# Analyse comparative de la sensibilité de méthodes pour l'étude de la prédiction de la fraction verte au travers de la résolution et de la diversité d'échelles spatiales

Mario Serouart

☐ CFR Angers ■ CFR Rennes

Année universitaire : 2019 - 2020

Spécialité :

INGÉNIEUR EN AGROALIMENTAIRE

Spécialisation :

SCIENCE DES DONNÉES

**Mémoire de fin d'études**

■ d'ingénieur de l'École nationale supérieure des sciences agronomiques, agroalimentaires, horticoles et du paysage (AGROCAMPUS OUEST), école interne de l'institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement

☐ de master de l'École nationale supérieure des sciences agronomiques, agroalimentaires, horticoles et du paysage (AGROCAMPUS OUEST), école interne de l'institut national d'enseignement supérieur pour l'agriculture, l'alimentation et l'environnement

☐ d'un autre établissement (étudiant arrivé en M2)

# Analyse comparative de la sensibilité de méthodes pour l'étude de la prédiction de la fraction verte au travers de la résolution et de la diversité d'échelles spatiales

Par : SEROUART Mario

Image composée en partie d'une illustration créée par rawpixel.com

***Soutenu à Rennes le 10 Septembre 2020***

***Devant le jury composé de :***

Président :

Maître de stage : Frédéric Baret

Enseignant référent : Marie-Pierre Etienne

Autres membres du jury :

Mathieu Emily | Enseignant - Chercheur

# Remerciements

C'est d'ores et déjà en préambule que j'annonce que ce stage de fin d'étude fut l'expérience professionnelle la plus agréable et la plus enrichissante, en terme de connaissances, que j'ai pu effectuer. De par le cadre, l'autonomie, la confiance accordée ainsi que la totale liberté quant à l'approche de la résolution de la problématique. Je ne peux que remercier tous les acteurs de cet épanouissement, à savoir :

Merci à Frédéric Baret m'ayant supervisé avec bienveillance, je suis extrêmement reconnaissant de la confiance accordée et des opportunités créées.

Merci à Raul Lopez-Lozano, pour son encadrement précieux et sa bonne humeur.

Merci à Etienne David, soutien inconditionnel aussi bien sur le plan professionnel que personnel.

Merci à Simon Madec, pour ses discussions enrichissantes et son expérience sur les jeux de données utilisés.

Merci à Maëva Labouyrie, pour sa patience pour avoir écouté toutes mes euphories de programmation mais aussi mes découvertes concernant les résultats de cette problématique.

Merci aux enseignants de la spécialisation Science des données d'AGROCAMPUS OUEST de nous avoir inculqué le savoir nécessaire à la compréhension d'outils et au developpement de compétences.

Merci à mes parents, pour m'avoir guidé et aidé de toutes les manières possibles afin d'atteindre la fin de ce cursus.

Enfin, je ne saurais citer tous les acteurs, proches ou lointains, m'ayant aidé à la réalisation de ce projet, cependant je les remercie, c'est un travail de toute une structure.

# Table des matières

## Table des figures



## Liste des tableaux

# 1 Introduction

La densité de végétation est une question vitale et importante dans un contexte de population croissante, en particulier lorsque le problème de l'accès à la ressource est en jeu. Cet indicateur est le reflet direct de l'état de santé d'une culture et de ce fait est donc largement étudié grâce à de nombreuses méthodes de quantifications, notamment ces dernières années par l'acquisition et l'analyse d'images RVB (Jin et al., 2017 ; Liu et al., 2017). En rendant la vision par ordinateur accessible, celle-ci permet aux chercheurs d'utiliser des techniques non invasives pour extraire les caractéristiques des plantes, utiles pour toutes sortes de travaux (sélection, études de rendements, études du sol, ...), et permet de remplacer à plus grande échelle le travail d'estimation laborieux d'antan généralement basé sur des études qualitatives. Les méthodes basées sur l'acquisition d'images via divers capteurs pour estimer la densité grâce aux traits phénotypiques, ont connu des progrès technologiques durant la dernière décennie. Ces progrès ont permis d'obtenir des capteurs de plus en plus précis, et des images RVB à haute résolution spatiale, augmentant la précision des différents indicateurs utilisés pour estimer la densité grâce à une meilleure perception des traits phénotypiques (Walter et al., 2015 ; L. Li et al., 2014 ; Spalding et al., 2013).

Cependant, l'avancée technologique mentionnée ci-dessus et la disponibilité de capteurs d'acquisition à partir d'un large panel conduisent à des protocoles d'acquisitions d'images et des paramètres de capteurs différents, influençant la précision de la densité. Ajoutons à cela des variétés de cultures toutes autant différentes, nécessitant de ce fait des paramètres spécifiques et une totale maîtrise des outils utilisés. Dans la littérature, le facteur limitant de la précision se trouve être la taille du pixel. Ce facteur a de nombreuses fois été démontré comme affectant l'estimation de la densité de végétation (Hengl, 2006 ; Jin et al., 2017). La plupart des méthodes actuelles utilisées estiment la densité en classifiant les pixels de l'image selon diverses classes afin d'en percevoir une fraction végétale, évaluant ainsi la proportion de plantes dans une culture (Y. Li et al., 2020).

Pour en revenir aux pixels, afin d'estimer au mieux la densité, le protocole et le capteur utilisés doivent tout deux permettre la détection des plus fines variations de traits phénotypiques. Pour cela, plus les pixels sont petits, plus les détails et donc implicitement les informations seront précises et permettront une détection accrue. Cela est d'autant plus utile sur les cultures étudiées dans ce document, à savoir, les cultures de riz et de blé présentant de fines feuilles, mais cela est généralisable aux premiers stades de développement de tout type de cultures, où de trop grossiers pixels peuvent surpasser la taille entière de la feuille d'intérêt (Mahlein, 2015). En plus du nombre et de la taille suffisants, afin de détecter les plus minimes variations de traits, ajoutons l'apparition d'un autre concept, que sont les pixels mixtes. Ils interviennent dans une image lorsque l'on modifie la taille des pixels en assemblant les pixels adjacents (Jones et al., 2014). Les pixels mixtes sont des pixels avec un mélange de couleurs et donc un mélange des classes à segmenter, ce qui rend la classification des pixels plus ambiguë par rapport aux pixels dits "purs" qui ne possèdent pas d'interprétations variables de par la facilité à les distinguer et donc de les classer (Hsieh et al., 2002). Ainsi l'obtention de fins pixels semblent être primordial au vu de l'impact sur la précision que cela engendre. Trois paramètres sont principalement utilisés dans les discussions sur la télédétection car ils influencent sur les détails qui peuvent être perçus dans les images (Porebski et al., 2008). Ces paramètres sont cités sous les termes de "taille", "résolution" et "échelle". La résolution peut être exprimée en nombre de pixels (souvent référée comme étant une taille d'image), les images de faibles résolutions ayant moins de pixels que celles de hautes résolutions. Et de ce fait, plus les images sont petites, plus les pixels sont gros (si l'on compare à une même largeur d'image) et moins ils sont nombreux. La taille fait référence à la taille absolue en pixels des éléments représentés sur le sol, en d'autres termes, aux dimensions géométriques de tout objet dans les images. Ces deux paramètres étant lié à un dernier paramètre qui est l'échelle de l'image. L'échelle spatiale sera définie ici comme, le ratio entre la distance entre deux points sur une image et la distance réelle entre ces deux mêmes points sur le terrain, en d'autres termes, la proportion que prend un objet dans l'image. Travailler à grande échelle, permet à la fois de capturer une gamme de petites et grandes structures visuelles, là où les petites

structures à trop petite échelle ne seront plus perceptibles, même avec de nombreux pixels sur des images à grande résolution, pour les raisons énoncées sur la discussion de la taille des feuilles.

Dans cette étude, la fraction verte ($FV$) a été utilisée pour montrer l'impact de la taille des pixels sur celle-ci. La fraction verte est une méthode classique pour évaluer le développement des cultures (Gitelson et al., 2002). Elle relate de la surface foliaire, plus précisément la fraction des pixels verts couvrant la surface au sol. C'est donc un indicateur très exploité dans l'analyse d'images en raison du caractère coloré de la chlorophylle (Marcial et al., 2018). Habituellement, les méthodes d'estimation sur des images sont faites par classification comme déjà énoncé, et donc séparation des pixels de végétation contrastant avec les pixels appartenant au sol. Le tout en utilisant des algorithmes de segmentation basés sur le pixel ou sur l'objet/la région (avec pour ce dernier la prise en compte d'un contexte plus ou moins important). C'est notamment le cas dans cette étude plusieurs algorithmes, dont quelques-uns usuellement utilisés, vont être comparés afin de relater les différents comportements de performances, face à la dégradation de résolution, ainsi que face à des comportements de variations d'échelles, impliquant une variation de la taille absolue des objets.

Par conséquent, l'étude sera principalement articulée en premier lieu, par l'estimation de la taille des éléments des images. Une fois ce paramètre connu, il pourra être utilisé comme référence pour l'étude de la sensibilité des méthodes d'estimation de la fraction verte à la résolution. Il sera également modifié pour la création de sous-ensembles d'éléments de même taille, dans l'étude de l'apprentissage des caractéristiques dépendantes de la taille, définies par Gluckman, 2006 ou, plus généralement, dans l'étude des informations relatives à la taille à des échelles spatiales.

# 2 Matériels et Méthodes

Dans la suite de ce mémoire, les données utilisées seront principalement des images. Un bref rappel de la définition d'une image est donc exposé ci-après. Une image est simplement une matrice de valeurs correspondant à une intensité lumineuse pour chaque pixel (exprimée entre 0 et 255). Deux types d'images seront utilisées. Les images en niveaux de gris, ayant une seule valeur pour chaque pixel, tandis que les images en couleurs en ont trois. En effet chaque pixel étant représenté par les valeurs d'intensité lumineuse du rouge, du vert et du bleu (RGB en anglais ou RVB en français). Ainsi, une image de taille/résolution 512 pixels par 512, a les dimensions [512*512*1] pour une image en niveaux de gris et [512*512*3] pour une image en couleurs. Ces dimensions seront appelées canaux.

## 2.1 Jeux de données

Les images proviennent de deux ensembles de données (datasets) sur le riz et le blé recueillis dans les champs de l'Institut des services agroécologiques durables, de l'Université de Tokyo, Japon (35°44'21.7"N, 139°32'31.9"E).

### 2.1.1 Riz

L'ensemble de données sur le riz contient une séquence chronologique d'images RVB (à travers différents jours, heures et conditions météorologiques) d'une seule variété de riz japonaise, à savoir le Kinmaze. Les images des séries chronologiques ont été acquises quotidiennement de Juin (20 jours après le semage) à Août 2019 (environ une semaine avant floraison), de 8h00 à 16h00. La hauteur d'acquisition des images a été fixée à 1,5 m au-dessus du sol. Le champ de vision des caméras était d'environ 138cm*96cm, avec une distance focale de 24 mm correspondant à une résolution d'image de 5184*3456 pixels, le capteur utilisé était le Canon EOS Kiss X5 avec un objectif EF-S18-55 mm.

Pour cette étude, 34 images ont été sélectionnées manuellement à partir des séries temporelles complètes des images prises au cours des premières phases de croissance et celles intermédiaires. Le tout en tenant compte des variations des conditions météorologiques et d'éclairage, intégrant

ainsi une diversité au sein des jeux de données. Finalement, chaque image a été soigneusement annotée en deux classes (végétation et arrière-plan/sol).

Voici les articles relatifs aux caractéristiques d'acquisitions des images de riz (Desai et al., 2019 ; Guo et al., 2017).

#### 2.1.2 Blé

Le protocole d'acquisition des images du blé est identique à celui du riz, sauf que le capteur a été fixé à 1,8 m au-dessus de la culture. Le cultivar de blé Japonica utilisé était le Kinunonami, de Mars à Mai 2012. Là encore, compte tenu des variations des conditions météorologiques et d'éclairage, 7 images de résolution 5184*3456 pixels ont été choisies entre les stades de croissance précoces et intermédiaires.

Aucun article ne pu être fourni, ces images ayant été réalisées à des fins personnelles.

### 2.2 Traitement des données

Le traitement des données est principalement composé de 4 étapes :

(1) En considérant les deux ensembles de données de la partie précédente, pour chaque image, des patchs sont réalisés, c'est-à-dire des sous-échantillons de taille 512*512 ; (2) Ensuite, à l'aide de variogrammes, la taille moyenne des éléments est déduite de chaque patch ; (3) Ainsi, des sous-ensembles de données contenant des éléments de même taille sont composés. Pour le riz, un sous-ensemble de données avec seulement des images d'éléments de 40 pixels ($px$) et un autre avec des images d'éléments de 120px. Pour le blé, un sous-ensemble de données avec des images d'éléments de 120px est créé ; (4) Enfin, une description de la façon dont la résolution a été artificiellement dégradée est présentée.

Un aperçu de l'ensemble du traitement des données est disponible sous forme de représentation visuelle, en fin de partie, dans la Figure 2.

#### 2.2.1 Creation de sous-datasets

##### 2.2.1.1 Creation de patchs

Limité par la mémoire du GPU, il n'aurait pas été possible d'entraîner le modèle avec les images originales de 5184*3456 pixels pour l'ensemble des jeux de données du riz et du blé.

Une méthode pour surmonter ce problème, et parallèlement augmenter notre ensemble de données, a été de découper les images originales en plusieurs patchs carrés (Pound et al., 2017). La taille d'image acceptable a été fixée à 512*512 selon la configuration de l'ordinateur. Cette taille permet de maintenir un niveau de détail assez élevé. Ainsi, après sélection, 200 images annotées ont été obtenues pour l'ensemble de données sur le riz. Puis, environ 50 images pour l'ensemble de données sur le blé, étant donné que moins d'images étaient disponibles, beaucoup d'entre elles étaient vides, remplies de terre aux premiers stades de développement. Ces carrés seront appelés et considérés comme "images" jusqu'à la fin du document, pour plus de commodité.

##### 2.2.1.2 Estimation de la taille des éléments et création des sous-datasets

Le variogramme est un outil couramment utilisé en géostatistiques. Il permet de rendre compte de la variabilité spatiale au sein d'une image à travers les structures et les variations des éléments contenus (tehrani et al., 2017 ; Gringarten et al., 2001). Ce module a été réalisé grâce à la bibliothèque Scikit-GStat sur Python (Mälicke et al., 2019). Cet outil permettrait de connaître la taille des éléments de l'image en pixels, à défaut de connaître la distance d'échantillonnage au sol (la distance physique au sol entre les centres de deux pixels voisins dans l'image, donc la taille des objets en physique en cm ou m), dépendant de plusieurs paramètres dont l'objectif, la focale et la hauteur du capteur. L'étude se focalisera donc sur des tailles d'éléments comparées à des tailles de pixels en absolu, et non pas exprimées en cm.

Le variogramme, basé sur la distance entre deux points de l'image, mesure la variabilité des valeurs entre ces points. Plus, ils sont proches, plus la variabilité (i.e covariance) est faible (i.e élevée) et donc plus ils sont susceptibles d'appartenir au même objet. En géologie, les points d'observations du sol étudié sont issus de forages, par mesure de coûts, ceux-ci ne peuvent être trop nombreux, c'est pour cela que pour produire une carte complète, les géologues ont souvent recours à de l'interpolation. Cependant, plus il y a de points disponibles, plus la variabilité mesurée est correcte. Dans notre cas, le coût du forage n'existe pas, autant de points d'observations que l'on souhaite peuvent être pris, la seule limite étant le temps de calcul. Une grille de 128*128 est alors apposée sur l'image, soit autant de points d'observations. On obtient ainsi des coordonnées (points d'intersection de la grille) et des valeurs, correspondantes à des valeurs de pixels comprises entre 0 et 255 (l'image passe du RVB à l'échelle de gris, pour avoir une unique variance). Ces données ponctuelles s'apparentent à un échantillonnage d'une variable spatialement continue (l'image étudiée) que l'on peut décrire comme un champ *"aléatoire"*. Ainsi, la qualité des relations statistiques entre les points mesurés ne dépend que du nombre de points sur la grille et donc de la mesure des valeurs associées, au détriment du temps de calcul.

Ensuite pour obtenir le variogramme, il nécessaire d'étudier la variable dépendante (la variance, dépendant de la distance entre les points), par rapport à une variable indépendante (les coordonnées spatiales). La variance est calculée à l'aide de ce que l'on appelle un estimateur. Les formules trouvées dans la bibliographie impliquent un principe général étant de calculer la différence au carré entre les valeurs des points mesurés couplés, le tout pondéré. Dans notre étude, il semblerait que l'estimateur Cressie réussisse mieux à gérer les valeurs élevées au vu des RMSE obtenus comparés aux autres estimateurs (Cressie et al., 1980). L'estimateur de Cressie est calculé à l'aide des équations suivantes pour toutes les paires de points mesurées, séparées par la distance $h$.

$$2\gamma(h) = \frac{\left(\frac{1}{N(h)}\sum_{i=1}^{N(h)}|x|^{0.5}\right)^4}{0.457 + \frac{0.494}{N(h)} + \frac{0.045}{N^2(h)}} \quad \text{avec} \quad x = Z(x_i) - Z(x_{i+h}) \tag{1}$$

Où $h$ est la distance entre une paire de points appelée décalage (lag), $N$ le nombre d'observations de paires de points de distance $h$, et $Z$ la valeur d'un point de coordonnée $xi$ . Il est à noter que chaque variance est calculée pour chaque distance $h$.

Un modèle théorique est par la suite construit, sa mise en œuvre est la suivante. Une fonction non linéaire est définie, et ajustée par la méthode classique des moindres carrés. Au vu des tests effectués et des RMSE obtenus, un modèle exponentiel semble satisfaire un bon ajustement et un chevauchement du point d'inflexion satisfaisant sur les images étudiées. Il est habituel de ne montrer que la moitié des distances h, car au-delà, la variance n'est plus interprétable à l'échelle des données. De plus, 100 points ont été projetés, pour avoir une courbe bien définie et l'estimation la plus précise de la valeur seuil, car elle reflétera la taille des objets dans l'image. Il y a un compromis à faire entre ces paramètres choisis. Plus le nombre de points projetés est important, plus la précision du point d'inflexion est grande, mais plus la variabilité est importante, ce qui rend le modèle théorique plus sensible.

En ce qui concerne l'interprétation, plus la valeur seuil est atteinte rapidement, plus la variance augmente rapidement lorsque les points se séparent, plus les points proches ne sont plus corrélés entre eux, et donc plus les objets dans l'image ont une faible taille (c'est-à-dire qu'ils sont petits). La taille estimée moyenne des éléments (en pixels) correspond à la valeur de la Portée effective dans la Figure 1, c'est à dire la valeur de l'abscisse au point d'inflexion, soit la distance h entre des paires de points pour laquelle ceux-ci ne sont plus corrélés.

Un aspect reste non résolu, la direction. Il est possible de choisir la direction des paires de points, afin de forcer le variogramme à augmenter la distance $h$ uniquement sur l'axe directionnel indiqué et ainsi augmenter la précision si tous les objets de l'image sont dans cette direction. Cependant, le problème appliqué aux plantes est discutable. Alors qu'en géologie, les sols des strates sont propices à une direction, ici plusieurs facteurs peuvent déplacer les feuilles dans n'importe quelle direction (stade de développement, espèce, vent, la prise de vue lors de l'acquisition,

...) de sorte que celles-ci ne se retrouvent jamais dans la même direction. Le choix d'une étude omnidirectionnelle s'est avéré plus efficace.

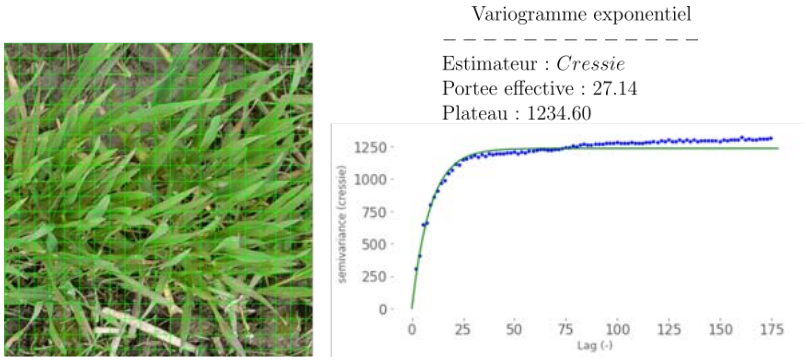


FIGURE 1: *Exemple d'une image et de son variogramme associé, une grille verte a été placée, les carrés sont de côtés égaux à la Portée effective estimée (soit 28 pixels), afin d'avoir une idée relative de la taille moyenne des éléments estimée face aux réels éléments.*

Une limite a été détectée au cours des différentes phases de test. Lorsque cet outil a été utilisé dans des rizières, toutes sortes d'objets présents dans le champ sont réfléchis dans l'eau, formant des ombres, faussant l'estimation de la Portée effective. Il a alors été décidé de poser le masque annoté associé afin de combler le fond et de se concentrer uniquement sur les feuilles.

Par conséquent, l'étude se concentrera sur trois ensembles de données, formés à partir des 250 images citées ci-dessus, parmi lesquelles les plus intéressantes furent choisies (pas de flou, complexité de l'arrière plan, luminosité variante, ...). Chaque sous-ensemble de données comportera une taille d'éléments définie, par taille nous définissons largeur de feuille moyenne. En ce qui concerne ces images, toutes les méthodes proposées n'auront pas de phase de d'entraînement, dans ce cas, seules les images tests seront étudiées. Pour le reste, la procédure de répartition entre les ensembles de données d'entraînement et de test est la suivante.

- Un premier sous-dataset riz , avec des feuilles de 40 pixels, contient 45 images d'entraînement et 16 images de test.
- Un second sous-dataset riz, avec des feuilles de 120 pixels, avec 35 images d'entraînement et 10 images de test.
- Le second dataset riz étant additionné à un dernier, pour l'ensemble de données sur le blé, avec des feuilles également de 120 pixels, il contient 25 images d'entraînement et 10 images de test.

Ainsi, au final l'ensemble des deux datasets additionnés contient 141 images : 105 utilisées par la machine pour l'entraînement (dont des images de validation correspondant à une répartition de 75/25%, pour l'optimisation des hyperparamètres, ceux-ci étant omis dans la rédaction mais bien pris en compte dans l'étude) et 36 pour le test. Les deux datasets à 120 pixels ont été fusionnés.

### 2.2.2 Simulation de la dégradation de la résolution

Pour évaluer les impacts de la résolution sur la fraction verte, les images sont artificiellement dégradées pour obtenir des images grossières avec des pixels de plus en plus grands, afin de simuler les différentes hauteurs d'acquisition des images ou l'utilisation de différents capteurs. Ce protocole, même si certains paramètres restent fixes (comme la distance focale) et ne reflètent donc pas parfaitement une simulation parfaite des différents capteurs, est réputé conférer une bonne simulation.

Le but était ici de créer plusieurs nouveaux ensembles de données à partir de ceux créés ci-dessus, en réduisant les images à une même taille d'éléments, afin d'évaluer l'étude de la résolution en fonction du pixel. Par exemple, un des nouveaux ensembles de données crées pourrait être un ensemble de données où toutes les images ont une taille d'élément de 40 pixels. Un des jeux de

données sur le riz comporte déjà des éléments de cette taille et se voit donc non modifié, mais l'autre jeu de données, à savoir 120 pixels blé **et** riz, verra ses images se réduire, au point où les éléments atteignent une taille de 40 pixels, estimé par le variogramme. Ce principe sera répété pour autant d'ensembles de données que nécessaire.

Ces redimensionnements d'images ont été effectués à l'aide d'un algorithme d'interpolation bicubique. Les interpolations cubiques ont été faites avec la bibliothèque Python OpenCV (Bradski, 2000). Il s'agit d'une méthode courante de dégradation des images, où la valeur du pixel de sortie est une moyenne pondérée sur les 16 pixels les plus proches en fonction de leur distance. Elle consiste à ajuster une série de polynômes cubiques aux valeurs de luminosité. Les images dégradées par interpolation cubique sont connues pour être plus lisses que les autres méthodes, comme celles des pixels les plus proches ou les méthodes bilinéaires (Cracknell, 2020).

La question sur le maintien de la taille d'images obtenues après un sous-échantillonnage (réduction de la résolution) pour amener les images à une même taille d'élément, ou alors un sur-échantillonnage successivement à l'étape précédente afin d'assurer un retour à la taille initiale de 512*512, reste en suspens.

Une analyse quantitative a montré que les deux méthodes étaient fiables et comparables. En effet, même si la taille des pixels revient à celle initiale en pleine résolution dans le second cas, il s'avère que le sur-échantillonnage étant calculé à partir de l'image réduite (composée de plus gros pixels) permet l'obtention d'informations similaires. Par interpolation bicubique, l'information d'un gros pixel va être pondérée pour le diviser en de nombreux petits pixels, au détriment de la qualité , par l'ajout de pixels mixtes.

Cependant certaines des méthodes d'estimation de la fraction verte présentées dans la section suivante sont fortement liées à la texture et réagissent donc mal au sur-échantillonnage en raison des pixels mixtes, alors que d'autres méthodes basées sur la finesse, le nombre de voisins proches et une segmentation, ne peuvent pas être faites sur de minuscules images et dans ces cas, le sur-échantillonnage, malgré les pixels mixtes, permet de meilleures performances. Un examen visant à clarifier la méthode de dégradation la plus appropriée sera disponible dans la section Résultats et Discussion.

Pour se concentrer uniquement sur la taille des pixels, il n'était pas possible d'utiliser plusieurs ensembles de données avec des résolutions différentes, ce qui aurait entraîné un biais qui ne pouvait pas être ignoré en raison des nombreux facteurs environnementaux sous-jacents non contrôlés. De ce fait il a été nécessaire de ne travailler uniquement qu'avec de mêmes images en simulant une dégradation. Pour finir, seul le processus de modification des images a été décrit, cependant les masques associés ont également subi les mêmes modifications. Ainsi la qualité des masques modifiés dépend directement de la qualité de la segmentation manuelle à haute résolution. Par manque de temps, les masques n'ont pas été manuellement annotés pour chaque dégradation, ce qui aurait été tout aussi ambiguë de part la classification manuelle des pixels mixtes. De plus, il a été montré en réalité, que les images étaient bien annotées avec cette méthode. De plus, ces masques ont été utilisés pour toutes les méthodes d'apprentissage, et donc tous les algorithmes ont appris sur ces mêmes masques, ce qui a permis de fixer et d'éviter un potentiel biais.

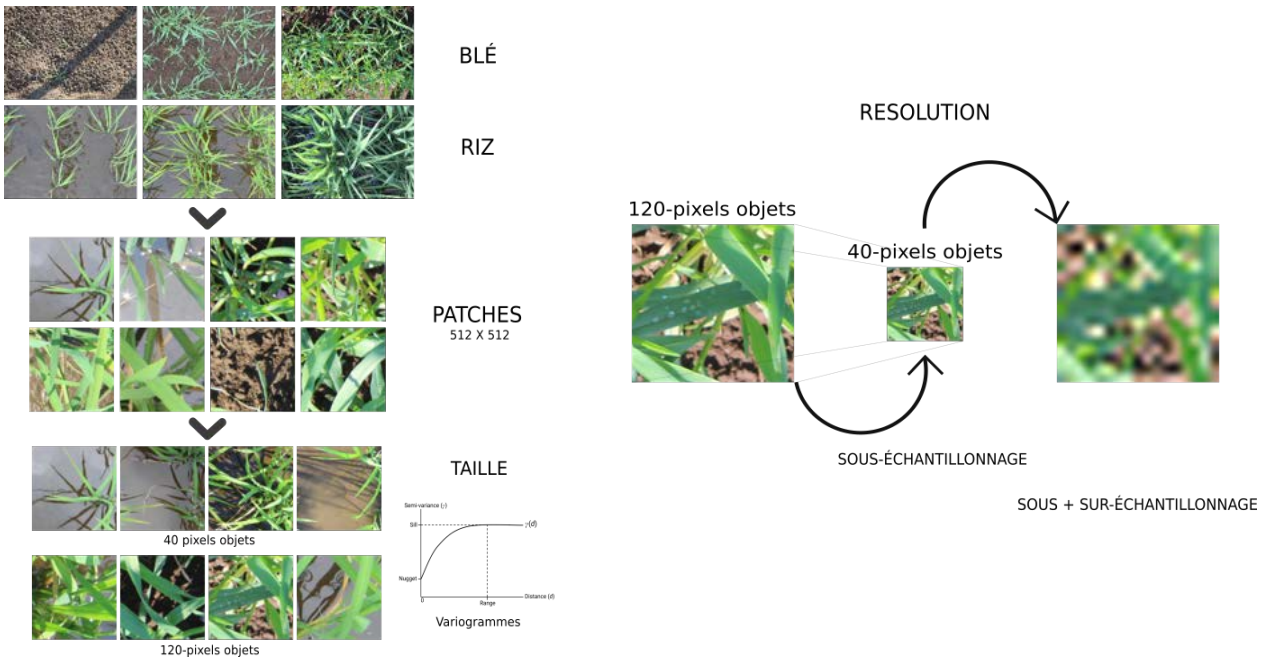


FIGURE 2: *Synthèse du traitement des données effectué.*

## 2.3 Description des algorithmes d'estimation de la Fraction Verte

Comme le mentionne l'article de Y. LI et al., 2020, la couleur est une caractéristique essentielle et largement utilisée qui affecte fortement la performance de segmentation des cultures. Après avoir examiné les algorithmes de segmentation des cultures dans leur benchmark, ils ont observé que presque tous les algorithmes utilisaient un espace de couleur unique, dans la plupart des cas RVB. Mais en raison de la complexité de l'environnement extérieur, les images sont facilement affectées par différentes illuminations, et la couleur ne peut donc pas être la seule caractéristique étudiée pour l'estimation de la fraction verte. C'est pourquoi nous avons décidé d'élargir le panel de description des caractéristiques en utilisant également des caractéristiques de textures. Pour des raisons de synthèse, cette section passe brièvement en revue les algorithmes utilisés pour l'estimation de la fraction verte, et décrit la théorie de chacun d'entre eux. Un aperçu des caractéristiques de chaque technique est disponible ci-dessous (Tableau 1).

| | **Attributs** | | | | |
|---|---|---|---|---|---|
| **Methodes** | Apprentissage | Traitement | Principe | Sortie | Temps de calcul |
| Neural Network (**NN**) | *OUI* | Moyenne spatiale | Regression | FV | *Rapide* |
| Convolutional Neural Network (**CNN**) | *OUI* | Convolution | Extraction textures image + Regression | FV | *Rapide (GPU)* |
| ExcessGreen (**ExG**) | *NON* | Indice de couleur | Classification | FV + Masque segmentation | *Rapide* |
| Random Forest (**RF**) | *OUI* | Haralick features | Classification | FV + Masque segmentation | *Lent* |
| U-net | *OUI* | Dé/Convolution | Extraction textures image + Classification | FV + Masque segmentation | *Rapide (GPU)* |

TABLE 1: *Caractéristiques des méthodes utilisées dans la prédiction de la Fraction verte.*

### 2.3.1 Regression Neural network

Keras a été utilisé avec le backend de Tensorflow (CHOLLET et al., 2015 ; ABADI et al., 2016).

Le but principal de cette méthode est de prédire une variable quantitative continue, en d'autres termes, d'obtenir une estimation directe de la fraction verte. Pour ce faire, à partir des canaux numériques moyens rouge, vert et bleu des images, un dataset de 3 variables prédictives est crée. La variable à prédire étant la fraction verte, celle-ci sera déduite des masques annotés, comme étant le ratio de pixels blancs sur le nombre total de pixels. Ce principe de mesure de valeur de fraction verte de référence sera un principe de base pour l'ensemble des méthodes exposées ci-après. À cette fin, l'étude de cette méthode se concentre sur l'utilisation de réseaux de neurones. Un modèle séquentiel a été utilisé selon le principe schématique général suivant

la Figure 3. Dans cette même figure les biais ne sont pas représentés pour un enjeu de clarté graphique.

Rappelons brièvement le fonctionnement, servant de socle commun pour de futures méthodes basées sur l'apprentissage profond. Chaque nœud présent dans le réseau possède une valeur, et chaque noeud est connecté à d'autres selon un coefficient, nommé poids, rendant l'ensemble du modèle ajustable. Selon le type de connectivité, un nœud, si il n'est pas dans la couche d'entrée, servant à initier le modèle aux données réelles, est la somme de la valeur des noeuds de la couche précédente et de leurs poids associés, à l'instar d'une équation de régression multiple classique. Via un système d'allers-retours (backpropagation) le réseau ajuste les poids liés aux variables/noeuds afin de minimiser la fonction de perte (fonction rendant compte de l'erreur face aux valeurs prédites et de références) et de s'assurer en apprenant, que l'on converge vers la vérité. En effet en modifiant ces coefficients, tout le réseau se voit changé (comme les valeurs des noeuds des couches dépendent de celles antérieures), on ajuste alors chaque poids pour que la régression soit la plus juste possible. Enfin une fonction d'activation pour chaque noeud, linéaire ou non, permet de le faire intervenir en le stimulant ou non, suivant son efficacité dans la prédiction.

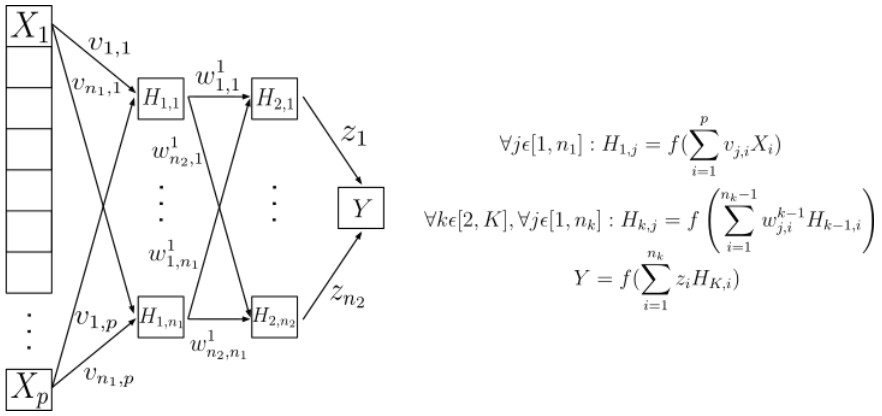


FIGURE 3: *Réseau de neurones schématisé d'après Machine Learning | Neural Network [Diapositives] (2019, Septembre) (*EMILY*, 2019).*

Où $f$ est une fonction d'activation, $v_{j,i}$ , $w^{k-1}_{j,i}$ et $z_i$ sont les poids des connections entre deux noeuds, dans deux couches consécutives, pour l'exemple.

NN est une méthode relativement simple et directe, facile à mettre en œuvre, qui ne nécessite pas de ressources informatiques importantes et qui est donc rapide à entraîner et à appliquer. L'inconvénient est qu'il ne s'agit pas d'une méthode de segmentation, et qu'elle ne fournit que la fraction verte comme résultat se basant sur une moyenne, qui par analogie avec des outils de statistiques communs, renvoie une méthode peu élaborée.

### 2.3.2 Regression Convolutional neural network

Par rapport à NN, le CNN utilise des couches de convolutions pour extraire de l'image RVB des traits caractéristiques de textures (features en anglais) qui devraient permettre a *priori*, une meilleure caractérisation de la fraction verte. En effet, la différence majeure avec le NN est que ce dernier ne présente aucune indication de localités relativistes, de part son architecture il ne prend pas en compte la complexité spatiale liant les éléments entre eux dans une image. Ici, aplanir l'image pleine résolution en un vecteur 1D (à alimenter dans un réseau de neurones) n'est pas suffisant. Les dépendances spatiales doivent être prises en compte, pour cela il est nécessaire de passer par une phase de traitement des données impliquant le concept de couches de convolution, expliqué ci-après, permettant d'extraire des features de l'image d'entrée (ou de la feature map de sortie des couches précédentes). Pour des raisons de synthétisation, il est théoriquement possible de reprendre les mêmes codes schématiques que le NN, puisque chaque noeud dans la couche d'entrée ne serait relié aux noeuds de la couche suivante, uniquement si il existe une dépendance

spatiale, résultant en des poids partagés, cependant dans la littérature il est choisi de présenter le CNN différemment. Mais il est à noter que le principe reste le même. Enfin ces traitements s'effectuent jusqu'au point où l'image multidimensionnelle est suffisamment transformée pour être convertie/aplatie vers un vecteur 1D. En effet, au travers des opérations de convolutions, le concept d'image se perd au profit du gain d'informations relativistes complexes, de ce fait il est usuel d'aplanir pour traiter ces nouvelles données. La dernière étape étant l'ajout d'une couche entièrement connectée, comme un NN classique, permettant la prédiction de la fraction verte en tant que variable continue. Une vue schématique du CNN mis en œuvre est présentée dans la Figure 4.

Pour davantage de détails, la couche de convolution est la couche qui permet d'extraire les caractéristiques de la carte des caractéristiques (feature map en anglais, étant soit l'entrée ou la feature map au rang $n+1$, ...). La couche de convolution est constituée d'un filtre défini, une petite matrice de poids pouvant être ajustable, glissant sur la feature map. Cette matrice prend en compte le pixel au centre comme zone d'intérêt, tout en attribuant aussi des poids aux pixels voisins, dans le but déjà énoncé, étant la notion de gain d'informations spatiales. À chaque patch, les produits scalaires entre les poids du filtre et la portion de l'image étudiée sont faits pour produire une nouvelle valeur de pixel, puis ainsi de suite pour chaque pixel, afin de former au final une nouvelle et complète feature map de sortie. En général, une couche de convolution transforme une matrice d'entrée en une pile de feature map. La profondeur de la pile dépendant du nombre de filtres définis pour une couche, où chaque filtre a des poids différents, spécialisé dans la détection de formes particulières afin de capter diverses informations. Après chaque couche de convolution, il est de coutume d'appliquer une couche d'activation pour introduire de la non-linéarité ainsi que pour accélérer le processus l'apprentissage, ReLU est presque toujours utilisé, cela applique la fonction $f(x) = max(0, x)$ sur les pixels et change toutes les valeurs négatives de l'élément convolué/feature map à 0.

Une couche de regroupement (appelée Pooling) est ensuite utilisée pour réduire la dimensionnalité/résolution de la feature map tout en conservant les informations utiles à apprendre pour chaque carte de la pile. Le plus courant est le Max pooling qui sélectionne la valeur maximale dans un patch sélectionné, par un principe de filtre similaire.

Cette séquence de couche de convolution et de max pooling est répétée autant de fois que nécessaire. Finalement la carte de sortie (de faible résolution, mais de profondeur de pile importante) est convertie/aplatie passant alors d'une donnée multidimensionnelle en un vecteur 1D, une couche entièrement connectée est ajoutée, le principe de la somme des poids a déjà été expliqué, tout le reste suit un NN classique. Dans les CNN un système de backpropagation est aussi appliqué, cependant le principe diffère légèrement des réseaux de neurones classiques, étant donné que les filtres et leurs valeurs associées pour chaque pixel sont les poids à ajuster, puisqu'à l'initialisation le modèle ne connaît pas les motifs à trouver, il doit alors s'adapter en modifiant/créant de nouveaux filtres avec des poids différents.

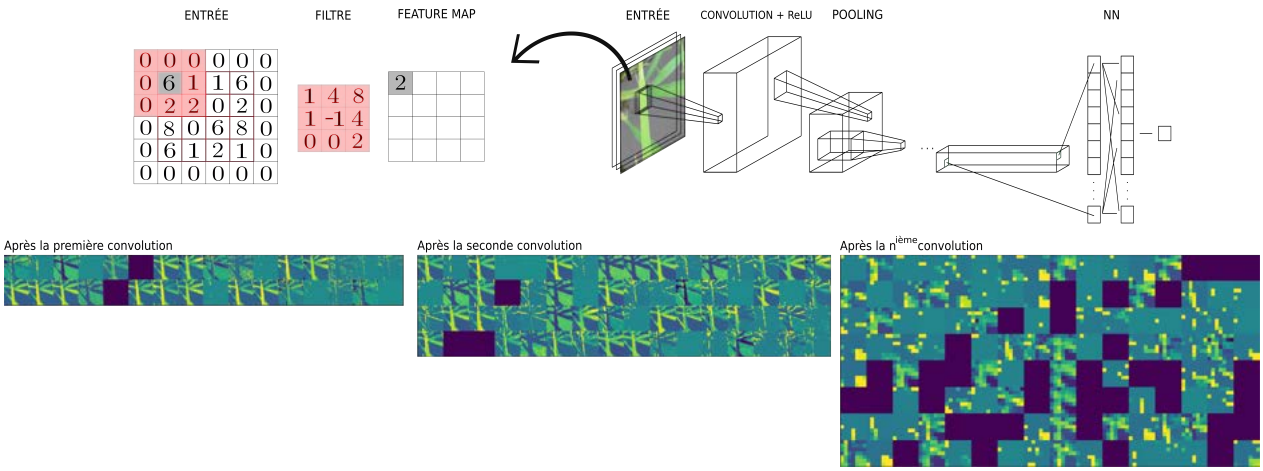


FIGURE 4: *Réseau neuronal convolutif.*

La méthode CNN (comme NN) ne calcule que la fraction verte, et à la différence des méthodes qui vont suivre, ne fournit pas de masque de segmentation.

### 2.3.3 ExcessGreen

Selon les travaux de Woebbecke, après avoir initialement testé différents indices de couleur de végétation, un en particulier a été sélectionné et défini comme le plus souvent efficace, appelé alors Excess Green (Woebbecke et al., 1995).

Pour chaque pixel de l'image, la formule $2g-r-b$ est appliquée rendant ainsi l'image davantage axée sur la végétation verte, selon

$$r = \frac{R'}{R'+G'+B'} \qquad g = \frac{G'}{R'+G'+B'} \qquad b = \frac{B'}{R'+G'+B'} \tag{2}$$

Où $R'$, $G'$ et $B'$ sont normalisés (valeurs entre 0 et 1, en divisant les pixels par 255).

Lorsque l'on veut limiter la subjectivité du sujet quant à la valeur seuil pour la conversion d'une image après indice vers une image entièrement binaire pour obtenir un masque de segmentation, la méthode d'Otsu est utilisée. Cette dernière établit un seuil optimal $S$ en explorant la totalité de l'histogramme de l'image modifiée en niveaux de gris par le biais d'un algorithme itératif (Otsu, 1985). Ainsi, les valeurs de pixels inférieures à ce seuil optimal sont considérées comme de l'arrière-plan et donc modifiées en valeur 0, inversement pour la végétation, tout cela étant visible dans la Figure 5. Ce seuil est défini comme optimal car il minimise la variance intra-classe (i.e. maximise la variance inter-classe) entre la végétation et le fond, selon $\sigma^2_{intra} = p_1(S) \times \sigma^2_1(S) + p_2(S) \times \sigma^2_2(S)$ .

Où $p_1(S)$, par exemple, est la probabilité de se situer dans la classe 1 pour un seuil $S$ étudié, calculé comme une somme de probabilités, pour chacun des niveaux de gris se situant dans la classe 1, i.e. avant le seuil $S$. Ces dernières probabilités étant elle-même calculées, pour chacun des niveaux de gris, comme le nombre de pixels prenant cette valeur de niveau, divisée par l'entièreté des pixels de l'image comme ceci, $P(s) = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(image(i,j)=s)}{\sum_{i=1}^{I}\sum_{j=1}^{J}(image(i,j))}$.

Même principe pour $\sigma^2_1(S)$ étant la variance de la classe 1 pour un seuil $S$ étudié. Calculée classiquement comme la somme, pour chaque niveau de gris de la classe 1, de la différence du niveau de gris étudié et de la moyenne de la classe 1, au carré, multiplié par la probabilité du niveau de gris étudié, le tout pondéré par $p_1(S)$, selon $\sigma^2_1(S) = \frac{\sum_{s=1}^{S}(s-\mu_1)^2 \times P(s)}{p_1}$ , avec $\mu_1 = \frac{\sum_{s=1}^{S} s \times P(s)}{p_1}$.

*(Pour plus de clarté et comme chaque facteur dépend du seuil étudié, les $(S)$ ont été omis).*

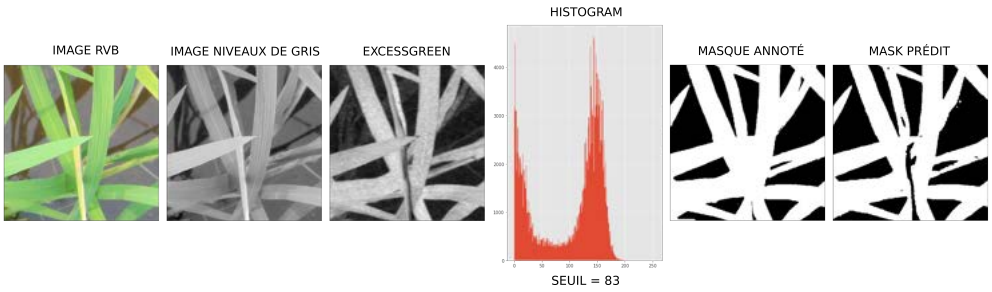


Figure 5: *Exemple du processus d'obtention d'une image transformée par indexation, puis application de la méthode d'Otsu pour binariser.*

Cette méthode de segmentation présente *a priori* plusieurs avantages. Il s'agit d'une méthode facile à calculer et à mettre en œuvre, qui nécessite peu de ressources de calcul, ne nécessite pas de formation spécifique, mais n'en est pour autant pas moins complète, elle fournit la fraction verte et un masque de segmentation. A proprement parlé, il ne s'agit pas d'une méthode d'apprentissage machine (machine learning), car elle ne nécessite aucun entraînement, étant donné que les valeurs seuils sont dérivées d'histogrammes appliquant un algorithme non paramétrique.

### 2.3.4 RandomForest

Une approche se basant sur la classification des pixels grâce à une forêts aléatoire utilisant les caractéristiques de Haralick extraites des matrices de co-occurrences calculées à partir d'images LBP (Local Binary Pattern), a également été évaluée pour estimer la fraction verte. Ces méthodes sont explicitées chacune d'entre elles schématiquement, ci-après, dans la Figure 6.

Contrairement à la classification traditionnelle des forêts aléatoires des images RVB, Porebski et al., 2008 proposent une approche basée sur la texture. Cette méthode suppose que les caractéristiques de texture peuvent être utiles pour séparer les feuilles du fond, car ces dernières ont une texture particulière déterminée par des caractéristiques directionnelles comme les nervures, les ondulations, etc.

Tout d'abord, les images d'entraînement sont transformées en images LBP, en suivant la méthode originale qui a été proposée pour la première fois par Ojala et al., 1996, afin d'exposer les textures présentes dans les images, permettant de souligner les variations locales. Le principe général est de comparer le niveau de luminance d'un pixel avec les niveaux de ses voisins plus ou moins proches selon un rayon modifiable. La différence entre le pixel central d'intérêt et les voisins est appliquée. Un motif binaire est apposé en fonction du signe de la région d'intérêt après la différence. Puis par puissance de 2 suivant un sens circulaire, le plus souvent dans le sens des aiguilles d'une montre, et global (à appliquer de la même manière pour tous les prochains pixels), le pixel central est résumé par l'addition de ces puissances de 2. Des informations relatives à des motifs réguliers dans l'image, c'est-à-dire des textures, sont alors rapportées et mises en valeur.

Il est donc intéressant d'extraire de ces images LBP texturées, des éléments qui mesurent la distribution des niveaux de gris et considèrent les interactions spatiales entre les pixels. Pour cela, l'utilisation des caractéristiques de Haralick, calculées à partir des matrices de co-occurrences des images LBP, est recommandée dans l'article.

Rappelons brièvement les propriétés de cette matrice. Celle-ci est carrée et de dimension $N$, le nombre de niveaux de gris dans la région d'intérêt. L'élément $[i, j]$ de la matrice est généré en comptant le nombre de fois qu'un pixel de valeur $i$ est adjacent à un pixel de valeur $j$. Comme les matrices mesurent l'interaction locale entre les pixels, *"pour réduire la sensibilité de la résolution, il est nécessaire de normaliser ces matrices en les divisant par le nombre total de cooccurrences"*. Chaque cellule est donc considérée comme la probabilité qu'un pixel de valeur $i$ soit adjacent à un pixel de valeur $j$.

En d'autres termes, les images LBP sont découpées en de nombreux morceaux, afin de se concentrer sur l'environnement proche du pixel central considéré pour un morceau donné. Ensuite, pour chacun de ces morceaux, les matrices de co-occurrences sont calculées, et les (14) caractéristiques de Haralick sont extraites (Haralick et al., 1973).

Ainsi, un grand nombre d'informations/caractéristiques pour chaque pixel à travers leurs environnements locaux sont disponibles, tout cela en connaissant leurs classes associées (végétation ou fond). C'est-à-dire autant d'informations disponibles comme entrées en tant que base d'entraînement dans une forêt aléatoire afin de par la suite, classer/prédire si tel pixel dans l'image test est défini comme une feuille ou non, de part ses caractéristiques d'Haralick, en prenant la prédiction majoritaire de tous les arbres obtenus en entraînement. Ces arbres étant classiquement effectués selon, un bootstrapping de l'ensemble du dataset de base pour chaque itération, puis une sélection aléatoire des variables/individus dans le but de les séparer en sous groupes pour chaque niveau d'un arbre de décision, puis une classification en fonction du critère de Gini visant a rendre ces sous groupes les plus purs possibles, i.e. où chaque sous groupe possède une unique classe. Dans cette étude, 500 morceaux ont été extraits pour chaque image d'entraînement, 300 arbres ont été réalisés avec une profondeur d'élagage de 20 afin de ne pas sur-ajuster, i.e. limiter la variance afin d'éviter une flexibilité du modèle trop importante. Pour améliorer le temps de calcul, la bibliothèque XGBoost a été utilisé lors de l'entraînement des forets aléatoires, cette bibliothèque permet l'utilisation du GPU, rendant la tâche plus rapide. Une parallélisation sur 10 cœurs avec la bibliothèque multiprocessing a aussi été implémenté pour l'extraction des textures pour des raisons similaires. (T. Chen et al., 2016 ; McKerns et al., 2012).

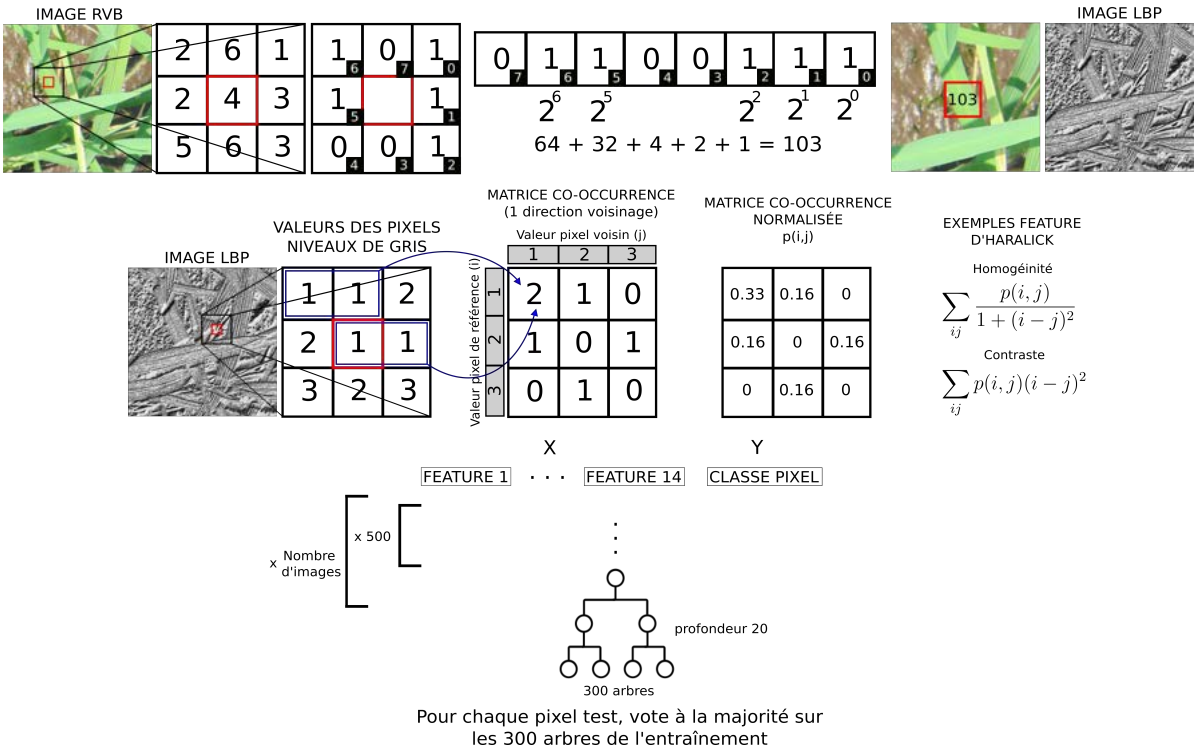


FIGURE 6: *Schématisation de calcul de la représentation LBP (1ère ligne), puis extraction des features d'Haralick (2ème ligne), enfin visualisation synthétique du jeu de données obtenu (3ème ligne).*

### 2.3.5 U-net

L'architecture du modèle U-net est largement utilisée pour la segmentation de classes binaires (RONNEBERGER et al., 2015). La théorie des réseaux de convolution ayant déjà été expliquée auparavant, pour plus de clarté celle-ci ne sera pas ré-exposée, bien que l'architecture U-net suive ce principe général à la différence que deux couches de convolutions se suivent au lieu d'une, c.f. Figure 7. Semblable, du moins pour la première partie, c'est-à-dire celle décrite par le document original, comme *"la voie de réduction permettant de saisir le contexte des images"* se comportant comme un CNN classique, où l'on indique si un motif a été trouvé. La deuxième partie a quant à elle possède une architecture particulière étant un *"chemin d'expansion symétrique"* qui permet une localisation précise des caractéristiques après détection de part la première branche, pour la segmentation fine de l'image. En effet, contrairement à un CNN où la profondeur de la pile de feature map est croissante, tout en diminuant la résolution pour uniquement capturer les caractéristiques principales. Ici pour la segmentation d'image, le but est de permettre au modèle de produire une prédiction sémantique en pleine résolution, revenant ainsi à la taille originale de l'image. Pour ce faire, la façon de reconstruire l'image est d'utiliser une structure d'encodage/décodage où le goulot d'étranglement serait la base de U-net. Par cette analogie, la technique consiste à sur-échantillonner la résolution d'une feature map. Là où le principe exposé précédemment diminue la résolution en résumant, la théorie inverse permettrait de sur-échantillonner en ajoutant des valeurs autour du pixel étudié grâce à des convolutions transposées.

Pour une convolution transposée, une valeur unique est prise dans la carte des caractéristiques à basse résolution et est multipliée à tous les poids d'un filtre, en projetant ces valeurs pondérées dans la feature map de sortie, puis en sommant les éléments se chevauchant. (DUMOULIN et al., 2016 ; LONG et al., 2015).

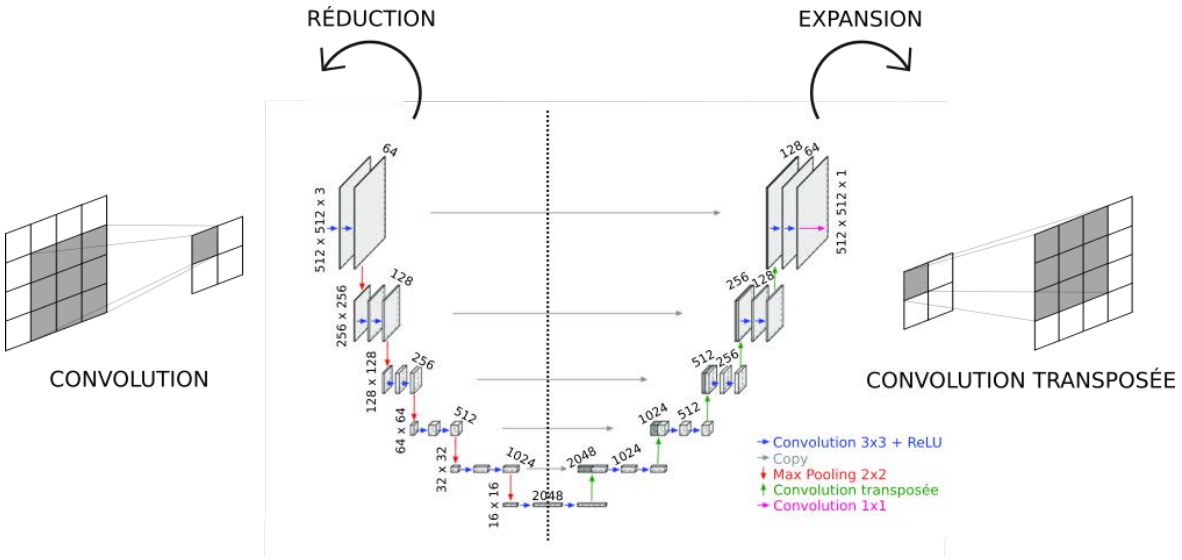


FIGURE 7: *Architecture et principe du modèle U-net.* ***Reproduit à partir de*** *« U-Net : Convolutional Networks for Biomedical Image Segmentation », par Ronneberger et al., 2015* ***et de*** *« Using the U-net convolutional network to map forest types and disturbance in the Atlantic rainforest with very high resolution images », par Wagner et al., 2019, Remote Sensing in Ecology and Conservation, Volume 5. (RONNEBERGER et al., 2015 ; WAGNER et al., 2019)*

Les valeurs dans la matrice d'entrée de la convolution transposée ne doivent pas nécessairement provenir de la matrice d'entrée de convolution initiale. Le principe primordial étant que les poids des filtres dans la partie expansion soient les mêmes, en version *disposition transposée*, que ceux de la voie de réduction. Au regard de la Figure 8, les valeurs obtenues après transposition ne sont pas les mêmes que celles de départ, en effet les convolutions sont des transformations linéaires. Plus précisément une transformation linéaire, dans notre cas, de $R^{16}$ à $R^4$, qui n'est pas inversible. A la différence des approches naïves, telles que les interpolations bicubiques, cette méthode à l'avantage de permettre un ré-échantillonage selon des filtres des convolutions résultants de la première branche d'U-net, et donc selon un apprentissage.

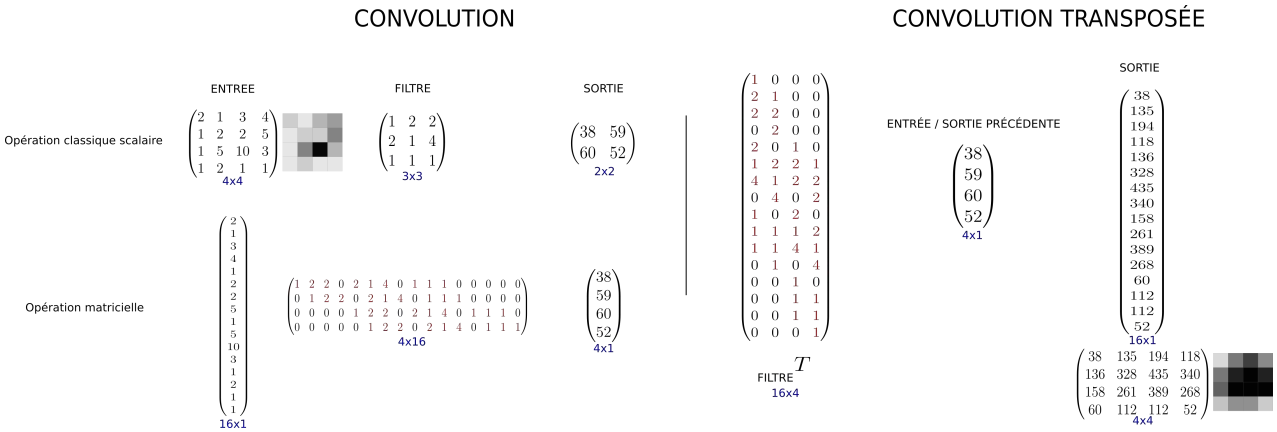


FIGURE 8: *Principe de la convolution transposée.*

L'objectif de l'architecture U-net est de procéder lentement à un sur-échantillonnage pour obtenir des segmentations fines et ainsi éviter de mauvaises qualités de représentations finales, ce qui serait le cas pour un sur-échantillonnage direct de la base du *"U"* vers la taille d'origine. De plus, dans la partie expansion, l'ajout des feature map de la partie réduction est effectué (exprimé par *copy* dans la Figure 7), celui-ci permettant l'ajout d'informations et donc de meilleurs résultats de segmentation.

Les auteurs de l'article ont montré que leur réseau peut être entraîné à partir de très peu

d'images tout en obtenant néanmoins de bons résultats. Pour cette méthode, la bibliothèque Python Segmentation models sous PyTorch a été utilisée (Yakubovskiy, 2020). L'activation du GPU a été possible (GeForce RTX 2080). L'avantage de cette bibliothèque est que les poids sont pré-entraînés sur le jeu de données 2012 ILSVRC ImageNet (Russakovsky et al., 2015) et permettent une convergence plus rapide et plus robuste des résultats. Les meilleures performances ont été observées pour l'architecture Efficientnet-b5, avec 28 millions de paramètres (Tan et al., 2019).

### 2.4 Métrique d'évaluation

Les performances ont été quantifiées en utilisant la mesure de la racine de l'erreur quadratique moyenne (RMSE) sur la prédiction de la fraction verte afin de travailler à même unité.

Selon $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(r_i - p_i)^2}$, où $N$ est le nombre d'images test, $r_i$ and $p_i$ respectivement la valeur de référence et la valeur prédite de la fraction verte, pour une image $i$.

Dans cette étude, la référence est toujours la fraction verte calculée à partir des masques haute résolution. En effet, les images/masques ont été dégradés comme vu précédemment et les prédictions associées pour les images de test ont été comparé à la fraction verte de référence haute résolution comme vu dans la Figure 9. Ainsi, nous vérifierons, l'efficacité de ces méthodes à haute résolution sans dégradation, puis nous nous concentrerons ainsi sur l'effet de la résolution et l'apparition de pixels mixtes sur les images face à la fraction verte de référence. Les résultats seront présentés sous forme de moyenne (avec prise en compte de l'écart-type) du RMSE sur 30 itérations lorsque des méthodes d'apprentissage sont impliquées (sauf pour RandomForest pour des questions de temps de calcul, avec seulement 5 itérations), une seule itération sinon sur ExG, l'indice étant identique puisque sans apprentissage.

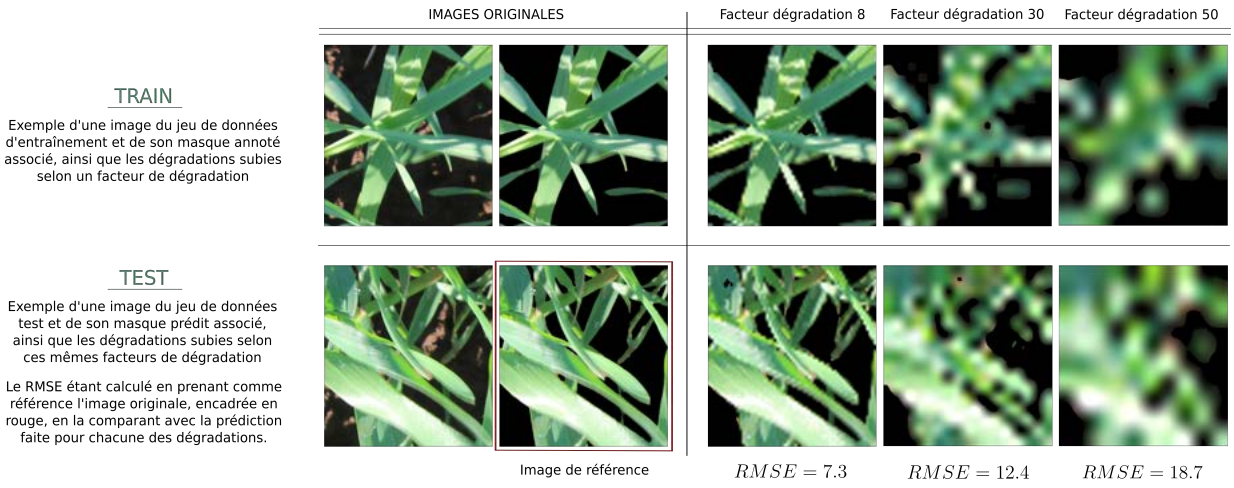


Figure 9: *Exemple de la dégradation, selon plusieurs facteurs, d'une image et d'un masque issues du dataset d'entrainement, ainsi que de la prédiction d'un masque pour une image test dégradée selon les mêmes facteurs. Le calcul du RMSE se faisant selon la formule ci-dessus entre le masque de référence haute résolution encadré en rouge et le masque prédit après dégradation.*

## 3 Résultats et Discussion

### 3.1 Ajustements et affinages des prédictions

#### 3.1.1 De meilleures prédictions à résolutions égales

Des mesures empiriques ont montré que la performance optimale des méthodes basées sur l'apprentissage est obtenue lorsque les images d'entraînement sont à la même résolution que celles du dataset test. Comme le montre la Figure 10, où pour chaque jeu de données d'entraînement, dégradé selon un certain facteur c'est-à-dire une image réduite selon un facteur puis sur-échantillonnée, une prédiction a été faite sur chaque jeu de de donnée test, dégradé

selon ces mêmes facteurs. La figure présente ces résultats en utilisant la méthode U-net, cependant un comportement similaire est observable pour les autres méthodes d'apprentissage, mais une seule méthode y est présentée pour des enjeux de synthétisation. Pour chacune des résolutions dégradées du jeu de données test en abscisses, la meilleure prédiction, *selon une tendance générale*, a été observée à la même résolution pour le jeu de données d'entraînement (la meilleure prédiction étant la valeur minimale du RMSE en tenant compte du croisement des écarts types sur 30 itérations). Le moment, où la résolution du jeu de données d'entraînement est égale à la résolution du jeu de données test, est schématisé par un point de la même couleur. Seuls 10 facteurs de dégradation sont indiqués pour des raisons de clarté graphique.

Ce résultat conforte le principe qu'il faille entraîner sur des images présentant, plus ou moins, les mêmes caractéristiques de dégradation, que les images test. Cela se traduit donc pour la suite de l'étude, par une mise à taille d'images d'entraînement et de test **similaires**, afin d'avoir une dégradation de résolution semblable. Puisque la taille de base des éléments pour chaque sous-dataset étant la même, il en résulte que lors de la réduction d'image, l'on obtient plus ou moins une taille finale d'image similaire selon la taille d'éléments cible, appliquant ainsi ce postulat de prédictions à résolutions égales.

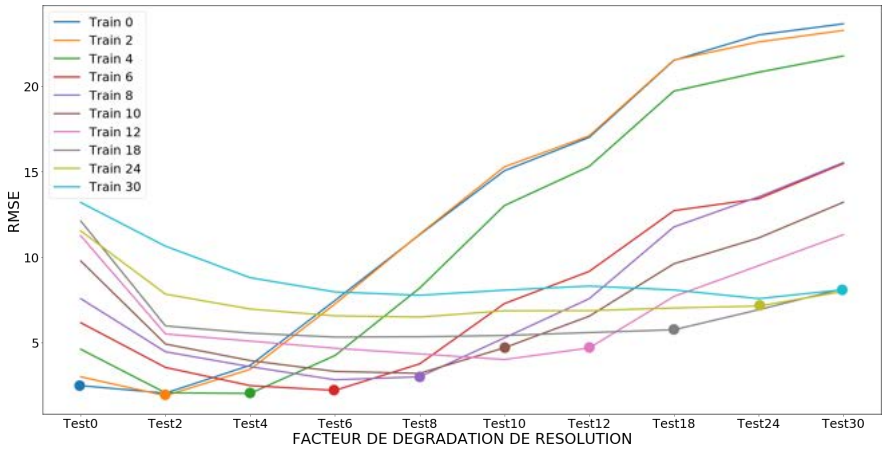


FIGURE 10: *Comparaisons de RMSE face à des datasets d'entrainement et test à des résolutions différentes.*

### 3.1.2 Comparaison de performances entre une dégradation de résolution au travers d'une réduction de l'image par sous-échantillonnage et par un retour à taille initiale par sur-échantillonnage.

Il a été énoncé dans la partie 2.2.2 que deux méthodes ont été utilisé pour dégrader la résolution. La première étant de réduire les images, par sous-echantillonnage, jusqu'à obtenir une taille d'éléments désirée. La seconde étant de reprendre cette image réduite, puis de la sur-échantilloner afin de revenir à la taille initiale. Des comportements différents dans les résultats obtenus furent observés selon la méthode utilisée.

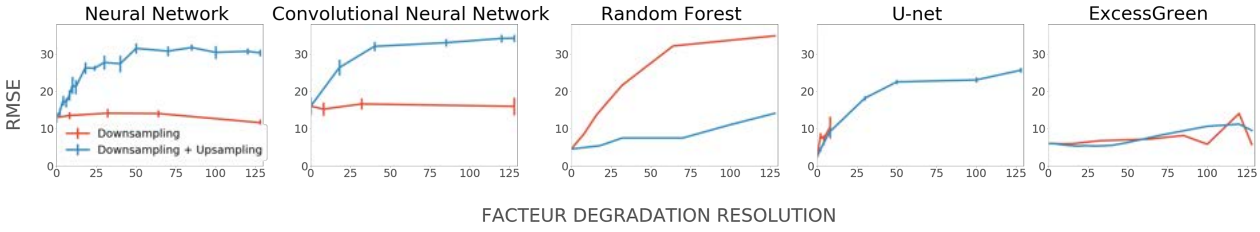


FIGURE 11: *Comparaisons de performances de RMSE entre une dégradation de résolution via souséchantillonnage ou sous+surechantillonnage pour chaque méthode.*

Trois comportements sont identifiables dans la Figure 11 :

- Un comportement similaire pour les deux méthodes, tel que ExcessGreen, ainsi le retour à la taille initiale via suréchantillonnage a été privilégié afin d'obtenir une courbe lisse.
- Le retour à la taille initiale via sur-échantillonnage permet d'obtenir de meilleurs résultats pour RandomForest, basée sur le nombre de voisins autour d'un pixel ainsi que le nombre de morceaux par image, il est aisément compréhensible que cela affecte la qualité de prédiction si l'image se réduit significativement. Pour U-net la raison est tout autre, la dimension de l'image en entrée est fixée comme étant au minimum de 64*64 pixels. Cette contrainte liée à l'architecture même de la méthode oblige à travailler avec des images ayant une taille suffisamment grande, et donc un retour à celle initiale. De plus se limiter à l'étude d'images dégradées au maximum à une résolution de 64*64 pixels ne permettait pas une vision suffisamment large pour l'étude effectuée.
- Enfin, pour les méthodes de régression directe, comme celles-ci se concentrent uniquement sur des textures comme la moyenne sur les trois canaux RVB ou les activations des couches pour prédire directement la fraction verte, et non la calculer au travers d'un masque de segmentation prédit basé sur le contexte proche et/ou global des objets. Il s'avère alors que sur-échantillonner après un sous-échantillonnage induit davantage de pixels mixtes, et de modifications de caractéristiques utilisées pour la régression. Cela impactant alors les variables explicatives des jeux de données, à l'instar de la moyenne des couleurs des images, montrant une différence significative entre la version sur et sous-échantillonnée, alors que ces mêmes variables doivent, par essence, rester fixes lors de la réduction d'images montrant une anomalie.

En résumé, là où sur-échantillonner permet pour les méthodes de segmentation de produire un résultat plus fin, car estimé à partir des masques à dimensions initiales produits. Pour les méthodes de régression simple, se basant sur une estimation directe sans masque, une réduction seule de l'image permet de conserver sans trop de variations les caractéristiques. Celles-ci qui se voyant perturbées à l'arrivée de pixels mixtes, cette introduction de pixels mixtes étant peu utile au vu de l'estimation directe ne nécessitant pas une résolution fine pour la formation d'un masque. Pour la suite de l'étude, les méthodes optimales pour chaque algorithme ont été comparées.

### 3.2 Précision des différentes méthodes pour la récupération de la fraction verte face à un changement de résolution

Selon la méthode expliquée dans la sous-partie Dégradation des images, les images des deux datasets contenant deux tailles distinctes d'éléments crées ont été réduites jusqu'à atteindre une taille d'éléments cible, dégradant ainsi la résolution initiale. Cette taille cible pouvant atteindre l'ordre du pixel, voire inférieur. Il est à noter que l'étude commence donc à des objets de taille 120 pixels, intervient alors à un moment donné le dataset 40 pixels, rendant l'étude biaisée, puisque non basée exactement sur le même dataset, du départ jusqu'à l'arrivée. Ce principe est connu et pris en compte, ainsi toutes les précautions seront prises lors de l'interprétation. Ce protocole ayant pour but en premier lieu, de simplifier l'interprétation en travaillant sur des données relatives à la taille du pixel, plutôt que sur d'abstraits facteurs de dégradation, ne prenant pas en considération la taille des éléments puisque dégradant toutes images quelque soit ses objets, à un même facteur. De plus cela permet de solidifier les résultats à très haute résolution, de par l'injection du dataset 40 pixels étant à pleine résolution. Par conséquent, nul RMSE ne peut être étudié et comparé entre la plage précédant et suivant le seuil 40 pixels, seulement les tendances face à une résolution décroissante.

Les résultats sont montrés dans la Figure 12 ci-contre. Les croix de couleurs représentant les valeurs prises sans l'injection du dataset 40 pixels, à une taille cible d'éléments de 40 pixels, des images 120 pixels.

On constate, en premier temps, une forte perturbation des valeurs de RMSE quand le seuil de taille d'éléments est inférieur au pixel, chose attendue de par le fait que des éléments entiers de taille moyenne inférieure à l'ordre du pixel sont résumés par des pixels de tailles supérieures

à ceux-ci, rendant l'information significativement bruitée avec une grossière représentation et détection des fines variations de traits phénotypiques, quasi-inexistantes dans ce type d'images.

Concernant la performance des méthodes. L'apprentissage profond semblerait permettre de très fins résultats sur des images hautes résolutions au vu de la valeur de départ et celle lors de l'addition d'images 40 pixels à résolution initiale. Cependant les performances sont assez sensibles à la résolution, même face à une faible dégradation, cette méthode ne semblant pas traiter les pixels mixtes convenablement. Cela s'explique de par l'architecture même de U-net, en prenant en compte la branche d'expansion, cela signifie que l'on sur-échantillonne deux fois (au départ pour dégrader la résolution, puis dans U-net) résultant en de nombreux pixels mixtes étant eux-mêmes calculés à partir de l'image dégradée en contenant, pour produire un masque. De par ce principe on ne peut pratiquer une infinie compression-décompression, sans altérer la qualité. Ainsi quand les images viennent à être dégradées modérément, soit la plage entre 20 et 1 pixel(s), les méthodes basées sur le pixel, telles que ExcessGreen et RandomForest, réussissent davantage à estimer fidèlement la fraction verte, de part une segmentation plus fine que la méthode basée sur la région/objet qu'est U-net, au vu des masques prédits montrés en exemple pour des tailles d'éléments de l'ordre d'un pixel.

Rappelons ici que les datasets sont les mêmes, là où ExG par essence, ne se base pas sur une approche supervisée, cela est pourtant le cas pour RF et U-net, montrant ainsi que deux méthodes en apparence similaire puisque basées sur les mêmes images d'entraînement, rendent des résultats très hétérogènes.

Finalement quand la taille des éléments d'étude est inférieure à la taille d'un pixel, à savoir une dégradation très avancée, une estimation directe par régression en se servant de la moyenne sur les trois canaux de couleurs des images semble satisfaire un résultat à la hauteur des méthodes basées sur le pixel et ce de part l'unique fait que la moyenne des couleurs est très peu modifiée lors de la dégradation de la résolution. Des résultats complémentaires non montrés sur la figure indiquent que lorsque nous continuons sous le seuil de 0.5 pixel, pendant un court moment, la méthode NN est significativement meilleure que les méthodes pixels, cependant elle croît par la suite, se situant au même niveau que ces dernières.

Enfin une méthode non mentionnée dans ces résultats, à savoir CNN, ne permet pas de bien estimer la fraction verte, et ce pour n'importe quelle résolution, en effet nous jugeons que de part la complexité des images, le modèle n'arrive pas à associer les features à une estimation quantitative de la fraction verte. Cela s'accentue au seuil inférieur au pixel, étant donné l'approche basée uniquement sur la texture que possède cette méthode.

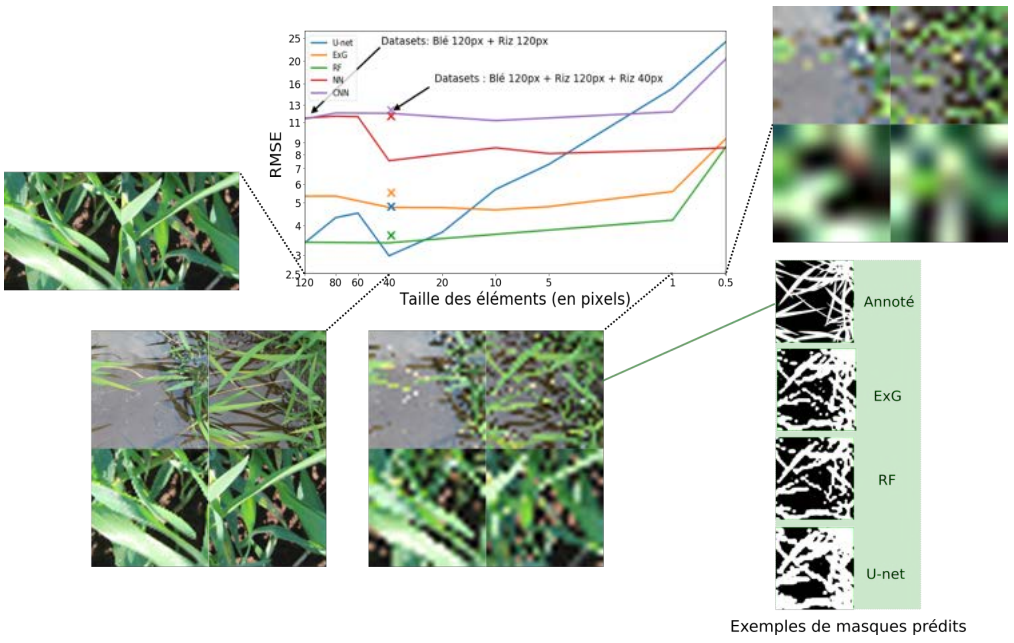


FIGURE 12: *Performances des différentes méthodes face à une résolution dégradante, exprimée en taille d'éléments en pixels.*

### 3.3 Reconnaissance de motifs selon l'échelle spatiale

Pour cette partie, nous avons utilisé exclusivement le dataset sur le riz et les sous-ensembles de données créés au point Création de sous-datasets, afin d'obtenir, à partir d'un ensemble de données initial de haute résolution, des objets de tailles/échelles différentes, à savoir des feuilles larges de 40 et 120 pixels.

Parler ici d'échelles différentes, n'est pas un abus de langage. En effet, même si le dispositif d'acquisition est fixe et ne varie pas, *a contrario* le développement foliaire varie. Ainsi notre échelle est à reconsidérer pour les deux sous-datasets, non seulement par le phénomène de croissance foliaire, mais aussi dû au fait que les feuilles au stade intermédiaire se situent davantage plus proches du capteur, de par leurs hauteurs, que celles au stade précoce rasant le sol, tout cela affectant l'échelle d'étude. Il n'est pas absurde alors de considérer que nous travaillons à deux échelles différentes, puisqu'il s'avère aussi que, pour un stade de développement défini, les feuilles situées au premier plan et au second plan font en moyenne la même taille biologiquement parlant, cependant elles ne se situent pas à la même distance du capteur.

De plus, le concept de taille d'éléments et d'échelle étant liés pour des objets physiques réels, ne pouvant faire varier l'un sans l'autre. En effet si l'on fait varier l'échelle d'un objet d'une taille définie, alors la taille absolue en pixels sur l'image de celui-ci se verra changée, et réciproquement, si l'on veut faire varier la taille de l'objet réel non modifiable, cela implique forcément un changement d'échelle.

Néanmoins on pourra considérer l'hypothèse que l'objet qu'est la feuille, est un objet très similaire dans nos images sélectionnées, et donc pourvu de caractéristiques semblables, pouvant être comparés à différentes échelles. Les modèles pourront reconnaître assez aisément que ce sont des objets similaires, non identiques.

Selon les interprétations faites ci-dessus, les deux méthodes les plus puissantes à très haute résolution sont U-net et RF. Nous allons donc nous concentrer sur celles-ci afin d'étudier l'impact de l'échelle sur l'estimation de la fraction verte dans les algorithmes machine learning.

| | *RandomForest* | | | *U − net* | | | |
|---|---|---|---|---|---|---|---|
| | *Train 40 − pixels* | *Train 120 − pixels* | *Mixte* | *Train 40 − pixels* | *Train 120 − pixels* | *Mixte* | *Mixte data augmentation* |
| *Test 40 − pixels* | **2.6 (0.02)** | **14.2 (1)** | **6.6 (0.02)** | **4.2 (0.6)** | **25.7 (3)** | **1.9 (0.3)** | **2.7 (0.4)** |
| *Test 120 − pixels* | **11.1 (1.4)** | **4.35 (0.01)** | **6 (0.01)** | **14.3 (0.9)** | **2.5 (0.3)** | **1.7 (0.3)** | **3.3 (0.7)** |

TABLE 2: *Comparaisons de résultats pour divers datasets en fonction de la taille des éléments pour des techniques machine learning. Résultats sous la forme de moyenne (écart-type).*

Pour ce faire et comme le montre le Tableau 2, pour les deux algorithmes, nous avons prédit la fraction verte pour une taille d'objet définie en entraînant à la fois sur des tailles d'objet similaires et de tailles différentes. Les ensembles de données sont chacun composés de 30 images d'entraînement et de 10 images test. De plus, la construction d'un jeu de données d'entraînement appelé "***mixte***" a été réalisée. Il s'agit simplement d'un ensemble de données d'entraînement contenant 15 images de feuilles de 120 pixels et 15 images de feuilles de 40 pixels. Celui-ci, comme les autres, sera testé sur les 20 images test mentionnées précédemment. Tous les hyper-paramètres sont égaux quel que soit le jeu de données utilisé. De plus, si aucune mention de Data augmentation n'est explicitée, alors aucune Data augmentation affectant la taille ou la résolution, par le biais d'un zoom par exemple, n'a été utilisée. L'augmentation du nombre de données a été implémenté en utilisant la bibliothèque albumentations (BUSLAEV et al., 2020).

Notons la présence d'un léger biais étant la sélection de 15 images de chaque dataset pour le jeu "***mixte***". Léger dans le sens où ces images proviennent d'un même environnement, rendant existante la variabilité au sein de chaque dataset, mais relativement contrôlée et donc négligeable, au point de pouvoir sélectionner 15 images aléatoirement sans pour autant influencer le modèle. Cela se confirme par le fait que de mêmes interprétations sont faites en utilisant les 15 autres images de chaque dataset dans le modèle mixte.

Incluons ces quelques lignes pour expliciter la réalisation de tests statistiques, vérification de

la normalité sur les résidus, ANOVA et enfin des analyses post-hoc (Tukey), afin de s'assurer de quelles sont les modalités des variables d'étude significativement différentes deux à deux. Il en résulte que chacune des observations a été testée, la significativité prouvée et les interprétations sont citées ci-après.

Une première observation est que l'entraînement sur des objets de tailles similaires à celle des images test à prédire donne de meilleurs résultats que la réciproque, c'est-à-dire un entraînement sur des images de tailles différentes des images de test. On peut donc supposer que les méthodes basent leur représentations sur des caractéristiques échelle-dépendantes, par échelle-dépendant entendons la capacité de ne reconnaître que les caractéristiques sur lesquelles le modèle a été entraîné, à l'échelle à laquelle il a été entraîné. Et de ce fait une sensibilité notoire dans la détection d'objets similaires comme déjà exprimé par Giladi et al., 2011.

Cela signifie-t-il qu'il faille faire correspondre, exactement, l'échelle des objets pour la reconnaissance de structures non-invariantes (les structures invariantes étant décrites ici comme des caractéristiques communes aux objets, quelle que soit la taille considérée) ? Il ne semble pas ; en effet, de meilleurs résultats sont obtenus, pour l'apprentissage profond, lorsque l'on utilise un ensemble de données d'échelles mélangées plutôt qu'un modèle apprenant sur des objets d'échelles identiques. Une variété d'échelles dans un modèle mixte semblerait alors améliorer les performances.

De ce fait, les représentations apprises par apprentissage profond vont-elles vraiment au-delà du stade du pixel comme prétendument ? Il semblerait, au vu du tableau, que la méthode U-net présente de meilleurs résultats pour le modèle "***mixte***" que la méthode RF, ne rendant pas meilleurs les résultats par l'ajout de variétés d'échelles. L'on a alors confirmation que la première base son apprentissage à l'échelle des régions tandis que la seconde à l'échelle des pixels. Une explication possible concernant la performance du modèle U-net serait que plus les informations dont nous disposons sont diverses, plus la représentation globale de l'objet étudié est meilleure. Tout cela grâce aux convolutions, réduisant la taille de l'image de la feature map de sortie à chaque convolution, intégrant alors une variation d'échelle au sein même du modèle. Par exemple, la taille d'une image contenant des objets de 120 pixels va être réduite successivement dans les convolutions, et donc indéniablement la taille de l'objet, permettant ainsi la prise en compte et l'apprentissage de caractéristiques à plusieurs échelles, puisque le filtre lui ne change pas de taille, pouvant servir par la suite, à mieux définir les objets de taille initiale 40 pixels. Ces mêmes convolutions se voyant dotés de filtres aux poids sensibles d'échelles différentes pour chaque image, augmentant ainsi l'efficacité des performances sur les images test. En d'autres termes, on force le modèle à reconnaître des caractéristiques communes (de part la similarité des objets étudiés malgré l'échelle), plutôt que de simplement se concentrer sur la détection d'objets de par leurs tailles, pouvant créer de faux positifs et réduire la performance.

Il est à spécifier de plus, que la méthode RF possède bien moins de degrés de liberté. En effet à mesure que l'entraînement du modèle U-net progresse, une augmentation du nombre de degrés de liberté de part l'utilisation de poids des filtres de convolution est à mentionner. Là où la force de RF à dégradation de résolution modérée était la finesse, impliquant un apprentissage et une prédiction au pixel, face à des représentations grossières de U-net basées sur l'apprentissage et la prédiction région/objet. Cela est aussi sa faiblesse à haute résolution, U-net se représentant les objets globaux beaucoup mieux et permettant ainsi la prise en compte de variations d'échelles et de contexte. Citons aussi un potentiel meilleur apprentissage des caractéristiques liées au background.

Qu'en est-il de la Data augmentation ? En général, lorsqu'on utilise de la Data augmentation (plus précisément de la résolution augmentation), l'on augmente le nombre d'images afin d'ajouter de la variété, chose qui pourrait s'avérer intéressant au vu des précédents résultats. Nous savons aussi qu'un motif échelle-dépendant ne peut être trouvé dans les images test que s'il a été appris pour la même échelle, d'après les résultats antérieurs. Si nous voulons avoir une chance de potentiellement trouver ces motifs au travers du réseau, il est nécessaire de redimensionner toutes les images selon le facteur d'échelle de chacune d'entre elles, un même nombre de fois, ce qui n'est habituellement pas fait dans la plupart des cas de Data augmentation, la norme étant de

travailler avec des plages aléatoires. Cette norme protocolaire a d'ailleurs été testé (par l'action de zooms ou l'ajout de bords noirs autour de l'image pour l'éloigner en gardant sa dimension initiale) et comme le montre le tableau, de moins bons résultats que la simple addition manuelle de variété sont à noter. En outre, ce processus entraîne bien souvent un zoom (*in* ou *out*) sur les images plutôt qu'une modification explicite de l'échelle des objets, ce qui conduit à une réduction de la résolution et, comme vu précédemment, les méthodes d'apprentissage profond, type U-net, sont assez sensibles, même à de faibles changements de résolution. Cela étant dit, il serait alors préférable, selon cette étude, d'implémenter directement et manuellement des images avec des objets de tailles différentes dans les ensembles de données.

# 4 Conclusion

En raison de la puissance du GPU et des résultats à haute résolution, l'apprentissage profond, type U-net, semble le plus approprié mais reste néanmoins très sensible aux dégradations de résolution. La classification des pixels par RF est également une valeur sûre, mais avec un temps de calcul d'inférence assez levé notamment dû à une non-compatibilité de la bibliothèque mahotas pour extraire les caractéristiques de Haralick avec le GPU, cette méthode semble donc perfectible. Cependant face à une dégradation de l'image modérée, les méthodes basées sur le pixel (ExG et RF) prennent le dessus et semblent minimiser la perte d'informations. Bien qu'elles soient plus efficaces, il s'agit toujours de méthodes présentant un écart assez important dû à la modification même de l'‘image de par la diminution de résolution (en moyenne et en valeur absolue $\tilde{5}\%$ en dehors des valeurs réelles). Enfin quand l'image est dégradée, au point d'atteindre des pixels plus larges que le taille des objets composant l'image, des méthodes de régression afin d'estimer directement la fraction verte, au travers de la moyenne des couleurs d'‘images, semblent satisfaire une estimation la plus fidèle, ou du moins du même ordre de grandeur que les méthodes pixels.

Les techniques utilisées ont fait en sorte que de nombreux paramètres furent contrôlés, pour cela un axe d'amélioration potentiel serait le traitement en aval des images dégradées. L'explosion des réseaux adverses pour la Super-résolution (technique visant à générer des textures similaires à celles initialement perdues durant le processus de dégradation), de type ESRGAN (Wang et al., 2019), et très récemment de l'intérêt de développer des méthodes alternatives (Menon et al., 2020) peuvent être un sujet d'étude quant à la question d'une reconstruction partielle de la qualité perdue lors de l'acquisition d'images à basse résolution. Puisse-t-elle être meilleure que l'approche de l'interpolation bicubique, ou au contraire complémentaire. Est-il possible de récupérer la résolution par des techniques de type GAN et donc d'étendre les champs d'application du Deep Learning sensibles à la résolution, au point de pouvoir amener le Deep Learning aux mêmes performances que les autres méthodes à basse résolution ?

## Références


ABADI, Martın et al., 2016. Tensorflow : A system for large-scale machine learning. In : *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, p. 265-283.

BRADSKI, G., 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

BUSLAEV, Alexander ; IGLOVIKOV, Vladimir I. ; KHVEDCHENYA, Eugene ; PARINOV, Alex ; DRUZHININ, Mikhail ; KALININ, Alexandr A., 2020. Albumentations : Fast and Flexible Image Augmentations. *Information* [online]. T. 11, n° 2, p. 125 [visité le 2020-08-05]. Disponible à l'adresse DOI : 10.3390/info11020125. Number : 2 Publisher : Multidisciplinary Digital Publishing Institute.

CHEN, Tianqi ; GUESTRIN, Carlos, 2016. XGBoost : A Scalable Tree Boosting System. In : *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA : ACM, p. 785-794. KDD '16. ISBN 978-1-4503-4232-2. Disponible à l'adresse DOI : 10.1145/2939672.2939785.

CHOLLET, Francois et al., 2015. *Keras*. GitHub. **urlalso** : https://github.com/fchollet/keras.

CRACKNELL, Arthur P. *Review article Synergy in remote sensing-what's in a pixel ? : International Journal of Remote Sensing : Vol 19, No 11* [online] [visité le 2020-07-03]. Disponible à l'adresse : https://www.tandfonline.com/doi/abs/10.1080/014311698214848.

CRESSIE, Noel ; HAWKINS, Douglas M., 1980. Robust estimation of the variogram : I. *Journal of the International Association for Mathematical Geology* [online]. T. 12, n° 2, p. 115-125 [visité le 2020-07-03]. ISSN 1573-8868. Disponible à l'adresse DOI : 10.1007/BF01035243.

DESAI, Sai Vikas ; BALASUBRAMANIAN, Vineeth N. ; FUKATSU, Tokihiro ; NINOMIYA, Seishi ; GUO, Wei, 2019. Automatic estimation of heading date of paddy rice using deep learning. *Plant Methods* [online]. T. 15 [visité le 2020-07-09]. ISSN 1746-4811. Disponible à l'adresse DOI : 10.1186/s13007-019-0457-1.

DUMOULIN, Vincent ; VISIN, Francesco, 2016. A guide to convolution arithmetic for deep learning.

EMILY, Mathieu, 2019. *Cours sur le Machine learning | Neural network [Diapositives]*. Département de statistique et informatique | Agrocampus Ouest, Rennes. **urlalso** : https://tice.agrocampus-ouest.fr/pluginfile.php/69214/mod_resource/content/4/ML6NN.pdf.

GILADI, Itamar ; ZIV, Yaron ; MAY, Felix ; JELTSCH, Florian, 2011. Scale-dependent determinants of plant species richness in a semi-arid fragmented agro-ecosystem : Scale-dependent plant diversity in an agro-ecosystem. *Journal of Vegetation Science* [online]. T. 22, n° 6, p. 983-996 [visité le 2020-07-31]. ISSN 11009233. Disponible à l'adresse DOI : 10.1111/j.1654-1103.2011.01309.x.

GITELSON, Anatoly ; KAUFMAN, Yoram ; STARK, Robert ; RUNDQUIST, Donald, 2002. Novel Algorithms for Remote Estimation of Vegetation Fraction. *Remote Sensing of Environment*. T. 80, p. 76-87. Disponible à l'adresse DOI : 10.1016/S0034-4257(01)00289-9.

GLUCKMAN, J., 2006. Scale Variant Image Pyramids. T. 1, p. 1069-1075. ISBN 978-0-7695-2597-6. Disponible à l'adresse DOI : 10.1109/CVPR.2006.265.

GRINGARTEN, Emmanuel ; DEUTSCH, Clayton, 2001. Teacher's Aide Variogram Interpretation and Modeling. *Mathematical Geology*. T. 33, p. 507-534. Disponible à l'adresse DOI : 10.1023/A:1011093014141.

GUO, Wei ; ZHENG, Bangyou ; DUAN, Tao ; FUKATSU, Tokihiro ; CHAPMAN, Scott ; NINOMIYA, Seishi, 2017. EasyPCC : Benchmark Datasets and Tools for High-Throughput Measurement of the Plant Canopy Coverage Ratio under Field Conditions. *Sensors (Basel, Switzerland)* [online]. T. 17, n° 4 [visité le 2020-07-31]. ISSN 1424-8220. Disponible à l'adresse DOI : 10.3390/s17040798.

HARALICK, Robert ; SHANMUGAM, K. ; DINSTEIN, Ih, 1973. Textural Features for Image Classification. *IEEE Trans Syst Man Cybern.* T. SMC-3, p. 610-621.

HENGL, Tomislav, 2006. Finding the right pixel size. *Computers & Geosciences* [online]. T. 32, n° 9, p. 1283-1298 [visité le 2020-07-30]. ISSN 0098-3004. Disponible à l'adresse DOI : 10.1016/j.cageo.2005.11.008.

HSIEH, Pi-Fuei ; LEE, Lou ; CHEN, Nai-Yu, 2002. Effect of spatial resolution on classification errors of pure and mixed pixels in remote sensing. *Geoscience and Remote Sensing, IEEE Transactions on.* T. 39, p. 2657-2663. Disponible à l'adresse DOI : 10.1109/36.975000.

JIN, Xiuliang ; LIU, Shouyang ; BARET, Frédéric ; HEMERLÉ, Matthieu ; COMAR, Alexis, 2017. Estimates of plant density of wheat crops at emergence from very low altitude UAV imagery. *Remote Sensing of Environment* [online]. T. 198, p. 105-114 [visité le 2020-07-30]. ISSN 0034-4257. Disponible à l'adresse DOI : 10.1016/j.rse.2017.06.007.

JONES, Hamlyn ; SIRAULT, Xavier, 2014. Scaling of Thermal Images at Different Spatial Resolution : The Mixed Pixel Problem. *Agronomy ISSN 2073-4395.* T. 4. Disponible à l'adresse DOI : 10.3390/agronomy4030380.

LI, Lei ; ZHANG, Qin ; HUANG, Danfeng, 2014. A Review of Imaging Techniques for Plant Phenotyping. *Sensors (Basel, Switzerland).* T. 14, p. 20078-20111. Disponible à l'adresse DOI : 10.3390/s141120078.

LI, Yanan ; HUANG, Ziyun ; CAO, Zhiguo ; LU, Hao ; WANG, Haihui ; ZHANG, Shuiping, 2020. Performance Evaluation of Crop Segmentation Algorithms. *IEEE Access.* T. 8, p. 36210-36225. ISSN 2169-3536. Disponible à l'adresse DOI : 10.1109/ACCESS.2020.2969451. Conference Name : IEEE Access.

LIU, Shouyang ; BARET, Frederic ; ANDRIEU, Bruno ; BURGER, Philippe ; HEMMERLE, Matthieu, 2017. Estimation of Wheat Plant Density at Early Stages Using High Resolution Imagery. *Frontiers in Plant Science* [online]. T. 8, p. 10 p. [Visité le 2020-07-30]. Disponible à l'adresse DOI : 10.3389/fpls.2017.00739.

LONG, Jonathan ; SHELHAMER, Evan ; DARRELL, Trevor, 2015. Fully Convolutional Networks for Semantic Segmentation. *arXiv :1411.4038 [cs]* [online] [visité le 2020-08-05]. Disponible à l'adresse : http://arxiv.org/abs/1411.4038.

MAHLEIN, Anne-Katrin, 2015. Plant Disease Detection by Imaging Sensors – Parallels and Specific Demands for Precision Agriculture and Plant Phenotyping. *Plant Disease* [online]. T. 100, n° 2, p. 241-251 [visité le 2020-07-03]. ISSN 0191-2917. Disponible à l'adresse DOI : 10.1094/PDIS-03-15-0340-FE. Publisher : Scientific Societies.

MÄLICKE, Mirko ; SCHNEIDER, Helge, 2019. *Scikit-GStat 0.2.6 : A scipy flavored geostatistical analysis toolbox written in Python.* Disponible à l'adresse DOI : 10.5281/zenodo.3531816.

MARCIAL, Mariana ; GONZÁLEZ-SANCHEZ, Alberto ; JIMÉNEZ, Sergio ; ONTIVEROS-CAPURATA, Ronald Ernesto ; OJEDA, Waldo, 2018. Estimation of vegetation fraction using RGB and multispectral images from UAV. *International Journal of Remote Sensing*, p. 1-19. Disponible à l'adresse DOI : 10.1080/01431161.2018.1528017.

MCKERNS, Michael M. ; STRAND, Leif ; SULLIVAN, Tim ; FANG, Alta ; AIVAZIS, Michael A. G., 2012. Building a Framework for Predictive Science. *arXiv :1202.1056 [cs]* [online] [visité le 2020-08-05]. Disponible à l'adresse : http://arxiv.org/abs/1202.1056. arXiv : 1202.1056.

MENON, Sachit ; DAMIAN, Alexandru ; HU, Shijia ; RAVI, Nikhil ; RUDIN, Cynthia, 2020. *PULSE : Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models.*

OJALA, Timo ; PIETIKÄINEN, Matti ; HARWOOD, David, 1996. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition* [online]. T. 29, n° 1, p. 51-59 [visité le 2020-07-03]. ISSN 0031-3203. Disponible à l'adresse DOI : 10.1016/0031-3203(95)00067-4.

OTSU, Nobuyuki, 1985. A Threshold Selection Method from Gray-Level Histograms, p. 5.

POREBSKI, A. ; VANDENBROUCKE, Nicolas ; MACAIRE, Ludovic, 2008. Haralick feature extraction from LBP images for color texture classification. In : p. 1-8. Disponible à l'adresse DOI : 10.1109/IPTA.2008.4743780.

POUND, Michael ; ATKINSON, Jonathan ; WELLS, Darren ; PRIDMORE, Tony ; FRENCH, Andrew, 2017. Deep Learning for Multi-task Plant Phenotyping. In : disponible à l'adresse DOI : 10.1109/ICCVW.2017.241.

RONNEBERGER, Olaf ; FISCHER, Philipp ; BROX, Thomas, 2015. U-Net : Convolutional Networks for Biomedical Image Segmentation. *arXiv :1505.04597 [cs]* [online] [visité le 2020-08-05]. Disponible à l'adresse : http://arxiv.org/abs/1505.04597. arXiv : 1505.04597.

RUSSAKOVSKY, Olga et al., 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*. T. 115, n° 3, p. 211-252. Disponible à l'adresse DOI : 10.1007/s11263-015-0816-y.

SPALDING, Edgar ; MILLER, Nathan, 2013. Image analysis is driving a renaissance in growth measurement. *Current opinion in plant biology*. T. 16. Disponible à l'adresse DOI : 10.1016/j.pbi.2013.01.001.

TAN, Mingxing ; LE, Quoc V., 2019. *EfficientNet : Rethinking Model Scaling for Convolutional Neural Networks*. Disponible à l'adresse arXiv : 1905.11946 [cs.LG].

TEHRANI, Mohammad maleki ; EMERY, Xavier ; MERY, Nadia, 2017. Indicator Variograms as an Aid for Geological Interpretation and Modeling of Ore Deposits. *Minerals*. T. 7, p. 241. Disponible à l'adresse DOI : 10.3390/min7120241.

WAGNER, Fabien ; IPIA, Alber ; TARABALKA, Yuliya ; LOTTE, Rodolfo ; FERREIRA, Matheus ; P.M, Aidar ; GLOOR, Manuel ; PHILLIPS, Oliver ; ARAGÃO, Luiz, 2019. Using the U-net convolutional network to map forest types and disturbance in the Atlantic rainforest with very high resolution images. *Remote Sensing in Ecology and Conservation*. T. 5. Disponible à l'adresse DOI : 10.1002/rse2.111.

WALTER, Achim ; LIEBISCH, Frank ; HUND, Andreas, 2015. Plant phenotyping : From bean weighing to image analysis. *Plant Methods*. T. 11, p. 14. Disponible à l'adresse DOI : 10.1186/s13007-015-0056-8.

WANG, Xintao ; YU, Ke ; WU, Shixiang ; GU, Jinjin ; LIU, Yihao ; DONG, Chao ; QIAO, Yu ; LOY, Chen Change, 2019. ESRGAN : Enhanced Super-Resolution Generative Adversarial Networks : Munich, Germany, September 8-14, 2018, Proceedings, Part V. In : p. 63-79. ISBN 978-3-030-11020-8. Disponible à l'adresse DOI : 10.1007/978-3-030-11021-5_5.

WOEBBECKE, D. ; MEYER, George ; BARGEN, K. ; MORTENSEN, David, 1995. Color Indices for Weed Identification Under Various Soil, Residue, and Lighting Conditions. *Transactions of the ASAE*. T. 38, p. 259-269. Disponible à l'adresse DOI : 10.13031/2013.27838.

YAKUBOVSKIY, Pavel, 2020. *Segmentation Models Pytorch* [https://github.com/qubvel/segmentation_models.pytorch]. GitHub.

**Diplôme :** Master

**Spécialité :** Science des données pour la biologie

**Enseignant référent :** Marie-Pierre Etienne

**Auteur(s) :** SEROUART Mario

**Date de naissance :** 06/08/1996

**Nb pages :** 20 **Annexe(s) :** 0

**Année de soutenance :** 2020

**Organisme d'accueil :** INRAE AVIGNON

**Adresse :**

228 route de l'Aérodrome
CS 40509
84914 Avignon cedex 9

**Maître de stage :** Frederic Baret

**Titre français :** Analyse comparative de la sensibilité de méthodes pour l'étude de la prédiction de la fraction verte au travers de la résolution et de la diversité d'échelles spatiales.

**Titre anglais :** Comparative analysis of the sensitivity of methods for studying the prediction of the green fraction through the resolution and diversity of spatial scales.

**Résumé :** Le phénotypage à haut débit se développe rapidement pour les applications de sélection variétale et d'agriculture de précision. La fraction verte, la fraction de pixels verts dans une image, est l'un des traits les plus utiles pour suivre le développement de la végétation qui peut être extrait d'images à haute résolution prises à partir d'une gamme de systèmes. Cependant, cette gamme, aussi variée soit-elle, se traduit par une diversité de résolution d'images et de taille d'éléments différentes, ce qui implique indéniablement une altération de la précision de la fraction verte estimée, en raison de la fraction de pixels mélangés, et des informations pertinentes sur les objets étudiés variant selon les échelles spatiales. Une première phase d'étude a consisté à estimer la taille moyenne des éléments composant l'image au travers de variogrammes, puis à simuler artificiellement une dégradation de la résolution spatiale des images. Les objectifs de l'étude proposée sont d'évaluer et de comparer les performances de différents types d'approches afin d'estimer la sensibilité relative de la prédiction de la fraction verte à la résolution et à la diversité des échelles spatiales. Il a été prouvé que les méthodes basées sur l'apprentissage profond, sont les plus efficaces à haute résolution. Pour une dégradation modérée les méthodes basées sur les pixels prennent le dessus et semblent minimiser la perte d'information. Enfin une résolution fortement dégradée, implique l'utilisation de méthodes de régression afin d'estimer directement la fraction verte. Une seconde étude portée sur l'apprentissage des motifs échelle-dépendants est aussi présentée, il en résulte qu'une insertion de diversité de taille d'objets permait de meilleurs résultats grace à une meilleure représentation de l'objet dans sa globalité pour l'apprentissage profond, à l'inverse pour les méthodes basées sur le pixel.

**Abstract :** High throughput phenotyping is developing rapidly for varietal selection and precision farming applications. The green fraction, the fraction of green pixels in an image, is one of the most useful features for tracking vegetation development that can be extracted from high-resolution images taken from a range of systems. However, this range, as varied as it may be, results in a diversity of image resolution and feature sizes, which undeniably implies an alteration in the accuracy of the estimated green fraction, due to the mixed pixel fraction, and the fact that relevant information on the objects under study varies at different spatial scales. A first study phase consisted in estimating the average size of the elements composing the image through variograms, then artificially simulating a degradation of the spatial resolution of the images. The objectives of the proposed study are to evaluate and compare the performance of different types of approaches in order to estimate the relative sensitivity of green fraction prediction to the resolution and diversity of spatial scales. Methods based on deep learning have been shown to be the most effective at high resolution. For moderate degradation, pixel-based methods take over and seem to minimize information loss. Finally, a strongly degraded resolution implies the use of regression methods to directly estimate the green fraction. A second study focused on the learning of scale-dependent patterns is also presented, it results that an insertion of object size diversity allowed better results thanks to a better representation of the object as a whole, for deep learning, and conversely for pixel-based methods.

**Mots-clés :** Vision par ordinateur, Apprentissage profond, Résolution spatiale, Imagerie RGB

**Key Words :** Computer Vision, Deep Learning, Spatial Resolution, RGB Imaging