Predicción estadística de las etapas fenológicas del colza (*Brassica Napus L)* a partir de datos meteorológicos y observaciones satelitales

Tesis de investigación

Elvia Julieth Arellano Ortiz

Encadrant de stage
Corentin Barbu, Inrae
Tuteur universitaire
Nicolas Delbart, Université de Paris

Année 2019 - 2020

Master Géographie et Sciences du Territoire, M2 parcours Géomatique et Télédétection Appliquées à l'Environnement

Resumen

Los cambios de estado fenológico de las plantas son importantes indicadores en la investigación agronómica. Sin embargo, la dificultad para la recolección de datos fenológicos a gran escala es un desafío actual. La utilización en conjunto de información espectral proveniente de imágenes satelitales y datos meteorológicos preprocesados, se perfila como una solución a tal desafío.

Por lo tanto, el principal objetivo de este trabajo es ajustar y evaluar diferentes modelos para predecir las fases fenológicas con la utilización de datos satelitales y productos meteorológicos. Para ello, se construyó un conjunto de datos para 8 estados fenológicos recolectados a partir de la base de datos **Vigicultures®** durante la campaña agrícola 2017 para parcelas de colza distribuidas en toda Francia. Ajustamos los modelos estadísticos utilizando los métodos de *Machine Learning* más utilizados para clasificar información categórica, como *Lasso-Multinomial*, *Random Forest (RF)* y *KNN*. La calidad de los modelos fue estimada usando sus matrices de confusión y su *accuracy* global. Los resultados obtenidos mostraron un potencial variable para acoplar los índices derivados de los productos de la teledetección con las variables meteorológicas. La aplicación de ambas fuentes de datos nos permite clasificar los estados fenológicos con una *accuracy* de 0.84 en el mejor modelo encontrado. Encontramos que una buena predicción de los estados fenológicos intermedios está relacionada principalmente con los datos meteorológicos, mientras que para los estados primaverales (floración), hay una fuerte importancia de índices espectrales como el *NDYI*. El hecho de tener en cuenta las variables espacio-temporales sólo mejora marginalmente el modelo de referencia. La diversidad de las fuentes de información es más importante que el preprocesamiento de la información antes de proporcionarla al modelo de *RF*. Aunque el modelo de referencia no tiene por objeto sustituir las observaciones in situ, puede ayudar en el proceso de toma de decisiones.

Palabras claves: Fenología, aprendizaje automático, clasificación, bosque aleatorio, Brassica napus, Copernicus, Sentinel-2, Modelización de cultivos, Cambio climático.

Abstract

Changes in the phenological state of plants are important indicators in agronomic research. However, the difficulty of collecting phenological data on a large scale is a current challenge. The joint use of spectral information from satellite images and pre-processed meteorological data appears to be a response to this challenge.

Therefore, the main objective of this work is to adjust and evaluate different models to predict phenological phases using satellite data and meteorological products. A dataset for 8 phenophases collected in the **Vigicultures®** database during the 2017 agricultural season has been built for rapeseed plots spread over the whole French territory. We fitted the statistical models using the most commonly used *Machine Learning* (ML) methods to classify categorical information, such as *Lasso-Multinomial, Random Forest (RF)* and *KNN*. The quality of the models was estimated using their confusion matrices and overall *accuracy*. The results obtained showed a variable potential for coupling indices derived from remote sensing products with meteorological variables. Crop stages are estimated with these models using several data sources: Sentinel 2 spectral data, meteorological data (Météo-France's SAFRAN model) and space-time data. With the reference model using meteorological and spectral data, we obtained an *accuracy* of 0.84 with almost only inversions between neighboring stages. We have studied the impact of modifications of this model as well as the impact of different variables on the quality of the prediction. We found that good prediction of intermediate phenological stages is mainly related to meteorological data, while for spring states (flowering) there is a strong importance of spectral indices such as NDYI. Taking into account spatio-temporal variables only marginally improves the reference model. The diversity of information sources is more important than preprocessing before providing it to the Random Forest model. Although the reference model is not intended to replace in-situ observations, it can assist in the decision-making process.

Keywords: Phenology, Machine learning, classification, Random Forest, rapeseed, Canola
Brassica napus, Copernicus, sentinel-2, Crop modeling, Climate change.

Agradecimientos

La experiencia de construir conocimiento es una aventura extrema y gratificante. Agradezco a las personas que de una u otra forma han participado en este eterno aprendizaje para seguir conociendo, descubriendo y avanzando en esta aventura que llamo vida.

En primer lugar, me gustaría agradecer a Corentin Barbu que me ha guiado durante esta práctica académica. Siempre supo estar disponible y me apoyó mucho durante esta formación en programación y estadística que debo admitir que fue un gran reto. Gracias también por las conversaciones durante la pausa del almuerzo.

Gracias también a Nicolas Delbart por su apoyo y ayuda en los momentos en que pensé que no sería posible empezar este proceso.

Un agradecimiento especial a todo el equipo de Inrae en Thiverval-Grignon por su amabilidad y bienvenida, y también a todos los amigos que conocí en residencia de estudiantes en Versalles.

Agradezco también al instituto técnico Terres Inovia por brindar el acceso a la base de datos Vigicultures para el Colza, a la unidad AgroClim del INRAE por brindar el acceso a los datos meteorológicos SAFRAN y a Mathieu Fauvel del CESBIO por proveer las extracciones hechas por *iota2* de los datos Sentinel-2.

Por último, un agradecimiento al Sr. Rivals por su apoyo diario durante las prácticas.

Introduction

Contexto General

Anualmente, el CNES (Centro Nacional de Estudios Espaciales) realiza una convocatoria de propuestas de investigación a los laboratorios espaciales para el desarrollo de temáticas derivadas de la observación remota de superficies terrestres. El proyecto TOSCA-PARCELLE es el resultado de una de esas convocatorias en las que el uso de imágenes satelitales es el insumo principal. Dicho proyecto busca fomentar los esfuerzos para aunar y capitalizar la cadena de tratamiento de *iota2* (Infrastructure pour l'Occupation des sols par Traitement Automatique).

Originalmente, *iota2* fue diseñado como un flujo de trabajo de clasificación para la cartografía de la cubierta terrestre a gran escala, sin embargo la versatilidad del algoritmo también permite realizar extracciones de información espectral en toda Francia al nivel de escala de la parcela agrícola, función de gran importancia para el desarrollo de este trabajo.

El uso de la información espectral extraída a partir de la utilización de *Iota2* permite que el Centro Nacional de Investigación en Agricultura y Medio Ambiente (INRAE) y el Instituto de Ciencias e Industrias de la Vida y del Medio Ambiente AgroparisTech co-construyan junto a los agricultores el futuro de una agricultura más sostenible.

Dentro de la Unidad Mixta de Investigación (UMR) en agronomía, el equipo de investigación crea herramientas que ayudan a la toma de decisiones. Las herramientas diseñadas, buscan mejorar el control biológico de bioagresores para disminuir el uso de productos fitosanitarios.

Es en este contexto que se inscribe esta pasantía, en donde se busca establecer un modelo de clasificación de las etapas fenológicas de los cultivos de gran interés agroalimentario. Lo anterior permite comprender cómo la presencia de bioagresores en determinadas etapas del desarrollo vegetal puede afectar el rendimiento final de las cosechas.

El seguimiento de las diferentes etapas del desarrollo de los cultivos se denomina Fenología (Beurs et Henebry 2005). La Fenología ha sido abordada científicamente desde diferentes escalas espaciales. Al nivel de la parcela, existen metodologías in situ que permiten determinar los estados fenológicos exactos de los cultivos (van Vliet et al. 2003). A escala local, el uso de vectores aéreos (UAV) equipados con instrumentos de medición (cámaras espectrales), permite analizar la vegetación a mayor escala sin comprometer la precisión de la información que alimenta los modelos (Berra, Gaulton, et Barr 2019). A escala regional y mundial, el uso de instrumentos de observación remota facilita el análisis de grandes zonas (bosques y campos) para determinar las tendencias y las respuestas de los cultivos a diferentes variables como el cambio climático, la calidad

del suelo, la presencia de estrés, entre otras (Heumann et al. 2007; Han et al. 2018; Brown et al. 2008).

Análisis Fenológico en Agricultura

En la agricultura, el análisis remoto del ciclo fenológico de los cultivos es una herramienta clave para, entre otras cosas, determinar el rendimiento y la respuesta de los campos a las variables externas, en particular a la presión de plagas y enfermedades de los cultivos. La incursión de la teledetección en la agricultura ha permitido considerar efectos específicos extrapolados a realidades más grandes con una menor inversión de recursos (X. Zhang, Friedl, et Schaaf 2009; Wardlow et Egbert 2008). El estudio de la fenología de las plantas mediante la observación remota ha sido ampliamente discutido en la literatura, ya que el lanzamiento de satélites equipados con sensores capaces de explotar la energía reflejada en las superficies terrestres ha permitido analizar el comportamiento de la vegetación ya sea en base a su clorofila, estructura o capacidad de retención de agua para deducir su estado fenológico (X. Zhang, Friedl, et Schaaf 2009).

Teledetección y Fenología

Sensores como MODIS a bordo de los satélites estadounidenses *Acqua y Terra* se han utilizado ampliamente para este fin (Fisher et Mustard 2007; Ahl et al. 2006). Sin embargo, en la actualidad es la misión Europea *Sentinel*, con su familia de satélites y las mejoras de los instrumentos, la que proporciona imágenes satelitales de alta resolución espacial y temporal (Jönsson et al. 2018; Vrieling et al. 2018). Desde la perspectiva de la teledetección, la estimación convencional de las mediciones fenológicas suele hacerse a partir de series temporales. Esta estimación suele tener tres pasos fundamentales: 1) limpieza de los datos y presentación de informes; 2) suavización de los datos y reconstrucción de los datos de las series temporales; y 3) extracción de las mediciones fenológicas generadas a partir de los datos de las series temporales reconstruidas (Zeng et al. 2020) .

***Machine Learning* y Fenología**

Existen también otros enfoques basados en la complementariedad ("acoplamiento") entre diferentes tipos de datos (Almeida et al. 2014). Estos enfoques pueden establecer modelos de predicción de las diferentes etapas de un fenómeno utilizando instrumentos de inteligencia artificial como el M*achine Learning* (ML) de cual es parte el Aprendizaje Profundo para identificar patrones (Czernecki, Nowosad, et Jabłońska 2018).

En el marco de esta pasantía, analizaremos el aporte de la información espectral, climatológica y espacio-temporal para la predicción de los estados fenológicos de cultivos de importancia agroecológica. Abordaremos dicha pregunta de investigación utilizando herramientas de clasificación con métodos de M*achine Learning*.

Determinaremos los cambios de cada etapa fenológica de una campaña agrícola de colza en parcelas distribuidas por toda Francia.

Inicialmente realizaremos la extracción de la información espectral de las 10 bandas de Sentinel-2, se calcularán los índices espectrales y se evaluará su potencial de clasificación en el estado de floración, posteriormente se acoplarán datos meteorológicos a la información espectral y finalmente se hará uso de métodos de *Machine Learning* como la regresión logística penalizada multinomial (LASSO), el *K-Nearest Neighbors* (KNN) y *Random Forest* (RF) para determinar el aporte de las variables temáticas a la determinación de patrones en los datos.

En este caso de aplicación, el uso de métodos de *Machine Learning*, nos permitirá conocer el aporte de la teledetección a la gestión sostenible de los bioagresores en cultivos de gran importancia agroalimentaria determinando la combinación adecuada de variables para la clasificación de estados fenológicos del Colza (*Brassica napus L.*).

1. Materiales y Métodos

La metodología se divide en 3 pasos. La primera sección describe las bases de datos utilizadas para la recuperación de la información utilizada. Presenta también, las regiones en donde se encuentran las parcelas. La segunda sección presenta los tres métodos de clasificación utilizados para la detección de los estados fenológicos. En la tercera parte se detalla la metodología utilizada para definir el aporte de los diferentes conjuntos de variables.

1.1. Materiales

1.1.1. Datos Agronómicos

Vigicultures®

Aplicación departamental de introducción de datos epidemiológicos para cultivos de campo (colza, trigo, girasol, etc.) implementada por institutos técnicos (Arvalis, Terre Inovia, ITB) (Simonneau, Chollet, et Gouwier 2013). **Vigicultures®** junto con la base de datos *VégéObs* reúne datos de vigilancia epidemiológica para obtener información en tiempo real sobre la presión de las plagas en los cultivos. Esta base de datos orquestada por el ministerio de agricultura y el ministerio del medio ambiente es una herramienta clave de prevención y análisis de riesgos para la creación de Boletines Fitosanitarios (BSV). Para nuestro caso de estudio, utilizamos los estados fenológicos de los cultivos que son registrados cada vez que una observación de plagas es realizada.

Estados fenológicos

El estado fenológico de las parcelas es establecido a partir de una clasificación propia, establecida en la base de datos **Vigicultures®.** Para el cultivo de colza se identificaron 28 estados fenológicos (Semis, A, B1, B2, B3, B4, B5, B6, B7, B8, B9, B10, "> 10 feuilles", C1, C2, D1, D2, E, F1, F2, G1, G2, G3, G4 - Floraison toujours en cours, Fin floraison, G4 - Floraison terminée, G5 et Hors culture). Los estados “G4 - Floraison Terminée” y “Hors de Culture” fueron descartados debido a su ambigüedad y su reducido número de observaciones.

A continuación se hace un paralelo de los estados **Vigicultures®** con la escala **BBCH** (Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie). La escala BBCH describe las etapas fenológicas de los cultivos utilizando criterios que relacionan la etapa de crecimiento con un código decimal (Meier 2001). El primer dígito indica la etapa principal de desarrollo (por ejemplo, 6 = floración), mientras que el segundo dígito se refiere a una etapa secundaria de crecimiento o al porcentaje de plantas en esa etapa.

Tabla 1. Paralelo entre la escala **Vigicultures®** y la escala **BBCH**

Escala **Vigicultures® original**	Escala **Vigicultures® Agrupada**	Escala **BBCH** (Meier 2001)
Semis, A	A	**fase 0:** Germinación, brotación, desarrollo de la yema.
B1, B2, B3, B4, B5, B6	B1 -B6	**fase 1:** Desarrollo de las hojas (tallo principal).
B7, B8, B9, B10, > 10 feuilles	B7 - B10>	**fase 2:** Formación de brotes laterales / (ahijamiento).
C1, C2	C	**fase 3:** Crecimiento longitudinal del tallo o crecimiento en roseta, desarrollo de brotes (retoños)/ encañado (tallo principal).
D1, D2	D	**fase 4:** Desarrollo de las partes vegetativas cosechables de la planta o de órganos vegetativos de propagación / embuchamiento.
E	E	**fase 5:** Emergencia de la inflorescencia (tallo principal).
F1, F2, G1, G2, G3, G4 - Floraison toujours en cours, G4 - Floraison terminée, G5	F-G	**fase 6:** Floración (tallo principal).
		fase 7: Desarrollo del fruto.
		fase 8: Coloración o maduración de frutos y semillas.
NA	NA	**fase 9:** Senescencia.

Registre Parcellaire Graphique (RPG))

Base de datos geográfica que se utiliza como referencia para la evaluación de las ayudas de la Política Agrícola Común (PAC) europea. La versión anónima contiene datos gráficos de parcelas (desde 2015) con su cosecha principal. Estos datos han sido producidos por el Organismo de Servicios y Pagos (SPA) desde 2007. La reutilización del RPG es gratuita para todos los usos, incluidos los comerciales, según los términos de la "licencia abierta"[1].

[1] https://www.data.gouv.fr/

Identificación de las parcelas de interés

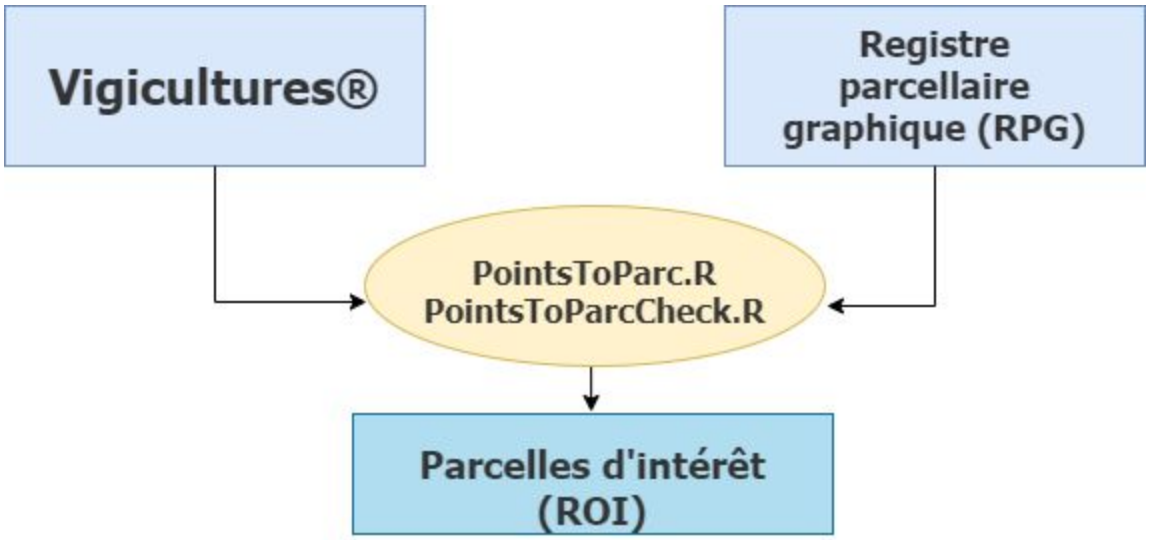


Fig 1. Diagrama general de preprocesamiento de las bases de datos agronómicas

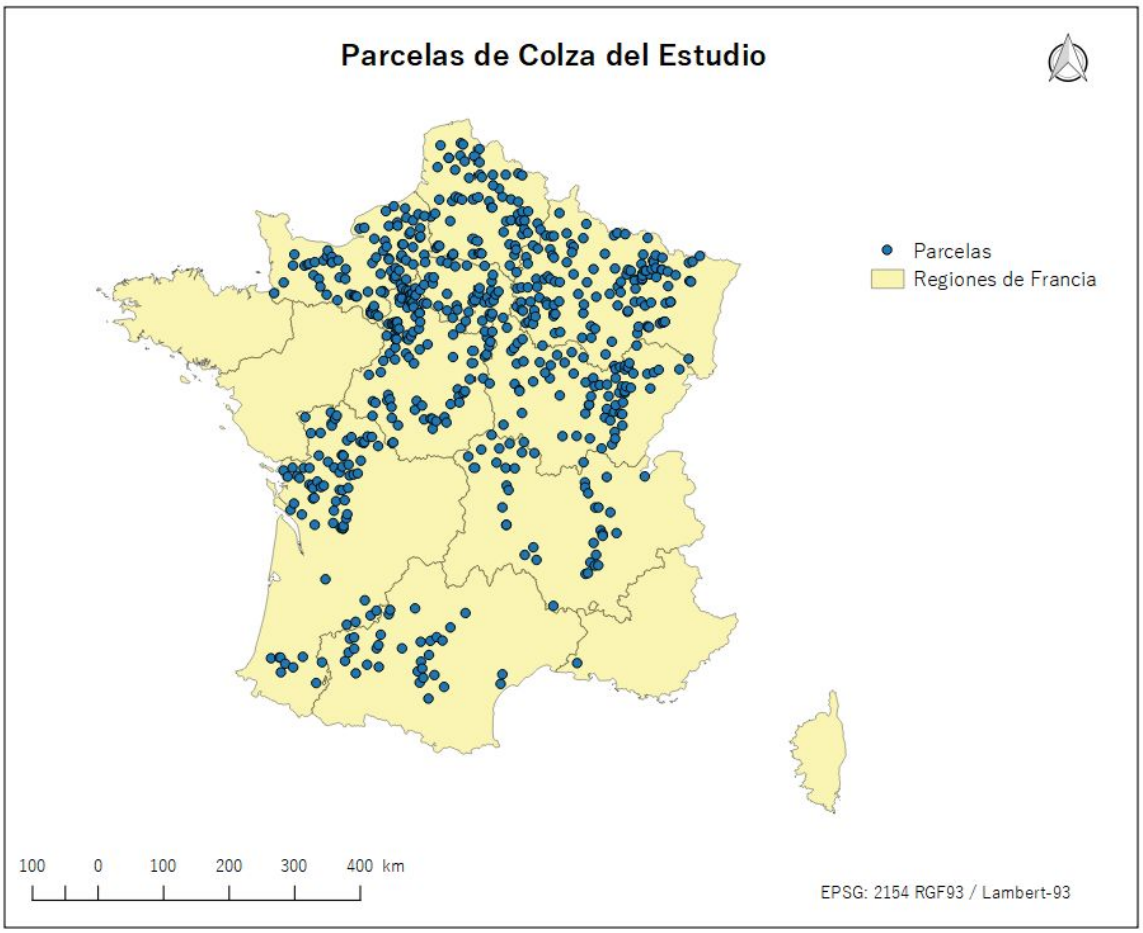


Fig 2. Mapa de las parcelas de interés en Francia

De la base de datos **Vigicultures®** es extraída la información de parámetros agrícolas (tipo de cultivo, estado fenológico observado, departamento, etc) relacionados a un punto GPS, esa información es fusionada con la información relacionada a la parcela registrada en la base de données RPG. Los polígonos resultantes demarcan las regiones de interés (ROI) para el posterior análisis con las imágenes satelitales y las variables climatológicas.

1.1.2. Datos Espectrales

Sentinel-2

El conjunto de satélites ópticos Sentinel-2 (2A et 2B) hace parte de la familia de satélites del proyecto espacial europeo para la observación remota de la tierra. Desde junio de 2015, las imágenes multiespectrales permiten analizar el ciclo de desarrollo y crecimiento de las plantas a una escala mundial. Con 13 bandas espectrales a una alta resolución espacial (4 bandas a 10m, 6 bandas a 20m et 3 bandas a 60m) y un tiempo de revisita de 5 días, su aplicación en agricultura es de las más documentadas (Zhang, Friedl, et Schaaf 2009).

Transformación de la Información Espectral

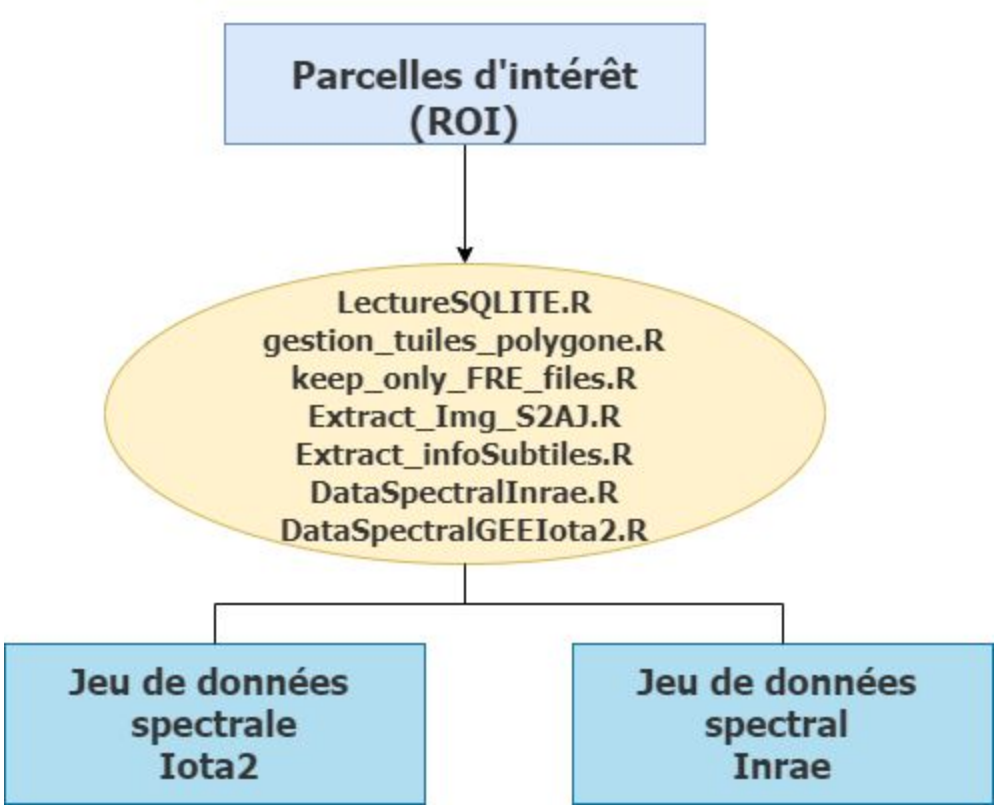


Fig 3. Diagrama general de preprocesamiento de la información espectral Sentinel-2

La información espectral es obtenida a partir de dos metodologías distintas. En ambas metodologías, las tiles Sentinel-2 de Nivel 2A fueron descargadas del Theia[2] Centro de datos terrestres (Hagolle [2016] 2020). Las adquisiciones corresponden a la campaña de cosecha de 2017 (entre 1 de Julio de 2016 y 25 de agosto 2017).

En la primera metodología, la extracción de datos espectrales se realizó utilizando *iota2* (Inglada et al. 2016) y *MAJA* (MACCS[3]-ATCOR[4] Joint Algorithm) desarrollada por el Centre National d'Etudes Spatiales (CNES) y el Centre d'Etudes Spatiales de la Biosphère (CESBIO), por un lado, y el Centro Aeroespacial Alemán (DLR), por otro. Las imágenes son ortorectificadas, corregidas atmosféricamente sin nubes y con detección de sombras (Baetens, Desjardins, et Hagolle 2019). Todas las adquisiciones fueron re-muestreadas para rellenar los espacios dejados por las nubes y las sombras (cada 10 días, comenzando en 2016-07-01 y terminando en 2017-08-25). Las 10 bandas de S2 (B2, B3, B4, B5, B6, B7,B8, B8A, B11 et B12) son recuperadas a una resolución espacial de 10 y 20 metros sin proceso de resampling.

En la segunda metodología, las adquisiciones fueron realizadas a partir de la herramienta *SEN2COR* (Muller-Wilm 2012). Las 10 bandas se presentan en dos formas: una forma, Reflectancia de Superficie corregida por efectos atmosféricos y ambientales (SRE_Bx.tif), otra forma, Reflectancia Plana que es adicionalmente corregida por efectos de pendiente (FRE_Bx.tif)[5]. Trabajaremos con los datos S2 L2A usando el producto FRE_Bx.tif. Las bandas fueron extraídas en su resolución original siendo después transformadas todas a 10 metros utilizando como parámetro para definir el nuevo valor de los píxeles, el método de Nearest Neighbor.

En ambos casos, las parcelas de interés recuperadas de acuerdo con las bases de datos agronómicas se asocian a la información espectral de las tiles relacionadas a su localización geográfica. Las imágenes satelitales son seleccionadas a partir de la fecha de observación de los diferentes estados fenológicos. Esta selección busca que la diferencia entre la fecha de observación del estado y la fecha de la información espectral sea entre 0 y 5 días antes de la observación in-situ.

Índices Espectrales

Las bandas espectrales fueron utilizadas para obtener los índices espectrales que se consideraron pertinentes para el análisis de los estados fenológicos en agricultura. En las siguientes tablas presentamos las bandas espectrales y los índices utilizados en este caso de estudio.

[2] https://theia.cnes.fr
[3] Multi-sensor Atmospheric Correction and Cloud Screening software (MACCS)

[4] Atmospheric Correction software (ATCOR)

[5] https://labo.obs-mip.fr/multitemp/sentinel-2/theias-sentinel-2-l2a-product-format/

Tabla 2. Bandas Espectrales Sentinel-2 utilizadas

Nombre	Resolución	Longitud de Onda	Descripción
B2	10 metros	496.6nm (S2A) / 492.1nm (S2B	Azul
B3	10 metros	560nm (S2A) / 559nm (S2B)	Verde
B4	10 metros	664.5nm (S2A) / 665nm (S2B)	Rojo
B5	10 metros	703.9nm (S2A) / 703.8nm (S2B)	Red Edge 1
B6	20 metros	740.2nm (S2A) / 739.1nm (S2B)	Red Edge 2
B7	20 metros	782.5nm (S2A) / 779.7nm (S2B)	Red Edge 3
B8	20 metros	835.1nm (S2A) / 833nm (S2B)	Infrarrojo Cercano
B8A	20 metros	864.8nm (S2A) / 864nm (S2B)	Red Edge 4
B11	20 metros	1613.7nm (S2A) / 1610.4nm (S2B)	SWIR 1
B12	20 metros	2202.4nm (S2A) / 2185.7nm (S2B)	SWIR 2

Tabla 3. Índices espectrales utilizados y sus fórmulas

Índices	Fórmula para Sentinel-2	Fuente
Normalized Difference Vegetation Index (NDVI)	$NDVI = \frac{B8 + B4}{B8 - B4}$	(Rouse et al. 1973)
Green Normalized Difference Vegetation Index (GNDVI)	$GNDVI = \frac{B8 - B3}{B8 + B3}$	(Gitelson, Kaufman, et Merzlyak 1996)
Normalized Difference Water Index (NDWI)	$NDWI = \frac{B3 - B8}{B3 + B8}$	(Gao 1996)
Normalized Difference Yellow Index (NDYI)	$NDYI = \frac{B3 - B2}{B3 + B2}$	(Sulik et Long 2016)

Normalized Difference Moisture Index (NDMI)	$NDMI = \frac{B8A - B11}{B8A + B11}$	(Sykas 2019)
Enhanced Vegetation Index (EVI)	$EVI = 2.5\left[\frac{B8 - B4}{B8 + 6B4 - 7.5B2 + 1}\right]$	(Liu et Huete 1995)
Structure Insensitive Pigment Index (SIPI)	$SIPI = \frac{B8 - B2}{B8 + B4}$	(Sykas 2019)
Soil Adjusted Vegetation Index (SAVI)	$SAVI = \frac{B8 - B4}{1.428\,(B8 + B4 + 0.428)}$	(Huete 1988)
Atmospherically Resistant Vegetation Index (ARVI)	$ARVI = \frac{B8 - 2B4 + B2}{B8 + 2B4 + B2}$	(Tanre, Holben, et Kaufman 1992)
Advanced Vegetation Index (AVI)	$AVI = [B8 * (1 - B4) * (B8 - B4)]^{1/3}$	(Roy, Sharma, et Jain 1996)
Bare Soil Index (BSI)	$BSI = \frac{(B11 + B4) - (B8 + B2)}{(B11 + B4) + (B8 + B2)}$	(Sykas 2019)
Moisture Stress Index (MSI)	$MSI = \frac{B11}{B8}$	(Rock, Williams, et Vogelmann 1985)

Tasseled Cap

Además de los índices espectrales mencionados arriba, la información espectral obtenida fue transformada a partir de la metodología de "*Tasseled Cap*".

Kauth, R. J. and Thomas, G. S. (1976) idearon una transformación de la información de las bandas espectrales para maximizar la información contenida en nuevos elementos de análisis. Es un método de compresión para reducir múltiples datos espectrales, concretamente de 6 bandas, en tres neo-canales que permiten comprender fenómenos importantes del desarrollo de cultivos en el espacio espectral (Kauth et Thomas 1976). Los neo-canales obtenidos después de la transformación son los siguientes:

Tabla 4. Transformaciones *Tasseled Cap*

Índices	Fórmula para Sentinel-2	Utilización
Brightness*[6] *Index	$BI = \sqrt{\frac{B4^2 + B3^2}{2}}$	Asociado a las variaciones de reflectancia del suelo.
***Greenness* (verdor):**	**Greenness**[7] = (-0.2848B2)+(-0.2435B3)+(-0.5436B4) +0.7243B8+0.0840B11+(-0.1800B12)	Correlacionado con el vigor de la vegetación
***Wetness* (Humedad)**	**Wetness**[8] = 0.1509B2+0.1973B3+0.3279B4 +0.3406B8+(-0.7112B11)+(-0.4572B12)	Influído por las bandas en el IR medio y tiene que ver con la humedad vegetal y del suelo.

1.1.3. Datos meteorológicos

AgroClim

AgroClim es una unidad al servicio de la comunidad del INRAE. Esta unidad gestiona la red agroclimática nacional del INRAE y la base de datos correspondiente. Su función es asegurar la trazabilidad de las observaciones dependientes del clima. AgroClim es también el punto de entrada único para que las unidades del INRAE obtengan datos meteorológicos de Météo-France[9] .

Los datos utilizados son producto del modelo de datos climatológicos elaborado por Météo-France, *SAFRAN* (Système d'Analyse Fournissant des Renseignements Atmosphériques à la Neige). Safran trabaja en regiones de clima homogéneo. Estas regiones tienen una forma irregular, y su superficie es normalmente inferior a 1.000 km^2. En cada región homogénea, Safran estima la variación de 8 parámetros climáticos (tabla 5) para cada clase de altitud de 300 m, sobre la base de todos los datos climáticos disponibles (estaciones meteorológicas, pero también análisis de modelos de pronóstico meteorológico a gran escala, como el modelo ARPEGE de Météo-France) (Lemaire 2015). Los análisis de temperatura, humedad, velocidad del viento y cobertura de nubes se producen cada 6 horas. El análisis de las precipitaciones se hace diariamente. Después de obtener los valores de las zonas, el análisis se interpola espacialmente en una cuadrícula regular de 8 km x 8 km.

[6] https://foodsecurity-tep.net/S2_BI
[7] https://www.indexdatabase.de/search/?s=tasselled+cap
[8] https://www.indexdatabase.de/search/?s=tasselled+cap
[9] https://www6.paca.inrae.fr/agroclim/

Tabla 5. Données spatialisées par le modèle Safran de Météo - France (Lemaire 2015)

Données disponibles	Période	Résolution de la maille
1. Temperaturas mínima, máxima y media a 2 m sobre el nivel del suelo (°C) ; 2. Humedad relativa media a 2 m sobre el suelo (en g.kg-1) ; 3. Fuerza media del viento a 10 m sobre el nivel del suelo (en m/s); 4. 4. Precipitación sólida (en mm) 5. Precipitación líquida (en mm) 6. Radiación infrarroja/solar (en J/cm^2) 7. Radiación atmosférica (en J.cm-2) 8. Evapotranspiración potencial (FTE mm), La fórmula de Penman-Monteith	1958 à aujourd'hui	8 km x 8 km

Transformación de la información meteorológica

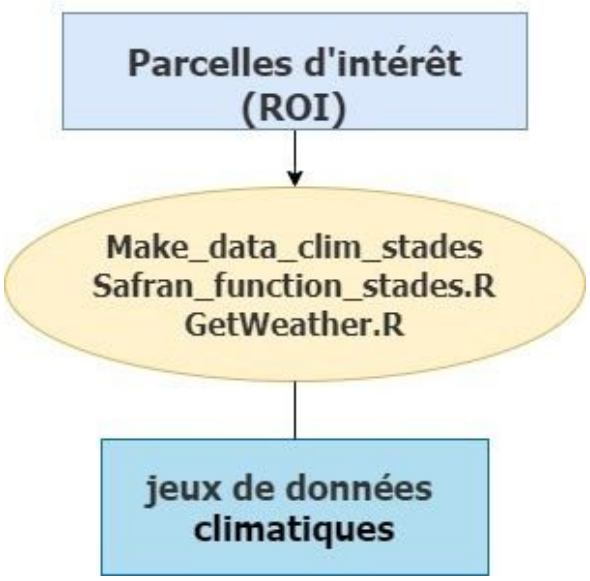


Fig 4. Diagrama general de preprocesamiento de las bases de datos meteorológicos

Para este estudio, consideramos todas las variables climatológicas obtenidas por el modelo SAFRAN. Adicionamos otra variable, los grados día de crecimiento acumulados (gdd), la cual está íntimamente relacionada a la evolución fenológica de los cultivos. El cálculo de esta variable basado en la siguiente fórmula:

$$GDD = (Tmax + Tmin)/2 - Tbase$$

Utilizamos la temperatura base de 5° según (Morrison, McVETTY, et Shaykewich 1989) y la función gdd() del paquete ***pollen***[10], basada en (Baskerville et Emin 1969).

[10] https://cran.r-project.org/web/packages/pollen/vignettes/gdd.html

Los datos diarios departamentales de las estaciones meteorológicas más próximas a las parcelas de interés fueron agrupados por semanas. La predicción de los estados se realizó con la información climatológica de las últimas 52 semanas a la fecha de observación in situ. Esta decisión está basada en la hipótesis empírica de que las variaciones de las condiciones meteorológicas durante al menos 10 meses pueden impactar el crecimiento de la planta de la siembra hasta la cosecha. Además, la información meteorológica es un proxy de la información temporal que podría ser útil para identificar si es tiempo de sembrar ya que las variaciones de temperatura, por ejemplo, permiten a un modelo como *Random Forest*, encontrar oscilaciones en la señal. Si consideramos los desafíos que presenta el cambio climático actual a los procesos de modelización, esta identificación de la temporada tiene una ventaja sobre la fecha de observación pues podemos adaptar la meteorología a un periodo específico del año, esto nos permite ajustar el modelo a otras regiones y otros años. Por otra parte, desde el punto de vista de preprocesamiento de los datos, si extraemos 10 meses para un estado fenológico es congruente hacerlo para todos los otros, para tener el mismo número de variables independientes por clase.

Construcción del juego de datos final

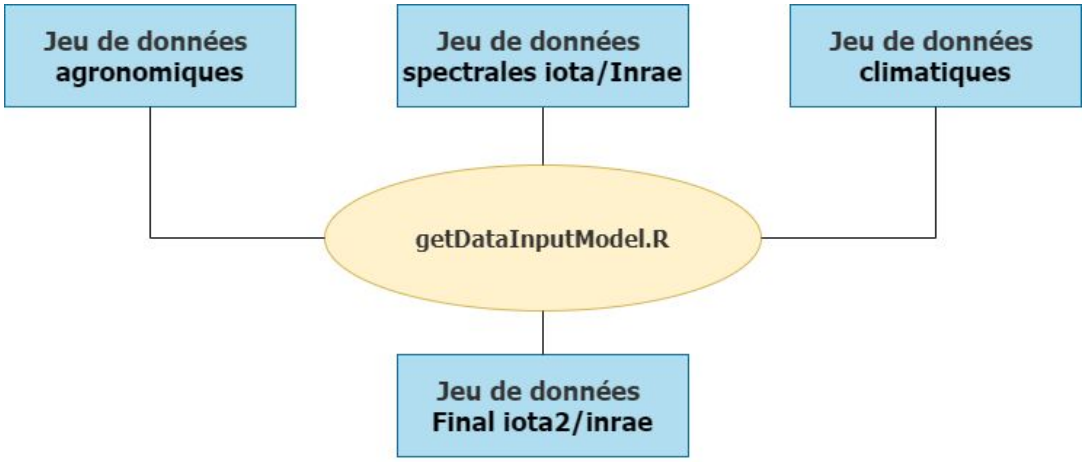


Fig 5. Diagrama general para la construcción del juego de datos final

Los datos climatológicos fueron fusionados con la información agronómica y espectral a partir del identificador único de cada parcela de interés. Al final del preprocesamiento se obtuvo el siguiente juego de datos:

Tabla 6. Composición final del juego de datos para las modelizaciones

Nb de parcelas	Nb de variables	Nb de observaciones
561	519 28 espectrales 491 Climáticas	3033

1.2. Métodos

1.2.1. Métodos de clasificación utilizados

Lasso Multinomial (GLM)

En 1996 Tibshirani elaboró el LASSO (Least Absolute Shrinkage and Selection Operator) que es un método que reduce a cero el coeficiente de regresión de variables menos impactantes. Asociado a una validación cruzada, permite el nivel de impacto adecuado y así, realiza una selección de variables. La idea es que el método LASSO minimiza la suma de los cuadrados residuales para los que la suma de las estimaciones (coeficientes) no es mayor que una cierta constante (Efendi et Ramadhan 2018). Dicho de otra manera, LASSO restringe la estimación a menos de una cierta constante (en este caso, usamos el lambda1se) de modo que algunas estimaciones son cero.

Para predecir variables categóricas múltiples, la utilización del modelo logit multinomial en el análisis de regresión para las respuestas de múltiples categorías no ordenadas es la más utilizada (Tutz, Pößnecker, et Uhlmann 2015).

En este caso, utilizamos el paquete ***glmnet*** para ajustar el modelo de referencia. El modelo permite determinar las variables más importantes en la clasificación de los estados fenológicos.

Multinomial Logistic Regression (MLR) - Redes Neuronales

La regresión multinomial es una extensión de la regresión logística binomial. El algoritmo nos permite predecir una variable dependiente categórica que tiene más de dos niveles (Hosmer et Lemeshow 1989). Como cualquier otro modelo de regresión, la salida multinomial puede predecirse usando una o más variables independientes. Las variables independientes pueden ser de tipo nominal, ordinal o continua.

El MLR aplica una transformación logarítmica no lineal que permite calcular la probabilidad de aparición de cualquier número de clases de una variable dependiente basándose en variables explicativas. A diferencia de los modelos de regresión lineal que utilizan los mínimos cuadrados como estimador, los coeficientes de la MLR se estiman típicamente utilizando la máxima probabilidad (Jeune et al. 2018).

Para esta modelización utilizamos el paquete ***nnet*** para ajustar el modelo multinomial a través de una red neuronal.

Ordinal Logistic Regression (OLR)

Uno de los modelos estadísticos más apropiados para el análisis de los datos con una variable de respuesta categórica es el modelo de regresión logística (Efendi et Ramadhan 2018). La regresión logística ordinal es una extensión del modelo de regresión logística simple. En la regresión logística simple, la variable dependiente es categórica y sigue

una distribución de Bernoulli. En la regresión logística ordinal la variable dependiente es ordinal, es decir, hay un ordenamiento explícito en las categorías (Ananth et Kleinbaum 1997).

El modelo de regresión logística ordinaria tienen en cuenta el orden de la variable dependiente categórica utilizando eventos acumulativos para el cálculo del logaritmo de las probabilidades (Ananth et Kleinbaum 1997). Esto significa que, a diferencia de la regresión logística simple, los modelos logísticos ordinales consideran la probabilidad de un evento y todos los eventos que están por debajo del evento focal en la jerarquía ordenada.

En este caso de estudio, una vez ordenada la variable categórica de los estados fenológicos, se usó la regresión logística ordinal para predecir los estados en función de las variables independientes. Esto nos permitirá determinar cuál de nuestras variables independientes (si alguna) tiene un efecto estadísticamente significativo en nuestra variable dependiente. El paquete utilizado en R fue ***ordinal***.

Random Forest (RF)

Los bosques aleatorios son una combinación de árboles de decisión. En este método de clasificación cada árbol depende de los valores de un vector aleatorio muestreado independientemente con la misma distribución para todos los árboles del bosque (Breiman 2001). El error de generalización para los bosques converge en un límite a medida que el número de árboles en el bosque se hace grande. El error de generalización de un bosque de clasificadores de árboles depende de la fuerza de los árboles individuales del bosque y de la correlación entre ellos (Boulesteix et al. 2012).

Random Forest es un algoritmo muy interesante para el manejo de información espectral y el acoplamiento con otras variables (como las climáticas, por ejemplo) (Muñoz et al. 2018). Presenta características como el funcionamiento eficaz con grandes conjuntos de datos, la capacidad de identificar patrones de asociación no lineales entre los predictores y la respuesta, además de manejar variables de predicción altamente correlacionadas (Kühnlein et al. 2014).

El algoritmo genera una estimación interna no sesgada del error de generalización (error OOB) y tiene la capacidad de determinar qué variables son importantes en la clasificación (Breiman 2001).

Los paquetes utilizados en R fueron ***RandomForest*** y ***Caret***. En la clasificación de estados fenológicos, el modelo de *Random Forest* fue estimado con **500 árboles**.

k-Nearest Neighbors (kNN)

El algoritmo de clasificación kNN se ha convertido en un método importante en la minería de datos y el aprendizaje de máquinas desde que fue propuesto en 1967 (Deng et

al. 2016). Para aplicar el método tradicional de kNN en grandes volúmenes de datos, las metodologías pueden ser a menudo categorizadas en dos partes, es decir, encontrar rápidamente las muestras más cercanas o seleccionar muestras representativas (o la eliminación de algunas muestras) para reducir el cálculo de kNN (Zhu, Zhang, et Huang 2014).

KNN es un algoritmo de clasificación estándar basado exclusivamente en la elección de la métrica de clasificación. Es "no paramétrico". Sólo debe establecerse la k, que es el número de vecinos a partir del cual se establecen las distancias. K es un valor entero especificado por el usuario. La elección óptima del valor depende en gran medida de los datos. En general, un valor mayor suprime los efectos del ruido, pero hace que los límites de la clasificación sean menos claros.

En este caso de estudio, el algoritmo fue utilizado en R a partir del paquete **Caret**[11], determinando como método de control la validación cruzada con 10 folds.

1.2.2. Detección del estado de Floración

A manera de test inicial, se realizó una primera clasificación binaria del estado de floración. Utilizamos un modelo basado en la capacidad predictiva de los índices espectrales. El método utilizado fue *Random Forest* y se ajustó para los estados fenológicos agrupados en 8 clases.

1.2.3. Condiciones de Referencia

El modelo de referencia es construido considerando los estados fenológicos en función de las variables climatológicas y espectrales (Figura 6). El dataset de datos espectrales utilizado es resultado de la cadena de tratamiento iota2 (primera metodología de extracción).

Inicialmente, ajustamos el modelo de referencia a partir de los cuatro métodos de clasificación seleccionados para este caso de estudio (*Lasso Multinomial*, *Ordinal Logistic Regression*, *Random Forest* y *K-Nearest Neighbors*). Después, evaluamos los cuatro métodos en función de la precisión y el tiempo de cálculo. Finalmente, seleccionamos el *Random Forest* para clasificar los estados fenológicos agrupados. Los índices fueron las variables espectrales utilizadas.

[11] https://cran.r-project.org/web/packages/caret/caret.pdf

Estados ~ Clima + Espectral

Donde:

clima = Temperatura Mínima
Temperatura Media
Temperatura Máxima
Precipitación
Evapotranspiracion
Velocidad media del viento
Radiación Solar
Grados día de crecimiento acumulados (gdd)
Humedad Relativa

Espectral = Normalized Difference Vegetation Index (NDVI)
Normalized Difference Water Index (NDWI)
Green Normalized Difference Vegetation Index (GNDVI)
Normalized Difference Yellow Index (NDYI)
Normalized Difference Moisture Index (NDMI)
Enhanced Vegetation Index (EVI)
Advanced Vegetation Index (AVI)
Soil Adjusted Vegetation Index (SAVI)
Moisture Stress Index (MSI)
Bare Soil Index (BSI)
Atmospherically Resistant Vegetation Index (ARVI)
Structure Insensitive Pigment Index (SIPI)

Figure 6. Modèle de référence

Las condiciones de referencia son establecidas como la línea base para evaluar y/o mejorar la clasificación en función del acoplamiento o no de otras variables temáticas. Esta línea base se establece para testar la variación de una variable a la vez y no todas las combinaciones de variables.

El problema de investigación está dividido en preguntas específicas que buscan ser resueltas por modificar una variable a la vez a partir de dichas condiciones de referencia. La selección de las condiciones de referencia está fundamentada en la experiencia del equipo de trabajo y en el soporte académico:

- ✓ El algoritmo *Iota2*, es una cadena de procesamiento para la producción operativa de mapas de la cubierta terrestre a partir de series temporales de imágenes de teledetección utilizando una clasificación supervisada (Inglada et al. 2016; Fauvel et al. 2020). Su versatilidad y nivel de precisión permite utilizarla en contextos diversos.
- ✓ La utilización de *índices espectrales* en agricultura ha sido una de las metodologías de análisis de mayor tendencia en las últimas tres décadas (Bolton et Friedl 2013). En particular, los indices normalizados de vegetacion como el *NDVI*, han sido fuertemente utilizados dadas sus ventajas de interpretacion al mejorar la discriminación entre el suelo y la vegetación, reduciendo el efecto del relieve en

la caracterización espectral de las diferentes cubiertas (Islam et Bala 2008; Bolton et Friedl 2013).

- ✓ El algoritmo de *Random Forest (RF)* es un método de clasificación con menor sensibilidad a la calidad de las muestras de entrenamiento y al sobreajuste (en comparación a otros métodos). Estas ventajas se deben al gran número de árboles de decisión producidos al seleccionar aleatoriamente un subconjunto de muestras de entrenamiento (Belgiu et Drăguţ 2016). Además de lo anterior, es un método ya utilizado por el equipo de investigación en el que se encuadra esta práctica.
- ✓ La selección de variables climáticas y espectrales con el objetivo de que el modelo sea reproducible a distintas escalas espaciales y en distintos lugares geográficos es una estrategia de generalización para modelizaciones futuras.
- ✓ Los estados fenológicos reagrupados en 8 clases hacen más precisa la tarea de clasificación. En este caso de estudio, la imprecisión de los datos in-situ y la limitación temporal de la información espectral y climática (una observación semanal) hace difícil distinguir correctamente los 26 estados. Finalmente el interés agronómico de esta clasificación se concentra en los estados más representativos del cultivo.

1.2.4. Comparación de modelos

Ajustamos diferentes modelos de clasificación para los 8 estados fenológicos agrupados registrados (ver tabla 1). Después comparamos dichos modelos con el modelo de referencia (condiciones de referencia).

Como primer paso, ajustamos el modelo de referencia utilizando los cuatro métodos de clasificación seleccionados para este estudio de caso (Lasso Multinomial, Regresión Logística Ordinal, *Random Forest* y *K-Nearest Neighbors*). A continuación, evaluamos los cuatro métodos en términos de *accuracy* y tiempo de cálculo. Finalmente, seleccionamos el *Random Forest*.

La idea fue luego crear modelos que buscan determinar la relevancia y/o importancia de los grupos de variables (espectrales, climatológicas y espacio-temporales) para la identificación de los estados fenológicos. Evaluamos el potencial predictivo de las variables temáticas de forma aislada, considerando modelos en los que a partir de un solo grupo de variables se pudiera identificar con precisión los estados. En este caso se utilizaron los siguientes modelos:

Estados ~ Bandas Espectrales

Estados ~ Índices espectrales

Estados ~ Tasseled Cap

Estados ~ Clima

Estados ~ Espacio-Tiempo*

Fig. 7 Modelos individuales de clasificación. * Fecha de observación *in-situ*, departamento

Después, analizamos el potencial de los índices espectrales con los datos interpolados a diez días (iota2) y los datos sin interpolación (inrae). Finalmente acoplamos variables espectrales, climáticas y espaciotemporales para determinar el potencial de clasificación en conjunto.

Estados ~ Índices espectrales + Espacio-Tiempo

Estados ~ Clima + Espacio-Tiempo

Estados ~ Índices + Clima + Espacio-Tiempo

Fig. 8 Modelos acoplados de clasificación

La evaluación de los diferentes modelos de clasificación se realizó a partir de sus matrices de confusión y las siguientes métricas:

Tableau 7. Métricas de evaluación de los modelos

Medida	Fórmula	Concepto
Average Accuracy **(Sokolova et Lapalme 2009)**	$\frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$ tp_i son verdaderos positivos, fp_i - falsos positivo, fn_i - falso negativo, y tn_i - verdadero negativo, respectivamente.	La eficacia media por clase de un clasificador
Coeficiente kappa de Cohen (McHugh 2012)	$k = \frac{(p_o - p_e)}{(1 - p_e)}$ p_o es la probabilidad empírica de acuerdo en la etiqueta asignada a cualquier muestra (la proporción de acuerdo observada), y p_e es el acuerdo esperado cuando ambos anotadores asignan etiquetas al azar. p_e Se estima utilizando un previo empírico por anotador sobre las etiquetas de clase.	La puntuación kappa es un número entre -1 y 1. Las puntuaciones superiores a 0,8 se consideran generalmente como un buen acuerdo; cero o menos significa que no hay acuerdo (etiquetas prácticamente aleatorias).
Out-of-bag (OOB) error **(Hastie, Tibshirani, et Friedman 2009)**	*Random Forest* se entrena utilizando la agregación de bootstrap, donde cada nuevo árbol se ajusta a partir de una muestra de bootstrap de las observaciones de entrenamiento $Z_i = (x_i, y_i)$. El error fuera de bolsa (OOB) es el error medio de cada árbol calculado utilizando las predicciones de los árboles que no contienen en su respectiva muestra de bootstrap. Esto permite que *Random Forest* se ajuste y valide mientras se está entrenando.	

2. Resultados

Los estados fenológicos de **Vigicultures®** determinadas in-situ se establecen como labels de clasificación. Los estados observados son la variable dependiente a predecir. Los perfiles espectrales de Sentinel-2 (S2) son promediados para cada una de las 561 parcelas estudiadas. En la primera parte, realizamos una clasificación binaria (presencia o ausencia de flores) para el estado de floración con el método *Random Forest*, considerando únicamente la información espectral. En la segunda parte, evaluamos en función de la precisión y el tiempo de cálculo, los cinco métodos de clasificación seleccionados. En la tercera parte realizamos clasificaciones teniendo en cuenta el acoplamiento entre las variables las espectrales, climatológicas y espacio-temporales utilizando como base el modelo de referencia. Evaluamos el potencial predictivo de cada uno de los modelos a partir de las métricas resultantes de las matrices de confusión. Finalmente, analizamos el impacto de dos factores en la clasificación: la agrupación de los estados fenológicos y la creación de un subconjunto de datos balanceados.

2.1. Clasificación Binaria del Estado de Floración con el método de *Random Forest*

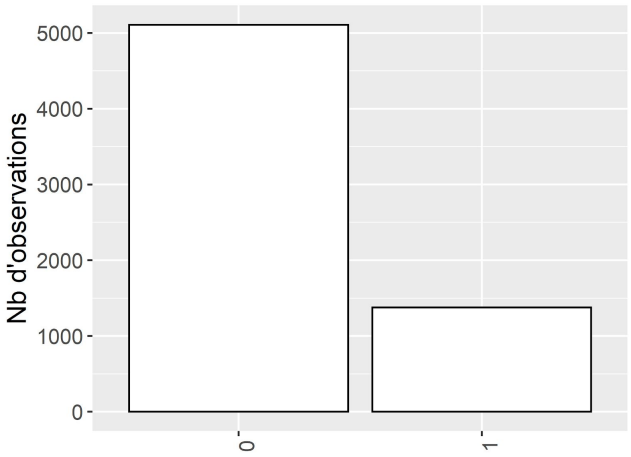
Modelo para la Floración

Realizamos un análisis preliminar para determinar la capacidad predictiva de las variables espectrales (índices) en una clasificación binaria, estado de floración (presencia o ausencia de flores).

Floración ~ Índices espectrales

Donde :
Indices espectrales = Normalized Difference Vegetation Index (NDVI), Normalized Difference Water Index (NDWI), Green Normalized Difference Vegetation Index (GNDVI), Normalized Difference Yellow Index (NDYI), Normalized Difference Moisture Index (NDMI), Enhanced Vegetation Index, Advanced Vegetation Index, Soil Adjusted Vegetation Index, Moisture Stress Index (MSI), Bare Soil Index, Atmospherically Resistant Vegetation Index, Structure Insensitive Pigment Index

Fig. 9 Modelo de floración



De acuerdo al gráfico, los datos se encuentran desbalanceados. De 6494 observaciones, tenemos 1376 (21%) en estado de floración y 5118 (79%) que no. Este desbalance en los datos se debe a que estamos confrontando un solo estado a los demás.

Fig. 10 Distribución de las observaciones para las clases binarias (Flor - no flor)

Los resultados del modelo de clasificación se presentan abajo:

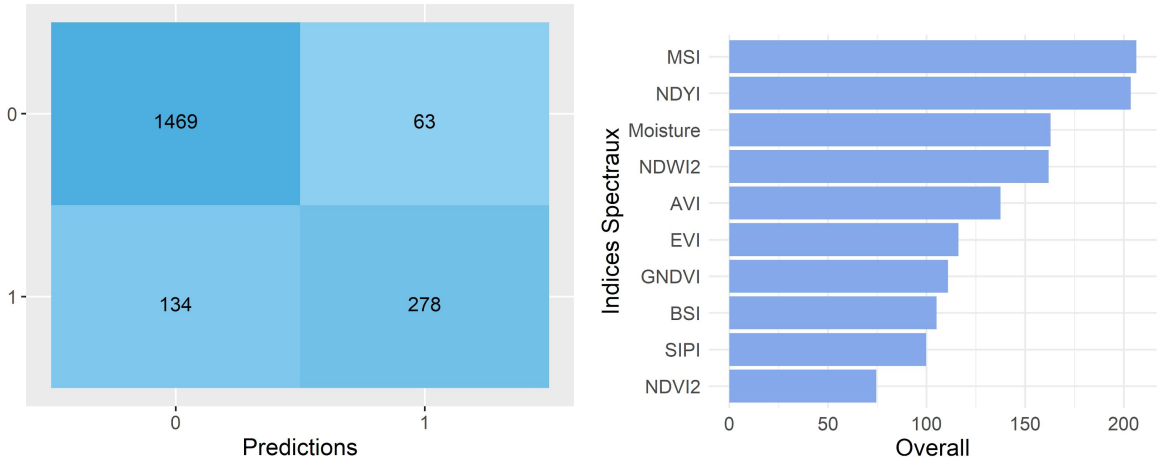


Fig 11. (Izquierda). Matriz de Confusión. (Derecha) Importancia de las variables

La matriz de confusión nos muestra la dificultad que presenta el modelo para determinar de manera adecuada el estado de floración cuando los datos están desbalanceados. Para este estado, la tasa de falsos positivos (elementos mal clasificados) es importante, sin embargo el modelo acierta en el 72.41% de los casos para la floración (ver tabla 8).

Tabla 8. Matriz de confusión binaria

	Predicciones	
Clases	0	1
0	95.52%	4.48%
1	27.59%	72.41%

En cuanto a las variables que mejor explican el modelo, índices espectrales como *Moisture Stress Index* (*MSI*), el *Normalized Difference Yellow Index* (*NDYI*) y el Normalized *Difference Water Index* (*NDWI*) son los que mejor explican la presencia o ausencia de flores en las observaciones analizadas.

Las métricas de evaluación del modelo nos muestran que para los datos de entrenamiento el OOB es inferior al 10%. La *accuracy* y el coeficiente kappa son 0.91 y 0.71 respectivamente. Para el conjunto de validación, la precisión disminuye en un 1% y el kappa en un 3%.

2.2. Clasificación Multi-estados

2.2.1. Estados Fenológicos Agrupados (8 estados)

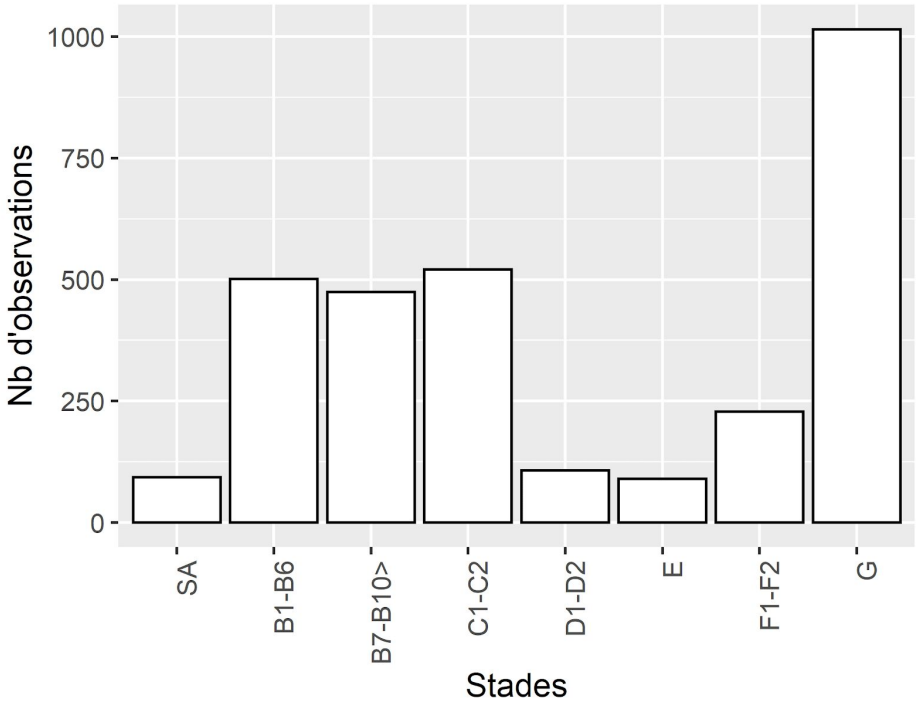


Fig. 12 Distribución de las observaciones para los estados fenológicos agrupados

El dataset se encuentra desbalanceado y el número de observaciones para cada estado difiere de manera observable (fig. 12). Sin embargo, los estados poco representativos (SA, D1-D2 y E) tienen más de 90 observaciones. Estados como el B1-B6, el B7-B10> y el C1-C2 son más homogéneos con cerca de 500 observaciones. Para el estado F1-F2 hay cerca de 250 observaciones. Finalmente para el estado final G, el número de observaciones aumenta. Debido a que la observación de los estados es extraída de **Vigicultures®**, podríamos considerar que el elevado número de observaciones para los estados finales de desarrollo del colza sea debido a la presencia de más bioagresores en

ese periodo fenológico. Es por lo anterior, que la identificación de estos estados tiene mayor importancia para nuestra problemática.

2.2.2. Modelos Estatísticos (Comparación de Métodos de clasificación)

Nos preguntamos si uno de los cuatro métodos de clasificación seleccionados podría ser más preciso al momento de predecir los estados fenológicos de las observaciones in-situ. Para ello, construimos un dataset basado en el modelo de referencia (ver fig. 6). Obtuvimos un juego de datos compuesto por 3029 observaciones y 428 variables. Este dataset se descompondrá aleatoriamente en dos datasets conformados por muestras diferentes. Un conjunto de entrenamiento compuesto por el 70% del dataset inicial y el otro 30% de las observaciones constituye el conjunto de prueba. Las figuras 13, 14, 16, 18 y 20 ilustran las matrices de confusión obtenidas en el dataset de prueba para cada uno de los clasificadores.

Lasso - Multinomial

Utilizamos Lasso en su modalidad multinomial y los resultados para el conjunto de prueba se presentan en la figura 13. La matriz de confusión, nos muestra una *accuracy* global del 85%. Observamos que las clases mejor clasificadas por el modelo son las clases B1-B6 (76.0%), B7-B10> (85.21%), C1-C2 (96.79%) y G (96.71%). Los errores entre clases siempre se dan por la vecindad entre estas (el estado anterior o la posterior al estado que se quiere predecir), a excepción de una observación clasificada como F1-F2, siendo su verdadera clase D1-D2. Observamos también que a menor número de observaciones, las clases vecinas tienen una mayor tendencia a ser confundidas. La clase SA se confunde con la B1-B6 y en el caso de la D1-D2 el modelo la predice como C1-C2.

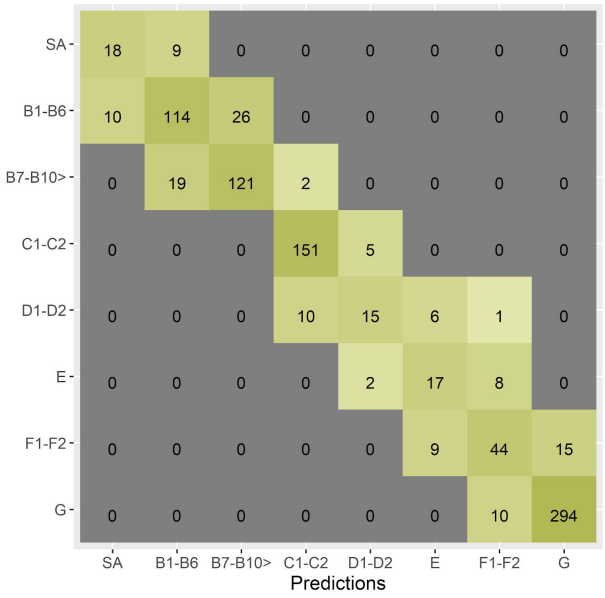


Fig. 13 Matriz de Confusión para el clasificador Lasso- Multinomial en el conjunto de prueba

LASSO nos permite a partir del coeficiente de *lambda1se* determinar el número de variables que explican el modelo sin sobre-ajustarlo (selección de variables), para nuestro caso de estudio, la figura 14 nos muestra como la precipitación (de las primeras semanas) y la humedad relativa (de las últimas semanas) son las variables más representativas. Sin embargo, índices espectrales como el *GNDVI*, el *MSI* y el *EVI* están presentes. Estos índices relacionados a la presencia de humedad y al contenido clorofílico de la planta nos permiten concluir que la reacción de la planta a condiciones hídricas determinadas, definen adecuadamente el estado fenológico de la misma.

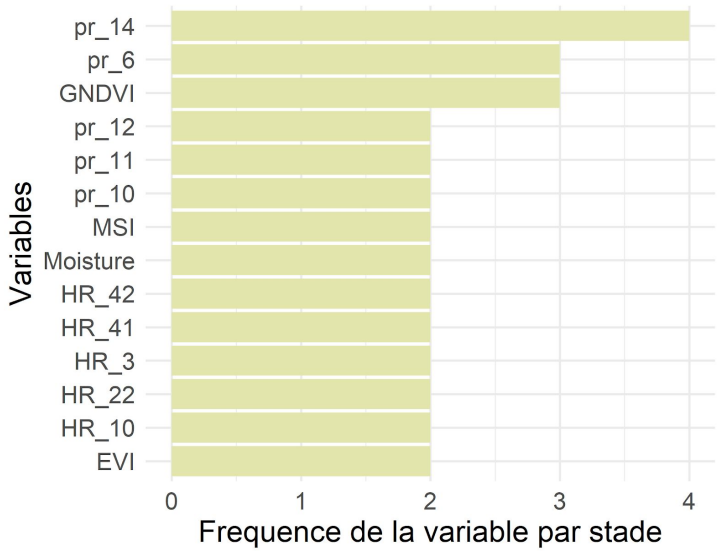


Fig. 14 Importancia de las variables explicativas del modelo Lasso-Multinomial

Ordinal Logistic Regression

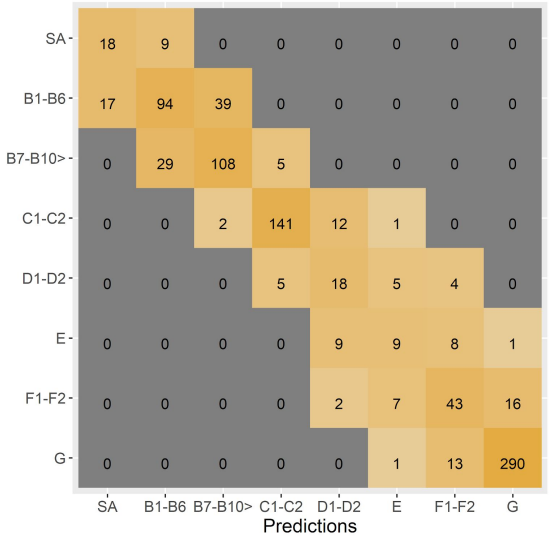


Fig. 15 Matriz de Confusión para el clasificador OLR en el conjunto de prueba

Decidimos clasificar los estados a partir de un modelo ordinal dado el carácter secuencial de los estado fenológicos (un estado precede al otro). Los resultados de la matriz de confusión son cercanos al modelo Lasso-multinomial. La *accuracy* por clase disminuye para las clases mejor predichas por el clasificador anterior. En este modelo observamos las siguientes *accuracy* B1-B6 (62.67%), B7-B10> (76.06%), C1-C2 (90.38%) y G (95.39%). Sin embargo, en estados con poco número de observaciones (SA, D1-D2 y E), el modelo confunde las clases vecinas. El modelo presenta más errores entre clases distantes que el modelo anterior, pues clasifica observaciones de estados con una distancia interclase de dos (D1-D2 como F1-F2, por ejemplo). Concluimos que ordenar las categorías fenológicas, disminuye la *accuracy* del modelo general (79%) en comparación a los resultado obtenidos por el modelo Lasso.

Multinomial Logistic Regression - Réseaux de neurones

El análisis de la matriz de confusión (fig. 16), nos muestra diferencias con relación a la modelización lasso-multinomial en el reconocimiento de los estados analizados uno por uno, a pesar de que la *accuracy* global sigue siendo aceptable (83%). Observamos que las clases mejor clasificadas por LASSO disminuyen al utilizar redes neuronales (B1-B6 (68.67%), B7-B10> (83.10%), C1-C2 (95.51%) y G (96.71%)). El modelo tiene tendencia a confundir con más facilidad las clases, aun cuando no son vecinas. Esta dificultad hace que, a pesar de tener una buena *accuracy*, el modelo sea menos eficiente que LASSO. Concluimos que en esta clasificación con datos desbalanceados, las redes neuronales ajustan el modelo de forma cercana a los modelos lineales usados precedentemente.

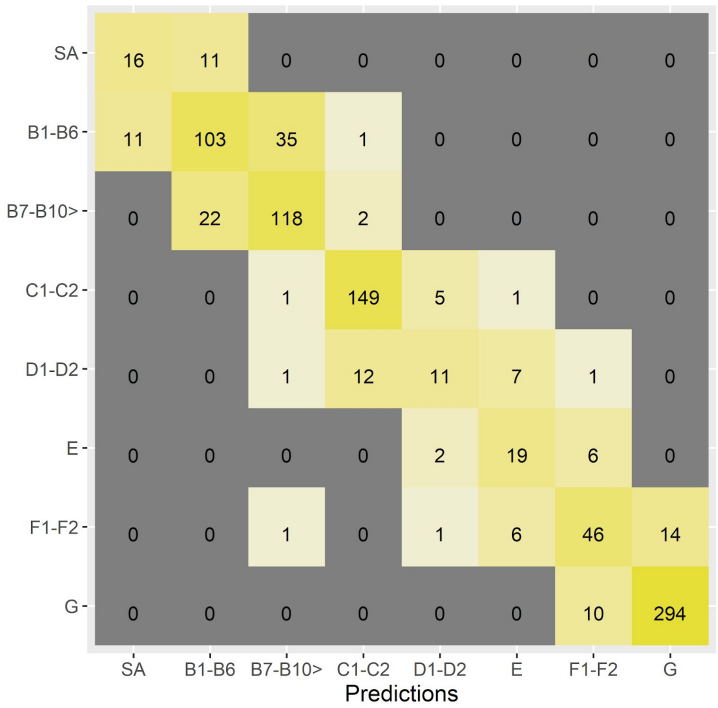


Fig. 16 Matriz de Confusión clasificador MLR- Lasso para el conjunto de Prueba

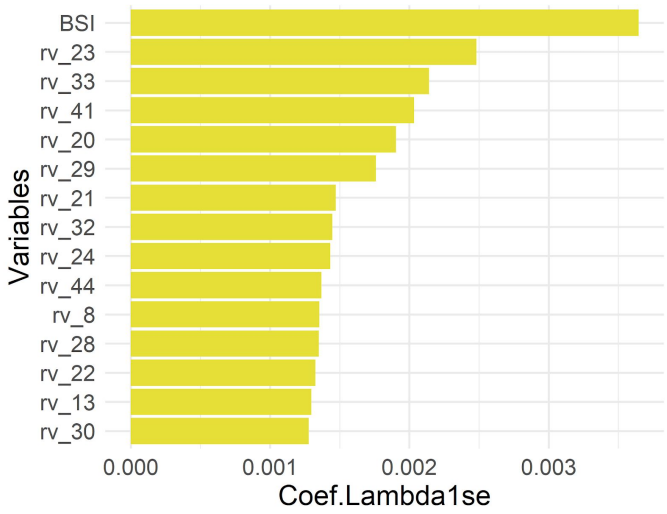


Fig. 17 Importancia de las variables explicativas del modelo MLR

En la fig. 17 podemos observar que las variables seleccionadas por el modelo para clasificar los estados fenológicos, son el índice *BSI* (*Bare Soil Index*) donde las bandas B2, B4, B8 y B11 están implicadas, así como la variable climática rayonnement Solaire a la mitad del año precedente a la fecha de observación in-situ del estado. Podríamos concluir que el modelo clasifica basado en condiciones de ausencia y/o presencia de vegetación (Índice *BSI*) y a la respuesta espectral del colza a la intensidad de la radiación solar.

Random Forest

Utilizamos un clasificador no lineal para determinar si este método representaba una mejora de la precisión en la predicción de los estados. Los resultados se presentan abajo en la matriz de confusión.

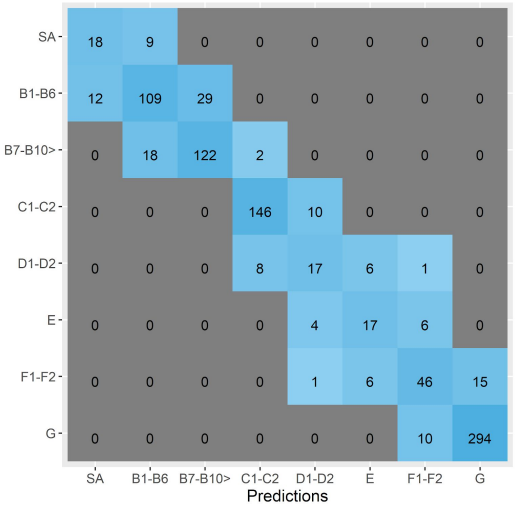


Fig. 18 Matriz de Confusión para el clasificador *Random Forest* en el conjunto de prueba

Observamos que los resultados son comparables a los métodos lineales ajustados anteriormente. La similitud en la clasificación con los modelos multinomiales es la más cercana. Con una *accuracy* general de 84%, identificamos que para las clases donde las observaciones son pocas, el clasificador continúa confundiendo la clase objetivo con sus vecinas (SA, D1-D2, E y F1-F2). Los estados D1-D2 y F1-F2 fueron clasificados fuera de las clases inmediatamente vecinas(a una distancia de dos clases). Para las clases mejor identificadas, los resultados continúan siendo adecuados. Para el estado B1-B6 (74.00%), B7-B10> (87.32%), C1-C2 (93.59%) y G (96.71%). Concluimos que el tipo de acercamiento (lineal o no lineal), no afecta drásticamente los resultados de la clasificación, sin embargo LASSO presenta una *accuracy* mejor (entre un 6% y un 1%) que los otros modelos.

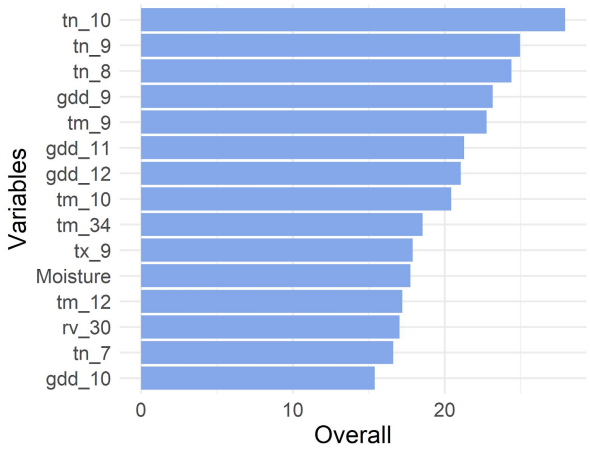


Fig. 19 Importancia de las variables explicativas del modelo RF

Cuando observamos la importancia de las variables que selecciona el modelo, podemos concluir que son las variables climáticas las que determinan la clasificación de una observación en un estado u otro, siendo las más relevantes en el método de *Random Forest*, las temperaturas (mínima y media), los grados días de desarrollo (gdd) del final del primer trimestre y la radiación solar de la mitad del año precedente a la observación in-situ. Los índices espectrales que evalúan la humedad del suelo y el estrés hídrico de la planta complementan el modelo.

k-Nearest Neighbors (kNN)

Ajustamos un modelo no paramétrico basado en distancias (euclidianas), para determinar si la precisión de los resultados de este acercamiento es comparable con los modelos anteriores.
En la matriz de confusión para este modelo (fig. 20) continuamos observando resultados cercanos a los anteriores. Con una precisión general cercana a Lasso-Multinomial (83.4%), una sola observación clasificada a más de una clase de distancia (D1-D2) y una clasificación muy acertada en los estados B1-B6 (71.33%), B7-B10 (82.39%), C1-C2 (94.87%) y G (95.72%) es un método interesante para la identificación de los estados fenológicos. Los estados con pocas observaciones continúan teniendo un número

importante de falsos positivos sin embargo el modelo las clasifica bien. Concluimos que con la elección de un clasificador simple se obtienen resultados similares a modelos más parametrizados sin embargo limitar la decisión de clasificación a la distancia podría ser poco reproducible y explicativo.

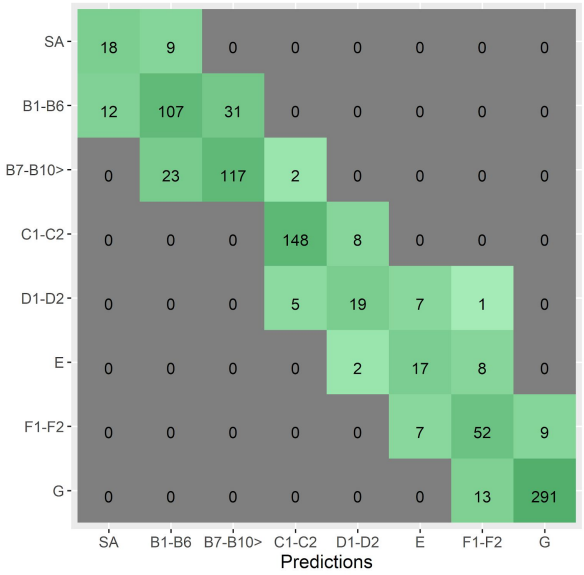


Fig. 20 Matriz de Confusión para el clasificador *K-Nearest Neighbors* en el conjunto de prueba

Finalmente, podemos concluir que al ajustar cinco modelos, cada uno con un acercamiento diferente, los clasificadores convergen en resultados cercanos. Las clases mejor clasificadas fueron la clase C1-C2 y la clase G. Estas clases presentan un buen número de observaciones in-situ y patrones climáticos y/o espectrales que permiten clasificarlos con facilidad sin obtener errores representativos. Observamos también que las etapas confundidas por nuestros modelos podrían tener similitudes en las variables climáticas y en su comportamiento espectral.

En la siguiente tabla, observamos a modo de resumen los cinco clasificadores evaluados. Si los valores de *accuracy* son similares con un máximo para el Lasso - Multinomial, los tiempos de cálculo para este modelo más complejo son también mucho más importantes que para los otros. Por otro lado, el modelo de *Random Forest* tiene el tiempo de estimación más bajo, a la vez que conserva una excelente capacidad de predicción que nos lleva a confirmar nuestra elección del modelo de *Random Forest* para la continuación de nuestras investigaciones sobre la importancia de las diversas variables explicativas.

Tabla 9. *Accuracy* y tiempo de cálculo para los modelos evaluados

Método	***Accuracy***	**Tiempo de Cálculo (s)**
Lasso - Multinomial	85.4%	1200
Multinomial Logistic Regression - Redes Neuronales	83.4%	26
Ordinal Logistic Regression	79.6%	60
Random Forest	84.2%	18
k-Nearest Neighbors	83.4%	60

2.3. Comparación de la capacidades predictivas aportadas por diferentes tipos de datos

2.3.1. Pretratamientos de las Bandas Espectrales (por extracción iota2)

En el ejercicio de determinar las variables más significativas para predecir el cambio de estado en el colza, se quiso identificar si la clasificación a partir de diferentes transformaciones de la información espectral (Bandas, Índices y *Tasseled Cap*) podrían mejorar el modelo de referencia. Comparamos la *Accuracy* y el error fuera de bolsa (OOB) de cada una de las transformaciones espectrales, así como el porcentaje de éxito de clasificación por clase en los datos de entrenamiento.

Fig. 21 *Accuracy* y OOB de cada modelo espectrales(conjunto de Entrenamiento)

En la fig. 21, para el caso de *Tasseled Cap* un OOB del 0.42, una precisión del 0.68 y un kappa de 0.59 lo posiciona como el menos performante. Para las bandas y los índices espectrales, las métricas de evaluación son cercanas. Con un OOB de 0.33, una precisión y un kappa de 0.70 y 0.60 respectivamente, la elección entre los índices y las bandas se reduce a efectos prácticos, como la facilidad de interpretación en el caso de los índices

donde la simplicidad de aplicación en el caso de la bandas. Por lo anterior, se podría argumentar que es mejor utilizar en orden descendente, las bandas, los índices y la metodología *Tasseled cap* para estudiar los estados del colza.

Al comparar la *accuracy* de los modelos espectrales con el modelo de referencia (Baseline), observamos una diferencia del 16% para bandas y del 26% para *Tasseled Cap*. Las variables espectrales clasifican cerca del 70% de las observaciones, sin embargo el modelo de referencia (índices + clima) continúa siendo el mejor clasificador (84%). La adjunción de los datos climáticos a los datos espectrales aporta bastante información.

Observando la siguiente tabla podemos apreciar que el porcentaje de acierto de los modelos espectrales para cada estado, en el conjunto de entrenamiento.

Tabla 10. Porcentaje de aciertos de cada modelo por cada estado Fenológico

Modelo	**SA**	**B1-B6**	**B7-B10**	**C1-C2**	**D1-D2**	**E**	**F1-F2**	**G**
Bandas	55.56%	67.01%	54.58%	62.19%	10.00%	29.17%	42.94%	85.69%
Índices	28.79%	71.79%	50.30%	76.99%	0.00%	15.87%	41.25%	89.73%
TassCap	16.67%	73.50%	23.80%	60.55%	1.33%	12.70%	36.25%	82.70%
BaseLine	54.55%	77.21%	82.53%	95.62%	49.33%	49.21%	71.88%	95.36%

Codigo de colores: Amarillo mejores clasificaciones. Azul segundas mejores clasificaciones

El modelo de referencia ofrece una mejor clasificación para todos los estados fenológicos exceptuando el primero. El segundo lugar es alcanzado por el modelo que considera las 10 bandas espectrales. El modelo de los índices espectrales sigue de cerca al de las bandas sin embargo en estados donde el número de observaciones es bajo, tiende a confundir los estados objectivos con las clases vecinas. El modelo *Tasseled Cap* solamente supera a los dos anteriores en el estado B1-B6. Es posible clasificar los estados exclusivamente a partir de información espectral pero es importante considerar el aporte de otras variables para afinar la clasificación.

2.3.2. Focalización en imágenes recientes (iota2-inrae) - Metodologías de extracción

Para analizar el impacto de la metodología de extracción de la información espectral, realizamos una clasificación a partir de los índices espectrales para ambos juegos de datos (iota2 y inrae).

La *accuracy* de las clasificaciones a partir de la metodología *iota2* es del 0.68 frente a la metodología inrae que es del 0.62. En cuanto al OOB, *iota2* identifica las clases con una reducción del 5% en comparación a inrae.

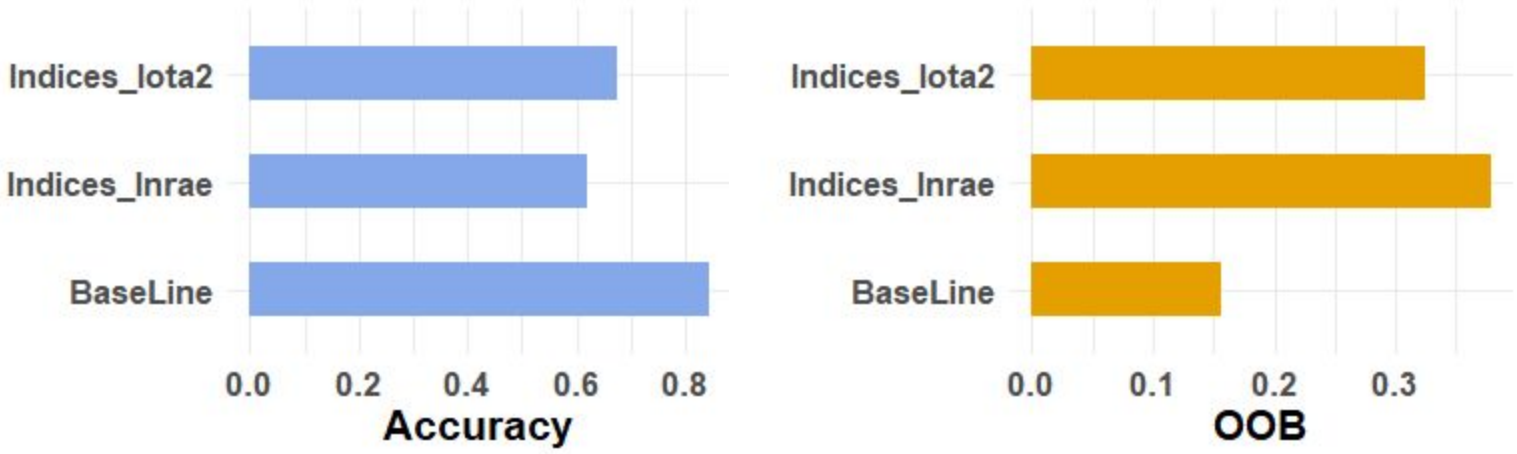


Fig. 22 *Accuracy* y OOB de cada metodología de extracción (conjunto de entrenamiento)

Cuando comparamos los dos clasificadores a partir del porcentaje de éxito por clase, observamos que *iota2* es el mejor. Las clases B1-B6, B7-B10, C1-C2 y G que presentan un número de observaciones importante (conjunto de entrenamiento: 351, 332, 365 y 711 observaciones respectivamente), son los estados mejor predichos por el modelo. En ambos casos el modelo no encuentra un patrón para clasificar el estado D1-D2.

Para ambos juegos de datos el estado D1-D2 es confundido por el estado C1-C2 (40% de las observaciones son clasificadas en la clase precedente). En la escala BBCH ambos estados corresponde al desarrollo de hojas (roseta) y de órganos vegetativos que al ser tan cercanos, resulta difícil diferenciar solamente con información espectral.

Tabla 11. Porcentaje de aciertos de cada modelo por cada estado Fenológico

Modelo	**SA**	**B1-B6**	**B7-B10**	**C1-C2**	**D1-D2**	**E**	**F1-F2**	**G**
Indices_Iota2	28.79%	71.79%	50.30%	76.99%	0.00%	15.87%	41.25%	89.73%
Índices_Inrae	31.82%	64.67%	36.75%	72.33%	0.00%	14.29%	38.75%	86.36%
BaseLine	54.55%	77.21%	82.53%	95.62%	49.33%	49.21%	71.88%	95.36%

Codigo de colores: Amarillo mejores clasificaciones. Azul segundas mejores clasificaciones

Podemos concluir que el uso de la cadena de tratamiento iota2, presenta mejores resultados en el estado actual de la cadena Inrae.

2.3.3. Variables climáticas vs. Variables Espacio-Temporales

Para determinar si las variables climáticas son más predictivas que las variables espacio temporales, comparamos ambos modelos teniendo las condiciones de referencia como base.

Las gráficas muestran resultados muy cercanos. Con una *accuracy* de 0.81 y un kappa de 0.77 para las variables espacio-temporales frente a una *accuracy* de 0.82 y 0.78 para las variables climáticas, la diferencia principal es el OOB siendo ligeramente mayor para el clasificador *Date_Dep* (0.19 vs 0.18).

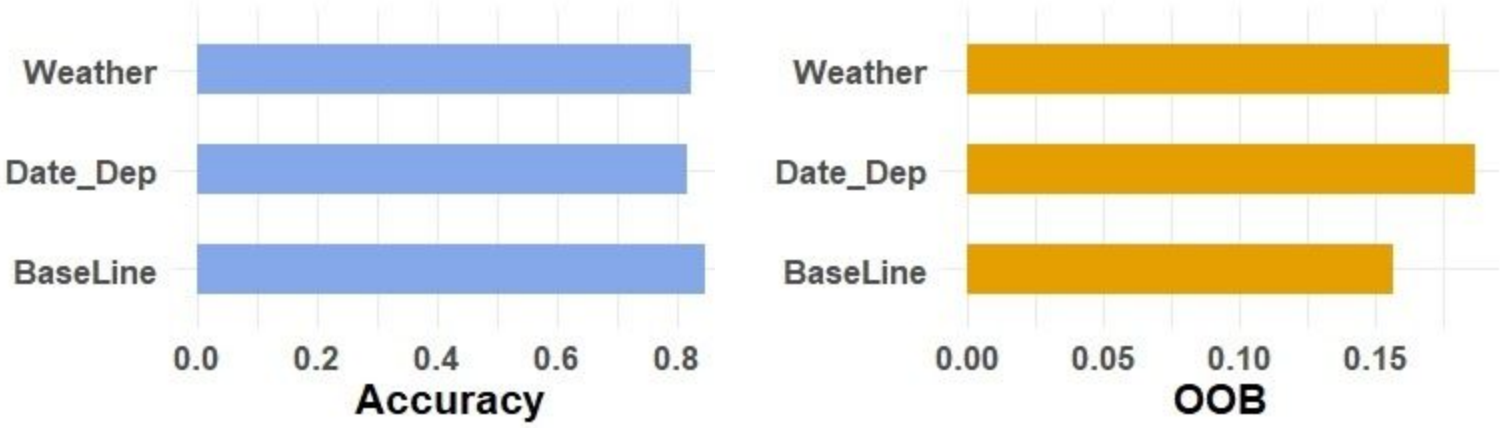


Fig 23. *Accuracy* y OOB modelos basados en variables climatológicas y espacio-temporales

A nivel de *accuracy* por clase, los estados B1-B6, B7-B10, C1-C2 y G presentan las mejores predicciones.

Tabla 12. Porcentaje de aciertos de cada modelo por cada estado Fenológico

Modelo	**SA**	**B1-B6**	**B7-B10**	**C1-C2**	**D1-D2**	**E**	**F1-F2**	**G**
Date_Dep	40.91%	76.64%	74.70%	97.81%	38.67%	31.75%	54.37%	97.61%
Weather	51.52%	72.36%	74.70%	95.89%	48.00%	46.03%	71.45%	95.64%
BaseLine	54.55%	77.21%	82.53%	95.62%	49.33%	49.21%	71.88%	95.36%

Codigo de colores: Amarillo mejores clasificaciones. Azul segundas mejores clasificaciones

El análisis de variables climáticas y espacio temporales continúan siendo, en general, ligeramente menos eficaces que el modelo de referencia al momento de la predicción. Sin embargo estados como C1-C2 y G son mejor clasificados por las variables espacio-temporales. Concluimos que después del modelo de referencia, son las variables climáticas las que mejor clasifican los estados fenológicos para la colza pero la pérdida de precisión por no utilizar la información espectral es leve. De misma manera, el conjunto de fechas y departamentos brinda una *accuracy* comparable a la información

meteorológica, aunque para estados específicos e importantes como aquello de la floración, el uso de las variables climáticas implica una diferencia importante (54.37% - 71.88%).

2.3.4. Combinación de información de diferentes variables temáticas

Nos preguntamos si la combinación de diferentes variables temáticas en un solo modelo podría mejorar la clasificación de los estados fenológicos. Construimos combinaciones que combinaban dos variables temáticas y excluían la tercera (clima+espacio-tiempo y índices espectrales + espacio-tiempo), para finalmente combinar las tres (clima + indices + espacio-tiempo) y comparamos su desempeño con las métricas de la *accuracy* y el OOB.

La figura 24. nos muestra *accuracy* bastantes cercanas entre los diferentes modelos. Los modelos en los que utilizamos la información espacio-temporal acoplada con las variables espectrales obtuvimos una *accuracy* de 0.81, pero cuando acoplamos variables espacio-temporales con variables climáticas, la *accuracy* incrementa en un 1%. Por otro lado, al acoplar los tres conjuntos de variables temáticas en un solo modelo (WIDD: Clima + Índices espectrales + Date + departamento) obtuvimos una *accuracy* muy cercana a aquella del modelo de referencia pero con un error de clasificación mayor (0.156 vs 0.157).

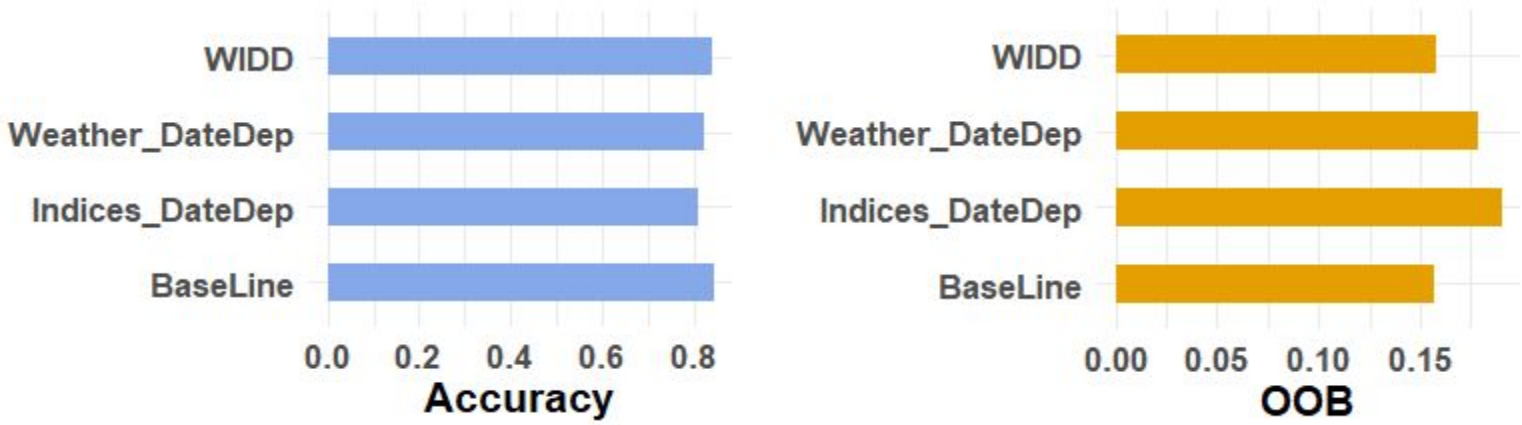


Fig 24. *Accuracy* y OOB modelos basados en combinación de variables espectrales, climatológicas y espacio-temporales

Finalmente al comparar el porcentaje de éxito en la clasificación por cada una de las clases observamos que en algunos estados fenológicos hay modelos que clasifican mejor o que los resultados son iguales al modelo de referencia. En clases como B7-B10 o F1-F2, el modelo *WIDD* predice las clases en el 82.53% y el 71,88% de los casos respectivamente. Por otro lado, en clases como C1-C2, D1-D2, E y G, los mejores resultados están distribuidos en los tres modelos. Podemos concluir que el acoplamiento entre las diferentes variables nos ofrece una mejora en la predicción de estados individuales pero que el modelo elegido como referencia sigue siendo un buen

modelo siendo para todas las etapas el mejor o el segundo mejor modelo (y sólo ligeramente).

Tableau 13. Porcentaje de aciertos de cada modelo por cada estado Fenológico

Modelo	**SA**	**B1-B6**	**B7-B10**	**C1-C2**	**D1-D2**	**E**	**F1-F2**	**G**
Weather_DateDep	53.03%	70.94%	74.40%	95.39%	49.13%	50.797%	71.36%	95.20%
Índices_DateDep	45.45%	76.07%	79.82%	96.44%	14.67%	25.40%	63.12%	96.22%
WIDD[12]	51.52%	76.92%	83.13%	96.34%	52.00%	47.62%	70.04%	95.14%
BaseLine	54.55%	77.21%	82.53%	95.62%	49.33%	49.21%	71.88%	95.36%

Codigo de colores: Amarillo mejores clasificaciones. Azul segundas mejores clasificaciones

Además, ya hemos visto que las variables climáticas son las que mayor peso tienen a la hora de clasificar los estados fenológicos. Concluimos que aunque los modelos anteriores ofrecen resultados cercanos al modelo de referencia, es este el más versátil para clasificaciones en las que se quiera predecir sin depender de variables de tiempo y espacio y así ampliar el espectro de utilización a otros lugares.

[12] WIDD = Weather + Índices +DateDep

2.5. Impacto de la agrupación y del remuestreo

2.4.1. Impacto de la agrupación de estados (26 estados)

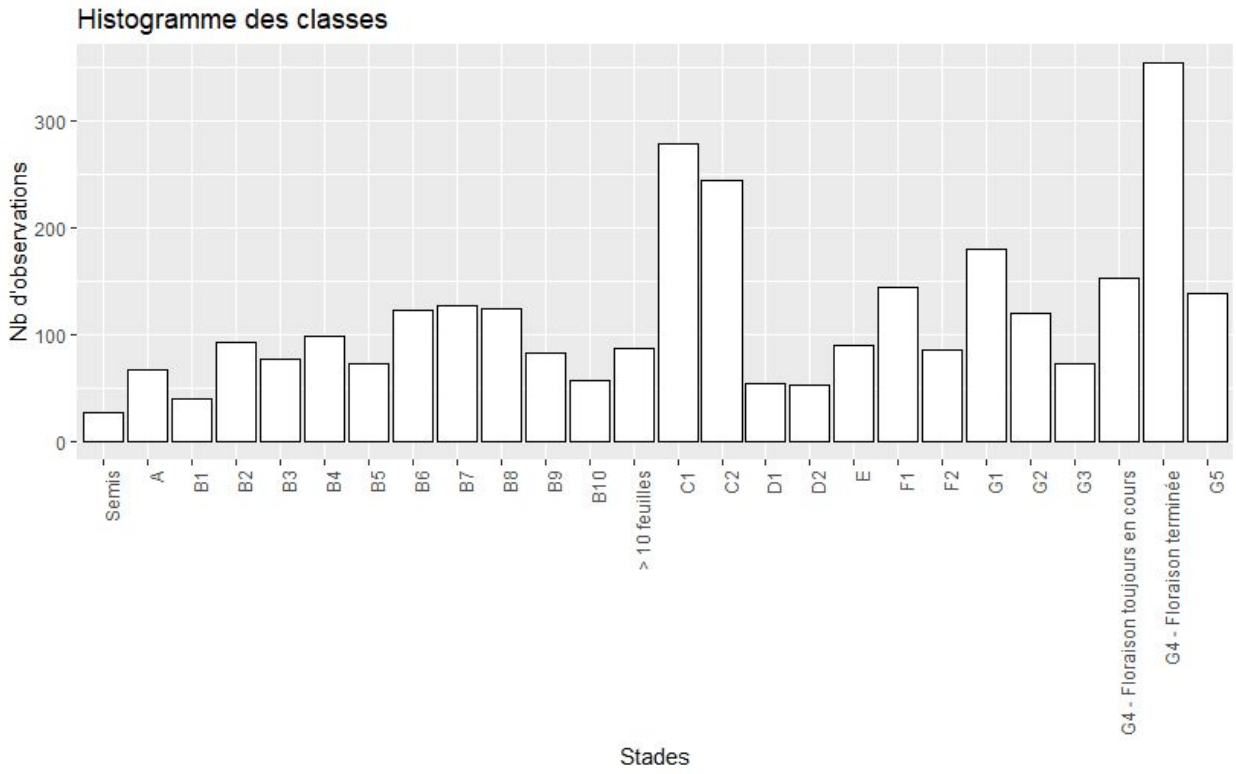


Fig 25. Distribución de las observaciones para los estados fenológicos no agrupados

Cuando observamos el número de observaciones de cada uno de los estados no agrupados, observamos una fuerte variabilidad. Estados fenológicos como el C1, el C2 el G1 y el G4 -Floraison terminée son los más representativos. Por su lado estados minoritarios como Semis, B1, D1 y D2 con un número de observaciones inferior a 50, presentan un gran desafío para los clasificadores utilizados. Nos preguntamos si un conjunto de datos fuertemente desbalanceado, podría ser bien clasificado utilizando las condiciones de referencia y el método de *Random Forest*.

La figura 26 nos muestra la comparación de las matrices de confusión para los estados agrupados y no agrupados. Observamos que para los estados iniciales (izquierda), el modelo confunde la clase objetivo con hasta 8 clases diferentes (estado B3), sin embargo estas 8 clases son todas consideradas como vecinas en el modelo de 8 clases y los errores son concentradas en gran medida en las clases más cercanas. A partir del estado C1 la cantidad de verdaderos positivos aumenta y la diferencia entre clases es mejor. La calidad de la clasificación, en relación con el modelo de 8 clases, puede incluso mejorarse de vez en cuando. Por ejemplo, el conjunto C1-C2 sólo ve 4 confusiones con D1-D2 en lugar de 10. El conjunto D1-D2 todavía tiene 8 confusiones con C1-C2 y aumenta de 6 a 9 sus confusiones con E pero no tiene ninguna confusión con el estado

más distante F1-F2. El estado E sólo admite confusiones con las subclases más cercanas (D2 y F1). El conjunto F1-F2 tampoco se confunde ya con el estado más lejano D1-D2. También es posible hacer distinciones claras dentro de las clases agrupadas en ciertos casos, como el muy amplio grupo de observaciones G: La oposición entre las tres primeras clases de G y las tres últimas es particularmente marcada. En general, la agrupación de los valores en torno a la diagonal es sorprendente y sugiere que una estimación a nivel de la clase inicial **Vigicultures®** seguiría siendo informativa, especialmente si se aumentara el número todavía pequeño de observaciones por clase. Sin embargo, las métricas de evaluación caen lógicamente con el aumento del número de clases. Con 26 estados fenológicos la *accuracy* general del modelo es inferior al 50% y el error de (OOB) es más que triplicada (0.15 vs 0.51).

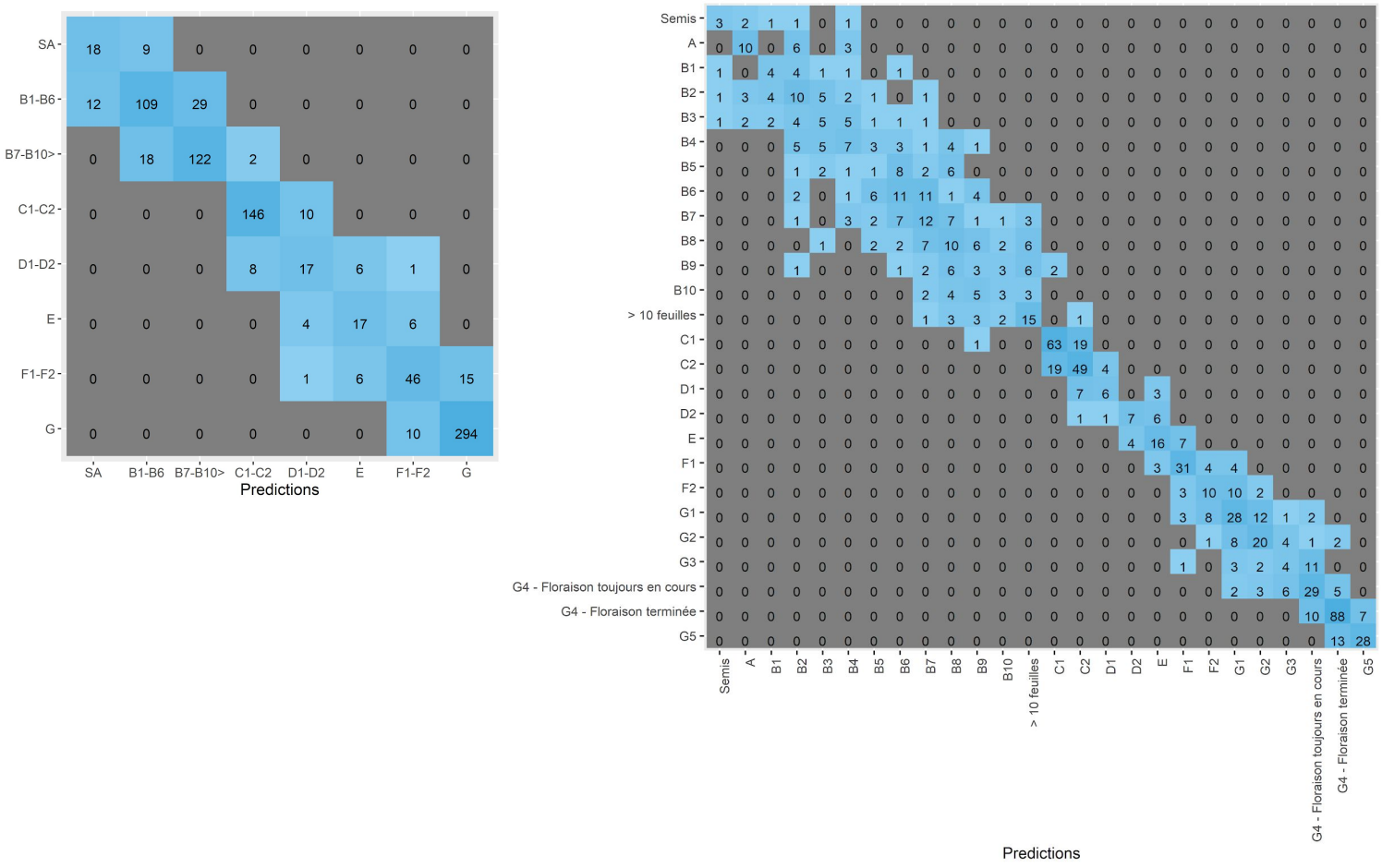


Fig 26. Matrices de confusión del modelo de referencia con estados fenológicos agrupados (izquierda) y los estados no-agrupados (derecha)

Concluimos que, la estrategia de agrupación nos ha permitido tener resultados sintéticos probablemente generalizables a una clasificación más fina que parece en parte posible en caso que se necesite.

2.4.2. Modelo de referencia con estrategia de remuestreo

Debido a la dificultad que encontramos para predecir las clases en las que el número de observaciones es considerablemente inferior, decidimos evaluar las condiciones de referencia en un conjunto de datos balanceado a partir de tres métodos de remuestreo. Inicialmente balanceamos el conjunto de datos realizando un proceso de *downsampling*. Para este método conservamos todos los casos de la clase minoritaria, y elegimos aleatoriamente una muestra con el mismo número de casos en las clases mayoritarias. Seguidamente, balanceamos los datos haciendo un upsampling en donde dejamos todas las instancias de la clase mayoritaria, y aumentamos el número de casos de las clases minoritaria muestreando con reemplazamiento. Finalmente, utilizamos la técnica *SMOTE*[13] que incluye *Upsampling y Downsampling* al mismo tiempo. Para mantener el uso de los conjuntos de entrenamiento/prueba lo aplicamos por separado a cada uno de los dos conjuntos.

Los resultados que nos muestra la figura 27 nos permite identificar que el mejor modelo es el balanceado a partir del método de upsampling. Con una *accuracy* de 0.98% para el conjunto usado para el entrenamiento del modelo, este modelo aumenta la performance de la clasificación de los estados fenológicos en un 14% en comparación al modelo de referencia. El error de clasificación se reduce drásticamente a un valor de 0.017 frente a un valor de 0.15 del modelo de referencia. La técnica híbrida *SMOTE* tiene una *accuracy* de 0.88 mejorando en un 4% la *accuracy* del modelo de referencia como también un OOB inferior (0.11 vs 0.15). Sin embargo utilizar la técnica de downsampling para balancear el conjunto datos reduciendo el número de observaciones desmejoran las predicciones.

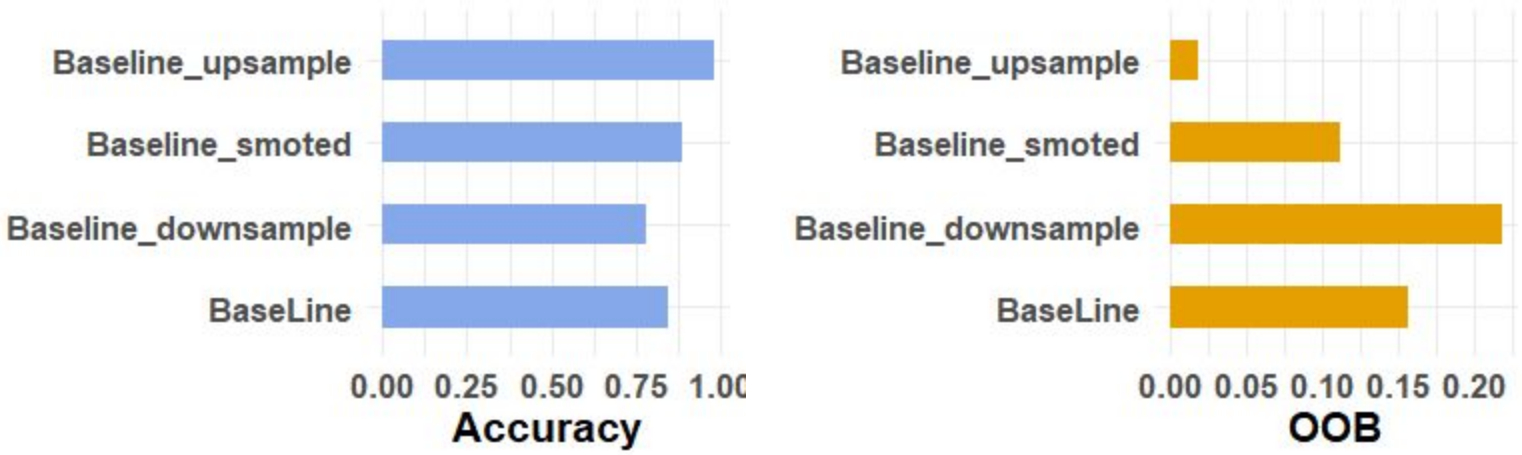


Fig 27. *Accuracy* y OOB del modelo de referencia con datos balanceados

Podríamos concluir que la utilización de métodos de remuestreo para balancear los datos mejora notoriamente la calidad de la clasificación con una inversión mínima en tiempo de cálculo. Sin embargo al evaluar el modelo en el conjunto de test encontramos una disminución en las métricas de evaluación. Para el conjunto de entrenamiento

[13] Synthetic Minority Oversampling Method

tenemos una *accuracy* de 0.98, un kappa de 0.98 y un OOB del 2%. Sin embargo, para el conjunto de test obtenemos una *accuracy* de 0.73 y un kappa de 0.69 que son inferiores a la estrategia sin remuestreo. La estrategia de remuestreo no parece entonces permitir *in fine* mejorar la predicción.

4. Discusión

La naturaleza subjetiva de las observaciones fenológicas terrestres siempre ha sido un problema en el estudio reciente de la fenología (Czernecki, Nowosad, et Jabłońska 2018). El desarrollo de métodos de clasificación para identificar patrones que ayuden a la toma de decisiones en el análisis del comportamiento de la vegetación de interés agrícola es el pilar de análisis de este problema de investigación.

Con el interés de determinar la importancia de las variables espectrales, climáticas y espacio-temporales en la identificación de los diferentes estados fenológicos de cultivos como el colza, evaluamos diferentes hipótesis. Inicialmente realizamos una clasificación binaria para el estado de floración en la que identificamos que índices espectrales como el *MSI*, el *NDYI* y el *NDWI* son elementos fundamentales para la clasificación de este estado. Las métricas de evaluación son adecuadas pero la incidencia en el desbalanceo de los datos dificulta la tarea de clasificación. Seguidamente, encontramos que al evaluar cinco métodos de clasificación los resultados son bastante cercanos. Esperábamos que el modelo *OLR* que tiene en cuenta el orden jerarquizado de las clases, fuera el más acertado ya que es el más próximo a la realidad (ordena las etiquetas en un orden de ocurrencia), pero métodos como el *Random Forest* se mostraron más performantes. Por otra parte, estudiando las diferentes posibilidades de agrupamiento entre variables temáticas, encontramos que las variables meteorológicas son decisivas para la clasificación y que en situaciones en las que las observaciones in-situ no estén disponibles o sean inconsistentes, un acoplamiento entre clima e índices espectrales permite predecir los estados fenológicos con un *accuracy* del 84% con muy pocos errores implicando clases muy distintas. Finalmente, el impacto del agrupamiento de las clases para mejorar el éxito de la clasificación, es una herramienta que permite priorizar los estados más importantes a estudiar. La utilización de técnicas de remuestreo de los datos mejora el *accuracy* aparente del modelo de referencia pero introducen un sobreajuste del modelo que resulta en una diferencia de *accuracy* entre el conjunto de entrenamiento y el conjunto de test cercana al 25%.

Los parámetros fenológicos de las imágenes satelitales multitemporales tienen el potencial de indicar el desarrollo del crecimiento de los cultivos en una gran región (Fisher et Mustard 2007; Zhong et al. 2011; Zhong, Gong, et Biging 2014; Li et al. 2014). Específicamente para la floración, al igual que d'Andrimont et al. (2020), identificamos que el índice *NDYI* captura el aumento en la coloración amarilla de las flores de colza en el banda espectral del verde(B3). El amarillo de los pétalos de colza se debe a su contenido en pigmentos carotenoides que absorben longitudes de onda de ~450 nm (Sulik et Long 2016). Las condiciones internas (la clorofila o la capacidad de retención de agua) y/o externas de humedad (suelo) también son importantes en la clasificación. Índices como el NDWI y el MSI en donde las bandas B8 (NIR), B8a(Red Edge 4) y B11(SWIR 1) están implicados. *Random Forest* se perfila como un algoritmo adecuado para clasificar los estados fenológicos de forma individual, sin embargo los análisis subsiguientes nos permitieron generalizar el modelo y aplicarlo a una clasificación multi-estados.

Para desarrollar una herramienta que ayude en el proceso de predicción de los estados fenológicos del colza, desarrollamos un enfoque basado en un modelo de referencia que selecciona los índices espectrales y las variables climáticas para la toma de decisiones. Utilizando ese modelo de referencia las comparaciones para determinar el mejor clasificador se hacen más fáciles. Los datos in-situ para nuestro caso de estudio son limitados, ya que obtener datos de campo fiables y exactos a una escala apropiada es un esfuerzo difícil (Fisher y Mustard 2007). El enfoque de clasificación basado en la clasificación de clases categóricas tiene ventajas cuando la disponibilidad de datos de realidad de terreno es limitada (Zhong et al. 2011).

Por lo anterior, la decisión de evaluar los clasificadores para comparar su rendimiento entre sí (benchmarking) permite establecer una referencia/orientación empírica para seleccionar los clasificadores más apropiados para problemas específicos (C. Zhang et al. 2017). Al igual que Lorena et al. (2011), encontramos que para estudios biogeográficos, RF es una técnica de modelización prometedora, debido a su alto rendimiento en conjuntos de datos compuestos por grandes números de variables diversamente independientes. Sin embargo, otros modelos de regresión multinomial basados en el Lasso o redes neuronales presentan *accuracy* considerablemente performantes. En cualquier caso, el compromiso entre eficiencia y velocidad justifica la elección por Inglada et al. (2016) del *Random Forest* como el algoritmo de referencia para la clasificación de la ocupación del suelo para *iota2* (Inglada et al. 2015).

Por otra parte, de acuerdo a Zhang, Friedl, et Schaaf (2009), se ha demostrado que los índices de vegetación proporcionan una mejor descripción del crecimiento de los cultivos y mejoran considerablemente la precisión de la clasificación de los mismos. Sin embargo, en nuestro caso, al analizar diferentes transformaciones de la información espectral, encontramos que las bandas espectrales individuales son variables igual de interesantes que los índices, en algoritmos como *Random Forest,* para diferenciar una clase de otra porque los árboles de decisión permiten combinaciones de variables que pueden llegar a ser igual de interesantes a las proporcionadas por los índices espectrales. Analizando la contribución de cada variable explicativa en el modelo final (modelo de referencia), encontramos una influencia pequeña/moderada, de la información espectral satelital. Por el contrario, las características meteorológicas son las más predictivas en el caso de las fases fenológicas otoñales,invernales y primaverales, lo que demuestra una relación con la temperatura de esos periodos. La influencia de la temperatura en el crecimiento de las plantas es definitivamente mayor en la primavera, cuando comienzan su ciclo de desarrollo después de la pausa de invierno (Pope et al. 2013; Springate et Kover 2014). Sin embargo, la información meteorológica también proporciona por su carácter cíclico una información sobre la temporalidad de la observación que es muy importante en la identificación del estadio adecuado como lo indica la eficiencia de los modelos con datos espacio-temporales.

Dentro de nuestro análisis, el proceso de clasificación no considera explícitamente la dimensión temporal (acercamiento del conjunto de datos como serie temporal). Sin

embargo, existen modelos que incorporan la fecha de observación como variable explicativa, que junto a información meteorológica podría dar resultados satisfactorios en términos de *accuracy* en estudios posteriores.

En cuanto a la metodología de extracción de la información espectral, cuando la escala espacial de análisis es elevada, la presencia de nubes y sombras en las imágenes satelitales son situaciones a considerar. Según Inglada et al. (2015) los datos interpolados tienden a reducir estos inconvenientes. Es por tal motivo, que los resultados obtenidos con la metodología *iota2* presentan mejores resultados que la metodología inrae en donde no consideramos el uso de la máscara de nubes propuesta por el producto theia. No obstante, la metodología *inrae* tiene la ventaja de poder ser utilizada en tiempo real mientras que el *iota2* aquí permite la interpolación utilizando la siguiente imagen aunque se tome mucho tiempo después.

Si bien en este estudio no realizamos un proceso de clustering estadístico, la decisión de agrupar estados basados en la experticia de equipo de investigación y la comparación con la escala BBCH, nos permitió obtener un modelo con una *accuracy* adecuada. Los resultados en las matrices de confusión nos muestran que a mayor número de clases (estados fenológicos) a predecir, la variabilidad y la presencia de datos atípicos aumenta, lo cual tiende a disminuir la eficiencia en la tarea de clasificación, pues las clases son confundidas entre sí con mayor frecuencia .

A pesar de las deficiencias de este enfoque y limitaciones que nos presentan los datos satelitales en la modelización de la fenología de las plantas, este acercamiento aún podría ser capaz de dar una aproximación fiable a las observaciones terrestres tradicionales, especialmente en lo que respecta a los finales del invierno (estados B7-B10> y C1-C2) y de la primavera (estados F1-F2 y G). Sin embargo, la tendencia a confundir estados vecinos es una variable que debe seguir siendo analizada en trabajos posteriores.

6. Límites y Dificultades

En el ejercicio de dar solución a la pregunta de la investigación, el tiempo es una de las variables más condicionantes, es por tal razón que considerar todas las posibilidades de acoplamiento entre variables temáticas no fue posible. Podría darse el caso de que una combinación no considerada se obtenga una precisión mejor a las obtenidas en nuestro análisis.

Por otro lado, de acuerdo al análisis bibliográfico, el estudio fenológico de los cultivos, en la mayoría de los casos es analizado como una serie temporal. Nosotros asumimos el riesgo de analizar la problemática desde un acercamiento diferente para establecer el potencial predictivo de las variables climáticas y espectrales sin considerar explícitamente la temporalidad del fenómeno. El acercamiento convencional fue abordado en el marco de otra pasantía en Toulouse en el CESBIO.

La calidad del conjunto de datos es una variable a considerar para futuros análisis, si bien **Vigicultures®** nos ofrece información fenológica su objetivo principal es la surveillance epidemiológica de los cultivos. Si bien estos datos no están orientados específicamente a la vigilancia de los estados y podrían ser imprecisos, ofrecen una oportunidad única de ajustar un modelo de predicción a un gran número de campos repartidos por toda Francia.

Hemos tenido dos tipos de dificultades durante la pasantía: las dificultades del proceso de investigación y las dificultades logísticas.

Las dificultades en el proceso de investigación están asociadas principalmente al procesamiento de la información espectral. En la segunda metodología de extracción (inrae, ver pág. 14), realizamos las correcciones atmosféricas pero no tuvimos en cuenta la máscara de nubes, el tiempo no permitió re-extraer y re-procesar los datos, sin embargo esta tarea es la más importante para poder comparar correctamente ambas metodologías (iota2/inrae) pues el interés de la metodología inrae es utilizar la imagen satelital más reciente. Este acercamiento en el que reducimos el número de observaciones podría ser presentar otra dificultad limitando la disponibilidad de los datos satelitales , sin embargo podría mejorar la pertinencia de la información usada.

En cuanto a las dificultades logísticas, iniciar un proceso de aprendizaje aplicado en una situación de crisis sanitaria mundial (COVID-19) dificulta la tarea en múltiples aspectos, siendo el más relevante el proceso administrativo en el marco de un confinamiento generalizado. sin embargo la buena comunicación y el esfuerzo mancomunado ayudaron a la solución de las dificultades en los tiempos adecuados. Por otra parte, esta situación atípica permitió ajustar los recursos internos de cada una de las partes para que el trabajo a distancia se vuelva una estrategia eficiente para el aprendizaje. Otra dificultad logística estuvo asociada al daño (corrupción material) del disco duro donde estaba almacenada todas las imágenes satelitales, lo que retrasó una

semana los procesos siguientes, sin embargo esto permitió probar y confirmar la eficiencia de la cadena de tratamiento.

Conclusión

El interés de nuestro acercamiento de clasificación radica en que una vez clasificados los estados fenológicos a partir del modelo de referencia, seamos capaces de establecer relaciones entre la cantidad de bioagresores y un estado fenológico determinado, lo que puede ayudar a la identificación de consecuencias en el rendimiento final de los cultivos. Aunque el estudio no alcanza a determinar el impacto de dichas relaciones, los resultados obtenidos establecen un primer paso importante para continuar construyendo conocimiento en esta área.

El estudio y el análisis de los resultados obtenidos nos permitieron proponer un modelo performante basado en el algoritmo de *Random Forest* para la clasificación de los estados fenológicos del colza a partir de variables meteorológicas y espectrales. Aunque el modelo no pueda completamente sustituir las observaciones *in situ*, sí puede ayudar en el proceso de adopción de decisiones y disminuir la dependencia al trabajo de campo para la obtención de información fenológica, especialmente cuando en los datos de archivo sobre bioagresores y rendimiento no se identifican las fechas de los cambios de fase fenológica sino que se dispone de imágenes de satélite.

Las perspectivas futuras de este trabajo se formulan desde tres frentes. En primer lugar, analizar el efecto de un agrupamiento aleatorio de los estados fenológicos sin considerar la agrupación de estados **Vigicultures®** a partir de la clasificación BBCH. La idea de realizar un proceso de clusterización automática (clasificación no supervisada) para agrupar las clases en las que el modelo tiene mayor dificultad en diferenciar podría ser un camino interesante a explorar. En segundo lugar, cuando existen clases excesivamente representadas se corre el riesgo de que el modelo aprenda demasiado en detrimento de las clases menos representadas. Para evitar este problema, se propone profundizar en la construcción de un conjunto de datos equilibrado donde la composición en cada estado sea casi idéntica. Finalmente, se propone ajustar un modelo que integre varios submodelos para cada estado fenológico y se evalúen sus resultados con los obtenidos hasta el momento.

Bibliographie

Ahl, Douglas E., Stith T. Gower, Sean N. Burrows, Nikolay V. Shabanov, Ranga B. Myneni, et Yuri Knyazikhin. 2006. « Monitoring Spring Canopy Phenology of a Deciduous Broadleaf Forest Using MODIS ». *Remote Sensing of Environment* 104 (1): 88‑95. https://doi.org/10.1016/j.rse.2006.05.003.

Almeida, Jurandy, Jefersson A. dos Santos, Bruna Alberton, Ricardo da S. Torres, et Leonor Patricia C. Morellato. 2014. « Applying Machine Learning Based on Multiscale Classifiers to Detect Remote Phenology Patterns in Cerrado Savanna Trees ». *Ecological Informatics*, Special Issue on Multimedia in Ecology and Environment, 23 (septembre): 49‑61. https://doi.org/10.1016/j.ecoinf.2013.06.011.

Ananth, C. V., et D. G. Kleinbaum. 1997. « Regression Models for Ordinal Responses: A Review of Methods and Applications. » *International Journal of Epidemiology* 26 (6): 1323‑33. https://doi.org/10.1093/ije/26.6.1323.

Andrimont, Raphaël d', Matthieu Taymans, Guido Lemoine, Andrej Ceglar, Momchil Yordanov, et Marijn van der Velde. 2020. « Detecting Flowering Phenology in Oil Seed Rape Parcels with Sentinel-1 and -2 Time Series ». *Remote Sensing of Environment* 239 (mars): 111660. https://doi.org/10.1016/j.rse.2020.111660.

Baetens, Louis, Camille Desjardins, et Olivier Hagolle. 2019. « Validation of Copernicus Sentinel-2 Cloud Masks Obtained from MAJA, Sen2Cor, and FMask Processors Using Reference Cloud Masks Generated with a Supervised Active Learning Procedure ». *Remote Sensing* 11 (4): 433. https://doi.org/10.3390/rs11040433.

Baskerville, G. L., et P. Emin. 1969. « Rapid Estimation of Heat Accumulation from Maximum and Minimum Temperatures ». *Ecology* 50 (3): 514‑17. https://doi.org/10.2307/1933912.

Belgiu, Mariana, et Lucian Drăguţ. 2016. « Random Forest in Remote Sensing: A Review of Applications and Future Directions ». *ISPRS Journal of Photogrammetry and Remote Sensing* 114 (avril): 24‑31. https://doi.org/10.1016/j.isprsjprs.2016.01.011.

Berra, Elias Fernando, Rachel Gaulton, et Stuart Barr. 2019. « Assessing Spring Phenology of a Temperate Woodland: A Multiscale Comparison of Ground, Unmanned Aerial Vehicle and Landsat Satellite Observations ». *Remote Sensing of Environment* 223 (mars): 229‑42. https://doi.org/10.1016/j.rse.2019.01.010.

Beurs, Kirsten M. De, et Geoffrey M. Henebry. 2005. « Land Surface Phenology and Temperature Variation in the International Geosphere–Biosphere Program High-Latitude Transects ». *Global Change Biology* 11 (5): 779‑90. https://doi.org/10.1111/j.1365-2486.2005.00949.x.

Bolton, Douglas K., et Mark A. Friedl. 2013. « Forecasting Crop Yield Using Remotely Sensed Vegetation Indices and Crop Phenology Metrics ». *Agricultural and Forest Meteorology* 173 (mai): 74‑84. https://doi.org/10.1016/j.agrformet.2013.01.007.

Boulesteix, Anne-Laure, Silke Janitza, Jochen Kruppa, et Inke R. König. 2012. « Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics ». *WIREs Data Mining and Knowledge Discovery* 2 (6): 493‑507. https://doi.org/10.1002/widm.1072.

Breiman, Leo. 2001. « Random Forests ». *Machine Learning* 45 (1): 5‑32. https://doi.org/10.1023/A:1010933404324.

Brown, Jesslyn F., Brian D. Wardlow, Tsegaye Tadesse, Michael J. Hayes, et Bradley C. Reed. 2008. « The Vegetation Drought Response Index (VegDRI): A New Integrated Approach for Monitoring Drought Stress in Vegetation ». *GIScience & Remote Sensing* 45 (1): 16‑46.

https://doi.org/10.2747/1548-1603.45.1.16.

Czernecki, Bartosz, Jakub Nowosad, et Katarzyna Jabłońska. 2018. « Machine Learning Modeling of Plant Phenology Based on Coupling Satellite and Gridded Meteorological Dataset ». *International Journal of Biometeorology* 62 (7): 1297‑1309. https://doi.org/10.1007/s00484-018-1534-2.

Deng, Zhenyun, Xiaoshu Zhu, Debo Cheng, Ming Zong, et Shichao Zhang. 2016. « Efficient KNN Classification Algorithm for Big Data ». *Neurocomputing*, Learning for Medical Imaging, 195 (juin): 143‑48. https://doi.org/10.1016/j.neucom.2015.08.112.

Efendi, Achmad, et Hafidz Wahyu Ramadhan. 2018. « Parameter Estimation of Multinomial Logistic Regression Model Using Least Absolute Shrinkage and Selection Operator (LASSO) ». In , 060002. East Java, Indonesia. https://doi.org/10.1063/1.5062766.

Fauvel, Mathieu, Mailys Lopes, Titouan Dubo, Justine Rivers-Moore, Pierre-Louis Frison, Nicolas Gross, et Annie Ouin. 2020. « Prediction of Plant Diversity in Grasslands Using Sentinel-1 and -2 Satellite Image Time Series ». *Remote Sensing of Environment* 237 (février): 111536. https://doi.org/10.1016/j.rse.2019.111536.

Fisher, Jeremy I., et John F. Mustard. 2007. « Cross-Scalar Satellite Phenology from Ground, Landsat, and MODIS Data ». *Remote Sensing of Environment* 109 (3): 261‑73. https://doi.org/10.1016/j.rse.2007.01.004.

Gao, Bo-cai. 1996. « NDWI—A Normalized Difference Water Index for Remote Sensing of Vegetation Liquid Water from Space ». *Remote Sensing of Environment* 58 (3): 257‑66. https://doi.org/10.1016/S0034-4257(96)00067-3.

Gitelson, Anatoly A., Yoram J. Kaufman, et Mark N. Merzlyak. 1996. « Use of a Green Channel in Remote Sensing of Global Vegetation from EOS-MODIS ». *Remote Sensing of Environment* 58 (3): 289‑98. https://doi.org/10.1016/S0034-4257(96)00072-7.

Hagolle, Olivier. (2016) 2020. *olivierhagolle/theia_download*. Python. https://github.com/olivierhagolle/theia_download.

Han, Qifei, Tiejun Wang, Yanbin Jiang, Richard Fischer, et Chaofan Li. 2018. « Phenological Variation Decreased Carbon Uptake in European Forests during 1999–2013 ». *Forest Ecology and Management* 427 (novembre): 45‑51. https://doi.org/10.1016/j.foreco.2018.05.062.

Hastie, Trevor, Sami Tibshirani, et Harry Friedman. 2009. *Elements of Statistical Learning Ed. 2*. Springer.

Heumann, B. W., J. W. Seaquist, L. Eklundh, et P. Jönsson. 2007. « AVHRR Derived Phenological Change in the Sahel and Soudan, Africa, 1982–2005 ». *Remote Sensing of Environment* 108 (4): 385‑92. https://doi.org/10.1016/j.rse.2006.11.025.

Hosmer, DW, et S Lemeshow. 1989. *Applied logistic regression*. New York: John Wiley & Sons. https://www.researchgate.net/profile/Andrew_Cucchiara/publication/261659875_Applied_Logistic_Regression/links/542c7eff0cf277d58e8c811e/Applied-Logistic-Regression.pdf.

Huete, A.R. 1988. « A Soil-Adjusted Vegetation Index (SAVI) ». *Remote Sensing of Environment* 25 (3): 295‑309. https://doi.org/10.1016/0034-4257(88)90106-X.

Inglada, Jordi, Marcela Arias, Benjamin Tardy, Olivier Hagolle, Silvia Valero, David Morin, Gérard Dedieu, et al. 2015. « Assessment of an Operational System for Crop Type Map Production Using High Temporal and Spatial Resolution Satellite Optical Imagery ». *Remote Sensing* 7 (9): 12356‑79. https://doi.org/10.3390/rs70912356.

Inglada, Jordi, Vincent Vincent, Marcela Arias, et Benjamin Tardy. 2016. *iota2-a25386*. Zenodo. https://doi.org/10.5281/zenodo.58150.

Islam, Akm Saiful, et Sujit Kumar Bala. 2008. « Assessment of Potato Phenological Characteristics Using MODIS-Derived NDVI and LAI Information ». *GIScience &*

Remote Sensing 45 (4): 454‑70. https://doi.org/10.2747/1548-1603.45.4.454.

Jeune, Wesly, Márcio Rocha Francelino, Eliana de Souza, Elpídio Inácio Fernandes Filho, Genelício Crusoé Rocha, Wesly Jeune, Márcio Rocha Francelino, Eliana de Souza, Elpídio Inácio Fernandes Filho, et Genelício Crusoé Rocha. 2018. « Multinomial Logistic Regression and Random Forest Classifiers in Digital Mapping of Soil Classes in Western Haiti ». *Revista Brasileira de Ciência do Solo* 42. https://doi.org/10.1590/18069657rbcs20170133.

Jönsson, Per, Zhanzhang Cai, Eli Melaas, Mark A. Friedl, et Lars Eklundh. 2018. « A Method for Robust Estimation of Vegetation Seasonality from Landsat and Sentinel-2 Time Series Data ». *Remote Sensing* 10 (4): 635. https://doi.org/10.3390/rs10040635.

Kauth, R J, et G S Thomas. 1976. « The Tasselled Cap - A Graphic Description of the Spectral-Temporal Development of Agricultural Crops as Seen by LANDSAT ». *Proceedings Second Ann. Symp. Machine Processing of Remotely Sensed Data.*, West Lafayette: Purdue University Lab. App. Remote Sensing., , 13.

Kühnlein, Meike, Tim Appelhans, Boris Thies, et Thomas Nauss. 2014. « Improving the Accuracy of Rainfall Rates from Optical Satellite Sensors with Machine Learning — A Random Forests-Based Approach Applied to MSG SEVIRI ». *Remote Sensing of Environment* 141 (février): 129‑43. https://doi.org/10.1016/j.rse.2013.10.026.

Lemaire, Jean. 2015. « Des données climatiques spatialisées pour un diagnostic de qualité Aurelhy, ETPQ,Safran et Digitalis », janvier.

Li, Qiangzi, Xin Cao, Kun Jia, Miao Zhang, et Qinghan Dong. 2014. « Crop type identification by integration of high-spatial resolution multispectral data with features extracted from coarse-resolution time-series vegetation index data ». *International Journal of Remote Sensing* 35 (16): 6076‑88. https://doi.org/10.1080/01431161.2014.943325.

Liu, Hui Qing, et Alfredo Huete. 1995. « A feedback based modification of the NDVI to minimize canopy background and atmospheric noise ». *IEEE Transactions on Geoscience and Remote Sensing* 33 (2): 457‑65. https://doi.org/10.1109/TGRS.1995.8746027.

Lorena, Ana C., Luis F. O. Jacintho, Marinez F. Siqueira, Renato De Giovanni, Lúcia G. Lohmann, André C. P. L. F. de Carvalho, et Missae Yamamoto. 2011. « Comparing Machine Learning Classifiers in Potential Distribution Modelling ». *Expert Systems with Applications* 38 (5): 5268‑75. https://doi.org/10.1016/j.eswa.2010.10.031.

McHugh, Marry L. 2012. « Interrater Reliability: The Kappa Statistic ». *Biochemia Medica*, 276‑82. https://doi.org/10.11613/BM.2012.031.

Meier, Uwe. 2001. *Growth Stages of Mono-and Dicotyledonous plants*. 2nd ed. Berlin,Germany: Federal Biological Research Centre for Agriculture and Forestry. https://www.politicheagricole.it/flex/AppData/WebLive/Agrometeo/MIEPFY800/BBCHengl2001.pdf.

Morrison, M. J., P. B. E. McVETTY, et C. F. Shaykewich. 1989. « The Determination and Verification of a Baseline Temperature for the Growth of Westar Summer Rape ». *Canadian Journal of Plant Science* 69 (2): 455‑64. https://doi.org/10.4141/cjps89-057.

Muller-Wilm, U. 2012. « Sentinel-2 MSI—Level 2A Products Algorithm Theoretical Basis Document ». European Space Agency. https://earth.esa.int/c/document_library/get_file?folderId=349490&name=DLFE-4518.pdf.

Muñoz, Paul, Johanna Orellana-Alvear, Patrick Willems, et Rolando Célleri. 2018. « Flash-Flood Forecasting in an Andean Mountain Catchment—Development of a Step-Wise Methodology Based on the Random Forest Algorithm ». *Water* 10 (11): 1519. https://doi.org/10.3390/w10111519.

Pope, Katherine S., Volker Dose, David Da Silva, Patrick H. Brown, Charles A. Leslie, et

Theodore M. DeJong. 2013. « Detecting Nonlinear Response of Spring Phenology to Climate Change by Bayesian Analysis ». *Global Change Biology* 19 (5): 1518‑25. https://doi.org/10.1111/gcb.12130.

Rock, B., D. Williams, et J. Vogelmann. 1985. « Field and Airborne Spectral Characterization of Suspected Acid Deposition Damage in Red Spruce (Picea Rubens) from Vermont ». *Machine Processing of Ren~otely Sensed Data Symposium*, 71‑81.

Rouse, J. W., Jr., R.H Haas, J.A Schell, et D.W Deering. 1973. « Monitoring vegetation systems in the Great Plains with ERTS ». *NASA SP-351 I* 3rd ERTS Symposium: 309‑17.

Roy, P. S., K. P. Sharma, et A. Jain. 1996. « Stratification of Density in Dry Deciduous Forest Using Satellite Remote Sensing Digital Data—An Approach Based on Spectral Indices ». *Journal of Biosciences* 21 (5): 723‑34. https://doi.org/10.1007/BF02703148.

Simonneau, Danièle, Didier Chollet, et Céline Gouwier. 2013. « Vigicultures, base d'information des BSV grandes cultures - Arvalis ». https://www.arvalisinstitutduvegetal.fr/. 2013. https://www.arvalisinstitutduvegetal.fr/86-des-bsv-grandes-cultures-sont-ecrits-a-partir-des-donnees-de-vigicultures--@/view-703-arvstatiques.html.

Sokolova, Marina, et Guy Lapalme. 2009. « A Systematic Analysis of Performance Measures for Classification Tasks ». *Information Processing & Management* 45 (4): 427‑37. https://doi.org/10.1016/j.ipm.2009.03.002.

Springate, David A., et Paula X. Kover. 2014. « Plant Responses to Elevated Temperatures: A Field Study on Phenological Sensitivity and Fitness Responses to Simulated Climate Warming ». *Global Change Biology* 20 (2): 456‑65. https://doi.org/10.1111/gcb.12430.

Sulik, John J., et Dan S. Long. 2016. « Spectral Considerations for Modeling Yield of Canola ». *Remote Sensing of Environment* 184 (octobre): 161‑74. https://doi.org/10.1016/j.rse.2016.06.016.

Sykas, Dimitris. 2019. « Spectral Indices with Multispectral Satellite Data ». GIS and Earth Observation University. 2019. https://www.geo.university/pages/spectral-indices-with-multispectral-satellite-data.

Tanre, D., B.N. Holben, et Y.J. Kaufman. 1992. « Atmospheric correction algorithm for NOAA-AVHRR products: theory and application ». *IEEE Transactions on Geoscience and Remote Sensing* 30 (2): 231‑48. https://doi.org/10.1109/36.134074.

Tutz, Gerhard, Wolfgang Pößnecker, et Lorenz Uhlmann. 2015. « Variable Selection in General Multinomial Logit Models ». *Computational Statistics & Data Analysis* 82 (février): 207‑22. https://doi.org/10.1016/j.csda.2014.09.009.

Vliet, Arnold J. H. van, Rudolf S. de Groot, Yvette Bellens, Peter Braun, Robert Bruegger, Ekko Bruns, Jan Clevers, et al. 2003. « The European Phenology Network ». *International Journal of Biometeorology* 47 (4): 202‑12. https://doi.org/10.1007/s00484-003-0174-2.

Vrieling, Anton, Michele Meroni, Roshanak Darvishzadeh, Andrew K. Skidmore, Tiejun Wang, Raul Zurita-Milla, Kees Oosterbeek, Brian O'Connor, et Marc Paganini. 2018. « Vegetation Phenology from Sentinel-2 and Field Cameras for a Dutch Barrier Island ». *Remote Sensing of Environment* 215 (septembre): 517‑29. https://doi.org/10.1016/j.rse.2018.03.014.

Wardlow, Brian D., et Stephen L. Egbert. 2008. « Large-Area Crop Mapping Using Time-Series MODIS 250 m NDVI Data: An Assessment for the U.S. Central Great Plains ». *Remote Sensing of Environment* 112 (3): 1096‑1116. https://doi.org/10.1016/j.rse.2007.07.019.

Zeng, Linglin, Brian D. Wardlow, Daxiang Xiang, Shun Hu, et Deren Li. 2020. « A Review of Vegetation Phenological Metrics Extraction Using Time-Series, Multispectral Satellite Data ». *Remote Sensing of Environment* 237 (février): 111511. https://doi.org/10.1016/j.rse.2019.111511.

Zhang, Chongsheng, Changchang Liu, Xiangliang Zhang, et George Almpanidis. 2017. « An Up-to-Date Comparison of State-of-the-Art Classification Algorithms ». *Expert Systems with Applications* 82 (octobre): 128‑50. https://doi.org/10.1016/j.eswa.2017.04.003.

Zhang, Xiaoyang, Mark A. Friedl, et Crystal B. Schaaf. 2009. « Sensitivity of vegetation phenology detection to the temporal resolution of satellite data ». *International Journal of Remote Sensing* 30 (8): 2061‑74. https://doi.org/10.1080/01431160802549237.

Zhong, Liheng, Peng Gong, et Gregory S. Biging. 2014. « Efficient Corn and Soybean Mapping with Temporal Extendability: A Multi-Year Experiment Using Landsat Imagery ». *Remote Sensing of Environment* 140 (janvier): 1‑13. https://doi.org/10.1016/j.rse.2013.08.023.

Zhong, Liheng, Tom Hawkins, Greg Biging, et Peng Gong. 2011. « A phenology-based approach to map crop types in the San Joaquin Valley, California ». *International Journal of Remote Sensing* 32 (22): 7777‑7804. https://doi.org/10.1080/01431161.2010.527397.

Zhu, Xiaofeng, Lei Zhang, et Zi Huang. 2014. « A Sparse Embedding and Least Variance Encoding Approach to Hashing ». *IEEE Transactions on Image Processing* 23 (9): 3737‑50. https://doi.org/10.1109/TIP.2014.2332764.

Anexos

Tablas de resultados de las modelizaciones con los datos de entrenamiento y los datos de test

Spectral

Ensemble de données	**OOB**	**Modèle** Entraînement		**Validation** Test		**Précision du modèle**	**Différence avec mod. de réf.** baseline = 0.84
		accuracy	kappa	accuracy	kappa		
Bandes	32.27%	0.68	0.59	0.70	0.62	0.68	0.16
Índices	32.50%	0.67	0.59	0.69	0.60	0.67	0.17
TassCap	52.35%	0.58	0.46	0.60	0.49	0.58	0.26

Inrae_Iota2 (Indices)

Ensemble de données	**OOB**	**Modèle** Entraînement		**Validation** Test		**Précision du modèle**	**Différence avec mod. de réf.** baseline = 0.84
		accuracy	kappa	accuracy	kappa		
Indices_Iota2	32.5%	0.68	0.59	0.69	0.60	0.68	0.17
Indices_Inrae	37.9%	0.62	0.52	0.65	0.56	0.62	0.22

Autres modèles

Ensemble de données	**OOB**	**Modèle** Entraînement		**Validation** Test		**Précision du modèle**	**Différence avec mod. de réf.** baseline = 0.84
		accuracy	kappa	accuracy	kappa		
Date_Dep	18.61%	0.81	0.77	0.82	0.77	0.81	0.03
Weathers	17.71%	0.82	0.78	0.82	0.77	0.82	0.02
WDD	17.80%	0.82	0.78	0.82	0.77	0.82	0.02
IDD	18.98%	0.81	0.76	0.83	0.79	0.81	0.03
IWDD	15.73%	0.84	0.80	0.85	0.81	0.84	0.00