Prédiction statistique des stades du colza (*Brassica Napus L.*) à partir de données météorologiques et d'observations satellitaires

Mémoire de Stage

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

Table de Matières

Résumé

Les changements de l'état phénologique des plantes sont des indicateurs importants dans la recherche agronomique. Cependant, la difficulté de collecter des données phénologiques à grande échelle est un défi actuel. L'utilisation conjointe d'informations spectrales provenant d'images satellites et de données météorologiques prétraitées apparaît comme une réponse à ce défi.

Par conséquent, l'objectif principal de ce travail est d'ajuster et d'évaluer différents modèles pour prédire les phases phénologiques à l'aide de données satellitaires et de produits météorologiques. Un jeu de données pour 8 phénophases collectées dans la base de données **Vigicultures®** au cours de la saison agricole 2017 a été construit pour des parcelles de colza réparties sur l'ensemble du territoire français. Nous avons ajusté les modèles statistiques en utilisant les méthodes de *Machine Learning* (ML) les plus couramment utilisées pour classer les informations catégorielles, telles que le *Lasso-Multinomial*, le *Random Forest* et le *KNN*. La qualité des modèles a été estimée à l'aide de leurs matrices de confusion et de leur *accuracy* globale. Les résultats obtenus ont montré un potentiel variable pour coupler les indices dérivés des produits de télédétection avec les variables météorologiques. Les stades de culture sont estimés avec ces modèles en s'appuyant sur plusieurs sources de données : les données spectrales Sentinel 2, des données météorologiques (modèle SAFRAN de Météo-France) et des données spatio-temporelles. Avec le modèle de référence mobilisant données météorologiques et spectrales, nous avons obtenu une *accuracy* de 0,84 avec presque uniquement des inversions entre stades voisins. Nous avons étudié l'impact de modifications de ce modèles ainsi que l'impact des différentes variables sur la qualité de la prédiction. Nous avons constaté qu'une bonne prédiction des stades phénologiques intermédiaires est principalement liée aux données météorologiques, tandis que pour les états printaniers (floraison), il y a une forte importance des indices spectraux tels que le *NDYI*. La prise en compte des variables spatio-temporelles n'améliorent que marginalement le modèle de référence. La diversité des sources d'information est plus importante que les pré-traitements avant de les fournir au modèle de Random Forest. Bien que le modèle de référence ne soit pas destiné à remplacer les observations in situ, il peut aider au processus de prise de décision.

Mots clés : Phénologie, *Machine Learning*, classification, *Random Forest*, Colza, *Brassica Napus*, Copernicus, Sentinel-2, modélisation des cultures, changement climatique.

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 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 pre-processing 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.

Remerciements

L'expérience de la construction de connaissances est une aventure extrême et enrichissante. Je suis reconnaissant aux personnes qui, d'une manière ou d'une autre, ont participé à cet apprentissage toujours renouvelé pour continuer à connaître, découvrir et avancer dans cette aventure que j'appelle la vie.

Je tiens tout d'abord à remercier Corentin Barbu qui m'a guidée et orientée dans cette pratique académique. Il a toujours su être disponible et m'a beaucoup soutenu pendant cette formation en programmation et statistiques qui, je dois l'admettre, a été un grand défi. Merci également pour les conversations qui ont eu lieu pendant la pause déjeuner.

Un merci tout particulier à toute l'équipe Inrae de Thiverval-Grignon pour leur gentillesse et leur accueil, ainsi qu'à tous les amis que j'ai rencontrés dans le salon des étudiants à Versailles.

Je remercie le CNES pour avoir pourvu au financement de ce stage, l'institut technique Terres Inovia pour avoir donné accès à la base de données Vigicultures pour le colza, l'unité AgroClim de l'INRAE pour avoir donné accès aux données météorologiques SAFRAN et Mathieu Fauvel du CESBIO pour avoir dirigé le projet PARCELLES financé ce stage et fourni des extractions iota2 des données Sentinel-2.

Merci également à Nicolas Delbart pour son soutien et son aide dans les moments où je pensais qu'il ne serait pas possible d'entamer ce processus.

Enfin, un merci à M. Rivals pour son soutien quotidien pendant le stage.

Introduction

Contexte Général

Chaque année, le CNES (Centre National d'Etudes Spatiales) lance un appel à propositions de recherche auprès des laboratoires spatiaux pour le développement de thèmes issus de la télédétection des surfaces terrestres. Le projet TOSCA-PARCELLE est le résultat d'un de ces appels dont l'utilisation d'images satellites est l'élément principal. Ce projet vise à promouvoir les efforts pour unifier et capitaliser la chaîne de traitement de **iota2** (Infrastructure pour l'Occupation des sols par Traitement Automatique).

À l'origine, iota2 a été conçu comme un flux de travail de classification pour la cartographie de l'occupation des sols à grande échelle, mais la polyvalence de l'algorithme permet également d'effectuer des extractions d'informations spectrales dans toute la France à l'échelle de la parcelle agricole qui nous intéresse ici.

L'utilisation des informations spectrales extraites de l'utilisation de Iota2 permet à l'Institut National de Recherche pour l'Agriculture et l'Environnement (INRAE) et à l'Institut des Sciences et Industries du Vivant et de l'Environnement AgroparisTech de co-construire avec les agriculteurs l'avenir d'une agriculture plus durable.

Au sein de l'Unité Mixte de Recherche (UMR) en agronomie, l'équipe de recherche crée des outils d'aide à la décision. Les outils conçus visent notamment à améliorer le contrôle biologique des bio-agresseurs afin de réduire l'utilisation des produits phytosanitaires.

C'est dans ce contexte que s'inscrit ce stage, dont l'objectif est d'établir un modèle de classification des stades phénologiques des cultures agroalimentaires. Cela nous permettra de comprendre comment la présence de bioagresseurs à certains stades du développement des plantes peut affecter le rendement final des cultures.

La surveillance des différents stades de développement des cultures est appelée phénologie (Beurs et Henebry 2005). La phénologie a été abordée scientifiquement à partir de différentes échelles spatiales. Au niveau des parcelles, il existe des méthodologies in situ pour déterminer les stades phénologiques exacts des cultures (van Vliet et al. 2003). A l'échelle locale, l'utilisation de vecteurs aériens (UAV) équipés d'instruments de mesure (caméras spectrales), permet d'analyser la végétation à une plus grande échelle sans compromettre l'*accuracy* des informations qui alimentent les modèles (Berra, Gaulton, et Barr 2019). À l'échelle régionale et mondiale, l'utilisation d'instruments d'observation à distance facilite l'analyse de vastes zones (forêts et champs) pour déterminer les tendances et les réactions des cultures à différentes variables telles que le changement climatique, la qualité des sols et la présence de stress, entre autres (Heumann et al. 2007; Han et al. 2018; Brown et al. 2008).

Analyse phénologique en agriculture

En agriculture, l'analyse à distance du cycle phénologique des cultures est un outil essentiel pour, entre autres, déterminer le rendement et la réponse des champs aux variables externes, en particulier à la pression des ravageurs et des maladies des cultures. L'incursion de la télédétection dans l'agriculture a permis de considérer des effets spécifiques extrapolés à des réalités plus larges avec moins d'investissement de ressources (X. Zhang, Friedl, et Schaaf 2009; Wardlow et Egbert 2008). L'étude de la phénologie des plantes par télédétection a été largement discutée dans la littérature, car le lancement de satellites équipés de capteurs capables d'exploiter l'énergie réfléchie par les surfaces terrestres a permis d'analyser le comportement de la végétation soit sur la base de sa chlorophylle, soit de sa structure ou de sa capacité de rétention d'eau pour en déduire son état phénologique (X. Zhang, Friedl, et Schaaf 2009).

Télédétection et phénologie

Des capteurs tels que le MODIS à bord des satellites américains *Acqua et Terra* ont été largement utilisés à cette fin (Fisher et Mustard 2007; Ahl et al. 2006). Cependant, c'est actuellement la mission européenne *Sentinel*, avec sa famille de satellites et ses améliorations d'instruments, qui fournit des images satellites à haute résolution dans l'espace et le temps (Jönsson et al. 2018; Vrieling et al. 2018). Du point de vue de la télédétection, l'estimation conventionnelle des mesures phénologiques est généralement faite à partir de séries temporelles. Cette estimation comporte généralement trois étapes principales : 1) le nettoyage des données et l'établissement de rapports ; 2) le lissage des données et la reconstruction des séries temporelles; et 3) l'extraction des mesures phénologiques générées à partir des données des séries temporelles reconstruites (Zeng et al. 2020).

Apprentissage automatique et phénologie

Il existe également d'autres approches basées sur la complémentarité ("couplage") entre différents types de données (Almeida et al. 2014). Ces approches peuvent établir des modèles prédictifs des différentes étapes d'un phénomène en utilisant des outils d'intelligence artificielle tels que le *Machine Learning* (ML) dont le *Deep Learning* (DL) fait partie afin d'identifier des modèles (Czernecki, Nowosad, et Jabłońska 2018).

Dans le cadre de ce stage, nous analyserons la contribution des informations spectrales, climatologiques et spatio-temporelles à la prédiction des états phénologiques des cultures d'importance agro-écologique. Nous aborderons cette question de recherche en utilisant des outils de classification avec des méthodes d'apprentissage automatique. Nous déterminerons l'évolution de chaque stade phénologique d'une campagne de colza dans des parcelles réparties sur l'ensemble du territoire français.

Dans un premier temps, nous extrairons les informations spectrales des 10 bandes Sentinel-2, calculerons les indices spectraux et évaluerons leur potentiel de

classification au stade de la floraison, puis nous couplerons les données météorologiques aux informations spectrales et enfin nous utiliserons des méthodes d'apprentissage automatique telles que la régression logistique de pénalités multinomiales (LASSO), les K- Nearest Neighbors (KNN) et la Random Forest (RF) pour déterminer la contribution des variables thématiques à la détermination des patrons dans les données.

Dans ce cas d'application, l'utilisation des méthodes d'apprentissage automatique, nous permettra de connaître la contribution de la télédétection à la gestion durable des bioagresseurs dans les cultures de grande importance agroalimentaire, en déterminant la combinaison appropriée de variables pour la classification des états phénologiques du colza (*Brassica napus L.*).

1. Matériels et Méthodes

La méthodologie est divisée en 3 étapes. La première section décrit les bases de données utilisées pour la recherche des informations utilisées. Elle présente également les régions où les parcelles sont situées. La deuxième section présente les méthodes de classification utilisées pour la détection des stades phénologiques. La troisième section détaille la méthodologie utilisée pour définir la contribution des différents ensembles de variables.

1.1. Matériels

1.1.1. Données Agronomiques

Vigicultures®

Application départementale d'introduction de données épidémiologiques pour les grandes cultures (colza, blé, tournesol, etc.) mise en œuvre par les instituts techniques (Arvalis, Terre Inovia, ITB) (Simonneau, Chollet, et Gouwier 2013). **Vigicultures®** et la base de données VégéObs collecte des données de surveillance épidémiologique pour obtenir des informations en temps réel sur la pression des ravageurs sur les cultures. Cette base de données orchestrée par le ministère de l'agriculture et le ministère de l'environnement est un outil essentiel de prévention et d'analyse des risques dans la création des Bulletins phytosanitaires (BSV). Pour notre étude de cas, nous avons utilisé les stades phénologiques des cultures qui sont enregistrés à chaque fois qu'une observation de ravageurs ou de maladies est faite.

Les stades phénologiques

L'état phénologique des parcelles est établi à partir d'une classification propriétaire établie dans la base de données **Vigicultures®**. Pour la culture du colza, 28 stades phénologiques ont été identifiés (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 de floraison, G4 - Floraison terminée, G5 et Hors culture). Les états "Floraison Terminée" et "Hors de Culture" ont été écartés en raison de leur ambiguïté et de leur faible nombre d'observations.

Voici un parallèle des états de **Vigicultures®** avec l'échelle BBCH (Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie) L'échelle BBCH décrit les stades phénologiques des cultures en utilisant des critères qui relient le stade de croissance à un code décimal (Meier 2001). Le premier chiffre indique le stade de développement principal (par exemple 6 = floraison), tandis que le deuxième chiffre se réfère à un stade de croissance secondaire ou au pourcentage de plantes à ce stade.

Tableau 1. Parallèle entre **Vigicultures®** et l'échelle **BBCH**

Échelle **Vigicultures® originale**	Regroupement de l'Échelle **Vigicultures®**	Échelle **BBCH** (Meier 2001)
Semis, A	A	**phase 0 :** Germination, germination, développement des bourgeons.
B1, B2, B3, B4, B5, B6	B1 -B6	**phase 1 :** Développement des feuilles (tige principale).
B7, B8, B9, B10, > 10 feuilles	B7 - B10>	**phase 2 :** Formation de pousses latérales / (tallage).
C1, C2	C	**phase 3 :** Croissance de la tige longitudinale ou de la rosette, développement des pousses (germes)/racines (tige principale).
D1, D2	D	**phase 4 :** Développement des parties végétatives récoltables de la plante ou des organes végétatifs de multiplication/encastrement.
E	E	**phase 5 :** Émergence de l'inflorescence (tige principale).
F1, F2, G1, G2, G3, G4 - Floraison toujours en cours, G4 - Floraison terminée, G5	F-G	**phase 6 :** Floraison (tige principale).
		phase 7 : Développement du fruit.
		phase 8 : Coloration ou maturation des fruits et des graines.
NA	NA	**phase 9 :** La sénescence.

Registre Parcellaire Graphique (RPG)

Base de données géographiques utilisée comme référence pour l'évaluation des aides de la politique agricole commune européenne (PAC). La version anonyme contient des données graphiques des parcelles (depuis 2015) avec leur récolte principale. Ces données sont produites par l'Agence des services et des paiements (SPA) depuis 2007. La réutilisation du RPG est gratuite pour toutes les utilisations, y compris commerciales, selon les termes de la "licence ouverte"[1].

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

Identification des parcelles d'intérêt

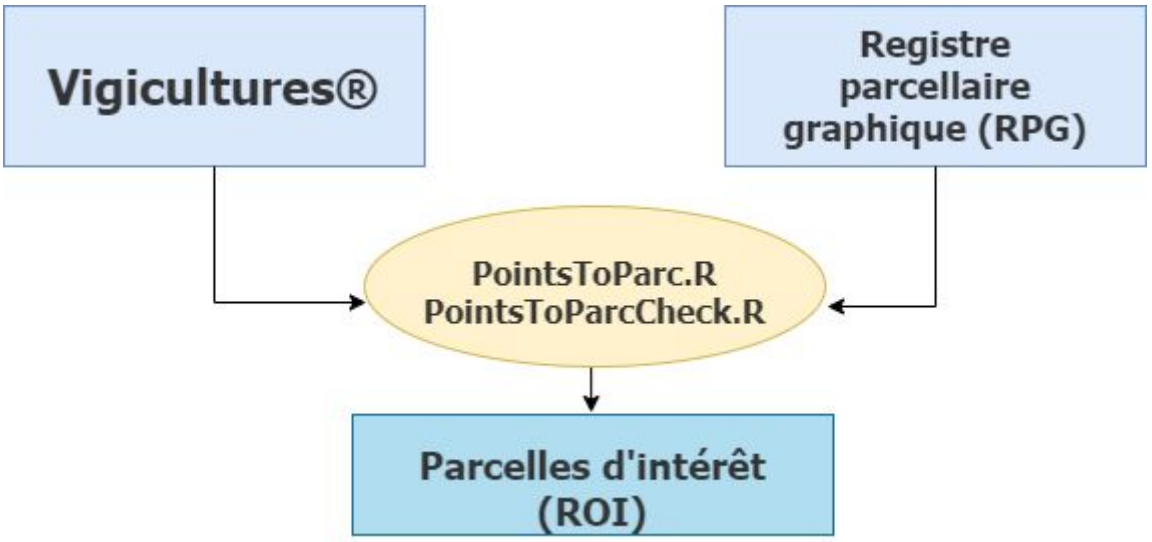


Fig. 1. Schéma général du prétraitement des données agronomiques

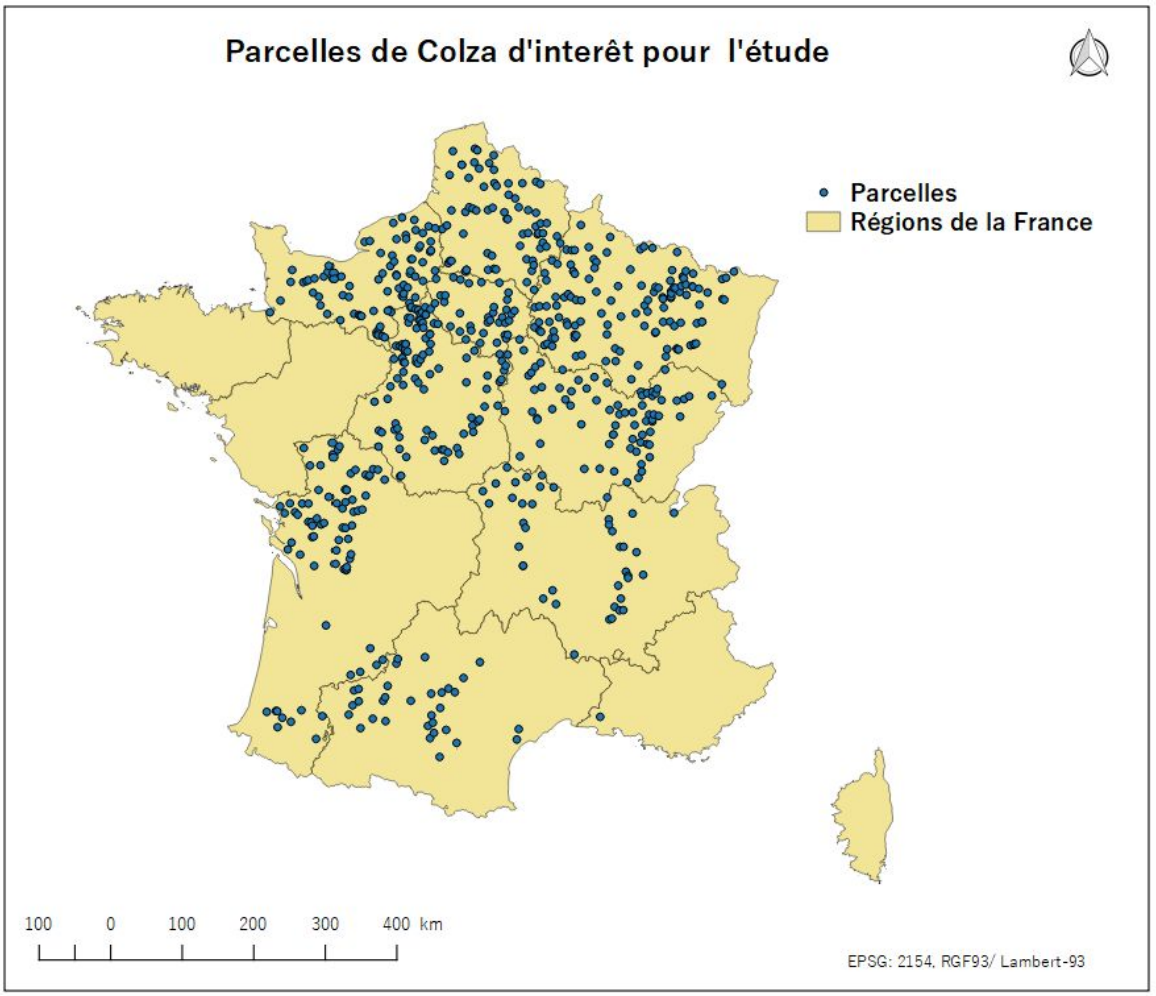


Fig 2. Cartes des parcelles d'intérêt en France

A partir de la base de données **Vigicultures®**, les informations sur les paramètres agricoles (type de culture, état phénologique observé, département, etc.) liés à un point GPS sont extraites et fusionnées avec les informations relatives à la parcelle enregistrée dans la base de données RPG. Les polygones résultants ont délimité les régions d'intérêt (ROI) pour une analyse ultérieure à l'aide d'images satellites et de variables climatiques.

1.1.2. Données Spectrales

Sentinel-2

Le réseau de satellites optiques Sentinel-2 (2A et 2B) fait partie de la famille des satellites d'observation terrestre à distance du projet spatial européen. Depuis juin 2015, les images multispectrales permettent d'analyser le développement et le cycle de croissance des plantes à l'échelle mondiale. Avec 13 bandes spectrales à haute résolution spatiale (4 bandes à 10m, 6 bandes à 20m et 3 bandes à 60m) et un temps de revisite de 5 jours, son application en agriculture est l'une des plus documentées (Zhang, Friedl, et Schaaf 2009).

Transformation de l'information spectrale

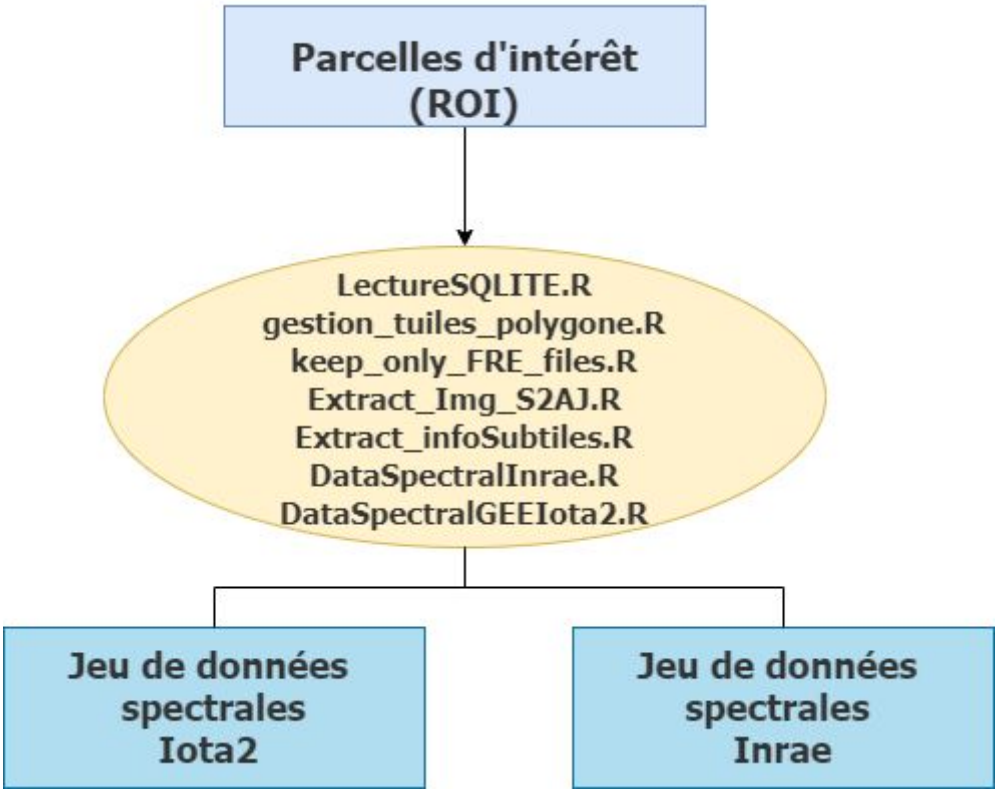


Fig 3. Diagramme général de prétraitement des informations spectrales de Sentinel-2

L'information spectrale est obtenue à partir de deux méthodologies différentes. Dans les deux méthodologies, les tuiles Sentinel-2 de niveau 2A ont été téléchargées à

partir du centre de données terrestres Theia[2] (Hagolle [2016] 2020). Les acquisitions correspondent à la saison de récolte 2017 (entre le 1er juillet 2016 et le 25 août 2017).

Dans la première méthodologie, l'extraction des données spectrales a été réalisée en utilisant **iota2** (Inglada et al. 2016) et **MAJA** (MACCS[3]-ATCOR[4] Joint Algorithm) développés par le Centre National d'Etudes Spatiales (CNES) et le Centre d'Etudes Spatiales de la Biosphère (CESBIO) d'une part, et le Centre Aérospatial Allemand (DLR) d'autre part. Les images sont orthorectifiées, corrigées atmosphériquement sans nuages et avec détection des ombres (Baetens, Desjardins, et Hagolle 2019). Toutes les acquisitions ont été ré-échantillonnées pour combler les lacunes laissées par les nuages et les ombres (tous les 10 jours, à partir du 2016-07-01 et jusqu'au 2017-08-25). Les 10 bandes utilisées de S2 (B2, B3, B4, B5, B6, B7, B8, B8A, B11 et B12) sont récupérées à une résolution spatiale de 10 et 20 mètres sans processus de rééchantillonnage.

Dans la deuxième méthodologie, les acquisitions ont été effectuées à l'aide de l'outil **SEN2COR** (Muller-Wilm 2012). Les 10 bandes sont présentées sous deux formes : une forme, la Réflectance de surface corrigée pour les effets atmosphériques et environnementaux (SRE_Bx.tif), une autre forme, la Réflectance plane qui est en outre corrigée pour les effets de pente (FRE_Bx.tif)[5]. Nous travaillerons avec les données S2 L2A en utilisant le produit FRE_Bx.tif. Les bandes ont été extraites dans leur résolution d'origine puis transformées à 10 mètres en utilisant pour définir la nouvelle valeur des pixels la méthode du plus proche voisin.

Dans les deux cas, les parcelles d'intérêt récupérées dans les bases de données agronomiques sont associées aux informations spectrales des tuiles liées à leur localisation géographique. Les images satellites sont sélectionnées à partir de la date d'observation des différents stades phénologiques. Cette sélection vise à ce que la différence entre la date d'observation de l'état et la date de l'information spectrale soit comprise entre 0 et 5 jours avant l'observation in-situ.

Indices spectraux

Des bandes spectrales ont été utilisées pour obtenir les indices spectraux considérés comme pertinents pour l'analyse des états phénologiques en agriculture. Dans les tableaux suivants, nous présentons les bandes spectrales et les indices utilisés dans cette étude de cas.

[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/

Tableau 2. Bandes spectrales Sentinel-2 utilisées

Nombre	Résolution	Longueur d'Onde	Description
B2	10 mètres	496.6nm (S2A) / 492.1nm (S2B	Bleu
B3	10 mètres	560nm (S2A) / 559nm (S2B)	Vert
B4	10 mètres	664.5nm (S2A) / 665nm (S2B)	Rouge
B5	10 mètres	703.9nm (S2A) / 703.8nm (S2B)	Red Edge 1
B6	20 mètres	740.2nm (S2A) / 739.1nm (S2B)	Red Edge 2
B7	20 mètres	782.5nm (S2A) / 779.7nm (S2B)	Red Edge 3
B8	20 mètres	835.1nm (S2A) / 833nm (S2B)	Proche Infrarouge
B8A	20 mètres	864.8nm (S2A) / 864nm (S2B)	Red Edge 4
B11	20 mètres	1613.7nm (S2A) / 1610.4nm (S2B)	SWIR 1
B12	20 mètres	2202.4nm (S2A) / 2185.7nm (S2B)	SWIR 2

Tableau 3. Indices spectraux utilisés et leurs formules

Indices	Formule pour Sentinel-2	Source
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

En plus des indices spectraux mentionnés ci-dessus, les informations spectrales obtenues ont été transformées à partir de la méthodologie "*Tasseled Cap*".

Kauth, R. J. et Thomas, G. S. (1976) ont imaginé une transformation de l'information des bandes spectrales pour maximiser l'information contenue dans les nouveaux éléments d'analyse. Il s'agit d'une méthode de compression permettant de réduire de multiples données spectrales, en l'occurrence 6 bandes, en trois néo-canaux, qui permettent de comprendre d'importants phénomènes de développement des cultures dans l'espace spectral (Kauth et Thomas 1976). Les néo-canaux obtenus après la transformation sont les suivants :

Tableau 4. Néo-canaux *Tasseled Cap*

Indices	Formule pour Sentinel-2	Utilisation
Brightness*[6] *Index	$BI = \sqrt{\frac{B4^2}{\frac{B3^2}{2}}}$	Associé aux variations de la réflectance du sol.
***Greenness*[7]**	**Greenness** = (-0.2848B2)+(-0.2435B3)+(-0.5436B4) +0.7243B8+0.0840B11+(-0.1800B12)	Corrélation avec la vigueur de la végétation.
***Wetness*[8]**	**Wetness** = 0.1509B2+0.1973B3+0.3279B4 +0.3406B8+(-0.7112B11)+(-0.4572B12)	Influencé par les bandes dans l'IR Moyen et lié à l'humidité des plantes et du sol.

1.1.3. Données météorologiques

AgroClim

AgroClim est une unité au service de la communauté INRAE. Cette unité gère le réseau agroclimatique national de l'INRAE et la base de données correspondante. Sa fonction est d'assurer la traçabilité des observations dépendantes du climat. AgroClim est également le point d'entrée unique des unités INRAE pour obtenir des données météorologiques de Météo-France[9].

Les données utilisées sont le produit du modèle de données climatologiques développé par Météo-France, *SAFRAN* (Système d'Analyse Fournissant des Renseignements Atmosphériques à la Neige). Safran travaille sur des régions climatiquement homogènes. Ces régions ont une forme irrégulière, leur surface est normalement inférieure à 1 000 km^2. Dans chaque région homogène, Safran estime la variation de 8 paramètres climatiques (tableau 5 ci-dessous) pour chaque classe d'altitude de 300 m, à partir de toutes les données climatiques disponibles (postes météorologiques, mais aussi des analyses des modèles de prévision du temps à grande échelle comme le modèle ARPEGE de Météo-France) (Lemaire 2015). Les analyses de température, humidité, vitesse du vent et nébulosité sont produites toutes les 6 heures. L'analyse des précipitations est faite au pas de temps journalier. Après avoir obtenu les valeurs pour les zones, l'analyse est interpolée spatialement sur une grille régulière 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/

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

Données disponibles	Période	Résolution de la maille
1. Températures minimales, maximales et moyennes à 2 m au-dessus du sol (en °C) ; 2. Humidité relative moyenne à 2 m au-dessus du sol (en g.kg-1) ; 3. Force moyenne du vent à 10 m au-dessus du sol (en m/s) 4. Précipitations solides (en mm) 5. Précipitations liquides (en mm) 6. Rayonnement infrarouge/solaire (en J/cm^2) 7. Rayonnement atmosphérique (en J.cm-2) 8. Évapotranspiration potentielle (ETP mm), formule de Penman-Monteith	1958 à aujourd'hui	8 km x 8 km

Transformation des informations météorologiques

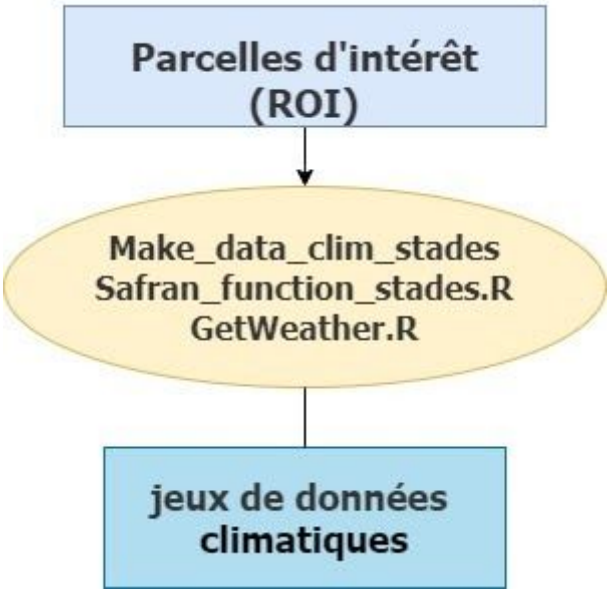


Fig. 4. Schéma général de prétraitement des bases de données météorologiques.

Pour cette étude, nous avons pris en compte toutes les variables climatologiques obtenues par le modèle SAFRAN. Nous avons ajouté une autre variable, le degré jour de croissance (gdd), qui est étroitement liée à l'évolution phénologique des cultures. Le calcul de cette variable est basé sur la formule suivante:

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

Nous utilisons la température base de 5° selon (Morrison, McVETTY, et Shaykewich 1989) et la fonction gdd() du paquet de ***pollen***[10], basée sur (Baskerville et Emin 1969).

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

Les données départementales quotidiennes des stations météorologiques les plus proches des parcelles d'intérêt ont été regroupées par semaine. La prédiction des états a été faite avec les informations climatologiques des 52 dernières semaines à la date d'observation in situ. Cette décision est basée sur l'hypothèse empirique que les variations des conditions météorologiques pendant au moins 10 mois peuvent avoir un impact sur la croissance des plantes, du semis à la récolte. En outre, les informations météorologiques sont une approximation des informations temporelles qui pourraient être utiles pour déterminer s'il est temps de planter, puisque les variations de température, par exemple, permettent à un modèle comme Random Forest de trouver des oscillations dans le signal. Si l'on considère les défis que le changement climatique actuel pose au processus de modélisation, cette identification de la saison a un avantage sur la date d'observation car elle permet d'adapter la météorologie à une période spécifique de l'année, ce qui permet d'ajuster le modèle à d'autres régions et d'autres années. D'autre part, du point de vue du prétraitement des données, si nous extrayons 10 mois pour un état phénologique, il est cohérent de le faire pour tous les autres, afin d'avoir le même nombre de variables indépendantes par classe.

Construction de l'ensemble de données final

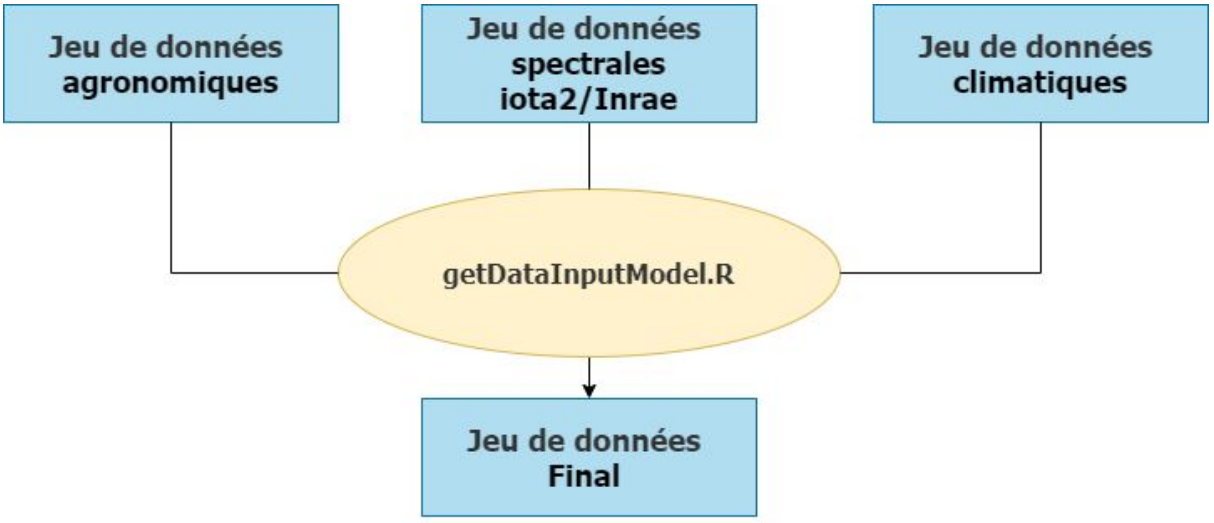


Fig. 5. Schéma général pour la construction de l'ensemble de données final

Les données climatologiques ont été fusionnées avec les informations agronomiques et spectrales correspondant à l'identifiant unique de chaque parcelle d'intérêt. À la fin du prétraitement, l'ensemble de données suivant a été obtenu :

Tableau 6. Composition finale de l'ensemble des données pour les modélisations

Nb de parcelles	Nb de variables	Nb d'observations
561	519 28 spectrales 491 climatiques	3033

1.2. Méthodes

1.2.1. Méthodes de classification utilisées

Lasso Multinomial (GLM)

En 1996, Tibshirani a développé le LASSO (Least Absolute Shrinkage and Selection Operator) qui est une méthode qui réduit à zéro le coefficient de régression des variables les moins impactantes. Associé à une validation croisée, il permet d'obtenir le niveau d'impact approprié et donc de faire une sélection de variables. L'idée est que la méthode LASSO minimise la somme des carrés résiduels pour lesquels la somme des estimations (coefficients) n'est pas supérieure à une certaine constante (Efendi et Ramadhan 2018). En d'autres termes, le LASSO limite l'estimation à moins d'une certaine constante (dans ce cas, nous utilisons lambda1se), de sorte que certaines estimations sont nulles.

Pour prédire les variables catégorielles multiples, l'utilisation du modèle logit multinomial dans l'analyse de régression pour les réponses de plusieurs catégories non ordonnées est la plus utilisée (Tutz, Pößnecker, et Uhlmann 2015). La régression multinomiale est une extension de la régression logistique binomiale. L'algorithme nous permet de prédire une variable catégorielle dépendante qui a plus de deux niveaux (Hosmer et Lemeshow 1989). Comme tout autre modèle de régression, la variable obtenue en sortie du modèle multinomial peut être prédite en utilisant une ou plusieurs variables indépendantes. Les variables indépendantes peuvent être nominales, ordinales ou continues.

Pour faire cette analyse multinomiale LASSO, nous utilisons le paquet ***glmnet*** pour ajuster le modèle de référence. Le modèle permet de déterminer les variables les plus importantes dans la classification des états phénologiques.

Régression logistique multinomiale (MLR) - Réseaux de neurones

Le MLR applique une transformation logarithmique non linéaire qui permet de calculer la probabilité d'occurrence d'un nombre quelconque de classes d'une variable dépendante sur la base de variables explicatives. Contrairement aux modèles de régression linéaire qui utilisent les moindres carrés comme critère, les coefficients du MLR sont généralement estimés en utilisant la probabilité maximale (Jeune et al. 2018).

Pour cette modélisation, nous utilisons le paquet ***nnet*** pour faire correspondre le modèle multinomial à un réseau de neurones.

Régression logistique ordinale (ORL)

Un des modèles statistiques les plus appropriés pour l'analyse des données avec une variable de réponse catégorielle est le modèle de régression logistique (Efendi et Ramadhan 2018). La régression logistique ordinale est une extension du modèle de régression logistique simple. Dans la régression logistique simple, la variable dépendante est catégorique et suit une distribution de Bernoulli. Dans la régression

logistique ordinale, la variable dépendante est ordinale, c'est-à-dire qu'il y a un ordre explicite dans les catégories (Ananth et Kleinbaum 1997).

Le modèle de régression logistique ordinal prend en compte l'ordre de la variable dépendante catégorielle en utilisant les événements cumulatifs pour le calcul du logarithme des probabilités (Ananth et Kleinbaum 1997). Cela signifie que, contrairement à la régression logistique simple, les modèles logistiques ordinaux considèrent la probabilité d'un événement et de tous les événements en dessous de l'événement focal en une hiérarchie ordonnée.

Dans cette étude de cas, une fois la variable catégorielle des états phénologiques ordonnée, la régression logistique ordinale a été utilisée pour prédire les états en fonction des variables indépendantes. Cela nous permettra de déterminer lesquelles de nos variables indépendantes (le cas échéant) ont un effet statistiquement significatif sur notre variable dépendante. Le paquet utilisé dans R était ***ordinal***.

Random Forest (RF)

Les "forêts aléatoires" (Random Forest) sont une combinaison d'arbres de décision. Dans cette méthode de classification, chaque arbre dépend des valeurs d'un vecteur aléatoire échantillonné indépendamment, avec la même distribution pour tous les arbres de la forêt (Breiman 2001). L'erreur de généralisation pour les forêts converge vers une limite à mesure que le nombre d'arbres dans la forêt augmente. L'erreur de généralisation d'un classificateur d'arbres forestiers dépend de la force des arbres individuels de la forêt et de la corrélation entre eux (Boulesteix et al. 2012).

Le Random Forest est un algorithme très intéressant pour la gestion des informations spectrales et le couplage avec d'autres variables (comme les variables climatiques, par exemple) (Muñoz et al. 2018). Il présente des caractéristiques telles qu'un fonctionnement efficace avec des jeux de données de grande taille, la capacité à identifier des relations non linéaires entre les prédicteurs et la réponse, et à traiter des variables prédictrices fortement corrélées (Kühnlein et al. 2014).

L'algorithme génère une estimation interne non biaisée de l'erreur de généralisation (erreur OOB) et a la capacité de déterminer quelles variables sont importantes dans la classification (Breiman 2001).

Les paquets utilisés dans R étaient ***RandomForest*** et ***Caret***. Dans la classification des stades phénologiques, le modèle du Random Forest a été paramétré avec *500 arbres*.

k-Nearest Neighbors (kNN)

L'algorithme de classification kNN est devenu une méthode importante dans l'exploration de données et le ML depuis qu'il a été proposé en 1967 (Deng et al. 2016). Pour appliquer la méthode traditionnelle kNN à de grands volumes de données, les méthodologies peuvent souvent être classées en deux parties, d'un côté trouver rapidement les échantillons les plus proches, ou sélectionner des échantillons représentatifs (ou éliminer certains échantillons) pour réduire l'estimation kNN (Zhu, Zhang, et Huang 2014).

Le k-NN est un algorithme de classification standard basé exclusivement sur le choix des mesures de classification. Il est "non-paramétrique". Seul le k, qui est le nombre de voisins à partir duquel les estimations sont établies, doit être fixé. K est une valeur entière spécifiée par l'utilisateur. Le choix optimal de la valeur dépend largement des données. En général, une valeur plus élevée supprime les effets du bruit, mais rend les résultats de la classification moins précis.

Dans cette étude de cas, l'algorithme a été utilisé dans R à partir du paquet **Caret**[11], en déterminant comme méthode de contrôle la validation croisée avec 10 plis (*10 folds*).

1.2.2. Détection de la floraison

Comme premier test, une première classification binaire de la phase de floraison a été effectuée. Nous avons utilisé un modèle basé sur la capacité prédictive des indices spectraux. La méthode utilisée était le Random Forest et elle a été ajustée pour les stades phénologiques regroupés en 2 classes. Les classes de floraison que nous opposons aux autres sont F1, F2, G1, G2, G3, G4 - Floraison toujours en cours.

1.2.3. Conditions de référence

Le modèle de référence est construit en considérant les états phénologiques en fonction des variables climatologiques et spectrales (figure 6 ci-après). L'ensemble de données spectrales utilisé est le résultat de la chaîne de traitement iota2 (première méthodologie d'extraction).

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

États ~ Climat + Spectral

Où :

Climat = Température minimale
Température moyenne
Température maximale
Précipitation
Evapotranspiration
Vitesse moyenne du vent
Rayonnement solaire
Degré jour de croissance (gdd)
Humidité relative

Spectral = 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)

Fig. 6. Modèle de référence

Les conditions de référence sont établies comme ligne de base pour évaluer et/ou améliorer la classification en fonction du couplage ou non d'autres variables thématiques. Cette base de référence est établie pour tester la variation d'une variable à la fois et non pas de toutes les combinaisons de variables.

Le problème de recherche est divisé en questions spécifiques dont on cherche les réponses en modifiant une variable à la fois à partir de ces conditions de référence. La sélection des conditions de référence est basée sur l'expérience de l'équipe de travail et sur le soutien universitaire :

- ✓ L'algorithme *Iota2* est une chaîne de traitement pour la production opérationnelle de cartes de l'occupation des sols à partir de séries temporelles d'images de télédétection en utilisant une classification supervisée(Inglada et al. 2016; Fauvel et al. 2020). Sa polyvalence et son niveau d'*accuracy* lui permettent d'être utilisé dans une variété de contextes.
- ✓ L'utilisation d'*indices spectraux* en agriculture a été l'une des méthodes d'analyse la plus populaire au cours des trois dernières décennies (Bolton et Friedl 2013). En particulier, les indices de végétation normalisés tels que le NDVI ont été largement utilisés en raison de leurs avantages interprétatifs pour améliorer la discrimination entre le sol et la végétation, réduisant l'effet

du relief sur la caractérisation spectrale des différentes couvertures terrestres (Islam et Bala 2008; Bolton et Friedl 2013).

- ✓ L'algorithme *Random Forest (RF)* est une méthode de classification moins sensible à la qualité des échantillons d'entraînement et au surajustement (par rapport à d'autres méthodes). Ces avantages sont dus au grand nombre d'arbres de décision produits par la sélection aléatoire d'un sous-ensemble d'échantillons de formation (Belgiu et Drăguţ 2016). De plus, il s'agit d'une méthode déjà utilisée par l'équipe de recherche dans laquelle ce travail est inscrit.
- ✓ La sélection des *variables climatiques et spectrales* dans le but de rendre le modèle reproductible à différentes échelles spatiales et dans différents lieux géographiques est une stratégie de généralisation pour la modélisation future.
- ✓ Les états phénologiques *regroupés en 8 classes* rendent la tâche de classification plus précise. Dans ce cas d'étude, l'imprécision des données in-situ et la limitation temporelle des informations spectrales et climatiques (une observation par semaine) ne permettent pas de distinguer correctement les 26 états. En définitive, l'intérêt agronomique de cette classification se concentre sur les états les plus représentatifs de la culture.

1.2.4. Comparaison des modèles

Nous avons ajusté différents modèles de classification pour les 8 états phénologiques groupés enregistrés (voir tableau 1). Nous comparons ensuite ces modèles avec le modèle de référence (conditions de référence).

Dans un premier temps, nous avons ajusté le modèle de référence à partir des quatre méthodes de classification sélectionnées pour cette étude de cas (Lasso multinomial, Régression logistique ordinale, Random Forest et K-Nearest Neighbor). Ensuite, nous avons évalué les quatre méthodes en fonction de l'*accuracy* et du temps de calcul. Enfin, nous avons sélectionné Random Forest.

L'idée était ensuite de créer des modèles qui cherchent à déterminer la pertinence et/ou l'importance des groupes de variables (spectrales, climatologiques et spatio-temporelles) pour l'identification des états phénologiques. Nous avons évalué le potentiel prédictif des variables thématiques de manière isolée, en considérant des modèles dans lesquels, à partir d'un seul groupe de variables, les états pouvaient être identifiés avec *précision*. Dans ce cas, les modèles suivants ont été utilisés :

États ~ Bandes spectrales

États ~ Indices spectraux

États ~ Tasseled Cap

États ~ Climat

États ~ Espace-Temps*

Fig. 7. Modèles individuels de classification. *Date de l'observation in-situ, département

Nous avons ensuite analysé le potentiel des indices spectraux avec les données interpolées sur dix jours (iota2) et les données non interpolées (INRAE). Enfin, nous avons couplé les variables spectrales, climatiques et spatio-temporelles pour déterminer le potentiel de classification global.

États ~ Indices spectraux + Espace-Temps

États ~ Climat + Espace-Temps

États ~ Indices spectraux + Climat + Espace-Temps

Fig. 8. Modèles couplés de classification

L'évaluation des différents modèles de classification a été réalisée sur la base de leurs matrices de confusion et des mesures suivantes :

Tableau 7. Mesures d'évaluation des modèles

Mesure	Formule	Concept
Average Accuracy **(Sokolova et Lapalme 2009)**	$\frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$ Pour chaque classe i, tp_i sont des vrais positifs, fp_i - des faux positifs, fn_i - des faux négatifs, et tn_i - des vrais négatifs, respectivement.	L'efficacité moyenne par classe d'un classificateur
Coefficient kappa de Cohen (McHugh 2012)	$k = \frac{(p_o - p_e)}{(1 - p_e)}$ p_o est la probabilité empirique d'accord sur l'étiquette attribuée à un échantillon (la proportion d'accord observé), et p_e est l'accord attendu lorsque les deux correcteurs attribuent des étiquettes au hasard. p_e est estimé en supposant une attribution aléatoire des étiquettes de classe.	Le coefficient kappa est un nombre compris entre -1 et 1. Les coefficients supérieurs à 0,8 sont généralement considérés comme un bon accord ; zéro ou moins signifie qu'il n'y a aucun accord (étiquettes pratiquement aléatoires).
Out-of-bag (OOB) error **(Hastie, Tibshirani, et Friedman 2009)**	Random Forest est entraîné en utilisant l'agrégation bootstrap, où chaque nouvel arbre est ajusté à partir d'un échantillon bootstrap des observations d'entraînement $Z_i = (x_i, y_i)$. L'erreur out-of-bag (OOB) est l'erreur moyenne calculée en utilisant les prédictions des arbres qui ne contiennent pas leur échantillon bootstrap respectif. Cela permet au Random Forest de s'ajuster et de se valider pendant l'entraînement.	

2. Résultats

Les étapes phénologiques de la **Vigicultures®** déterminées in situ sont établies sous forme d'étiquettes de classification. Les stades observés in-situ sont la variable dépendante à prédire. Les profils spectraux Sentinel-2 (S2) sont moyennés pour chacune des 561 parcelles étudiées. Dans la première section, nous effectuons une classification binaire (présence ou absence de fleurs) pour l'état de floraison avec la méthode Random Forest, en ne considérant que les informations spectrales. Dans la section 2.2, nous évaluons les cinq méthodes de classification sélectionnées en termes d'*accuracy* et de temps de calcul. Le modèle de référence est la base de cette évaluation. Dans la section 2.3, nous effectuons des classifications en tenant compte du couplage entre les variables spectrales, climatologiques et spatio-temporelles sur la base du modèle de référence. Nous évaluons le potentiel prédictif de chacun des modèles à partir des mesures résultant des matrices de confusion. Enfin, nous analysons l'impact de deux facteurs dans la classification : le regroupement des états phénologiques et la création d'un sous-ensemble de données équilibrées.

2.1. Classification binaire de l'état de floraison avec la méthode *Random Forest*

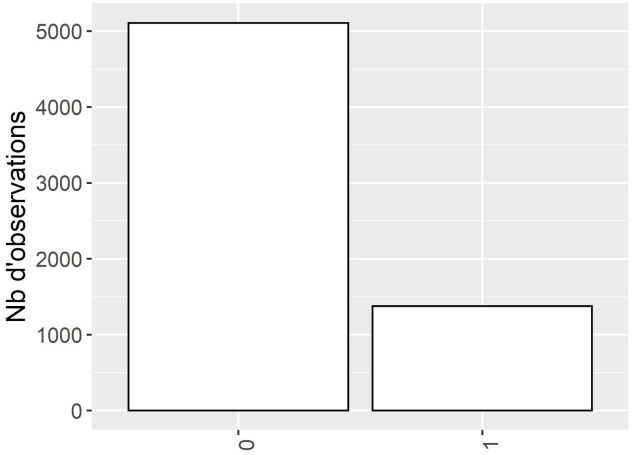
Modèle de floraison

Nous avons effectué une analyse préliminaire pour déterminer la capacité prédictive des variables spectrales (indices) à réaliser une classification binaire de l'état de floraison (présence ou absence de fleurs).

Floraison ~ Indices spectraux

Où :
Indices spectraux = 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. Modèle de Floraison



Selon le graphique, les données sont déséquilibrées. Sur 6494 observations, nous en avons 1376 (21%) au stade de la floraison et 5118 (79%) qui ne le sont pas. Ce déséquilibre dans les données est dû au fait que nous ne confrontons qu'un stade à tous les autres.

Fig. 10. Distribution des observations pour les classes binaires (Fleur - pas de fleur)

Les résultats du modèle de classification sont présentés ci-dessous :

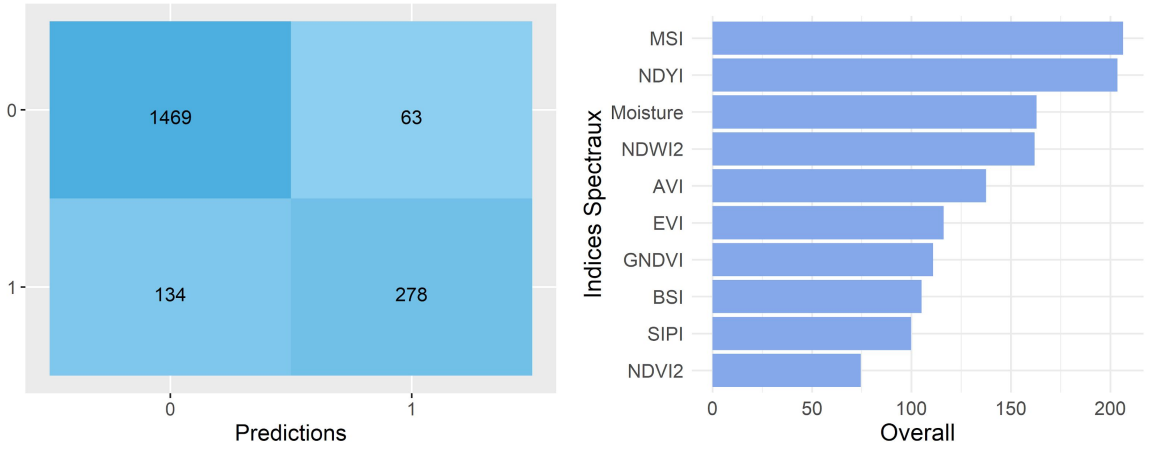


Fig 11. (à gauche) Matrice de confusion. (à droite) Importance des variables

La matrice de confusion nous montre la difficulté du modèle à déterminer correctement l'état de floraison lorsque les données sont déséquilibrées. Pour cet état, le taux de faux positifs (éléments inexactement classés comme fleuris) est important, cependant le modèle est correct dans 72,41% des cas pour la floraison (voir tableau 8 ci-dessous).

Tableau 8. Matrice de confusion binaire

	Prédictions	
Classes	0	1
0	95.52%	4.48%
1	27.59%	72.41%

Quant aux variables qui expliquent le mieux le modèle, les indices spectraux tels que l'indice de stress hydrique (MSI), l'indice Normalized Difference Yellow Index (NDYI) et l'indice Normalized Difference Water Index (NDWI) sont ceux qui expliquent le mieux la présence ou l'absence de fleurs dans les observations analysées.

Les paramètres d'évaluation du modèle nous montrent que pour les données de formation, l'OOB est inférieur à 10%. L'*accuracy* et le coefficient kappa sont respectivement de 0,91 et 0,71. Pour l'ensemble de validation, l'*accuracy* diminue de 1 % et le kappa de 3 %.

2.2. Classification multi-états

2.2.1. Stades phénologiques groupés (8 États)

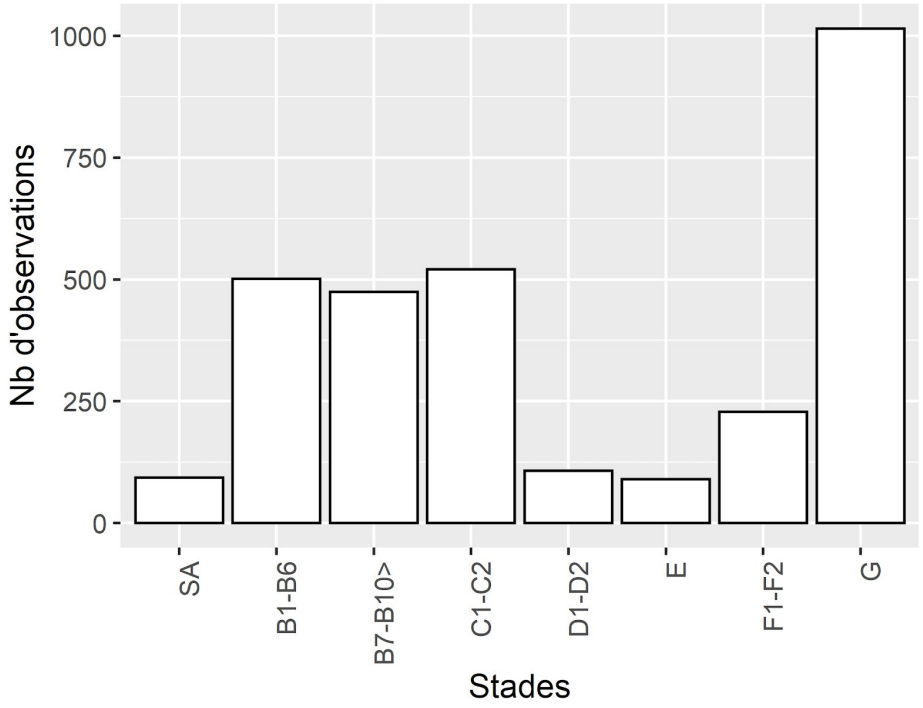


Fig. 12. Distribution des observations pour les stades phénologiques groupés

L'ensemble de données est déséquilibré et le nombre d'observations pour chaque état diffère de manière observable (fig. 12 ci-dessus). Cependant, les stades phénologiques moins représentatifs (SA, D1-D2 et E) font l'objet de plus de 90 observations chacun. Les stades tels que B1-B6, B7-B10> et C1-C2 sont plus représentés avec environ 500 observations. Pour le stade F1-F2, il y a environ 250 observations. Enfin, pour le stade final G recueille le plus d'observations de loin (plus de 1000). Comme l'observation des stades est extraite de **Vigicultures®**, on pourrait considérer que le

nombre élevé d'observations pour les stades finaux de développement du colza est dû à la présence d'un plus grand nombre de bioagresseurs dans cette période phénologique. Par conséquent, l'identification de ces stades est de grande importance pour notre problématique.

2.2.2. Modèles statistiques (comparaison des méthodes de classification)

Nous nous demandons si l'une des cinq méthodes de classification considérées pourrait mieux prédire les états phénologiques des observations in-situ. Pour cela, nous nous sommes sur les conditions de références (voir fig. 6, 3029 observations et 428 variables). Cet ensemble de données sera décomposé de manière aléatoire en deux ensembles le premier d'entraînement et le second de test, respectivement de 70% et 30% des observations. Les figures 13, 14, 16, 18 et 20 illustrent les matrices de confusion obtenues dans l'ensemble de données de test pour chacun des classificateurs.

Lasso Multinomial (GLM)

Nous avons utilisé le Lasso dans son mode multinomial et les résultats de l'ensemble de tests sont présentés sur la figure 13 ci-dessous. La matrice de confusion nous montre comment se détaille l'*accuracy* globale de 85%. On observe que les classes les mieux prédites par le modèle sont les classes B1-B6 (76,0%), B7-B10> (85,21%), C1-C2 (96,79%) et G (96,71%). Les erreurs entre les classes se produisent entre classes voisines dans le temps (l'état d'avant ou d'après l'état observé), à l'exception d'une observation classée F1-F2, alors que sa véritable classe est D1-D2. On observe également que plus le nombre d'observations est faible, plus les classes voisines ont tendance à être confondues, en revanche, les deux classes qui ont le plus grand nombre d'observations, G et C1-C2 sont les mieux prédites. La classe SA est confondue avec B1-B6 et dans le cas de D1-D2, le modèle la prédit comme C1-C2.

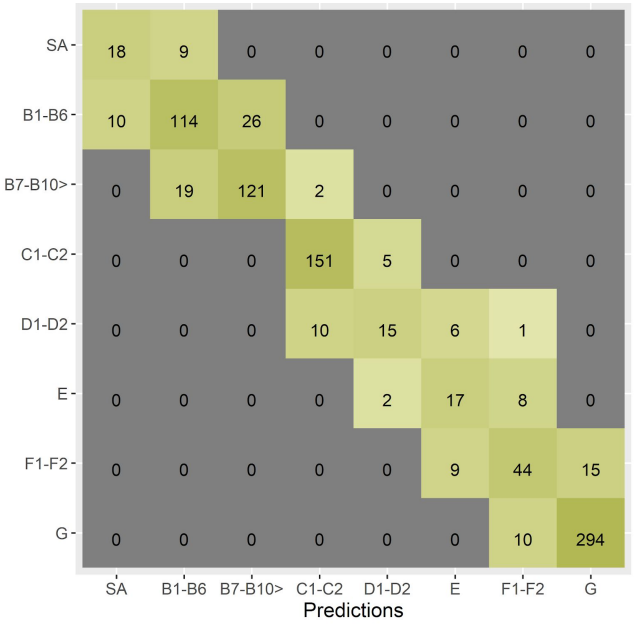


Fig. 13. Matrice de confusion pour le classificateur MLR - Lasso Multinomial pour l'ensemble de test

Le LASSO nous permet de déterminer le nombre de variables qui expliquent le modèle à partir du coefficient lambda en évitant le sur-ajustement (sélection de variables). Pour notre étude, la figure 14 indique la fréquence de sélection de chaque variable sur l'ensemble des 8 stades. Les précipitations (des premières semaines) et l'humidité relative (des dernières semaines) sont les variables les plus sélectionnées. Cependant, des indices spectraux tels que le GNDVI, le MSI et l'EVI sont aussi présents. Ces indices liés à la présence d'humidité et à la teneur en chlorophylle de la plante nous permettent de conclure que la réaction de la plante à certaines conditions hydriques définit adéquatement l'état phénologique de la plante.

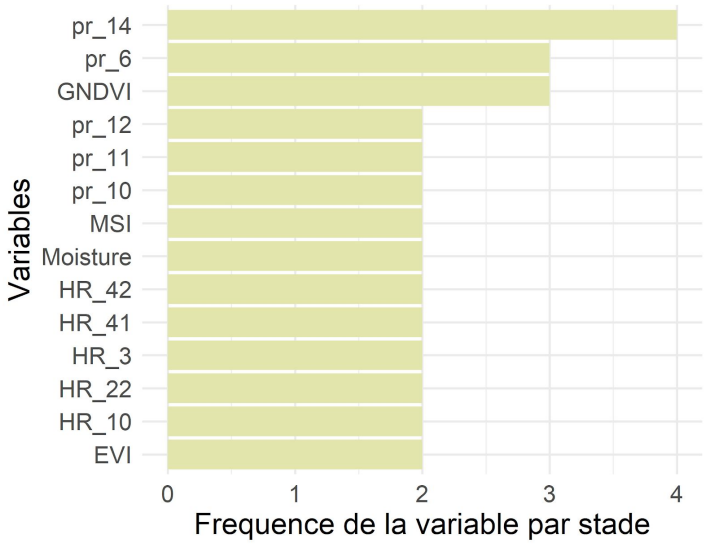


Fig. 14. Importance des variables explicatives du modèle MLR - Lasso Multinomial

Régression logistique ordinale (OLR)

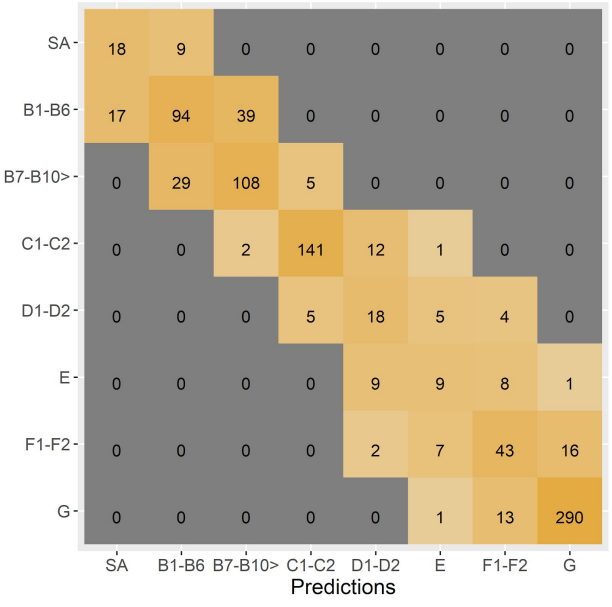


Fig. 15. Matrice de confusion pour le classificateur OLR dans l'ensemble de test

Nous avons décidé de classer les états à partir d'un modèle ordinal étant donné le caractère séquentiel des états phénologiques (un état précède l'autre). Les résultats de la matrice de confusion sont proches du modèle multinomial. L' *accuracy* par classe diminue pour les classes les mieux prédites par le classificateur précédent. Dans ce modèle, nous observons les *accuracy* suivantes : B1-B6 (62,67%), B7-B10> (76,06%), C1-C2 (90,38%) et G (95,39%). Cependant, dans les états où le nombre d'observations est faible (SA, D1-D2 et E), le modèle confond moins les classes voisines. Cependant, le modèle présente des erreurs plus graves car il intervertit es classes plus distantes que le modèle précédent : il classe les observations des états avec une distance interclasse de deux (D1-D2 comme F1-F2, par exemple). Si l'ordonnancement des catégories phénologiques aurait dû améliorer la précision, il est possible que le passage d'une régression LASSO à une régression logistique plus simple pénalise le modèle aboutissant à une *accuracy* (79%) nettement inférieur à celle du modèle précédent.

Multinomial Logistic Regression (MLR) - Réseaux de neurones

L'analyse de la matrice de confusion (fig. 16 ci-après), montre des différences par rapport à la modélisation LASSO multinomiale dans la reconnaissance des états analysés un par un, même si l'*accuracy* globale est encore acceptable (83%). On observe que les classes les mieux prédites par Lasso diminuent lorsqu'on utilise des réseaux de neurones (B1-B6 (68,67%), B7-B10> (83,10%), C1-C2 (95,51%) et G (96,71%)). Le modèle tend à confondre les classes plus facilement, même lorsqu'elles ne sont pas voisines. Cette difficulté rend le modèle, malgré sa bonne *accuracy*, moins efficace que le Lasso. Nous concluons que dans cette classification avec des données non équilibrées, les réseaux de neurones ajustent le modèle de manière proche même si légèrement inférieure, au modèle utilisé LASSO multinomial.

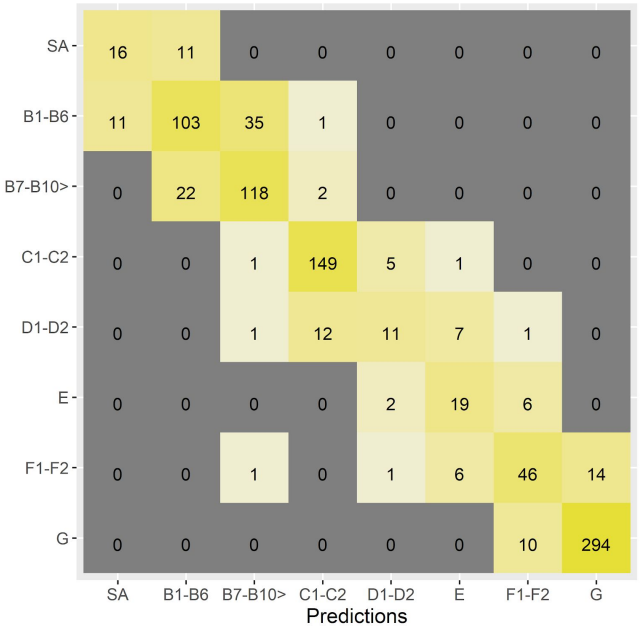


Fig. 16. Matrice de Confusion classificateur MLR - Lasso pour l'ensemble de test

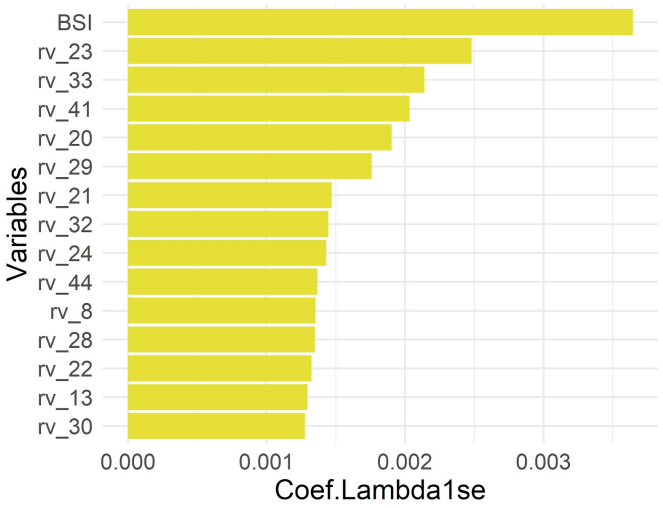


Fig. 17. Importance des variables explicatives du modèle MLR

Sur la figure 17 ci-dessus, on peut voir que les variables sélectionnées par le modèle pour classer les états phénologiques, sont le BSI (Bare Soil Index) où les bandes B2, B4, B8 et B11 sont concernées, ainsi que la variable climatique Rayonnement Solaire au milieu de l'année précédant la date d'observation in situ de l'état. Nous avons pu conclure que le modèle classifie en fonction des conditions d'absence et/ou de présence de la végétation (indice BSI) et de la réponse spectrale du colza à l'intensité du rayonnement solaire.

Random Forest (RF)

Nous avons utilisé un classificateur non linéaire pour déterminer si cette méthode représentait une amélioration de l'*accuracy* de la prédiction des états phénologiques. Les résultats sont présentés ci-dessous dans la matrice de confusion.

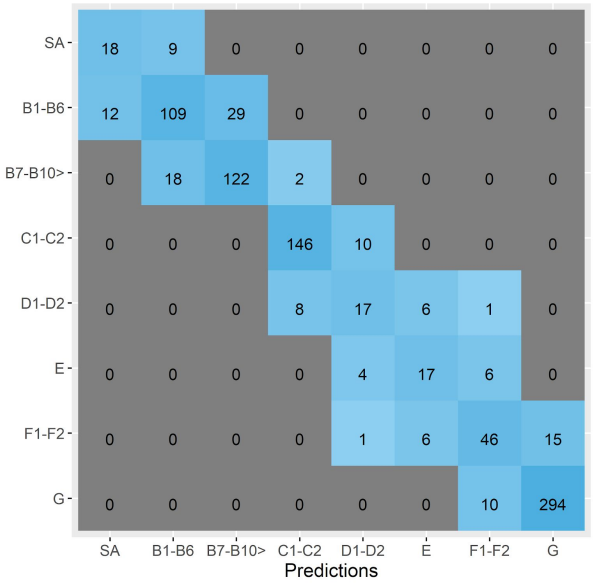


Fig. 18. Matrice de confusion pour le classificateur Random Forest dans l'ensemble de test

Nous constatons que les résultats sont comparables à ceux des méthodes linéaires ajustées ci-dessus. La similarité est grande avec la classification multinomiale par LASSO. Avec une *accuracy* globale de 84%, nous avons constaté que pour les classes où les observations sont peu nombreuses, le classificateur continue à confondre la classe cible avec ses voisines immédiates (SA, D1-D2, E et F1-F2) et il n'y a que 2 erreurs avec des classes non immédiatement voisines. Toutefois, pour les classes mieux identifiées, les résultats continuent d'être adéquats. Pour les états B1-B6 (74,00%), B7-B10> (87,32%), C1-C2 (93,59%) et G (96,71%). Nous concluons que le type d'approche (linéaire ou non linéaire), n'affecte pas de manière drastique les résultats de la classification.

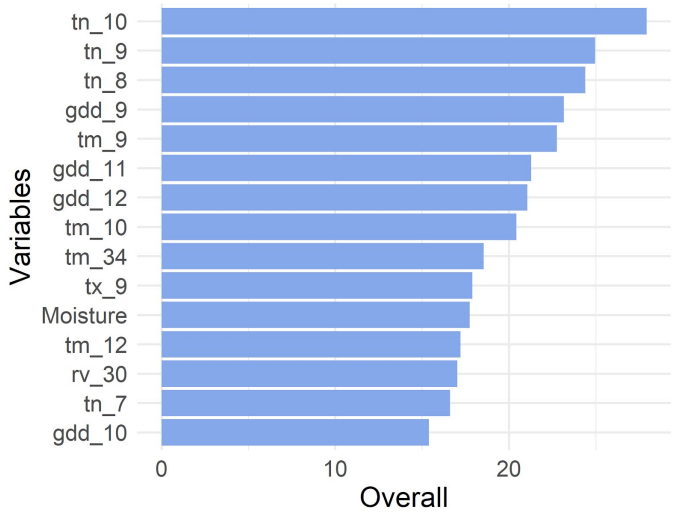


Fig. 19. Importance des variables explicatives du modèle RF

Lorsque l'on observe l'importance des variables sélectionnées par le modèle, on peut conclure que ce sont les variables climatiques qui déterminent la classification d'une observation dans un état ou un autre, les plus pertinentes dans la méthode Random Forest étant les températures (minimales et moyennes), le degré jour de croissance (gdd) de la fin du premier trimestre et le rayonnement solaire au milieu de l'année précédant l'observation in-situ. Des indices spectraux évaluant l'humidité du sol et le stress hydrique des plantes complètent le top 16 des variables les plus importantes pour la classification par ce modèle.

k-Nearest Neighbors (kNN)

Nous avons ajusté un modèle non paramétrique basé sur les distances (euclidiennes) afin de déterminer si l'*accuracy* des résultats de cette approche était comparable aux modèles précédents.

Dans la matrice de confusion de ce modèle (fig. 20 ci-dessous), nous continuons à voir des résultats proches de ceux des modèles précédents. L'*accuracy* globale était proche du Lasso Multinomial (83,4%), une seule observation classifiée à plus d'une classe de distance (D1-D2) et une classification très précise dans les états B1-B6 (71,33%), B7-B10 (82,39%), C1-C2 (94,87%) et G (95,72%) est une méthode intéressante pour l'identification des états phénologiques. Les États ayant peu d'observations

continuent à avoir un nombre important de faux positifs. Nous concluons qu'avec le choix d'un classificateur simple, on obtient des résultats similaires à ceux de modèles plus complexes.

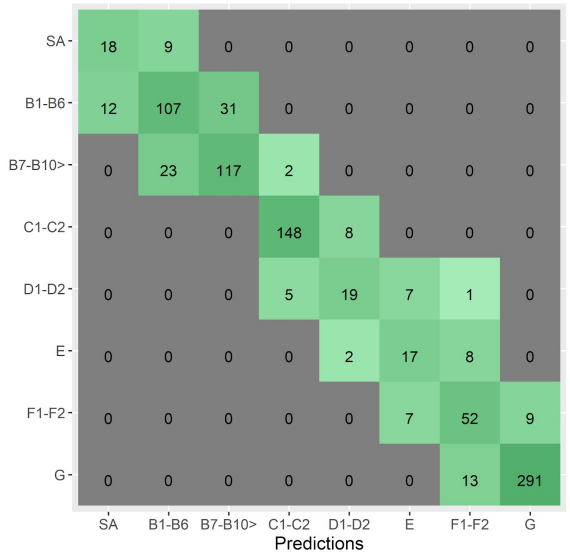


Fig. 20. Matrice de confusion pour le classificateur k-Nearest Neighbor dans l'ensemble de test

Enfin, nous pouvons conclure qu'en ajustant cinq modèles, chacun avec une approche différente, les classificateurs convergent vers des résultats proches. Les classes les mieux prédites étaient les classes C1-C2 et G. Elles présentent un bon nombre d'observations in-situ et de schémas climatiques et/ou spectraux qui permettent de les classer facilement sans obtenir d'erreurs importantes.

Dans le tableau suivant, nous observons en résumé les cinq classificateurs testés. si les valeurs d'accuracy sont similaires avec un maximum pour le Lasso Multinomial, les temps de calcul pour ce modèle plus complexe sont aussi beaucoup plus important que pour les autres. En revanche, le modèle Random Forest est celui qui a le temps d'estimation le plus faible tout en conservant une excellente capacité prédictive ce qui nous amène à confirmer notre choix du modèle Random Forest pour la suite de nos investigations sur le rôle des différentes variables explicatives.

Tableau 9. *Accuracy* et temps de calcul pour les modèles évalués

Méthode	**Accuracy**	**Temps de Calcul** (s)
Lasso - Multinomial	85.4%	1200
Multinomial Logistic Regression - réseaux de neurones	83,4%	26
Ordinal Logistic Regression	79,6%	60
Random Forest	84,2%	18
k-Nearest Neighbors	83,4%	60

2.3. Comparaison de modèles basés sur différents types de variables prédictives.

2.3.1. Pré-traitements des bandes Spectrales (sur extraction iota2)

Dans l'exercice de détermination des variables les plus significatives pour prédire le changement d'état phénologique du colza, nous avons voulu identifier si la classification à partir de différentes transformations de l'information spectrale (Bandes, Indices et *Tasseled Cap*) pouvait améliorer le modèle de référence. Nous avons comparé l'*accuracy* et l'OOB de chacune des transformations spectrales ainsi que le taux de réussite de la classification par classe dans les données d'entraînement.

Fig. 21. *Accuracy* et OOB de chaque modèle spectral (ensemble d'entraînement)

Dans la figure 21, pour le *Tasseled Cap*, une OOB de 0,42, une *accuracy* de 0,68 et un kappa de 0,59 le positionnent comme le moins performant. Pour les bandes spectrales et les indices, les mesures d'évaluation sont proches. Avec un OOB de 0,33, une *accuracy* et un kappa de 0,68 et 0,59 respectivement, le choix entre les indices et les bandes est réduit à des raisons pratiques, comme la facilité d'interprétation dans le cas des indices ou la simplicité de mise en oeuvre pour les bandes.
En comparant l'*accuracy* des modèles spectraux avec le modèle de référence (Baseline), on observe une différence de 16% pour les bandes et de 26% pour *Tasseled Cap*. Les variables spectrales classifient environ 70% des observations, cependant le modèle de référence (indices + climat) continue d'être le meilleur classificateur (84%). L'ajout de données climatiques aux données spectrales fournit un grand nombre d'informations.

En observant le tableau suivant, nous pouvons apprécier le pourcentage de réussite des modèles spectraux pour chaque état, dans l'ensemble de formation.

Tableau 10. Pourcentage d'occurrences de chaque modèle pour chaque état phénologique

Modèle	**SA**	**B1-B6**	**B7-B10**	**C1-C2**	**D1-D2**	**E**	**F1-F2**	**G**
Bandes	55.56%	67.01%	54.58%	62.19%	10.00%	29.17%	42.94%	85.69%
Indices	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%

Code couleurs : le jaune correspond aux meilleures classifications. Le bleu correspond aux secondes meilleures classifications.

Le modèle de référence offre une meilleure classification pour tous les états phénologiques sauf pour le premier. La deuxième place est en général atteinte par le modèle qui considère les 10 bandes spectrales. Le modèle des indices spectraux suit de près celui des bandes, cependant dans les états où le nombre d'observations est faible, il tend à confondre les états objectifs avec les classes voisines. Le modèle *Tasseled Cap* ne dépasse les deux précédents que dans l'état B1-B6. Il semble possible de classer les états exclusivement à partir d'informations spectrales mais il est important de considérer la contribution d'autres variables pour affiner la classification.

2.3.2. Focus sur les images récentes (iota2-inrae) - Méthodes d'extraction

Afin d'analyser l'impact de la méthodologie d'extraction des informations spectrales, nous avons effectué une classification à partir des indices spectraux pour les deux ensembles de données (iota2 et inrae).

L'*accuracy* des classifications de la méthodologie iota2 est de 0,68 par rapport à la méthodologie inrae qui est de 0,62. Comme pour l'OOB, iota2 identifie les classes avec une réduction de 5% par rapport à l'inrae.

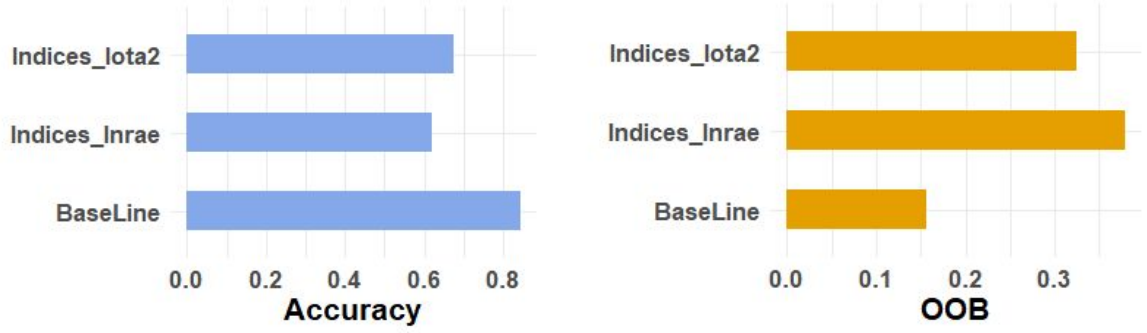


Fig. 22. *Accuracy* et OOB de chaque méthodologie d'extraction (ensemble d'entraînement)

Lorsque l'on compare les deux classifications en fonction du taux de réussite par classe, on constate que iota2 est le meilleur. Les classes B1-B6, B7-B10, C1-C2 et G,

qui présentent un nombre important d'observations (ensemble d'entraînement : 351, 332, 365 et 711 observations respectivement), sont les états les mieux prédits par le modèle. Dans les deux cas, le modèle ne trouve pas de patron pour classifier l'état D1-D2.

Pour les deux ensembles de données, l'état D1-D2 est confondu avec l'état C1-C2 (40 % des observations sont classées dans la classe précédente). Dans l'échelle BBCH, les deux états correspondent au développement des feuilles (rosette) et des organes végétatifs qui, étant si proches, sont difficiles à différencier uniquement avec des informations spectrales.

Tableau 11. Pourcentage d'occurrences de chaque modèle pour chaque état phénologique

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%
Indices_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%

Code couleurs : le jaune correspond aux meilleues classifications. Le bleu correspond aux secondes meilleures classifications.

Nous pouvons conclure que l'utilisation de la chaîne de traitement iota2 présente de meilleurs résultats en l'état actuel de la chaîne Inrae.

2.3.3. Variables climatiques vs. Variables Spatio-Temporelles

Pour déterminer si les variables climatiques sont plus prédictives que les variables spatio-temporelles, nous avons comparé les deux modèles en nous basant sur les conditions de référence.

Les graphiques montrent des résultats très proches. Avec une *accuracy* de 0,81 et un kappa de 0,77 pour les variables spatio-temporelles contre une *accuracy* de 0,82 et 0,78 pour les variables climatiques, la principale différence est que l'OOB est légèrement supérieure pour le classificateur *Date_Dep* (0,19 contre 0,18).

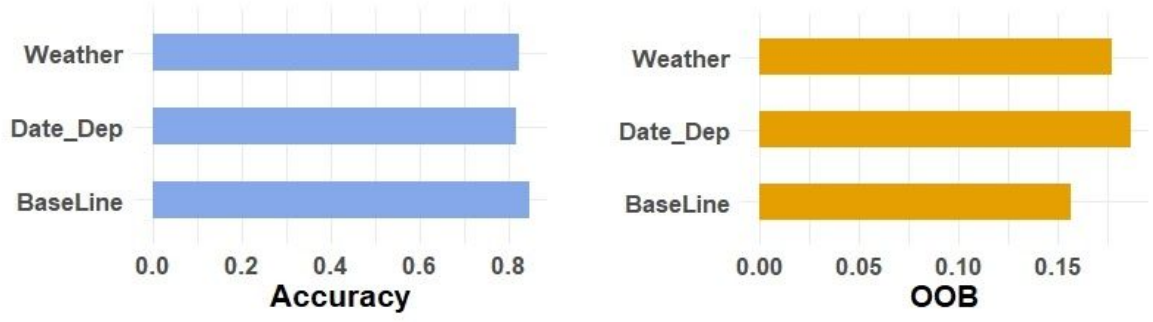


Fig 23. *Accuracy* et OOB des modèles basés sur les variables climatologiques et spatio-temporelles

Au niveau de l'*accuracy* par classe, les états B1-B6, B7-B10, C1-C2 et G présentent les meilleures prédictions.

Tableau 11. Pourcentage d'occurrences de chaque modèle pour chaque état phénologique

Modèle	**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%

Code couleurs : le jaune correspond aux meilleues classifications. Le bleu correspond aux secondes meilleures classifications.

L'analyse des variables climatiques et spatio-temporelles continue à être, en général, légèrement moins efficace que le modèle de référence au moment de la prédiction. Cependant, des états tels que C1-C2 et G sont mieux classés par des variables spatio-temporelles. Nous concluons qu'après le modèle de référence, ce sont les variables climatiques qui permettent le mieux de classer les états phénologiques du colza, mais la perte d'*accuracy* due à la non utilisation de l'information spectrale est faible. De même, l'ensemble des dates et des départements fournit une *accuracy* souvent comparable à celle des informations météorologiques, même si pour des stades précis et important comme celui de la floraison, il l'utilisation des variables climatiques induit une différence importante (54.37% - 71.88%)

2.3.4. Combinaison d'information de différentes variables thématiques

Nous nous interrogeons sur le fait que la combinaison de différentes variables thématiques dans un seul modèle puisse améliorer la classification des états phénologiques. Nous avons construit des combinaisons qui combinent deux variables thématiques et excluaient la troisième (climat + espace-temps et indices spectraux + espace-temps), pour finalement combiner les trois (climat + indices + espace-temps) et comparer leurs performances avec l'*accuracy* et les mesures OOB.

La figure 24 ci-dessous montre une *accuracy* assez proche entre les différents modèles. Les modèles dans lesquels nous avons utilisé les informations spatio-temporelles couplées aux variables spectrales ont obtenu une *accuracy* de 0,81, mais lorsque nous avons couplé les variables spatio-temporelles aux variables climatiques, l'*accuracy* a augmenté de 1%. En revanche, lorsque nous avons couplé les trois ensembles de variables thématiques dans un seul modèle (WIDD : Climate + Spectral Indices + Date + Département), nous avons obtenu une *accuracy* très proche de celle du modèle de référence mais avec une erreur de classification plus importante (0,156 contre 0,157).

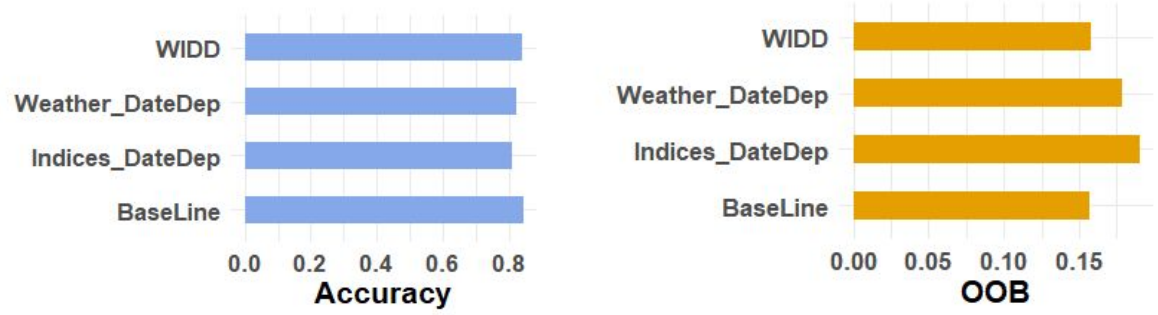


Fig 24. *Accuracy* et OOB des modèles basés sur les combinaisons de variables spectrales, climatologiques et spatio-temporelles

Enfin, en comparant le taux de réussite de la classification pour chacune des classes, nous observons que dans certains états phénologiques, il existe des modèles qui classent mieux ou que les résultats sont égaux au modèle de référence. Dans des classes telles que B7-B10 ou F1-F2, le modèle WIDD prédit les classes dans 82,53% et 71,88% des cas respectivement. En revanche, dans les classes telles que C1-C2, D1-D2, E et G, les meilleurs résultats sont répartis dans les trois modèles. Nous pouvons conclure que le couplage entre les différentes variables nous offre une amélioration de la prédiction des états individuels mais que le modèle choisi comme référence est toujours un bon modèle étant pour tous les stades soit le meilleur soit le deuxième meilleur modèle (et de peu).

Tableau 12. Pourcentage d'occurrences de chaque modèle pour chaque état phénologique

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%	50.80%	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%

Code couleurs : le jaune correspond aux meilleues classifications. Le bleu correspond aux secondes meilleures classifications.

De plus, nous avons vu que les variables climatiques ont le plus grand poids dans la classification des états phénologiques. Nous concluons que, bien que les modèles précédents offrent des résultats proches du modèle de référence, celui-ci est le plus polyvalent pour les classifications dans lesquelles on veut prédire sans dépendre des variables de temps et d'espace et ainsi étendre le spectre d'utilisation à d'autres lieux.

[12] WIDD = Weather + Indices +DateDep

2.4. Impact du regroupement et du rééchantillonnage

2.4.1. Impact du regroupement des états phénologiques (26 états)

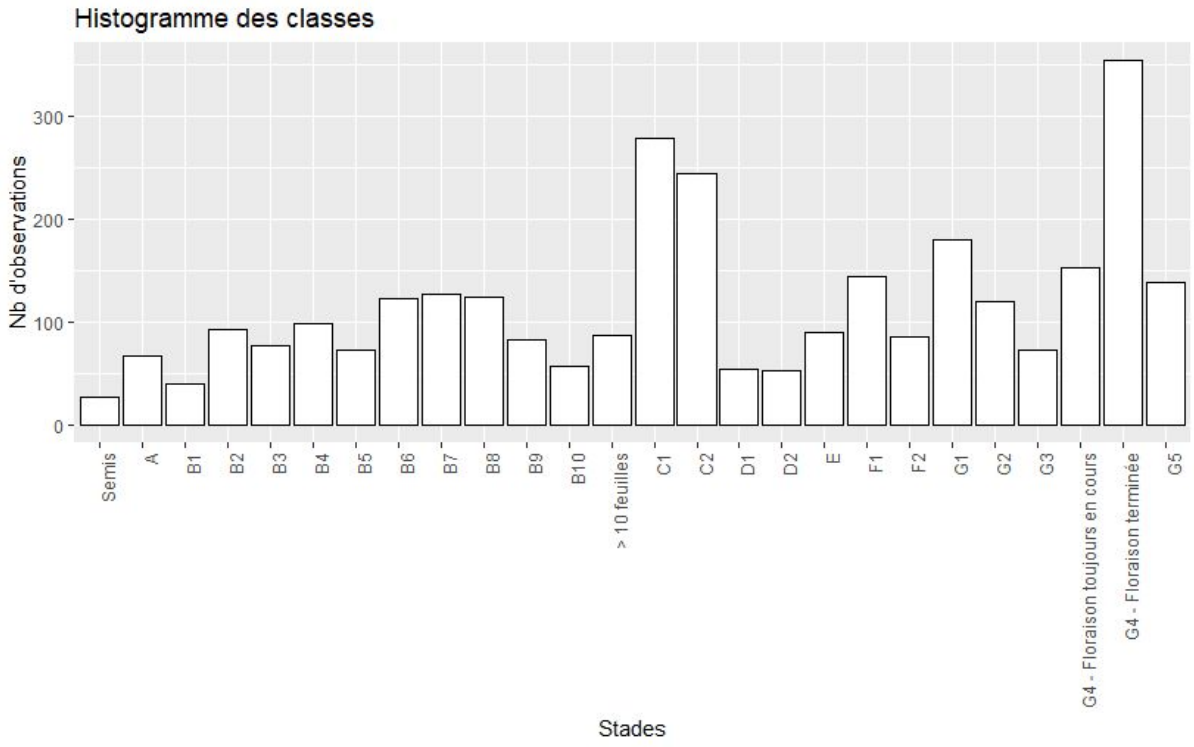


Fig 25. Distribution des observations pour les stades phénologiques non groupés

Lorsque nous examinons le nombre d'observations de chacun des états non groupés, nous constatons une forte variabilité. Les états phénologiques tels que C1, C2, G1 et G4-Floraison terminée sont les plus représentatifs. D'autre part, les états minoritaires tels que les Semis, B1, D1 et D2 avec un nombre d'observations inférieur à 50, présentent un grand défi pour les classificateurs utilisés. Nous nous interrogeons sur le fait qu'un ensemble de données fortement déséquilibré puisse être bien classé en utilisant les conditions de référence et la méthode du Random Forest.

La figure 26 ci-après montre la comparaison des matrices de confusion pour les états groupés et non groupés. On observe que pour les états initiaux (à gauche), le modèle confond la classe cible avec jusqu'à 8 classes différentes (état B3), cependant ces 8 classes sont toutes considérées comme voisines dans le modèle à 8 classes et les erreurs sont largement concentrées dans les classes les plus voisines. À partir de l'état C1, le nombre de vrais positifs augmente et la différenciation entre les classes est meilleure. La qualité de classification, rapportée au modèle à 8 classes peut même être améliorée ponctuellement. Par exemple, l'ensemble C1-C2 ne voit que 4 confusions avec D1-D2 au lieu de 10. L'ensemble D1-D2 a toujours 8 confusions avec

C1-C2 et augmente de 6 à 9 ses confusions avec E mais n'a aucune confusion avec le stade plus distant F1-F2. Le stade E n'admet que des confusions avec les sous-classes les plus proches (D2 et F1). L'ensemble F1-F2 n'a plus non plus de confusion avec le stade distant D1-D2. Des distinctions nettes à l'intérieur des classes regroupées sont aussi possibles dans certain cas, c'est notamment le cas du groupe très nombreux des observations G: L'opposition entre les trois premières classes de G et les trois dernières est particulièrement marquée. Dans l'ensemble, le regroupement des valeurs autour de la diagonale est marquant et laisse supposer qu'une estimation au niveau de la classe initiale Vigicultures resterait informative, surtout si le nombre encore faible d'observation par classe venait à être augmenté. Cependant, les paramètres d'évaluation chutent logiquement avec l'augmentation du nombre de classes. Avec 26 états phénologiques, l'*accuracy* globale du modèle est inférieure à 50% et l'erreur OOB est plus que triplée (0,15 contre 0,51).

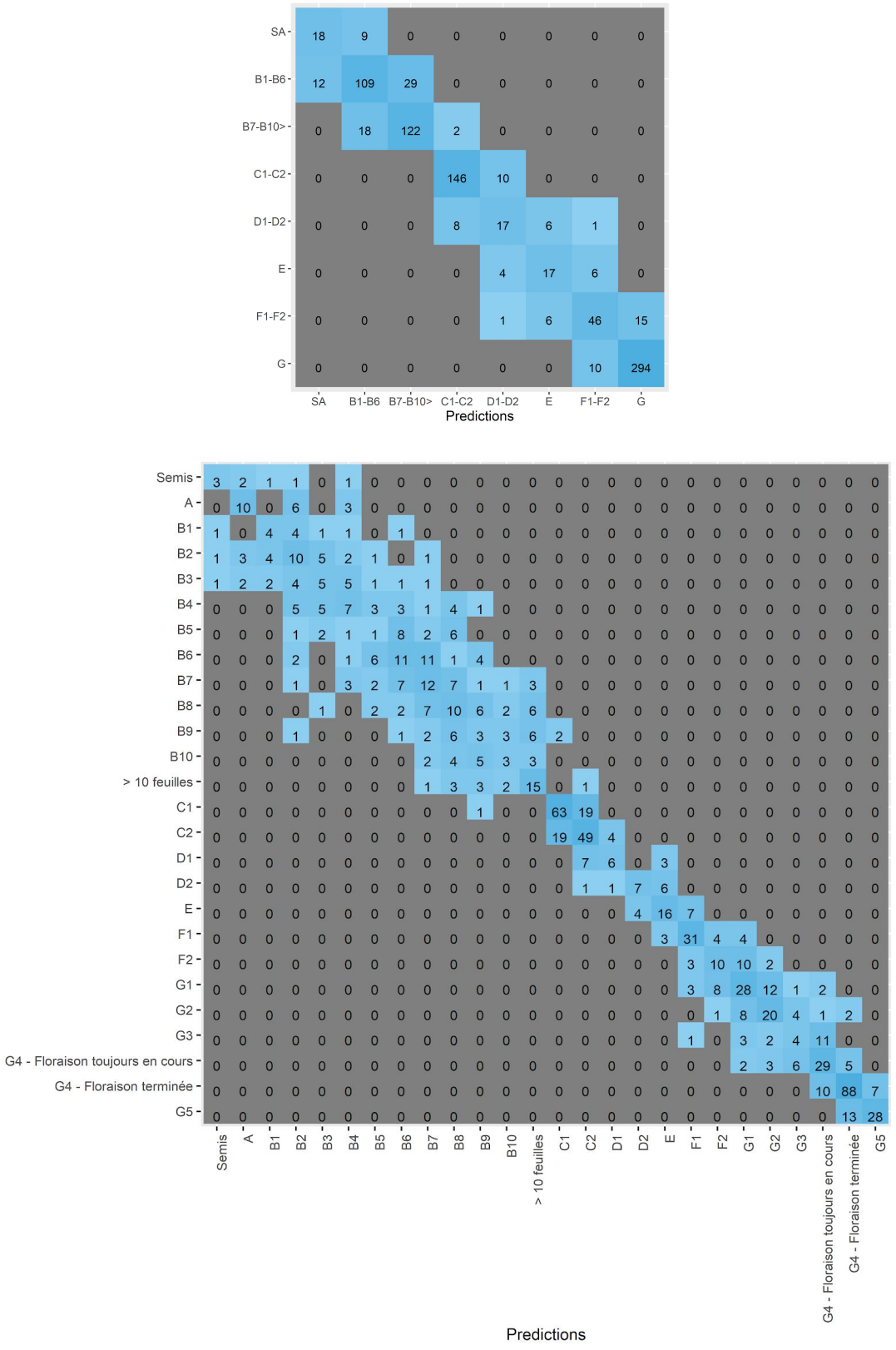


Fig 26. Matrices de confusion du modèle de référence avec les états phénologiques groupés (en haut) et les états non groupés (en bas)

Nous concluons que la stratégie de regroupement nous a permis d'avoir des résultats synthétiques probablement généralisables à un classement plus fin qui semble en partie possible en cas de besoin.

2.4.2. Modèle de référence avec stratégies de ré-échantillonnage

En raison de la difficulté que nous avons rencontrée pour prédire les classes où le nombre d'observations est considérablement plus faible, nous avons décidé d'évaluer les conditions de référence dans un ensemble de données équilibré en utilisant trois méthodes de rééchantillonnage. Nous avons d'abord équilibré l'ensemble des données en effectuant un processus de sous-échantillonnage. Pour cette méthode, nous retenons tous les cas de la classe minoritaire et choisissons au hasard un échantillon avec le même nombre de cas dans les classes majoritaires. Ensuite, nous équilibrons les données par un suréchantillonnage où nous laissons tous les cas de la classe majoritaire, et nous augmentons le nombre de cas dans les classes minoritaires par un échantillonnage avec remplacement. Enfin, nous utilisons la technique SMOTE[13] qui comprend à la fois le sur-échantillonnage et le sous-échantillonnage. Pour maintenir l'utilisation des ensembles d'entraînement/tests, nous l'avons appliquée séparément à chacun des deux ensembles.

Les résultats présentés à la figure 27 (ci-dessous) nous permettent de déterminer que le meilleur modèle est celui qui est équilibré et qui repose sur la méthode de l'échantillonnage ascendant. Avec une *accuracy* de 0,98% pour l'ensemble utilisé pour l'entraînement du modèle ce modèle augmente la performance de la classification des états phénologiques de 14% par rapport au modèle de référence. L'erreur de classification est réduite de façon drastique à une valeur de 0,017 par rapport à une valeur de 0,15 dans le modèle de référence. La technique hybride *SMOTE* a une *accuracy* de 0,88 améliorant de 4 % l'*accuracy* du modèle de référence ainsi qu'une OOB plus faible (0,11 contre 0,15). Cependant, l'utilisation de la technique de sous-échantillonnage pour équilibrer l'ensemble des données en réduisant le nombre d'observations réduira les prédictions.

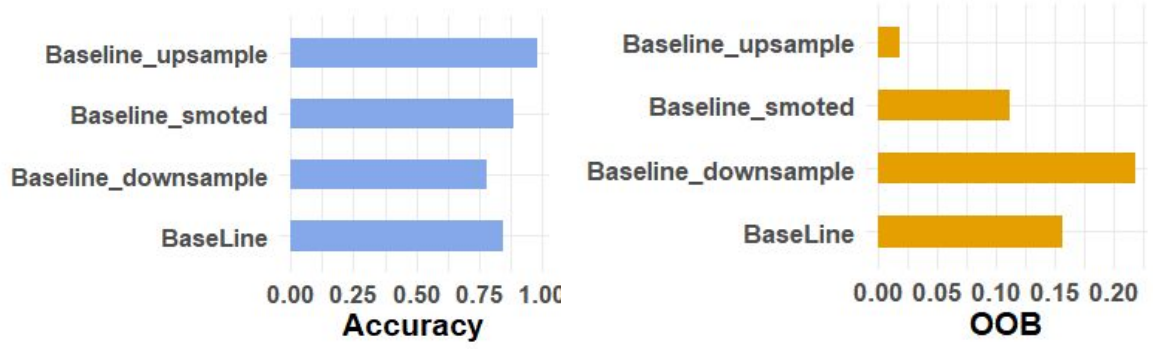


Fig 27. *Accuracy* et OOB du modèle de référence avec données équilibrées

[13] Synthetic Minority Oversampling Method

Nous pourrions conclure que l'utilisation de méthodes de rééchantillonnage pour équilibrer les données améliore considérablement la qualité de la classification avec un investissement minimum en temps de calcul. Cependant, lors de l'évaluation du modèle dans l'ensemble de test, nous avons constaté une diminution des paramètres d'évaluation. Pour la série d'entraînement, nous avons une *accuracy* de 0,98, un kappa de 0,98 et un OOB de 2 %. Cependant, pour l'ensemble de test, nous obtenons une *accuracy* de 0,73 et un kappa de 0,69 qui sont en fait inférieur à la stratégie sans ré-échantillonnage. La stratégie de ré-échantillonnage ne semble donc pas permettre *in fine* d'amélioration de la prédiction.

3. Discussion

La nature subjective des observations phénologiques terrestres a toujours été un problème dans l'étude récente de la phénologie (Czernecki, Nowosad, et Jabłońska 2018). Le développement de méthodes de classification pour identifier des modèles d'aide à la décision dans l'analyse du comportement de la végétation d'intérêt agricole est le pilier de l'analyse de cette problématique de recherche.

Dans l'intérêt de déterminer l'importance des variables spectrales, climatiques et spatio-temporelles dans l'identification des différents états phénologiques de cultures telles que le colza, nous avons évalué différentes hypothèses. Au départ, nous avons effectué une classification binaire de l'état de floraison dans laquelle nous avons identifié que les indices spectraux tels que le *MSI*, le *NDYI* et le *NDWI* sont des éléments fondamentaux pour la classification de cet état. Les paramètres d'évaluation sont adéquats, mais l'impact sur le déséquilibre des données rend la tâche de classification difficile. Ensuite, nous avons constaté que, bien qu'en évaluant cinq méthodes de classification, les résultats sont assez proches. Nous nous attendions à ce que le modèle *OLR*, qui prend en compte l'ordre hiérarchique des classes, soit le plus précis puisqu'il est le plus proche de la réalité (il trie les étiquettes dans un ordre d'occurrence), mais des méthodes telles que le *Random Forest* se sont révélées plus performantes. D'autre part, en étudiant les différentes possibilités de regroupement entre les variables thématiques, nous avons constaté que les variables météorologiques sont déterminantes pour la classification et que dans les situations où les observations in situ ne sont pas disponibles ou sont incohérentes, un couplage entre les indices climatiques et spectraux permet de prédire les états phénologiques avec une *accuracy* de 84% avec très peu d'erreurs impliquant des classes très différentes. Enfin, l'impact du regroupement des classes pour améliorer le succès de la classification, est un outil qui permet de hiérarchiser les états les plus importants à étudier. L'utilisation de techniques de rééchantillonnage des données améliore l'*accuracy* apparente du modèle de référence mais introduit un sur-ajustement du modèle qui se traduit par une différence d'*accuracy* entre l'ensemble d'entraînement et l'ensemble de test proche de 25 %.

Les paramètres phénologiques des images satellites multi-temporelles peuvent indiquer le développement de la croissance des cultures dans une grande région (Fisher et Mustard 2007; Zhong et al. 2011; Zhong, Gong, et Biging 2014; Li et al. 2014). En ce qui concerne spécifiquement la floraison, comme d'Andrimont et al. (2020), nous avons identifié que l'indice *NDYI* saisit l'augmentation de la coloration jaune des fleurs de colza dans la bande spectrale verte (B3). Le jaune des pétales de colza est dû à sa teneur en pigments caroténoïdes qui absorbent des longueurs d'onde de ~450 nm (Sulik et Long 2016). Les conditions d'humidité interne (chlorophylle ou capacité de rétention d'eau) et/ou externe (sol) sont également importantes dans la classification. Indices tels que NDWI et MSI où les bandes B8 (NIR), B8a (Red Edge 4) et B11 (SWIR 1) sont concernées. Le Random Forest est un algorithme approprié pour classer les états phénologiques individuellement,

cependant, des analyses ultérieures nous ont permis de généraliser le modèle et de l'appliquer à une classification multi-états.

Afin de mettre au point un outil d'aide au processus de prédiction des états phénologiques du colza, nous avons développé une approche basée sur un modèle de référence qui sélectionne des indices spectraux et des variables climatiques pour la prise de décision. Grâce à ce modèle de référence, les comparaisons visant à déterminer le meilleur classifieur sont facilitées. Les données in-situ pour notre étude de cas sont limitées, car l'obtention de données de terrain fiables et précises à une échelle appropriée est un effort difficile (Fisher y Mustard 2007). L'approche basée sur les classifications catégorielles présente des avantages lorsque la disponibilité des données sur la réalité du terrain est limitée (Zhong et al. 2011).

Par conséquent, la décision d'évaluer les classificateurs pour comparer leurs performances entre eux (benchmarking) permet une référence/orientation empirique pour sélectionner les classificateurs les plus appropriés pour des problèmes spécifiques (C. Zhang et al. 2017). De la même manière que Lorena et al. (2011), nous constatons que pour les études biogéographiques, le RF est une technique de modélisation prometteuse, en raison de ses performances élevées sur des ensembles de données composés d'un grand nombre de variables diverses et indépendantes. Cependant, d'autres modèles de régression multinomiale basés sur le Lasso ou des réseaux de neurones présentent une *accuracy* considérablement élevée. Le compromis entre efficacité et rapidité justifie en tout cas le choix par Inglada et al. (2016) du Random Forest comme algorithme de référence pour la classification de l'occupation du sol par iota2.

D'autre part, selon Zhang, Friedl, et Schaaf (2009), il a été démontré que les indices de végétation permettent de mieux décrire la croissance des cultures et améliorent considérablement l'*accuracy* de la classification des cultures. Cependant, dans notre cas, en analysant les différentes transformations des informations spectrales, nous avons constaté que les bandes spectrales individuelles sont des variables tout aussi intéressantes que les indices dans des algorithmes tels que Random Forest pour différencier une classe d'une autre parce que les arbres de décision permettent des combinaisons de variables qui peuvent être tout aussi intéressantes que celles fournies par les indices spectraux. En analysant la contribution de chaque variable explicative dans le modèle final (modèle de référence), on constate une influence faible à modérée de l'information spectrale du satellite. Au contraire, les caractéristiques météorologiques sont les plus prédictives dans le cas des stades phénologiques de l'automne, de l'hiver et du printemps, ce qui montre une relation avec la température de ces périodes. L'influence de la température sur la croissance des plantes est nettement plus importante au printemps, lorsque les plantes commencent leur cycle de développement après la pause hivernale (Pope et al. 2013; Springate et Kover 2014). Cependant, en raison de sa nature cyclique, l'information météorologique fournit également des informations sur la temporalité de l'observation, ce qui est très important pour l'identification de l'étape appropriée, comme l'indique l'efficacité des modèles avec des données spatio-temporelles.

Dans le cadre de notre analyse, le processus de classification ne prend pas explicitement en compte la dimension temporelle (approximation de l'ensemble de données en tant que série temporelle). Cependant, il existe des modèles qui intègrent la date d'observation comme variable explicative, ce qui, associé aux informations météorologiques, pourrait donner des résultats satisfaisants en termes d'*accuracy dans des* études ultérieures.

Concernant la méthodologie d'extraction des informations spectrales, lorsque l'échelle spatiale d'analyse est élevée, la présence de nuages et d'ombres sur les images satellites sont des situations à prendre en compte. Selon Inglada et al. (2015), les données interpolées tendent à réduire ces inconvénients. Pour cette raison, les résultats obtenus avec la méthodologie iota2 présentent de meilleurs résultats que la méthodologie inrae où l'on ne tient pas compte de l'utilisation du masque de nuage proposé par le produit theia. Cependant, la méthodologie inrae a l'avantage de pouvoir être utilisée en temps réel alors que iota2 permet ici une interpolation à partir de l'image suivante même si elle est prise beaucoup plus tard.

Bien que dans cette étude nous n'ayons pas effectué de regroupement statistique, la décision de regrouper les états phénologiques en fonction de l'expertise de l'équipe de recherche et de la comparaison avec l'échelle BBCH, nous a permis d'obtenir un modèle d'une *accuracy* adéquate. Les résultats des matrices de confusion montrent que plus le nombre de classes (états phénologiques) à prédire est élevé, plus la variabilité et la présence de valeurs aberrantes sont importantes, ce qui tend à diminuer l'efficacité de la tâche de classification, puisque les classes sont plus fréquemment confondues entre elles.

Malgré les lacunes de cette approche et les limites présentées par les données satellitaires dans la modélisation de la phénologie des plantes, cette approche pourrait encore être en mesure de donner une approximation fiable des observations traditionnelles au sol, notamment en ce qui concerne la fin de l'hiver (états B7-B10> et C1-C2) et le printemps (états F1-F2 et G). Cependant, la tendance à confondre les états voisins est une variable qui doit continuer à être analysée dans les travaux ultérieurs.

4. Limites et Difficultés

Dans l'exercice de résolution de la question de recherche, le temps est l'une des variables les plus conditionnantes, et pour cette raison, il n'a pas été possible d'envisager toutes les possibilités de couplage entre les variables thématiques. Il se peut qu'une combinaison non prise en compte permettra d'obtenir une meilleure *accuracy* que celles obtenues dans notre analyse.

D'autre part, selon l'analyse bibliographique, l'étude phénologique des cultures, dans la plupart des cas, est analysée comme une série temporelle. Nous avons pris le risque d'analyser le problème selon une approche différente pour établir le potentiel prédictif des variables climatiques et spectrales sans considérer explicitement la temporalité du phénomène. L'approche conventionnelle a été adoptée dans le cadre d'un autre stage réalisé à Toulouse au CESBIO.

La qualité de l'ensemble des données est une variable à considérer pour les analyses futures, bien que **Vigicultures®** nous fournisse des informations phénologiques, son principal objectif est le suivi épidémiologique des cultures. Si ces données ne sont pas spécifiquement orientées vers le suivi des stades et pourraient être imprécises, elles offrent une opportunité unique d'ajuster un modèle de prédiction à un grand nombre de champs répartis dans toute la France.

Nous avons aussi eu deux grands types de difficultés pendant le stage. : les difficultés du processus de recherche et les difficultés logistiques.

Les difficultés dans le processus de recherche sont principalement liées au traitement des informations spectrales. Dans la seconde méthodologie d'extraction (inrae, voir page 14), nous avons effectué les corrections atmosphériques mais nous n'avons pas pris en compte le masque nuageux, faute de temps nous n'avons pas pu ré-extraire et re-traiter les données, cependant cette tâche est la plus importante pour pouvoir comparer correctement les deux méthodologies (iota2/inrae) car l'intérêt de la méthodologie inrae est d'utiliser l'image satellite la plus récente. Cette approche dans laquelle nous réduisons le nombre d'observations pourrait présenter une autre difficulté en limitant la disponibilité des données satellitaires, mais elle pourrait améliorer la pertinence de l'information utilisée.
En ce qui concerne les difficultés logistiques, le lancement d'un processus d'apprentissage tel que celui de ce stage dans une situation de crise sanitaire mondiale (COVID-19) rend la tâche difficile à de nombreux égards, le plus important a été les difficultées administratives dans le cadre d'un confinement généralisé. Toutefois, une bonne communication et des efforts conjoints ont permis de résoudre les difficultés à temps. Cependant, cette situation atypique a permis d'ajuster les ressources internes de chacune des parties afin que le travail à distance devient une stratégie efficace pour l'apprentissage. Une autre difficulté logistique a été la casse (corruption matérielle) du disque dur où étaient stockées toutes les images satellites, ce qui a retardé d'une semaine les processus suivants, mais cela a permis de tester et confirmer l'efficacité de la chaîne de traitement.

Conclusion

L'intérêt de notre approche de classification réside dans le fait qu'une fois les stades phénologiques classifiés à partir du modèle de référence, nous sommes en mesure d'établir des relations entre la quantité de bioagresseurs et un stade phénologique donné, ce qui peut aider à identifier les conséquences sur le rendement final de la culture. Bien que l'étude ne soit pas en mesure de déterminer l'impact de ces relations, les résultats obtenus constituent une première étape importante pour continuer à développer les connaissances dans ce domaine.

L'étude et l'analyse des résultats obtenus nous ont permis de proposer un modèle de performance basé sur l'algorithme de Random Forest pour la classification des états phénologiques du colza à partir de variables météorologiques et spectrales. Bien que le modèle ne puisse pas pleinement remplacer les observations *in situ*, il peut aider au processus de prise de décision et réduire la dépendance à l'égard du travail sur le terrain pour obtenir des informations phénologiques, notamment lorsque sur des données d'archives sur les bioagresseurs et les rendements les dates de changement de stades phénologiques n'ont pas été identifiés mais que des images satellites sont disponibles.

Les perspectives de ce travail sont formulées sur trois fronts. D'abord, analyser l'effet d'un regroupement aléatoire d'états phénologiques sans tenir compte du regroupement des états de Vigicultures de la classification BBCH. L'idée d'effectuer un processus de regroupement automatique (classification non supervisée) pour regrouper les classes dans lesquelles le modèle a plus de difficultés à se différencier pourrait être une piste intéressante à explorer. Deuxièmement, lorsqu'il y a des classes surreprésentées, il y a un risque que le modèle en apprenne trop au détriment des classes moins représentées. Pour éviter ce problème, il est proposé d'approfondir la construction d'un ensemble de données équilibré où la composition dans chaque état est presque identique. Enfin, il est proposé d'adapter un modèle qui intègre plusieurs sous-modèles pour chaque état phénologique et d'évaluer ses résultats avec ceux obtenus jusqu'à présent.

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.

Annexe

Tableaux de résultats des modélisations avec les données d'entraînement et les données 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
Indices	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