L'Institut Agro Rennes-Angers
☐ **Site d'Angers X Site de Rennes**

Année universitaire : 2022.- 2023 Spécialité : Science des données	**Mémoire de fin d'études** X d'ingénieur de l'Institut Agro Rennes-Angers (Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement) ☐ de master de l'Institut Agro Rennes-Angers (Institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement) ☐ de l'Institut Agro Montpellier (étudiant arrivé en M2) ☐ d'un autre établissement (étudiant arrivé en M2)

Prédiction de la date de floraison de pommiers au printemps à partir d'images drones et de spectre NIRS en période de sénescence.

Par : Niels André

Soutenu à Rennes le 07/09/2023

Devant le jury composé de :

Présidents : Marie-Pierre Etienne (Enseignant-chercheur), Magalie Houée-Bigot (Enseignant-chercheur)

Maître de stage : Romain Fernandez, chercheur CIRAD (en distanciel)

Enseignant référent : Marie-Pierre Etienne (Enseignant-chercheur)

Les analyses et les conclusions de ce travail d'étudiant n'engagent que la responsabilité de son auteur et non celle de l'Institut Agro Rennes-Angers

Fiche de confidentialité et de diffusion du mémoire

Confidentialité

☒ Non ☐ Oui si oui : ☐ 1 an ☐ 5 ans ☐ 10 ans

Pendant toute la durée de confidentialité, aucune diffusion du mémoire n'est possible (1).

Date et signature du **maître de stage** (2) : 27 Juillet 2023 Evelyne Costes
(ou de l'étudiant-entrepreneur)

UMR 1334 AGAP Institut
Cirad - INRAE - L'Institut Agro Montpellier
Université de Montpellier
Amélioration Génétique et Adaptation des Plantes méditerranéennes et tropicales
CIRAD TA A-108/01 - Avenue Agropolis
34398 MONTPELLIER CEDEX 5 - FRANCE
http://umr-agap.cirad.fr

A la fin de la période de confidentialité, sa diffusion est soumise aux règles ci-dessous (droits d'auteur et autorisation de diffusion par l'enseignant à renseigner).

Droits d'auteur

L'auteur(3) **Nom Prénom** ---André Niels---

autorise la diffusion de son travail (immédiatement ou à la fin de la période de confidentialité)

☒ Oui ☐ Non

Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☒ la diffusion papier et électronique du mémoire (joindre dans ce cas la fiche de conformité du mémoire numérique et le contrat de diffusion)

(Facultatif) ☒ accepte de placer son mémoire sous licence Creative commons CC-By-Nc-Nd (voir Guide du mémoire Chap 1.4 page 6)

Date et signature de l'**auteur** : 27/07/2023

Autorisation de diffusion par le responsable de spécialisation ou son représentant

L'enseignant juge le mémoire de qualité suffisante pour être diffusé (immédiatement ou à la fin de la période de confidentialité)

☐ Oui ☐ Non

Si non, seul le titre du mémoire apparaîtra dans les bases de données.

Si oui, il autorise

☐ la diffusion papier du mémoire uniquement(4)

☐ la diffusion papier du mémoire et la diffusion électronique du résumé

☐ la diffusion papier et électronique du mémoire

Date et signature de l'**enseignant** :

(1) L'administration, les enseignants et les différents services de documentation de l'Institut Agro Rennes-Angers s'engagent à respecter cette confidentialité.
(2) Signature et cachet de l'organisme
(3).Auteur = étudiant qui réalise son mémoire de fin d'études
(4) La référence bibliographique (= Nom de l'auteur, titre du mémoire, année de soutenance, diplôme, spécialité et spécialisation/Option)) sera signalée dans les bases de données documentaires sans le résumé

Table des matières

Introduction..... 1
I. Contexte du stage..... 1
1. Les organismes d'accueil.....1
2. Le projet scientifique FruitFlow..... 2
3. Questions de recherche et objectif du stage.....2
II. Matériel et données.....3
1. Caractéristiques du verger.....3
2. Données drone..... 4
3. Données NIRS.....5
III. Méthodes et résultats : Exploitation des données.....6
1. Prédiction de la date de floraison à partir des données NIRS..... 6
a. Travail antérieur à mon arrivée..... 6
b. Concaténation des données des différentes dates.....7
c. Utilisation de l'outil Pinard..... 8
2. Prédiction de la date de floraison à partir des données drone.....9
a. Travail antérieur à mon arrivée..... 9
b. Visualisation temporelle des donnés drone..... 9
c. Segmentation des arbres..... 11
d. Prédiction de la date de floraison.....15
e. Vérification manuelle de l'hypothèse..... 16
3. Exploitation combinée NIRS/Drone..... 17
a. Simulation de données multispectrales via le spectre NIRS.....18
b. Prédiction du NDVI moyen à partir du spectre NIRS..... 19
c. Prédiction multimodale NIRS/drone..... 20
IV. Discussion..... 20
V. Conclusion et perspectives.....21
Bibliographie.....22

J’aimerais tout d’abord remercier Romain Fernandez, mon encadrant principal au cours de ce stage, qui m’a accompagné de manière quotidienne et m’a beaucoup appris sur la manipulation d’images.

J’aimerais aussi remercier Vincent Segura et Frédéric Boudon qui m’ont suivi de façon hebdomadaire dans mon avancée et m’ont aidé dans les directions à prendre ainsi que tous les encadrants qui m’ont suivi de façon mensuelle.

J’adresse mes remerciements à toute l’équipe Phenomen, qui a amené une bonne ambiance au courant de mon stage, par des pique-nique, des évènements et surtout durant les pauses café.

Merci aux stagiaires Phenomen, avec qui, pour la plupart, j’ai partagé le bureau. Merci à Pierre en particulier, avec qui j’ai partagé un bureau pendant 2 mois durant lesquels nous avons été assidus et nous sommes entraidés.

Enfin, merci à Marie-Pierre Etienne, qui était mon encadrante durant ce stage. Elle a pris le temps de se renseigner sur l’avancée de mon stage, sur les éventuels problèmes et m’a guidé dans la direction à prendre pour ce rapport.

Introduction

Le dérèglement climatique mondial dû aux activités humaines fait de la décennie passée la plus chaude depuis plus de 100 000 ans. Les modèles de prédiction des températures indiquent un réchauffement global de 1.5°C par rapport à l'ère pré-industrielle, qui sera atteint d'ici 2030-2035, peu importe les mesures de réduction d'émissions de CO2 appliquées. (**IPCC,** 2023).

Cette augmentation globale des températures et ce dérèglement climatique affectent la phénologie de nombreuses espèces végétales, qui fleurissent plus tôt, notamment en France (***J.Legave et al.***, 2008), ce qui expose ces fleurs aux périodes de gel (***Djordjevic B et al.***, 2018).

Dans une optique de pouvoir continuer à produire des pommes dans les différentes régions européennes avec leurs spécificités climatiques actuelles et à venir, les généticiens se doivent d'identifier les génotypes présentant des résistances aux stress environnementaux, dans le but de les sélectionner et de créer de nouvelles variétés adaptées aux différentes régions de production.

Malheureusement, notre capacité de sélection est limitée par nos connaissances des mécanismes génétiques et moléculaires impliqués dans la résistance aux stress environnementaux. Afin de mieux caractériser ces mécanismes et leur impact sur la production, il faut réussir à caractériser les moments clés du développement phénologique de l'arbre que sont la phase de pleine croissance végétative (au printemps) et la phase d'entrée en dormance (à l'automne). En effet, le projet FRUITFLOW formule l'hypothèse qu'il y a une relation entre l'état de l'arbre à l'automne et sa floraison au printemps suivant, et souhaite l'observer sur de nombreux génotypes. La caractérisation manuelle de ces deux phases phénologiques est donc très fastidieuse et demande une certaine expertise. Dans ce contexte, le projet FRUITFLOW s'intéresse à caractériser une population de diversité de pommiers comprenant 250 variétés, à haut débit à l'aide de différentes méthodes : l'imagerie aéroportée et la spectroscopie infrarouge (NIRS). L'imagerie est réalisée à l'aide d'un drone (UAV) équipé d'une caméra RGB et d'une caméra multispectrale (MS), ce qui permet de collecter rapidement une grande quantité de données avec peu de main-d'œuvre et à moindre coût par rapport à d'autres méthodes telles que l'utilisation d'autres véhicules aériens ou de satellites. Les données NIRS sont tirées des feuilles de ces mêmes arbres, qui ont été prélevées à quelques jours près de l'acquisition au drone, et analysées au laboratoire.

L'objectif de mon stage était d'utiliser et comparer ces deux types de données pour prédire la date de floraison des pommiers (années n) à partir de données acquises au cours de la saison de végétation précédente (année n-1), à l'aide de méthodes de machine learning et de deep learning.

I. Contexte du stage

Dans cette première partie je vais présenter les différents organismes d'accueil avec lesquels j'ai travaillé, le projet FruitFlow dans lequel s'inscrit mon stage, les différentes questions de recherche qui ont été posées avant mon arrivée, et l'objectif de mon stage.

1. Les organismes d'accueil

Les organismes d'accueil avec lesquels j'ai travaillé sont au nombre de 3, pour un total de 6 encadrants :

- L'équipe AFEF de l'UMR AGAP Institut qui a mandaté mon stage, m'a fourni les images drone et les données NIRS et m'a aidé dans mon analyse sur les données NIRS.
- L'équipe Phenomen de l'UMR AGAP Institut qui m'a accueilli dans ses locaux, m'a très bien intégré, m'a suivi tout au long de mon stage et m'a aidé dans mon approche de traitement d'image.
- L'équipe ICAR de l'UMR LIRMM qui a contribué aux consultations sur les techniques d'analyse d'image.

Le stage a été encadré par : Evelyne Costes (INRAE, directrice de recherche, AGAP-AFEF), Romain Fernandez (CIRAD, Chercheur AGAP-Phenomen) spécialisé dans le traitement et l'analyse d'image, Frédéric Boudon (CIRAD, Chercheur, AGAP-Phenomen) spécialisé dans la modélisation, le traitement et analyse de données phénotypage de plantes, Fernando Andrés-Lalaguna (INRAE, chargé de recherche, AGAP-AFEF), Vincent Segura (INRAE, chargé de recherche, AGAP-DAAV), ainsi qu'Emmanuel Faure (CNRS, Chercheur, LIRMM-ICAR) spécialisé dans le deep learning.

2. Le projet scientifique FruitFlow

La pomme est le 3e fruit le plus produit au monde par poids (***Beed F et al.,*** *2021*) après les pastèques et les bananes, ce qui en fait le 1er fruit produit dans des conditions climatiques tempérées.

Le cycle des espèces fruitières tempérées est synchronisé avec celui des saisons, qui, avec le dérèglement climatique, sont très changeantes : des hivers plus courts et moins froids, et des étés plus longs et plus chauds (**Wang et al.**, 2021).
Ces changements affectent la phénologie des plantes et donc leur date de floraison (***J.Legave et al.***, 2008), ce qui peut induire des problèmes de fécondation et donc diminuer la production de pommes (Jin-yong PU et al., 2008).

Dans le but de conserver une production répondant à la demande mondiale, il faut sélectionner des variétés de pommiers ayant de plus faibles besoins en froid (low chilling). Il est donc crucial de comprendre comment les différentes variétés de pommier réagissent aux changements de température, et c'est là un des objectifs du projet FRUITFLOW (FACCE-JPI - ERA-NET SusCrop 2, coordonné par Fernando Andres, équipe AFEF, AGAP Institut).

Le projet FRUITFLOW réalise ses recherches sur 2 axes :

- La caractérisation à haut débit des arbres fruitiers sur toute la saison de végétation, depuis la pleine croissance végétative au printemps jusqu'à l'entrée en dormance en automne.
- L'identification d'allèles et des marqueurs moléculaires associés aux caractères d'intérêt dans le but de les utiliser en programmes de sélection de nouvelles variétés adaptées au changement climatique.

Les méthodes de caractérisation à haut débit sélectionnées sont la NIRS (Near InfraRed Spectroscopy) et l'imagerie aéroportée (drone RGB/MS). Mon stage s'inscrit donc dans le premier axe du projet FRUITFLOW, pour prédire la date de floraison des pommiers à l'aide de ces deux types de données.

3. Questions de recherche et objectif du stage

A mon arrivée, des travaux sur le même projet avaient déjà eu lieu (un stage en 2022 mais seulement sur l'imagerie aéroportée et le travail d'une post-doctorante uniquement sur les données NIRS), tentant de répondre à la question biologique suivante :

- Peut-on prédire la date de floraison de pommiers au printemps à l'aide de données datant de la saison de végétation précédente ?

Les premiers résultats obtenus étaient très limités sur les données imagerie aéroportée et prometteurs sur les données NIRS. Différentes pistes d'améliorations étaient envisagées en sortie de ces travaux, en particulier la combinaison de ces 2 sources d'informations.

Pour mon stage, la problématique biologique a donc été découpée en 3 questions méthodologiques :

- Peut-on prédire la date de floraison à l'aide de données NIRS ?
- Peut-on prédire la date de floraison à l'aide de données imagerie aéroportée (drone) ?
- Peut-on coupler ces deux types de données afin d'améliorer la prédiction ?

Pour répondre à ces différentes questions, l'objectif général de mon stage était donc de prédire les dates de floraison des pommiers à l'aide des données NIRS et des données drones.

II. Matériel et données

Dans cette partie nous allons présenter de quelle manière les données ont été récoltées, avec quels outils et quelles sont leurs caractéristiques.

1. Caractéristiques du verger

Les données ont été récoltées sur un verger implanté à l'Unité Expérimentale Diascope d'INRAE située à Mauguio (34130; 43.611433, 3.978552), près de Montpellier (Figure 1). Ce verger est constitué d'une "Core collection" de pommiers maximisant la diversité génétique avec un minimum d'individus. Dans le cas présent, cette collection est constituée exclusivement de variétés cultivées de pommes à couteau plantées pour la plupart en 2014-2015 (Lassois et al., 2016).

Il est constitué de 250 variétés, avec quatre arbres pour chaque variété, pour un total de 1100 pommiers répartis sur 11 rangs de 100 individus. L'espacement inter-rang est de 5 mètres et l'espacement intra-rang est de 2 mètres.

Ces individus (plantés en 2014-2015) ne sont pas tous intégrés à l'étude. Les individus non considérés sont :

- Ceux morts avant l'hiver 2021 ou qui n'ont pas fleuri au printemps 2022
- Les individus hors-type (phénotype différent de celui attendu pour ce génotype)

- Les pommiers récemment plantés (le 11e rang ainsi que des remplacements des arbres morts)
- Les pommiers qui n'apparaissent pas dans une capture de drone (pour les données drone)
- Les pommiers qui n'ont pas de spectre NIRS suite à un problème en laboratoire ou de récolte (pour les données NIRS).

Les effectifs exacts seront détaillés dans les 2 sous-parties suivantes.

Figure 1 : *Emplacement du verger à l'UE Diascope, Mauguio, France vu du ciel (Source : Google Earth).*

2. Données drone

Le drone utilisé pour l'acquisition des données est un drone de la marque Mikrokopter. Pour l'étude, la vitesse a été régulée entre 10 et 14 km/h à 24 mètres du sol. Le drone a capturé environ 500 clichés par prise, tous géoréférencés en WGS84 (longitude, latitude, altitude) grâce au GPS interne du drone. L'ensemble des vols a eu lieu entre 11h et 12h pour essayer d'avoir une luminosité uniforme tout au long de la séquence d'acquisition, pour tous les pommiers. 674 individus sont intégrés à notre étude, avec 214 génotypes représentés et une moyenne de 3,15 individus par génotype.

Pour l'expérimentation menée en 2021, il y a eu 4 périodes d'acquisition : juin (développement végétatif maximal) , septembre, octobre et novembre (entrée en dormance). Pour chacune des acquisitions, le drone a fait 2 vols. Lors de chaque vol, il embarquait l'une des deux caméras utilisées pour l'étude (RGB et MS pour multispectral). Sur le sol étaient disposés des GCP (Ground Control Point), des objets de forme circulaire géoréférencés au centimètre près. Ces points de contrôle facilitent et rendent plus précis le géoréférencement des images lors de la reconstruction d'une orthomosaïque assemblant toutes les images, comparé à l'utilisation simple du GPS embarqué sur le drone.
Les orthomosaïques géoréférencées sont reconstruites à l'aide du logiciel Agisoft Metashape. Deux orthomosaïques, une RGB et une MS sont créées pour chaque mois.

La caméra RGB est une Sony Alpha 5100 à 3 capteurs, d'une résolution de 6000x4000 pixels tandis que la caméra MS (multispectrale) est une AirPhen V3 à 6 capteurs, d'une résolution de 1280x960 pixels. Ses 6 capteurs ont respectivement leur bande passante dans : le bleu (450 nm), le vert (530 et 570 nm) , le rouge (675 nm) , le rouge lointain (730 nm) et le proche infrarouge (850 nm). Les images RGB correspondent à des photographies "standards". Les images multispectrales sont fréquemment utilisées dans la littérature, en

particulier pour calculer des indices de végétation, notamment le NDVI (Normalized Difference Vegetation Index) et le NDRE (Normalized Difference Reg Edge index) (Équations 1a et 1b), utilisés entre autres par ***Davidson C et al.*** *(2022)* dans un contexte similaire.

Équations 1a et 1b : $NDVI = \frac{(PIR - R)}{(PIR + R)}$ $NDRE = \frac{(PIR - RE)}{(PIR + RE)}$

Dans ces équations, “PIR” désigne le proche infrarouge (850 nm), “R” le rouge (675 nm) et “RE” le rouge lointain (730 nm). Les valeurs de réflectance étant strictement positives, le NDVI et le NDRE sont compris entre -1 et 1.

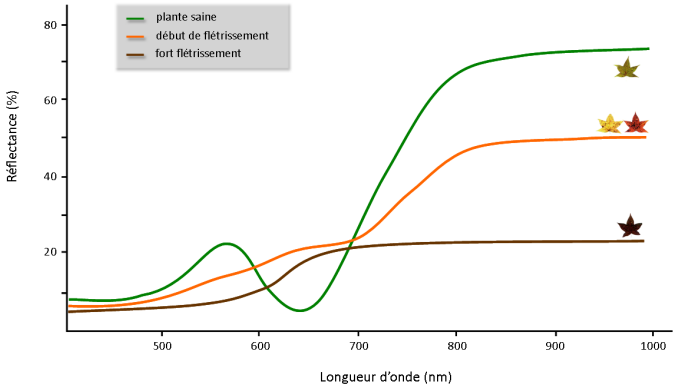


Figure 2 : *Courbes de réflectance représentatives d'une plante saine, en cours de flétrissement ou complètement flétrie (Source : UVED).*

Ces indices de végétation sont utilisés dans la littérature pour caractériser la vigueur des plantes, car l'activité photosynthétique se traduit par une absorption de la lumière dans le domaine du rouge. La sénescence induit une augmentation de la réflectance des feuilles dans la gamme des rouges, et une diminution des valeurs de NDVI et NDRE, qui sont de l'ordre de 0.4 à 0.6 pour une feuille saine, et diminuent jusqu'à atteindre 0 ou des valeurs négatives pour une feuille qui n'est plus “en fonctionnement”. Cette caractéristique nous intéresse pour caractériser la sénescence des arbres, et essayer de la relier au jour de floraison.

3. Données NIRS

Les données NIRS sont issues, pour chaque arbre, de 16 disques foliaires prélevés sur 16 feuilles (1 disque par feuille), ce sont des prélèvements d'un disque de surface foliaire de 1,3cm de diamètre. Les disques foliaires ont été récoltés aux mêmes périodes que les acquisitions de drone. Ces feuilles ont été sélectionnées pour ne pas avoir été malade, ne pas être fripées ou venir d'un rejet du porte-greffe. De plus, elles ont été prises de façon à être représentatives de l'arbre, donc 2 feuilles en haut, 4 feuilles à mi-hauteur et 2 feuilles en bas de l'arbre de chaque côté de celui-ci.
Les disques foliaires ont ensuite été réunis, directement congelés dans l'azote liquide sur le terrain, puis lyophilisés pendant 72h avant d'être broyées et passées au spectromètre pour obtenir un spectre NIRS par arbre, par date (4 dates).

Un exemple des spectres obtenus est donné en Figure 3 :

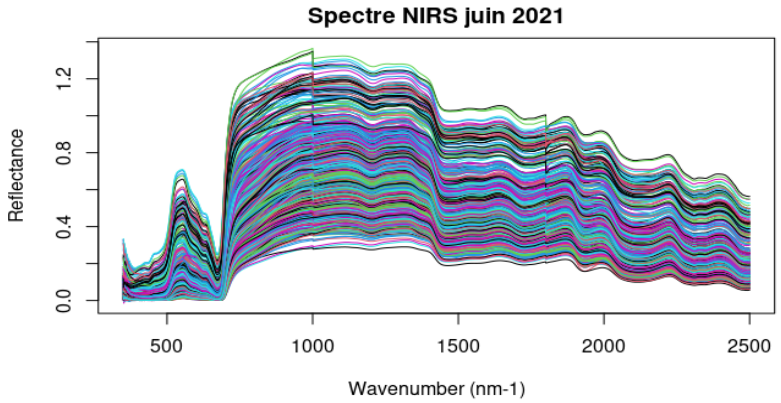


Figure 3 : *Spectres NIRS des 702 individus du mois de juin 2021.*

Chaque date n'a pas le même nombre d'individus, les individus ayant des spectres NIRS pour chacune des dates et ayant fleuri sont au nombre de 534, avec 212 génotypes représentés et une moyenne de 2,52 individus par génotype.

En résultat de cette collecte de données, nous disposons donc de l'ensemble de données suivant:

- Les images des pommiers récoltées en juin, septembre, octobre et novembre 2021 (n = 674)
- De spectres NIRS des pommiers récoltés aux mêmes dates (n = 534)
- Des dates de floraison des pommiers au printemps 2022

III. Méthodes et résultats : Exploitation des données

Dans cette partie, nous allons aborder comment les données avaient déjà été exploitées avant mon arrivée, comment j'ai construit de nouvelles démarches pour leur exploitation, les traitements que j'ai réalisés au niveau imagerie et comment j'ai réussi à améliorer la prédiction NIRS.

Lors de nos prédictions, nous utiliserons la MSE (Mean Square Error ; Équation 2a) et sa racine carrée, la RMSE. La MSE est utilisée pour calculer le pourcentage de variance capturée (Équation 2b), en comparant la MSE de nos prédictions avec la variance du jeu de données à prédire. La RMSE est utilisée pour évaluer l'erreur moyenne de prédiction, en jours.

Équations 2a et 2b : $MSE = 1\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad Var_{capt} = (1 - \frac{MSE}{var}) * 100$

Où n désigne le nombre d'individus, Y_i les valeurs observée et $\hat{Y}_i$ les valeurs prédites. "var" est la variance de la variable à prédire (le jour de floraison), mesurée dans le jeu de données complet. La variance capturée est utilisée comme indicateur de performance pour évaluer nos pipelines de prédiction du jour de floraison, le but étant de maximiser le pourcentage de variance capturée.

1. Prédiction de la date de floraison à partir des données NIRS

a. Travail antérieur à mon arrivée

Avant mon arrivée au Cirad, les données NIRS avaient déjà été analysées par Vincent Segura et Amy Watson (post-doc de l'équipe AFEF), qui avaient réalisé une régression par moindres carrés partiels (PLSR : Partial Least Squares Regression) sur R. Ils avaient en amont procédé à différents traitement des spectres :

- Ajusté les sauts liés au changement de détecteur lors de l'acquisition du spectre
- Retiré quelques individus ayant un spectre anormal (outlier)
- Retiré l'extrémité bruitée du signal (les 50 premiers nm)

Une fois cela réalisé, ils créaient différentes modalités de spectres prétraités pour extraire au total 5 "features" :

- Le spectre brut
- Le spectre normalisé
- Le spectre "detrend" (permet d'enlever la tendance sous-jacente au spectre)
- La dérivée première
- La dérivée seconde

Cela permettait, pour chaque date, de réaliser une régression PLS. La PLSR utilise une combinaison linéaire des variables explicatives tout en réduisant leur dimensionnalité. Cette réduction de dimensionnalité se fait en identifiant des composantes latentes qui sont des combinaisons linéaires des prédicteurs originaux. Ces composantes latentes sont construites de sorte à maximiser la covariance entre les variables explicatives et la variable réponse. La PLS est une méthode reconnue et validée par la communauté scientifique dans le cadre de l'analyse de spectres NIRS (**Cheng J-H et al.,** 2017).

La PLSR réalisée analysait chacune des features indépendamment des autres et qui (après cross validation à 5 folds) retournait 2 sorties : la feature qui a permis d'obtenir la meilleure RMSE et la RMSE correspondant. Le meilleur résultat était obtenu avec la dérivée première du spectre du mois de juin. La RMSE était de 7,92, ce qui correspond à 59,79% de variance capturée avec 702 individus.

Les données des différentes dates n'avaient pas été analysées avec les mêmes individus (nombre d'individus différent pour chaque date). La question se posait donc d'homogénéiser les individus de différentes dates pour pouvoir les comparer, mais aussi de tenter de concaténer les dates pour tenter d'améliorer les résultats prédictifs.

b. Concaténation des données des différentes dates

Pour plus de facilité lors de la manipulation des données et du code, j'ai décidé de passer au langage Python et d'analyser uniquement le spectre brut. J'ai codé une première fonction me retournant le nombre de composantes optimales pour la PLS, en testant de 1 à k composantes (k choisi), s'arrêtant si il n'y a pas d'amélioration de la RMSE depuis plus de m composantes (early stopping avec m au choix). J'ai aussi réalisé une seconde fonction me permettant de prédire avec plus de répétitions. Elle retourne une RMSE moyen et une variance capturée moyenne plus représentatives (car plus de répétitions) ainsi que le graphique représentant les valeurs réelles en fonction des valeurs prédites.

J'ai comparé les différentes dates entre elles afin de voir l'évolution de la RMSE au fil du

temps. Les mesures obtenues pour chacune des dates sont présentées dans le Tableau 1 (avec le même nombre d'individus n= 534) :

Jour observation	Juin	Septembre	Octobre	Novembre
RMSE	**7.739**	**7.643**	**8.246**	**8.894**
Variance capturée	**51%**	**52,2%**	**44,4%**	**35,3%**

Tableau 1 : RMSE et Variance capturée pour chaque mois.

On observe (Tableau 1) globalement une diminution de la variance capturée (sauf en septembre) au cours des mois. Nous nous sommes posé la question suivante : **Malgré cette diminution d'information prédictive au cours des mois, y a t'il de l'information supplémentaire dans ceux-ci, qui pourrait contribuer à améliorer la prédiction ?**

Dans le but de répondre à cette question, nous avons décidé de concaténer les spectres et de réaliser la PLSR sur différentes concaténations :

Jour observation	Juin	Juin/Septembre	Juin/Septembre /Octobre	Juin/Septembre /Octobre/Novembre
RMSE	**7.739**	**7.310**	**7.238**	**7.282**
Variance capturée	**51%**	**56,3%**	**57,2%**	**56,6%**

Tableau 2 : RMSE et Variance capturée pour différentes combinaisons de mois concaténés

On remarque que la qualité de prédiction augmente lorsqu'on augmente les données disponibles (Tableau 2). Nous pouvons donc conclure que chacun des mois apporte une information complémentaire, qui va en s'amenuisant au fur et à mesure de la sénescence (Tableau 1).

c. Utilisation de l'outil Pinard

Dans le but d'améliorer les résultats, nous avons utilisé Pinard (**G.Beurier *et al.,*** 2022*)*, un package Python développé dans le but d'analyser les spectres NIRS en réalisant différents preprocessing sur les donnés afin d'avoir plus de features que simplement le spectre brut. Pinard permet de faire de l'augmentation de données et tester différentes méthodes de machine learning.
Nous nous sommes servi d'un set de preprocessing appelé "smallset" recommandé directement par Gregory Beurier. Ce set contient les transformations suivantes :

- La transformation *identité* qui consiste à utiliser les valeurs brutes du spectre sans aucune modification.
- La *normalisation standard* qui vise à mettre toutes les données du spectre sur la même échelle en soustrayant la moyenne et en divisant par l'écart type du spectre.
- Le filtre *Savitzky Golay* qui réduit le bruit et les fluctuations dans le spectre NIRS en ajustant localement des polynômes de régression.
- La décomposition en *ondelette de Haar* qui divise le spectre en différentes bandes de fréquences afin de mieux isoler les caractéristiques importantes du signal.
- Un filtre de *detrending* qui vise à supprimer une éventuelle tendance non linéaire

présente dans le spectre NIRS.

Cela nous donne un total de 5 features de spectre que la PLS va analyser simultanément plutôt qu'individuellement comme réalisé auparavant sur R, afin de prédire la date de floraison. Augmenter le nombre de features augmente l'information sur les spectres et donc peut permettre une meilleure prédiction par la régression PLS. Voici donc les résultats que nous avons obtenus sur le tableau 3:

	Juin	Tous les mois	Juin - Pinard	Tous les mois - Pinard
RMSE	**7.739**	**7.282**	**7.739**	**6,827**
Variance capturée	**51%**	**56,6%**	**51%**	**61,9%**

Tableau 3 : RMSE et Variance capturée sans Pinard versus avec Pinard.

Grâce à ces nouvelles features, la variance capturée a augmenté d'environ 5%, cumulé à la concaténation des différents mois, nous avons augmenté la variance capturée d'un total de 10,9% pour les 534 individus.

Pour conclure, nous avons réussi à mieux prédire la date de floraison lorsque nous utilisons l'information de spectres NIRS de plusieurs dates, ce qui signifie qu'il y a de l'information utile tout au long de la sénescence de l'arbre pour prédire sa date de floraison au printemps suivant. De plus, la combinaison de différentes transformations du spectre améliore la prédiction du modèle.

2. Prédiction de la date de floraison à partir des données drone

a. Travail antérieur à mon arrivée

L'année dernière, un autre stage avait été réalisé sur la même problématique générale avec uniquement les données drone. Ce stage avait été réalisé par Dan Busnach, étudiant en master 2 d'économétrie à l'AMSE (école d'économie Aix-Marseille).

Lors de son stage, Dan a d'abord exploré les données afin de voir la distribution des dates de floraison des 674 pommiers. L'écart-type des dates de floraison de l'ensemble de la population est de 11,08 jours, et l'écart-type intra-génotype est bien moins élevé puisqu'il est en moyenne sur les 214 génotypes observés de 1.82 jours. Le génotype est donc un déterminant majeur de la date de floraison, avec 97,26% de la variance totale expliquée par le génotype (et seulement 2,74% par l'individu).

Dans un second temps, des "patchs" ont été extraits des orthomosaïques représentant chaque arbre. Un patch est une image carrée, centrée sur chaque individu en utilisant leur localisation indiquée par un centroïde géoréférencé de l'arbre. La taille des patch a été choisie pour minimiser l'environnement (pelouse, etc.) et les arbres voisins.Les patchs MS avaient du coup des tailles de 200x200 pixels, et les patchs RGB des tailles de 300x300 pixels (différence dûe à la résolution de chacune des caméras). Ces patchs ont permis d'individualiser les arbres et d'extraire un NDVI et un NDRE moyens pour chaque date et pour chaque arbre.

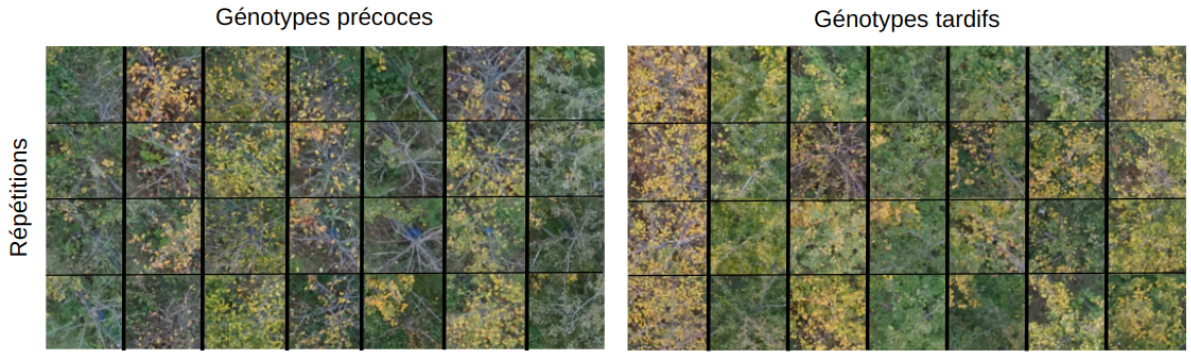
Différentes méthodes de machine learning ont été testées pour prédire la date de floraison à partir du NDVI et NDRE moyens, telles que les méthodes Lasso, Random Forest ou encore Gradient Boosting. Pour le deep learning, une architecture inspirée de l'encodeur de ResUnet a été réalisée pour extraire automatiquement des caractéristiques complexes et prédire les dates de floraison.

Après avoir essayé ces nombreuses méthodes, le meilleur résultat a été obtenu en simplifiant les dates de floraison individuelles en 6 périodes de floraison via un clustering. Pour chaque individu la période de floraison est prédite via un réseau de neurones puis une régression linéaire est utilisée pour prédire une date de floraison moyenne pour chaque période. Une RMSE de 10,01 a été obtenue. Cela représente un peu moins de 20% de la variance capturée. L'erreur de prédiction est quasiment égale à l'écart-type du jeu de données à 1 jour près, c'est à dire que la prédiction est à peine meilleure qu'un modèle qui prédirait la moyenne de la date de floraison, indépendamment des images.

b. Visualisation temporelle des donnés drone

Dans mon stage nous avons adopté une approche différente sur l'analyse des images. Nous avons voulu d'abord passer par une étape de visualisation experte des données pour former une hypothèse biologique à partir de ce que l'on observait. J'ai donc visualisé un certain nombre d'arbres, en les regroupant par génotype, et en comparant leurs dates de floraison au printemps suivant. Le but était de voir si l'on pouvait observer une tendance générale différente pour des génotypes qui fleurissent tôt et des génotypes qui fleurissent tard.

J'ai considéré les génotypes dont les 4 individus avaient été observés, pour avoir un échantillon bien représentatif de ces génotypes. Parmi ces génotypes, j'ai sélectionné les 7 génotypes les plus précoces (Figure 4a) et les 7 génotypes les plus tardifs (Figure 4b) pour lesquels j'ai réalisé une série temporelle (image 4D avec le temps comme 4e dimension) . Le mois de novembre de cette série temporelle est représenté en Figure 4 car c'est le seul où nous avons pu observer une différence visuelle notable entre nos 2 groupes d'individus (données non montrées).



Figures 4a et 4b : *Les 7 génotypes les plus précoces (à gauche) et les 7 génotypes les plus tardifs (à droite). Les quatre patchs en colonne correspondent aux 4 arbres du même génotype, les 7 patchs en ligne correspondent aux sept génotypes choisis.*

Ces images représentent un total de 14 génotypes et 56 individus, ce qui correspond à environ 1/10e de la population. A cette date, les génotypes qui ont fleuri de façon plus précoce (Figure 4a) ont moins de feuilles que ceux qui ont fleuri tardivement (Figure 4b).
L'hypothèse suivante a donc été formulée :

Les arbres qui perdent leurs feuilles plus tôt à l'automne, ont tendance à fleurir plus tôt au printemps suivant.

Pour tester cette hypothèse, j'ai essayé d'extraire explicitement le pourcentage de feuilles au cours du temps et l'évolution du NDVI ainsi que d'autres caractéristiques, pour les confronter avec la date de floraison. La mesure du pourcentage de feuilles et du NDVI moyen des arbres se sont révélés difficiles à cause d'éléments des images qui "brouillent" le signal. Notamment, les patchs des arbres contiennent des parties des arbres voisins, et des morceaux de pelouse qui peuvent facilement être confondus avec la couronne de l'arbre. Ces éléments apportent des informations contradictoires aux informations recherchées (Figure 5).

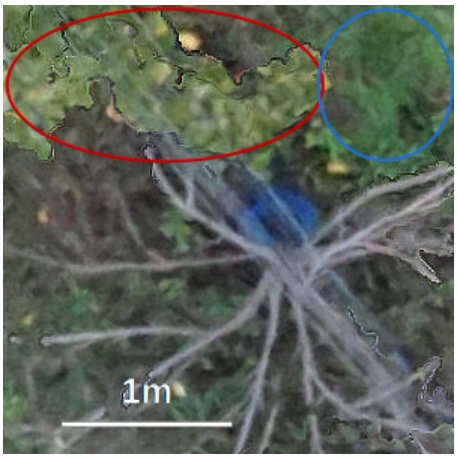


Figure 5 : *Exemple d'un patch d'arbre avec la pelouse et du feuillage d'un arbre voisin*

Pour corriger ce problème, nous avons mis en place un pipeline de segmentation automatique des arbres pour les distinguer du sol (segmentation sémantique), et les individualiser (segmentation d'instances) qui nous permet d'avoir la segmentation des couronnes individuelles des arbres. Une fois cette segmentation réalisée, nous pourrons extraire des indices de végétation plus précis, ainsi qu'estimer l'indice foliaire de l'arbre au cours du temps, dans le but de répondre à notre hypothèse.

c. Segmentation des arbres

Voici le schéma fonctionnel de la solution proposée pour segmenter les couronnes individuelles :

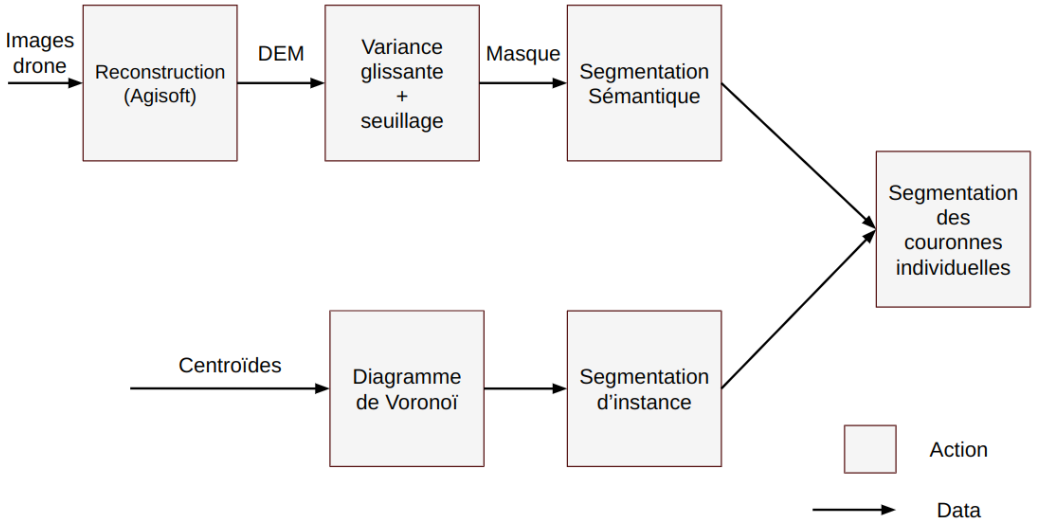
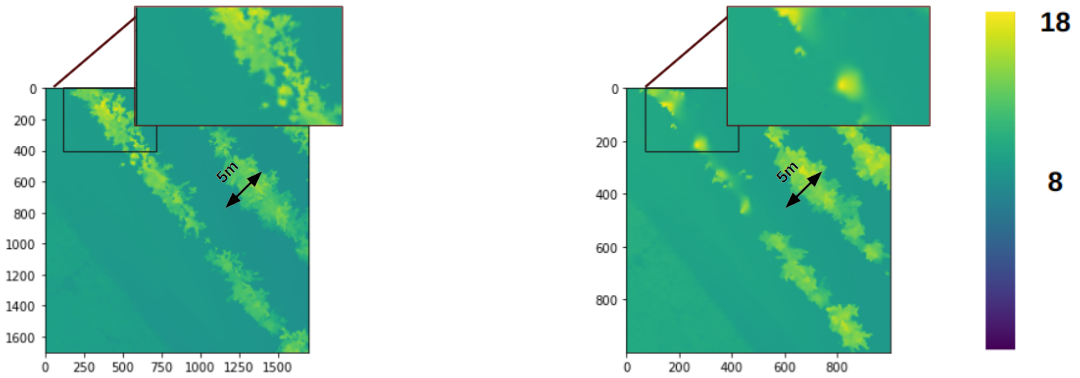


Figure 6 : *Schéma fonctionnel de la solution proposée et de l'ordre de réalisation*

i. Segmentation sémantique

1. Le Dense Elevation Model

Lors de la reconstruction de l'orthomosaïque, le logiciel Agisoft Metashape construit parallèlement une orthoimage nommée “Dense Elevation Model” ou DEM, qui indique en tout point l’élévation du terrain. Ce DEM pourrait nous permettre de détecter les pixels correspondant aux arbres, ou à l’environnement (sol). Après un certain nombre d’essais, j’ai choisi de ne travailler qu’avec les DEM provenant des orthomosaïques RGB étant donné que leur qualité était meilleure (Figures 7a et 7b). Cela est probablement dû au fait que la résolution de la caméra RGB est supérieure à celle de la caméra MS, et que l’algorithme de reconstruction d’Agisoft travaille plus difficilement avec les cartes multispectrales, dans lesquelles les objets sont plus difficiles à reconnaître.



Figures 7a et 7b : *DEM provenant de l’image RGB (à gauche) et de l’image MS (à droite)*

2. Application de la variance glissante et du seuil

Le sol du verger étant relativement plat, j’ai identifié les points de sol en calculant la carte de variance locale, puis j’ai seuillé cette carte pour ne sélectionner que les points dont le voisinage varie beaucoup en altitude. En effet, les points d’élévation correspondant à l'arbre varient bien plus que ceux correspondants au sol. L’opération de seuillage de la carte de variance du DEM produit un masque binaire qu’on peut appliquer sur les images pour sélectionner les arbres. Cependant les dimensions du DEM, les coordonnées à l’origine et les tailles de pixel ne sont pas les mêmes que celles des ortho-mosaïques RGB et multispectrale. Il faut donc ré-échantillonner ce masque pour pouvoir l’appliquer aux autres données.

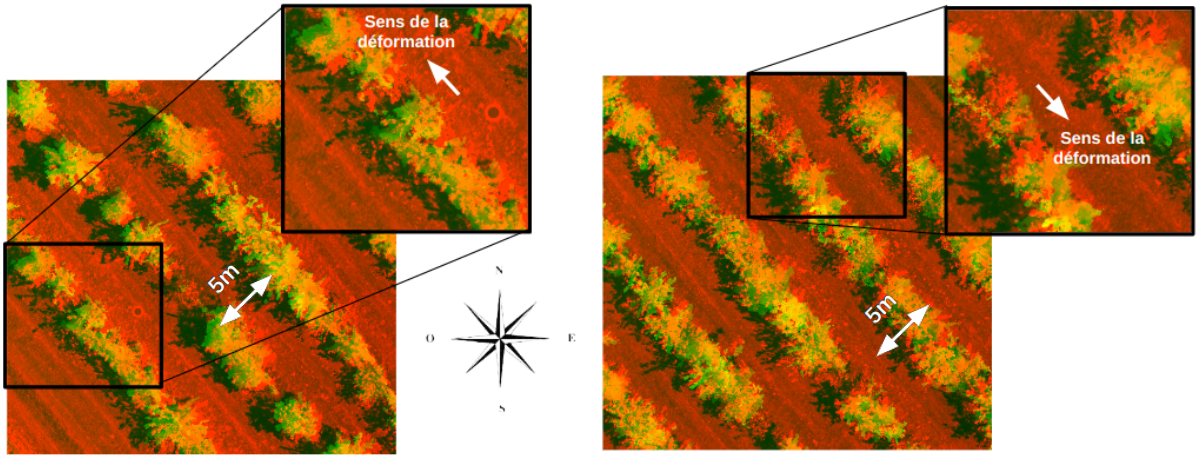
3. Rééchantillonnage et segmentation sémantique

La première approche de traitement d’image que nous avons imaginée pour résoudre ce problème était de redimensionner et remettre à l’échelle le DEM sur l’orthomosaïque à l’aide d’une matrice de passage et du package SimpleITK. Le package nous a permis de corriger l’origine grâce aux tailles de pixel relatives de chacune des 2 images et leurs coordonnées à l’origine. Nous avons ensuite redimensionné le DEM sur l’orthomosaïque grâce au rapport de taille de pixel entre les 2 images, via une interpolation bilinéaire. Ces 2 opérations peuvent être résumées par la matrice de passage qui décrit le changement de système de coordonnées (Équation 3).

$$Equation\ 3 :\ P\ = \begin{pmatrix} \frac{Voxel\ size\ \ x1}{Voxel\ size\ x2} & 0 & t_x \\ 0 & \frac{Voxel\ size\ \ y1}{Voxel\ size\ y2} & t_y \\ 0 & 0 & 1 \end{pmatrix}$$

$$Coords_2 = P * Coords_1$$

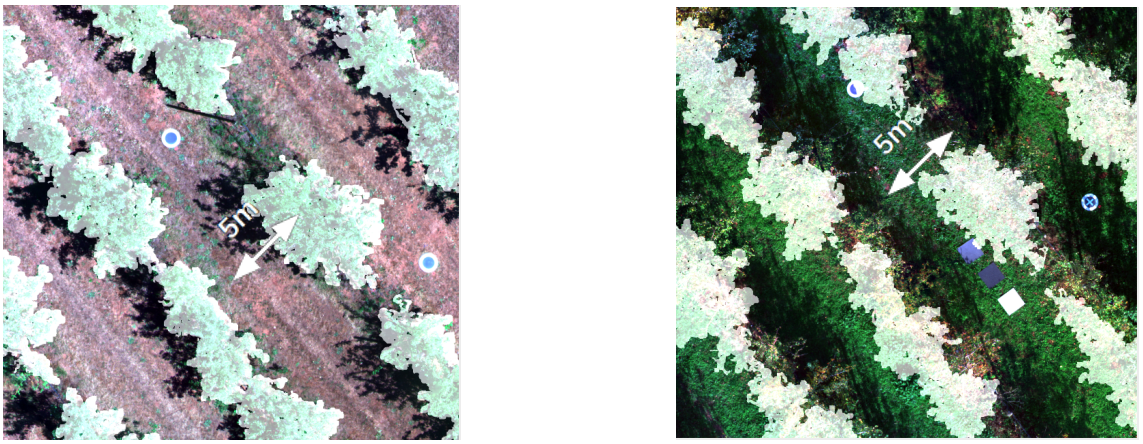
Avec t_x, t_y les coordonnées d'origine de l'espace 1 projetées dans l'espace 2. Afin de vérifier l'alignement des différentes sources de données, je les ai superposées en leur attribuant chacune un canal différent (Figures 8a et 8b).



Figures 8a et 8b : *DEM (vert) superposé sur l'ortomosaïque (rouge), sur les premiers rangs du verger (à gauche) et sur les derniers (à droite).*

Après superposition, on remarque un léger décalage des deux images, décalage qui s'inverse au fil des rangs. Le DEM est décalé vers le Nord-Ouest par rapport à l'image sur les premiers rangs (Figure 8a), et il est décalé vers le Sud-Est sur les derniers rangs (Figure 8b). J'ai essayé de faire une correction supplémentaire en estimant une transformation affine à partir de points de contrôle identifiés manuellement, mais sans succès. Nous avons alors travaillé sur un recalage automatique complémentaire, en estimant un champ de vecteurs de déformation.

J'ai utilisé Fijiyama (**Fernandez R. et al.*, 2020***), un plugin de recalage d'image sur le logiciel ImageJ, afin de corriger cette déformation, en estimant d'abord une transformation rigide, puis par un champ de vecteurs. Après ce recalage, j'ai pu constater un meilleur alignement, de l'ensemble de la parcelle (Figures 9a et 9b).



Figures 9a et 9b : *DEM (blanc) superposé sur l'ortomosaïque (RGB), en juin (à gauche) et en novembre (à droite).*

Le DEM s'aligne très bien en juin, toutes les feuilles sont prises en compte par celui-ci. En revanche, en novembre, l'alignement marche moins bien. Cela est certainement dû aux

ombres qui posent souci à Fijiyama, mais aussi aux erreurs de reconstruction qui font que nous avons DEM et une orthomosaïque de moins bonne qualité. Nous avons gardé ces masques afin de continuer notre segmentation.

ii. Segmentation d'instance

Une fois obtenue la segmentation sémantique de l'image (associer une étiquette à chacun des pixels), nous avons besoin d'individualiser les arbres, et donc de construire une segmentation d'instance.

1. Centroïdes et diagramme de Voronoï

Pour individualiser les arbres, nous réalisons une segmentation du masque global des arbres, que nous divisons en sous-zones, une zone par arbre. Etant donné que nous possédons les coordonnées de leurs centroïdes, nous utilisons ces coordonnées pour calculer le diagramme de Voronoï. Le diagramme de Voronoï permet de décomposer un espace métrique à partir d'un ensemble discrets de points que l'on nomme "germes". Il se construit en déterminant, pour cet ensemble de germes, les médiatrices de chaque couple de germes. Un point d'une médiatrice appartient donc à une frontière de Voronoï s'il est équidistant d'au moins deux germes et qu'il n'y a pas de distance plus petite entre ce point et un autre germe de l'ensemble.

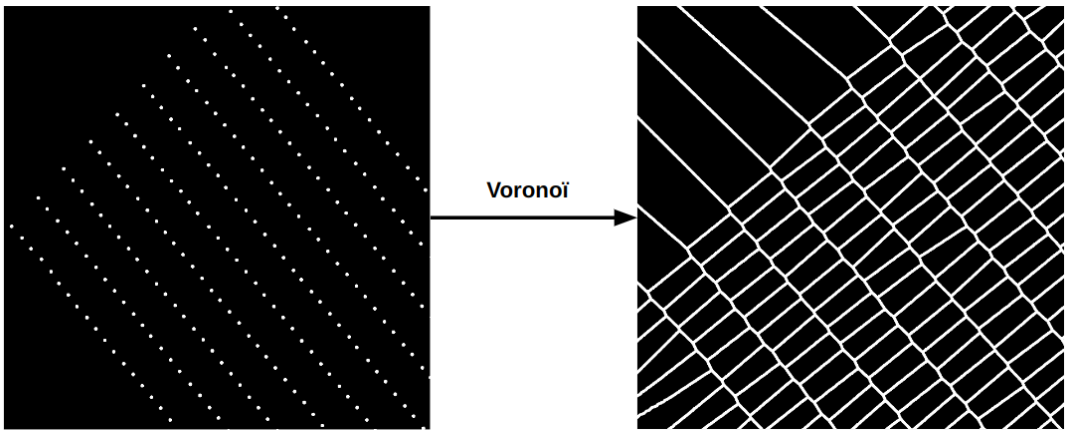


Figure 10 : *Création du Diagramme de Voronoï à partir centroïdes de nos arbres*

2. Segmentation d'instance

Une fois la diagramme de Voronoï appliqué, nous obtenons une segmentation d'instance de nos arbres (Figure 10) :

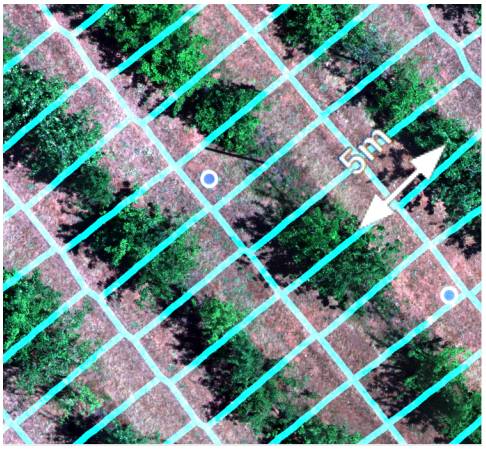
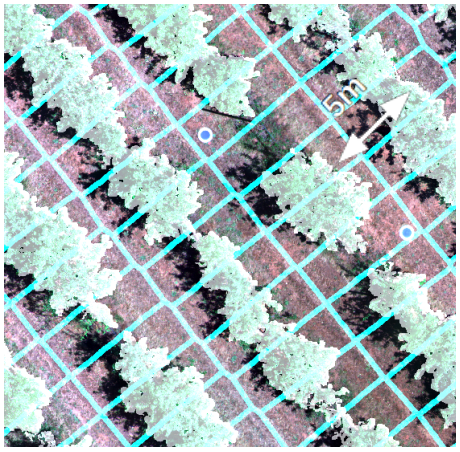


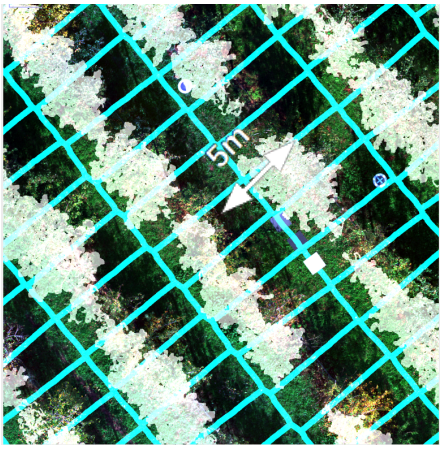
Figure 11 : *Segmentation d'instance des arbres (cyan) réalisée grâce au diagramme de Voronoï superposée sur l'orthomosaïque (RGB) en juin.*

Le résultat n'est pas parfait, mais même à l'œil nu, il est parfois difficile de déterminer où est la frontière entre deux arbres. Le diagramme aurait pu être légèrement amélioré si j'avais réalisé un tri dans les centroïdes initiaux en retirant les arbres morts qui ont été considérés ici. En effet, un arbre mort laisse la place à ses voisins pour se développer au niveau racinaire et aérien.

iii. Segmentation des couronnes individuelles

Une fois la segmentation d'instance combinée avec la segmentation sémantique, nous obtenons une segmentation des couronnes individuelles pour chacun des arbres (Figures 12a et 12b).





Figures 12a et 12b : *Segmentations sémantique (blanc) et instantielle (cyan) superposées sur l'ortomosaïque (RGB), en juin (à gauche) et en novembre (à droite)*

d. Prédiction de la date de floraison

Nous nous appuyons sur la segmentation des couronnes individuelles des arbres pour extraire certains attributs et servir de variables explicatives pour prédire la date de floraison. En complément des 6 canaux de la caméra multi-spectrale, j'ai rajouté le NDVI comme 7e canal, puis j'ai utilisé la fonction "regionprops table" de scikit image pour extraire les 22 attributs suivants pour chaque date :

- Aire de l'individu

- Moyenne des 7 canaux (calculé sur l'aire segmentée)
- Écart type des 7 canaux sur la segmentation de l'arbre (calculé sur l'aire segmentée)
- Médiane des 7 canaux sur la segmentation de l'arbre (calculé sur l'aire segmentée)

Au total nous avons donc 88 features par arbre. Pour évaluer la validité des features, nous avons calculé des features relatives en prenant juin comme référence et en faisant l'hypothèse que le NDVI et l'aire des feuilles ne ferait que baisser avec le temps.

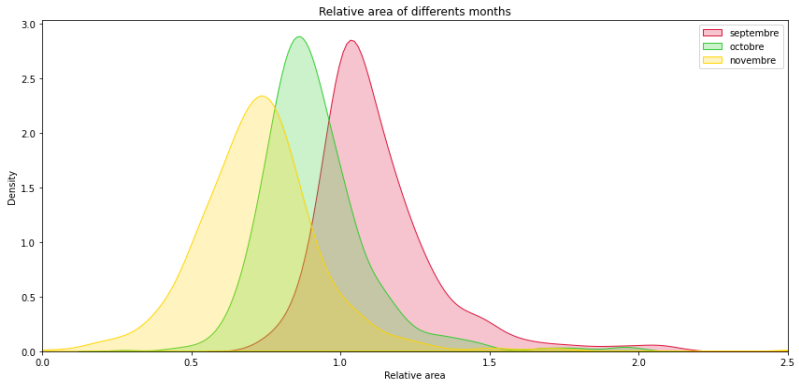


Figure 13 : *Graphique de densité de l'aire relative des mois de septembre, octobre et novembre par rapport à juin*

Nous pouvons voir (Figure 13) que l'aire relative a bien une tendance décroissante au niveau de la distribution. On observe une légère augmentation de l'aire en septembre relativement à juin, qui pourrait être expliquée par la croissance des arbres.

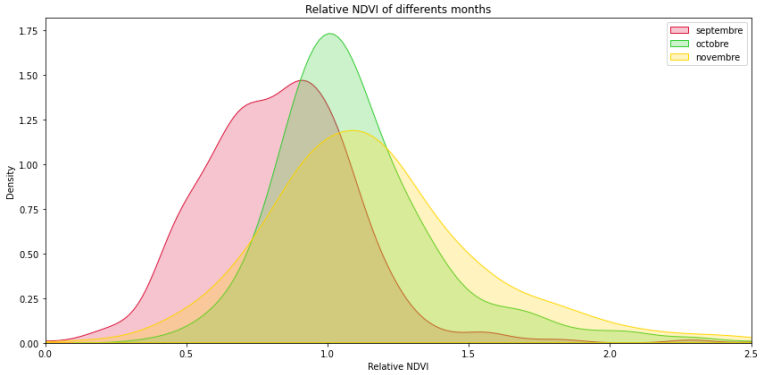


Figure 14 : *Graphique de densité du NDVI relatif des mois de septembre, octobre et novembre par rapport à juin*

La figure (Figure 14) montre probablement un défaut, car les arbres perdent leurs feuilles, et les feuilles qui perdent leur activité photosynthétique devraient contribuer à une perte de signal NDVI. En revanche, à l'automne, le gazon devient très vert, le décalage vers des valeurs positives du NDVI observé en octobre/novembre est un probable indicateur de la présence de gazon inter-rang dans la segmentation.

Malgré cela nous avons essayé de faire de la prédiction de la date de floraison par une régression Random Forest à l'aide des 88 features sur les 574 arbres. Les résultats sont plutôt mauvais, car nous ne capturons que 6% de la variance totale, avec un RMSE de 10.75.

e. Vérification manuelle de l'hypothèse

Les résultats de la partie précédente ne permettent pas de conclure sur l'hypothèse initiale de l'impact de l'état du feuillage de l'arbre au moment de la sénescence sur sa date de floraison au printemps suivant. Afin de tester cette hypothèse, nous avons "court-circuité" l'étape de segmentation automatique (Figure 15). Pour cela, nous avons annoté manuellement le pourcentage de feuilles et de NDVI relatif par rapport à juin de chaque arbre pour chaque date.

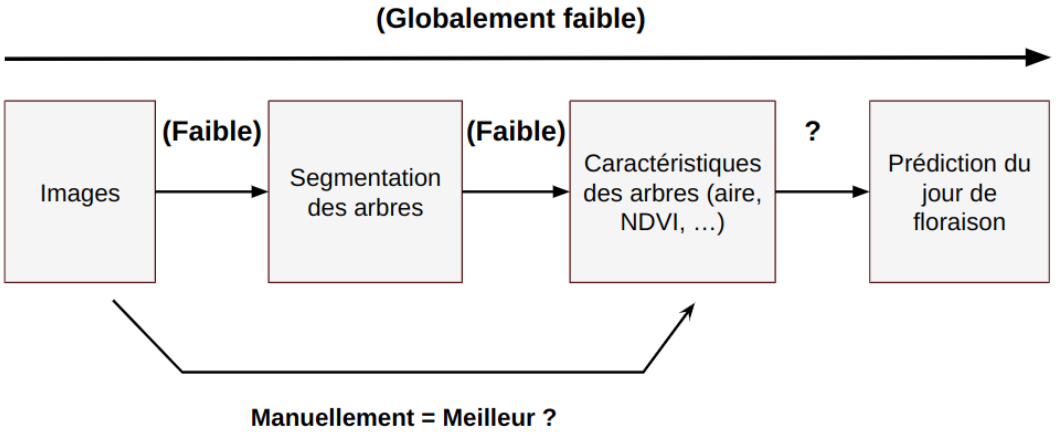


Figure 15 : *Schéma du cheminement suivi et de l'idée d'annoter manuellement pour tester notre hypothèse initiale*

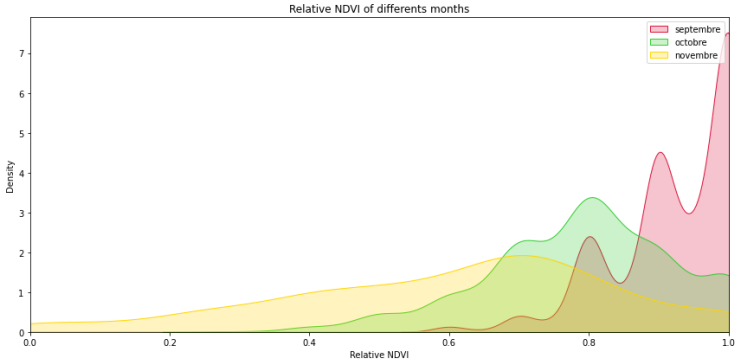


Figure 16 : *Graphique de densité du NDVI relatif des mois de septembre, octobre et novembre par rapport à juin*

La répartition de NDVI suit une évolution attendue (Figure 16), avec une diminution progressive.

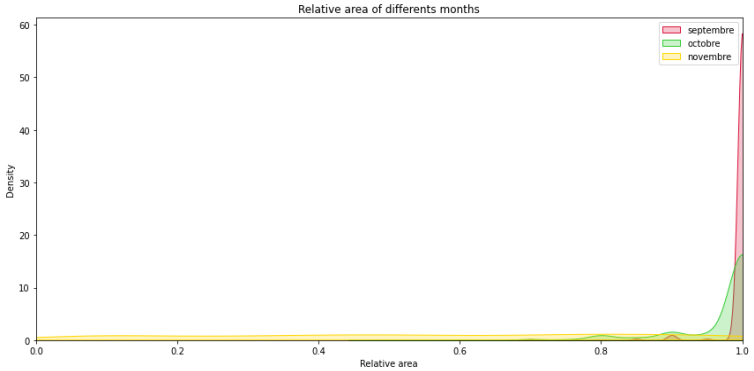


Figure 17 : *Graphique de densité de l'aire relative des mois de septembre, octobre et novembre par rapport à juin*

La répartition des aires (Figure 17) montre que les arbres perdent progressivement leurs feuilles et qu'ils sont répartis de façon assez homogène en novembre.

Nous avons utilisé la prédiction Random Forest sur ces données afin de prédire la date de floraison. Les résultats de variance capturées ont été très mauvais, plus encore qu'avec nos features extraites suite à la segmentation automatique de nos arbres.

Nous en avons conclu que notre hypothèse initiale de l'impact des feuilles restantes au mois de novembre sur le jour de floraison au printemps suivant ne peut être validée sur la base de ces essais. En effet, nous pensions voir une certaine tendance à l'œil nu sur un échantillon de nos arbres, nous avons annoté la totalité nos arbres de cette façon, mais nous n'arrivons pas à prédire la date de floraison avec un prédicteur simple. Cela ne plaide pas pour notre hypothèse initiale, cependant les éléments en l'état ne permettent pas de conclure.

3. Exploitation combinée NIRS/Drone

Le NIRS et l'imagerie drone sont deux outils différents pour mesurer la réflectance du feuillage à différentes longueurs d'ondes. Dans cette partie nous essayons de construire des ponts entre ces données. Nous développons deux expériences pour documenter les similarités entre les données drone et NIRS, d'abord en observant les corrélations entre les données NIRS et les valeurs multispectrales observées en imagerie drone, puis en essayant de prédire la valeur du NDVI moyen observé par drone, par régression à partir du spectre NIRS. Enfin, nous essayons de mobiliser les deux informations en parallèle pour réaliser une prédiction multimodale en nous appuyant sur une combinaison NIRS drone.

a. Simulation de données multispectrales via le spectre NIRS

Nous avons formulé l'hypothèse que les données NIRS et drone étaient liées puisque dans les deux cas nous tentons de capturer les caractéristiques du feuillage des arbres, et ce, au même moment dans la saison de végétation. Pour tester cette hypothèse, nous essayons de faire un lien entre ces données. Pour cela nous simulons le signal que capturerait la caméra multispectrale si les données d'entrée étaient un spectre NIRS. Pour créer ce signal multispectral artificiel, nous imitons la fonction de transfert de la caméra multispectrale pour chaque longueur d'onde capturée, en nous basant sur les spécifications, qui indiquent la longueur d'onde centrale et le FWHM (Full width at half maximum, largeur à mi-hauteur du pic d'amplitude maximal du signal) (Figure 18).

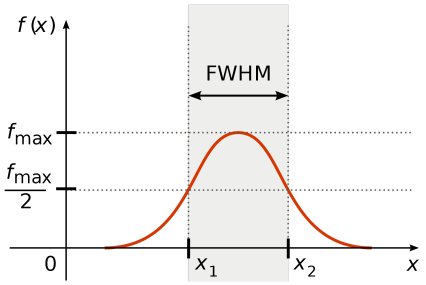


Figure 18 : *Schéma représentatif du FWHM. source : Wikipédia*

Dans le cas de la caméra multispectrale utilisée qui est une Airphen V3, le FWHM est de 10nm pour les 6 longueurs d'ondes. Nous avons donc appliqué un filtre gaussien avec un FWHM de 10 nm sur chacune de nos 6 longueurs d'onde de la caméra sur le spectre NIRS

afin d'obtenir un signal MS simulé (MS_sim). Pour avoir une seule valeur par arbre par canal, nous avons réalisé la somme des valeurs du spectre, pondérées par les valeurs de la fonction de transfert. Nous répétons ce calcul pour les 6 canaux de la caméra multispectrale (Figure 19).

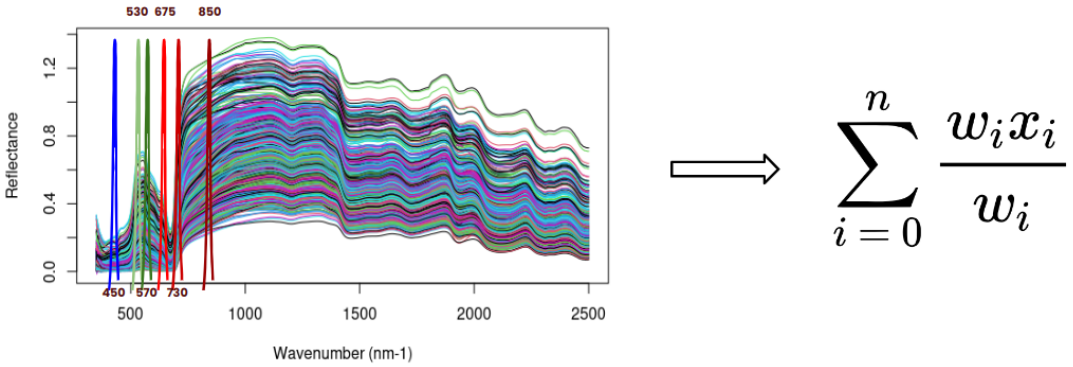


Figure 19 : *Application du FWHM sur le spectre NIRS et somme pondérée appliquée pour chacun des 6 canaux MS simulés*

Une fois cette somme pondérée réalisée, nous avons une valeur simulée du signal MS (un vecteur à 6 valeurs) par arbre. Dans l'objectif de voir si les données MS et NIRS coïncident, nous essayons de mettre en correspondance cette valeur MS_sim. Pour cela, nous calculons la corrélation entre le MS_sim de l'arbre et le MS de chaque pixel de l'arbre observé en drone (Figure 20).

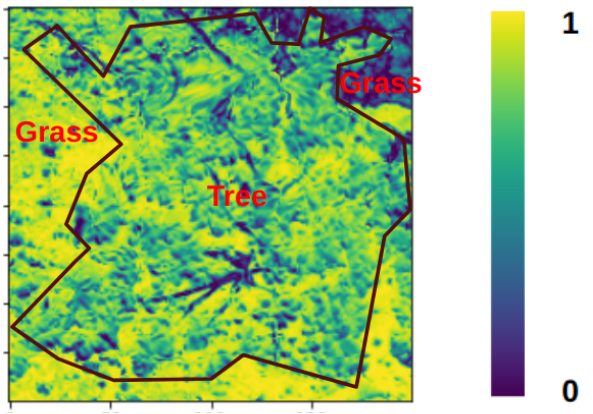


Figure 20 : *carte de la corrélation entre les 6 canaux MS et les 6 valeurs MS simulées à partir du NIRS pour chacun des pixels.*

De manière surprenante, les zones qui corrèlent le mieux ne correspondent pas à l'arbre mais à l'herbe. Les zones qui ne corrèlent pas très bien correspondent parfois aussi à de l'herbe, et l'arbre, lui, corrèle de manière limitée. A titre de vérification, nous avons vérifié si les longueurs d'onde sélectionnées sur la base des spécifications de la caméra MS étaient bien celles qui correspondaient au signal NIRS. Nous avons donc répété le calcul de la carte de corrélation en décalant de 10 nm en 10 nm les longueurs d'ondes des 6 canaux MS (figure 21).

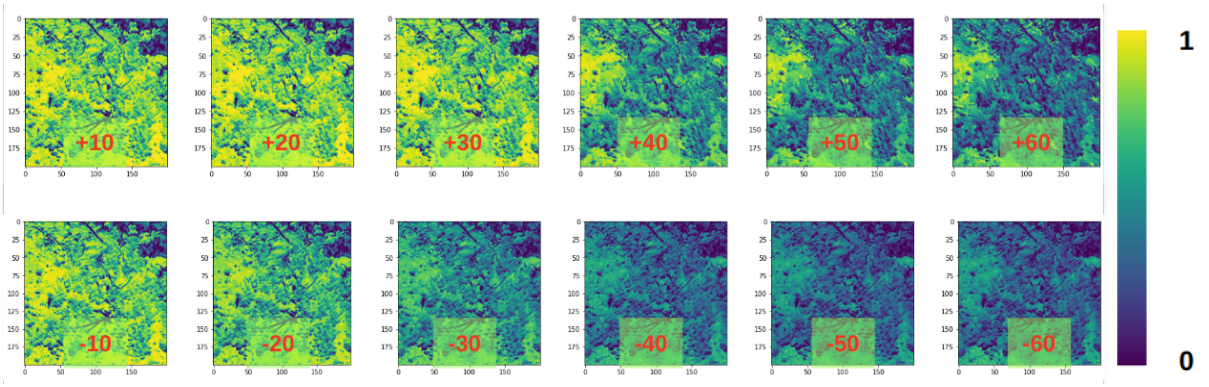


Figure 21 : *Patch de corrélation entre les 6 canaux MS et les 6 valeurs MS simulées à partir du NIRS en décalant de 10 nm le centre de chaque canal à chaque patch.*

Plus on décale nos canaux, moins les pixels des patchs corrèlent, ce qui plaide pour une bonne définition des canaux dans cette méthode. Cette vérification faite, nous concluons que cette approche ne permet pas de corréler les données NIRS et les données drone. Nous n'arrivons pas à déterminer pourquoi les données NIRS et drone sont aussi décorrélées alors qu'elles représentent le même individu, on suppose que de nombreux paramètres entrent en jeu : traitement de la feuille (lyophilisation) avant la prise de son spectre NIRS, influence de l'architecture de l'arbre et des conditions météo sur la construction de l'image.

b. Prédiction du NDVI moyen à partir du spectre NIRS

Nous avons développé une seconde approche pour faire un lien entre les données, en essayant de prédire la valeur moyenne du NDVI d'un arbre à partir de son spectre NIRS, et donc de ne pas limiter le signal NIRS à 6 bandes, mais d'utiliser la totalité de celui-ci. De plus, le spectre nous donne certainement une information sur la santé de l'arbre, donc sur son NDVI. Si nous pouvons prédire le NDVI à partir du spectre NIRS, alors il y a un certain lien entre les données.

Pour cela, j'ai essayé de prédire par régression PLS, pour chaque date, le NDVI des arbres à l'aide de leur spectre NIRS. J'ai réalisé cela avec comme variable réponse les features extraites de nos images segmentées (la surface de la couronne, ou le NDVI moyen) ou bien la valeur du NDVI au niveau du centre du patch d'arbre (Tableau 4).

Variance capturée	Juin	Septembre	Octobre	Novembre
Features	-5,5%	15,20%	3,58%	29,60%
Patchs	9,50%	17,67%	11,88%	25,86%

Tableau 4 : Variance capturée lors de la prédiction du NDVI moyen des centres de patchs ou des features à l'aide du spectre NIRS correspondant (même mois, même individu).

On observe globalement une assez mauvaise prédiction (sauf en novembre, mais nous considérons que 30% est assez bas), avec plus de régularité prédictive sur les centres des patchs que sur les features extraites de nos images segmentées. Encore une fois nous ne pouvons pas conclure d'où vient cette différence entre nos deux types de données.

c. Prédiction multimodale NIRS/drone

Afin de répondre à la troisième question méthodologique de mon stage, même si les données drones semblent difficiles à utiliser, nous avons réalisé la prédiction du jour de floraison à partir de la combinaison multimodale des données disponibles NIRS et drone.

Pour cela, nous avons ajouté nos 88 features extraites de la segmentation automatique au spectre NIRS afin de voir si les résultats prédictifs étaient meilleurs.

La variance capturée est de 62,4% pour une RMSE de 6.78 pour la PLSR de Pinard. Il y a eu une augmentation de la variance capturée de 0,5%. Cela reste négligeable mais on peut en conclure qu'il pourrait y avoir de l'information utile dans les images aériennes, non redondantes avec les données NIRS, mais la méthode d'analyse utilisée ne permet de les exploiter que très faiblement.

IV. Discussion

Durant ce travail, nous avons amélioré les résultats obtenus précédemment dans le projet FruitFlow. Nous avons montré qu'il était possible de prédire plus précisément la date de floraison de pommiers à partir des données acquises pendant la sénescence, spectres NIRS et images drone.

En particulier sur les spectres NIRS, nous avons capturé 62,4% de la variance de la date de floraison, contre 51% avant ce travail (performances mesurées en cross-validation). Pour améliorer les résultats, nous avons d'abord montré que les spectres observés à chacune des dates d'observation apportent des informations complémentaires (qui diminuent au fur et à mesure de la sénescence). Nous avons combiné ces informations en entraînant un régresseur PLS à travailler sur ces données concaténées. Nous avons développé la nouvelle solution en Python, et mobilisé la librairie Pinard qui nous a permis une meilleure exploitation de l'ensemble des transformations des spectres.

Nous avons ensuite travaillé sur l'imagerie drone. Suite aux difficultés rencontrées sur le stage précédent (en 2022), nous avons testé une hypothèse de travail simple : le lien entre la dynamique de sénescence et la date de floraison, pour prédire des dates de floraison par imagerie. Nous avons construit un pipeline permettant d'extraire des caractéristiques temporelles des arbres : dynamique de sénescence et vigueur mesurée comme le NDVI moyen. Ce pipeline n'a pas permis d'améliorer l'état des prédictions.
Nous avons identifié plusieurs points faibles, et des pistes pour aller au-delà de ces difficultés. La reconstruction par Agisoft (effectuée en amont de notre travail) produit des artefacts qui rendent difficile l'analyse automatique des images, la reconstruction est particulièrement mauvaise sur les orthomosaïques multispectrales. Cette reconstruction sur Agisoft Metashape pourrait combiner MS et RGB, afin de bénéficier de la définition des images RGB pour reconstruire un ensemble MS+RGB facile à superposer et plus facile à exploiter.
Aussi, le pipeline d'analyse à partir des orthoimages et des DEM est sûrement perfectible : en particulier la segmentation par sélection des zones à forte variance semble peu sélective et les arbres sont mélangés avec des portions de gazon (mesures non vérifiées). Enfin, notre pipeline s'appuie sur la reconstruction avec l'usage du DEM pour la segmentation, mais nous aurions peut-être moins de difficultés en utilisant les images d'origine du drone, non reconstruites en orthomosaïque (ce qui apporte cette déformation), qui pourraient être segmentées par une voie plus "standard" (réseau de neurones Unet).

Finalement, l'hypothèse de lien entre la date de floraison, la géométrie et la dynamique de la sénescence observées en vue aérienne (chute des feuilles et NDVI moyen de la couronne) a été testée à partir des paramètres extraits automatiquement et manuellement, et n'a pu être validée. Pour la suite, il pourrait être envisagé d'avoir des features plus détaillées. Peut être que les arbres vu de dessus impliquent des difficultés (distinguer le gazon et autre). Peut-être qu'en caractérisant mieux l'architecture de l'arbre, comme ***Abdollahnejad A et al.*** *(2018)* on pourrait capturer plus de variance.
Enfin nous avons tenté de relier les données NIRS et les données multispectrales. Pour cela nous avons généré un signal multispectral à partir des données NIRS, et essayer de prédire le NDVI à l'aide des spectres NIRS. Ces deux méthodes ne nous ont pas permis d'établir de lien évident entre ces données de vue aérienne et les données NIRS. De nombreuses approches ont été essayées (non abordées dans le rapport) pour essayer de lier le signal MS simulé au signal MS capturé, mais elles n'ont fait que confirmer le fait que ces deux types de données sont très différents.

V. Conclusion et perspectives

Comme nous avons pu le voir, la prédiction de la date de floraison de pommiers afin de réaliser la caractérisation à haut débit à l'aide d'une drone est difficile, mais possible à moyen débit à partir de données NIRS collectées par des opérateurs. Notre hypothèse de l'état des feuilles et le taux de feuilles restantes au mois de novembre reste à ce jour non validée, mais l'image possède probablement un potentiel informatif que nous n'avons pas réussi à extraire, probablement en raison des différentes difficultés mentionnées ci-dessus.
L'utilisation simultanée de spectres NIRS acquis à différentes dates, traités avec l'outil Pinard permet des résultats prédictifs meilleurs qu'en prédisant à partir d'une seule date (à nombre d'individus comparables).
Côté images RGB et MS acquises par drone, si de nouvelles prises de données viennent à être réalisées en respectant les conditions que j'ai énumérées dans la discussion, le traitement et l'exploitation de ces dernières serait beaucoup plus simple et efficace.

Mon code a été déposé sur un dépôt GitHub (**André N., 2023**), ce qui permet l'utilisation ultérieure de mes fonctions pour traiter les images ou analyser les spectres NIRS

Bibliographie

Abdollahnejad A, Panagiotidis D, Surový P. Estimation and extrapolation of tree parameters using spectral correlation between UAV and Pléiades data. *Forests.* 2018 Feb 11;9(2):85.

Boiarskii B, Hasegawa H. Comparison of NDVI and NDRE indices to detect differences in vegetation and chlorophyll content. *J. Mech. Contin. Math. Sci.* 2019;4:20-9.

Cheng JH, Sun DW. Partial least squares regression (PLSR) applied to NIR and HSI spectral data modeling to predict chemical properties of fish muscle. *Food engineering reviews.* 2017 Mar;9:36-49.

Davidson C, Jaganathan V, Sivakumar AN, Czarnecki JM, Chowdhary G. NDVI/NDRE prediction from standard RGB aerial imagery using deep learning. *Computers and Electronics in Agriculture.* 2022 Dec 1;203:107396.

Djordjevic B, Djurovic D, Vulic T, Zec G. INFLUENCE OF SPRING FROST ON APPLE FLOWER BUDS AT VARIOUS DEVELOPMENTAL STAGES. *Journal of Agricultural, Food and Environmental Sciences*, JAFES. 2018 Dec 1;72(3):72-5.

Beed F, Taguchi M, Telemans B, Kahane R, Le Bellec F, Sourisseau JM, Malézieux E, Lesueur-Jannoyer M, Deberdt P, Deguine JP, Faye E. *Fruits et légumes*. Opportunités et défis pour la durabilité des petites exploitations agricoles.

Fischer C, Fedrigotti VM. 'An Apple A Day'… Is Going Away. What Can We Do to Stop the Decline in Per Capita Apple Consumption?. *American Journal of Biomedical Science & Research.* 2020 Sep 10;10(3):226-7.

Lassois L, Denancé C, Ravon E, Guyader A, Guisnel R, Hibrand-Saint-Oyant L, Poncet C, Lasserre-Zuber P, Feugey L, Durel CE. Genetic diversity, population structure, parentage analysis, and construction of core collections in the French apple germplasm based on SSR markers. *Plant Molecular Biology Reporter.* 2016 Aug;34:827-44.

Legave JM, Farrera I, Alméras T, Calleja M. Selecting models of apple flowering time and understanding how global warming has had an impact on this trait. *The Journal of Horticultural Science and Biotechnology.* 2008 Jan 1;83(1):76-84.

Panagiotidis D, Abdollahnejad A, Surový P, Chiteculo V. Determining tree height and crown diameter from high-resolution UAV imagery. *International journal of remote sensing*. 2017 May 19;38(8-10):2392-410.

Wang J, Guan Y, Wu L, Guan X, Cai W, Huang J, Dong W, Zhang B. Changing lengths of the four seasons by global warming. *Geophysical Research Letters*. 2021 Mar 28;48(6):e2020GL091753.

Jin-yong PU, Xiao-ying YA, Xiao-hong YA, Yan-ping XU, Wei-tai WA. Impacts of climate warming on phonological period and growth of apple tree in Loess Plateau of Gansu province. *Chinese Journal of Agrometeorology*. 2008 Apr 10;29(02):181.

Wei-tai W et al. (2008) Impacts of Climate Warming on Phonological Period and Growth of Apple Tree in Loess Plateau of Gansu Province. Chinese Journal of Agrometeorology

André N (2023) Stage sur la prévision de date de floraison de pommiers. GitHub. https://github.com/NielsAdr/Stage_deepflowering

Beurier G, Cornet D, Rouan L (2022) Pinard: a Pipeline for Nirs Analysis ReloadeD. https://pypi.org/project/pinard/0.9.5/

Fernandez R (2020) Fijiyama. ImageJ Wiki : https://imagej.github.io/plugins/fijiyama

Geves (2019) : Glossaire ressources phytogénétiques : https://www.geves.fr/wp-content/uploads/20190611_Glossaire_RPG_v3_df.pdf

IPCC (2023) **:** AR6 Synthesis Report**:** Climate Change. https://www.ipcc.ch/report/ar6/syr/

UVED (2008) : Les propriétés optiques des feuilles. https://e-cours.univ-paris1.fr/modules/uved/envcal/html/vegetation/caracteristique-vegetation/proprietes.html

Wikipédia (2023) : Largeur à mi-hauteur. https://fr.wikipedia.org/wiki/Largeur_%C3%A0_mi-hauteur

L'INSTITUT agro Rennes Angers	Diplôme : Ingénieur agronome Spécialité : Science des données Spécialisation / option : Statistiques Bayésiennes Enseignant référent : Marie-Pierre Etienne

Auteur(s) : Niels André Date de naissance* : 17/02/1999	Organisme d'accueil : Inrae Occitanie Montpellier Adresse : 2 place Pierre Viala, 34060 Montpellier
Nb pages : 29 Annexe(s) : 0	Maître de stage : Romain Fernandez
Année de soutenance : 2023	

Titre français : Prédiction de la date de floraison de pommiers au printemps à partir d'images drones et de spectre NIRS en période de sénescence.

Titre anglais : Prediction of spring apple blossom date from drone images and NIRS spectra during senescence.

Résumé (1600 caractères maximum) :

Cette étude vise à prédire la date de floraison des pommiers en combinant des données NIRS et des images drone à la sénescence précédent le printemps.

Les spectres NIRS ont été collectés à différentes dates, dans un premier temps, les données NIRS ont été utilisées date par date pour prédire la date de floraison. Ensuite, une prédiction PLSR a été réalisée en combinant les spectres de plusieurs dates à l'aide de l'outil Pinard. Cette approche a permis d'améliorer la prédiction de la date de floraison en capturant 62,4% de la variance de la variable à prédire.

Les images drone ont été traitées pour extraire des caractéristiques temporelles des arbres. Cependant, des problèmes de reconstruction d'orthomosaïques et de modèles d'élévation ont été rencontrés, rendant difficile l'extraction des caractéristiques souhaitées. Une tentative de corrélation entre les données NIRS et drone a été effectuée en simulant un signal multispectral à partir des spectres NIRS ainsi qu'une prédiction d'un indice de végétation de l'arbre (NDVI calculé sur les données drone) à partir de leurs spectres respectifs. Cependant, aucun lien fort n'a été mesuré entre les deux types de données.

En conclusion, l'utilisation des données NIRS a montré des résultats prometteurs dans la prédiction de la date de floraison. Cependant, l'exploitation des images drone s'est avérée plus complexe en raison des problèmes de reconstruction et de segmentation. Des ajustements seront nécessaires pour mieux exploiter les données drone dans de futures recherches.

Abstract (1600 caractères maximum) :

This study aims to predict the flowering date of apple trees by combining NIRS data and drone imagery at senescence before spring.

NIRS spectra were collected at different dates, and NIRS data were first used on a date-by-date basis to predict the flowering date. Then, a PLSR prediction was made by combining the spectra from several dates using the Pinard tool. This approach improved the prediction of flowering date, capturing 62.4% of the variance.

Drone images were processed to extract temporal characteristics of trees. However, problems with orthomosaic reconstruction and elevation models were encountered, making it difficult to extract the desired features. An attempt was made to correlate NIRS and drone data by simulating a multispectral signal from the NIRS spectra and a prediction of a tree vegetation index (NDVI calculated on the drone data) from their respective spectra. However, no link was found between the two types of data.

In conclusion, the use of NIRS data showed promising results in predicting flowering date. However, the use of drone images proved more complex due to reconstruction and segmentation problems. Adjustments will be needed to better exploit drone data in future research.

Mots-clés : Prédiction, date de floraison, pommiers, imagerie drone, spectres NIRS

Key Words : Prediction, flowering date, apple trees, drone imagery, NIRS spectra

* *Elément qui permet d'enregistrer les notices auteurs dans le catalogue des bibliothèques universitaires*