Facultad de Ciencia y Tecnología

Universidad Autónoma de Entre Ríos

Licenciatura en Sistemas de Información

Tesina de Grado

TELEDETECCIÓN Y MACHINE LEARNING APLICADOS AL SEGUIMIENTO DE CULTIVOS DE SOJA

Autor

Frank Hernán Valentín

Director

Mag. Diaz, Santiago R.

Docentes

Ing. Vera Mónica

Lic. Bonadeo Pamela

Lic. Cian Damián

Junio 2026

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

DEDICATORIA

A mi abuelo “Coco”, por haber sido guía y ejemplo de quien quiero ser.

AGRADECIMIENTOS

Gracias a los docentes de la Facultad de Ciencia y Tecnología de la Universidad Autónoma de Entre Ríos, por su aporte a mi formación y por los conocimientos que sentaron las bases de mi desarrollo profesional.

Gracias a mi director Santiago, por hacerse un espacio para orientarme en el desarrollo de esta tesina, a pesar de tener muchas otras actividades.

Gracias a mis amigos, dentro y fuera de la universidad, por permitirme momentos de diversión y despeje de este laborioso proceso. En especial, quiero agradecer a Ignacio y a Bernardo, a quienes en múltiples ocasiones pedí opinión sincera sobre este trabajo.

Gracias a mi familia, por sobre todo a mis padres Silvina y Héctor, por su amor incondicional, por permitirme una buena formación, por apoyarme siempre, y por acompañarme y preocuparse por mí en todo momento.

RESUMEN

La presente investigación trata sobre la predicción de una serie temporal del Índice Normalizado de Diferencia de Vegetación, obtenido a partir de imágenes satelitales de cultivos de soja en la región centro-norte de la provincia de Santa Fe, Argentina. El estudio se desarrolló sobre un lote ubicado en la localidad de San Justo de dicha provincia, tomando la campaña 2020-2021 para el entrenamiento de dos modelos con diferentes enfoques: un modelo estadístico tradicional SARIMA y una red neuronal recurrente con celdas Long Short-Term Memory. Luego, se utilizó la campaña 2024-2025 para comprobar la eficacia de los modelos en la detección de anomalías conocidas de antemano.

Los resultados obtenidos demostraron que la red neuronal, además de obtener mayor precisión en la predicción de la serie temporal frente al modelo SARIMA, logró identificar las anomalías, mientras que este último no lo consiguió.

Como aporte metodológico, se propone un procedimiento para el seguimiento del NDVI en cultivos de soja que comprende las siguientes etapas: análisis del área de estudio, extracción y transformación de los datos, ampliación de la serie temporal, ajuste de los modelos y aplicación de estos para la detección de anomalías.

Palabras clave: Teledetección, Índice Normalizado de Diferencia de Vegetación, Machine Learning, Redes Neuronales, SARIMA

INTRODUCCIÓN

En el ámbito de los Sistemas de Información, el análisis avanzado de datos y el uso de técnicas de aprendizaje automático se han consolidado como herramientas clave para la generación de conocimiento y el apoyo a la toma de decisiones en muchos sectores.

Un ejemplo de esta tendencia se da en el ámbito agrícola, donde el uso de técnicas de teledetección ha posibilitado el desarrollo de series temporales de índices de vegetación que describen la dinámica de los cultivos sin necesidad de relevamientos in situ. Entre estos índices, el Índice Normalizado de Diferencia de Vegetación (Normalized Difference Vegetation Index, NDVI) se ha consolidado como un indicador robusto del estado fenológico y del vigor de la vegetación (Farias et al., 2023). No obstante, su uso plantea desafíos asociados a la calidad de las observaciones (afectadas por nubes, atmósfera y resolución espacial) y a la variabilidad inherente de los sistemas agrícolas .

En paralelo, los avances en ciencia de datos y aprendizaje automático ofrecen nuevas metodologías para analizar estas series temporales. Los modelos estadísticos clásicos resultan útiles para capturar patrones de tendencia y estacionalidad; sin embargo, presentan limitaciones al momento de representar relaciones complejas o cambios abruptos en los datos. Las redes neuronales recurrentes del tipo LSTM (Long Short-Term Memory) emergen como una alternativa que aprovecha dependencias de largo plazo (Nielsen, 2019), y resulta prometedora para el modelado del crecimiento de cultivos.

Por otra parte, al ser la soja uno de los principales cultivos de la región pampeana y del centro-norte de Argentina, con un rol estratégico en la economía nacional, disponer de herramientas que permitan monitorear su crecimiento y predecir su comportamiento ante condiciones ambientales adversas resulta fundamental para la planificación agrícola y la toma de decisiones.

Por estos motivos, se decidió realizar una investigación centrada en la comparación de la eficacia de un modelo estadístico clásico y una red neuronal en una tarea de modelado de series temporales del NDVI, para su uso en la detección de anomalías sobre cultivos de soja; y para permitir su replicación se propone el siguiente objetivo general:

Diseñar una metodología para el monitoreo del Índice Normalizado de Diferencia de Vegetación, aplicada a cultivos de soja en el centro-norte de la provincia de Santa Fe.

Además, para lograrlo se formulan los siguientes objetivos particulares:

- Estudiar técnicas de teledetección aplicadas al análisis de cultivos.
- Analizar las etapas fenológicas de la soja en base al Índice Normalizado de Diferencia de Vegetación (NDVI).
- Estudiar Machine Learning enfocado en modelos de series temporales.
- Evaluar el rendimiento de modelos de Machine Learning sobre una serie temporal del Índice Normalizado de Diferencia de Vegetación (NDVI) de cultivos de soja, en el centro-norte de la provincia de Santa Fe.
- Diseñar una metodología para el monitoreo del Índice Normalizado de Diferencia de Vegetación, adaptado a las condiciones del centro-norte de Santa Fe.

De este modo, el trabajo busca aportar conocimiento y herramientas aplicables en la agricultura de precisión, contribuyendo a un uso más eficiente de los recursos y a la toma de decisiones basada en evidencia.

CAPÍTULO 1 - ANÁLISIS DE CULTIVOS UTILIZANDO TELEDETECCIÓN

En este capítulo se presenta la técnica de teledetección, utilizada para obtener datos en formato de imágenes desde sensores remotos, y su uso para el cálculo de índices de vegetación empleados para estimar el vigor de distintos tipos de cultivos. El objetivo del capítulo es comprender cómo se obtienen estos índices a partir de datos abiertos provistos por agencias espaciales para ser utilizados por el público general.

1.1 - Teledetección

Se trata de una técnica para obtener información sobre objetos o áreas mediante sensores ubicados a distancia, lo que implica que se hace sin contacto físico directo con el elemento estudiado (SEGEMAR, s. f.).

En la agricultura se encuentran ejemplos como cámaras térmicas usadas para detectar problemas en el riego, sensores LiDAR (*Light Detection and Ranging*) que utilizan luz láser para crear modelos detallados del terreno en tres dimensiones, o sensores espectrales o sensores multiespectrales que miden la reflectancia de la radiación electromagnética en distintas longitudes de onda para inferir propiedades de los cultivos. Los sensores utilizados para la teledetección pueden ser fijos, montados en el suelo, o estar colocados sobre drones, aviones, o incluso satélites.

Aplicada en cultivos, esta técnica proporciona información confiable sobre los factores que afectan al desarrollo de la planta, permitiendo al productor tomar decisiones basadas en datos, en un período de tiempo corto desde que ocurren. Además, mediante el cálculo de diferentes índices, permite evaluar la evolución de la cubierta vegetal y el crecimiento de los cultivos, como también monitorear la incidencia de enfermedades en las plantas, o incluso detectar el rendimiento potencial de un lote, por lo que constituye una herramienta valiosa para determinar el estado de los cultivos de una manera no destructiva, de bajo costo y con una velocidad de adquisición de datos relativamente alta (Farias et al., 2023).

Como se mencionó, una de las formas de obtención de los datos es a través de sensores ubicados en satélites. En esta línea, agencias espaciales como la Agencia Espacial Europea (European Space Agency, ESA) o la Administración Nacional de Aeronáutica y el Espacio (National Aeronautics and Space Administration, NASA), llevan a cabo programas de observación de la tierra donde se obtiene este tipo de información que luego se brinda a los usuarios por medio de plataformas, en su mayoría, de acceso libre y gratuito.

Uno de estos programas es el llamado Copérnico, llevado a cabo por la Comisión Europea en conjunto con la ESA desde 2014. Este programa tiene como

objetivo la producción de grandes volúmenes de datos provenientes de satélites y otras fuentes de observación en tierra para su utilización en la monitorización del planeta.

Satélites Sentinel

Los satélites Sentinel constituyen una serie de misiones satelitales del programa Copérnico dedicadas al monitoreo del clima terrestre y a la generación de registros continuos de observación de la Tierra, dando continuidad a décadas de datos recopilados por misiones satelitales previas. En el marco de estas misiones, se han puesto en órbita distintos satélites que proporcionan observaciones sistemáticas del planeta y servicios de información para una amplia variedad de aplicaciones. En particular, las misiones Sentinel ofrecen imágenes ópticas[1] y de radar que permiten observar la superficie terrestre de manera continua, lo que resulta fundamental para el monitoreo del uso del suelo y de la vegetación (European Commission, s.f.).

La información generada por el Programa Copérnico es utilizada en múltiples áreas, entre ellas el monitoreo del clima, del ambiente marino, de la atmósfera, la gestión de emergencias y aplicaciones de seguridad. De esta forma, este programa nutre sistemas de información con datos reales con el objetivo de apoyar estudios científicos, negocios basados en el análisis de datos, políticas públicas o actividades de gestión ambiental y territorial.

Con el objetivo de obtener una comprensión rápida de la capacidad de cada satélite y sus usos, en la Tabla 1 se listan características de cada misión junto con sus aplicaciones más comunes.

[1] Una imagen óptica es generada a partir de la reflexión, refracción o difracción de ondas de luz por diferentes tipos de lentes o espejos

Tabla 1

Tabla comparativa entre misiones Sentinel del Programa Copérnico

Misión	Satélite	Aplicaciones
Sentinel-1	● Instrumentos: SAR[2] en banda C (5.405 GHz), que garantiza la obtención de datos tanto de día como de noche y en prácticamente cualquier condición climática. ● Resoluciones espaciales 5m×5m, 5m×20m, 25×100m (según modo de adquisición). ● Tiempo de revisita: 6 días utilizando los dos satélites de la misión.	● Monitoreo de superficies terrestres y oceánicas. ● Detección de inundaciones. ● Identificación de deformaciones del terreno. ● Monitoreo de hielo marino. ● Gestión de emergencias climáticas.
Sentinel-2	● Instrumentos: sensor óptico multiespectral, que trabaja con 13 bandas del espectro electromagnético. ● Resolución espacial: 10m×10m (para bandas B2, B3, B4, B8), 20m×20m (para bandas B5, B6, B7, B8a, B11, B12) y 60m×60m (para bandas B1, B9, B10). ● Tiempo de revisita: 5 días con la constelación.	● Monitoreo de vegetación y cultivos. ● Detección de cambios en la cobertura del suelo. ● Estudios de sedimentos en el mar. ● Monitoreo forestal y ambiental. ● Identificación de fallas geológicas.

[2] Un radar de apertura sintética (Synthetic Aperture Radar, SAR) es un instrumento de recolección activa de datos (a diferencia de una imagen óptica que es un método pasivo) que emite un pulso de energía y luego registra la cantidad de energía reflejada tras interactuar con la superficie de la Tierra. En ellos, la resolución espacial que tendrán las imágenes generadas están ligadas a la longitud de onda en la que trabaja el sensor (NASA, s.f.).

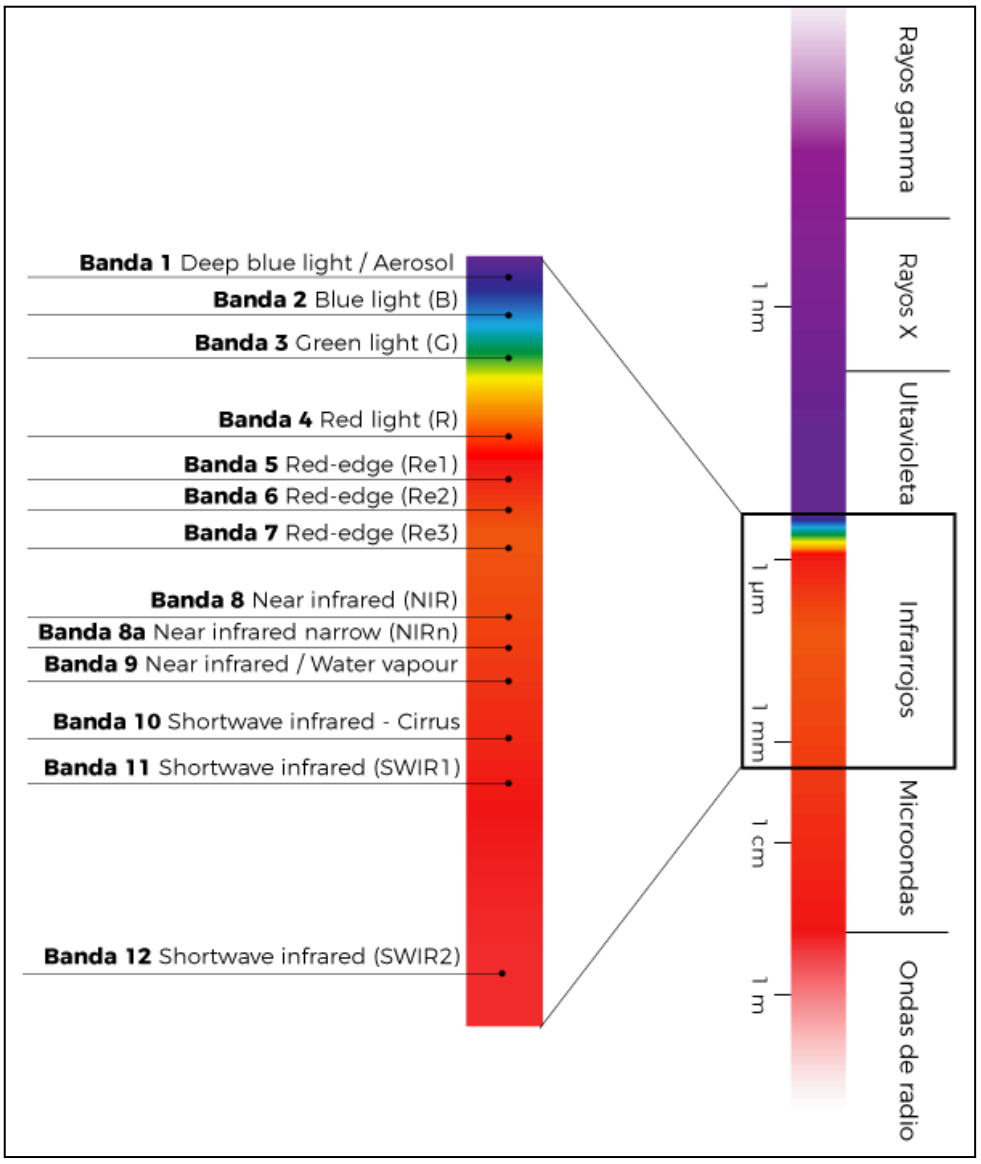
Misión	Satélite	Aplicaciones
Sentinel-3	● Instrumentos: Instrumento de color para océanos y tierra, que mide el color de la superficie; Radiómetro de medición de temperatura de la superficie del mar y de la tierra, que mide la temperatura de la superficie; Altímetro SAR, que mide distancia hasta la superficie . ● Resolución espacial: 300m×300m o 500m×1 km, según instrumento. ● Tiempo de revisita: menor a 2 días.	● Monitoreo de océanos y clima. ● Monitoreo de la temperatura de superficie del mar y tierra. ● Detección del color del océano, altimetría marina y dinámica oceánica. ● Monitoreo ambiental.
Sentinel-4	● Instrumentos: Espectrómetro de Imagen Hiperespectral, que descompone la luz para identificar diferentes gases. ● Resolución espacial: 8km×8km. ● Tiempo de revisita: casi en tiempo real (1h).	● Monitoreo de calidad del aire y composición atmosférica. ● Detección de gases contaminantes y aerosoles.
Sentinel-5	● Instrumentos: Espectrómetro ultravioleta, visible, infrarrojo cercano e infrarrojo de onda corta, que mide la luz reflejada por la atmósfera. ● Resolución espacial: 7km×7km. ● Tiempo de revisita: 1 día.	● Monitoreo de la calidad del aire. ● Monitoreo de clima. ● Monitoreo de la capa de ozono.

Nota. Esta tabla expone características y aplicaciones comunes de cada misión del Programa Copérnico. Información adaptada de Sentinel-1, por Copernicus (s. f.), https://sentiwiki.copernicus.eu/web/sentinel-1; Sentinel-2, por Copernicus (s. f.), https://sentiwiki.copernicus.eu/web/sentinel-2; Sentinel-3, por Copernicus (s. f.), https://sentiwiki.copernicus.eu/web/sentinel-3; Sentinel-4, por European Space Agency (s. f.), https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-4; Sentinel-5, por European Space Agency (s. f.), https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-5

El sensor MSI que portan ambos satélites de la misión Sentinel-2 trabaja en diferentes bandas del espectro electromagnético; para una mejor visualización de esto, la Figura 1 plasma en qué espacio de la longitud de onda se encuentra cada banda.

Figura 1

Bandas de Sentinel-2



Nota. Esta figura muestra donde se encuentra en el espectro de luz cada banda del satélite Sentinel-2. Adaptado de *Teledetección. Edu-Sat*.Edusat. (s. f.) https://www.edu-sat.com/teledeteccion/?lang=es

Combinando bandas específicas del espectro de onda electromagnética se obtienen productos para diferentes aplicaciones como imágenes en color natural (combinando las bandas B4, B3, B2), imágenes para geología (combinando B12, B11, B2) que resaltan fallas geológicas, formaciones rocosas y formaciones geológicas, imágenes para el estudio del mar (combinando B4, B3, B1) que permiten identificar

sedimentos suspendidos en el agua, e imágenes de índices de vegetación (combinando B8, B4) que permiten distinguir áreas con follaje denso de áreas urbanas o con cuerpos de agua (GIS Geography, 2025).

Los datos obtenidos en todas estas misiones están disponibles a través de distintas plataformas de procesamiento geoespacial como Google Earth Engine (GEE). Esta plataforma es un servicio de la empresa Google que permite desarrollar algoritmos para el análisis de datos geoespaciales a gran escala y en la nube. GEE cuenta con un amplio catálogo de datos de distintas fuentes que es accesible desde su plataforma web o a través de su interfaz de programación de aplicaciones (API), entre éstos se proveen a los usuarios los datos capturados por los sensores de las misiones Sentinel.

1.2 - Índices de Vegetación

Mediante la combinación matemática de los valores correspondientes a diferentes bandas del espectro electromagnético registradas por sensores remotos, se obtienen los denominados Índices de Vegetación. El objetivo de estos índices es resaltar la presencia, condición y vigor de la vegetación, aprovechando las diferencias en cómo las plantas reflejan o absorben la radiación en determinadas longitudes de onda (principalmente en el rojo y el infrarrojo cercano). Estos índices permiten reducir la influencia de otros elementos de la superficie, como suelo o agua, y facilitar el análisis de la cobertura vegetal a partir de imágenes satelitales (Jensen, 2007).

Un índice muy utilizado por investigadores actualmente es el Índice Normalizado de Diferencia de Vegetación, o NDVI (por sus siglas en inglés, Normalized Difference Vegetation Index), el cual tiene las ventajas de permitir el monitoreo de cambios en el crecimiento de la vegetación de manera interanual, y de una fórmula matemática que reduce el ruido multiplicativo proveniente de elementos ajenos a la vegetación. Aunque posee también las desventajas de no contemplar el vigor de la vegetación más allá de un punto de saturación, y de verse afectado por el fondo sobre el que se encuentran las imágenes de la vegetación, como puede ser el suelo visible entre las hojas de una planta (Jensen, 2007).

Este índice se basa en el comportamiento espectral característico de la vegetación verde, que absorbe gran parte de la radiación en la banda roja (RED) para la fotosíntesis y refleja fuertemente en el infrarrojo cercano (Near Infrared, NIR) debido a la estructura interna de las hojas (Jensen, 2007), y se calcula como la diferencia entre las mediciones de la banda de infrarrojo cercano y las de la banda de luz roja:

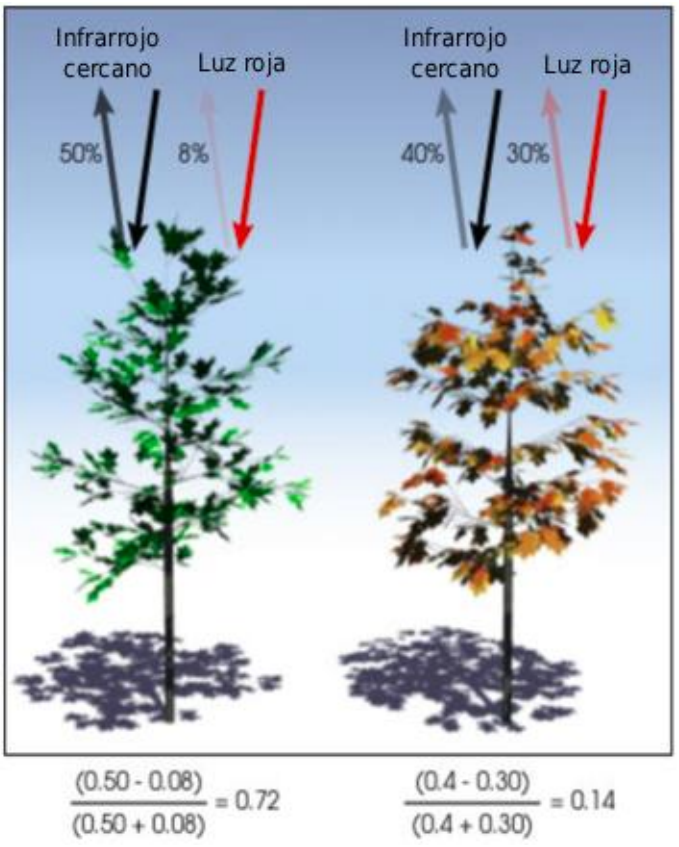
$$NDVI = \frac{(NIR - Red)}{(NIR + Red)}$$

La adquisición de datos en las bandas RED y NIR constituye un elemento clave en aplicaciones de agricultura de precisión orientadas al monitoreo de vegetación debido a que los índices obtenidos con combinaciones de ellas se utilizan para evaluar el vigor vegetal y distinguir áreas con vegetación saludable de aquellas afectadas por estrés o de superficies con escasa o nula cobertura vegetal.

La Figura 2 ejemplifica el uso del NDVI mediante una situación en la cual existen dos árboles, el izquierdo se encuentra en buen estado de salud, por lo tanto sus hojas son de color verde y reflejan más luz en el espectro infrarrojo cercano que luz roja; mientras que el de la derecha posee alguna limitación, por eso sus hojas tienen un color marrón o amarillento y reflejan un poco menos luz del espectro infrarrojo cercano y bastante más luz roja. Debajo de los árboles se muestra el cálculo del NDVI para cada uno, obteniendo un índice mayor para el árbol con hojas verdes.

Figura 2

Ejemplo de cálculo de NDVI



Nota. Ejemplo de cálculo de NDVI en un árbol sano y uno seco. Adaptado de *Geographic information systems and remote sensing: Innovative tools for plant health*, por Haggag, Wafaa & Ali, Rafat & Al-Ansary, Noran. (2023). https://www.researchgate.net/figure/Normalized-difference-vegetation-index-NDVI-reflects-the-photosynthetic-activity-a_fig1_375867555

Los valores del NDVI varían entre −1 y 1. Valores negativos suelen asociarse con superficies de agua, nieve o nubes, mientras que valores cercanos a 0 corresponden generalmente a suelos sin cobertura o a vegetación muy escasa. En cambio, valores positivos altos indican la presencia de vegetación densa y saludable, caracterizada por una fuerte reflectancia en el infrarrojo cercano (Jensen, 2007).

En la Figura 3 se observa una imagen del NDVI calculado para un lote de muestra. Este tipo de representación permite evaluar el vigor de la vegetación y analizar la variabilidad espacial presente dentro del cultivo. Dado que el NDVI se encuentra relacionado con variables como la biomasa, el contenido de clorofila y el estado nutricional de las plantas, su utilización facilita la identificación de sectores con comportamientos diferenciales que podrían requerir un análisis más detallado. Asimismo, las imágenes derivadas de este índice se emplean para la elaboración de mapas de variabilidad espacial, los cuales constituyen una herramienta de apoyo para la toma de decisiones en el manejo de los cultivos (Farias et al., 2023).

Figura 3

Ejemplo de Imagen del NDVI

Al calcular el NDVI desde imágenes satelitales se deben tener en cuenta diferentes fuentes de error posibles, como la presencia de nubes o sombra, la existencia de gases o vapor de agua que alteran la captación de las bandas RED y NIR, las cuales representan las regiones espectrales de la luz roja y del infrarrojo cercano sobre las que se fundamenta el cálculo del NDVI; ruido en la captación de imágenes del sensor, o el efecto de la reflectancia de algún objeto ubicado en el piso debajo de las hojas del cultivo que podría alterar el NDVI en etapas tempranas del cultivo si no es tenido en cuenta en el preprocesamiento de las imágenes.

CAPÍTULO 2 - CULTIVO DE SOJA

Este capítulo tiene como objetivo realizar una breve descripción del proceso de crecimiento de la planta de soja, con el objetivo de comprender sus estadíos fenológicos, y las variables que lo afectan. Además, se introduce la relación entre el NDVI y los períodos fenológicos de la soja.

2.1 - Evolución

El proceso de desarrollo de la planta de soja contempla desde la germinación de la semilla hasta que la planta está lista para la cosecha. Durante este período, existen diferentes factores que afectan el crecimiento, algunos no controlados por los productores agropecuarios como factores climáticos; y otros sí, como los tipos de insumos utilizados, y fechas de siembra. Al trabajar con cultivos, es fundamental conocer la terminología asociada al desarrollo de las plantas para interpretar correctamente las instrucciones. Por ejemplo, cuando un fabricante de pesticidas indica el momento adecuado de aplicación, es necesario identificar con precisión la etapa fenológica[3] correspondiente; de lo contrario, no se logrará el efecto esperado (Fehr & Caviness, 1977).

En cultivos de soja, la descripción de las etapas fenológicas más utilizada actualmente es la escala propuesta por Fehr y Caviness[4], que consta de dos descripciones separadas para las etapas principales, la de vegetación (V) y la de reproducción (R). A su vez, éstas etapas se subdividen formando la escala que se observa en la Figura 4, y que se describe a continuación.

Por un lado, los estadíos vegetativos, que comprenden las siguientes etapas:

- Emergencia (VE): etapa en la que ya se observan los cotiledones, que son las primeras hojas de la planta, por encima de la superficie del suelo.
- Cotiledonar (VC): hojas unifoliadas desenrolladas lo suficiente como para que los bordes de la hoja no se toquen.
- Primer nudo (V1): hojas completamente desarrolladas en los nudos unifoliados.
- Segundo nudo (V2): hoja trifoliada, por encima de los nudos unifoliados, completamente desarrollada.

[3] Las etapas fenológicas son las distintas fases de crecimiento y desarrollo que componen el ciclo de vida de una planta, están determinadas por cambios visuales en su estructura y su duración es influida por factores biológicos, climáticos y genéticos. Comarca Agrícola. (2024, octubre 10). ¿Qué son las etapas fenológicas en los cultivos y por qué son importantes?. https://comarcaagricola.com/comarca-agricola-04/

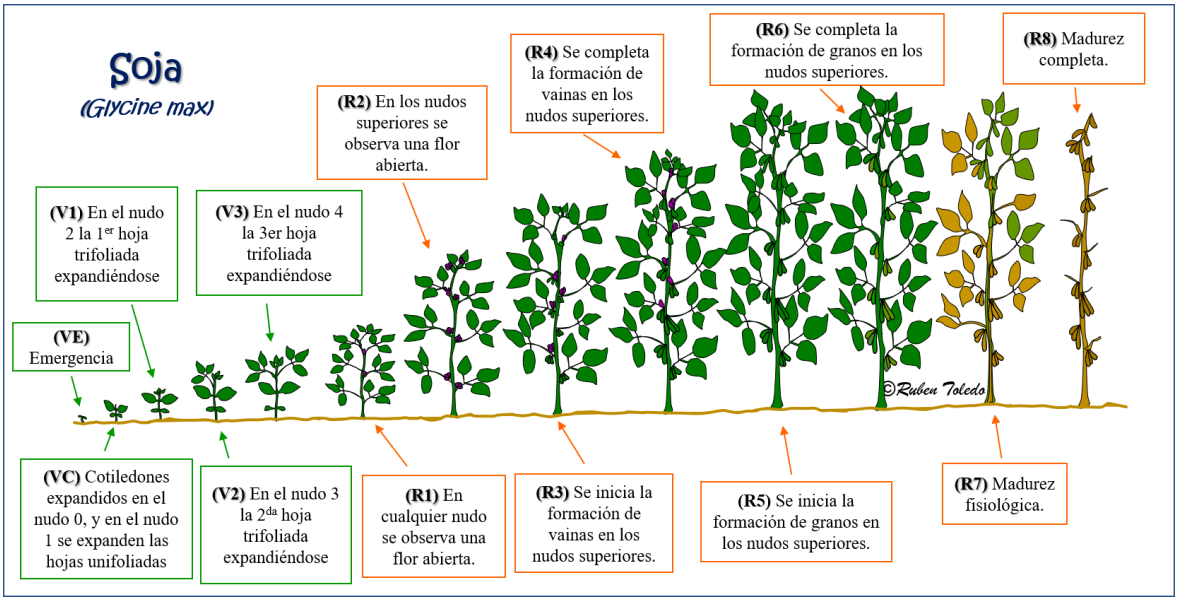
[4] Fehr, W. R., & Caviness, C. E. (1977). *Stages of soybean development* (Special Report 80). Iowa State University of Science and Technology, Cooperative Extension Service, and Agriculture and Home Economics Experiment Station.

- Tercer nudo (V3): tres nudos en el tallo principal con hojas totalmente desarrolladas, comenzando con los nudos unifoliados.
- N-ésimo nudo (Vn): N nudos en el tallo principal con hojas completamente desarrolladas, comenzando con los nudos unifoliados.

Los nudos son marcas que dejan las hojas al caer del tallo principal de la planta y se utilizan para determinar su estadío ya que, a diferencia de las hojas, son permanentes.

Figura 4

Representación de la clave fenológica de soja



Nota. La figura es una representación de cada etapa de la escala de Fehr y Caviness. Tomado de *Representación de la clave fenológica de soja,* por Toledo, R., 2023, ResearchGate, https://doi.org/10.13140/RG.2.2.31548.82564

Por su parte, los estadíos reproductivos comprenden las etapas de:

- Comienzo de la floración (R1): existe una flor abierta en cualquier nudo del tallo principal.
- Plena floración (R2): existe una flor abierta en alguno de los dos nudos superiores del tallo principal que posea una hoja completamente desarrollada.
- Vaina inicial (R3): existe una vaina de 5 mm de largo en uno de los 4 nudos superiores del tallo principal que posea una hoja completamente desarrollada.
- Vaina completa (R4): existe una vaina de 2 cm de largo en uno de los 4 nudos superiores del tallo principal que posea una hoja completamente desarrollada.

- Semilla inicial (R5): existe una semilla de 3 mm de largo en una vaina de uno de los 4 nudos superiores del tallo principal que posea una hoja completamente desarrollada.
- Semilla completa (R6): existe una vaina que contiene una semilla verde que llena su cavidad, en uno de los 4 nudos superiores del tallo principal que posea una hoja completamente desarrollada.
- Comienzo de madurez (R7): existe una vaina normal en el tallo principal que ha alcanzado su color normal de madurez.
- Madurez completa (R8): 95% de las vainas han alcanzado su color normal a la madurez.

Esta escala representa el desarrollo de una planta de soja particular, para evaluar el crecimiento de un lote entero se debe tomar el período en el que esté al menos la mitad de una muestra representativa de todo el lote.

La escala también permite distinguir según su hábito de crecimiento las variedades de soja determinada e indeterminada, cuyos patrones de desarrollo difieren significativamente. En la determinada, la floración ocurre casi al mismo tiempo entre la parte superior e inferior de la planta, por lo que el desarrollo de las vainas y semillas es prácticamente igual en toda la planta. Por otra parte, la soja indeterminada cuenta con menos de la mitad de su altura final cuando comienza la floración, por lo que el desarrollo de las vainas y semillas es siempre más avanzado en la parte inferior de la planta (Fehr & Caviness, 1977).

2.2 - Manejo

Para garantizar un crecimiento adecuado de la planta, y por consiguiente alcanzar un mayor rendimiento, se debe prestar atención a los distintos factores que influyen en este proceso. Además de elementos no controlados relacionados al clima y la capacidad productiva del suelo, la cantidad de granos que produce una planta de soja responde a variables que los agricultores deben planificar inteligentemente: "… la fecha de siembra, densidad, elección de cultivares y nutrición tienen una importancia preponderante en los resultados que obtiene el productor" (Morla et al., 2022, p. 69).

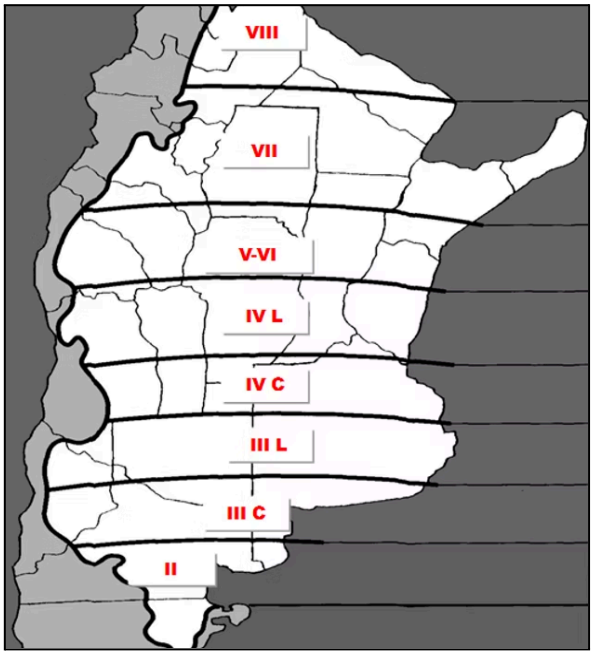
Los distintos cultivares de soja se agrupan en los llamados Grupos de Madurez (GM), que van del 00 al X en todo el mundo, y utilizan la duración de la etapa de emergencia a la floración como característica de agrupamiento. Los GM inferiores al V de ciclo corto se denominan bajos o menores por tener ciclo de desarrollo más corto, y son todos de hábito de crecimiento indeterminado, mientras que los GM de ciclo largo y superiores son llamados altos o mayores, y contemplan tanto genotipos de hábito de crecimiento indeterminado como determinado (Toledo, 2018). Asimismo, se los clasifica

como GM productivos y GM estables: cuánto más bajo es el grupo más corto es el ciclo desde la siembra hasta la cosecha, por lo que más productivo es; y cuanto más alto el GM más se alarga el ciclo, lo que permite una mayor utilización de recursos y más margen de recuperación ante condiciones adversas, por lo tanto el rendimiento obtenido es más estable.

En Argentina se utilizan sólo los GM en el rango II a VIII, siendo posible sembrar los GM IV a VIII en la Región Norte (por encima de los 30º de latitud sur); mientras que en la Región Pampeana Norte (entre los 30º y 36º de latitud sur) se emplean los GM IV al GM VI, con posibilidad de usar cultivares de ciclo largo de GM III en el sur de la región y cultivares de GM VIII en el norte; y en la Región Pampeana Sur (al sur de los 36º de latitud sur), se siembran genotipos de los GM II al IV (Toledo, 2017). La Figura 5 revela las franjas resultantes de la división del país por grupo de madurez utilizado.

Figura 5

Franjas latitudinales de adaptación



Nota: La figura muestra las franjas resultantes de separar por grupo de madurez la zona productiva de soja en Argentina. Tomado de *Grupos de Madurez y Fecha de siembra*, por Toledo, R., 2017, SOJA Su ecofisiología y manejo, https://toledoruben.wixsite.com/cultivodesoja/fs-y-gm.

Otro factor importante para que la planta de soja obtenga buen rendimiento es la fecha de siembra: “existe una ventana climática extensa para la siembra del cultivo de soja, desde las siembras anticipadas (principios de octubre) hasta fechas ultra tardías

(finales de enero). Las decisiones de manejo del cultivo son centrales para poder hacer el uso más eficiente de los recursos ambientales en esta amplia ventana agroclimática." (Salvagiotti et al., 2010, p. 152).

Si la siembra se realiza al comienzo de la ventana de clima apto (octubre-noviembre), se le llama soja de primera; mientras que si se siembra a fines de ella (diciembre-enero) y le sucede a un cultivo invernal, que generalmente es trigo, se le llama soja de segunda. Los beneficios fundamentales de la soja de segunda son que posibilita dos cosechas en un mismo año y la oportunidad de siembra sobre rastrojo, práctica que retrasa el surgimiento de malezas y reduce la necesidad de fertilizantes.

Ajustar la fecha de siembra al grupo de madurez utilizado es crucial para garantizar la disponibilidad de los recursos necesarios y para minimizar el estrés durante etapas críticas como la de floración o la de emergencia de semillas, con el objetivo de maximizar la productividad del cultivo. En el mercado actual se encuentran variedades adaptadas a distintos tipos de ambiente: desde las que son completamente insensibles al fotoperíodo, pasando por las que requieren períodos más largos, utilizadas en latitudes altas; hasta variedades creadas para latitudes bajas que poseen alta sensibilidad y florecen en pocos días (Toledo, 2018).

En cuanto a variables climatológicas, la temperatura regula el ciclo fenológico completo. Si bien no se observa mucha variabilidad en la respuesta de cada grupo de madurez a la temperatura, la principal diferencia radica en los tiempos requeridos para floración entre los grupos de madurez mayores y los menores, que es inferior en estos últimos.

A diferencia de lo que se podría esperar, la relación entre la duración de cada etapa con la temperatura no sigue una relación lineal. Según Toledo (2018), el crecimiento se acelera progresivamente cuando las temperaturas superan un umbral mínimo (temperatura base), alcanzando su máxima velocidad en un rango óptimo. No obstante, si la temperatura excede ese rango, la tasa de desarrollo comienza a decaer, hasta detenerse por completo al llegar a la temperatura máxima.

Otro elemento importante en este proceso es el fotoperíodo, que es el tiempo de luz diario al que está expuesta la planta, e influye principalmente en la inducción de la floración, y a diferencia de la temperatura sólo actúa desde que aparecen las primeras hojas trifoliadas (V1-V2) hasta el comienzo de la madurez (R7).

Estas dos variables, la temperatura y el fotoperíodo, trabajan en conjunto ya que la respuesta al fotoperíodo se ve afectada por la temperatura. En ambientes fríos se reduce la sensibilidad al fotoperíodo, lo que produce que aumente la duración de la etapa vegetativa y se retrase la floración.

Más causantes de variaciones en el desarrollo del cultivo, y por lo tanto en su rendimiento final son la profundidad y densidad de siembra, el uso de fertilizantes que cubran los nutrientes faltantes en el ambiente, y la ocurrencia de plagas, malezas, enfermedades o condiciones climáticas adversas como sequías, inundaciones y fuertes vientos que produzcan el vuelco de las semillas.

Para medir la superficie que ocupan las hojas de una planta, denominada área foliar, se utiliza el Índice de Área Foliar (IAF), que se calcula como:

$$IAF = \frac{Área\ Foliar \times Densidad\ de\ Población}{Área\ con\ Sombra}$$

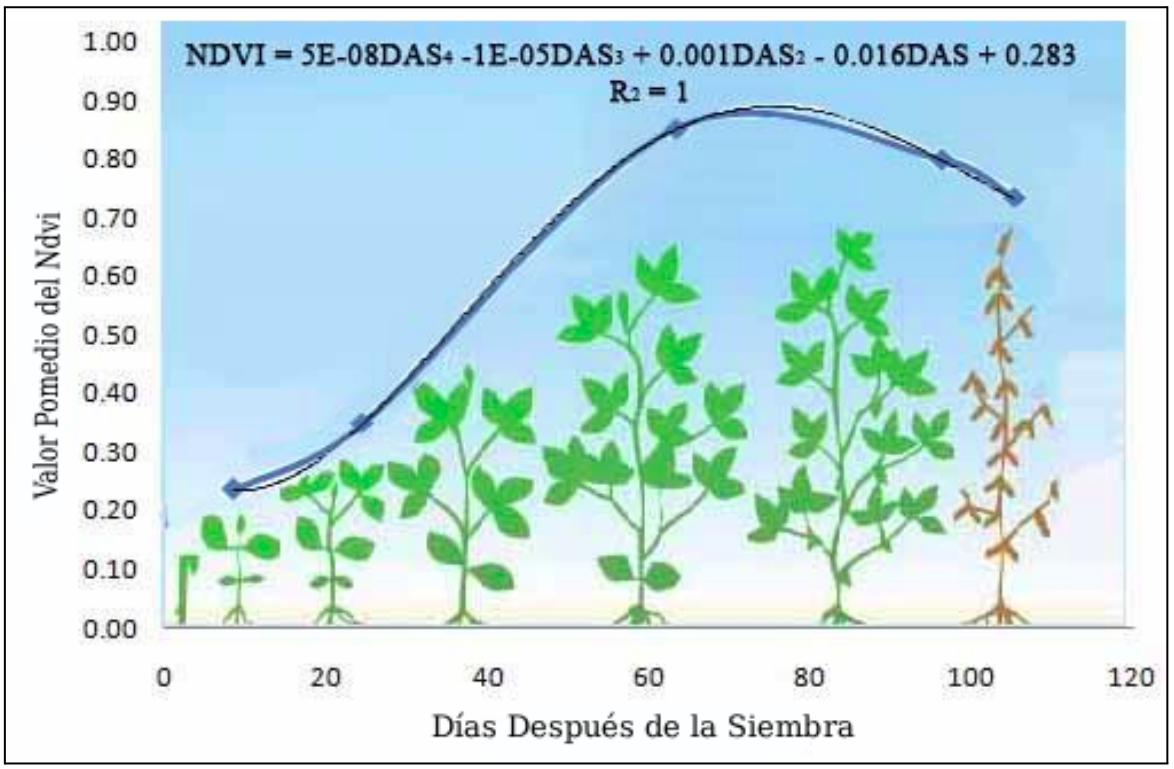
Cuando el IAF de la planta alcanza un valor específico, las hojas ya capturan la mayor parte de la luz solar recibida; por lo que luego de pasar este umbral, si el IAF sigue incrementando la tasa de crecimiento del cultivo ya no lo hace. Si el cultivo no logra alcanzar este valor crítico, la captación de radiación será menor y por lo tanto también lo serán el crecimiento y el rendimiento final de la planta (Toledo, 2018).

Distintas investigaciones sugieren que utilizando el índice NDVI en reemplazo del IAF se obtienen los mismos resultados. Por ejemplo, Bajocco et al. (2022) afirman que el IAF posee una correlación no lineal con el NDVI, por lo que este último índice es también un predictor del área foliar de la planta de soja, y por lo tanto del estadío fenológico en el que se encuentra.

La Figura 6 muestra la relación entre distintos valores de este índice, con la cantidad de días después de la siembra de la planta, en ella se observa que el índice comienza en un valor aproximado a 20, luego sube a medida que la planta crece hasta un valor aproximado de 90, para terminar bajando nuevamente al final del ciclo cuando la planta se seca.

Figura 6

Valor del NDVI según madurez de la planta



Nota. La figura presenta la relación entre los valores promedio del índice NDVI y el período fenológico en el que se encuentra cada día después de la fecha de siembra.

Adaptado de *MONITORAMENTO DA FENOLOGIA DA SOJA IRRIGADA USANDO PERFIS DE SÉRIE TEMPORAL DE NDVI.* Bariani, C & Kersten, Diogo & Victoria, N & Carlesso, R & Petry, Mirta & Peripolli, Mariane (2015).

https://www.researchgate.net/figure/Figura-3-Grafico-dos-valores-medio-para-o-perfil-temporal-de-NDVI-para-o-ciclo-da-soja-e_fig2_293488294

Si bien una figura como esta no permite determinar el período fenológico exacto en el que se encuentra la planta, utilizar un índice calculado a partir de imágenes satelitales permite realizar un seguimiento a distancia del crecimiento del cultivo; esto evita tener que concurrir presencialmente a los lotes, ahorrando costos y tiempo.

CAPÍTULO 3 - SERIES TEMPORALES APLICADAS EN TELEDETECCIÓN

En este capítulo se abordan fundamentos de las series temporales, los modelos estadísticos y la detección de anomalías, junto con su aplicación en teledetección a partir del análisis de la evolución del NDVI en el tiempo.

3.1 - Series Temporales

Una serie de tiempo, o serie temporal, es un conjunto de valores de una variable obtenidos en intervalos de tiempo regulares y ordenados en orden cronológico. El análisis de series de tiempo consiste en extraer información estadística de dichos conjuntos para determinar hechos pasados o predecir el comportamiento futuro (Nielsen, 2019).

Construir series de tiempo del NDVI, índice presentado en la Sección 1.2, permite realizar un análisis sobre el crecimiento de los cultivos de soja, ya sea para estudiar la curva de crecimiento en el pasado, o predecir cómo será en el futuro.

3.2 - Filtro Savitzky-Golay

Como se mencionó al final de la Sección 1.2, los valores del NDVI capturados por el satélite Sentinel-2 suelen contener ruido correspondiente a imágenes con nubosidad alta; una técnica para minimizar este ruido, y por lo tanto suavizar la curva que los representa, es utilizar un filtro de regresión polinomial local propuesto por Savitzky y Golay (1964).

Esta herramienta funciona tomando los valores en una ventana simétrica de amplitud fija alrededor de un centro y aplicando a cada punto dentro de la ventana la ecuación del filtro, que utiliza coeficientes calculados aproximando los valores de la serie por un polinomio mediante mínimos cuadrados[5] (Savitzky y Golay, 1964).

La ecuación del filtro es:

$$y_j^* = \frac{1}{N} \sum_{i=-m}^{m} C_i \, Y_{j+i}$$

donde:

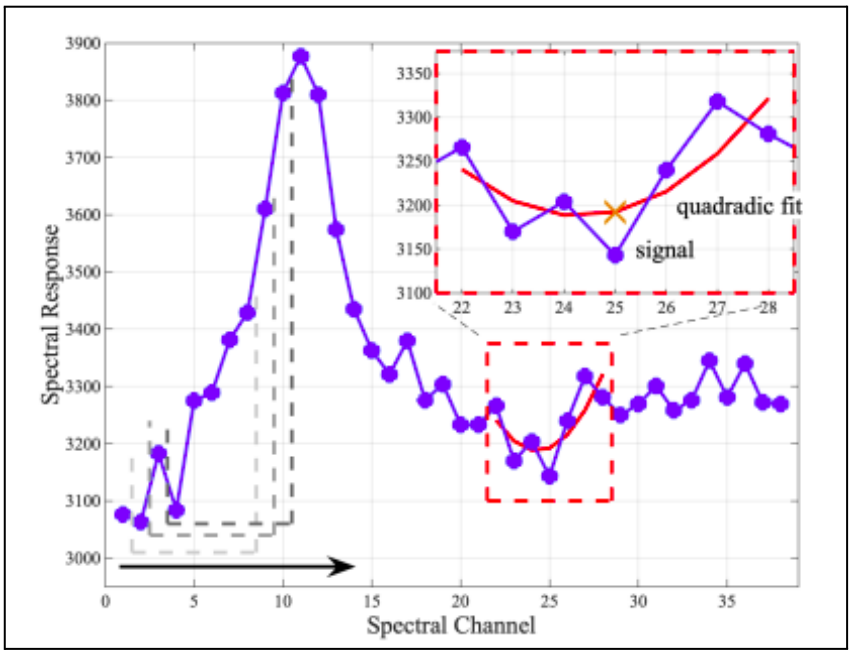
- y_j^* es el valor suavizado calculado por el filtro,
- N es un factor de normalización
- m es la cantidad de puntos a cada lado del centro de la ventana,
- C_i son los coeficientes obtenidos de la aproximación por mínimos cuadrados,
- Y_{j+i} son los valores reales de los datos.

[5] El método de mínimos cuadrados aproxima una función polinómica a los valores de la serie calculando los coeficientes que minimizan la sumatoria de los cuadrados de la diferencia entre los valores reales y los que toma la función en cada paso temporal.

En la figura 7 se ejemplifica este proceso mostrando una curva espectral ruidosa en violeta y, en rojo en el recuadro de la esquina superior derecha, el resultado suavizado obtenido de la aplicación del filtro Savitzky-Golay con una ventana de 7 pasos temporales y un polinomio de grado 2.

Figura 7

Aplicación de filtro Savitzky-Golay



Nota. Figura de ejemplo de aplicación del filtro Savitzky-Golay. Tomado de *Manned-Unmanned Teaming Applied To HEMS Missions: A Path Planning Approach Based On The Pilot's Workload Assessment*. Roncolini, F., & Quaranta, G. (2024). https://www.researchgate.net/figure/Savitzky-Golay-filter-with-a-window-width-of-7-and-a-polynomial-order-of-2-The-filter_fig6_382104894

3.3 - Modelo SARIMA

En el análisis estadístico de series temporales, un modelo muy utilizado es el de Media Móvil Integrada Autorregresiva Estacional (Seasonal Autoregressive Integrated Moving Average, SARIMA), debido a la facilidad de entendimiento de sus parámetros, que no requieren un conjunto de datos muy grande para obtener buenos resultados, y que existen métodos automatizados para determinar los mejores parámetros (Nielsen, 2019).

Según Nielsen (2019), un modelo SARIMA es un modelo ARIMA estacionalizado, el cual a su vez consta de tres partes:

Componente autorregresiva (AR): establece que el valor actual de la serie depende de sus valores pasados. Es decir, modela la variable como una combinación lineal de observaciones anteriores.

Componente de media móvil (MA): modela el valor actual como una función de los errores pasados, en vez de utilizar los valores de la variable estudiada, capturando el efecto de perturbaciones aleatorias previas sobre la serie.

Componente integrada (I): consiste en aplicar diferenciación[6] a la serie una o más veces con el objetivo de eliminar tendencias y volverla una serie estacionaria, condición necesaria para el modelado.

Los modelos ARIMA están representados por tres parámetros: el parámetro “p” de la componente AR, que representa la cantidad de valores de pasos temporales hacia atrás tomados para la autorregresión; el parámetro “q” de la componente MA, que indica la cantidad de errores de pasos temporales hacia atrás que contempla el modelo; y la cantidad de veces que se aplica la diferenciación (I) representada por el parámetro “d” (Nielsen, 2019).

Cuando se asume que los datos siguen tendencias estacionales, se añaden al modelo tres parámetros P, D, y Q, que representan los mismos componentes del modelo ARIMA pero teniendo en cuenta la estacionalidad. Por lo que un modelo SARIMA tendrá un total de 7 parámetros, expresados como ARIMA(p, d, q) (P, D, Q, m), siendo m el número de pasos temporales contenidos en cada ciclo estacional.

A pesar de sus beneficios, la simplicidad del modelo no le permite manejar datos con correlaciones no lineales. Además de que un aumento en la cantidad de datos utilizados para el cálculo de sus parámetros no necesariamente implica una mejora en su rendimiento (Nielsen, 2019). Por estas razones es que en algunos casos es recomendable probar otro tipo de modelos y comprobar si se obtiene mayor rendimiento.

3.3.1 - Determinación de Uso de SARIMA

Un método fácil para determinar si una serie de tiempo dada logra ser descrita por un modelo con los componentes AR y MA descritos en la Sección 3.3, o para encontrar los valores de sus parámetros, es analizar sus funciones de Autocorrelación (Autocorrelation Function, ACF) y Autocorrelación Parcial (Partial Autocorrelation Function, PACF).

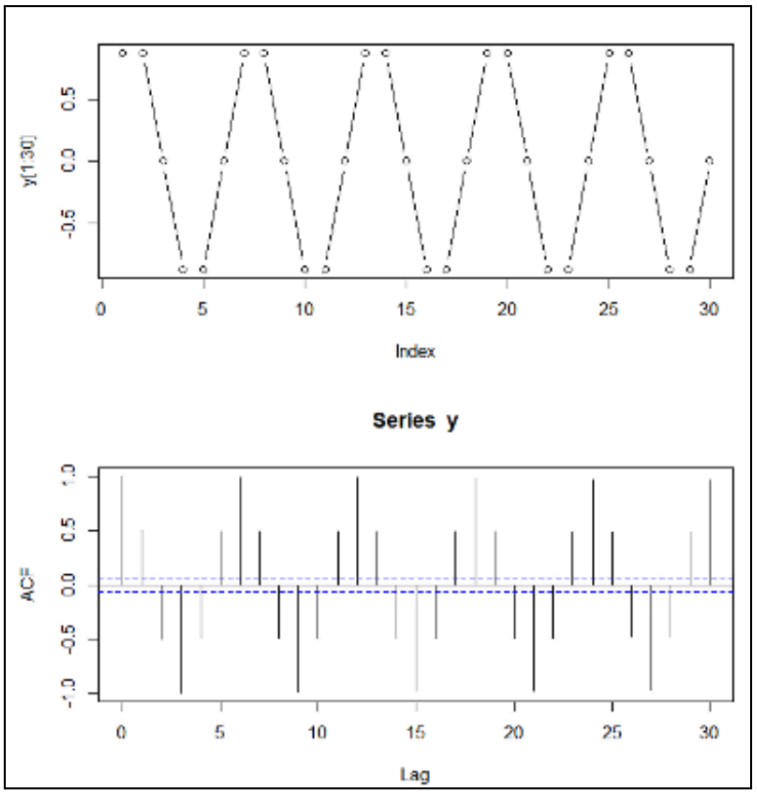
La ACF proporciona información sobre cómo se correlacionan linealmente un valor de la serie con uno de sus rezagos en el tiempo, permitiendo analizar cuánto ayudan los valores pasados a predecir los futuros (Nielsen, 2019).

[6] Este proceso de diferenciación convierte la serie temporal en una nueva serie temporal de la diferencia entre los valores originales de puntos temporales contiguos.

La figura 8 exhibe un ejemplo de una función sinusoidal con un período cercano a 10 y su ACF. En ella se observa cómo a medida que aumenta la cantidad de pasos temporales hacia atrás tomados, a los que se les llama “lag”, el valor de la ACF crece o decrece según la correlación entre el valor del paso temporal actual y el lag tomado. Por ejemplo, para el lag 6 se observa que la ACF toma un valor cercano a 1, lo que indica que los valores de la serie separados por seis períodos tienden a moverse en el mismo sentido (crecen o decrecen a la vez), reflejando la naturaleza cíclica de la serie.

Figura 8

Gráfico de una función sinusoidal y su ACF



Nota. Tomado de *Practical time series analysis: Prediction with Statistics & Machine Learning* (p. 92), A. Nielsen, 2019, O’Reilly Media.

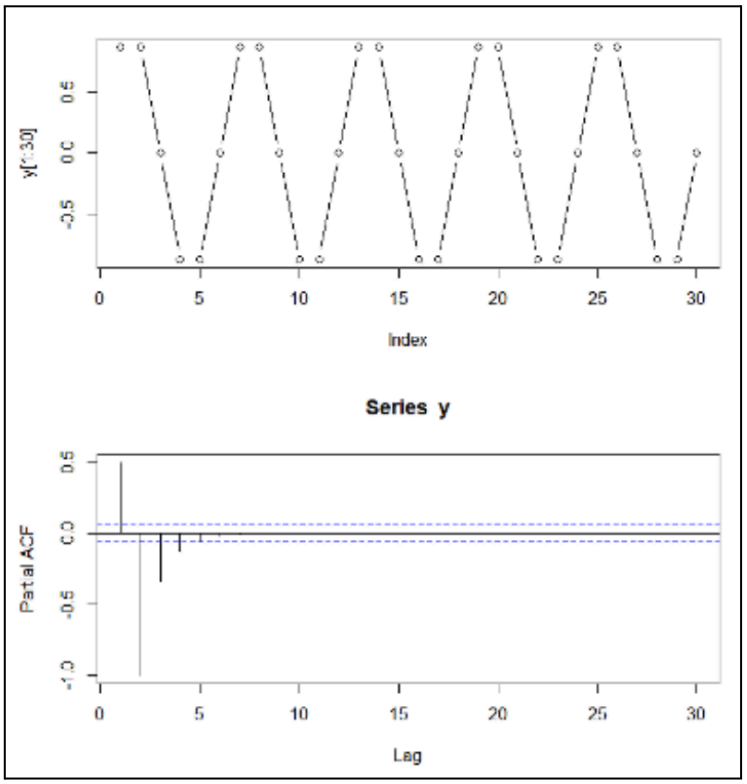
La PACF, a diferencia de la ACF que tiene en cuenta todos los valores intermedios entre los dos puntos temporales tomados para el cálculo, mide la correlación entre un valor de la serie y uno de sus rezagos, eliminando el efecto de los rezagos intermedios.

Un ejemplo de esto se evidencia en la Figura 9, que representa la misma función sinusoidal de la figura anterior junto con su gráfico de la PACF. Observando el lag 2, se aprecia un valor negativo cercano a -1, lo que indica una correlación inversa entre los valores de la serie separados por dos pasos temporales. Sin embargo, a diferencia de la

ACF, esta correlación no es directa, ya que la PACF ajusta por el efecto del lag 1, aislando únicamente la dependencia directa entre dichos pasos temporales. Este comportamiento permite identificar relaciones más específicas en la serie, sin la influencia de correlaciones intermedias.

Figura 9

Gráfico de un proceso estacional sin ruido y su PACF



Nota. Tomado de *Practical time series analysis: Prediction with Statistics & Machine Learning* (p. 93), A. Nielsen, 2019, O'Reilly Media.

En la práctica, los gráficos de estas funciones son de mayor utilidad que el cálculo entre pasos temporales ya que permiten identificar visualmente, de manera más sencilla, qué componentes debe contener el modelo que las define.

Ante esta situación, se analiza el gráfico de la ACF (Fig. 8) y si en él se observa decrecimiento gradual, el modelo debe incluir un componente AR(p); mientras que si se produce un corte abrupto en los datos, sugiere un modelo MA(q) donde el valor de "q" será el número de paso temporal donde se produce el corte.

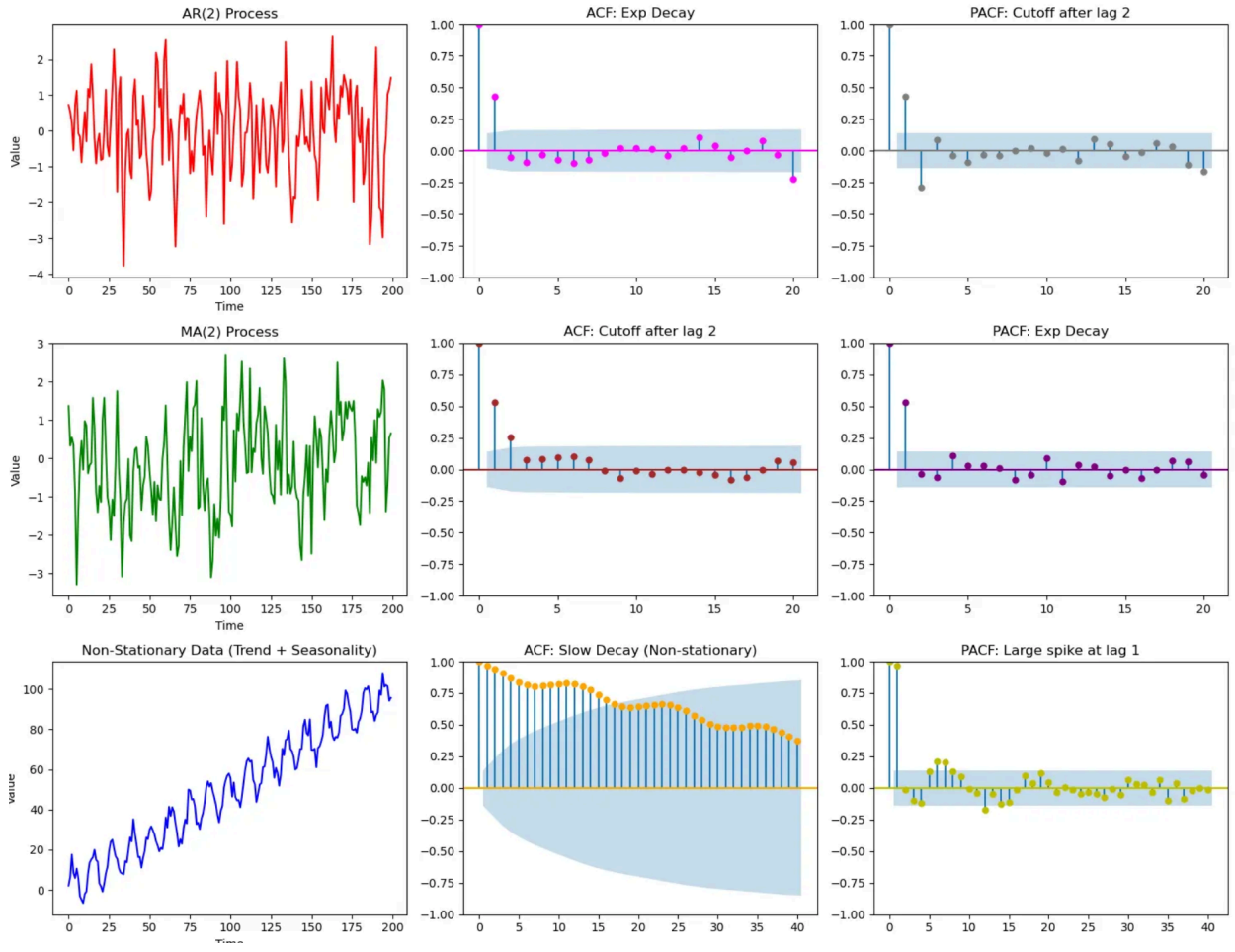
Por el contrario, en el gráfico de la PACF (Fig. 9), un decrecimiento gradual de los valores implica que el modelo debería contener un componente MA(q), y un corte abrupto significa que el modelo debe contener un componente AR(p) donde el valor de p será el número de paso temporal donde se produce el corte.

Como último caso, si en los gráficos de ambas funciones los valores decrecen gradualmente sin cortes abruptos implica que ambos componentes están presentes, por lo que sugiere un modelo ARMA(p, q); en este caso se deben realizar otro tipo de pruebas para determinar los valores adecuados de los parámetros p y q (Nielsen, 2019).

La Figura 10 ejemplifica los tres casos con diferentes series temporales y sus gráficas de ACF y PACF. En ella, las áreas celestes en los gráficos de ACF (segunda columna) y PACF (tercera columna) representan un intervalo de confianza del 95%, los valores que se encuentren dentro de este área no tienen importancia estadística por que lo son considerados cero.

Figura 10

Análisis de ACF y PACF para diferentes series temporales



Nota. Esta figura contiene el gráfico de tres series temporales, la primera es de proceso AR(2), la segunda es de un proceso MA(2), y la tercera de un conjunto de datos no estacionales; junto con la ACF y PACF para cada una de las tres. Estos gráficos ayudan a comprender cómo obtener visualmente valores para los parámetros de modelos AR y MA. Tomado de *ACF and PACF in Time Series Analysis,* Prathik C., 2025, Medium. https://medium.com/@prathik.codes/acf-and-pacf-in-time-series-analysis-c7d32ac8dc39

Para la primera serie temporal (en color rojo) se logra determinar visualmente que la ACF va decayendo en cada lag; mientras que la PACF cuenta con un corte abrupto luego del lag 2, indicando que responde a un proceso representable por un modelo AR(2).

En la segunda serie temporal (en color verde) ocurre que el corte abrupto es encontrado en la ACF, luego del lag 2; mientras que la PACF decae exponencialmente, indicando un proceso descrito por un modelo MA(2).

Por último, en la última serie temporal que representa un conjunto de datos no estacional (en color azul), tanto la ACF como la PACF decaen en cada lag ya sea de manera lenta o abrupta; esto indica que el conjunto representa un proceso expresable por un ARMA(p, q), y requiere una diferenciación para conseguir la estacionariedad en los datos.

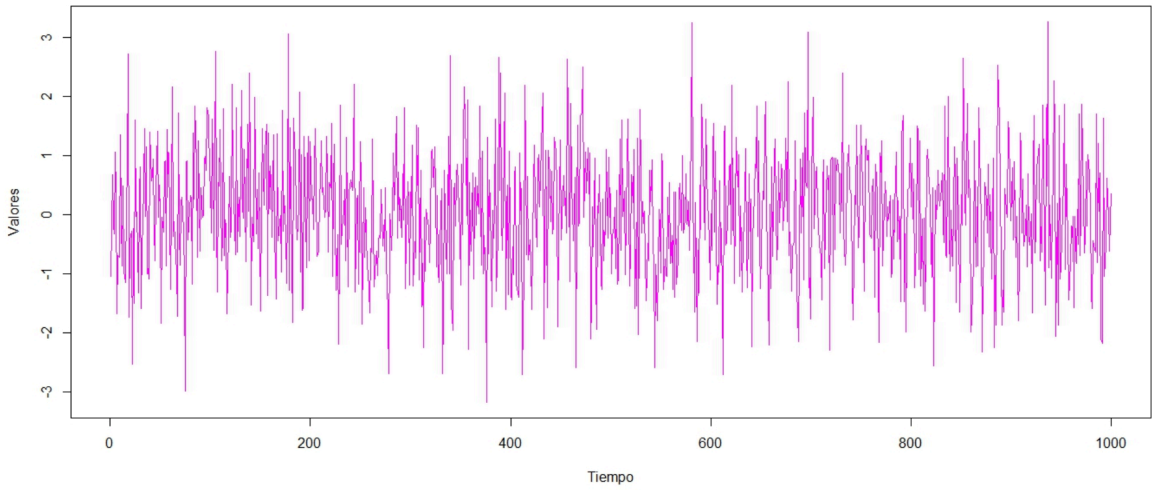
En la práctica, el modelo ARMA es poco utilizado en comparación con el modelo ARIMA por lo que se suele aplicar diferenciación “d” veces para convertir la serie no estacionaria en una estacionaria, y de este modo convertir el modelo ARMA a ARIMA. Además, si la serie no sólo presenta dependencias entre pasos temporales cercanos, sino también patrones repetitivos estacionales o periodicidad a largo plazo, el modelo a utilizar es SARIMA, donde el parámetro de periodicidad m se debe conocer de antemano y P, D, y Q se determinarán de la misma manera que p, d, y q, pero esta vez tomando los rezagos estacionales de la serie para realizar la evaluación.

Finalmente, para asegurar que una serie temporal consigue ser descrita como combinaciones lineales de sus valores y errores pasados por un modelo SARIMA, se debe comprobar que el gráfico de los residuos del modelo, que es la diferencia entre el valor observado y el valor predicho, no sigue un patrón identificable sino que es ruido blanco; esto es, que los residuos del modelo son una señal aleatoria y no guardan correlación estadística (Fernandez Ruiz, V).

Para comprender la estructura de una función de ruido blanco, la Figura 11 contiene un ejemplo de este tipo de estructura. En ella se percibe que para cada paso temporal, el valor que toma la serie no tiene correlación con valores pasados sino que es producido por un proceso con resultado aleatorio.

Figura 11

Ejemplo de Ruido Blanco



Nota. Esta figura ejemplifica una serie de tiempo de ruido blanco, donde se aprecia que los valores no tienen correlación, sino que son aleatorios. Tomado de Ruido Blanco, RPubs, por Fernández Ruiz, V. (2018). https://rpubs.com/vicferu/394452

Si el gráfico de residuos de un modelo SARIMA tiene forma de ruido blanco, implica que el modelo explica la totalidad de la parte predecible de la serie, siendo lo que no alcanza a explicar causado por sucesos imprevisibles. Mientras que si el gráfico de residuos del modelo no contiene esta forma, se debe aplicar otro tipo de modelo ya que SARIMA no logrará captar en su totalidad la dinámica de los datos.

3.3.2 - Criterios de Selección de Modelos Estadísticos

Para la comparación de distintos modelos estadísticos, a fin de seleccionar el más adecuado, se emplean criterios de información. Estos criterios evalúan la relación entre la bondad de ajuste, que mide que tan bien se ajusta el modelo a los datos reales, y la complejidad del modelo, penalizando los modelos con mayor cantidad de parámetros para evitar el sobreajuste (Géron, 2023).

Entre los criterios empleados para este propósito se presenta el Criterio de Información de Akaike (Akaike Information Criterion, AIC) y el Criterio de Información Bayesiano (Bayesian Information Criterion, BIC), cuyas ecuaciones son:

$$AIC = 2p - 2\,log(\widehat{\ell})$$

$$BIC = log(m)p - 2\,log(\widehat{\ell})$$

donde:

- p es el número de parámetros aprendidos por el modelo; en general, el parámetro que se busca minimizar,

- $\widehat{\ell}$ es el valor máximo de la función de verosimilitud; en general, el parámetro que se busca maximizar,
- m es el número de puntos en la serie temporal.

Generalmente, ambos criterios seleccionan el mismo modelo como el mejor. En caso que esto no suceda, el modelo seleccionado por BIC tiende a ser de menos parámetros, mientras que el seleccionado por AIC tiende a ser el de mayor precisión.

En la práctica, se emplean herramientas que automatizan la selección de estos parámetros. Un ejemplo es la librería pmdarima del lenguaje Python, la cual implementa algoritmos de búsqueda como “auto_arima”, permitiendo estimar de forma eficiente los órdenes del modelo a partir de criterios de información como AIC o BIC. Esto es útil en cuando se requiere explorar configuraciones de parámetros posibles, ya que reduce la intervención manual y acelera el proceso de modelado, aunque puede requerir un mayor costo computacional.

3.4 - Detección de Anomalías en Series Temporales

Un aporte fundamental en estadística es el Teorema del Límite Central, el cual establece que, al tomar una muestra aleatoria simple[7] de tamaño *n* de una población con media μ, y varianza σ^2 conocida, la distribución de la media muestral $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ se aproxima a una distribución normal $N(\mu,\frac{\sigma 2}{n})$ cuando *n* es suficientemente grande, incluso si la población original no sigue una distribución normal; donde la noción de “tamaño suficiente” depende de la forma de la distribución original (Navidi, 2006). La relevancia de este teorema radica en que permite asumir un comportamiento aproximadamente normal en los datos bajo condiciones generales.

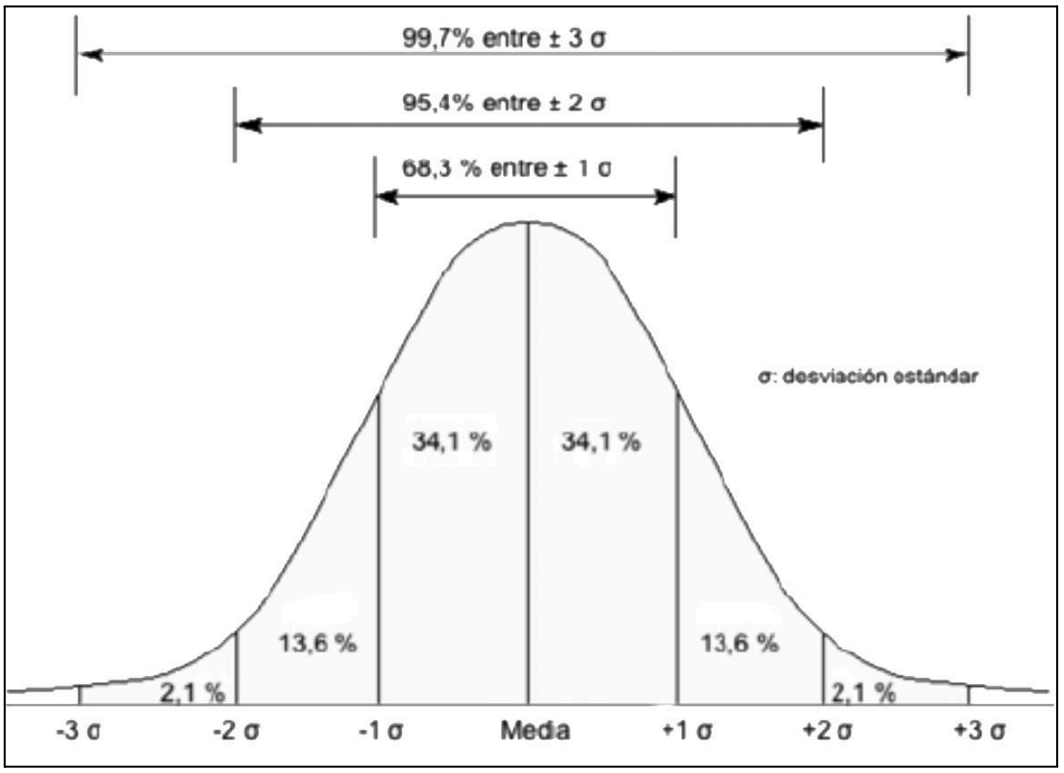
A partir de esta aproximación, es posible definir criterios para la detección de valores atípicos en un conjunto de datos. En particular, en uno con distribución aproximadamente normal, una observación se clasifica como anómala si se desvía significativamente de su media.

Un criterio ampliamente utilizado es la regla de las tres desviaciones estándar, según la cual un valor *x* se considera anómalo si cumple con $|x - \mu| > 3\sigma$, como lo grafica la Figura 12.

[7] Una muestra aleatoria simple está formada por elementos de una población en la que cada colección de n elementos tiene la misma probabilidad de ser elegida para la muestra (Navidi, 2006).

Figura 12

Distribución normal y sus porcentajes respecto de la desviación estándar.



Nota. Esta figura ilustra una distribución normal y el porcentaje de los datos contenidos en los rangos de 1, 2, y 3 desviaciones estándar de la media. Tomado de *Estadística Descriptiva e Inferencial*, por Romero Aroca, P., Garcia, C., Gonzalez Lopez, J., (2013) *https://www.researchgate.net/figure/Distribucion-normal-y-sus-porcentajes-respecto-de-la-desviacion-estandar_fig1_275021043*

Este umbral se basa en que, en una distribución normal, aproximadamente el 99,7% de los valores se encuentran dentro del intervalo [μ - 3σ, μ + 3σ], por lo que resulta poco frecuente observar valores fuera de este rango (Navidi, 2006).

CAPÍTULO 4 - INTELIGENCIA ARTIFICIAL APLICADA A SERIES TEMPORALES

En el presente capítulo se exponen conceptos de Inteligencia Artificial (IA), adentrándose en una de sus ramas más utilizadas actualmente, el Aprendizaje Automático. Se estudia cómo funciona una Red Neuronal Recurrente, utilizando un tipo de neurona específica para el aprendizaje automático en series temporales.

Actualmente no existe una definición única de IA, Russell y Norvig (2004) mencionan que a lo largo de la historia distintos autores han seguido uno de cuatro enfoques para su definición: sistemas que piensan como humanos, sistemas que actúan como humanos, sistemas que piensan racionalmente, y sistemas que actúan racionalmente. Un ejemplo de este último enfoque es la definición del Grupo de Expertos de Alto Nivel en Inteligencia Artificial, que la describe como “sistemas que muestran un comportamiento inteligente al analizar su entorno y tomar acciones –con cierto grado de autonomía– para lograr objetivos específicos …” (High-Level Expert Group on Artificial Intelligence [AI HLEG], 2018, p.1).

4.1 - Machine Learning

Una de las ramas del campo de la IA más utilizadas en la actualidad es el Aprendizaje Automático o Machine Learning (ML). En el área de ML, el objetivo es desarrollar algoritmos que “aprendan” de los datos y construyan modelos que representen sus características, para así poder realizar predicciones o tomar decisiones, sin estar programados específicamente para cada tarea.

Una definición más formal es la propuesta por Mitchell (1997), quien establece que un sistema computacional “aprende” a partir de una experiencia *E*, respecto de una tarea *T* y una medida de rendimiento *R*, si su desempeño en la tarea *T*, evaluado mediante *R*, mejora con la experiencia *E*. Esto es, un programa aprende cuando es capaz de incrementar progresivamente su rendimiento en una tarea específica a medida que acumula experiencia al resolverla de forma autónoma en múltiples ocasiones.

Según el tipo de entrenamiento que se utiliza para el ajuste de los modelos se agrupa a los sistemas de ML como aprendizaje supervisado, aprendizaje no supervisado, y aprendizaje semisupervisado, entre muchas otras categorías (Géron, 2023).

El aprendizaje es supervisado cuando el conjunto de entrenamiento incluye los datos que se toman como “solución” a la tarea, llamados etiquetas. Ejemplos de tareas que utilizan este tipo de entrenamiento son la clasificación y predicción. Mientras que en el aprendizaje no supervisado, las etiquetas no están incluidas en el conjunto de entrenamiento, sino que el algoritmo aprende cuáles son por su cuenta; este tipo de

aprendizaje incluye tareas como el agrupamiento, la reducción de dimensionalidad, o la detección de anomalías. Y por último, el aprendizaje semisupervisado es una categoría intermedia donde sólo una fracción mínima del conjunto de entrenamiento posee etiqueta, mientras que el resto de ellas son generadas por el sistema.

4.1.1 - Consideraciones para un correcto entrenamiento de modelos de Machine Learning

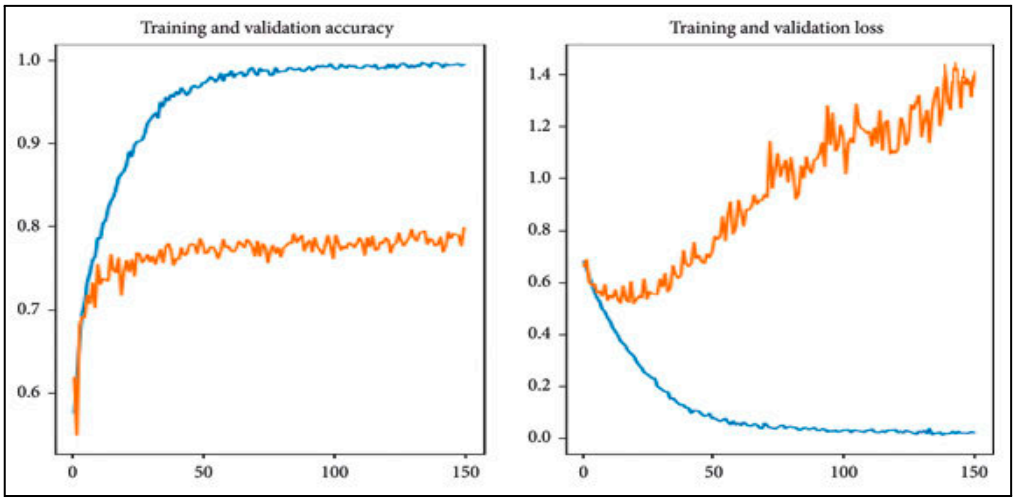
Al momento de trabajar con técnicas de aprendizaje automático se debe considerar la cantidad mínima de datos necesarios, y que éstos no sean de baja calidad; ya que cantidades insuficientes de datos, o datos que no representan la realidad, harán que los modelos de ML no logren describir los datos reales de manera precisa, resultando en un modelo ineficiente.

Además, al momento de entrenar los modelos se debe tener en cuenta no realizar un sobreajuste a los datos de entrenamiento, esto se produce cuando el modelo se entrena para representar demasiado bien las características del conjunto de entrenamiento pero al llevarlo al conjunto real no obtiene el mismo rendimiento.

La Figura 13 ilustra un caso de sobreajuste mediante las curvas de exactitud (izquierda) y pérdida (derecha), para los conjuntos de entrenamiento (azul) y validación (naranja) de un modelo.

Figura 13

Sobreajuste visualizado en curvas de exactitud y pérdida



Nota. Esta figura grafica un ejemplo de sobreajuste del modelo a los datos de entrenamiento. Tomado de *Deep CNN and Deep GAN in Computational Visual Perception-Driven Image Analysis*. https://www.researchgate.net/figure/Overfitting-a-Accuracy-curve-after-dropout-b-Loss-curve-after-dropout_fig9_350793697

Se observa que, a medida que avanza el entrenamiento, la exactitud en el conjunto de entrenamiento aumenta de forma sostenida, mientras que la exactitud en validación se estabiliza. De manera complementaria, la pérdida en entrenamiento decrece progresivamente, en contraste con la pérdida en validación, que muestra una tendencia creciente. Este comportamiento evidencia que el modelo se ajusta en exceso a los datos de entrenamiento, capturando patrones específicos que no se generalizan adecuadamente a datos nuevos. En un entrenamiento adecuado, se esperaría que las curvas de entrenamiento y validación presenten formas similares, sin divergencias significativas.

Otro problema es el subajuste del modelo, que constituye el caso opuesto al sobreajuste y ocurre cuando el modelo es demasiado simple para capturar adecuadamente las características subyacentes de los datos (Géron, 2023). En esta situación, tanto la curva de exactitud en el entrenamiento como la de validación se mantienen en valores bajos, mientras que la pérdida permanece relativamente alta en ambos conjuntos, sin mostrar mejoras significativas a lo largo del entrenamiento. Esto indica que el modelo no logra aprender los patrones relevantes del problema.

Para evitar estos dos últimos problemas se suele reservar una parte del conjunto de entrenamiento para realizar una validación, generalmente un 20% del conjunto inicial. A este subconjunto se le llama Conjunto de validación, y es importante que se sea utilizado hasta después que termine el ajuste del modelo para verificar el resultado del ajuste. De este proceso se obtiene una medición del rendimiento del modelo sobre datos no vistos con anterioridad, por lo que permite tomar acciones de manera correspondiente, para prevenir un sobreajuste o un subajuste.

Al momento de entrenar un modelo por primera vez, no se conoce cuáles serán los hiperparámetros que lograrán el mejor rendimiento. Por esto, es que generalmente se recurre a alguna técnica de optimización de hiperparámetros, las cuales consisten en realizar una búsqueda de los parámetros que optimizan los resultados del modelo. Existen diferentes técnicas para realizar esta búsqueda, entre las que destacan por su simplicidad de implementación la búsqueda en red, que realiza un barrido de todas las combinaciones posibles de hiperparámetros; y la búsqueda aleatoria, que sigue una lógica similar a la búsqueda en red pero no respeta un orden específico en los valores de hiperparámetros probados, sino que son escogidos de forma aleatoria.

4.1.2 - Métricas de error de modelos de Machine Learning

Para realizar una comparación entre modelos, es necesario evaluar su desempeño mediante métricas de error. Estas métricas son medidas cuantitativas que permiten estimar qué tan cerca se encuentran los valores predichos por un modelo respecto de los valores reales, proporcionando un criterio objetivo para comparar distintos enfoques. En términos generales, valores más bajos de estas métricas indican un mejor desempeño del modelo, ya que reflejan una menor diferencia entre las predicciones realizadas y los datos reales.

Una de las métricas más utilizadas es el Error Cuadrático Medio (Mean Squared Error, MSE), definido como el promedio de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo. Matemáticamente, se expresa como $\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ donde n es la cantidad de datos, y el valor observado y $\widehat{y}$ el valor predicho por el modelo (Hodson, 2022). Esta métrica, al utilizar el cuadrado de la diferencia, pondera con mayor peso los errores grandes por lo que es muy sensible a valores atípicos. Por esta razón es que en ocasiones se elige utilizar la Raíz del Error Cuadrático Medio (Root Mean Squared Error, RMSE), que devuelve el error en la misma unidad que los datos del conjunto, mejorando la interpretabilidad de los resultados. La ecuación del RMSE es $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}$ (Hodson, 2022).

4.2 - Aumentación de Datos

Una solución a los problemas de falta de datos, y datos de baja calidad, expuestos en la sección 4.1.1, es la creación de datos artificiales. Este proceso se llama aumentación de datos, y se define como una técnica de generación de nuevos datos sintéticos a partir de los datos existentes, con el objetivo de incrementar la cantidad y diversidad del conjunto de datos disponible. Esta estrategia es utilizada cuando los datos son escasos, y permite mejorar la capacidad de los modelos de predicción para capturar patrones y realizar estimaciones más precisas (Pugnaire, 2023).

Aunque al momento de utilizarla se debe prestar especial atención a que los datos creados mantengan coherencia con las características del conjunto de datos original. Es fundamental que las nuevas muestras no distorsionen los patrones reales presentes en los datos; de lo contrario, el modelo podría aprender representaciones erróneas, afectando negativamente su capacidad de generalización y precisión en la predicción.

Como expone Pugnaire (2023), operaciones tradicionales utilizadas para la aumentación de datos en series temporales son:

- Time warping: que introduce variaciones en la velocidad de la serie temporal, modificando de manera controlada la progresión temporal sin alterar su duración.
- Convolución: que aplica una operación matemática entre la serie y una ventana de pasos temporales definida, permitiendo resaltar patrones o características específicas mediante la combinación de valores vecinos.
- Pooling: que reduce la resolución temporal agrupando pasos en intervalos y aplicando una función (como el cálculo de la media, máximo o mínimo) para obtener una representación más generalizada de cada grupo.
- Introducción de ruido: añadir perturbaciones aleatorias a cada punto de la serie, siguiendo una distribución estadística, con el fin de incrementar la variabilidad de los datos.

Todas las operaciones listadas mantienen la cantidad original de pasos temporales (Pugnaire, 2023), por lo que son de especial utilidad en series temporales ya que permiten aplicar la aumentación de datos sin romper la estructura temporal original.

CAPÍTULO 5 - REDES NEURONALES Y DEEP LEARNING

Este capítulo introduce los fundamentos de las redes neuronales artificiales y el aprendizaje profundo, partiendo de un primer modelo básico y explicando cómo las neuronas artificiales procesan información mediante sumas ponderadas y funciones de activación no lineales. Luego se aborda la evolución hacia arquitecturas más complejas con múltiples capas, enfocándose redes neuronales recurrentes especialmente diseñadas para el análisis de series temporales. Finalmente, se presentan las celdas Long Short-Term Memory como una solución avanzada que permite capturar dependencias de largo plazo mediante mecanismos internos de memoria.

5.1 - Redes Neuronales Artificiales

Inspirados en el funcionamiento de las neuronas biológicas, Warren McCulloch y Walter Pitts (1943) propusieron un modelo computacional basado en lógica proposicional, capaz de representar funciones mediante la interconexión de múltiples unidades, al que denominaron Redes Neuronales Artificiales (RNA). Una forma de interpretar las RNA es como grafos dirigidos, donde cada nodo representa una neurona y cada conexión posee un valor asociado, denominado peso sináptico. Estos pesos se ajustan automáticamente durante el proceso de entrenamiento en función del error cometido por la red al verificar su salida con los datos reales, por lo que se dice que aprende a partir de datos.

Posteriormente, Frank Rosenblatt (1958) continuó desarrollando esta idea al introducir el modelo conocido como perceptrón. Este modelo calcula primero una suma ponderada de las entradas, expresada mediante la siguiente ecuación:

$$\sum_{j=0}^{n} W_j \cdot a_j + b$$

donde a_j es el valor de entrada de la neurona, W_j sus pesos sinápticos asociados, y *b* es un término de sesgo. Luego, se aplica una función de activación para determinar la salida de la neurona.

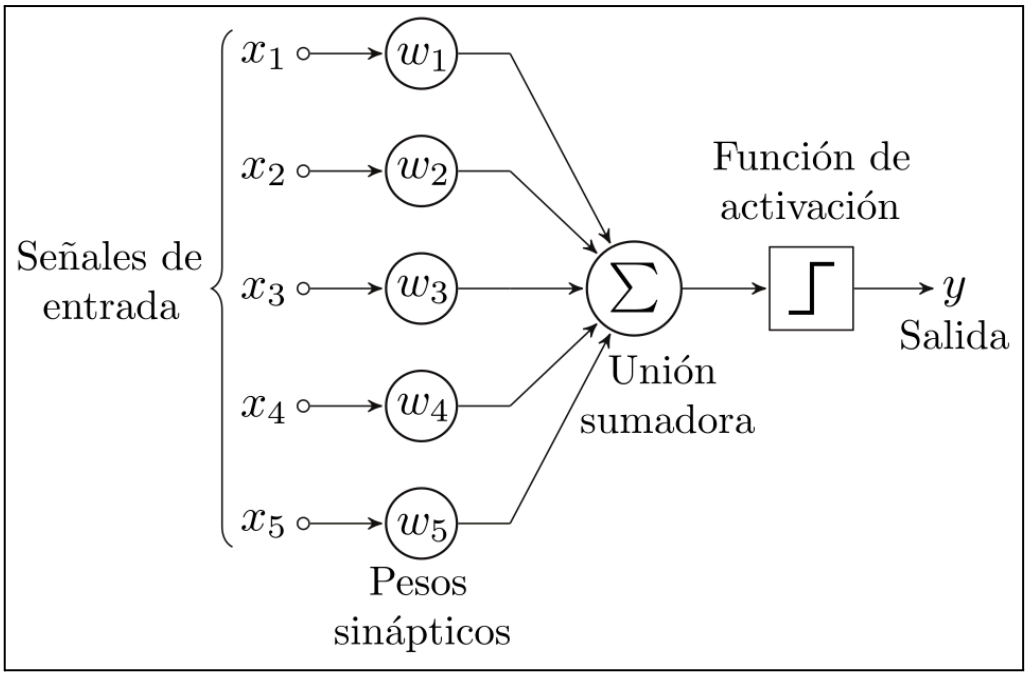
El término de sesgo se agrega a cada neurona de la red, y está conectado a una entrada fija con valor -1 (Russell, S., Norvig, P., 2004). Este término añade flexibilidad a la neurona, permitiéndole activarse aún si los valores de entrada no logran alcanzar el umbral de la función de activación.

La Figura 14 es una representación gráfica de una red neuronal simple, llamada Perceptrón Simple, que consta de una sola neurona con cinco entradas ***x***, cinco pesos sinápticos ***w***, y que produce una única salida ***y*** determinada por una función de activación. En esta imagen, y en la bibliografía en general, el término de sesgo es

obviado ya que se da por sentado que el lector los tendrá en cuenta aunque no sean explícitamente representados en el gráfico.

Figura 14

Perceptrón Simple



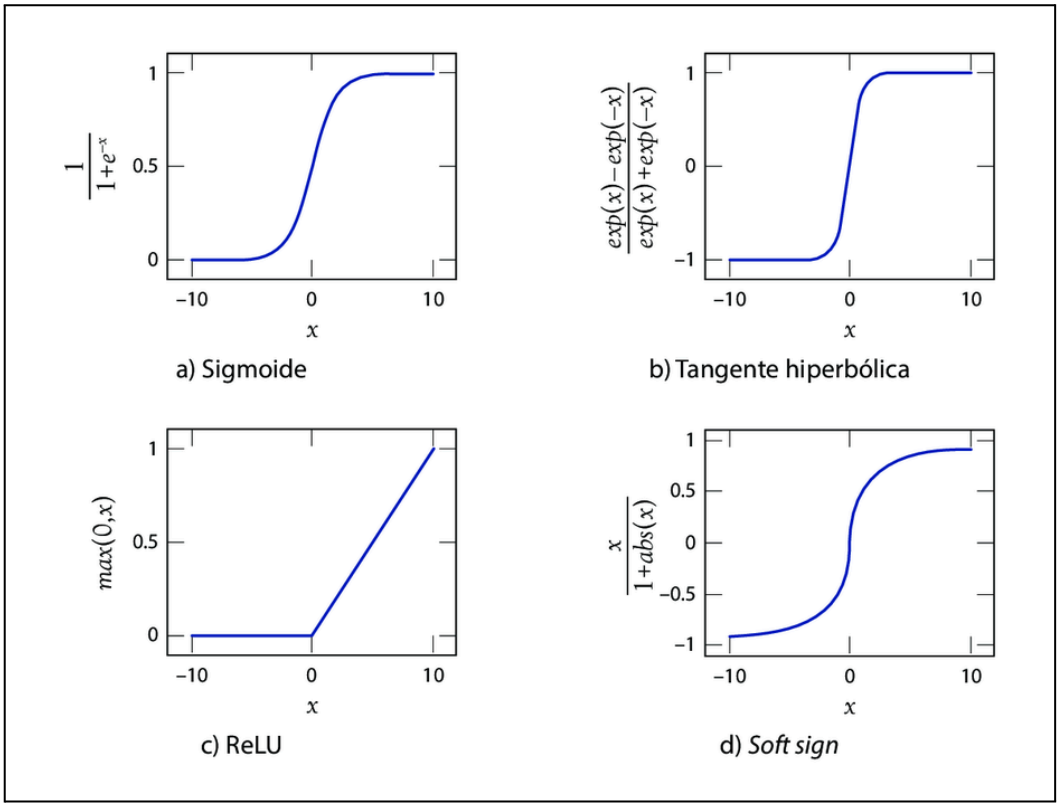
Nota. Esta figura grafica una red neuronal conocida, llamada Perceptrón Simple. Tomado de *Perceptrón 5 unidades.* Cartas, A. (2015). Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Perceptr%C3%B3n_5_unidades.svg

La función de activación determina la salida de cada neurona, tomando ésta valor positivo si la función de activación también lo toma. Estas funciones son diseñadas para que el valor de salida tome un valor cercano a uno cuando las entradas de la neurona sean las correctas, y un valor cercano a cero cuando sean erróneas. Además deben tener una activación no lineal, para permitir que la red aprenda patrones más complejos que los representados en un plano o hiperplano (Russell, S., Norvig, P., 2004).

Ejemplos de funciones de activación ampliamente utilizadas en la literatura se presentan en la Figura 15. En la esquina superior izquierda la función Sigmoide con ecuación $\frac{1}{1+e^{-x}}$, en la esquina superior derecha la función Tangente Hiperbólica con ecuación $\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$, en la esquina inferior izquierda la función ReLU que en cada *y* toma el valor máximo entre 0 y *x*, y en la esquina inferior derecha la función Soft Sign con ecuación $\frac{x}{1+|x|}$.

Figura 15

Funciones de activación



Nota. La figura presenta cuatro funciones de activación más utilizadas actualmente para redes neuronales artificiales. Tomado de *Las neuronas artificiales y su papel central en la inteligencia artificial*, Garduno-Alvarado, Tzolkin & Troncoso, Feliú & Wolf Iszaevich, Gunnar. (2025). https://www.researchgate.net/figure/Figura-3-Algunas-funciones-de-activacion_fig1_387969679

Este tipo de modelo puede extenderse a arquitecturas más complejas, compuestas por una capa de entrada, una capa de salida y una o más capas intermedias denominadas capas ocultas, formadas por múltiples neuronas de distintos tipos. Las capas en las que cada neurona mantiene conexiones con todas las neuronas de la capa adyacente se conocen como capas completamente conectadas. Asimismo, cuando una red posee múltiples capas ocultas es denominada Red Neuronal Profunda, por lo que a su área de estudio se la conoce como Deep Learning (DL) o aprendizaje profundo.

El entrenamiento de los modelos de DL se basa en la optimización de sus pesos de conexión con el objetivo de minimizar el error de predicción. Para ello se utiliza un algoritmo llamado retropropagación, el cual se enmarca dentro de los métodos de aprendizaje supervisado definidos en la Sección 4.1, y consiste en hacer una pasada por las neuronas de la red tomando los valores actuales de los pesos, y generando con ellos un resultado de salida la red. Con esa salida y el valor real de la etiqueta predicha se calcula una medida del error de la red, que es utilizada para ajustar los pesos mediante

una ecuación que permite ajustar sólo los que influyeron en el error de la salida. Este proceso de dos etapas se repite varias veces, ajustando en cada iteración los pesos de conexión para que logren captar los patrones en los datos, y cuando se les de como entrada un dato nuevo puedan relacionarlo a los ya aprendidos (Géron, 2023).

La retropropagación hace uso del concepto de vector gradiente del error para el ajuste de los pesos de conexión. Este vector $\nabla f = (\frac{\delta f}{\delta x}, \frac{\delta f}{\delta y}, \frac{\delta f}{\delta z})$, se compone de las derivadas parciales del error de la red con respecto al peso asociado a cada neurona . Al usar el negativo de este vector -∇f, se obtiene la dirección máxima de disminución del error de la red. Haciendo uso de este concepto, el algoritmo ajusta de a pasos pequeños los valores de los pesos en la dirección de este vector; la magnitud de estos pasos es determinada por un parámetro llamado tasa de aprendizaje, que es configurable a la hora de entrenar la red (Mitchell, 1997).

5.2 - Redes Neuronales Recurrentes

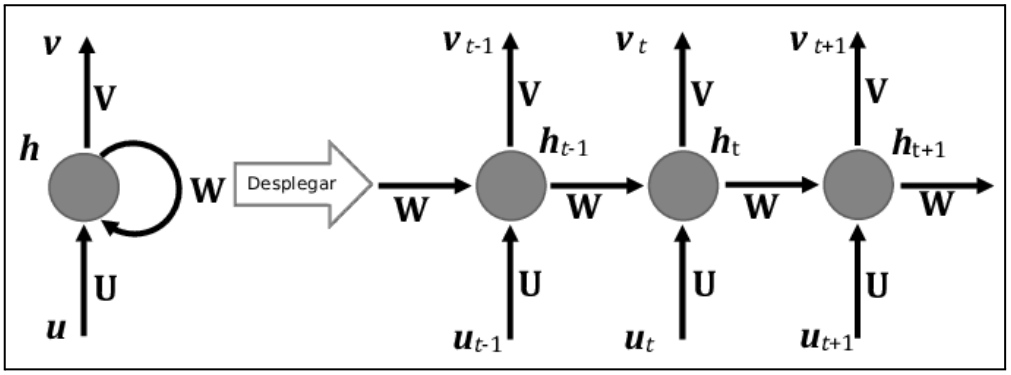
La aplicación de métodos de DL al análisis de series temporales constituye una línea de investigación relativamente reciente. Sin embargo, estos métodos han demostrado capacidad suficiente para modelar comportamientos no lineales complejos presentes en las series de tiempo analizadas (Nielsen, 2019). Las redes neuronales empleadas para este tipo de tareas se denominan Redes Neuronales Recurrentes (RNR).

Las RNR difieren de las RNA presentadas en la sección 5.1 en que las conexiones entre las neuronas pueden no solo ser hacia la capa siguiente, sino también hacia sí misma, como se observa en la Figura 16.

Este esquema muestra una RNR compuesta por una celda recurrente (izquierda) y su despliegue en el tiempo (derecha). En él U, V y W representan matrices de pesos asociadas respectivamente a la entrada u, al estado oculto h, y a la salida v. El despliegue en el tiempo indica que, en cada paso temporal t, la red recibe como entrada tanto el vector actual u_t, como el estado oculto del paso anterior h_{t-1}, el cual actúa como memoria del sistema.

Figura 16

Red Neuronal Recurrente Simple



Nota. Red Neuronal Recurrente con una sola neurona (izquierda) y su despliegue en el tiempo (derecha). Adaptado de *A recurrent neural network-accelerated multi-scale model for elasto-plastic heterogeneous materials subjected to random cyclic and non-proportional loading paths. Wu, Ling & Nguyen, Van Dung & Nanda Gopala, Kilingar & Noels, Ludovic* (2020). https://www.researchgate.net/figure/Recurrent-Neural-Network-with-a-hidden-state-passing-trough-the-input-series-U-V-and-W_fig2_342625605

Gerón (2023) clasifica las RNR según el manejo de sus entradas y salidas en:

- Redes Secuencia a Secuencia: caracterizadas por utilizar, para cada paso temporal, la salida actual y la secuencia de datos de entrada original como entrada en el paso temporal siguiente.
- Redes Secuencia a Vector: caracterizadas por tomar una secuencia de datos de entrada e ignorar las salidas de los pasos temporales intermedios, utilizando únicamente la salida del último de ellos.
- Redes Vector a Secuencia: caracterizadas por introducir a la red todos los datos de entrada a la vez en un vector, en lugar de una secuencia que incorpora un dato por paso temporal.
- Red Codificador-Decodificador: son un caso específico de RNR donde se utiliza una red Secuencia a Vector seguida inmediatamente de una red Vector a Secuencia.

Cada uno de estos tipos de RNR funciona mejor para un conjunto específico de problemas, siendo las de Secuencia a Secuencia las más utilizadas tradicionalmente para trabajar con series temporales como las presentadas en la Sección 3.1.

La capacidad de las neuronas de una RNR para incorporar información proveniente de estados anteriores en el procesamiento de nuevas entradas les confiere una memoria interna temporal, motivo por el cual suelen denominarse celdas de memoria. Esta característica permite capturar dependencias y patrones de corto plazo presentes en una serie temporal. Sin embargo, cuando se trabaja con secuencias

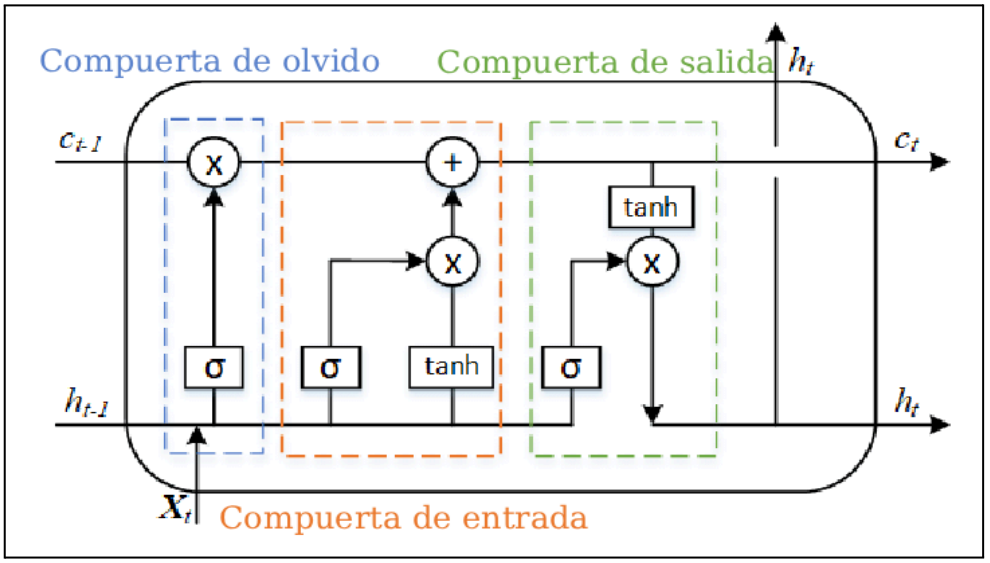
extensas, las RNR convencionales presentan limitaciones para preservar la información relevante a lo largo de numerosos pasos temporales. Como consecuencia, la influencia de los estados más antiguos tiende a disminuir progresivamente, dificultando el aprendizaje de dependencias de largo plazo y reduciendo la capacidad de la red para utilizar información proveniente de etapas tempranas de la secuencia.

5.3 - Celda Long Short Term Memory

Para resolver el problema de la pérdida de memoria de la red, se crearon distintos tipos de celdas de memoria más complejas como la Celda de Memoria a Corto Plazo Duradera (Long Short-Term Memory, LSTM). Estas celdas poseen compuertas lógicas internas que le permiten decidir qué datos se desechan y cuáles se conservan en cada paso temporal (Géron, 2023). La Figura 17 grafica una de estas celdas.

Figura 17

Celda LSTM



Nota. Esta figura representa una celda Long Short-Term Memory y sus componentes internos. En ella, σ representa la función sigmoide logística, que produce valores en el intervalo [0,1]; *tanh* corresponde a la función tangente hiperbólica, que genera valores en el intervalo [−1,1]. El símbolo × indica una multiplicación elemento a elemento entre vectores, mientras que el símbolo + representa una suma elemento a elemento. Las líneas horizontales superiores corresponden al estado de celda c_t, y las inferiores al estado oculto h_t. Por último las flechas indican el flujo de los datos entre los distintos componentes.

Adaptado de *The Prediction of Sea State Parameters by Deep Learning Techniques using Ship Motion Data*. Mittendorf, Malte & Nielsen, Ulrik & Bingham, Harry. (2022). https://www.researchgate.net/figure/LSTM-cell-according-to-Hochreiter-and-Schmidhuber-1997_fig1_358931161

En esta figura se aprecian la compuerta de olvido (componentes dentro del área punteada azul), que es la encargada de decidir que se borra del estado a largo plazo en cada paso temporal; la compuerta de entrada (componentes dentro del área naranja), que decide qué partes del estado a corto plazo y las entradas se sumarán al estado a largo plazo; y la compuerta de salida (componentes dentro del área punteada verde), que decide qué parte del estado a largo plazo se utiliza como salida en cada paso temporal. Gerón (2023) explica que estas compuertas dotan a la red de dos estados que actualiza a la vez en cada paso temporal analizado, uno a corto plazo que se utilizará como entrada en el paso temporal siguiente, y otro a largo plazo que mantendrá guardados los patrones más complejos de la serie temporal.

CAPÍTULO 6 - METODOLOGÍA PARA EL SEGUIMIENTO DE CULTIVOS

Sobre la base teórica expuesta en los capítulos anteriores, en el presente apartado se propone una metodología para un correcto monitoreo del índice NDVI, introducido en la Sección 1.2, aplicado a cultivos de soja; cumpliendo así el objetivo general de esta investigación.

La metodología que se propone es luego puesta en práctica haciendo uso de las herramientas de teledetección exhibidas en la Sección 1.1 para el cálculo del índice, y creando con él series temporales, presentadas en la Sección 3.1, las cuales son utilizadas para el ajuste de un modelo estadístico SARIMA, presentado en la Sección 3.3, y una Red Neuronal Recurrente, presentadas en la Sección 5.2.

Metodología propuesta

La metodología consta de los siguientes pasos:

- A) Análisis del área estudiada: primeramente se debe caracterizar la región identificando eventos ambientales que influyen en la dinámica de crecimiento de los cultivos de soja como los listados en la Sección 2.2; incluyendo clima, precipitaciones, vientos, tipo de suelo, y fenómenos climáticos anómalos.
- B) Extracción y transformación de datos: seguido, se debe obtener las imágenes satelitales, filtrando las que contengan nubosidad superior al 20%; con ellas calcular luego el NDVI promedio por fecha de observación, y crear una serie temporal de este índice sobre el lote analizado; además, se debe aplicar a dicha serie un filtro de reducción de ruido como Savitzky-Golay, definido en la Sección 3.2, para permitir un apropiado ajuste de los modelos.
- C) Aumento de datos: en caso de que la serie de tiempo del NDVI obtenida no contenga una cantidad suficiente de pasos temporales para el correcto ajuste de los modelos, se deben generar datos adicionales mediante técnicas de aumentación de datos, como las presentadas en la Sección 4.2.
- D) Ajuste de modelos: luego se debe utilizar la serie temporal creada para el ajuste de los modelos y contrastar las predicciones generadas por ellos contra una serie temporal real del NDVI del lote, de manera de determinar su eficacia en el modelado del crecimiento del cultivo tomando como comparación métricas como las expuestas en la Sección 4.1.2.

- E) Implementación para detección de anomalías: por último, utilizando la serie temporal de NDVI generada por los modelos como base para establecer la dinámica normal de crecimiento en el lote, se debe determinar un valor para el umbral de observaciones consideradas anómalas, técnica estudiada en la Sección 3.4, y de esta forma lograr identificar desviaciones en el crecimiento del cultivo de soja.

6.1 - Ejecución de la metodología propuesta

En esta sección se aplica la metodología propuesta paso por paso, ejemplificando posibles herramientas a utilizar y resultados esperados, y demostrando su efectividad en el cumplimiento del objetivo general de la investigación.

A - Análisis del área estudiada

Primeramente se analizó el área donde se encuentra el lote para una mejor comprensión de los resultados de los modelos. En esta investigación, se tomó como área de estudio un lote de 66 hectáreas de uso agrícola, ubicado a las afueras de la ciudad de San Justo, provincia de Santa Fe, Argentina.

La ciudad se ubica en la región centro-norte de la provincia, donde el clima transiciona de subtropical a templado, con temperaturas que no alcanzan valores extremos ni de frío ni de calor. En cuanto a precipitaciones, la media histórica desde el año 1960 al año 2007 fue de 1130 mm; siendo los meses más húmedos de noviembre a abril y los más secos de mayo a octubre. Y con respecto a vientos, dominan los provenientes del Sur y Sureste. Por último, la ciudad presenta un suelo medianamente profundo, con buen drenaje, con paisajes ligeramente ondulados. Estas características hacen que la capacidad productiva de las tierras sea considerada como alta (Municipalidad de San Justo & RAMCC, 2019).

Sin embargo, la productora agropecuaria “Don Mariano S.R.L”, quien trabaja el lote estudiado, explicó que durante la campaña 2024-2025 la zona sufrió una anomalía climatológica del tipo sequía durante los meses diciembre y enero. Este evento ocasionó que aún habiendo elegido una fecha de siembra en el mes de diciembre, generalmente correcta para cultivos de soja de segunda, se vea afectado el período de floración de la planta (etapas R1-R3 según la escala presentada en la Sección 2.1). Más aún, cuando el cultivo alcanzó las etapas de relleno de granos (etapas R4-R6 según la escala presentada en la Sección 2.1) ocurrió un fenómeno de helada. Estos dos eventos en conjunto ocasionaron que el rendimiento del lote para esta campaña haya sido considerablemente inferior a comparación de años anteriores.

Más aún, la productora agropecuaria expresó que el fenómeno de sequía estuvo presente en la zona también en las tres campañas anteriores, por lo que se decidió tomar la campaña 2020-2021 como referencia de un desarrollo normal de cultivo de soja en la región. Por esto, se utilizó la campaña 2020-2021 para la elaboración de un conjunto de datos representativo de una campaña normal para el entrenamiento de los modelos que serán luego evaluados, con el objetivo de comparar la precisión de cada uno al momento de identificar las anomalías conocidas de la campaña 2024-2025.

B - Extracción y transformación de los datos

Seguido al análisis de área, se obtuvieron las imágenes satelitales del lote, con las cuales se creó una serie temporal del NDVI, índice expuesto en la Sección 1.2, para ambas campañas con el objetivo de identificar temporalmente las anomalías antes mencionadas.

Las imágenes fueron analizadas en la plataforma GEE, introducida en la Sección 1.1, y recuperadas de la colección con identificador *COPERNICUS/S2_SR_HARMONIZED* que contiene imágenes adquiridas por los satélites Sentinel-2, presentado en el mismo capítulo. La elección de estos satélites se basó en que su sensor multiespectral permite obtener información de la vegetación y cobertura de suelo con alta resolución espacial, y un tiempo de revisita corto. Además, la versión armonizada de las imágenes minimiza las variaciones entre imágenes de distintos años, lo que facilita el estudio temporal de las características del objeto analizado (Google Earth Engine, s.f.).

Dado que las imágenes producidas por Sentinel-2 tienen un tiempo de revisita de 5 días, y habiendo limitado el análisis a los meses en los cuales se llevan a cabo las campañas de soja (diciembre-abril), se esperarían aproximadamente 36 observaciones por campaña. Sin embargo, para reducir el ruido en la serie temporal se decidió filtrar las imágenes con nubosidad superior al 20%. Luego del filtrado, el número de observaciones utilizables por campaña se redujo a 20. El producto de este proceso fueron dos conjuntos de imágenes multiespectrales representativas de cada campaña. A continuación, se calculó la media del NDVI en el lote para cada imagen, y se crearon dos series temporales con sus valores, concepto formulado en la Sección 3.1.

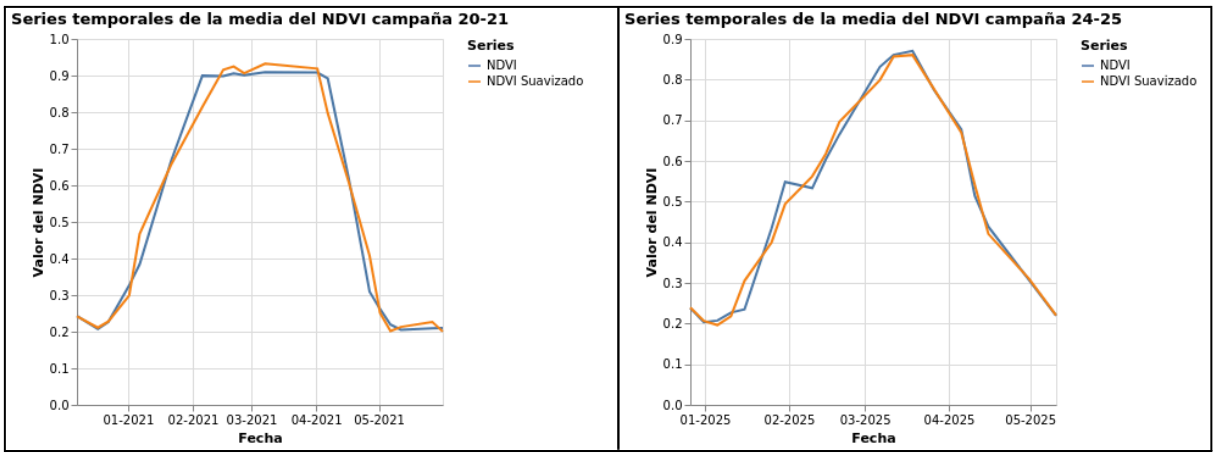
Como última instancia en el procesamiento de estos datos, se aplicó a ambas series el filtro Savitzky-Golay explicado en la Sección 3.2; de esta manera se obtuvieron dos series temporales suavizadas, para un mejor funcionamiento de los modelos utilizados posteriormente.

La Figura 18 contiene los gráficos de las series del NDVI original (curvas azules) y del NDVI suavizado por el filtro (curvas naranjas) para ambas campañas. Puntualmente

en la campaña 2024-2025, se aprecia el resultado de la aplicación del filtro en el mes 02-2025, donde la curva original presenta una falla y la curva suavizada una aproximación a la original sin esa falla.

Figura 18

Gráficos de series temporales del NDVI en las campañas analizadas

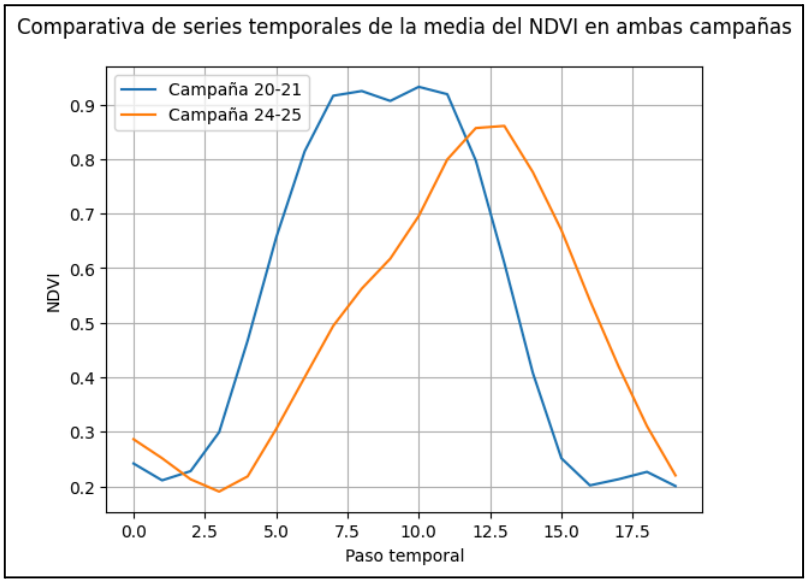


Como se mencionó previamente, a partir de la información provista por la productora agropecuaria responsable del manejo del lote durante las campañas 2020-2021 y 2024-2025, se identificaron dos anomalías climáticas adversas ocurridas en esta última: una sequía inicial seguida de un episodio de helada. Ambos fenómenos incidieron de manera significativa en el crecimiento del cultivo, afectando negativamente su desarrollo fisiológico. Como consecuencia, al momento de alcanzar la madurez completa, las plantas presentaron una baja formación de granos, lo que se tradujo en una reducción significativa del rendimiento del lote en comparación con campañas previas.

La Figura 19 expone las series temporales del NDVI de ambas campañas superpuestas, permitiendo identificar visualmente las anomalías descritas.

Figura 19

Superposición de NDVI de campañas 2020-2021 y 2024-2025



La curva del NDVI en la campaña 2020-2021, en azul, determina el comportamiento normal del desarrollo de las plantas de soja en el lote. En esta curva, el crecimiento del NDVI se da aceleradamente en los primeros dos meses en conjunto con los períodos de emergencia y floración de las plantas (VE-R4), quedando estancado en un valor aproximado de 0.9 cuando las plantas alcanzan el período de llenado de granos (R5-R6), y decreciendo en el último mes cuando alcanzan la madurez completa (R7-R8).

Por el contrario, en la curva de la campaña 2024-2025, en naranja, los estadíos de crecimiento y floración (VE-R4) toman casi el doble de pasos temporales hasta que el NDVI llega a su valor máximo, que es menor a 0.9 lo que indica que se cuenta con menos plantas o plantas menos vigorosas que en la campaña 2020-2021. Esta sección de la curva se corresponde con la primera anomalía comentada por la productora agropecuaria.

Adicionalmente, la curva de la campaña 2024-2025 exhibe la segunda anomalía, que afectó el período de llenado de granos (R5-R6) de las plantas, y en el la curva del NDVI se visualiza en el poco tiempo en que se mantiene en su valor máximo, comenzando un período de madurez completa (R7-R8) antes de tiempo.

C - Aumento de la serie temporal

Debido a que el conjunto de datos de entrenamiento creado es limitado, se optó por aumentarlo mediante las cuatro técnicas presentadas en la Sección 4.2: time warping, convolución, pooling, e introducción de ruido. En este punto se debe tener en cuenta que la creación de series artificiales a partir de variaciones de la serie original puede introducir

sesgos o características irreales al conjunto de datos utilizado para entrenar los modelos; sin embargo, al utilizar el mismo conjunto artificial en ambos modelos, la comparación entre ellos tiene validez ya que cualquier cambio introducido al conjunto afectará a ambos modelos. En este sentido, los resultados obtenidos deben interpretarse teniendo en cuenta esta limitación y se utilizan sólo con fines comparativos, en un escenario real es preferible utilizar datos auténticos de diferentes campañas.

El proceso de aumentación de datos resultó en 30 series temporales (Figura 20). Para el modelo SARIMA, estas se concatenaron (Figura 21) conformando una serie extendida de 60 puntos, lo que permitió al modelo disponer de una base de datos suficiente para una correcta estimación de parámetros. Por su lado, para la red LSTM se utilizaron las 30 series de forma independiente, aprovechando su capacidad de aprendizaje por lotes. Esta diferenciación asegura que cada modelo reciba la información en el formato óptimo para su estructura.

Figura 20

Gráfico de la serie de NDVI original y las series aumentadas

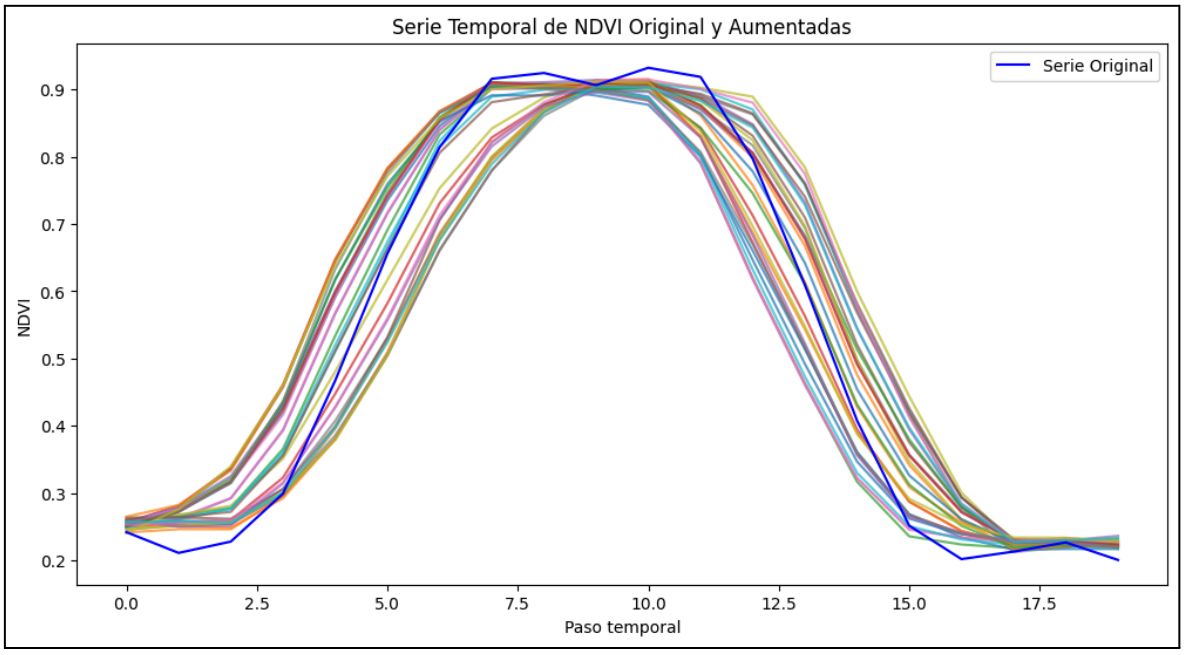
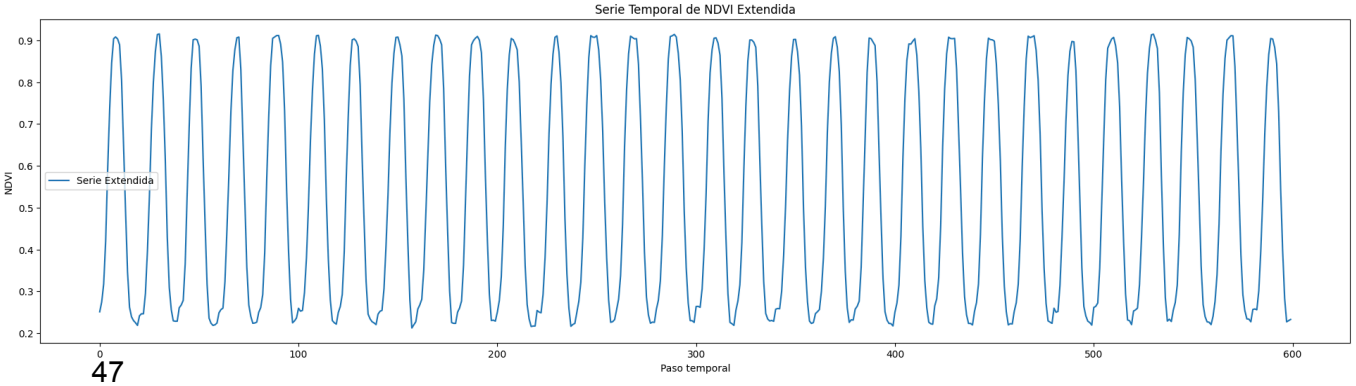


Figura 21

Gráfico de la serie de NDVI extendida



D - Ajuste de modelos

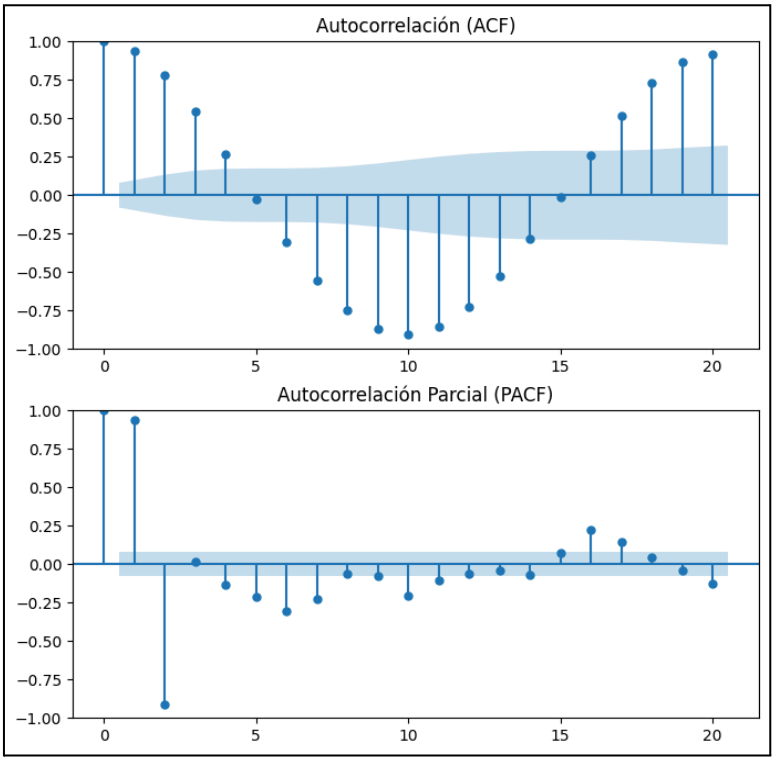
Para este punto de la aplicación de la metodología se tomó la decisión de trabajar con dos modelos con bases diferentes para comprobar cuál presenta mayor precisión. Por un lado se utilizó la serie extendida para entrenar un modelo SARIMA (introducido en la Sección 3.3), el cual es puramente estadístico; y por otro, una RNR con celdas LSTM (conceptos expuestos en las secciones 5.2 y 5.3) que utiliza aprendizaje automático para el ajuste de sus parámetros.

D.1 - SARIMA

Para la evaluación del modelo SARIMA, en primer lugar se graficaron las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) de la serie extendida, ambas introducidas en la Sección 3.3.1, con el objetivo de analizar sus características. Como se observa en la Figura 22, la ACF exhibe un comportamiento oscilante con una periodicidad aproximada de 20 rezagos, evidenciada por la alternancia entre valores positivos y negativos y por la repetición de dicho patrón. Este período coincide con la longitud de cada una de las series temporales individuales, compuestas por 20 pasos temporales, por lo que se incorpora un componente estacional de período s = 20. Por su parte, la función PACF no presenta un corte abrupto claramente definido, aunque muestra correlaciones significativas en los primeros rezagos, lo que sugiere la presencia de componentes autorregresivos de bajo orden.

Figura 22

Gráficos de ACF y PACF de la serie temporal extendida



Según los métodos de determinación de uso de SARIMA estudiados en la Sección 3.3.1, los gráficos sugieren el uso de un modelo SARIMA(p,d,q)(P,D,Q)s.

Al analizar las funciones ACF y PACF, se observó que la dependencia temporal de corto plazo se concentra en los primeros rezagos. Por esto, y para evitar un sobreajuste, se acotaron parámetros no estacionales *p* y *q* al rango [0, 5]. Asimismo, dado que en los gráficos la estacionalidad no presenta un patrón muy complejo, se limitaron los parámetros estacionales *P* y *Q* al rango [0, 3]. Por su parte, los parámetros *d* y *D* se configuraron para ser detectados automáticamente por la función.

Seguido, se utilizó el algoritmo de búsqueda automática de parámetro "auto_arima", presentado en la Sección 3.3.2. Las restricciones aplicadas a los parámetros redujeron el espacio de búsqueda, y por consiguiente el tiempo y poder de cómputo requeridos por el algoritmo, manteniendo únicamente configuraciones coherentes con la estructura observada en la serie.

Este proceso determinó que la combinación de parámetros que minimiza los valores de los criterios AIC y BIC, presentados en la Sección 3.3.2, con valores de -3217 y -3187 respectivamente corresponde al modelo **ARIMA(2,0,1)(1,0,1)[20]**. La minimización de estos criterios indica que esta configuración constituye la mejor aproximación entre los modelos evaluados, logrando un adecuado balance entre precisión y complejidad del modelo.

En la componente no estacional del modelo, los valores obtenidos fueron *p*=2, lo que indica la inclusión de un término autorregresivo de segundo orden; *d*=0, evidenciando que no fue necesaria la diferenciación de la serie; y *q*=1, incorporando un término de media móvil de primer orden.

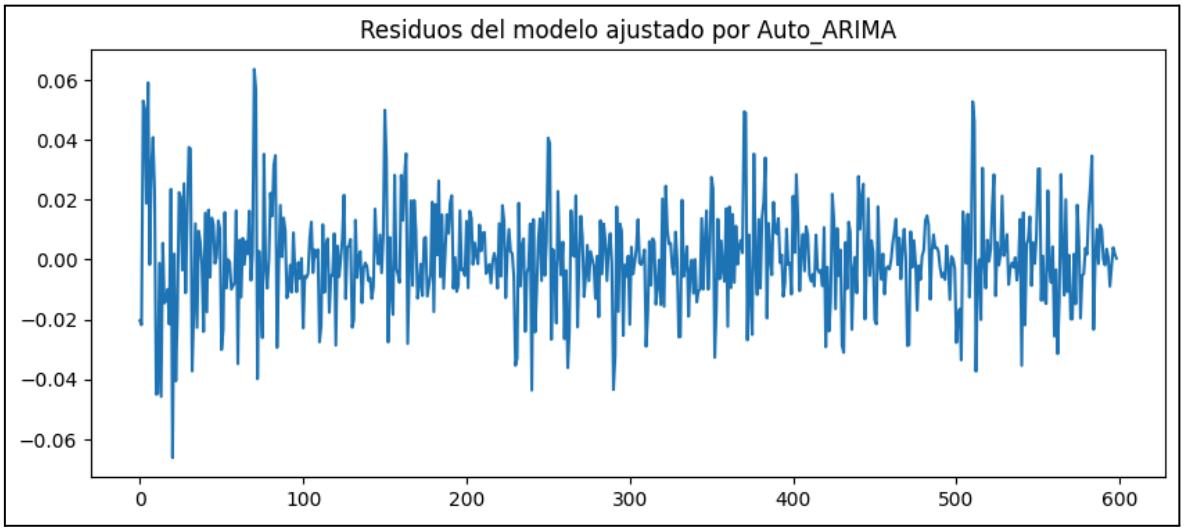
En cuanto a la componente estacional, se obtuvieron los valores *P*=1, que representa un término autorregresivo estacional de primer orden; *D*=0, indicando que no fue requerida diferenciación estacional; y *Q*=1, incorporando un término de media móvil estacional de primer orden. El período estacional se fijó en *s*=20, en concordancia con la cantidad de observaciones temporales de la serie correspondiente a la campaña 2020-2021.

Posteriormente, se graficaron los residuos del modelo para identificar si el modelo consigue representar la serie extendida. Como se mencionó en la Sección 3.3.1, para determinarlo se deben analizar los residuos del modelo e identificar que no sigan un patrón concreto.

En el gráfico de estos residuos, apreciables en la Figura 23, no se observa que sigan un patrón concreto sino que se presentan como ruido blanco, lo que sugiere que el modelo SARIMA consigue representar la serie estudiada como combinaciones lineales de sus valores y errores pasados.

Figura 23

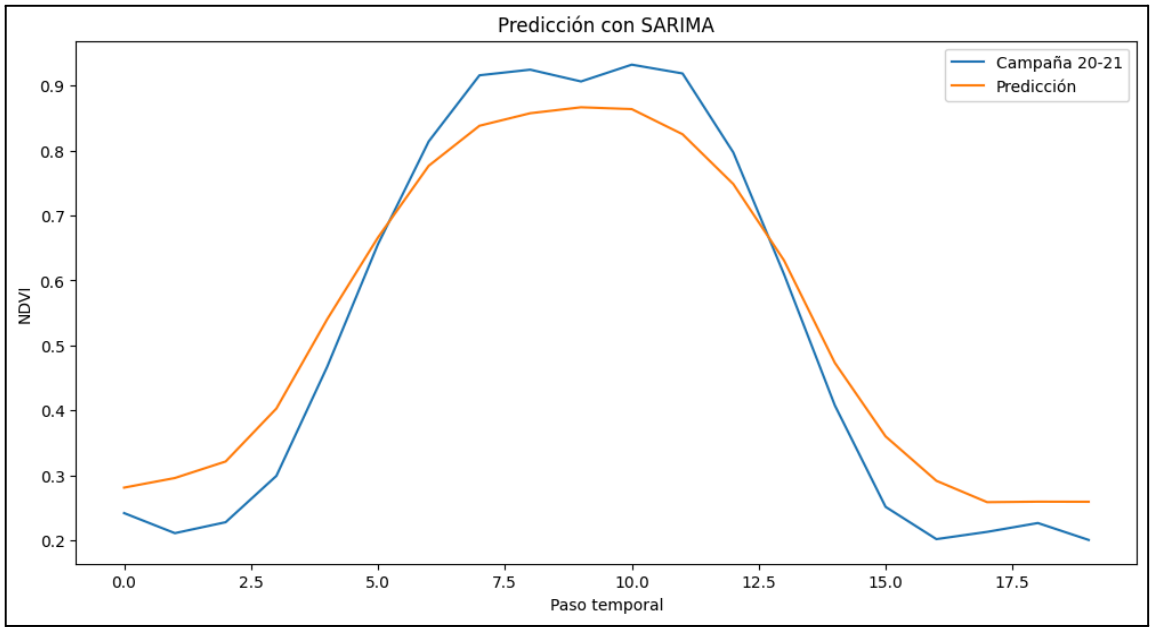
Gráfico de residuos del modelo SARIMA



Con los parámetros del modelo establecidos, se realizó una predicción de la serie temporal de la campaña 2020-2021; el resultado se exhibe en la Figura 24. En ella se observa que la serie predicha por este modelo estadístico se asemeja en su forma a la real utilizada para su ajuste.

Figura 24

Gráfico de predicción de serie temporal del NDVI con SARIMA



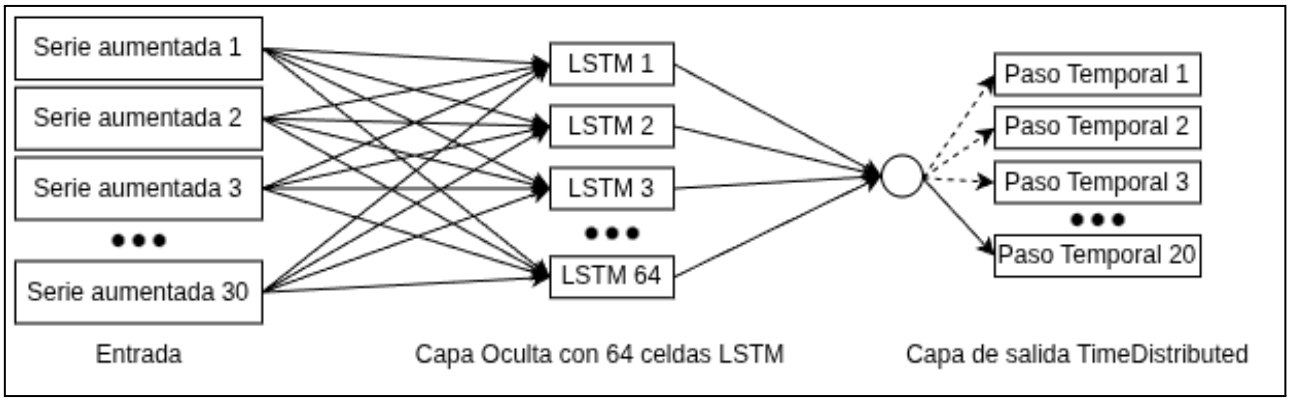
Como métrica de evaluación del error del modelo se utilizó el RMSE, definido en la Sección 4.1.2, para el cual se obtuvo un valor de 0.068.

D.2 - Red Neuronal Recurrente con Celdas LSTM

Aunque el error obtenido por el modelo SARIMA no es considerado alto, utilizar un enfoque de Machine Learning como el estudiado en el Capítulo 4, permite en algunos casos obtener mejores resultados en la predicción para series temporales. Por ello, se creó y evaluó también una RNR, un tipo de modelo de DL presentado en la Sección 5.2, compuesta por celdas LSTM, introducidas en la Sección 5.3, cuya arquitectura final se aprecia en la Figura 25.

Figura 25

Arquitectura final modelo LSTM



Según las clasificaciones expuestas en la Sección 5.2, se trata de un modelo secuencia a secuencia, que consta de:

- Una entrada que toma un conjunto de series de valores del NDVI, de 20 pasos temporales cada una.
- Una capa completamente conectada con 64 celdas LSTM que genera como salida un vector de 20 dimensiones para cada uno, determinado por la cantidad de pasos temporales de las series del conjunto de entrada.
- Una capa de salida que aplica una red completamente conectada con una sola neurona a cada uno de los 20 pasos temporales de las series de manera independiente, transformando las 64 dimensiones de la capa anterior de vuelta en una sola dimensión, y generando, en un mismo resultado, una nueva serie de 20 pasos temporales futuros predichos por la red.

De igual manera que en el modelo SARIMA, para automatizar el proceso de búsqueda de los parámetros que producen el mejor resultado para la RNR con celdas LSTM, se realizó una búsqueda en red de hiperparámetros, método presentado en la Sección 4.1.1, utilizando el conjunto de las 30 series aumentadas creadas en la en la

Sección C de la metodología. De este conjunto se dividió un 80% de los datos para el entrenamiento de los modelos y el 20% restante se utilizó para validación.

Para eficientizar el tiempo de ejecución y consumo de recursos en la búsqueda en red, sólo se hicieron 50 pasadas de entrenamiento por el conjunto de datos, para cada una de las combinaciones de tres valores posibles para los siguientes hiperparámetros:

- Cantidad de celdas LSTM: se evaluó con valores 16, 32, 64.
- Tasa de aprendizaje: se evaluó con valores 0.0025, 0.005, 0.01.

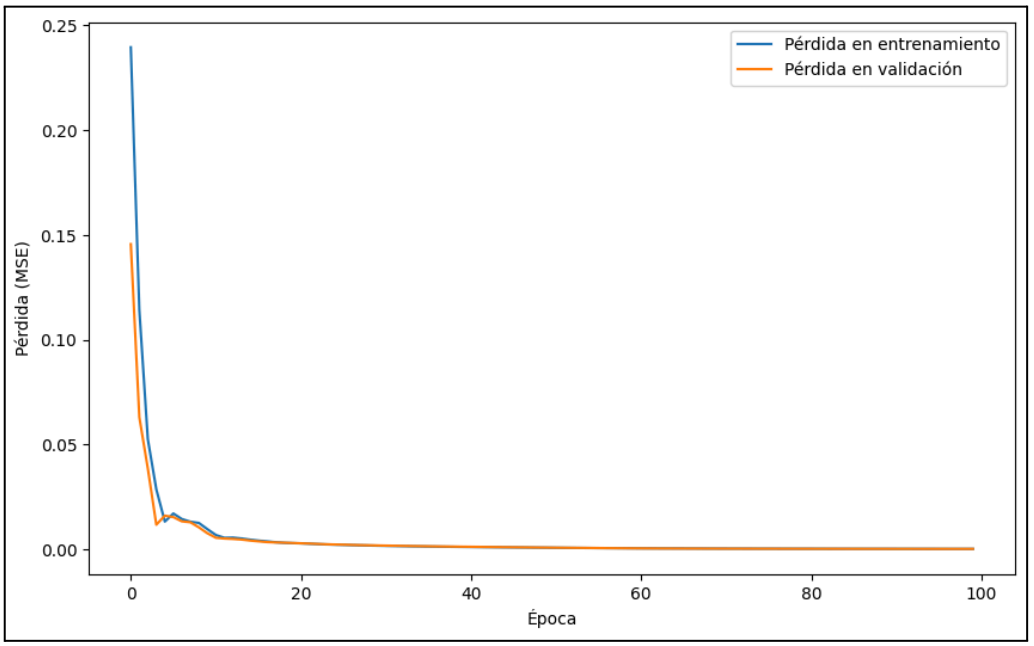
Las combinaciones de hiperparámetros se analizaron con la métrica de error RMSE presentada en la Sección 4.1.2, siendo la combinación que obtuvo el mínimo valor en esta métrica, de alrededor de 0.03 para el conjunto de validación, la siguiente:

- Cantidad de celdas LSTM: 64.
- Tasa de aprendizaje: 0.005.

Para acabar de ajustar el modelo seleccionado, éste fue entrenado durante 100 pasadas más por el conjunto de datos de entrenamiento, y luego se evaluaron las funciones de pérdida sobre este conjunto y sobre el de validación. La Figura 26 grafica estas funciones.

Figura 26

Funciones de pérdida en entrenamiento y validación del modelo LSTM



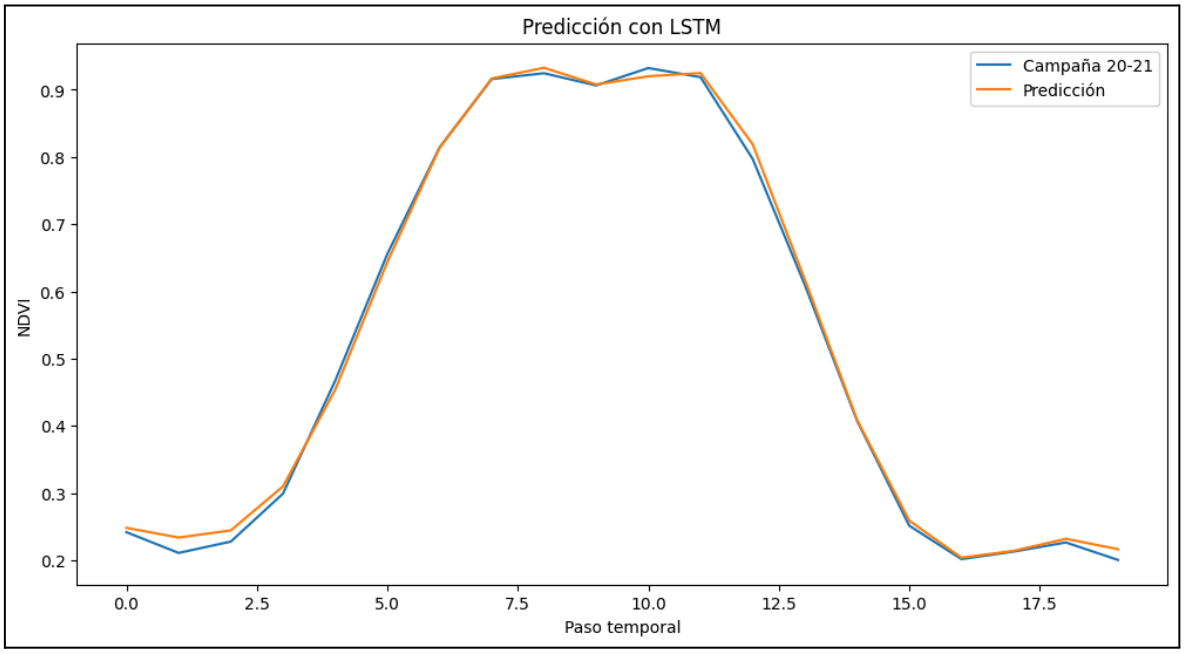
Como se observa, ambas funciones siguen una curva casi idéntica, decreciendo a medida que pasan las épocas; lo que indica que no se produjo un sobreajuste a los datos de entrenamiento sino que el modelo generaliza bien el comportamiento de la serie.

Finalmente, de igual forma que con el modelo SARIMA, se realizó una predicción de la serie temporal de la campaña 2020-2021 con este modelo de Machine Learning,

obteniendo un RMSE final menor a 0.01. La Figura 27 superpone la serie de la campaña 2020-2021 y la serie predicha por este modelo, en ella se visualiza la mejora en la precisión con respecto al modelo estadístico evaluado anteriormente.

Figura 27

Gráfico de predicción de serie temporal del NDVI con una RNR



Con el objetivo de contrastar las evaluaciones de precisión de cada modelo, se calculó la misma métrica de evaluación del error utilizada para evaluar el modelo SARIMA; consiguiendo esta RNR un RMSE de 0.013, valor significativamente menor al obtenido por el modelo estadístico.

E - Implementación de los modelos para detección de anomalías

Como último paso en la metodología, se implementaron ambos modelos para determinar cuál demuestra mayor eficacia al momento de identificar valores anómalos en series temporales de NDVI sobre un caso real.

Tal como se expuso en la Sección 3.4, una observación es considerada anómala cuando se ubica fuera del intervalo definido por tres desvíos estándar respecto de la media, el cual, bajo el supuesto de normalidad, abarca aproximadamente el 99,7% de los valores del conjunto de datos. En este contexto, para tomar un valor de NDVI igual a 0.4 como umbral de referencia para la detección de anomalías en este caso, se debió establecer un desvío estándar de 0.133 a fin de definir dicho intervalo de tolerancia de ±3 desvíos estándar.

Una vez definido el desvío estándar, se construyeron los intervalos de tolerancia para cada modelo ajustado en la Sección D de la metodología. De esta manera, la identificación de anomalías se realizó evaluando para cada paso temporal si los valores observados en la campaña 2024-2025 exceden los límites establecidos por los intervalos de tolerancia.

En la Figura 28 se ilustra este procedimiento utilizando el modelo SARIMA, mientras que en la Figura 29 se presenta el mismo análisis para la RNR. En ambas figuras, el área pintada de gris representa el intervalo de tolerancia, mientras que la curva en color azul representa las predicciones generadas por cada modelo a partir de su entrenamiento sobre la campaña 2020-2021, y la curva en color naranja corresponde a los valores observados de NDVI para la campaña 2024-2025. Por último, los puntos rojos representan las anomalías detectadas.

Figura 28

Detección de anomalías utilizando el modelo SARIMA

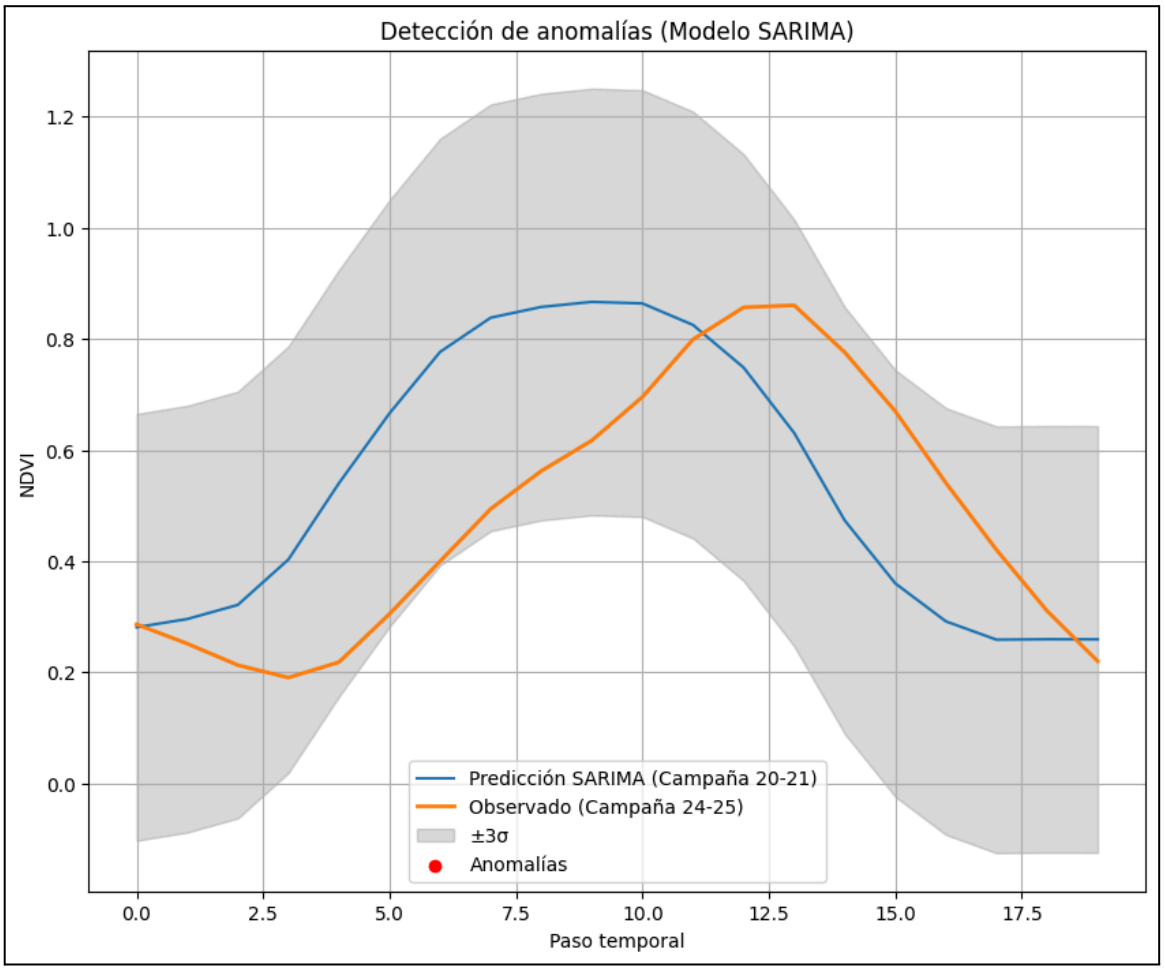
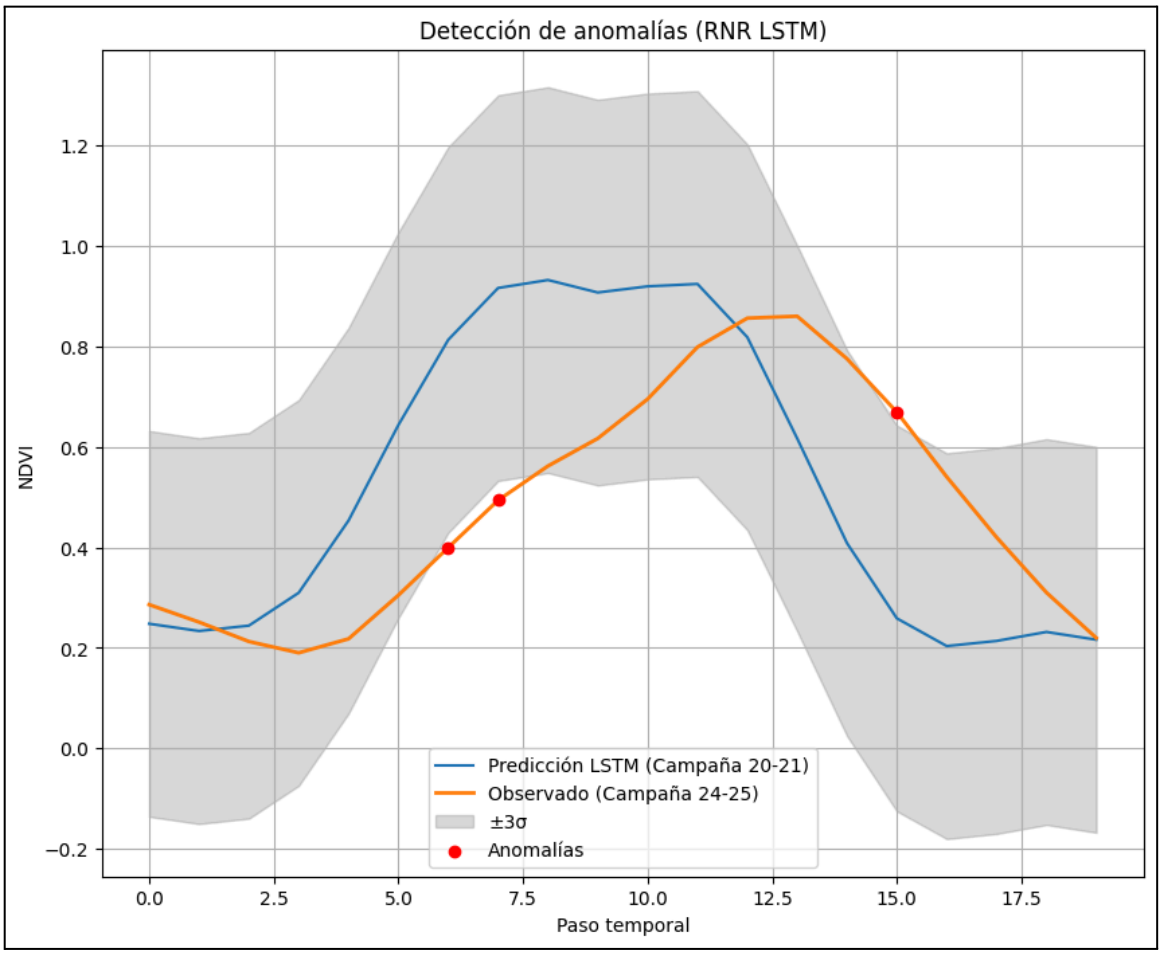


Figura 29

Detección de anomalías utilizando la RNR con celdas LSTM



La Figura 29 exhibe tres puntos temporales marcados en rojo, en contraste con la Figura 28 que no presenta ninguno. Por ello se determina que la red RNR obtiene un desempeño superior al modelo SARIMA en la identificación de anomalías en cultivos de soja en la región centro-norte de la provincia de Santa Fe, para la campaña 2024-2025; ya que manteniendo un desvío estándar constante para ambos casos, la RNR evidencia una mayor sensibilidad en la detección de desviaciones respecto al NDVI observado, logrando identificar tres pasos temporales con valores anómalos del índice. En cambio, el modelo SARIMA no detecta anomalías bajo este mismo criterio, lo que sugiere una menor capacidad para capturar variaciones abruptas en la serie analizada.

CONCLUSIÓN

La presente investigación tuvo como objetivo general diseñar una metodología para el monitoreo del Índice Normalizado de Diferencia de Vegetación aplicada a cultivos de soja en el centro-norte de la provincia de Santa Fe. Dicho objetivo fue alcanzado mediante la propuesta presentada en el Capítulo 6, que consta de cinco pasos secuenciales: análisis del área estudiada, extracción y transformación de los datos, aumento de la serie temporal, ajuste de modelos, e implementación para la detección de anomalías.

En relación a los objetivos particulares, el primero “estudiar técnicas de teledetección aplicadas al análisis de cultivos” fue abordado en el Capítulo 1, donde se presentaron los fundamentos de la teledetección, las misiones del Programa Copérnico y las características del satélite Sentinel-2 como fuente de datos para el cálculo de índices de vegetación. El segundo objetivo “analizar las etapas fenológicas de la soja en base al NDVI” fue desarrollado en el Capítulo 2, donde se describió la escala de Fehr y Caviness y se estableció la relación entre los valores del NDVI y los distintos períodos de desarrollo de la planta. El tercer objetivo “estudiar Machine Learning enfocado en modelos de series temporales” fue cubierto en los Capítulos 3, 4 y 5, que presentaron respectivamente los modelos estadísticos SARIMA, los fundamentos del aprendizaje automático y la arquitectura de redes neuronales recurrentes con celdas LSTM. Los objetivos cuarto y quinto “evaluar el rendimiento de los modelos sobre una serie temporal de NDVI de cultivos de soja” y “diseñar una metodología para el monitoreo del Índice Normalizado de Diferencia de Vegetación, adaptado a las condiciones del centro-norte de Santa Fe.” fueron cumplidos en el Capítulo 6, donde se propuso la metodología y se aplicó sobre un caso real, comparando los resultados obtenidos por los modelos.

De los resultados del ajuste de los modelos se concluye que, para el lote analizado en la localidad de San Justo, con cultivos de soja sembrados en diciembre-enero, la red neuronal recurrente con celdas LSTM obtuvo un RMSE de 0.013, valor significativamente menor al RMSE de 0.068 obtenido por el modelo SARIMA, lo que indica una mayor precisión en la predicción de la serie temporal del NDVI. Asimismo, en la etapa de detección de anomalías, utilizando el mismo desvío estándar como umbral para ambos modelos, la red neuronal identificó tres pasos temporales con valores anómalos correspondientes a las alteraciones climáticas conocidas de la campaña 2024-2025, mientras que el modelo SARIMA no detectó ninguno bajo el mismo criterio.

No obstante, estos resultados deben interpretarse teniendo en cuenta las limitaciones del estudio. En primer lugar, el conjunto de entrenamiento se construyó a partir de una única campaña de referencia, considerada representativa de un desarrollo normal del cultivo según la información provista por la productora agropecuaria

responsable del lote. En segundo lugar, dado que dicho conjunto contaba con solo 20 observaciones, fue necesario recurrir a técnicas de aumentación de datos para generar las muestras adicionales requeridas por los modelos. Si bien esta estrategia permitió la comparación entre ambos enfoques en igualdad de condiciones, las series temporales artificiales no amplían la información real disponible sino que la replican con variaciones controladas; por lo tanto, los resultados obtenidos tienen validez comparativa pero no equivalen a los que se obtendrían con datos auténticos de múltiples campañas. En un escenario real es preferible utilizar datos de diferentes campañas para el entrenamiento.

A pesar de estas limitaciones, la metodología propuesta constituye un aporte aplicable en la agricultura de precisión, ya que permite construir un modelo representativo de la dinámica de crecimiento de un lote específico y utilizarlo para identificar desviaciones respecto al comportamiento esperado, optimizando la toma de decisiones de productores agrícolas e ingenieros agrónomos ante la ocurrencia de eventos climáticos adversos.

FUTURAS LÍNEAS DE INVESTIGACIÓN

Como futuras líneas de investigación se propone utilizar la metodología propuesta para comprobar su eficacia en:

- otro tipo de cultivos de hoja verde, como trigo, sorgo o maíz,
- otras regiones del país con distintas variables ambientales,
- redes neuronales recurrentes con arquitectura diferente a la utilizada.
- otros tipos de modelos de Machine Learning,
- la utilización de modelos de machine learning para la detección de la fecha de siembra de un cultivo dada su curva de NDVI,
- la creación de una metodología para ofrecer el modelo creado como un servicio a clientes.

BIBLIOGRAFÍA

Bajocco, S., Ginaldi, F., Savian, F., Morelli, D., Scaglione, M., Fanchini, D., Raparelli, E., & Bregaglio, S. U. M. (2022). *On the use of NDVI to estimate LAI in field crops: Implementing a conversion equation library*. Remote Sensing, 14(15), 3554. https://doi.org/10.3390/rs14153554

Copernicus. (s. f.). Sentinel-1. SentiWiki. https://sentiwiki.copernicus.eu/web/sentinel-1

Copernicus. (s. f.). Sentinel-2. SentiWiki. https://sentiwiki.copernicus.eu/web/sentinel-2

Copernicus. (s. f.). Sentinel-3. SentiWiki. https://sentiwiki.copernicus.eu/web/sentinel-3

European Commission. (s. f.). *Copernicus programme. Copernicus SentiWiki*. https://sentiwiki.copernicus.eu/web/copernicus-programme

European Space Agency. (s. f.). Sentinel-4. Sentinel Online. https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-4

European Space Agency. (s. f.). Sentinel-5. Sentinel Online. https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-5

Farias G.D., Bremm C., Bredemeier C., de Lima Menezes J., Alves LA., Tiecher T., Martins A.P., Fioravanço G.P., da Silva G.P., de Faccio Carvalho P.C. (2023). *Normalized Difference Vegetation Index (NDVI) for soybean biomass and nutrient uptake estimation in response to production systems and fertilization strategies*. Front. Sustain. Food Syst. 6:959681. doi: 10.3389/fsufs.2022.959681

Fehr, W. R., & Caviness, C. E. (1977). *Stages of soybean development* (Special Report 80). Iowa State University of Science and Technology, Cooperative Extension Service, and Agriculture and Home Economics Experiment Station.

Fernández Ruiz, V. (2018). Ruido Blanco. RPubs. https://rpubs.com/vicferu/394452

Géron, A. (2023). *Aprende Machine Learning con Scikit-Learn, Keras y TensorFlow: conceptos, herramientas y técnicas para conseguir sistemas inteligentes* (3.ª ed.). Anaya Multimedia.

GIS Geography. (2025). *Sentinel 2 Bands and Combinations*. https://gisgeography.com/sentinel-2-bands-combinations/

Google Earth Engine. (s. f.). *Harmonized Sentinel-2 MSI: MultiSpectral Instrument, Level-2A (COPERNICUS/S2_SR_HARMONIZED).* Google Earth Engine Data Catalog. Recuperado el 20 de abril de 2026 de https://developers.google.com/earth-engine/datasets/catalog/COPERNICUS_S2_SR_HARMONIZED

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, M. del P. (2014). *Metodología de la investigación* (6ª ed.). McGraw-Hill Education. https://dialnet.unirioja.es/servlet/libro?codigo=775008

High-Level Expert Group on Artificial Intelligence (2018). *A definition of AI: Main capabilities and scientific disciplines*. https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

Hodson, T. O. (2022). *Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not.* Geoscientific Model Development, 15, 5481–5487. https://doi.org/10.5194/gmd-15-5481-2022

Jensen, J. R. (2007). *Remote sensing of the environment: An Earth resource perspective (2nd ed.)*. Pearson Prentice Hall.

McCulloch, W. S., & Pitts, W. (1943). *A logical calculus of the ideas immanent in nervous activity*. Bulletin of Mathematical Biophysics, 5(4), 115–133. https://doi.org/10.1007/BF02478259

Mitchell, T. M. (1997). *Machine Learning. McGraw-Hill*. https://www.cs.cmu.edu/~tom/files/MachineLearningTomMitchell.pdf

Morla, F. D., Sosa-Daniele, M. F., Marcellino, N., & Cerioni, G. A. (2022). *Aspectos de la ecofisiología y manejo del cultivo de soja*. Revista Ab Intus FAV-UNRC, 10(5), 68-74. https://doi.org/10.5281/zenodo.7484767

Municipalidad de San Justo & Red Argentina de Municipios frente al Cambio Climático (RAMCC). (2019). *Plan de Acción Climática de San Justo*. https://sanjusto.gov.ar/files/plan_de_accion_climatica.pdf

NASA. (s. f.). Synthetic aperture radar (SAR). Earthdata. Recuperado el 14 de marzo de 2026 de https://www.earthdata.nasa.gov/learn/earth-observation-data-basics/sar

Navidi, W. (2006). *Estadística para ingenieros y científicos*. McGraw-Hill Interamericana.

Nielsen, A. (2019). *Practical time series analysis. Prediction with Statistics & Machine Learning*. O'Reilly Media.

Ramos Pugnaire, J. M. (2023). *Data augmentation for time series*. Universidad Pontificia Comillas.

Rosenblatt, F. (1958). *The perceptron: A probabilistic model for information storage and organization in the brain*. Psychological Review, 65(6), 386–408. https://doi.org/10.1037/h0042519

Russell, S. J.; Norvig, P; (2004). *INTELIGENCIA ARTIFICIAL, UN ENFOQUE MODERNO*. *Segunda edición*. Pearson Educación.

Salvagiotti, F., Enrico, J. M., Bodrero, M., & Bacigaluppo, S. (2010). *Producción de soja y uso eficiente de los recursos*. INTA EEA Oliveros. Para Mejorar la Producción*, 45*, 151-154. https://www.academia.edu/57787170/Producci%C3%B3n_de_soja_y_uso_eficiente_de_los_recursos

Savitzky, A., & Golay, M. J. E. (1964). *Smoothing and differentiation of data by simplified least squares procedures*. Analytical Chemistry*, 36*(8), 1627–1639. https://doi.org/10.1021/ac60214a047

Servicio Geológico Minero Argentino (SEGEMAR). (s. f.). *Sensores remotos*. https://www.argentina.gob.ar/economia/segemar/sensores-remotos

Toledo, R. (2017). *Grupos de Madurez y Fecha de siembra*. SOJA Su ecofisiología y manejo. Recuperado el 28 de julio de 2025, de https://toledoruben.wixsite.com/cultivodesoja/fs-y-gm

Toledo, R. (2018). *Ecofisiología y generación de rendimiento de soja*. Cereales y Oleaginosas, FCA, UNC. https://www.researchgate.net/publication/384289635_Aspectos_de_la_ecofisiologia_y_la_generacion_de_rendimiento_de_soja_Glycine_max_L