# LE DÉFI DE L'ESTIMATION ROBUSTE DES TRAITS AVEC L'APPRENTISSAGE PROFOND SUR DES IMAGES RVB A HAUTE RÉSOLUTION

## *THE CHALLENGE OF ROBUST TRAIT ESTIMATES WITH DEEP LEARNING ON HIGH RESOLUTION RGB IMAGES*

Thèse de Étienne **DAVID**, [1]

Analysé par Guy **WAKSMAN**[2]

Directeur de thèse : Frédéric **BARET** (UMR EMMAH - Environnement Méditerranéen et Modélisation des Agro-Hydrosystèmes, INRAE)
Encadrant : Benoit **de SOLAN** (Arvalis)

La thèse de E. David est tout à fait impressionnante pour différentes raisons. Nous savons tous la complexité des métiers de l'obtention et de la multiplication végétale pour ce qui concerne les espèces annuelles comme pour les arbustives et les arbres. L'obtention variétale s'appuie bien entendu sur les meilleures techniques (génétique, biologie, biochimie, biostatistique) mais reste – pour le néophyte qu'est l'auteur de ces lignes – un grand art, avec un côté quasi-magique. La « magie » ne suffisant pas, nous voyons, dans cette thèse, que les technologies de capture et de traitement d'images haute résolution viennent à son secours.

Les défis et enjeux du phénotypage à haut débit (HTPP) restant encore peu connus, il parait cependant nécessaire en préalable à l'analyse de cette thèse de rappeler et souligner l'importance de ces approches qui passent maintenant au « Deep Learning », approche déjà utilisée dans d'autres domaines innovants de l'industrie.

Nous connaissons tous les contraintes nombreuses qui pèsent sur l'agriculture, liées au changement climatique (stress biotiques), aux interactions avec d'autres organismes (insectes et pathogènes), aux contraintes économiques (coûts des intrants) et enfin réglementaires et sociétales directes ou indirectes. Une des solutions consiste à sélectionner plus rapidement et mieux, et dans un raccourci extrême, passer à la sélection génomique.

Les outils sont presque tous là : séquençage, génétique d'association, utilisation de la variabilité génétique et des mutants (voire en les générant avec les nouvelles techniques d'édition de type Crispr/cas), analyses statistiques, *etc*. Or ces analyses doivent reposer sur un très grand nombre et une diversité de données, seuls capables de fiabiliser les résultats d'essais au champ, et donc du processus de sélection.

Pour être « idéales », ces données de champs doivent être nombreuses, sur de nombreux critères, parfois recueillies au cours du temps, à différents stades, *etc*. On en comprend immédiatement la charge de travail considérable qui grève l'acquisition des données. Mais il

[1] Thèse de doctorat de l'Université d'Avignon, INRAE – UMR EMMAH 114 (Environnement Méditerranéen et Modélisation des Agro-Hydrosystèmes), et ARVALIS – Institut du Végétal, Ecole doctorale 536 -Spécialité Sciences Agronomiques, soutenue le 2 novembre 2021

[2] Membre de l'Académie d'agriculture de France, section 9 « Agrofournitures »

y a plus problématique : certaines données ne sont pas des mesures, mais des notations selon une échelle (par exemple des symptômes de maladie), fréquemment de 0 à 9. Chaque personne va interpréter ce qu'elle voit pour attribuer une note, et parfois cette interprétation varie selon le jour ou le moment de la journée. Bien évidemment, il existe aussi une variabilité de lecture entre deux notateurs.

La « robotisation » de ces mesures et notations est un plus indéniable en diminuant la charge d'acquisition des données, et de bons algorithmes vont pouvoir réduire ou éliminer les variations évoquées précédemment et dues aux opérateurs.

L'intérêt, pour l'amélioration génétique, du phénotypage haut débit (HTPP) en grandes cultures, est confirmé (Chapitre 1), même si le lien avec les enjeux environnementaux, qui est mis en avant, n'est sans doute pas plus fort que le lien avec les enjeux de rendement, ou de régularité de ce dernier, par exemple. L'outil HTPP ne parait pas en effet dédié strictement à l'utilisation qui en est faite dans cette thèse. En tous cas, ce chapitre permet de s'initier au HTPP.

A la base de la première partie de cette thèse, il y a un « jeu » de données (des images !) « décrivant » 16 000 plantes collectées sur plusieurs sites expérimentaux. La possibilité d'appliquer les mêmes traitements (ou modèles) par les mêmes méthodes de l'intelligence artificielle, sur des données issues des différents sites, a été validée (Chapitre 2). Preuve de l'intérêt des méthodes d'exploration des données mises en œuvre, en particulier les méthodes « d'apprentissage profond ». Il est rassurant d'observer que l'ajout de quelques images annotées dans la base d'entrainement (partie des données sur laquelle les modèles sont mis au point) améliore grandement les performances.

Les travaux rapportés dans cette thèse ont abouti à la publication de deux bases de données d'images annotées en accès libre (Open Data), sous un format désormais largement adopté et accepté comme état de l'art au sein de la communauté du phénotypage. Les promoteurs de cette approche espèrent que ces données seront valorisées par d'autres acteurs de l'agriculture numérique. Nous pouvons en douter, mais pourquoi pas ? Une innovation en tout cas.

Dans la seconde partie de cette thèse, des jeux d'images de blé, provenant de dizaines d'organismes dans 12 pays distincts, ont été rassemblés et constituent un ensemble d'une grande diversité de pratiques agricole, de protocoles d'acquisition ou d'environnement. Il contient plus de 275 000 instances d'épis de blé annotées : un corpus intéressant pour l'évaluation des méthodes de détection et comptage des épis (Chapitre 3).

Cette évaluation a fait l'objet d'un concours sur deux plateformes numériques, dans le but de comparer les modèles de détection d'objet reconnus comme les plus efficaces. L'engagement de « la communauté du phénotypage » dans ce concours a été très significatif, et les solutions d'apprentissage profond proposées par les participants ont été très variées. Un tel résultat n'aurait pas été atteint sans l'organisation de ce concours (Chapitre 4).

De tels concours permettent ainsi - et ceci est une innovation étonnante - des partages sur les algorithmes les plus efficaces et ont donc une vocation pédagogique puisqu'ils permettent une montée collective en compétence.

Enfin, la thèse présente les résultats d'un travail sur l'évaluation des méthodes de « machine learning » pour la détection et le comptage d'objets dans différents contextes (Chapitre 5).

Ainsi la thèse de E. David contribue aussi bien au progrès du phénotypage haut débit qu'au développement des méthodes d'intelligence artificielle sur deux aspects au moins : l'idée de proposer des jeux de données et d'organiser une compétition pour le traitement et l'interprétation des résultats de ce traitement. Cette thèse pose aussi la question du passage du traitement de données d'entrainement au traitement de données « grandeur nature dans la vraie vie ».

Les outils de comptage des plantes développés et présentés au chapitre 1, et de comptage des épis de blé présentés au chapitre 3, sont aujourd'hui utilisés par Arvalis.

M. E. David a publié des articles dans Plant Phenomics et à la Conférence internationale sur l'apprentissage automatique (ICML). Il a soumis d'autres manuscrits à Frontiers in Plant Science, ainsi qu'au European Journal of Agronomy et à Gigascience.

En résumé, cette thèse montre toute la démarche mise en place pour le phénotypage haut débit : les outils, les critères, l'utilisation de données test (plusieurs milliers de photos), puis de données de validation (plusieurs dizaines de milliers de photos) partagées pour une optimisation des algorithmes, avec l'utilisation possible du « Deep Learning » pour affiner la précision et la robustesse des modèles utilisés.

Cette approche devrait permettre de réaliser un phénotypage à haut débit rapide et efficace, qualitativement et quantitativement à la hauteur des technologies de séquençage et de génomique.

La qualité et la quantité des travaux de cette thèse très originale et novatrice autorisent que cette analyse figure sur le site de l'Académie d'agriculture de France, à titre de valorisation.

Ci-après le résumé en français de cette thèse écrite en anglais.

**Le défi de l'estimation robuste des traits avec l'apprentissage profond sur des images RVB à haute résolution**

Le phénotypage à haut débit des plantes, notamment dans le cadre d'acquisitions en plein champ, repose sur l'interprétation de données issues de différents capteurs mis en œuvre sur des vecteurs variés tels que des tracteurs, des robots ou des drones. Initialement, ces données étaient interprétées à l'aide d'algorithmes de télédétection exploitant la résolution spectrale du signal.

Mais depuis 2015, les progrès du "Deep Learning", basé sur l'entrainement à partir d'exemples, ont permis des résultats prometteurs pour mesurer des traits essentiels comme le taux de couverture ou le comptage de plantes ou d'organes. Ces algorithmes utilisent des couches de convolution apprises, permettant de tirer parti de l'organisation spatiale du signal.

L'avantage de ces méthodes est qu'elles sont basées sur des capteurs Rouge-Vert-Bleu (RVB), qui sont beaucoup moins coûteux que les imageurs multi- ou hyperspectraux. Cependant, les algorithmes de « Deep Learning » sont sensibles aux changements de la distribution entre les données utilisées pour l'entrainement et les données prédites.

En pratique, des erreurs de prédiction, variables et non prédictibles d'un site à l'autre, peuvent être observées.

L'objectif de la thèse est de comprendre les causes de ces variations et de proposer des solutions pour des estimations de traits phénotypiques fiables en utilisant le « Deep Learning ». L'étude porte sur la détection de plantes et d'organes à partir d'images RVB haute résolution acquises sur le terrain. Nos travaux ont d'abord porté sur la constitution de bases de données d'images diversifiées provenant de différents lieux et stades de développement pour l'émergence de plantes (maïs, betterave, tournesol) et les épis de blé, ce qui a permis la publication de deux bases de données annotées, regroupant 27 sessions d'acquisition pour le drone et 47 pour la détection d'épis.

Ces jeux de données démontrent la différence de performances entre les résultats publiés et les nôtres en raison du changement de distribution.

Pour dépasser les limites des méthodes habituelles, nous avons organisé deux concours de données, les Global Wheat Challenges, en 2020 et 2021, qui nous ont permis d'obtenir des solutions entraînées pour la robustesse sur un jeu de données différent de celui de l'entraînement.

L'analyse des solutions a montré l'importance des stratégies d'entraînement pour la robustesse, au-delà des architectures utilisées. Nous avons également montré que ces solutions peuvent être déployées efficacement en remplacement du comptage manuel.

Enfin, nous avons démontré l'inefficacité des fonctions d'entraînement conçues pour l'entraînement robuste. Notre travail ouvre la perspective d'une meilleure évaluation du « Deep Learning » dans le contexte du phénotypage à haut débit et donc de la confiance dans son utilisation en conditions réelles.

https://hal.archives-ouvertes.fr/tel-03431192/