# MapBiomas Argentina

# Soja de Primera y Soja de Segunda

# (ATBD)

## Documento de Bases Teóricas del Algoritmo

**Colección 3 – Agricultura: Soja**
**Mapas anuales 2000–2026**
**Versión 1.0**
*Octubre de 2026*

# 1.- Introducción

## 1.1 Alcance y propósito de este documento

Este documento describe las bases teóricas, el marco metodológico y las consideraciones operativas utilizadas para generar las capas de soja de MapBiomas Argentina Colección 3: una colección de mapas anuales de 30 m de resolución con las clases soja de primera y soja de segunda y la extensión espacial de la superficie agrícola de Argentina para el período 2000–2026.

Al igual que otros ATBD de MapBiomas, el propósito de este documento no es especificar línea por línea qué se realizó, sino más bien proveer una guía conceptual de las acciones y decisiones involucradas en el mapeo. Para cada etapa del flujo de trabajo, se explica qué se mide, por qué se mide de esa forma y qué se entrega a la etapa siguiente. El valor exacto de cada parámetro, las rutas de los *assets* y el orden de ejecución residen en los scripts de Google Earth Engine (GEE) del repositorio del proyecto (Sección 8), los cuales constituyen la referencia para cualquiera que necesite reproducir o adaptar el método.

La evaluación de la exactitud del producto es una actividad independiente que, al momento de escribir este documento, no ha finalizado. Por lo tanto, aquí solo se presenta una comparación con estadísticas oficiales (Sección 7); la evaluación formal será documentada por separado.

## 1.2 La soja en Argentina: de primera y de segunda

La soja (*Glycine max*) es el cultivo dominante de la agricultura argentina. Según los mapas que se describen en este documento, se sembraron en promedio 17,8 Mha por año entre 2000 y 2025, equivalente al 46 % de la superficie agrícola mapeada por MapBiomas Argentina. En Argentina la soja se cultiva, a grandes rasgos, bajo dos sistemas de producción diferentes:

- **Soja de primera:** Soja sembrada en primavera (octubre–diciembre) en lotes sin un cultivo previo en la misma campaña.
- **Soja de segunda:** Soja sembrada a fines de primavera o principios de verano (diciembre–enero) inmediatamente después de la cosecha de un cultivo de invierno o la senescencia de un cultivo de cobertura. Su serie temporal de NDVI muestra dos ciclos dentro de una misma campaña agrícola: el cultivo de invierno —o cultivo de cobertura— y, tras un breve valle del NDVI, la soja.

Separar ambos tipos es importante porque la soja de segunda implica un doble cultivo (dos cosechas por año), un uso de insumos diferente y dinámicas distintas de cobertura del suelo. También es técnicamente exigente: el ciclo de la soja de segunda es más corto y está desplazado en el tiempo, y la actividad fotosintética de otros cultivos de verano, como el maíz, el sorgo o el girasol, se solapa temporalmente con la de la soja.

## 1.3 Objetivos y novedades en la Colección 3

La Colección 3 incorpora, por primera vez en MapBiomas Argentina, una capa de tipos de cultivo anidada dentro de la clase Agricultura: la plataforma ofrece ahora *Agricultura › Soja*, con los dos tipos de soja como clases independientes de la leyenda. Los objetivos de esta primera capa de soja fueron:

1. Producir una serie anual consistente que cubra toda la superficie agrícola del país desde 2000 hasta 2026, compatible con la colección de cobertura del suelo de MapBiomas Argentina;
2. Distinguir la soja de primera de la de segunda cada año;
3. Minimizar el esfuerzo humano requerido para la recolección de datos de entrenamiento,

de modo que la serie pueda extenderse y reprocesarse anualmente; y
4. Mantener la consistencia temporal de la serie a pesar del cambio de sensores disponibles a lo largo del período (solo Landsat antes de 2016, y la colección Harmonized Landsat Sentinel-2 de 2016 en adelante).

Cuatro decisiones metodológicas responden a estos objetivos y estructuran este documento: (i) una selección automática de muestras de entrenamiento basada en la Deformación Temporal Dinámica Ponderada por el Tiempo (TWDTW, por sus siglas en inglés; Maus et al. 2016) que convierte 80 firmas de referencia verificadas manualmente en cientos de miles de píxeles de entrenamiento; (ii) una clasificación multinivel (*Random Forest*) a dos escalas, primero por carta topográfica y luego por zona agroecológica; (iii) un conjunto de procedimientos que garantizan la continuidad en el período 2000–2015, cuando no existían datos de Sentinel-2; y (iv) reglas fenológicas y espectrales que separan la soja de primera de la de segunda y remueven el maíz, principal fuente de errores de comisión.

## 1.4 Unidad de mapeo, período y campaña agrícola

La unidad de mapeo es la campaña agrícola, que en Argentina se asumió que transcurre desde julio del año inicial (*y*) hasta junio del año siguiente (*y+1*) y potencialmente incluye un cultivo de invierno y uno de verano. Por convención, cada mapa anual se etiqueta con el año en que inicia la campaña: la banda “classification_2000” representa la campaña 2000/2001 y “classification_2025” la campaña 2025/2026. Por lo tanto, los 26 mapas anuales cubren el período 2000–2026.

Dentro de la campaña, las fechas se expresan como día del año (DOY) contado desde el 1 de enero del año *y*, de modo que las fechas del segundo año calendario (*y+1*) toman valores superiores a 365 (p. ej., DOY 400 ≈ 4 de febrero del año siguiente).

## 1.5 Área de estudio y máscara agrícola

La capa de soja está restringida a los píxeles clasificados como cultivos anuales (clase 19) por la colección de cobertura y uso del suelo (LULC; MapBiomas Argentina 2026) en el año correspondiente. Así, cada clasificación anual se enmascara con la clase 19 de modo que la soja es siempre un subconjunto de la agricultura mapeada por la colección de cobertura. Para la selección de muestras de entrenamiento de la clase "otros cultivos", la máscara se amplía levemente para incluir la clase 15 (pasturas), asegurando que las pasturas sembradas y los cultivos forrajeros estén representados en la clase que no es soja.

Dado que el año calendario de cobertura del suelo que mejor describe el momento de ocurrencia de un cultivo de verano es el año de finalización de la campaña agrícola, se utiliza la máscara del año *y+1* para la campaña *y/y+1*; para la última campaña se utiliza el año de cobertura más reciente disponible.

## 1.6 Conjuntos de datos de entrada

**Tabla 1.** Conjuntos de datos de entrada y su rol en el flujo de trabajo.

| Conjunto de datos | Período utilizado | Rol en el flujo de trabajo |
|---|---|---|
| **Harmonized Landsat Sentinel-2 (HLS) v2.0: HLSL30 y HLSS30.** | 2016–2026 | Serie densa decadal de NDVI para TWDTW; índices espectrales mensuales y bandas de pico de estación para el clasificador; reglas fenológicas. |
| **Sentinel-2 MSI Nivel-2A, armonizado (COPERNICUS/S2_SR_HARMONIZED)** | 2016–2026 | Atributos de Posición del Borde Rojo (*Red Edge Position*, REP) para el clasificador |
| **Landsat 5 TM, 7 ETM+ y 8 OLI, Colección 2 Nivel-2** | 2000–2015 | Índices espectrales mensuales y bandas de pico de estación para el clasificador; *proxy* de REP; NDVI para reglas fenológicas. |
| **Reflectancia de superficie Landsat completada con MODIS (GF-SG; Chen et al. 2021)** | 2000–2015 | Serie densa decadal de NDVI para TWDTW. |
| **MapBiomas Argentina LULC (Colección 3)** | 2000–2026 | Máscara de agricultura (clase 19), máscara de muestreo (clases 19 y 15). |
| **Mapas de soja del año 2021 de GLAD (Song et al. 2021); Mapa Nacional de Cultivos del INTA (MNC) 2020/21 y 2023/24; Mapa de maíz 2021 de ESA WorldCereal (Van Tricht et al. 2023)** | Campañas de referencia | Capa de convergencia utilizada únicamente para ubicar puntos candidatos de muestreo de soja de primera y de segunda (Sección 3.2). |
| **Zonas PAS (Bolsa de Cereales de Buenos Aires)** | — | Unidades espaciales del segundo nivel de clasificación y de las reglas de post-procesamiento. |
| **Grilla de MapBiomas Argentina (cartas)** | — | Grilla de procesamiento. |

**Datos ópticos y máscaras de calidad:** Las escenas HLS se filtran por cobertura de nubes de la escena y se enmascaran píxel a píxel con la banda Fmask, excluyendo nubes, sombras de nubes y píxeles adyacentes a nubes; en los pasos de post-procesamiento, los píxeles con reflectancia muy alta en el rojo (B4 > 0,8) también se descartan como nubes residuales. Las escenas Sentinel-2 utilizadas para REP se filtran por cobertura de nubes de la escena (< 20 %) y por la probabilidad de nubes por píxel (< 20 %). Las escenas Landsat Colección 2 se escalan

a reflectancia de superficie y se enmascaran con QA_PIXEL (nube dilatada, cirrus, sombra, nube y nieve), descartando reflectancias fuera del rango físico (0–1).
**Armonización entre sensores:** En el período solo-Landsat, las reflectancias de Landsat 8 OLI se transforman al dominio de Landsat 7 ETM+ banda por banda utilizando coeficientes entre sensores publicados (Roy et al. 2016). En la colección HLS las imágenes Landsat 8/9 y Sentinel-2 se disponen en una grilla común de 30 m con reflectancia armonizada (Claverie et al. 2018).

### 1.7 Plataforma de cómputo y grilla de procesamiento

Todo el flujo de trabajo se ejecuta en Google Earth Engine (Gorelick et al. 2017). Los pasos basados en imágenes corren sobre la grilla de trabajo de MapBiomas Argentina, correspondiente a las cartas del IGN. Para evitar efectos de borde entre cartas, incrementar la diversidad y asegurar la existencia de los datos de entrenamiento, los pasos por carta leen la carta de interés junto con sus cartas vecinas: las distancias, muestras y modelos se calculan sobre esta vecindad, y el resultado se exporta únicamente para la carta central. El segundo nivel de clasificación y las reglas de post-procesamiento corren sobre cada una de las zonas PAS.

# 2.- Visión general del flujo de trabajo

El flujo de trabajo combina una medida de similitud basada en dinámica temporal del NDVI (utilizada para generar datos de entrenamiento de forma automática) con un clasificador supervisado y un conjunto de reglas de decisión. Se ejecuta en tres etapas, utilizando cada etapa lo generado por la etapa anterior:

1. **Selección de muestras de entrenamiento:** Un conjunto reducido de series temporales de NDVI de referencia de soja de primera y de segunda se compara, píxel a píxel y año a año, con la serie temporal de NDVI de cada píxel agrícola usando TWDTW. Los píxeles muy similares a la referencia se convierten en muestras de soja; los píxeles claramente disímiles se convierten en muestras de otros cultivos.
2. **Clasificación supervisada:** Un algoritmo de *Random Forest* (RF), entrenado con esas muestras sobre un espacio de atributos multitemporal, clasifica soja en una primera instancia por carta (un modelo binario para cada tipo de soja) y luego por zona PAS (un modelo de tres clases: soja de primera, soja de segunda, otros cultivos).
3. **Post-procesamiento:** Reglas de decisión, un filtro de Posición del Borde Rojo y un filtro espacial remueven las confusiones restantes, principalmente con maíz y otros cultivos de verano, y eliminan píxeles o parches pequeños y aislados.

La idea subyacente de este flujo de trabajo es combinar un método que es muy bueno para detectar patrones temporales de NDVI pero que es muy sensible a las muestras de referencia utilizadas (es decir, TWDTW) con un clasificador que puede explotar múltiples atributos espectrales a la vez, pero requiere abundantes muestras de entrenamiento bien distribuidas en el espacio (es decir, Random Forest). Usar TWDTW para generar las muestras de entrenamiento del Random Forest evita el costo (económico, logístico y operativo) que representa contar con muestras de entrenamiento obtenidas a campo para tantos años y un área tan extensa, a la vez que relaja los requerimientos de imágenes satelitales y permite generalizar los resultados.

# 3.- Selección de muestras de entrenamiento

### 3.1 Fundamentos

La Deformación Dinámica del Tiempo (DTW, Dynamic Time Warping) mide la disimilitud entre dos series temporales tras encontrar la alineación óptima, permitiendo desplazamientos y dilataciones en el tiempo. En cultivos agrícolas, esta posibilidad es muy útil porque, en un sitio, la fecha de siembra puede cambiar entre años debido a diferencias en las condiciones meteorológicas y, entre sitios, debido a diferencias climáticas, entre otras causas. TWDTW (Maus et al. 2016) agrega una penalización que crece con la diferencia de tiempo entre las observaciones emparejadas, evitando alineaciones que serían poco realistas, como emparejar un máximo del NDVI estival con uno invernal. La distancia TWDTW es, por lo tanto, pequeña para un píxel cuya curva de NDVI tiene la forma y la temporalidad del cultivo de referencia, y grande en caso contrario. Este enfoque ha mostrado un buen desempeño en el mapeo de cultivos con series de Sentinel-2 (Belgiu & Csillik 2018).
Así, aquí TWDTW se utiliza como generador de muestras: solo se utilizan los extremos de la distribución de distancias, donde la decisión es menos incierta, dejando la resolución del centro ambiguo al *Random Forest*.

## 3.2 Series temporales de referencia

En una primera instancia, las series temporales de referencia se generaron de manera interactiva utilizando la serie temporal HLS del período del 1/7/2020 al 30/6/2021 y del 1/7/2023 al 30/6/2024 para soja de primera y de segunda respectivamente. Las coordenadas (latitud, longitud) candidatas de soja se seleccionaron a partir de una capa de convergencia: píxeles mapeados como soja simultáneamente por el producto de soja sudamericano de GLAD (Song et al. 2021) y por el Mapa Nacional de Cultivos del INTA, y no mapeados como maíz por ESA WorldCereal (Van Tricht et al. 2023; sólo soja de primera) o el Mapa Nacional de Cultivos (solo soja de segunda). Para la soja de segunda, la capa de convergencia se restringió además a píxeles con dos ciclos de crecimiento en la campaña. Dentro de estas áreas se inspeccionaron visualmente múltiples series temporales de NDVI de manera que la elección del punto sea lo más representativa posible.
El resultado fue un conjunto de 41 series temporales de referencia de soja de primera y 39 de soja de segunda distribuidas en toda la superficie agrícola. A cada carta con más de ca. 1.000 ha de agricultura (clase 19 de la colección 3) se le asignó la serie temporal de referencia más cercana (priorizando agrupamientos longitudinales), obteniendo tablas donde se especifica qué serie temporal de referencia corresponde a cada una de las 122 cartas (primera) y 123 cartas (segunda). Las series temporales de referencia difirieron en su longitud:

- **Primera:** 18 observaciones en pasos de 10 días (≈ 6 meses), centradas en el ciclo de la soja.
- **Segunda:** 36 observaciones en pasos de 10 días (una campaña agrícola completa), de modo que la serie temporal contiene tanto el cultivo de invierno previo como la soja.

Para cada carta y tipo de cultivo, se almacena también una tabla de fechas clave derivadas de la firma: el primer y último DOY del ciclo de la soja y, para la soja de segunda, el inicio y el fin del valle entre la cosecha del cultivo de invierno y la emergencia de la soja, ambos identificados visualmente. Estas fechas se utilizan posteriormente para adaptar las ventanas temporales de algunos atributos al calendario local (Sección 4.2).

## 3.3 Series temporales de NDVI

TWDTW requiere series temporales regulares y sin datos faltantes. Para cada carta y campaña, la serie de NDVI desde el 10 de julio del año *y* hasta el 1 de julio del año *y+1* se construye de la siguiente manera:

- **2016–2026 (HLS):** Se utilizan en conjunto las observaciones de NDVI de HLS L30 y S30

enmascaradas por nubes. Los valores faltantes se completan píxel a píxel mediante interpolación lineal entre la observación válida anterior y la posterior en una grilla regular de 10 días, y la serie se promedia luego en ventanas de 10 días. Los valores se almacenan como enteros (NDVI × 1000).

- **2000–2015 (Landsat + MODIS):** Debido a la insuficiente resolución temporal de las imágenes Landsat —agravada por las frecuentes ocurrencias de datos faltantes por presencia de nubes, por la falla del corrector de líneas de escaneo de Landsat 7 y el período entre la salida de funcionamiento de Landsat 5 y la disponibilidad de imágenes de Landsat 8—, las series decadales de NDVI se generaron mediante la metodología propuesta por Chen et al. (2021). Resumidamente, esta metodología (GF-SG) primero completa los vacíos de la serie Landsat con la serie de NDVI de MODIS, previamente ajustada mediante correspondencia de forma *(shape matching)* y corrección, para generar una serie sintética Landsat-MODIS. Luego aplica un filtro de Savitzky–Golay ponderado que elimina el ruido residual, por ejemplo, debido a errores en la detección de nubes.

## 3.4 Distancia TWDTW

Para cada píxel, la distancia TWDTW entre la serie temporal del NDVI *x* y la del NDVI de referencia *r* de su carta se calcula de forma restringida al período abarcado por la serie de referencia (desde su primera hasta su última fecha —tal como se establece en la tabla de la Sección 3.2— con un margen de dos días). El costo de emparejar la observación *i* del píxel con la observación *j* de la referencia es la diferencia euclidiana de sus valores de NDVI más un peso temporal (w) que se calcula mediante una función logística:

$$c(i, j) = | x(i) - r(j) | + w(g),$$

donde

$$w(g) = 1 / (1 + \exp[-a * (g - b)])$$

y *g* es la diferencia absoluta, en días, entre el DOY de ambas observaciones, beta (*b*) = 50 días es la tolerancia temporal (punto medio de la logística) y alfa (*a*) = 0,1 controla la pendiente del modelo logístico. El costo acumulado de la alineación óptima es la distancia TWDTW. Como resultado de este proceso se exporta una imagen de distancia por carta, campaña y tipo de soja. El cálculo es el mismo en ambos períodos de sensores (2000–2015 y 2016–2026); solo cambia la serie de NDVI de entrada.

## 3.5 De las distancias a las muestras de entrenamiento

Las imágenes de distancia se convierten en muestras de entrenamiento utilizando umbrales adaptativos en lugar de valores fijos ya que el valor absoluto de las distancias varía con la longitud del período considerado (18 o 36 observaciones de 10 días) y con el período de sensores, entre otros factores:

1. La imagen de distancia se enmascara con la máscara de muestreo (clases 19 y 15 de MapBiomas) y se calculan los percentiles de su distribución sobre el conjunto de la carta y sus cartas vecinas.
2. Los píxeles con una distancia menor al percentil 20 (P20) se etiquetan como soja (de primera o de segunda, según corresponda).
3. Los píxeles con una distancia mayor al percentil 60 (P60) se etiquetan como otros cultivos. La franja entre P20 y P60 se deja sin etiquetar.
4. Solo se conservan los píxeles pertenecientes a parches grandes y homogéneos: se calculan las componentes conectadas (conectividad 4) para cada etiqueta y se descartan

los parches menores a una superficie mínima (entre 10 y 100 ha, según el tipo de soja y el período de sensores). Esto favorece muestras dentro de los lotes y elimina píxeles mixtos en los bordes.

5. Se extrae una muestra aleatoria estratificada sobre la vecindad (típicamente de 1.500 a 3.000 píxeles por clase y por carta).

Este procedimiento se aplica de manera independiente para cada carta y cada campaña, de manera que el conjunto de muestras de entrenamiento refleje los cambios interanuales del área sembrada con soja y otros cultivos a lo largo del tiempo. Las 80 series temporales de referencia (41+39) son los únicos datos seleccionados manualmente en todo el flujo de trabajo.

# 4.- Clasificación supervisada

## 4.1 Fundamentos

Random Forest (Breiman 2001) es un ensamble de árboles de decisión entrenados a partir de muestras seleccionadas con la técnica de bootstrap y subconjuntos aleatorios de los atributos. Random Forest permite explotar espacios de atributos de alta dimensión con colinealidad y se ve poco afectado por una fracción moderada de muestras mal etiquetadas. Los modelos utilizados en GEE tuvieron 300 árboles.

## 4.2 Espacio de atributos (*feature space*)

El clasificador (Random Forest) recibe como entrada un espacio de atributos con más información espectral que la usada con el TWDTW. Cada tipo de soja tiene su propia ventana temporal: noviembre–abril para soja de primera (el ciclo estival) y la campaña completa julio–junio para soja de segunda (para capturar el cultivo de invierno previo). El espacio de atributos, de unas 64 capas, según la ventana, contiene cuatro grupos de variables:

- **Composiciones mensuales** (agregados de 30 días) de índices de vegetación y agua: EVI2, LSWI, NIRv y kNDVI, más DCNI (un índice de borde rojo, solo Sentinel-2) en el período HLS, y RENDVI, LICI y NRLI en el período Landsat (reemplazando las bandas del borde rojo de Sentinel 2 por las NIR y SWIR1 de Landsat) .
- **Bandas espectrales de pico de estación:** A partir de un mosaico de calidad (función qualityMosaic de GEE) se identifica la fecha de máximo NRLI (índice de borde rojo/clorofila), que corresponde al máximo desarrollo del dosel. Para ese momento se obtienen las siguientes bandas e índices: verde, rojo, borde rojo (Sentinel-2 B5 y B6), SWIR1 y SWIR2, los índices REP, NRLI, RENDVI, DCNI y LICI, y el tono (*hue*) de la composición RGB.
- **Posición del Borde Rojo (*Red Edge Position*, REP):** Mínimo, mediana y máximo de REP durante el verano (10 de diciembre–15 de marzo para soja de primera; la ventana local de soja derivada de las firmas de referencia para soja de segunda). REP es el atributo que mejor separa la soja del maíz (Sección 5.2).
- **NDVI del valle (solo soja de segunda):** Valores de NDVI decadales durante el valle entre la cosecha del cultivo de invierno y la emergencia de la soja, con fechas adaptadas a cada carta a partir de su firma de referencia. Describen la transición del doble cultivo explícitamente.

En el período Landsat (2000–2015) se reproduce la misma estructura con las bandas disponibles en Landsat 5/7/8: las seis bandas ópticas en el pico de NRLI —el pico de NRLI obtenido mediante la función qualityMosaic de GEE— y un *proxy* de REP calculado a partir de las bandas NIR, RED y SWIR1 de Landsat.

**Tabla 2.** Índices espectrales utilizados: fórmulas y referencias.

| Índice | Fórmula | Referencia |
|---|---|---|
| **NDVI** | (NIR - Red) / (NIR + Red) | Tucker (1979) |
| **EVI2** | 2,5 * (NIR - Red) / (NIR + 2,4 * Red + 1) | Jiang et al. (2008) |
| **LSWI** | (NIR - SWIR1) / (NIR + SWIR1) | Xiao et al. (2004) |
| **NIRv** | NDVI * NIR | Badgley et al. (2017) |
| **kNDVI** | tanh(NDVI²) | Camps-Valls et al. (2021) |
| **REP (Sentinel-2)** | 705 + 35 * [ (B4 + B7)/2 - B5 ] / (B6 - B5) | Frampton et al. (2013) |
| **RENDVI (Sentinel-2)** | (B8 - B6) / (B8 + B6) | Gitelson & Merzlyak (1994) |
| **LICI (Sentinel-2)** | (B6 / B5) - (B3 - B4) / (B3 + B4) | Li et al. (2020) |
| **NRLI (Sentinel-2)** | (B8 - RENDVI) * LICI | Lyu et al. (2026) |
| **DCNI refined (Sentinel-2)** | ((((B6-B5)*144)/7)-B5)/((B5 - B4) * ((((B6 - B5)*144)/7)-B4 + 0,03)) | Han & Chai (2025) |
| **HUE (Tono, Sentinel-2)** | Componente H de la conversión RGB a HSV (usando B4, B3, B2) | Smith (1978) |

## 4.3 Nivel 1: Clasificación binaria por carta

Para cada carta y campaña, se entrenan dos modelos binarios de RF independientes sobre la vecindad de la carta con las muestras de la Sección 3.5: soja de primera vs. otros cultivos y soja de segunda vs. otros cultivos (hiperparámetros: 300 árboles, 9 variables por división). Cada modelo se aplica a la carta, se enmascara con la máscara agrícola de MapBiomas (clase 19) y se exporta. Los dos mapas binarios se combinan en un mapa anual preliminar con valores 1 (soja de primera) y 2 (soja de segunda).

## 4.4 Nivel 2: Reclasificación de tres clases por zona PAS

En los mapas del Nivel 1 suelen aparecer discontinuidades en los límites de las cartas. Ello se debe a que los modelos de cada carta se entrenan de manera independiente, con muestras diferentes. Clasificar a un segundo nivel generalmente atenúa estos artefactos. Entonces, para cada zona PAS (Panorama Agrícola Semanal, Bolsa de Cereales de Buenos Aires), utilizada en

el Mapa Nacional de Cultivos (INTA), y campaña, se entrenó un Random Forest con tres clases (soja de primera, soja de segunda y otros cultivos) utilizando el mapa del Nivel 1 como etiquetas: los píxeles mapeados como primera o segunda proveen las clases de soja, y los píxeles agrícolas restantes (clases 19 y 15) proveen la clase otros cultivos. Se extrae una muestra estratificada de 2.500 píxeles por clase en cada zona. El espacio de atributos sigue la Sección 4.2 con una ventana común de noviembre a mayo.
De esta manera, la clasificación de segundo nivel, al utilizar un conjunto de entrenamiento más amplio y diverso, extraído de una zona agroecológica completa, suaviza los errores locales del primer nivel y produce una clasificación única y mutuamente excluyente de ambos tipos de soja.
Por último, se realizaron dos excepciones de procesamiento. En primer lugar, en zonas donde algunos meses estuvieron persistentemente nublados, el espacio de atributos se restringió a los períodos más informativos, descartándose aquellos con datos faltantes. En segundo lugar, en la campaña 2013/14 se registró un número inusualmente bajo de escenas Landsat, particularmente en algunos paths. Debido a ello, para ese año y para cada zona PAS se realizaron clasificaciones por paths y luego se unieron en un mosaico para completar el área de la zona PAS analizada.

# 5.- Post-procesamiento

## 5.1 Reglas fenológicas

Se generaron reglas de decisión empíricas, basadas en la serie de NDVI, para refinar la separación entre tipos de soja y eliminar píxeles cuya fenología fuera incompatible con la soja. Estas reglas se aplicaron solamente a los píxeles clasificados como soja (de primera o de segunda), por zona PAS, con umbrales seleccionados mediante inspección de perfiles típicos en cada región:

- **Reasignación por NDVI invernal:** Se asumió que un píxel de soja cuyo NDVI máximo entre agosto y mediados de diciembre superase 0,65 tuvo un cultivo de invierno activo y se lo asignó a soja de segunda; en caso contrario, se asignó a soja de primera.
- **Momento del máximo NDVI de soja de primera:** Para ser clasificado como soja de primera, un píxel debe alcanzar su pico de NDVI después de mediados de enero (DOY > 380) con un valor de pico ≥ 0,60.
- **Momento del pico de soja de segunda:** En la ventana estival (mediados de diciembre a abril), la soja de segunda debe alcanzar su pico entre el DOY 396 y 485 (febrero–abril) con un valor ≥ 0,60.
- **Senescencia:** Entre enero y junio, el NDVI debe caer por debajo de 0,50 al menos 40 días después de haberlo superado; este filtro permite excluir a los píxeles que no evidencian un período de senescencia (p. ej., cultivos perennes).
- **Duración del ciclo:** El período con NDVI ≥ 0,30 no debe superar los 150 días. En las zonas norte (Chaco y NOA), donde coexisten cultivos de verano de distintas duraciones, se aplica una regla más estricta: ciclos superiores a 120 días o sin senescencia se descartan, y los ciclos cortos se aceptan solo si su REP máximo es inferior a 731 nm.
- **Consistencia del NDVI primaveral:** Un píxel etiquetado como soja de segunda cuyo NDVI medio entre septiembre y noviembre sea inferior a un umbral específico del año (0,52–0,65, calibrado de forma anual de manera empírica) se asume que no tuvo cultivo de invierno y se lo reasigna a soja de primera.

## 5.2 Filtro espectral: Posición del Borde Rojo (*Red Edge Position*)

El maíz es la principal fuente de errores de comisión debido a la similitud de su dinámica temporal

del NDVI con la de la soja. Ambos cultivos difieren, sin embargo, en el contenido de clorofila y la estructura del dosel. Se ha documentado que estas diferencias generan un desplazamiento del borde rojo (es decir, el rápido aumento de la reflectancia entre el rojo y el infrarrojo cercano) hacia longitudes de onda más largas en el maíz (You et al. 2023). La Posición del Borde Rojo, calculada con las bandas de borde rojo de Sentinel-2 (Frampton et al. 2013), captura esta diferencia.

- **2016–2026:** El REP máximo entre diciembre y marzo se calcula a partir de las observaciones HLS S30 (cobertura de nubes de escena < 10 %). Píxeles de soja con REP máximo superior a 731 nm son removidos.
- **2000–2015:** Landsat carece de bandas de borde rojo. Para extender el filtro, se entrenó un clasificador auxiliar de Random Forest (con 100 árboles) en la campaña 2018/2019 para predecir —de manera binaria— si el REP máximo de un píxel supera los 728 nm, utilizando la mediana del conjunto de atributos de las imágenes Landsat de diciembre a marzo como predictores y el REP calculado con las imágenes HLS S30 como variable de respuesta. Las muestras de entrenamiento se tomaron de los puntos de validación aleatorios de MapBiomas Argentina dentro de cuatro regiones de muestreo agrícola. El modelo ajustado, guardado como asset de GEE, se aplicó a cada campaña desde 2000 hasta 2015, removiendo los píxeles de soja predichos con REP alto.

El REP modelado es un proxy y resulta menos discriminante que el medido; esta es una de las razones principales por las que se espera que la exactitud de la serie sea menor antes de 2016 (Sección 9).

## 5.3 Filtro espacial

Finalmente, un filtro espacial remueve píxeles aislados y pequeños grupos típicos de las clasificaciones basadas en píxeles. Para cada año, se calcula el número de píxeles conectados de la misma clase (conectividad 8 en GEE); los píxeles pertenecientes a grupos de 18 píxeles o menos (≈ 1,6 ha) se reemplazan por la moda de su vecindad de 3 × 3. La aplicación de este filtro es la última operación antes de la publicación.

## 5.4 Integración

La serie final se ensambla a partir de las clasificaciones por zona PAS de cada campaña, tras aplicar los filtros fenológicos, espectrales y espaciales. Los valores iguales a la clase otros cultivos se establecen como *no-data*, dejando una capa de dos clases por año, enmascarada por la clase agricultura de MapBiomas del año correspondiente.

---

# 6.- Productos

## 6.1 Capa de soja

El producto es una imagen multibanda única con una banda por campaña (classification_2000 … classification_2025) a 30 m, en la que cada píxel toma uno de los valores de la Tabla 3. Los píxeles fuera de las clases de soja quedan enmascarados.

**Tabla 3.** Valores de píxel y clases de la capa de soja.

| Valor de píxel | Clase | Definición |
|---|---|---|
| **1** | **Soja de primera** | Soja sin un cultivo previo en la misma campaña agrícola. |
| **2** | **Soja de segunda** | Soja sembrada tras la cosecha de un cultivo de invierno en la misma campaña. |

A partir de esta capa, la plataforma deriva la superficie anual de cada tipo de soja por territorio (país, provincia, departamento y otros territorios MapBiomas) y la frecuencia de soja, es decir, el número de campañas en las que un píxel fue mapeado como soja a lo largo de la serie.

## 6.2 Principales cifras de la serie

**Tabla 4.** Principales cifras de la serie 2000–2025.

| Indicador | Valor |
|---|---|
| **Superficie media de soja 2000–2025 (primera + segunda)** | 17,8 Mha (46 % de la superficie agrícola) |
| **Superficie media de soja de primera / de segunda** | 14,4 Mha / 3,4 Mha |
| **Superficie de soja en la campaña 2000/01** | 14,1 Mha (41 % de la superficie agrícola) |
| **Superficie de soja en la campaña 2025/26** | 19,2 Mha (49 % de la superficie agrícola) |
| **Mínimo de la serie** | 11,4 Mha en 2002/03 (33 % del área agrícola total) |
| **Máximo de la serie** | 22,6 Mha en 2016/17 (58 % del área agrícola total) |
| **Provincias con mayor superficie media** | Buenos Aires (5,3 Mha), Córdoba (4,0 Mha), Santa Fe (2,1 Mha) |
| **Departamentos con mayor superficie media** | Río Cuarto (394 kha, Córdoba), General Roca (254 kha, Córdoba), General López (204 kha, Santa Fe) |
| **Mayor frecuencia media de soja (departamentos)** | Necochea (Buenos Aires): 15 campañas; Lobería (Buenos Aires), Constitución (Santa Fe) y Caseros (Santa Fe): 14,5 |

## 6.3 Dónde se encuentran los productos

La capa de soja está disponible en la plataforma de MapBiomas Argentina en *Agricultura › Soja*, donde puede navegarse por año, cruzarse con territorios y descargarse como

estadísticas:

- https://plataforma.argentina.mapbiomas.org/
- https://argentina.mapbiomas.org/

El asset integrado está almacenado en el proyecto de GEE de MapBiomas Argentina, bajo projects/mapbiomas-argentina/assets/LAND-COVER/COLLECTION-3/GENERAL/SOY/. La copia pública y las descargas en formato GeoTIFF siguen las convenciones de nomenclatura de la red MapBiomas.

# 7.- Comparación con estadísticas oficiales

Como primer control de consistencia, la superficie anual de soja se comparó con las estimaciones de superficie sembrada de la Secretaría de Agricultura, Ganadería y Pesca (SAGyP). Ambas series muestran la misma trayectoria general: una rápida expansión entre 2000 y mediados de la década de 2000, una meseta de alrededor de 18–21 Mha entre 2006 y 2018, y una contracción posterior. MapBiomas reproduce la mayor parte de la variabilidad interanual, observándose las mayores diferencias en las primeras campañas de la serie (2000–2002), en 2016/17 (por encima de SAGyP) y en 2019–2023 (por debajo de SAGyP).

No se espera que ambas series coincidan exactamente: SAGyP reporta superficie sembrada por unidad administrativa, mientras que el mapa detecta doseles de soja que desarrollan una fenología reconocible a 30 m. Lotes fallidos o resembrados, parcelas muy pequeñas y lotes afectados por sequías severas (como en la campaña 2022/23) están subrepresentados en el mapa, mientras que los errores de clasificación y las diferencias en la máscara agrícola pueden generar sobreestimaciones. Se contempla una evaluación formal de la exactitud basada en datos de referencia independientes, la cual será informada por separado.

# 8.- Reproducibilidad

Todos los pasos están implementados como scripts de JavaScript en el Editor de Código de GEE, organizados en el repositorio users/mapbiomas-arg/soja-col1 y en el Github de Mapbiomas. Dos módulos compartidos contienen las constantes (grilla, vecindades, nombres de bandas, umbrales) y las funciones (interpolación de series temporales, TWDTW, cálculo de índices, composiciones mensuales, armonización) utilizadas por cada script. Los productos intermedios (mosaicos decadales, imágenes de distancia, muestras, clasificaciones por carta y por zona, máscaras REP) se almacenan como *assets* de GEE, lo que permite volver a ejecutar cada paso de forma independiente. La Tabla 5 detalla la secuencia.

**Tabla 5.** Secuencia de pasos, scripts y productos del flujo de trabajo.

| Paso | Script (familia) | Producto |
|---|---|---|
| **0** | 000-Exporta_mosaico | Serie decadal de NDVI |

| Paso | Script (familia) | Producto |
|---|---|---|
| | | interpolada a 10 días por carta (HLS). |
| **1** | 001-Selecciona muestra / scripts de firmas de referencia | Tablas de firmas de referencia (primera y segunda) y fechas clave. |
| **2** | 000-Exporta_distancias (HLS ≥ 2016; Landsat+MODIS < 2016) | Imágenes de distancia TWDTW por carta, campaña y tipo de cultivo de soja. |
| **3** | 004-Clasifica_vfinal SOY1 / SOY2 (versiones HLS y solo-Landsat) | Clasificaciones binarias de Nivel 1 por carta. |
| **4** | Exportación de muestras por zona PAS (FS_PAS) | Muestras de entrenamiento de tres clases por zona y campaña. |
| **5** | Clasifica por zona PAS (incluyendo versión por *path* para 2013/14) | Clasificaciones de Nivel 2 por zona. |
| **6** | Reglas fenológicas por zona PAS (variantes norte y centro–sur) | Clasificaciones filtradas 2016–2026. |
| **7** | Filtro REP (HLS ≥ 2016; REP modelado < 2016) e integración con regla NDVI primaveral | Serie anual de soja. |
| **8** | Filtro Espacial SOJA | Capa final publicada. |

# 9.- Limitaciones y desafíos conocidos

- **Dos períodos de sensores:** Antes de 2016 la clasificación se apoya únicamente en Landsat, con menor densidad de observaciones, sin bandas de borde rojo y con un REP modelado. Las campañas posteriores a 2003 están afectadas por la falla del corrector de líneas de escaneo de Landsat 7, y las campañas 2011/12 y 2012/13 por el vacío entre Landsat 5 y Landsat 8; la campaña 2013/14 fue particularmente compleja debido a la inexistencia de imágenes en ciertos paths. Los usuarios deben esperar una menor exactitud y un mayor ruido interanual en el período 2000–2015.
- **Dependencia de las firmas de referencia:** Todos los datos de entrenamiento derivan de 80 firmas seleccionadas en dos campañas (2020/2021 para soja de primera y 2023/2024 para soja de segunda). Regiones o años cuyo calendario se aparte de esas campañas (siembras muy tempranas o muy tardías, resiembras tras sequías o inundaciones) pueden quedar subrepresentados en el muestreo.
- **Confusión con otros cultivos de verano:** El maíz constituye la principal fuente de confusión, especialmente el maíz tardío sembrado tras un cultivo de invierno; el girasol, el

sorgo, el algodón y el poroto también pueden confundirse localmente, en particular en el norte, donde los calendarios agrícolas son más diversos, o en los años más recientes (p. ej., girasol).

- **Definición de la soja de segunda:** La clase soja de segunda asume un cultivo de invierno detectable en la serie de NDVI. Cultivos de servicio, rastrojos enmalezados o cultivos forrajeros con un vigoroso crecimiento invernal pueden inducir comisiones de soja de segunda, mientras que cultivos de invierno con bajo vigor pueden generar omisiones.
- **Herencia de la máscara agrícola:** La soja solo puede mapearse allí donde la colección de cobertura del suelo mapea cultivos anuales; los errores de omisión de dicha clase se propagan a la capa de soja.
- **Umbrales empíricos:** Los umbrales fenológicos y de REP se fijaron mediante inspección de expertos en lugar de ajustarse sobre datos independientes, y algunos varían según la zona o el año.

---

# 10.- Líneas de mejora priorizadas

1. **Análisis de la exactitud:** Estimar la exactitud por clase y ajustar el cálculo de superficies siguiendo, en lo posible, los lineamientos de Olofsson et al. (2014).
2. **Series temporales de referencia adicionales y multianuales:** Agregar series temporales de campañas contrastantes (años húmedos y secos) y de otras regiones generaría resultados más robustos ante potenciales cambios en las fechas de siembra u otras prácticas de manejo en el tiempo y en el espacio.
3. **Mapear maíz:** Entrenar el segundo nivel con una clase explícita de maíz en lugar de un filtro REP post-clasificación permitiría abordar la principal confusión directamente en la etapa de clasificación.
4. **Espacio de atributos armonizado entre períodos de sensores:** Utilizar el mismo conjunto de atributos compatibles con Landsat en ambos períodos, o evaluar otras metodologías para rellenar la serie Landsat de 2000 a 2015, reduciría la discontinuidad observada en 2015/2016.
5. **Umbrales más objetivos:** Reemplazar los umbrales fenológicos ajustados manualmente por valores estimados por zona a partir de los datos de referencia.
6. **Extensión a otros cultivos:** El mismo marco metodológico TWDTW + RF puede aplicarse a maíz, trigo y girasol, lo que también mejoraría la separación de la soja al mapear explícitamente a otros cultivos de verano.

# Referencias


- Badgley, G., Field, C. B., & Berry, J. A. (2017). Canopy near-infrared reflectance and terrestrial photosynthesis. *Science Advances*, 3(3), e1602244.
- Belgiu, M., & Csillik, O. (2018). Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. *Remote Sensing of Environment*, 204, 509–523.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Camps-Valls, G., et al. (2021). A unified vegetation index for quantifying the terrestrial biosphere. *Science Advances*, 7(9), eabc7447.
- Chen, Y., Cao, R. Y., Chen, J., Liu, L. C., & Matsushita, B. (2021). A practical approach to reconstruct high-quality Landsat NDVI time-series data by gap filling and the Savitzky–Golay filter. *ISPRS Journal of Photogrammetry and Remote Sensing*, 180, 174–190.

- Claverie, M., et al. (2018). The Harmonized Landsat and Sentinel-2 surface reflectance data set. *Remote Sensing of Environment*, 219, 145–161.
- Frampton, W. J., Dash, J., Watmough, G., & Milton, E. J. (2013). Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 82, 83–92.
- Gitelson, A., & Merzlyak, M. N. (1994). Quantitative estimation of chlorophyll-a using reflectance spectra: Experiments with autumn chestnut and maple leaves. *Journal of Photochemistry and Photobiology B: Biology*, 22(3), 247–252.
- Gorelick, N., et al. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, 18–27.
- Han, M., & Chai, J. (2025). An efficient and transferable remote sensing spectral index for regional corn mapping. *Science of Remote Sensing*, 100308.
- Jiang, Z., Huete, A. R., Didan, K., & Miura, T. (2008). Development of a two-band enhanced vegetation index without a blue band. *Remote Sensing of Environment*, 112(10), 3833–3845.
- Li, D., et al. (2020). Improved estimation of leaf chlorophyll content of row crops from canopy reflectance spectra through minimizing canopy structural effects and optimizing off-noon observation time. *Remote Sensing of Environment*, 248, 111985.
- Lyu, D., Lai, C., Zhu, B., Zhen, Z., & Song, K. (2026). An Efficient Remote Sensing Index for Soybean Identification: Enhanced Chlorophyll Index (NRLI). *Remote Sensing*, 18(2), 278.
- MapBiomas Argentina (2026). Collection 3 of the annual land use and land cover maps of Argentina. https://argentina.mapbiomas.org/
- Maus, V., Câmara, G., Cartaxo, R., Sanchez, A., Ramos, F. M., & de Queiroz, G. R. (2016). A time-weighted dynamic time warping method for land-use and land-cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(8), 3729–3739.
- Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148, 42–57.
- Roy, D. P., et al. (2016). Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity. *Remote Sensing of Environment*, 185, 57–70.
- Smith, A. R. (1978). Color gamut transform pairs. *SIGGRAPH Computer Graphics*, 12(3), 12–19.
- Song, X.-P., et al. (2021). Massive soybean expansion in South America since 2000 and implications for conservation. *Nature Sustainability*, 4, 784–792.
- Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2), 127–150.
- Van Tricht, K., et al. (2023). WorldCereal: a dynamic open-source system for global-scale, seasonal, and reproducible crop and irrigation mapping. *Earth System Science Data*, 15, 5491–5515.
- Xiao, X., et al. (2004). Modeling gross primary production of temperate deciduous broadleaf forest using satellite images and climate data. *Remote Sensing of Environment*, 91(2), 256–270.
- You, N., Dong, J., Li, J., Huang, J., & Jin, Z. (2023). Rapid early-season maize mapping without crop labels. *Remote Sensing of Environment*, 290, 113496.