para este proyecto se van a tener en cuenta estas imágenes, de libre acceso por la ESA, con los canales de polarización VV y VH, esto es, emisión vertical y recepción tanto vertical como horizontal. Se ha elegido la polarización vertical ya que es la que el programa de los satélites de Sentinel-1 ofrece sobre Europa, y, además, 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.

Las imágenes SAR pueden ser adquiridas con distintos formatos. El que se va a utilizar aquí es el formato Ground Range Detected (GRD) *¿Resolución?* 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 para este estudio, ya que no era uno de los parámetros clave, y los píxeles que presenta la imagen son aproximadamente cuadrados [20].

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. Una vez finalizado este proceso, las imágenes están preparadas para tratar su información información de manera más sencilla.

2.3.4 Regresión aplicada a la estimación

Una vez presentados los parámetros clave para determinar el estado de desarrollo de un cultivo, más concretamente de los arrozales, y la información útil extraíble de las imágenes SAR, se va a exponer cómo se relacionan estos parámetros entre sí, haciendo uso de la regresión. 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.

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.

La información con la que se va a trabajar presenta unas limitaciones temporales debido al reciente desarrollo del programa Copérnico. Aún así, parece válida para una aproximación suficientemente fiable. Los datos con los que se cuenta para este proyecto parten del año 2016 en adelante, cuando el periodo de revista de Sentinel-1 pasó a ser de 6 días, con lo que la resolución temporal de cada ciclo comenzó a ser aceptable. Al ser los periodos de cultivo del arroz de aproximadamente 4 meses, se realizan 2 veces al año, número no muy amplio para disponer de suficientes ciclos de crecimiento desde el año 2016 hasta hoy. El alza en el número de periodos para general el modelo recae en la división de la información en distintas parcelas a estudiar, por lo que por cada época de cultivo se dispone de más de un ciclo de desarrollo

a tener en cuenta para el modelo. En el caso de no disponer de suficiente información, esta se podría generar de manera estadística, partiendo de la original.

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, según la varianza que esta información tuviera, por lo que puede ser utilizado para predicción, en este caso, del estado fenológico del nuevo cultivo observado.

A esto se ha de sumar la implicación de las series temporales en la estimación de estos parámetros, ya que sabiendo el instante del ciclo en el que se encuentra la nueva información obtenida y cuál fue su estado anterior, la predicción resulta mucho más sencilla y fiable que considerando solamente la información de entrada sin ningún contexto. Para ello se haría uso de los modelos dinámicos.

Bibliografía

[1] ESA. *Sentinel-1: Radar mission*, 2014. URL `https://www.youtube.com/watch?v=FJWzLxdSMyA`.

[2] ESA. *Sentinel-1 constellation*, 2014. URL `https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Sentinel-1/Satellite_constellation`.

[3] de Ingeniería, R.A. *Diccionario español de ingeniería*, 2014. URL `http://diccionario.raing.es/es`.

[4] Vicente-Guijalba, F., Martinez-Marin, T., and Lopez-Sanchez, J.M. *Dynamical approach for real-time monitoring of agricultural crops*, 2014.

[5] 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.

[6] 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.

[7] ESA. *Copernicus overview*, 2016. URL `https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Overview4`.

[8] Earth Observing System (EOS). *Sentinel 1:*, 2014. URL `https://eos.com/sentinel-1/`.

[9] Russell, S.J. and Norvig, P. *Artificial Intelligence A Modern Approach*, 2010.

[10] Flach, P. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, 2012.

[11] Fürnkranz, J. and Hüllermeier, E. *Preference Learning: An Introduction*, 2011.

[12] Decoste, D. and Scholkopf, B. *Training invariant support vector machines. Kluwer Academic Publishers*, 2002.

[13] Joachims, T. *Text categorization with support vector machines: Learning with many relevant features. Machine Learning: ECML-98. Lecture Notes in Computer Science*, 1998.

[14] Barghout, L. *Spatial-taxon information granules as used in iterative fuzzy-decision-making for image segmentation. Granular Computing and Decision-Making*, 2015.

[15] MAITY, A. *Supervised classification of radarsat-2 polarimetric data for different land features. CoRR*, 2016. URL https://dblp.org/rec/journals/corr/Maity16.bib.

[16] CASTILLO, F.E. and SENTÍS, F.C. *Agrometeorología*, 2001.

[17] 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.

[18] 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.

[19] 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.

[20] ESA. *Missions: Sentinel -1. data products*, 2020. URL https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-1/data-products.