

**Pour l'obtention du diplôme de Master 2, Ingénieur  
Agronome, Spécialisation GEEFT : Gestion  
Environnementale des Écosystèmes et Forêts Tropicales**



Amélioration de la surveillance épidémiologique du  
« greening » des agrumes infectés par le Huanglongbing à La  
Réunion par Spectroscopie Proche Infra-Rouge

Soutenu publiquement le 8 juillet 2021  
À AgroParisTech, Centre de Montpellier

Auteur : Edouard SORIN

Encadrante de stage : Virginie RAVIGNE (CIRAD - UMR PVBMT)

Co encadrant : Olivier PRUVOST (CIRAD - UMR PVBMT)

Enseignant référent : Eric MARCON (AgroParisTech)

Examinateur : Simon TAUGOURDEAU (CIRAD - UMR SELMET)

Stage effectué du : 30/09/2020 au 02/04/2021 au :

CIRAD (Centre de coopération Internationale en Recherche Agronomique pour le Développement), Pôle de Protection des Plantes, 7 chemin de l'Irat, Ligne Paradis, 97410 Saint Pierre-La Réunion



**Résumé :** Le **Huanglongbing (HLB)** est une maladie bactérienne causant le dépérissement des agrumes. Il y a une résurgence de cette maladie sur l'île de la Réunion depuis 2012, ce qui menace les vergers d'agrumes. Les techniques de détections classiques de surveillance de la maladie par PCR sont coûteuses, longues et ne peuvent pas être réalisées à grande échelle. C'est dans ce contexte qu'intervient la technologie de l'analyse par **imagerie spectrale**. Cette technologie se base sur l'analyse de la signature spectrale qu'émet un support en réponse à une exposition lumineuse. Cette méthode est non destructive, en plus d'être relativement peu coûteuse. Les analyses ont permis de mettre en évidence une influence assez faible des parcelles et des variétés sur les longueurs d'onde où se détecte le HLB. Cela rend l'effet du HLB détectable sans bruits et donc utilisable pour faire du diagnostic. S'ajoute à cela un traitement des données d'imagerie spectrales par **apprentissage supervisé** dans le but de prédire le statut des arbres vis-à-vis de la maladie au sein des parcelles d'agrumes. Ce modèle est prometteur avec une qualité de la prédiction de 92.6% pour la méthode des Moindres Carrés Partiels (PLS) sur une base d'apprentissage de 8400 spectres de réflectance.

**Mots clés :** Apprentissage supervisé, Epidémiologie, Forêts Aléatoires (RF), HLB = Huanglongbing, Machine à Vecteurs de Support (SVM), Régression par les Moindres Carrés Partiels (PLS), Spectroscopie, Télédétection.

**Abstract:** **Huanglongbing (HLB)** is a bacterial disease causing dieback in citrus fruits. There has been a resurgence of this disease on Reunion Island since 2012, which threatens citrus orchards. The conventional detection techniques for disease surveillance by PCR are expensive, time consuming and cannot be carried out on a large scale. This is where the technology of **spectral imaging** analysis comes in. This technology is based on the analysis of the spectral signature that a medium emits in response to light exposure. This method is non-destructive, in addition to being relatively inexpensive. The analyzes made it possible to highlight a fairly weak influence of the plots and varieties on the wavelengths where HLB is detected. This makes the effect of HLB detectable without noise and therefore usable for diagnosis. Added to this is a processing of spectral imaging data by **machine learning** in order to predict the status of trees vis-à-vis the disease within citrus plots. This model is promising with a prediction quality of 92.6% for the Partial Least Squares (PLS) method on a training basis of 8400 reflectance spectra.

**Keywords:** Epidemiology, Machine learning, HLB = Huanglongbing, Partial Least Squares regression (PLS), Random Forests (RF), Remote sensing, Spectroscopy, Support Vector Machine (SVM).



# Table des matières

<b>Table des matières</b>	<b>v</b>
<b>Remerciements</b>	<b>vii</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Contexte général . . . . .	1
1.2 Etat des connaissances sur la maladie . . . . .	3
1.3 La résurgence du HLB à la Réunion . . . . .	5
<b>2 Matériel et Méthodes</b>	<b>7</b>
2.1 Identification de la maladie . . . . .	7
2.2 Choix des arbres . . . . .	8
2.3 Spectroscopie à main . . . . .	11
2.4 Traitement des données par analyses statistiques et apprentissage supervisé . . . . .	14
2.5 Forêts Aléatoires (RF) . . . . .	15
2.6 Machine à Vecteurs de Support (SVM) . . . . .	17
2.7 Régression par les Moindres Carrés Partiels (PLS)	18
2.8 Performance des classifications . . . . .	19
2.9 Amélioration du protocole de terrain . . . . .	19
<b>3 Résultats</b>	<b>21</b>
3.1 Influence du lieu d'échantillonnage et des variétés sur le spectre de réflectance global . . . . .	21
3.2 Effet du statut HLB sur les spectres de réflectance	23
3.3 L'approche par arbre de décision . . . . .	26
3.4 Comparaison des performances des trois méthodes de prédiction du statut HLB à partir des spectres réflectance . . . . .	28
3.5 Prédiction du statut HLB par Régression par les Moindres Carrés Partiels (PLS) . . . . .	29
3.6 Amélioration du protocole de terrain pour le choix du nombre de feuilles par arbre . . . . .	31
3.7 Amélioration du protocole de terrain pour le choix du nombre de mesures de réflectance par feuille . .	32
<b>4 Discussion</b>	<b>33</b>
<b>5 Conclusion</b>	<b>37</b>

<b>6 Annexe</b>	<b>39</b>
6.1 Annexe 1 : Importation des données brutes . . . . .	40
6.2 Annexe 2 : Fonction Matrice de confusion . . . . .	41
6.3 Annexe 3 : Matrice de confusion des 3 méthodes de machin learning . . . . .	44
6.4 Annexe 4 : Fonction nombre de feuille . . . . .	45
6.5 Annexe 5 : Prédiction du nombre optimal de feuille	46
6.6 Annexe 6 : Fonction nombre de répétition SPIR par feuille . . . . .	47
6.7 Annexe 7 : Prédiction du nombre de répétition op- timal . . . . .	48
<b>Bibliographie</b>	<b>49</b>

# Remerciements

Je tiens à remercier toutes les personnes qui ont contribué au bon déroulement de ce stage :

- Virginie RAVIGNE, mon encadrante de stage pour son encadrement, sa gentillesse , ses conseils et l'autonomie qu'elle m'a laissée au cours de ce stage afin que je développe mes propres axes de recherche sur cette thématique ;
- Frédéric CHIROLEU, chercheur en biostatistiques, co-auteur de ce rapport, et Thuy-Trang CAO, pour leur grande aide à la réalisation des scripts R et toutes les choses apprises lors de ces rendez-vous ;
- Olivier PRUVOST, pour son aide lors des analyse qPCR et son encadrement au laboratoire ;
- Karine BOYER, pour m'avoir conseillé et formé aux techniques de laboratoire ;
- Ismaël HOUILLON, doctorant, pour ses conseils et ses jeux de mots douteux ;
- Elisa PAYET, technicienne à la FDGDON, pour avoir pris le temps de nous communiquer leurs résultats des détections du HLB et les numéros des agriculteurs ;
- Raphaël SOLESSE, ainsi que Emmanuel TILLARD, pour leurs conseils, sur le pilotage de drones et la formation à l'utilisation des instruments de mesure de spectrométrie ;
- Louis-Axel EDOUARD RAMBAUT, pour son aide sur les scripts R et ses discussions pertinentes ;
- Claire MELOT, stagiaire au CIRAD, pour sa bonne humeur et son aide aux analyses de laboratoire tardives ;
- Le CIRAD, l'équipe du 3P de Saint-Pierre et en particulier les agents de l'UMR PVBMT (Unité Mixte de Recherche Peuplements Végétaux et Bioagresseurs en Milieux Tropicaux) ;
- Les stagiaires des Kazz pour leur amitié, les mangues et toutes les choses partagées ensemble qui ont rendu ce stage inoubliable !



# CHAPITRE 1

## Introduction

### 1.1 Contexte général

La culture d'agrumes est la principale culture fruitière au monde.<sup>1</sup> Cette culture fait face à une crise sans précédent à l'échelle mondiale due au Huanglongbing, couramment appelée HLB ou « greening », une maladie bactérienne qui est la plus dévastatrice des maladies des agrumes.<sup>2</sup> Le HLB se propage rapidement et a un impact dévastateur sur la production d'agrumes dans les principales régions productrices d'agrumes du monde telles que la Chine, le Brésil, l'Inde et les Etats-Unis.<sup>3</sup> Cette maladie a été identifiée pour la première fois en Chine il y a un siècle et depuis, 51 des 140 principaux pays producteurs d'agrumes ont été infectés par la maladie.<sup>4</sup> Cependant, d'importantes régions productrices d'agrumes comme la région méditerranéenne et l'Australie sont pour le moment épargnées par cette maladie.<sup>5</sup> Par ailleurs, *Trioza erytreae*, un des insectes vecteurs de la maladie a été trouvé dans la péninsule ibérique et pose une menace immédiate de propagation du HLB dans la région méditerranéenne. En Guadeloupe le HLB a décimé la quasi-totalité des vergers et fait chuter la production agrumicole de 70% au début de l'année 2020.<sup>6</sup> Le HLB est donc devenu un problème d'échelle mondiale dont la surveillance est devenue primordiale afin de limiter ses futures propagations.<sup>7</sup>

<sup>1</sup>A. Comte (2013). « Apport des marqueurs SSRs nucléaires, des InDel mitochondriaux et de la diversité allélique de gènes candidats pour la tolérance à la salinité ». fr. THESE DE DOCTORAT EN SCIENCES AGRONOMIQUES Spécialité : Sciences de la Production Végétale. Tunis : INSTITUT NATIONAL AGRONOMIQUE DE TUNISIE.

<sup>2</sup>É. A. S. Moriya et al. (juin 2019). « DETECTING CITRUS HUANGLONGBING IN BRAZILIAN ORCHARDS USING HYPERSPECTRAL AERIAL IMAGES ». en. In : *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W13, p. 1881-1886. DOI : [10.5194/isprs-archives-XLII-2-W13-1881-2019](https://doi.org/10.5194/isprs-archives-XLII-2-W13-1881-2019). (Visité le 19/04/2021).

<sup>3</sup>X. Deng et al. (août 2020). « Detection of Citrus Huanglongbing Based on Multi-Input Neural Network Model of UAV Hyperspectral Remote Sensing ». en. In : *Remote Sensing* 12.17, p. 2678. DOI : [10.3390/rs12172678](https://doi.org/10.3390/rs12172678). (Visité le 09/04/2021).

<sup>4</sup>Moriya et al. (2019). Cf. note 2.

<sup>5</sup>A. P. Gutierrez et L. Ponti (déc. 2013). « Prospective Analysis of the Geographic Distribution and Relative Abundance of Asian Citrus Psyllid (Hemiptera : Liviidae) and Citrus Greening Disease in North America and the Mediterranean Basin ». en. In : *Florida Entomologist* 96.4, p. 1375-1391. DOI : [10.1653/024.096.0417](https://doi.org/10.1653/024.096.0417). (Visité le 06/04/2021).

<sup>6</sup>C. Morillon (2020). *Huanglongbing the citrus disease*. en.

<sup>7</sup>N. Wang (mai 2019). « The Citrus Huanglongbing Crisis and Potential Solutions ». en. In : *Molecular Plant* 12.5, p. 607-609. DOI : [10.1016/j.molp.2019.03.008](https://doi.org/10.1016/j.molp.2019.03.008). (Visité le 06/04/2021).

L'un des problèmes concernant la surveillance et la lutte contre cette maladie est que les différents travaux de recherche apportant des éléments concernant la lutte contre cette maladie ont été menés sur des paysages homogènes de grandes surfaces de vergers en Californie, en Floride ou au Brésil.<sup>8</sup> Ces travaux sont donc difficilement applicables à des paysages très hétérogènes comme à la Réunion.<sup>9</sup> En effet, l'île de la Réunion est un territoire de 2 512 km<sup>2</sup> qui possède un fort gradient altitudinal, de 0 à 3070m d'altitude. S'ajoutent à cela les fortes différences d'hygrométrie entre la côte au vent à l'est et la côte sous le vent à l'ouest.<sup>10</sup> L'hétérogénéité éco-climatique étant importante sur ce territoire, la diversité des paysages est donc un aspect fondamental pour la surveillance de cette maladie.

<sup>8</sup>H. A. Narouei-Khandan et al. (mars 2016). « Global climate suitability of citrus huanglongbing and its vector, the Asian citrus psyllid, using two correlative species distribution modeling approaches, with emphasis on the USA ». en. In : *European Journal of Plant Pathology* 144.3, p. 655-670. doi : [10.1007/s10658-015-0804-7](https://doi.org/10.1007/s10658-015-0804-7). (Visité le 28/03/2021).

<sup>9</sup>T. R. Gottwald (2010). « Current Epidemiological Understanding of Citrus Huanglongbing ». en. In : *U.S. Department of Agriculture* 48.1, p. 39-119.

<sup>10</sup>C. Guilloteau (juill. 2018). *Utilisation de la connaissance du paysage agricole pour l'accompagnement des réseaux d'épidémiosurveillance : application au greening des agrumes à la Réunion*. Rapport de stage. Paris : AgroParisTech, p. 39.

<sup>11</sup>J. Leung (2014). *La production fruitière à La Réunion*. fr. fr.

<sup>12</sup>Guilloteau (2018). Cf. note 10.

<sup>13</sup>Ibid.

Depuis 2012, le département de la Réunion avec son agriculture traditionnelle en agrumiculture est menacé par la résurgence de cette maladie (V. Ravigné, comm. pers.). A la Réunion les vergers d'agrumes sont fragmentés et ne représentent une superficie totale que de 300 ha environ.<sup>11</sup> Les vergers sont donc répartis en petites parcelles présentes sur des altitudes comprises entre 0 et 1200m.<sup>12</sup> Parce que la filière est peu organisée, les vergers d'agrumes sont mal répertoriés par les services de l'Etat et un travail important de l'équipe d'accueil ces dernières années a consisté à reconstruire le parcellaire agrumicole en combinant photo-interprétation d'images aériennes et vérifications sur le terrain. On dénombre, à l'issue de ce travail 682 vergers d'agrumes dont 248 infectés par le HLB.<sup>13</sup> Depuis 2015, la DAAF (Direction de l'Alimentation, de l'Agriculture et de la Forêt) organise une surveillance du parcellaire agrumicole et est chargée de l'arrachage des plants contaminés. La DAAF a délégué à la FDGDON (Fédération Départementale des Groupements de Défense contre les Organismes Nuisibles) les prélèvements de matériel végétal ainsi que les tests officiels de détection de la maladie. En parallèle, l'ANSES (Agence Nationale de Sécurité Sanitaire de l'Alimentation, de l'Environnement et du Travail) en charge de la mise au point de méthodes officielles de détection est en coopération avec le CIRAD (Centre de Coopération Internationale en Recherche Agronomique pour le Développement) au travers de son UMR PVBMT (Unité Mixte de Recherche Peuplements Végétaux et Bioagresseurs en Milieu Tropical) pour l'amélioration de la détection et de la surveillance de la maladie.

## 1.2 Etat des connaissances sur la maladie

Le HLB est une maladie causée par la bactérie *Candidatus Liberibacter* associée à trois sous-espèces (*americanus*, *africanus* et *asiaticus*).<sup>14</sup> C'est la sous-espèce *Candidatus Liberibacter asiaticus* que l'on retrouve sur l'île de la Réunion.<sup>15</sup> Cette bactérie de type Gram- (non cultivable) se développe dans les tissus riches en glucides du phloème. Cela va entraîner des dépôts de callose dans le phloème et obstruer les vaisseaux, affaiblissant, à terme, l'arbre et entraînant sa mort.<sup>16</sup> Le HLB est spécifique des agrumes et essentiellement transmis par les psylles qui sont des insectes de l'ordre des hémiptères. Deux espèces de psylles sont connues pour véhiculer la bactérie responsable de la maladie : *Diaphorina citri* et *Trioza erytreae*. Seule la première a jusqu'ici été identifiée dans l'épidémie qui sévit actuellement à la Réunion. Ces deux espèces de psylles réalisent une partie de leurs cycles de développement sur les arbres de la famille des Rutacées dont font partie les agrumes cultivés. La bactérie est transmise d'un arbre à un autre via la salive de l'insecte en se nourrissant de la sève élaborée circulant dans le phloème.<sup>17</sup> Ce mécanisme de transmission de la maladie peut s'étendre sur une distance de 25 à 50 m autour de l'arbre infecté. Cette distance peut atteindre les 1,5 km en cas de vent, ce qui favorise grandement la dispersion de la maladie.<sup>18</sup> Par ailleurs, en plus de la transmission par les psylles, le HLB peut aussi être véhiculé par les greffes et les plantes en pépinière déjà contaminées et vendues aux agriculteurs et particuliers, ce qui fait de la gestion des pépinières un élément clé de la lutte contre la propagation de la maladie. Le HLB affecte tous les agrumes, mais à des degrés de sensibilités différentes. La maladie provoque un dépérissement des arbres avec des symptômes visibles ressemblant à un stress hydrique ou une carence comme le jaunissement asymétrique des feuilles et des flushs (jeunes feuilles, figure 1.1<sup>19</sup>).<sup>20</sup> Ces symptômes traduisent un dépérissement de l'arbre avec des fruits petits et de forme asymétrique.<sup>21</sup>



FIG. 1.1 : Symptôme du HLB sur flush, feuille, fruit et sur l'arbre

<sup>14</sup>Wang (2019). « The Citrus Huanglongbing Crisis and Potential Solutions », cf. note 7, p. 1.

<sup>15</sup>B. Aubert (1989). « Le greening une maladie infectieuse des agrumes d'origine bactérienne transmise par des homoptères psyllidés Stratégies de lutte développées à l'île de la Réunion, Circonstances épidémiologiques en Afrique Asie et modalités d'intervention ». THÈSE DE DOCTORAT EN SCIENCES AGRONOMIQUES Spécialité : Sciences de la Production Végétale. INRA Bordeaux : Université Bordeaux 2.

<sup>16</sup>J. Bové (2006). « HUANGLONGBING : A DESTRUCTIVE, NEWLY-EMERGING, CENTURY-OLD DISEASE OF CITRUS ». en. In : *Journal of Plant Pathology* 88.1, p. 7-37.

<sup>17</sup>Narouei-Khandan et al. (2016). Cf. note 8.

<sup>18</sup>Bové (2006). Cf. note 16.

<sup>19</sup>Guilloteau (2018). Cf. note 10.

<sup>20</sup>Bové (2006). Cf. note 16.

<sup>21</sup>T. R. Gottwald (1989). « Preliminary Analysis of Citrus Greening (Huanglungbin) Epidemics in the People's Republic of China and French Reunion Island ». en. In : *Phytopathology* 79.6, p. 687. DOI : 10.1094/Phyto-79-687. (Visité le 05/03/2021).

<sup>22</sup>Gottwald (1989). « Preliminary Analysis of Citrus Greening (Huanglongbing) Epidemics in the People's Republic of China and French Réunion Island », cf. note 21, p. 3.

<sup>23</sup>Aubert (1989). « Le greening une maladie infectieuse des agrumes d'origine bactérienne transmise par des homoptères psyllidés Stratégies de lutte développées à l'île de la Réunion, Circonstances épidémiologiques en Afrique Asie et modalités d'intervention », cf. note 15, p. 3.

<sup>24</sup>Bové (2006). « HUANGLONGBING : A DESTRUCTIVE, NEWLY-EMERGING, CENTURY-OLD DISEASE OF CITRUS », cf. note 16, p. 3.

<sup>25</sup>Ibid.

<sup>26</sup>Wang (2019). « The Citrus Huanglongbing Crisis and Potential Solutions », cf. note 7, p. 1.

Cependant, les symptômes ne permettent pas l'identification de la maladie avec certitude.<sup>22</sup> En effet, un jaunissement des feuilles peut également être causé par des carences en zinc ou en magnésium.<sup>23</sup> Par ailleurs, la difficulté de la détection de cette maladie réside en partie dans la période de latence qui peut durer 6 à 12 mois et durant laquelle les arbres sont parfaitement asymptomatics tout en transmettant la maladie.<sup>24</sup> Il a été montré que dans un verger avec 30% d'arbres présentant des symptômes, 50% des arbres sont déjà infectés.<sup>25</sup> De plus, il est pour le moment difficile pour les biologistes de bien comprendre le HLB. Les défis actuels résident principalement dans le fait que *Candidatus Liberibacter* ne peut être cultivée par aucune méthode *in vitro* en laboratoire. Par conséquent, les analyses moléculaires et génétiques traditionnelles ont une portée limitée pour l'étude de cette bactérie.<sup>26</sup>

## 1.3 La résurgence du HLB à la Réunion

Le HLB avait fait l'objet d'une importante campagne d'arrachage dans les années 1970-1980 en parallèle d'une lutte chimique basée sur des produits aujourd'hui interdits (notamment des antibiotiques). Ces méthodes accompagnées d'une lutte biologique et d'un replantage de plants sains ont permis d'éradiquer cette première épidémie, une réussite unique au monde.<sup>27</sup> La lutte biologique a reposé sur l'introduction de parasitoïdes des psylles permettant un contrôle des populations de psylles vecteurs de la maladie, encore actifs aujourd'hui.<sup>28</sup> Cependant, depuis 2012, de nouveaux cas de HLB ont été observés sur l'île. Les parasitoïdes étant toujours présents, aucune pullulation de psylles n'est relevée et c'est uniquement *D. citri* qui est observé.<sup>29</sup> Si la cause de cette résurgence ne vient pas d'une pullulation de la population de psylles, elle pourrait provenir de vergers comportant des arbres malades non arrachés lors de la première épidémie, d'un relâchement des pratiques en pépinière ou d'une introduction accidentelle.<sup>30</sup>

En l'absence de traitement, l'arrachage précoce des arbres malades est pour l'instant le moyen le plus efficace de lutte contre cette épidémie. Cependant, cette méthode est coûteuse en moyens humains et financiers. L'arrachage lui-même est une procédure complexe et après plantation, un verger ne rentre en production qu'au bout de plusieurs années (4 à 6 ans), ce qui constitue un manque à gagner important. Les aides financières sont très limitées et ne concernent actuellement que la replantation. Dès lors, l'acceptabilité de la lutte par arrachage est mauvaise. De plus, la technique officielle de détection et surveillance de la maladie par PCR est coûteuse, de réalisation laborieuse et peut donc difficilement être déployée à grande échelle. De ce fait, la stratégie de surveillance officielle actuelle qui est d'analyser 60 arbres à l'hectare ne permet pas aux agriculteurs de savoir précisément le statut de chaque arbre dans leur parcelle, ce qui diminue encore l'acceptabilité des mesures d'arrachage. Permettre aux agriculteurs de disposer d'une information sur le statut de chaque arbre dans leur parcelle est un levier essentiel pour les amener à prendre les décisions difficiles que l'épidémie leur impose.

D'autres moyens de surveillance doivent donc être mis en place. C'est dans ce contexte qu'intervient la technologie de l'analyse par imagerie spectrale. Cette technologie se base sur l'analyse de la signature spectrale mesurée en réflectance qu'émet un support en réponse à une exposition lumineuse.<sup>31</sup> Cette technologie présente plusieurs avantages prometteurs pour la détection du HLB. En effet, cette méthode est non destructive pour le support traité, est relativement peu coûteuse par rapport à la surface couverte et l'efficacité de l'analyse est importante à condition de trouver un calibrage optimal pour détecter la maladie.<sup>32</sup>

<sup>27</sup>Gottwald (1989). « Preliminary Analysis of Citrus Greening (Huanglongbing) Epidemics in the People's Republic of China and French Reunion Island », cf. note 21, p. 3.

<sup>28</sup>B. Aubert et al. (1996). « A Case Study of Huanglongbing (Greening) Control in Reunion ». en. In : *Thirteenth IOCVC Conference* 9.1, p. 3.

<sup>29</sup>Guilloteau (2018). *Utilisation de la connaissance du paysage agricole pour l'accompagnement des réseaux d'épidémi-o-surveillance : application au greening des agrumes à la Réunion*, cf. note 10, p. 2.

<sup>30</sup>Ibid.

<sup>31</sup>S. Sankaran et al. (fév. 2013). « Huanglongbing (Citrus Greening) Detection Using Visible, Near Infrared and Thermal Imaging Techniques ». en. In : *Sensors* 13.2, p. 2117-2130. DOI : 10.3390/s130202117. (Visité le 09/04/2021).

<sup>32</sup>S. Sankaran et R. Ehsani (nov. 2011). « Visible-near infrared spectroscopy based citrus greening detection : Evaluation of spectral feature extraction techniques ». en. In : *Crop Protection* 30.11, p. 1508-1513. DOI : 10.1016/j.cropro.2011.07.005. (Visité le 04/09/2020).

Réalisée par des spectromètres à main, cette technique pourrait en principe permettre d'augmenter significativement les cadences de détection de la maladie et ainsi fournir à bas coût une information complète aux agriculteurs touchés quant à l'état de leur verger. Des tentatives de détecter le HLB par ses signatures spectrales existent dans la littérature scientifique<sup>33</sup> et elles sont encourageantes mais le contexte réunionnais est particulier. Les nombreuses variétés locales n'ont jamais été testées dans la littérature, même les plus importantes (tangor, mandarine zanzibar, orange navel, citron quatre saisons...) car les recherches sont principalement concentrées sur les citronniers, les orangers et les mandariniers<sup>34 35</sup>. Comme les agrumes sont souvent une culture secondaire dans des exploitations centrées sur une autre culture (canne, maraîchage), les arbres sont souvent cultivés de façon non optimale, peu technique et donc peuvent souffrir de carences nutritives et de déficit hydrique. Ces éléments font que les signatures spectrales du HLB pourraient être moins claires à la Réunion que dans les conditions testées ailleurs. Au cours de ce stage, je répondrai ainsi à la question suivante : **Peut-on développer une méthode de la télédétection par imagerie spectrale fiable pour améliorer la surveillance épidémiologique du Huanglongbing (HLB) à La Réunion ?**

L'objectif global va donc être de prédire le statut positif ou négatif au HLB des arbres directement avec des outils de spectro-métrie.

Pour résoudre cette problématique, je traiterai les objectifs spécifiques suivants :

- 1) Les différentes parcelles et variétés analysées ont elles une influence sur les résultats visant à prédire l'état sanitaire des arbres ?
- 2) Quelles longueurs d'onde sont les plus discriminantes pour la recherche de la maladie par spectrométrie ?
- 3) Peut-on connaître l'état sanitaire d'un arbre via une analyse par spectrométrie ? Sous quel degré de précision ?

## CHAPITRE 2

# Matériel et Méthodes

### 2.1 Identification de la maladie

Pour la mise au point de la méthode, les feuilles analysées doivent avoir un statut connu vis-à-vis de la maladie. Pour ce faire, la méthode de détection par PCR (Polymerase Chain Reaction) est pour l'instant la plus utilisée pour surveiller l'évolution de la maladie. Cette méthode permet de détecter l'ADN de la bactérie à partir d'échantillons de feuilles.<sup>1</sup> L'inconvénient de cette méthode est qu'elle ne permet pas le traitement d'un volume important d'échantillons et ne donne pas une réponse immédiate sur le prélèvement effectué sur le terrain. La PCR est une technique reconnue de multiplication de l'ADN d'un échantillon. Le but est de dupliquer les échantillons d'ADN prélevés à chaque cycle de réaction via l'ADN polymérase. Ainsi, avec un seul brin d'ADN de départ, des millions de répliques sont produites à chaque cycle de réplication. La qPCR repose sur le même principe avec des améliorations de la technique de départ. Le principe est toujours de multiplier l'ADN de l'échantillon, mais avec cette fois une quantification du nombre de cycles PCR nécessaires à la détection de la maladie. Le nombre de cycles de réplication ou "Cycle Threshold" (CT) correspond au nombre de cycles PCR à partir duquel le produit de la PCR est détectable. Ainsi, plus le CT est faible plus il y a d'ADN présent initialement et inversement. Les échantillons sains et malades peuvent donc être distingués selon leur valeur de CT selon un seuil fixé. La méthode officielle actuelle basée sur la PCR classique est actuellement en cours de remplacement par une méthode basée sur la qPCR et qui utilisera le seuil de 36 cycles de détection maximum pour identifier un échantillon comme étant porteur de la maladie. C'est ce seuil de 36 cycles ou CT36 qui sera utilisé dans cette étude pour statuer de l'état malade ou sain des arbres. Chaque échantillon est testé 3 fois (on parle de triplicats techniques) et chaque plaque de 96 puits de qPCR contient trois témoins négatifs et deux témoins positifs. Lorsque les témoins donnent des résultats inattendus ou que les triplicats ne sont pas cohérents, les échantillons sont repassés.

<sup>1</sup>Gottwald (2010). « Current Epidemiological Understanding of Citrus Huanglongbing », cf. note 9, p. 2.

## 2.2 Choix des arbres

Le choix des parcelles échantillonnées a été conditionné par la recherche d'un équilibre entre les arbres malades et sains sur chaque parcelle basée sur les données de la surveillance officielle et du Cirad. Le défi est que les statuts des arbres ne sont pas connus à l'avance et il peut être compliqué de faire un design équilibré en amont de la récolte de données. L'objectif a donc été dans un premier temps de trouver les agriculteurs qui possèdent les trois variétés d'agrumes sur lesquelles se base cette étude, à savoir les citrons (*Citrus limon*), les mandarines zanzibar (*Citrus reticulata*) et les tangors (*Citrus reticulata x sinensis*). Dans un deuxième temps, pour chacun de ces agriculteurs il a fallu identifier précisément le statut des arbres.

Concernant les caractéristiques altitudinales et pluviométriques des parcelles, celle de Mr Hoarau et Mr Pothin échantillonnées à Petite-Ile sont comprises dans une échelle de pluviométrie identique la parcelle de Mr Gonthier échantillonnée dans les hauts du Tampon à savoir entre 1430 et 1680 mm de pluie par an.<sup>2</sup> Leurs altitudes respectives sont cependant très différentes avec une différence de 1000 m et plus entre la parcelle du Tampon et les autres (figure 2.1). La parcelle de Mr Barret située au niveau de la mer est la parcelle qui reçoit le moins de précipitations avec seulement 930 à 960 mm de pluie par an.<sup>3</sup>

<sup>2</sup>C. Equipe Artists - UR AïDA (2021). METEOR. URL : <https://smartis.re/METEOR> (visité le 08/04/2021).

<sup>3</sup>Ibid.

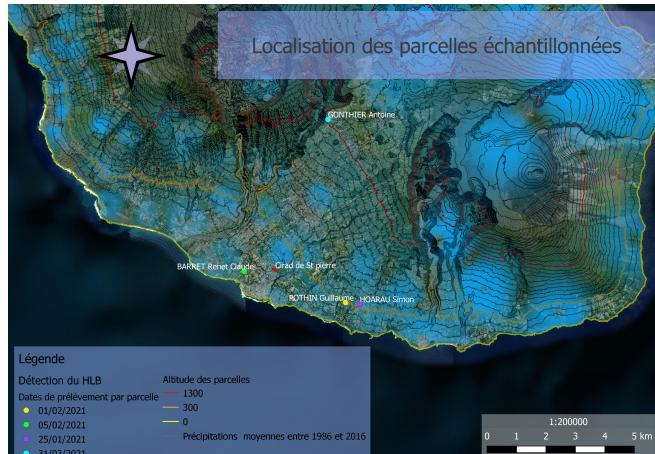


FIG. 2.1 : Carte présentant les parcelles échantillonnées, leurs dates de prélèvement ainsi que la pluviométrie et les altitudes

<sup>4</sup>Guilloteau (2018). *Utilisation de la connaissance du paysage agricole pour l'accompagnement des réseaux d'épidémi-o-surveillance : application au greening des agrumes à la Réunion*, cf. note 10, p. 2.

L'échantillonnage a commencé dans la commune de Petite-Île, qui est la plus grosse zone de production d'agrumes de l'île (environ 25% de la surface en agrumes).<sup>4</sup> Cette zone est aussi très touchée par le HLB (près de deux parcelles sur trois) avec par conséquent, des prospections importantes de la part de la DAAF et de la FDGDON qui fournissent des données précieuses pour l'échantillonnage. Les parcelles de Mr Hoarau et Mr Pothin échantillonnées à Petite-Île possèdent les trois variétés recherchées mais ont une répartition différente entre les arbres malades et sains.

Dans la parcelle de Mr Hoarau l'échantillonnage s'est basé sur une première cartographie des statuts HLB des arbres réalisée dans le cadre d'une expérimentation de la FGDGON pour le développement d'une autre méthode moléculaire. 42 arbres ont ainsi été sélectionnés dont 19 arbres se sont révélés sains et 23 arbres se sont révélés infectés par le HLB (figure 2.2).



FIG. 2.2 : Échantillonnage réalisé dans la parcelle de Mr Hoarau en 2021

A l'inverse dans la parcelle de Mr Pothin, l'échantillonnage des arbres s'est basé sur les dires d'expert de l'agriculteur et trois arbres sains pour 39 arbres infectés par le HLB ont été trouvés (figure 2.3).

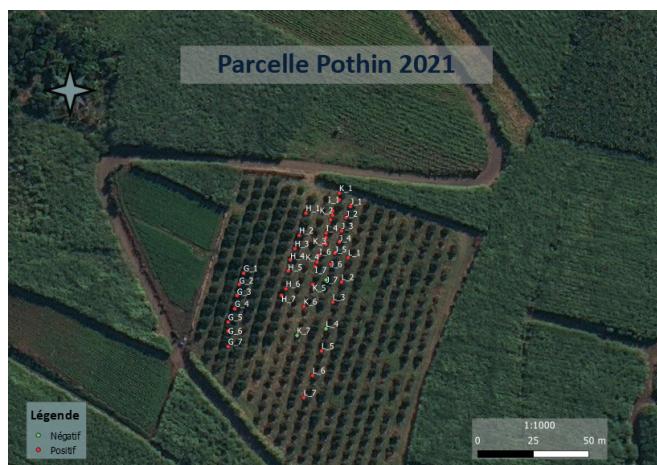


FIG. 2.3 : Échantillonnage réalisé dans la parcelle de Mr Pothin en 2021

La parcelle de Mr Barret située à Saint Pierre et constituée uniquement de citronniers a été échantillonnée de manière exhaustive. Cette parcelle fut la plus analysée en laboratoire. Les 156 arbres de la parcelle ont d'abord été regroupés par pools de 5 puis certains arbres ont été analysés individuellement. Au final 10 citronniers sains et quatre infectés par le HLB ont été trouvés (figure 2.4).



FIG. 2.4 : Échantillonnage réalisé dans la parcelle de Mr Barret en 2021

A l'issue de cet échantillonnage, un grand nombre d'arbres infectés chez Mr Pothin et Barret ont été trouvés, il y avait donc un déficit d'arbres sains. La parcelle de Mr Gonthier située au Tampon a donc été trouvée pour pallier à ce déficit. Cette parcelle compte une répartition égale entre les trois variétés et dont des lots d'arbres ont été testés négatifs aux HLB en novembre 2020 lors de la surveillance officielle. Pour chaque variété 14 arbres ont été échantillonnés (figure 2.5) afin de rééquilibrer l'ensemble du jeu de données. Par manque de temps, ces arbres n'ont pas été re-testés en qPCR et ont été considérés comme négatifs. Cet échantillonnage n'est pas idéal, dans la mesure où les facteurs "agriculteur" et "variété" ne sont pas croisés, et une attention particulière sur ce point devra être portée dans l'interprétation des résultats.

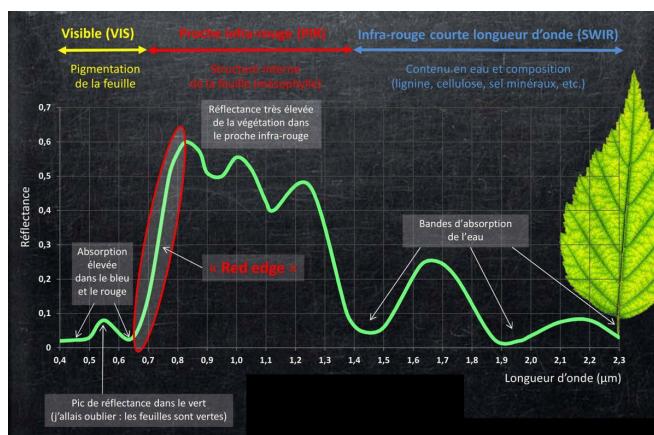


FIG. 2.5 : Échantillonnage réalisé dans la parcelle de Mr Gonthier en 2021

Au final, au cours de cette étude 140 arbres ont été échantillonnés et numérotés de A1 à T7.

## 2.3 Spectroscopie à main

L'objectif de l'utilisation de la spectroscopie est la discrimination des arbres sains et malades de façon rapide et clairement identifiable en passant l'outil de détection sur la parcelle. La détection de la maladie sur les plantes repose sur la capacité des feuilles à absorber, réfléchir ou transmettre la lumière dans différentes longueurs d'onde mesurées en nanomètre (nm). Ces caractéristiques sont liées à la composition biochimique des feuilles. Les variations de mesure de spectre peuvent donc être expliquées dans les différents domaines du spectre par la teneur en pigment des feuilles dans le visible (400-750 nm), par la structure cellulaire avec l'essentiel des vaisseaux dont le phloème dans le proche infrarouge (750-1250 nm) et par la teneur en eau dans le "Short-wave infrared" ou SWIR (1250-2500 nm) comme indiqué sur la (figure 2.6<sup>5</sup>).<sup>6</sup> Les bandes spectrales du visible (bleu, vert, rouge, "red-edge") sont les plus utilisées dans la détection de maladie des feuilles.<sup>7</sup> La qualité des résultats dépend de la résolution du capteur utilisé. Dans le visible (400-750 nm), la coloration des feuilles est due à la teneur en chlorophylles, caroténoïdes et anthocyanes. Les chlorophylles (de type a et b) captent la lumière dans des longueurs d'ondes correspondant au bleu et au rouge. Les caroténoïdes captent la lumière exclusivement dans le bleu et les anthocyanes la captent exclusivement dans le vert. Vers la fin du domaine visible (700 nm) il n'y a plus de captage de lumière aussi important et la lumière est quasiment entièrement réfléchie dans cette zone appelée "red edge". Ensuite, la zone du proche infrarouge (750-1250 nm) correspond principalement à la structure cellulaire interne des feuilles qui est aussi peu affectée en cas de maladie. Cette zone est utile pour la discrimination des feuilles entre feuillus et résineux qui ont des structures de tissus différentes. Enfin, dans le SWIR (1250-2500 nm), c'est principalement la teneur en eau des feuilles qui est mesurée. Plus la réflectance est faible, plus la teneur en eau est grande, ce qui est utile pour caractériser un stress hydrique chez la plante.<sup>8</sup>



<sup>5</sup>F. De Blomac (mars 2014). *Hyperspectral*. fr-FR. Section : 3D. URL : <https://decryptageo.fr/hyperspectral/> (visité le 21/06/2021).

<sup>6</sup>A. Comar (2013). « Etude des interactions feuille/lumière et de leurs implications pour le phénotypage haut débit au champ ». fr. THÈSE DE DOCTORAT EN SCIENCES AGRONOMIQUES Spécialité : Sciences de la Production Végétale. Avignon : Université d'Avignon.

<sup>7</sup>A. Mishra et al. (2007). « Spectral Characteristics of Citrus Greening (Huanglongbing) ». en. In : 2007 Minneapolis, Minnesota, June 17-20, 2007. T. 1. 073056. Minneapolis : American Society of Agricultural and Biological Engineers, p. 10. DOI : 10.13031/2013.24163. (Visité le 18/04/2021).

<sup>8</sup>J. L. Albetis de la Cruz (2018). « Potentiel des images multispectrales acquises par drone dans la détection des zones infectées par la Flavescence dorée de la vigne ». fr. THÈSE DE DOCTORAT EN SCIENCES AGRONOMIQUES Spécialité : Sciences de la Production Végétale. Toulouse : Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier).

FIG. 2.6 : Schéma de la variation de la réflectance de la végétation

<sup>9</sup>Albetis de la Cruz (2018). « Potentiel des images multispectrales acquises par drone dans la détection des zones infectées par la Flavescence dorée de la vigne », cf. note 8, p. 11.

Au cours de cette étude, la détection du statut infecté en HLB des arbres va être réalisé par la méthode de la Spectroscopie Proche Infra Rouge (SPIR) à main qui est catégorisée comme étant de la proxidétection.<sup>9</sup> Dans le cas de la proxidétection, l'acquisition des données se fait par contact direct ou à quelques centimètres de l'objet cible à l'échelle de la feuille ou de la plante (figure 2.7).

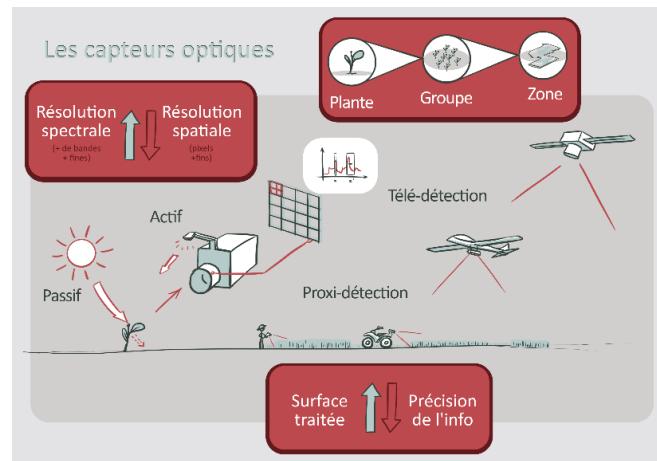


FIG. 2.7 : Schéma d'application des capteurs optiques (Source personnelle)

<sup>10</sup>Comar (2013). « Etude des interactions feuille/lumière et de leurs implications pour le phénotypage haut débit au champ », cf. note 6, p. 11.

Ainsi, pour l'acquisition des données, un spectromètre de terrain (ASD fieldspec Pro) a été utilisé en laboratoire (figure 2.8). L'appareil est utilisé afin de mesurer la réflectance, une mesure correspondant à la proportion (comprise entre 0 et 1) de lumière réfléchie par la surface d'un matériau qui est ici une feuille d'agrume.<sup>10</sup> Afin de mesurer cette réflectance, la feuille disposée sur un disque blanc en Spectralon® qui renvoie quasiment toute la lumière transmise. Cette lumière transmise par l'appareil traverse d'abord la feuille puis est réfléchie et retraverse la feuille avant d'atteindre le capteur. Le Spectralon® est régulièrement nettoyé au papier de verre pour conserver sa valeur de réflectance constante au cours du temps. De plus, les mesures sont réalisées à température ambiante et un blanc est effectué toutes les 15 minutes en réalisant une mesure de SPIR sur le support en Spectralon® seul. L'appareil mesure une bande de longueur d'onde allant de 350 à 2500 nm, la courbe caractérisant la réflectance en fonction des longueurs d'onde mesurées est appelée spectre de réflectance.<sup>11</sup> Pour chaque mesure, le spectre de réflectance correspond à une moyenne de 50 mesures de réflectance effectuées par l'appareil entre 350 et 2500 nm.

<sup>11</sup>Ibid.

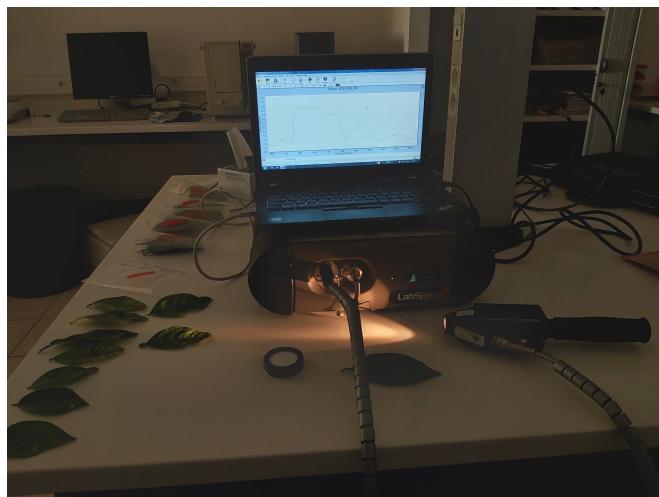


FIG. 2.8 : Photographie d'un spectromètre proche infrarouge à main

Sur chaque arbre 10 feuilles sont prélevées de manière aléatoire autour de l'arbre. Le nombre de 10 feuilles à échantillonner provient d'un compromis entre le temps d'analyse de ces feuilles et de la précision des résultats recherchés. Ce compromis n'aurait pas pu être possible sans les travaux d'un précédent stagiaire, Nathan Créquy qui avait prélevé trois lots de 10 feuilles sur 14 arbres échantillonnés et sur lesquelles des tests sur la baisse de précision ont pu être effectués avant d'aboutir au nombre idéal de feuilles à prélever.<sup>12</sup> Sur chacune des 10 feuilles prélevées par arbre, six mesures de réflectance sont réalisées afin d'avoir des spectres de réflectance issus des différentes zones de la feuille. Ce nombre de mesures optimal a été calculé en amont via les données issues de l'étude de Nathan Créquy qui avait 10 mesures SPIR par feuille et dont les paramètres de précision furent comparés au moyen de tests statistiques pour chaque mesure effectuée par feuille.<sup>13</sup> En fonction des longueurs d'onde étudiées, les capteurs utilisés sont différents et regroupés en deux catégories, les capteurs multispectraux et hyperspectraux. Les capteurs multispectraux permettent d'enregistrer des bandes spectrales qui ne sont pas contiguës et qui peuvent être ciblées, par exemple : le proche infrarouge, le bleu, le rouge et le vert. A l'inverse, les capteurs hyperspectraux enregistrent plus ou moins toutes les bandes spectrales (en fonction de la résolution de l'appareil) de façon contiguë dans un intervalle de longueur d'onde compris entre 350 et 2500 nm.<sup>14</sup> Au cours de cette étude, seul le capteur hyperspectral sera utilisé sur la partie SPIR avec six mesures SPIR par feuille sur les 10 feuilles prélevées par arbre. Le jeu de données (annexe 1) est donc composé des 140 arbres échantillonnés soit 1400 feuilles échantillonées et analysées en SPIR. Cependant, le nombre réel de données comprenant les 6 répétitions SPIR par feuilles échantillonées est de 8400.

<sup>12</sup>N. Créquy (août 2020). *Prise en compte de la structure du paysage pour la surveillance épidémiologique du HLB sur l'île de La Réunion et exploration d'une nouvelle méthode de diagnostic par Spectroscopie Proche Infra-Rouge*. Rapport de stage. Rennes : Agrocampus Ouest, p. 34.

<sup>13</sup>Ibid.

<sup>14</sup>C. Bertiaux (2015). « Mise en place de la spectroscopie proche infrarouge selon les approches Quality By Design et Lean Six Sigma pour le pilotage en ligne de l'humidité d'une poudre ». fr. THÈSE DE DOCTORAT EN SCIENCES PHARMACEUTIQUE. Bordeaux : Université Bordeaux 2 U.F.R. DES SCIENCES PHARMACEUTIQUES.

## 2.4 Traitement des données par analyses statistiques et apprentissage supervisé

<sup>15</sup>R Core Team (2021). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Tous les traitements statistiques sont réalisés sous R.<sup>15</sup>

Les données utilisées pour les analyses sont les valeurs de réflectance issues de la SPIR à main allant de 350 à 2500 nm.

Dans un premier temps, l'objectif va être de caractériser les sources de variations des spectres de réflectance, et en particulier, l'influence des parcelles et des variétés sur ces derniers. Ainsi pour chaque longueur d'onde, une décomposition de la variance est mise en œuvre via une ANOVA à un facteur où la variable expliquée est la réflectance observée pour cette longueur d'onde et la variable explicative est la variété ou la parcelle. Cette analyse produit pour chaque longueur d'onde une valeur du F de Fisher correspondant au rapport entre la variance intermodalités (entre parcelles ou entre variétés) et la variance résiduelle (au sein des parcelles ou des variétés). Plus le F de Fisher est élevé, plus la variable testée (parcelle ou variété) affecte la réflectance. Dans un second temps, une ANOVA à trois facteurs (statut HLB, variété, parcelle) est réalisée afin de montrer l'influence du statut HLB en interaction avec la variété et la parcelle sur le spectre de réflectance global.

Une fois démontrée l'influence ou la non-influence de ces variables de nuisance et l'effet qu'a le statut HLB sur le spectre de réflectance, il est possible d'envisager une prédiction sur le statut HLB des arbres à partir des valeurs de réflectance. Pour ce faire, trois types de méthodes de classification par apprentissage supervisé ont été mises en œuvre : les forêts aléatoires (RF pour « Random Forest »), la machine à vecteurs de support (SVM pour « Support Vector Machine ») et la régression par les moindres carrés partiels (PLS pour « Partial Least Square »).

Un algorithme d'apprentissage supervisé se compose à la fois de données d'entrée “training set” (75% de la base d'apprentissage) et de données cibles “test set” (25% de la base d'apprentissage).<sup>16</sup> Les données d'entrée comprennent des spectres de réflectances associés à un statut HLB. Elles permettent à l'algorithme de « s'entrainer » à reconnaître le statut HLB des arbres à partir des spectres de réflectance. Les données cibles sont composées de spectres pour lesquels le statut HLB est connu mais ne peut être utilisé par l'algorithme car ces données sont utilisées pour la vérification et pour éviter la redondance d'utilisation d'une donnée dans l'apprentissage. Elles permettent de mesurer la performance de l'algorithme. Après s'être assuré de la bonne performance de l'algorithme, on peut l'utiliser pour prédire le statut HLB d'arbres dont on ne dispose que du spectre de réflectance.<sup>17</sup>

<sup>16</sup>J. Tuszynski (2020). *caTools : Tools : Moving Window Statistics, GIF, Base64, ROC AUC, etc.* R package version 1.18.0.

<sup>17</sup>J. Knaus (2015). *snowfall : Easier cluster computing (based on snow)*. R package version 1.84-6.1.

## 2.5 Forêts Aléatoires (RF)

Les RF est une technique utilisant un algorithme d'apprentissage supervisé couplé avec un arbre de décision<sup>18 19</sup>. Un arbre de décision est une construction de choix souvent binaire qui permet de prendre une décision. Cet algorithme permet de créer une forêt d'arbres de décision qui permet d'améliorer la généralisation (sur un paramètre recherché) de l'ensemble du modèle. La RF combine la simplicité de lecture des arbres de décision ainsi que la robustesse de l'apprentissage supervisé qui améliore la précision de la prédiction.<sup>20</sup> La première étape de la construction du RF est la création d'un jeu de données pour chaque arbre de décision (figure 2.9<sup>21</sup>). Ce jeu de données se compose des lignes des données originelles piochées aléatoirement (la même ligne pouvant être sélectionnée plusieurs fois comme la ligne 2 de l'exemple en rouge). Ce procédé augmente la variété des arbres de décision ce qui rend l'algorithme encore plus robuste.<sup>22</sup>

Jeu de donné d'origine				Jeu de donné d'un arbre de décision			
4	80	1	220 000	3	70	0	190 000
3	70	0	190 000	4	60	0	170 000
2	40	1	140 000	3	70	1	200 000
4	60	0	170 000	3	70	0	190 000
3	70	1	200 000				

<sup>18</sup>A. Liaw et M. Wiener (2002). « Classification and Regression by randomForest ». In : *R News* 2.3, p. 18-22.

<sup>19</sup>M. Borkovec et N. Madin (2019). *gparty : 'ggplot' Visualizations for the 'partykit' Package*. R package version 1.0.0.

<sup>20</sup>F. Moutarde (2017). « Arbres de Décision et Forêts Aléatoires ». fr. In : *MINES ParisTech*, p. 20.

<sup>21</sup>Ibid.

<sup>22</sup>Ibid.

FIG. 2.9 : Schéma 1 de construction d'un jeu de données d'un arbre de décision en RF

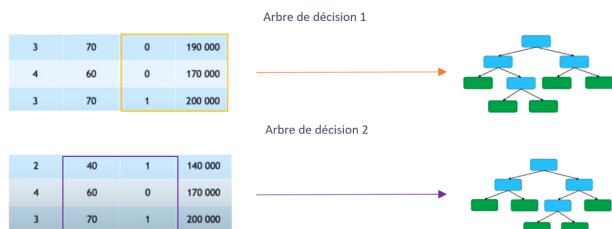
<sup>23</sup>Moutarde (2017). « Arbres de Décision et Forêts Aléatoires », cf. note 20, p. 15.

<sup>24</sup>Ibid.

FIG. 2.10 : Schéma 2 de construction d'un arbre de décision en RF

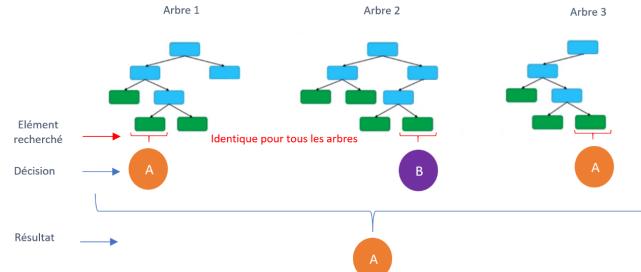
<sup>25</sup>Ibid.

La deuxième étape va être la création de chaque arbre de décision sur une partie différente (aléatoire et non exclusive) des données et des variables à l'intérieur de leurs jeux de données (figure 2.10, rectangles de couleurs<sup>23</sup>). L'intérêt est d'avoir des estimateurs différents dans chaque arbre de décision pour avoir un cheminement différent.<sup>24</sup>



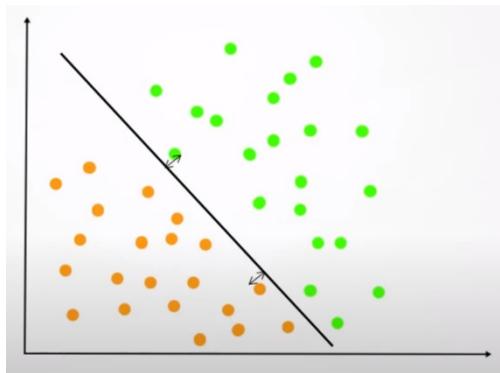
La troisième étape consiste à répéter l'étape 1 et 2 afin de créer autant d'arbres de décision que l'on souhaite dans le but d'obtenir un ensemble (forêt) d'arbres de décision générés aléatoirement (figure 2.11<sup>25</sup>). Le résultat de la RF va correspondre à la moyenne des décisions de l'ensemble des arbres pour l'élément recherché.

FIG. 2.11 : Schéma de construction d'une Forêts Aléatoire



## 2.6 Machine à Vecteurs de Support (SVM)

La SVM est aussi un algorithme d'apprentissage supervisé qui permet de faire des prédictions sur des variables qualitatives ou quantitatives.<sup>26</sup> L'objectif du SVM va consister à séparer deux classes de données, ici (infecté ; sain) pour pouvoir ensuite établir une généralisation dans la prédition quand on ne connaîtra pas la nature de l'échantillon (infecté ; sain). Pour séparer ces deux classes, l'idée est de maximiser les distances entre les échantillons avec un hyperplan séparateur (appelé aussi « support vector ») (figure 2.12<sup>27</sup>).



<sup>26</sup>D. Meyer et al. (2020). e1071 : *Misc Functions of the Department of Statistics, Probability Theory Group (Formerly : E1071)*, TU Wien. R package version 1.7-4.

<sup>27</sup>C. Cortes et V. Vapnik (sept. 1995). « Support-vector networks ». en. In : *Machine Learning* 20.3, p. 273-297. DOI : 10.1007/BF00994018. (Visité le 04/09/2020).

FIG. 2.12 : Schéma de construction d'un hyperplan dans le cadre du SVM

L'apprentissage supervisé va permettre de déterminer l'équation de l'hyperplan qui séparera les jeux de données des individus sains de ceux des malades. Cela va ensuite permettre la prédition pour les jeux de données à tester.

## 2.7 Régression par les Moindres Carrés Partiels (PLS)

Pour finir la PLS est une méthode statistique permettant de traiter une variable binaire à expliquer (ici le statut HLB infecté ou sain) à partir de variables explicatives nombreuses (ici les mesures de réflectance pour chaque longueur d'onde).<sup>28</sup>

<sup>28</sup>B.-H. Mevik et al. (2020). *pls : Partial Least Squares and Principal Component Regression*. R package version 2.7-3.

<sup>29</sup>H. Tenehaus (1999). « L'approche PLS ». In : *Revue de statistique appliquée* 47.2, p. 37.

<sup>30</sup>Ibid.

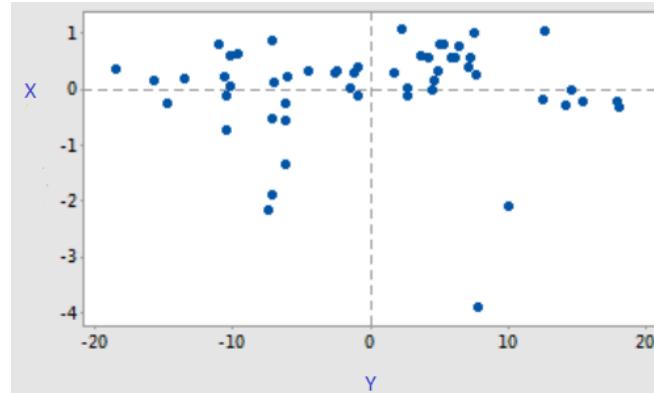


FIG. 2.13 : Tableau représentant les coefficients de régression en fonction de X et Y

L'apprentissage supervisé va permettre de déterminer les équations de régression PLS nécessaires à prédire les jeux de données à tester.

## 2.8 Performance des classifications

En apprentissage supervisé la matrice de confusion mesure la qualité d'un système de prédiction (table 2.1).

TAB. 2.1 : Matrice de confusion

	Négatif confirmé	Positif confirmé
Négatif prédit	Vrai Négatif (VN)	Faux Négatif (FN) (Erreur de type 2)
Positif prédit	Faux positif (FP) (Erreur de type 1)	Vrai positif (VP)

De cette matrice, plusieurs paramètres de performance sont calculés dont :

- La qualité de la prédiction (“Accuracy”) =  $(VP + VN) / (VP + VN + FN + FP)$  met en évidence les erreurs de la prédiction qu'elles soient de type 1 ou de type 2
- La précision (“Precision”) =  $VP / (VP + FP)$  met en évidence les erreurs de type 1

L'erreur de type 1 est problématique dans le cadre de la maladie, car elle implique l'arrachage d'un arbre sain alors prédit comme étant positif au HLB

- La sensibilité (“Sensitivity”) =  $VP / (VP + FN)$  met en évidence les erreurs de type 2

L'erreur de type 2 impacte quant à elle la détection de la maladie, l'arbre est prédit comme étant négatif à la maladie alors qu'il est en réalité malade.

Ces paramètres permettent de rendre compte de l'efficacité de la prédiction et de pouvoir choisir la meilleure méthode de prédiction.

Ces méthodes ont été choisies à partir de la bibliographie où la qualité de la prédiction est par exemple de 85% pour SVM dans l'article de Sindhuja Sankaran<sup>31</sup> et 96.4% pour PLS dans l'article de Xiaoling Deng.<sup>32</sup>

## 2.9 Amélioration du protocole de terrain

Afin d'améliorer le protocole de terrain pour de futures analyses, la question de savoir s'il était possible d'alléger le protocole de SPIR à main (i.e., faire moins de feuilles par arbre, moins de mesures de réflectance par feuille) sans perdre en performances de classification s'est posée. Pour ce faire, 100 jeux de données ont été simulés et dégradés par rééchantillonnage dans le vrai jeu de données et analysés avec les trois méthodes statistiques présentées.

<sup>31</sup>Sankaran et al. (2013). « Huanglongbing (Citrus Greening) Detection Using Visible, Near Infrared and Thermal Imaging Techniques », cf. note 31, p. 5.

<sup>32</sup>Deng et al. (2020). « Detection of Citrus Huanglongbing Based on Multi-Input Neural Network Model of UAV Hyperspectral Remote Sensing », cf. note 3, p. 1.



## CHAPITRE 3

# Résultats

### 3.1 Influence du lieu d'échantillonnage et des variétés sur le spectre de réflectance global

Dans un premier temps, il est intéressant d'observer les données brutes des spectres de réflectance en fonction de leurs variétés. La figure 3.1 montre les spectres de réflectance moyens des variétés citron (en vert), tangor (en orange) et zanzibar (en violet).

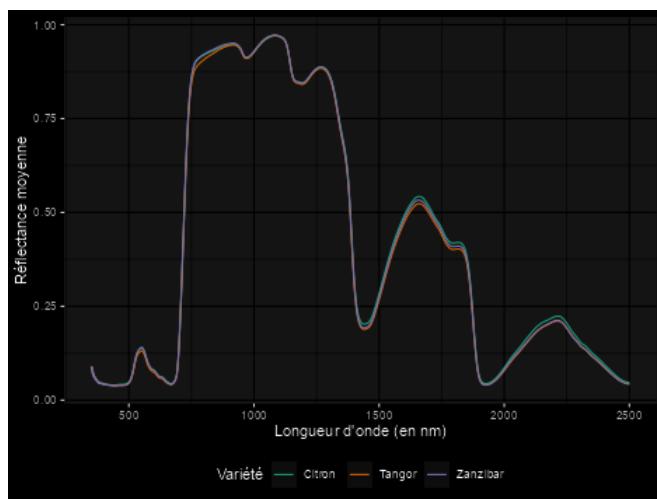


FIG. 3.1 : Spectre moyen en fonction de la variété pour les arbres négatifs aux HLB

Les spectres moyens des variétés échantillonnées ne semblent pas avoir de comportements significativement différents suivant les variétés sur l'ensemble des longueurs d'onde.

En complément, le calcul du F de Fisher via l'ANOVA est donc mis en œuvre afin de montrer un effet potentiel du lieu d'échantillonnage et des variétés sur le spectre de réflectance global (figure 3.2).

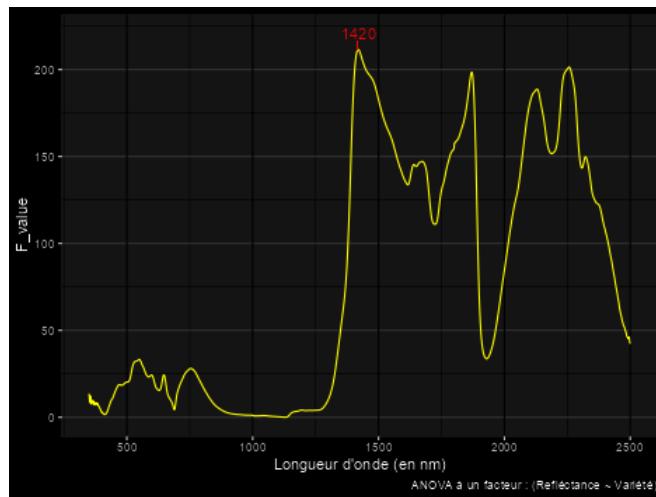


FIG. 3.2 : Valeur du F de Fisher pour chaque longueur d'onde montrant l'influence du facteur variété sur le spectre de réflectance global

L'influence des variétés sur les spectres de réflectance est importante à partir de 1400 nm avec un maximum à 1420 nm (valeur  $F = 230$ ). Ce maximum est suivi de deux autres pics à  $F = 200$  à 1800 nm et 2300 nm, tous compris dans la partie du spectre correspondant aux infrarouges courtes longueurs d'onde.

Concernant l'influence de la parcelle sur le spectre de réflectance global, celle-ci a une forte influence (figure 3.3).

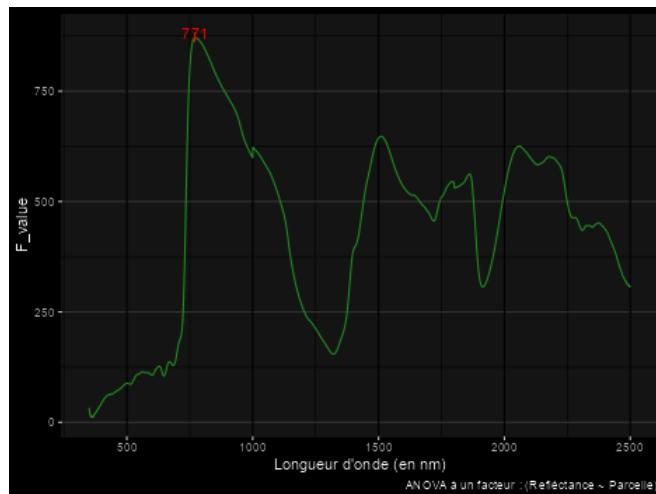


FIG. 3.3 : Valeur du F de Fisher pour chaque longueur d'onde montrant l'influence du facteur parcelle sur le spectre de réflectance global

En effet, dans le “Red edge” l'influence du facteur parcelle est à son maximum à 771 nm (valeur  $F = 875$ ) et a une influence qui reste assez élevée (autour de  $F = 625$ ) autour des longueurs d'onde 1500 nm et 2100 nm.

## 3.2 Effet du statut HLB sur les spectres de réflectance

La répartition des différentes variétés dans le jeu de données en fonction de leurs statuts HLB est assez équilibrée (table 3.1).

TAB. 3.1 : Répartition des variétés dans le jeu de données

Citron.Negatif	25
Citron.Positif	24
Tangor.Negatif	23
Tangor.Positif	23
Zanzibar.Negatif	25
Zanzibar.Positif	20
Total	140

La figure 3.4 montre les spectres de réflectance moyens des arbres positifs au HLB (en rouge) et des arbres négatifs (en vert). On voit des différences nettes de spectre en fonction du statut HLB.

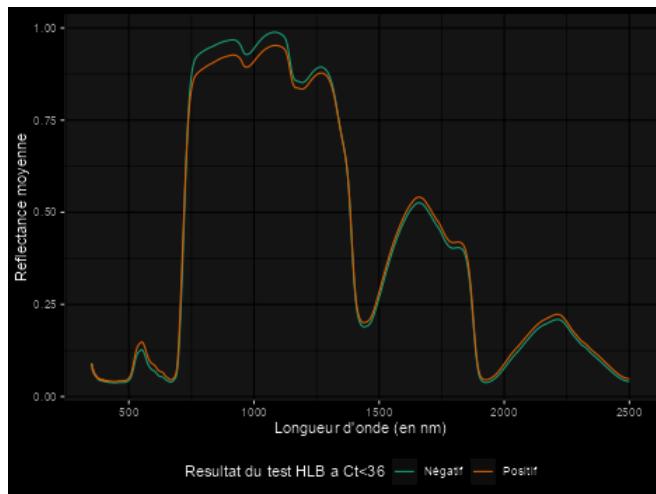


FIG. 3.4 : Spectre moyen en fonction du statut HLB des arbres

Sur certaines parties du spectre des différences de réflectance apparaissent encore plus nettement (figure 3.5).

Sur les longueurs d'onde de 400 à 680 nm les arbres positifs au HLB ont une réflectance légèrement plus élevée que les arbres sains.

Par ailleurs, une séparation moins nette s'observe dans la tranche de longueurs d'onde comprise entre 700 et 1400 nm (figure 3.6).

La différence qui paraissait importante sur la figure 3.5 est à nuancer si l'on s'intéresse aux spectres individuels. En effet, il ne semble pas se dégager de tendance claire vis-à-vis du statut sur cette partie du spectre.

C'est pourquoi il est difficile d'affirmer avec certitude l'effet de la maladie sur les spectres de réflectance en fonction du statut HLB en utilisant seulement les données brutes.

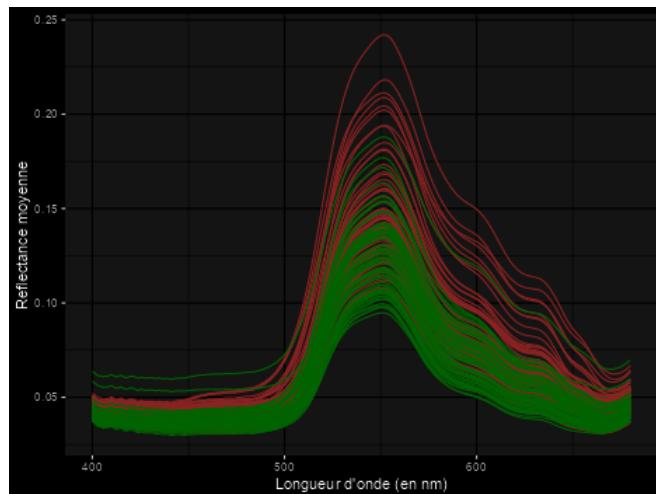


FIG. 3.5 : Spectres individuels d'arbres positifs (en rouge) et négatifs (en vert) au HLB pour les longueurs d'onde de 400 à 680 nm

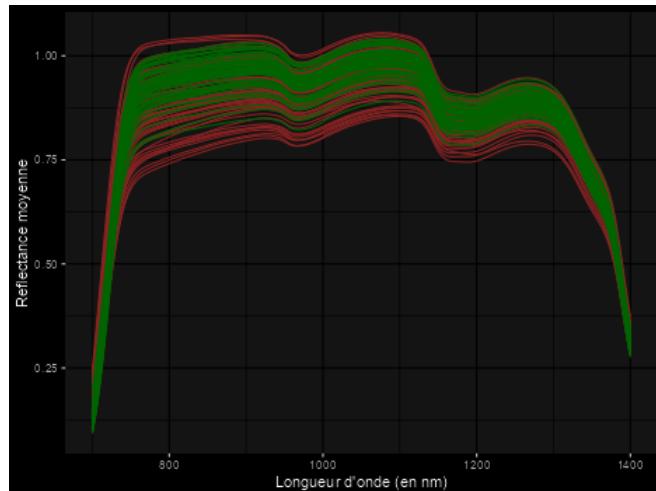


FIG. 3.6 : Spectres individuels d'arbres positifs (en rouge) et négatifs (en vert) au HLB pour les longueurs d'onde de 700 à 1400 nm

Ainsi, il est intéressant de savoir s'il y a une influence de la maladie sur les spectres de réflectance et sur quelles longueurs d'onde se situe cet effet (figure 3.7).

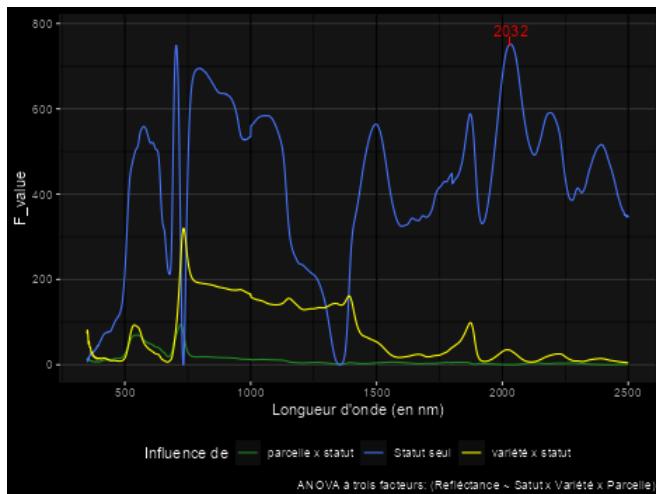


FIG. 3.7 : Valeur du F de Fisher montrant l'influence du statut seul en interaction avec la variété et la parcelle sur le spectre de réflectance

Globalement le statut HLB affecte le spectre de réflectance dans trois zones du spectre : dans le début du “Red edge” autour de 700 nm, dans le proche infrarouge autour de 800 nm et dans le “Short-wave infrared” autour de 2000 nm avec un pic à 2032 nm (valeur  $F = 751$ ). Les effets du statut sont influencés dans le “Red edge” et le proche infrarouge par le type de variété sur laquelle les feuilles ont été prélevées. Cependant le pic de cette influence correspond à une zone où la valeur de Fisher est quasi-nulle pour l'effet du statut sur la détection de la maladie par la réflectance. Cette influence des variétés est donc comprise entre les valeurs de Fisher de 200 à 150 en excluant la zone de creux. Ces valeurs sont assez faibles si on les compare aux effets du statut qui avoisine les 700 et plus. Les trois variétés ont donc un impact plutôt faible sur la détection de la maladie via la mesure de la réflectance. Cet impact est encore plus minime (valeur  $F = 30$ ) au niveau du pic d'influence de la maladie à 2032 nm correspondant à la quantité d'eau dans la feuille. Cet impact est encore plus faible pour l'influence de la parcelle sur le statut HLB. Celle-ci n'a qu'une très faible influence sur le statut HLB au niveau des spectres dans visible et le “Red edge” (valeur  $F = 98$ ) et une influence quasi-nulle dans les autres longueurs d'onde.

### 3.3 L'approche par arbre de décision

Une autre approche pour mettre en évidence l'effet du statut sur la détection de la maladie par la réflectance est la représentation en arbres de décision (figure 3.8).

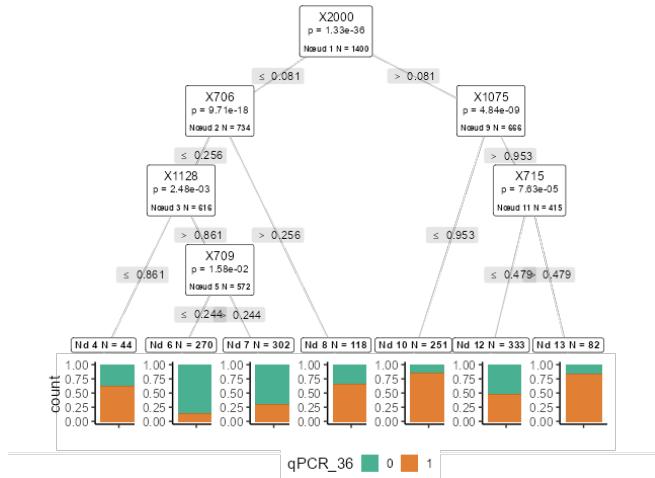


FIG. 3.8 : Arbre de décision sur la réflectance des données globales par rapport au statut HLB

Cette autre approche apporte des résultats cohérents avec le F de Fisher sur les longueurs d'onde où le statut HLB a le plus d'influence. Le premier nœud de décision à la longueur d'onde 2000 nm est très significatif ( $p$  value 1.32e-36) et se divise entre les zones du “Red edge” à 706 nm pour 774 feuilles et du proche infrarouge à 1075 nm pour les 666 feuilles restantes. Sur les 774 feuilles du deuxième nœud ( $p$  value 9.71e-18), 572 sont des feuilles issues d'arbres négatifs aux HLB à plus de 75% environ en additionnant la part des effectifs négatifs des nœuds 6 et 7. Par ailleurs, sur les 666 feuilles du nœud 9 ( $p$  value 4.84e-9), 333 sont globalement issues d'arbres positifs au HLB à plus de 75% environ en additionnant la part des effectifs positifs des nœuds 10 et 13. Cela montre qu'il est possible d'identifier depuis le spectre de réflectance, les arbres sains et malades avec un taux d'erreur d'environ 75%. L'intérêt de cette méthode en plus de donner les longueurs d'onde discriminantes pour la maladie, est la valeur de réflectance pour laquelle chaque longueur d'onde discriminante est clivante.

Enfin, en prenant uniquement la longueur d'onde la plus clivante (2000 nm) selon l'arbre de décision, il est possible de se rendre compte de l'hétérogénéité des distributions à l'intérieur de celle-ci (figure 3.9).

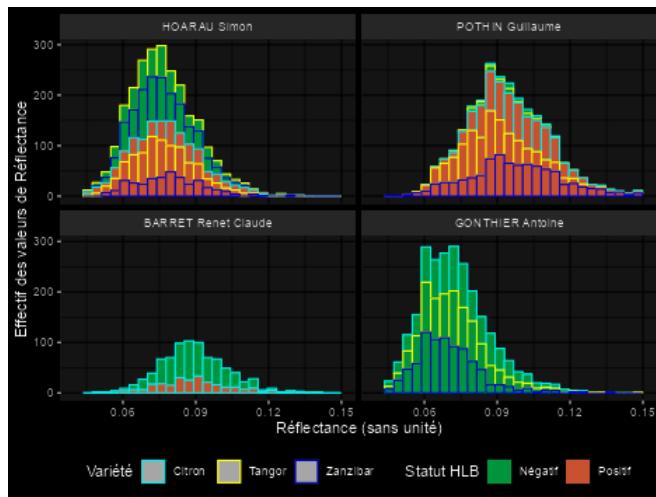


FIG. 3.9 : Distribution des réflectances pour la longueur d'onde 2000 nm

Malgré l'échantillonnage différent selon les parcelles, les valeurs de réflectance sont distribuées de façon quasi-normale bien que centrées différemment. Chez Mrs. Barret et Pothin, le centre est quasiment sur 0.09 alors que chez Gonthier et Hoarau celui-ci est environ à 0.07. Concernant les variétés, les Zanzibars semblent avoir les valeurs de réflectances les moins élevées. Malgré cette hétérogénéité dans la répartition des variables, qui est quasiment identique pour toutes les autres longueurs d'onde, cela n'a finalement pas un énorme impact sur la discrimination de la réflectance en fonction du statut.

### 3.4 Comparaison des performances des trois méthodes de prédiction du statut HLB à partir des spectres réflectance

Pour prédire le statut des arbres à partir des spectres de réflectance, les trois méthodes d'analyses statistiques en apprentissage supervisé sont mises en œuvre. Pour chacune de ces méthodes, les paramètres de performance sont mesurés (table 3.2) après 100 simulations en calculs parallèles présentés en annexes 2 et 3.

TAB. 3.2 : Paramètres de performance de la prédiction du statut HLB par les trois méthodes d'apprentissage supervisé

Paramètres de performance	Moyenne	Ecart type
Accuracy PLS	93.3	1.9
Accuracy RF	77.7	3.0
Accuracy SVM	84.4	2.2
Precision PLS	88.7	3.1
Precision RF	71.6	3.1
Precision SVM	82.6	3.4
Sensitivity PLS	98.9	1.4
Sensitivity RF	94.0	1.8
Sensitivity SVM	85.6	2.4

Au vu de ces simulations, la méthode de régression par les moindres carrés partiels (PLS) est la plus robuste avec une qualité de prédiction avoisinant les 93.3% et avec l'écart type le plus faibles d'environ 1.9. Cette méthode est aussi très robuste pour minimiser les erreurs de type 2 (sensibilité à 98.9%) et de type 1 (précision à 88.7%) avec là aussi des écarts types plus faible que les deux autres méthodes. Vient ensuite la méthode de la machine à vecteurs de support (SVM) avec une qualité de prédiction d'environ 84.4% et avec l'écart type le plus élevé avoisinant les 2.7. La méthode des forêts aléatoires (RF) est la moins robuste avec une qualité de prédiction de 77.7% approximativement.

### 3.5 Prédiction du statut HLB par Régression par les Moindres Carrés Partiels (PLS)

Concrètement, en utilisant la meilleure méthode, qui est celle des moindres carrés partiels, le statut des arbres peut donc être prédit précisément (figure 3.10).

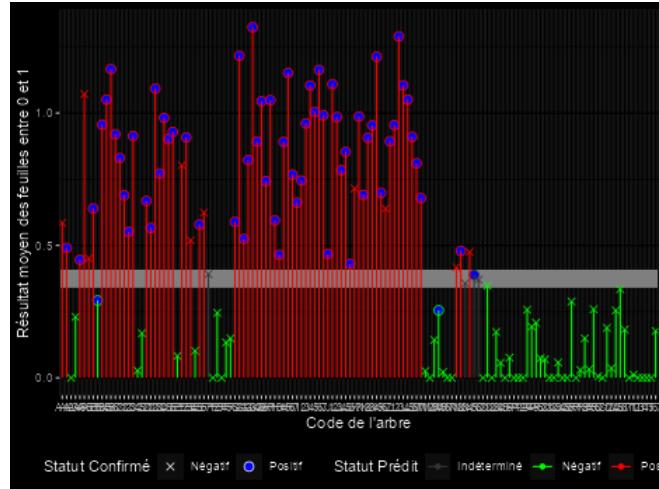


FIG. 3.10 : Prédiction du statut HLB par Régression par les Moindres Carrés Partiels

Cette représentation graphique est une aide à la décision pour déterminer le statut de chaque arbre à partir des résultats de la PLS. La prédiction par PLS donne une valeur correspondant à la moyenne des statuts prédits (compris entre 0 et 1 en ordonnées) des dix feuilles d'un arbre pour chaque arbre (en abscisse). Cela indique pour chaque arbre s'il est plutôt positif ou négatif, mais ça ne permet pas de trancher sur le statut de l'arbre. Cette prédiction se base donc sur deux seuils choisis via les observations de prédiction effectuées. Ces seuils sont compris entre 0.4 pour les arbres positifs qui seraient au-dessus de cette valeur (en rouge), 0.35 pour les arbres négatifs qui seraient en dessous de cette valeur (en vert) et les arbres indéterminés (en gris) qui seraient compris entre ces deux valeurs. Une comparaison entre le statut réel des arbres est confirmé par qPCR (avec un croix pour les négatifs et un rond pour les positifs), permet ensuite de réaliser la matrice de confusion (table 3.3).

TAB. 3.3 : Matrice de confusion de la méthode de Régression par les Moindres Carrés Partiels

	Négatif confirmé	Positif confirmé
Négatif prédit	57	2
Positif prédit	10	62

Sur cette prédiction, 62 arbres ont été prédits correctement positifs et 57 correctement négatifs. 10 arbres sont des faux positifs (erreur de type 1) et 2 sont des faux négatifs (erreur de type 2).

TAB. 3.4 : Paramètres de performance du statut HLB par Régression par les Moindres Carrés Partiels

Accuracy	90.8
Precision	86.1
Sensitivity	96.9

La prédiction par la méthode de régression par PLS a l'avantage d'avoir une sensibilité élevée (96.9%) ce qui minimise les erreurs de type 2, ce qui est très utile pour l'identification de la maladie sur une parcelle (table 3.4).

### 3.6 Amélioration du protocole de terrain pour le choix du nombre de feuilles par arbre

La méthode de la SVM est la méthode dont l'estimation des performances est la plus rapide à exécuter. Elle est donc utilisée pour tester s'il est possible d'alléger le protocole d'échantillonnage (figure 3.11).

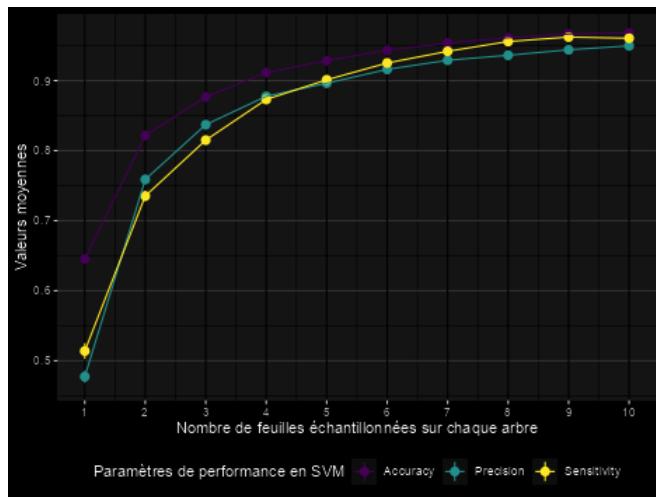


FIG. 3.11 : Prédiction des paramètres de SVM en fonction du nombre de feuilles échantillonnées sur chaque arbre, obtenu après avoir fait la moyenne de 1000 SVM

Après 1000 simulations, les paramètres de performance de la méthode SVM sont respectivement de 95% pour la précision (Precision), 96% pour la sensibilité (Sensitivity) et 97% pour la qualité globale de la prédiction (Accuracy) pour un échantillonnage de 10 feuilles. La précision et la sensibilité passent en dessous des 90% avec 5 feuilles prélevées. Avec 8 feuilles échantillonées, la qualité de la prédiction ainsi que la sensibilité ne baissent pas en dessous de 95%.

### 3.7 Amélioration du protocole de terrain pour le choix du nombre de mesures de réflectance par feuille

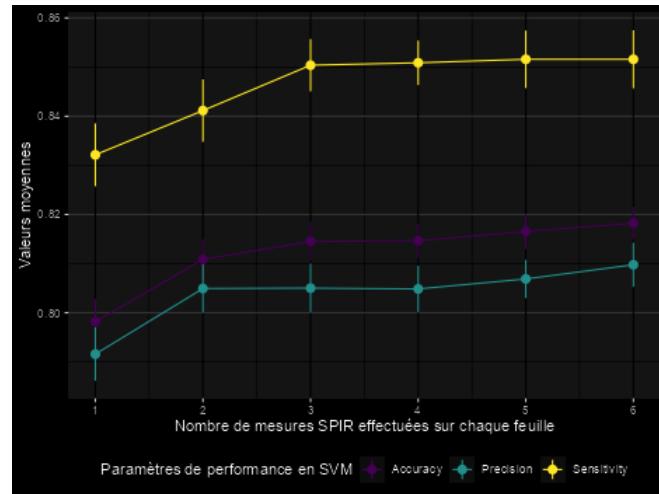


FIG. 3.12 : Prédiction des paramètres de SVM en fonction du nombre de répétition SPIR par feuille échantillonnés sur chaque arbre, obtenu après avoir fait la moyenne de 100 SVM

Après 100 simulations, les paramètres de performance de la méthode SVM sont respectivement de 85% pour la sensibilité (Sensitivity), 82% pour la qualité globale de la prédiction (Accuracy) et 81% pour la précision (Precision) pour un passage de 6 répétitions SPIR (figure 3.12). Les écarts types étant assez marqués, au bout de 3 répétitions SPIR les paramètres de performance sont très peu changeants.

## CHAPITRE 4

# Discussion

L'objectif de cette étude était de caractériser les sources de variations résultant de l'influence des parcelles, des variétés et du statut HLB sur les spectres de réflectance. Au vu des résultats, de par une valeur F de Fisher élevée (plus de 500 en moyenne), le lieu d'échantillonnage semble avoir une influence significative sur les valeurs de réflectance de 750 à 2500 nm (figure 3.2). Dans une moindre mesure, la variété a aussi une influence sur les valeurs de réflectance de 1400 à 2500 nm (figure 3.3). Aucun de ces deux facteurs ne semble affecter les longueurs d'onde dans le visible de 350 à 700 nm et les effets des variétés et du lieu d'échantillonnage sur le statut sont assez négligeables sur cette zone et dans l'infrarouge courte longueur d'onde (de 1400 à 2500 nm). Le statut seul a donc une influence sur cette partie du spectre avec une influence modérée des variétés et du lieu d'échantillonnage (figure 3.7). Ainsi, dans le visible et le début du "Red edge", cette influence peut s'expliquer par une moins bonne absorbance de la lumière par la chlorophylle des feuilles à cause de la maladie. Dans le proche infrarouge autour de 800 nm le statut a aussi une influence importante sur la réflectance avec en parallèle une influence importante du lieu d'échantillonnage sur cette partie du spectre (figure 3.3). Le proche infrarouge correspondant à la structure interne de la feuille a un effet sur le spectre de réflectance global dû à la maladie. En effet, le HLB affecte la structure cellulaire et en particulier celle du phloème, là où la bactérie se développe. Celle-ci va entraîner des nécroses et obstruer les vaisseaux. Ces dégradations vont donc altérer la réflectance de la lumière de cette partie de la feuille ce qui se voit lors de l'analyse des spectres. Pour finir, c'est dans l'infrarouge courte longueur d'onde autour de 2000 nm que s'observe la plus grande valeur F de Fisher (figure 3.7) mais c'est aussi une zone du spectre où le lieu d'échantillonnage (figure 3.3) et les variétés (figure 3.2) dans une moindre mesure ont un effet sur la réflectance globale. L'effet du statut sur cette zone peut s'expliquer par le fait que le HLB affecte grandement l'alimentation en eau de la feuille par l'obstruction des vaisseaux.<sup>1</sup>

<sup>1</sup>Bové (2006). « HUANGLONG-BING : A DESTRUCTIVE, NEWLY-EMERGING, CENTURY · OLD DISEASE OF CITRUS », cf. note 16, p. 3.

En plus de l'approche par la variance via l'ANOVA, l'approche par arbre de décision a permis de mettre en évidence d'une autre manière les longueurs d'onde discriminantes les plus clivantes. Ainsi, au vu de ces deux approches pour détecter plus facilement la maladie, trois tranches de longueurs d'onde sont à favoriser, autour de 700 nm, 1000 nm et 2000 nm (figure 3.8). Ces longueurs d'onde trouvées sont similaires à celles présentes dans la littérature pour la détection du HLB.<sup>2</sup>

<sup>2</sup>Mishra et al. (2007). « Spectral Characteristics of Citrus Greening (Huanglongbing) », cf. note 7, p. 11.

<sup>3</sup>Sankaran et al. (2013). « Huanglongbing (Citrus Greening) Detection Using Visible, Near Infrared and Thermal Imaging Techniques », cf. note 31, p. 5.

<sup>4</sup>Deng et al. (2020). « Detection of Citrus Huanglongbing Based on Multi-Input Neural Network Model of UAV Hyperspectral Remote Sensing », cf. note 3, p. 1.

La réalisation d'un modèle de prédiction du statut HLB par arbre à partir du spectre de réflectance s'est révélée être pertinente à bien des égards. Le modèle de prédiction a permis de connaître l'état sanitaire des arbres via trois méthodes d'analyse des spectres de réflectance. Ces trois méthodes ont des qualités de prédiction qui varient de 92.6% pour la méthode PLS à 85.8% et 76.4% pour la méthode SVM et RF (table 3.2). Ces valeurs sont là aussi proches de celles présentes dans la bibliographie concernant la qualité de la prédiction avec 85% pour le SVM dans l'article de Sindhuja Sankaran<sup>3</sup> et 96.4% pour la PLS dans l'article de Xiaoling Deng.<sup>4</sup> La méthode PLS étant la meilleure, elle permet de minimiser les erreurs de type 2 pour une meilleure identification de la maladie sur les parcelles (table 3.2). Cela peut s'expliquer dans la mesure où l'un des principaux intérêts de l'algorithme PLS est d'augmenter en performance et en fiabilité proportionnellement avec l'augmentation du nombre de variables explicatives permettant de prédire les composantes du modèle. C'est donc la méthode PLS qui est à favoriser pour les prédictions du statut HLB.

Finalement, cette étude confirme le haut potentiel qu'offre l'utilisation de la Spectroscopie Proche Infrarouge via une classification par PLS et SVM avec une approche orientée à la feuille pour la caractérisation de l'infection au HLB. Cette caractérisation est possible grâce aux seuils de décision trouvés. Ces deux seuils sont de 0.4 et plus pour les arbres malades à 0.35 et moins pour les arbres sains avec une classe indéterminée entre ces deux seuils. Ces analyses ont aussi permis d'améliorer le protocole de terrain en préconisant un échantillonnage minimal de huit feuilles par arbres (figure 3.11) et trois répétitions SPIR par feuille (figure 3.12) pour un bon compromis entre le temps d'échantillonnage et la précision du modèle. L'utilisation de la proxidétection dans la surveillance du HLB doit tout de même être couplée à un travail de terrain avec analyse en qPCR pour la validation du statut des arbres.

Le modèle de prédiction réalisée dans ce rapport pourra, à terme, être utilisé comme un outil de détection décisif dans l'amélioration de surveillance épidémiologique du HLB à La Réunion, surtout si le modèle est alimenté avec de nouveaux échantillons dans le but d'agrandir la base d'apprentissage. En effet, ce modèle est pour l'instant valable uniquement pour trois variétés d'agrumes sur les 27 que compte la Réunion. Afin d'optimiser ce modèle, il serait intéressant de l'enrichir avec davantage de variétés sur des lieux d'échantillonnages différents. Cela permettrait de répéter les analyses effectuées en comparant les résultats et les performances des différentes méthodes d'apprentissage supervisé testées pour améliorer le modèle.



## CHAPITRE 5

# Conclusion

Face à la résurgence actuelle du HLB à la Réunion, le potentiel d'un diagnostic HLB par SPIR, plus rapide que les analyses en laboratoire et peu onéreux est prometteur.

La construction de ce modèle de détection permettra de prédire le statut HLB des arbres sur différentes parcelles et d'en évaluer l'efficacité. Ce modèle permettra d'accentuer l'effort de surveillance sur les foyers de production les plus à risque où un échantillonnage qPCR classique n'est pas envisageable à grande échelle au regard des ressources disponibles. Le modèle pouvant déterminer assez précisément quels arbres sont infectés au HLB au sein d'une parcelle, cela permettra d'accompagner au mieux les agriculteurs dans la lutte contre la maladie.

Ainsi, prédire le statut HLB des parcelles d'agrumes par apprentissage supervisé est une solution prometteuse bien que la base de données d'apprentissage doit être étoffée pour garantir une prédiction plus juste à long terme.



## CHAPITRE 6

# Annexe

## 6.1 Annexe 1 : Importation des données brutes

```

> library(tidyr) # pivot longer & pivot_wider
> # Importation des donnees SPIR Global #####
> data_SPIR_Ed <- read.table(file = "SPIR_Global.csv",
+   header = T, sep = ";", stringsAsFactors = T, row.names = 1,
+   na.strings = c("", "NA"), dec = ",")
> # Creation des colonne 'code' de data_SPIR_Ed
> code_lab0 <- rownames(data_SPIR_Ed)
> names(data_SPIR_Ed)[1] = c("code_variete")
> names(data_SPIR_Ed)[2] = c("code_agri")
> data_SPIR_Ed$code_nbr_rep <- factor(substr(code_lab0,
+   8, 8))
> data_SPIR_Ed$code_ech_arbre <- factor(substr(code_lab0,
+   1, 2))
> data_SPIR_Ed$code_ech_feuille <- factor(substr(code_lab0,
+   1, 3))
> data_SPIR_Ed$code_rep_feuille <- factor(paste(data_SPIR_Ed$code_ech_feuille,
+   data_SPIR_Ed$code_nbr_rep, sep = ""))
> rm(code_lab0)
> data_SPIR_Ed <- data_SPIR_Ed[!is.na(data_SPIR_Ed$X350),
+   ]
> # Importation et preparation des resultats de la
> # Qpcr Globaux #####
> data_Qpcr_Ed <- read.table(file = "qPCR_global_Ed.csv",
+   header = T, sep = ";", stringsAsFactors = T, row.names = 1,
+   na.strings = c("", "NA"), dec = ",")
> # on stocke les noms d'echantillons positifs selon
> # nos 2 seuils de Ct (cycle de qPCR ), a savoir
> # moins de 32 cyclces et moins de 36 cycles qPCR
> seuils <- c(32, 36)
> trueP <- lapply(seuils, function(x) unique(data_Qpcr_Ed$Sample.Name[which(data_Qpcr_Ed$C..Mean <
+   x & data_Qpcr_Ed$C..SD < 1)]))
> names(trueP) <- paste("seuil", seuils, sep = ".")
> data_SPIR_Ed[paste("qPCR", seuils, sep = "_")] <- lapply(trueP,
+   function(x) as.numeric(data_SPIR_Ed$code_ech_arbre %in%
+     x))
> # on cherche quels sont les code_ech_arbre qui se
> # trouvent dans le vecteur x x reprenant
> # automatiquement les noms des arbres positifs
> # selon le seuil choisi
> select.lambda <- grep("^X", names(data_SPIR_Ed))
> data_SPIR_Ed <- data_SPIR_Ed[, c(names(data_SPIR_Ed)[-select.lambda],
+   names(data_SPIR_Ed)[select.lambda])]
> rm(data_Qpcr_Ed, select.lambda, seuils)
> data_SPIR_Ed[c("qPCR_32", "qPCR_36")] <- lapply(data_SPIR_Ed[c("qPCR_32",
+   "qPCR_36")], factor)
> # Format_long #####
> select.lambda <- grep("^X", names(data_SPIR_Ed))
> data_long_Ed <- pivot_longer(data = data_SPIR_Ed, cols = select.lambda,
+   values_to = "reflectance", names_to = "lambda")
> data_long_Ed$lambda <- as.numeric(gsub("X", "", data_long_Ed$lambda))
> data_long_Ed <- data_long_Ed[!is.na(data_long_Ed$reflectance),
+   ]

```

## 6.2 Annexe 2 : Fonction Matrice de confusion

```

> fct_ConfusionMatrix <- function(nb.rep, seuil.ct, list.feuilles) {
+   maliste <- lapply(list.feuilles, function(feuille) {
+     list_svm <- feuille[sample(1:length(feuille$code_ech_feuille),
+       nb.rep), ]
+     # on fait ça pour tte les feuilles mais tjrs en
+     # choisissant le nombre de rep tire aleatoirement
+     code <- grep("^code_ech", names(list_svm))
+     qPCR <- grep("^qPCR_", names(list_svm))
+     sortie <- cbind.data.frame(unique(list_svm[c(code,
+       qPCR])), matrix(apply(list_svm[-(which(colnames(list_svm) ==
+         "code_variete")):(which(colnames(list_svm) ==
+         "qPCR_36"))], 2, mean), nr = 1, dimnames = list(NULL,
+         names(list_svm)[-which(colnames(list_svm) ==
+           "code_variete")):(which(colnames(list_svm) ==
+             "qPCR_36")))))
+     sortie
+   })
+   test_ed <- do.call(rbind, maliste) # on colle toute les listes créées précédemment
+   test_ed <- test_ed[-(which(colnames(test_ed) ==
+     "qPCR_32"))]
+   test_ed[[paste0("qPCR_", seuil.ct)]] <- as.numeric(as.character(test_ed[[paste0("qPCR_",
+     seuil.ct)]]))
+   decoup <- sample.split(test_ed[, paste0("qPCR_",
+     seuil.ct)], SplitRatio = 0.5) # on decoupe le jeu de donné en training set et test set
+   training_set <- test_ed[decoup, ]
+   test_set <- test_ed[!decoup, ]
+   # Criteres choisi pour la separation des donnees
+   crit.pos <- 0.4
+   crit.neg <- 0.35
+   trueP$seuil.32 <- NULL
+   # Prediction RF #####
+   model_rf_36 <- randomForest(x = training_set[, 
+     grep("X", names(training_set))], y = training_set[[paste0("qPCR_",
+       seuil.ct)]], ntree = 100)
+   # Prediction sur les feuilles de la base
+   # d'apprentissage
+   rf_pred <- test_set
+   rf_pred$rf_pred_36 <- predict(model_rf_36, newdata = test_set,
+     decision.values = T)
+   # Conversion des resultats de la prediction en
+   # numerique
+   rf_pred$rf_pred_36 = as.numeric(as.character(rf_pred$rf_pred_36))
+   # Moyennage des resultats de la prediction pour
+   # chaque arbres
+   rf_pred_arbres_36 = aggregate(rf_pred_36 ~ code_ech_arbre,
+     data = rf_pred, mean, na.rm = T)
+   rf_pred_arbres_36$crit <- 0.5
+   rf_pred_arbres_36$crit[rf_pred_arbres_36$rf_pred_36 >=
+     crit.pos] <- 1
+   rf_pred_arbres_36$crit[rf_pred_arbres_36$rf_pred_36 <=
+     crit.neg] <- 0
+   # Creation d'une nouvelle colonne des resultats issu
+   # de la qPCR, pour comparer à la valeurs predite
+   rf_pred_arbres_36[paste("qPCR", seuil.ct, sep = "_")] <- lapply(trueP,
+     function(x) as.numeric(rf_pred_arbres_36$code_ech_arbre %in%
+       x))
+   # Résultats de la prédiction sous forme de matrice
+   # de confusion
+   rf_pred_arbres_36$crit <- factor(rf_pred_arbres_36$crit,
+     levels = c(0, 1))
+   rf_confusion_matrix_36 <- ftable(qPCR_36 ~ crit,
+     data = rf_pred_arbres_36)
+   rf_confusion_matrix_36 <- as.matrix(rf_confusion_matrix_36)
+   TP_rf <- rf_confusion_matrix_36[1, 1]
+   TN_rf <- rf_confusion_matrix_36[2, 2]
+   FN_rf <- rf_confusion_matrix_36[2, 1]

```

```

+
FP_rf <- rf_confusion_matrix_36[1, 2]
# Calcul des parametres rf de la matrice de
# confusion
Accuracy_rf36 <- ((TP_rf + TN_rf)/(TP_rf + TN_rf +
FN_rf + FP_rf) * 100)
Precision_rf36 <- ((TP_rf/(TP_rf + FP_rf) * 100))
Sensitivity_rf36 <- ((TP_rf/(TP_rf + FN_rf) * 100))
Parametre_rf_36 <- rbind(Accuracy_rf36, Precision_rf36,
Sensitivity_rf36)
# Prediction SVM #####
model_SVM_36 <- svm(y = training_set[, paste0("qPCR_",
seuil.ct)], x = training_set[, grep("^X", names(training_set))],
type = "C-classification", kernel = "linear")
# Prediction sur le test_set
svm_pred <- test_set
svm_pred$svm_pred_36 <- predict(model_SVM_36, newdata = test_set[, grep("^X", names(test_set))], decision.values = T)
# Conversion des resultats de la prediction en
# numerique
svm_pred$svm_pred_36 = as.numeric(as.character(svm_pred$svm_pred_36))
# Moyennage des resultats de la prediction pour
# chaque arbres
svm_pred_arbres_36 = aggregate(svm_pred_36 ~ code_ech_arbre,
data = svm_pred, mean, na.rm = T)
# Critere choisi sur les observations graphiques
svm_pred_arbres_36$crit <- 0.5
svm_pred_arbres_36$crit[svm_pred_arbres_36$svm_pred_36 >=
crit.pos] <- 1
svm_pred_arbres_36$crit[svm_pred_arbres_36$svm_pred_36 <=
crit.neg] <- 0
# Creation d'une nouvelle colonne des resultats issu
# de la qPCR, pour comparer à la valeurs predite
svm_pred_arbres_36[paste("qPCR", seuil.ct, sep = "_")] <- lapply(trueP,
function(x) as.numeric(svm_pred_arbres_36$code_ech_arbre %in%
x))
# Résultats de la prédition sous forme de matrice
# de confusion
svm_pred_arbres_36$crit <- factor(svm_pred_arbres_36$crit,
levels = c(0, 1))
svm_confusion_matrix_36 <- ftable(qPCR_36 ~ crit,
data = svm_pred_arbres_36)
svm_confusion_matrix_36 <- as.matrix(svm_confusion_matrix_36)
TP_svm <- svm_confusion_matrix_36[1, 1]
TN_svm <- svm_confusion_matrix_36[2, 2]
FN_svm <- svm_confusion_matrix_36[2, 1]
FP_svm <- svm_confusion_matrix_36[1, 2]
# Calcul des parametres sum de la matrice de
# confusion
Accuracy_svm36 <- ((TP_svm + TN_svm)/(TP_svm +
TN_svm + FN_svm + FP_svm) * 100)
Precision_svm36 <- ((TP_svm/(TP_svm + FP_svm) *
100))
Sensitivity_svm36 <- ((TP_svm/(TP_svm + FN_svm) *
100))
Parametre_svm_36 <- rbind(Accuracy_svm36, Precision_svm36,
Sensitivity_svm36)
# Prediction PLS #####
model_pls_36 <- plsr(training_set[[paste0("qPCR_",
seuil.ct)]], ~ ., data = training_set[, grep("^X",
names(training_set))], scale = TRUE, validation = "CV")
# Prediction
pls_pred <- test_set
# Prediction sur les feuilles de la base
# d'apprentissage
pls_pred$pls_pred_36 <- predict(model_pls_36, newdata = test_set[, grep("^X", names(test_set))], decision.values = T,
ncomp = 100)
# Conversion des resultats de la prediction en
# numerique

```

```

+
+   pls_pred$pls_pred_36 = as.numeric(as.character(pls_pred$pls_pred_36))
+   # Moyennage des resultats de la prediction pour
+   # chaque arbres
+   pls_pred_arbres_36 = aggregate(pls_pred_36 ~ code_ech_arbre,
+       data = pls_pred, mean, na.rm = T)
+   # Parametrage pour presentation graphique et la
+   # matrice de confusion
+   pls_pred_arbres_36$crit <- 0.5
+   pls_pred_arbres_36$crit[pls_pred_arbres_36$pls_pred_36 >=
+       crit.pos] <- 1
+   pls_pred_arbres_36$crit[pls_pred_arbres_36$pls_pred_36 <=
+       crit.neg] <- 0
+   # Creation d'une nouvelle colonne des resultats issu
+   # de la qPCR, pour comparer à la valeurs predite
+   pls_pred_arbres_36[paste("qPCR", seuil.ct, sep = "_")] <- lapply(trueP,
+       function(x) as.numeric(pls_pred_arbres_36$code_ech_arbre %in%
+           x))
+   # Résultats de la prédiction sous forme de matrice
+   # de confusion
+   pls_pred_arbres_36$crit <- factor(pls_pred_arbres_36$crit,
+       levels = c(0, 1))
+   pls_confusion_matrix_36 <- ftable(qPCR_36 ~ crit,
+       data = pls_pred_arbres_36)
+   pls_confusion_matrix_36 <- as.matrix(pls_confusion_matrix_36)
+   TP_pls <- pls_confusion_matrix_36[1, 1]
+   TN_pls <- pls_confusion_matrix_36[2, 2]
+   FN_pls <- pls_confusion_matrix_36[2, 1]
+   FP_pls <- pls_confusion_matrix_36[1, 2]
+   # Calcul des parametres pls de la matrice de
+   # confusion
+   Accuracy_pls36 <- ((TP_pls + TN_pls)/(TP_pls +
+       TN_pls + FN_pls + FP_pls) * 100)
+   Precision_pls36 <- ((TP_pls/(TP_pls + FP_pls) *
+       100))
+   Sensitivity_pls36 <- ((TP_pls/(TP_pls + FN_pls) *
+       100))
+   Parametre_pls_36 <- rbind(Accuracy_pls36, Precision_pls36,
+       Sensitivity_pls36)
+   parametre_36 <- rbind(Parametre_rf_36, Parametre_svm_36,
+       Parametre_pls_36)
+   All_parametre_36 <- as.vector(parametre_36)
+   names(All_parametre_36) <- c(rownames(parametre_36))
+   All_parametre_36
+
}

```

### 6.3 Annexe 3 : Matrice de confusion des 3 méthodes de machin learning

```

> # Library #####
> library(pls) # package pls
> library(randomForest) # package RF
> library(e1071) # package SVM
> library(tidyr) # pivot longer & pivot wider
> library(snowfall) # Utilisation du calcul paralell pour optimiser la vitesse de calcul
> # Importation des fonctions utiles
> source(file = "Fct_ConfusionMatrix_ct36.R") # Annex 2
> nb.simu <- 100 # Minimu 1000 simu
> rep.max <- 6 # nombre de repetition SPIR sur les feuilles , maximum 6
> Tirage <- split(data_SPIR_Ed, data_SPIR_Ed$code_ech_feuille,
+ drop = T) # drop = T pour enlever les tiroirs vides !!
> # Calcul parralelle
> sfInit(parallel = T, cpus = 4) # optimisation des processeurs sur les 4 coeurs
> sfLibrary(caTools) # la library des packages utilisés
> sfLibrary(pls)
> sfLibrary(randomForest)
> sfLibrary(e1071)
> sfLibrary(caret)
> sfExport("fct_ConfusionMatrix", "Tirage", "rep.max",
+ "nb.simu", "trueP") # les éléments extérieur à la fonction
> T1 <- Sys.time() # information sur le temps que met l'operation a se realiser
> # res.sum.36 <- sfClusterApplySR(rep(1:rep.max,
> # each = nb.simu), fct_sum , seuil.ct = 36 ,
> # list.feuilles= Tirage , restore = F, perUpdate =
> # 6 ) # restore = T seulement si ça plante !
> res.ML.36 <- sfClusterApplyLB(rep(rep.max, each = nb.simu),
+ fct_ConfusionMatrix, seuil.ct = 36, list.feuilles = Tirage)
> # fct_sum on remplis les 3 arguments qui sont :
> # nb.rep, seuil.ct, list.feuilles
> T2 <- Sys.time()
> sfStop() # stop l'utilisation du sfInit aux autres lignes de codes
> difftime(T2, T1) # information sur le temps qu'à mis l'operation
> intermed.36 <- as.data.frame(do.call(rbind, res.ML.36))
> # permet de basculer de la liste à la data.frame
> # pour le resultat issu de sfClusterApplyLB
> ML_global.36 <- pivot_longer((intermed.36), cols = 1:9,
+ names_to = "critere", values_to = "valeurs")
> # On moyenne tout les parametres pour chaque type
> # de machin learning
> ML.36 <- aggregate(valeurs ~ critere, ML_global.36,
+ mean)
> names(ML.36)[2] <- "Moyenne"
> # Meme chose avec l'écart type
> ML.36$et <- aggregate(valeurs ~ critere, ML_global.36,
+ sd)$valeurs
> ML.36

```

## 6.4 Annexe 4 : Fonction nombre de feuille

```

> fct_feuille <- function(nb.feuille, seuil.ct, list.arbres) {
+   maliste <- lapply(list.arbres, function(arbre) {
+     if (nb.feuille < nrow(arbre))
+       list_svm <- arbre[sample(1:length(arbre$code_ech_arbre),
+                               nb.feuille), ] else list_svm <- arbre
+     # on fait ça pour tte les feuilles mais ttrs en
+     # choisissant le nombre de rep tire aleatoirement
+     code <- grep("code_ech", names(list_svm))
+     qPCR <- grep("qPCR", names(list_svm))
+     sortie <- cbind.data.frame(unique(list_svm[c(code,
+                                               qPCR])), matrix(apply(list_svm[-(which(colnames(list_svm) ==
+                                               "code_ech_feuille")):which(colnames(list_svm) ==
+                                               "qPCR_36")], 2, mean), nr = 1, dimnames = list(NULL,
+                                               names(list_svm)[-which(colnames(list_svm) ==
+                                               "code_ech_feuille")]:which(colnames(list_svm) ==
+                                               "qPCR_36")))))
+     sortie
+   })
+   test_ed <- do.call(rbind, maliste) # on colle toute les listes créées précédemment
+   decoup <- sample.split(test_ed[, paste0("qPCR_",
+                                         seuil.ct)], SplitRatio = 0.75) # on decoupe le jeu de donné en training set et test set
+   training_set <- test_ed[decoup, ]
+   test_set <- test_ed[!decoup, ]
+   res.svm <- svm(y = training_set[, paste0("qPCR_",
+                                             seuil.ct)] # ici on prend les seuil donc sois 32 sois 36 , avec 'paste0' colle sans separateurs
+   ,
+   x = training_set[, grep("^X", names(training_set))] # ici on prend ttes les longueurs d'ondes
+   ,
+   type = "C-classification", kernel = "linear")
+   svm_pred <- predict(res.svm, newdata = test_set[, grep("^X", names(test_set))])
+   svm.confusion <- confusionMatrix(data = svm_pred,
+                                     reference = test_set[, paste0("qPCR_", seuil.ct)])
+   sortie <- c(svm.confusion$overall[1], svm.confusion$byClass[1],
+              svm.confusion$byClass[3]) # combine les 3 parametres recherches
+   names(sortie) <- c("Accuracy", "Sensitivity", "Precision") # renomer
+   sortie
+ }

```

## 6.5 Annexe 5 : Prédiction du nombre optimal de feuille

```

> rm(list = ls()) # nettoyage des listes de l'environnement de travail
> # Library #####
> library(pls) # PLS
> library(ggplot2) # Package ggplot pour graphiques
> library(ggdark) # Met un style de graphique ggplot en noir
> library(ggpubr) # Utilisation de la fonction ggarrange qui permet de coller 2 graphiques
> library(caTools) # sample.split
> library(tidyr) # pivot longer & pivot wider
> library(snowfall) # Utilisation du calcul paralell pour optimiser la vitesse de calcul
> # Importation des fonctions utiles
> source(file = "Scripts/Prediction/Fct_Feuilles.R") # fonction feuille avec SVM
> # Importation du jeu de donnee Global
> load("Sauvegardes_objet_R.data/Jeux de donnee/data_SPIR_Ed.Rdata")
> # I) Parametres SVM Ct<36 #####
> nb.simu <- 1 # Minimu 1000 simu Time difference of 15.39098 hours pour 1000 simu
> nb.feuille <- 10 # nombre de feuilles echantillonees par arbre , maximum 10
> seuil.ct <- 36
> intermed.arbre <- aggregate(reflectance ~ code_ech_feuille +
+   code_ech_arbre + qPCR_32 + qPCR_36 + lambda, data = data_long_Ed,
+   mean)
> mean.arbre <- pivot_wider(intermed.arbre, names_from = "lambda",
+   values_from = "reflectance", names_prefix = "X")
> Tirage <- split(mean.arbre, mean.arbre$code_ech_arbre,
+   drop = T) # drop = T pour enlever les tiroirs vide
> ## I.a) Calcul parallele #####
> sfInit(parallel = T, cpus = 4) # optimisation des processeurs sur les 4 coeurs
> sfLibrary(caTools) # la library des packages utilises
> sfLibrary(e1071)
> sfLibrary(pls)
> sfLibrary(caret)
> sfExport("fct_feuille", "Tirage", "nb.feuille", "nb.simu") # les elements exterieur a la fonction
> T1 <- Sys.time() # information sur le temps que met l'operation a se realiser
> # res.svm.36 <- sfClusterApplySR(rep(1:rep.max,
> # each = nb.simu), fct_sum , seuil.ct = 36 ,
> # list.feuilles= Tirage , restore = F, perUpdate =
> # 6 ) # restore = T seulement si ça plante !
> res.svm.36 <- sfClusterApplyLB(rep(1:nb.feuille, each = nb.simu),
+   fct_feuille, seuil.ct = 36, list.arbres = Tirage)
> # fct_sum on remplsis les 3 arguments qui sont :
> # nb.rep, seuil.ct, list.feuilles
> T2 <- Sys.time()
> sfStop() # stop l'utilisation du sfInit aux autres lignes de codes
> difftime(T2, T1) # information sur le temps qu'a mis l'operation
> ## I.b) Enregistrement des criteres de precision
> ## #####
> # load('Sauvegardes_objet_R.data/SVM_ct36_6rep_100simu_3cpu_sfClusterApplyLB.Rdata')
> intermed.36 <- as.data.frame(do.call(rbind, res.svm.36))
> # permet de basculer de la liste à la data.frame
> # pour le resultat issu de sfClusterApplyLB
> data_global.36 <- pivot_longer((intermed.36), cols = 1:3,
+   names_to = "critere", values_to = "valeurs")
> data_global.36$nb.rep <- rep(1:nb.feuille, each = (nb.simu *
+   3))

```

## 6.6 Annexe 6 : Fonction nombre de répétition SPIR par feuille

```

> fct_svm <- function(nb.rep, seuil.ct, list.feuilles) {
+   maliste <- lapply(list.feuilles, function(feuille) {
+     # On cree une list qui piochera au hazard 1 feuille
+     # avec le nombre de rep correspondant
+     if (nb.rep < nrow(feuille))
+       list_svm <- feuille[sample(1:length(feuille$code_ech_feuille),
+         nb.rep), ] else list_svm <- feuille
+     # on fait ça pour tte les feuilles mais tjrs en
+     # choisissant le nombre de rep tire aleatoirement
+     code <- grep("code_ech", names(list_svm))
+     qPCR <- grep("qPCR", names(list_svm))
+     sortie <- cbind.data.frame(unique(list_svm[c(code,
+       qPCR])), matrix(apply(list_svm[-(which(colnames(list_svm) ==
+         "code_nbr_rep"):which(colnames(list_svm) ==
+         "qPCR_36")]], 2, mean) # on moyenne la valeur des rep pour chaque feuille
+ ,
+       nr = 1, dimnames = list(NULL, names(list_svm)[-which(colnames(list_svm) ==
+         "code_nbr_rep"):which(colnames(list_svm) ==
+         "qPCR_36")))))
+     sortie
+   })
+   test_ed <- do.call(rbind, maliste) # on colle toute les listes créées précédemment
+   decoup <- sample.split(test_ed[, paste0("qPCR_",
+     seuil.ct)], SplitRatio = 0.75) # on decoupe le jeu de donné en training set et test set
+   training_set <- test_ed[decoup, ]
+   test_set <- test_ed[!decoup, ]
+   res.svm <- svm(y = training_set[, paste0("qPCR_",
+     seuil.ct)] # ici on prend les seuil donc sois 32 sois 36 , avec 'paste0' colle sans separateurs
+ ,
+     x = training_set[, grep("^X", names(training_set))] # ici on prend ttes les longueurs d'ondes
+ ,
+     type = "C-classification", kernel = "linear")
+   svm_pred <- predict(res.svm, newdata = test_set[, 
+     grep("^X", names(test_set))])
+   svm.confusion <- confusionMatrix(data = svm_pred,
+     reference = test_set[, paste0("qPCR_", seuil.ct)])
+   sortie <- c(svm.confusion$overall[1], svm.confusion$byClass[1],
+     svm.confusion$byClass[3]) # combine les 3 parametres recherches
+   names(sortie) <- c("Accuracy", "Sensitivity", "Precision")
+   sortie
+ }

```

## 6.7 Annexe 7 : Prédiction du nombre de répétition optimal

```

> # Library #####
> library(e1071) # SVM
> library(pls) # PLS
> library(ggplot2) # Package ggplot pour graphiques
> library(ggdark) # Met un style de graphique ggplot en noir
> library(ggpubr) # Utilisation de la fonction ggarrange qui permet de coller 2 graphiques
> library(caTools) # sample.split
> library(tidyr) # pivot longer & pivot wider
> library(caret) # fonction confusionMatrix
> library(snowfall) # Utilisation du calcul paralell pour optimiser la vitesse de calcul
> # Importation des fonctions utiles
> source(file = "Scripts/Prediction/Fct_SVM.R") # pour calculer avec fonction svm
> # source(file = 'Scripts/Prediction/Fct_PLS.R') #
> # pour calculer avec fonction pls
> # Importation du jeu de donnee Global
> load("Sauvegardes_objet_R.data/Jeux de donnee/data_SPIR_Ed.Rdata")
> # I) Parametres SVM Ct<36 #####
> nb.simu <- 100 # nombre de simulation
> rep.max <- 6 # nombre de repetition SPIR sur les feuilles , maximum 6
> Tirage <- split(data_SPIR_Ed[, -c(1:2)], data_SPIR_Ed$code_ech_feuille,
+ drop = T) # drop = T pour enlever les tiroirs vides !
> ## I.a) Calcul parallele #####
> sfInit(parallel = T, cpus = 4) # optimisation des processeurs sur les 4 coeurs
> sfLibrary(caTools) # la library des packages utilisés
> sfLibrary(e1071) # pour fct_svm
> sfLibrary(pls) # pour fct_pls
> sfLibrary(caret)
> sfExport("fct_svm", "Tirage", "rep.max", "nb.simu") # les elements exterieur a la fonction
> T1 <- Sys.time() # information sur le temps que met l'operation a se realiser
> res.svm.36 <- sfClusterApplyLB(rep(1:rep.max, each = nb.simu),
+ fct_svm, seuil.ct = 36, list.feuilles = Tirage)
> # fct_svm on remplis les 3 arguments qui sont :
> # nb.rep, seuil.ct, list.feuilles
> T2 <- Sys.time()
> sfStop() # stop l'utilisation du sfInit aux autres lignes de codes
> difftime(T2, T1) # information sur le temps qu'a mis l'operation
> ## I.b) Enregistrement des criteres de precision
> ## #####
> # load('Sauvegardes_objet_R.data/SVM_ct36_6rep_100simu_3cpu_sfClusterApplyLB.Rdata')
> intermed.36 <- as.data.frame(do.call(rbind, res.svm.36))
> # permet de basculer de la liste à la data.frame
> # pour le resultat issu de sfClusterApplyLB
> data_global.36 <- pivot_longer((intermed.36), cols = 1:3,
+ names_to = "critere", values_to = "valeurs")
> data_global.36$nb.rep <- rep(1:rep.max, each = (rep.max *
+ nb.simu/2))

```

# Bibliographie

- ALBETIS DE LA CRUZ, J. L. (2018). « Potentiel des images multispectrales acquises par drone dans la détection des zones infectées par la Flavescence dorée de la vigne ». fr. THÈSE DE DOCTORAT EN SCIENCES AGRONOMIQUES Spécialité : Sciences de la Production Végétale. Toulouse : Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier) (cf. p. 11, 12).
- AUBERT, B., M. GRISONI et M. VILLEMIN (1996). « A Case Study of Huanglongbing (Greening) Control in Réunion ». en. In : *Thirteenth IOCVC Conference* 9.1, p. 3 (cf. p. 5).
- AUBERT, B. (1989). « Le greening une maladie infectieuse des agrumes d'origine bactérienne transmise par des homoptères psyllidés Stratégies de lutte développées à l'île de la Réunion, Circonstances épidémiologiques en Afrique Asie et modalités d'intervention ». THÈSE DE DOCTORAT EN SCIENCES AGRONOMIQUES Spécialité : Sciences de la Production Végétale. INRA Bordeaux : Université Bordeaux 2 (cf. p. 3, 4).
- BERTAUX, C. (2015). « Mise en place de la spectroscopie proche infrarouge selon les approches Quality By Design et Lean Six Sigma pour le pilotage en ligne de l'humidité d'une poudre ». fr. THÈSE DE DOCTORAT EN SCIENCES PHARMACEUTIQUE. Bordeaux : Université Bordeaux 2 U.F.R. DES SCIENCES PHARMACEUTIQUES (cf. p. 13).
- BORKOVEC, M. et N. MADIN (2019). *ggparty : 'ggplot' Visualizations for the 'partykit' Package*. R package version 1.0.0 (cf. p. 15).
- BOVÉ, J. (2006). « HUANGLONGBING : A DESTRUCTIVE, NEWLY-EMERGING, CENTURY-OLD DISEASE OF CITRUS ». en. In : *Journal of Plant Pathology* 88.1, p. 7-37 (cf. p. 3, 4, 33).
- COMAR, A. (2013). « Etude des interactions feuille/lumière et de leurs implications pour le phénotypage haut débit au champ ». fr. THÈSE DE DOCTORAT EN SCIENCES AGRONOMIQUES Spécialité : Sciences de la Production Végétale. Avignon : Université d'Avignon (cf. p. 11, 12).
- COMTE, A. (2013). « Apport des marqueurs SSRs nucléaires, des InDel mitochondriaux et de la diversité allélique de gènes candidats pour la tolérance à la salinité ». fr. THÈSE DE DOCTORAT EN SCIENCES AGRONOMIQUES Spécialité : Sciences de la Production Végétale. Tunis : INSTITUT NATIONAL AGRONOMIQUE DE TUNISIE (cf. p. 1).
- CORTES, C. et V. VAPNIK (sept. 1995). « Support-vector networks ». en. In : *Machine Learning* 20.3, p. 273-297. doi : 10.1007/BF00994018. (Visité le 04/09/2020) (cf. p. 17).
- CRÉQUY, N. (août 2020). *Prise en compte de la structure du paysage pour la surveillance épidémiologique du HLB sur l'île de La Réunion et exploration d'une nouvelle méthode de diagnostic par Spectroscopie Proche Infra-Rouge*. Rapport de stage. Rennes : Agrocampus Ouest, p. 34 (cf. p. 13).
- DE BLOMAC, F. (mars 2014). *Hyperspectral*. fr-FR. Section : 3D. URL : <https://decryptageo.fr/hyperspectral/> (visité le 21/06/2021) (cf. p. 11).
- DENG, X., Z. ZHU, J. YANG, Z. ZHENG, Z. HUANG, X. YIN, S. WEI et Y. LAN (août 2020). « Detection of Citrus Huanglongbing Based on Multi-Input Neural Network Model of UAV Hyperspectral Remote Sensing ». en. In : *Remote Sensing* 12.17, p. 2678. doi : 10.3390/rs12172678. (Visité le 09/04/2021) (cf. p. 1, 6, 19, 34).
- EQUIPE ARTISTS - UR AIDA, C. (2021). *METEOR*. URL : <https://smartis.re/METEOR> (visité le 08/04/2021) (cf. p. 8).
- GOTZWALD, T. R. (1989). « Preliminary Analysis of Citrus Greening (Huanglungbin) Epidemics in the People's Republic of China and French Reunion Island ». en. In : *Phytopathology* 79.6, p. 687. doi : 10.1094/Phyto-79-687. (Visité le 05/03/2021) (cf. p. 3-5).
- GOTZWALD, T. R. (2010). « Current Epidemiological Understanding of Citrus Huanglongbing ». en. In : *U.S. Department of Agriculture* 48.1, p. 39-119 (cf. p. 2, 7).
- GUILLOTEAU, C. (juill. 2018). *Utilisation de la connaissance du paysage agricole pour l'accompagnement des réseaux d'épidémio-surveillance : application au greening des agrumes à la Réunion*. Rapport de stage. Paris : AgroParisTech, p. 39 (cf. p. 2, 3, 5, 8).
- GUTIERREZ, A. P. et L. PONTI (déc. 2013). « Prospective Analysis of the Geographic Distribution and Relative Abundance of Asian Citrus Psyllid (Hemiptera : Liviidae) and Citrus Greening Disease in North America and the Mediterranean Basin ». en. In : *Florida Entomologist* 96.4, p. 1375-1391. doi : 10.1653/024.096.0417. (Visité le 06/04/2021) (cf. p. 1).
- KNAUS, J. (2015). *snowfall : Easier cluster computing (based on snow)*. R package version 1.84-6.1 (cf. p. 14).
- LEUNG, J. (2014). *La production fruitière à La Réunion*. fr (cf. p. 2).
- LIAW, A. et M. WIENER (2002). « Classification and Regression by randomForest ». In : *R News* 2.3, p. 18-22 (cf. p. 15).
- MEVIK, B.-H., R. WEHRENS et K. H. LILAND (2020). *pls : Partial Least Squares and Principal Component Regression*. R package version 2.7-3 (cf. p. 18).
- MEYER, D., E. DIMITRIADOU, K. HORNIK, A. WEINGESSEL et F. LEISCH (2020). *e1071 : Misc Functions of the Department of Statistics, Probability Theory Group (Formerly : E1071)*, TU Wien. R package version 1.7-4 (cf. p. 17).
- MISHRA, A., R. EHSANI, WON SUK LEE et GENE ALBRIGO (2007). « Spectral Characteristics of Citrus Greening (Huanglongbing) ». en. In : *2007 Minneapolis, Minnesota, June 17-20, 2007*. T. 1. 073056. Minneapolis : American Society of Agricultural and Biological Engineers, p. 10. doi : 10.13031/2013.24163. (Visité le 18/04/2021) (cf. p. 11, 34).

- MORILLON, C. (2020). *Huanglongbing the citrus disease*. en (cf. p. 1).
- MORIYA, É. A. S., N. N. IMAI, A. M. G. TOMMASELLI, A. BERVEGLIERI, E. HONKAVAARA, M. A. SOARES et M. MARINO (juin 2019). « DETECTING CITRUS HUANGLONGBING IN BRAZILIAN ORCHARDS USING HYPERSPECTRAL AERIAL IMAGES ». en. In : *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-2/W13, p. 1881-1886. DOI : [10.5194/isprs-archives-XLII-2-W13-1881-2019](https://doi.org/10.5194/isprs-archives-XLII-2-W13-1881-2019). (Visité le 19/04/2021) (cf. p. 1).
- MOUTARDE, F. (2017). « Arbres de Décision et Forêts Aléatoires ». fr. In : *MINES ParisTech*, p. 20 (cf. p. 15, 16).
- NAROUEI-KHANDAN, H. A., S. E. HALBERT, S. P. WORNER et A. H. C. van BRUGGEN (mars 2016). « Global climate suitability of citrus huanglongbing and its vector, the Asian citrus psyllid, using two correlative species distribution modeling approaches, with emphasis on the USA ». en. In : *European Journal of Plant Pathology* 144.3, p. 655-670. DOI : [10.1007/s10658-015-0804-7](https://doi.org/10.1007/s10658-015-0804-7). (Visité le 28/03/2021) (cf. p. 2, 3).
- R CORE TEAM (2021). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria (cf. p. 14).
- SANKARAN, S. et R. EHSANI (nov. 2011). « Visible-near infrared spectroscopy based citrus greening detection : Evaluation of spectral feature extraction techniques ». en. In : *Crop Protection* 30.11, p. 1508-1513. DOI : [10.1016/j.cropro.2011.07.005](https://doi.org/10.1016/j.cropro.2011.07.005). (Visité le 04/09/2020) (cf. p. 5, 6).
- SANKARAN, S., J. MAJA, S. BUCHANON et R. EHSANI (fév. 2013). « Huanglongbing (Citrus Greening) Detection Using Visible, Near Infrared and Thermal Imaging Techniques ». en. In : *Sensors* 13.2, p. 2117-2130. DOI : [10.3390/s130202117](https://doi.org/10.3390/s130202117). (Visité le 09/04/2021) (cf. p. 5, 19, 34).
- TENEHAUS, H. (1999). « L'approche PLS ». In : *Revue de statistique appliquée* 47.2, p. 37 (cf. p. 18).
- TUSZYNSKI, J. (2020). *caTools : Tools : Moving Window Statistics, GIF, Base64, ROC AUC, etc.* R package version 1.18.0 (cf. p. 14).
- WANG, N. (mai 2019). « The Citrus Huanglongbing Crisis and Potential Solutions ». en. In : *Molecular Plant* 12.5, p. 607-609. DOI : [10.1016/j.molp.2019.03.008](https://doi.org/10.1016/j.molp.2019.03.008). (Visité le 06/04/2021) (cf. p. 1, 3, 4).



