Mise au point de méthodes hybrides combinant apprentissage machine et modélisation physique pour l'estimation de traits foliaires par spectroscopie

F. de Boissieu[1] D. Ienco[1] J.-B. Féret[1]

[1]TETIS, Irstea, AgroParisTech, CIRAD, CNRS, Université Montpellier, Montpellier, France

florian.deboissieu@irstea.fr

1 Introduction

Le contenu biochimique des feuilles présente un intérêt pour nombre d'applications dans les domaines de l'agriculture et de l'écologie. Parmi les paramètres associés à cette composition chimique, la masse surfacique foliaire (Leaf Mass per Ares, LMA) et le contenu surfacique en eau (Equivalent Water Thickness, EWT) sont particulièrement importants pour l'étude fonctionnelle des individus et des écosystèmes, et la compréhension des mécanismes adaptatifs aux changements climatiques [1], [2]. La mise au point de méthodes opérationnelles pour la mesure précise de ces traits foliaires peut s'appuyer sur la spectroscopie combinée à différentes approches basées sur des outils d'analyse de données tels que les méthodes statistiques et l'apprentissage machine [3], [4], ou des outils de modélisation physique [5], [6]. Chaque type d'approche présente des avantages et inconvénients. Les performances de l'apprentissage machine dépendent fortement des données d'apprentissage et présentent une capacité de généralisation limitée, alors que les méthodes physiques sont plus généralisables mais nécessite des méthodes de résolution couteuses en temps de calcul.

L'approche hybride permet théoriquement de combiner la robustesse des méthodes physiques et la rapidité d'exécution des méthodes statistiques ou d'apprentissage machine [7], [8]. Cette approche consiste à entrainer un modèle de régression sur des données de simulation pour ensuite l'appliquer à des données expérimentales. La mise au point de cette base de données simulées nécessite cependant un certain nombre de précautions afin de prendre en compte les incertitudes associées au modèle physique et d'éviter le sur-apprentissage. Une stratégie consiste à appliquer un bruit aux données simulées. Une autre est de sélectionner les domaines spectraux sur lequel le modèle est le plus fiable.

L'objectif de ce papier est d'évaluer le potentiel de méthodes hybrides pour l'estimation des paramètres EWT et LMA. Nous nous appuyons sur le modèle PROSPECT [5] pour générer des bases de données d'apprentissage auxquelles nous appliquons différents niveaux de bruit. La base de données simulées est utilisée pour ajuster des modèles de régression à partir de méthodes d'apprentissage machine. Enfin, les modèles sont évalués sur des données expérimentales.

2 Données

Une base de données de propriétés biophysiques foliaires de structure (paramètre *N* de PROSPECT), de EWT et de LMA est générée par tirage aléatoire de 10 000 échantillons sur la distribution normale multivariées proposée par [8]. Les spectres correspondants de réflectance et de transmittance directionnelles hémisphériques sont simulés à l'aide de PROSPECT dans le domaine infrarouge allant de 900 nm à 2400 nm.

Les résultats de [8] montrent que l'ajout de bruit gaussien aux propriétés optiques simulées permet d'améliorer l'applicabilité de modèles de régression définis a l'aide d'indices spectraux sur des données expérimentales. Dans le cas de modèles hybrides, nous testons l'hypothèse selon laquelle un niveau de bruit approprié appliqué aux propriétés optiques peut permettre de réduire l'influence de domaines pour lesquels le modèle n'est pas suffisamment réaliste, et limiter les risques de sur-apprentissage. Dans le cadre de cette étude, la base de propriétés optiques simulées est déclinée avec différents niveaux de bruit gaussien correspondant à 0%, 0.5%, 1%, 2%, 3%, 4%, et 5% de bruit absolu.

Le travail s'appuie également sur deux bases de données expérimentales disponibles en ligne, ANGERS et LOPEX [9], [10]. Ces bases de données sont constituées de 306 et 317 échantillons foliaires respectivement. Elles comprennent les mesures de LMA et EWT ainsi que les propriétés de réflectance et transmittance directionnelles hémisphériques mesurées de 400nm à 2500 nm. Seul le domaine de 900 nm à 2400 nm est considéré ici car l'influence des paramètres EWT et LMA sur le signal dans le domaine visible et proche infrarouge est négligeable, et le rapport signal sur bruit des données optiques diminue fortement au-delà de 2400 nm.

3 Méthodes

La version la plus récente du modèle PROSPECT [5] est utilisée dans le cadre de cette étude. L'estimation de la chimie foliaire est réalisée à l'aide de l'algorithme d'optimisation Sequential Quadratic Programming implémenté dans la fonction Matlab *fmincon*, qui est une procédure d'optimisation itérative non linéaire avec contraintes, cherchant à minimiser la fonction de mérite *M* :

$$M(N, EWT, LMA) = \sum_{\lambda=\lambda_1}^{\lambda_n} \left[\left(R_\lambda - \hat{R}_\lambda\right)^2 + \left(T_\lambda - \hat{T}_\lambda\right)^2 \right] \quad (1)$$

Avec $R_\lambda$et T_λ la réflectance et transmittance mesurées à la longueur d'onde λ, et $\hat{R}_\lambda$ et $\hat{T}_\lambda$ leur équivalent simulé par PROSPECT. Une étude en préparation montre que l'utilisation d'un domaine spectral réduit de 1800 nm à 2400 nm permet d'améliorer significativement les performances pour l'estimation des constituants foliaires. Cette inversion est donc effectuée pour les deux intervalles spectraux de 900 nm à 2400 nm, et de 1800 nm à 2400 nm [11].

Deux algorithmes d'apprentissage machine sont utilisés pour ajuster des modèles de régression: l'algorithme Epsilon Support Vector Regression (SVR, [12]) et l'algorithme Random Forest (RF [13]). Le modèle SVR est entrainé avec les paramètres par défaut, i.e. un noyau RBF, ε=0.1, C=1, $\gamma=1/N_X$ avec N_X le nombre de variables d'entrées du modèle. Le modèle RF est entrainé avec une forêt de 400 arbres. Les modèles sont entrainés sur les variables centrées normées.

Ces algorithmes sont classiquement entrainés a partir de données expérimentales. Les modèles de régression sont ajustés sur une base de données et évalué sur la seconde, et inversement. Pour les modèles hybrides, ces algorithmes sont entrainés sur les données simulées et évaluées sur les données expérimentales.

Les performances d'apprentissage des modèles hybrides sont testés en fonction du niveau de bruit, i.e. chaque base de données prise indépendamment ou cumulée aux bases de données de niveaux de bruit inférieur. Comme pour l'inversion du modèle physique, deux intervalles spectraux sont aussi testés.

4 Résultats

Les modèles de régression sont comparées en fonction de leur score en Root Mean Square Error. La Figure 1 présente les performances en fonction niveau de bruit (ponctuel ou cumulé), de l'intervalle de longueur d'ondes d'apprentissage et du type de model (RF et SVR). On notera tout d'abord que les performances des RF sont meilleures dans l'ensemble que celles des SVR. En se concentrant sur les modèles RF, les performances des modèles d'estimation de EWT sont relativement insensibles à l'intervalle de longueur d'onde considéré. Ce n'est pas le cas pour l'estimation du LMA qui montre une réduction de la RMSE de 50% lorsque les données d'entrées sont restreintes à l'intervalle λ=[1800, 2400] nm. La Figure 1 montre également que l'ajout de bruit apporte peu voire dégrade les performances de régression SVR. Pour les modèles RF, les conclusions sont plus mitigées. On observe une légère réduction de la RMSE avec l'ajout de bruit cumulé une fois le meilleur intervalle de longueurs d'ondes identifié.

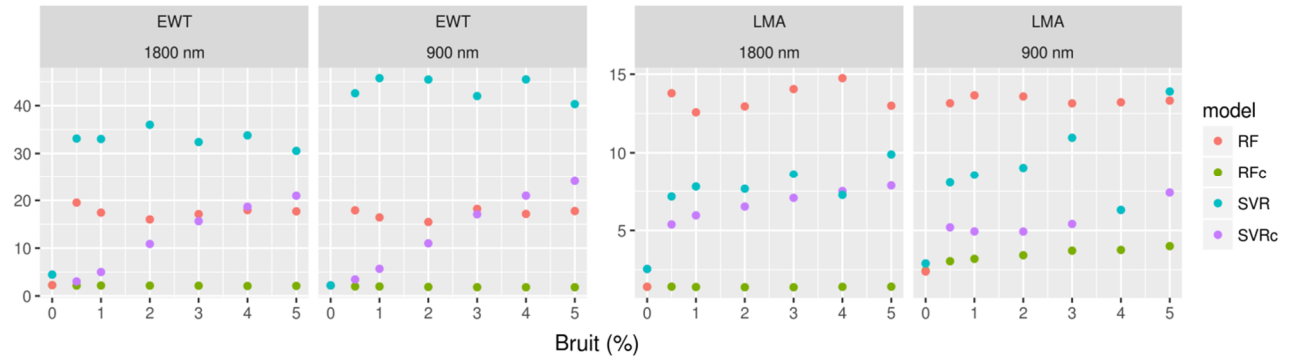

Figure 1: RMSE des modèles hybrides en fonction de l'algorithme et du niveau de bruit des données d'apprentissage. L'indice 'c' signifie cumulé. En en-tête, la borne inférieur de l'intervalle spectral est précisée.

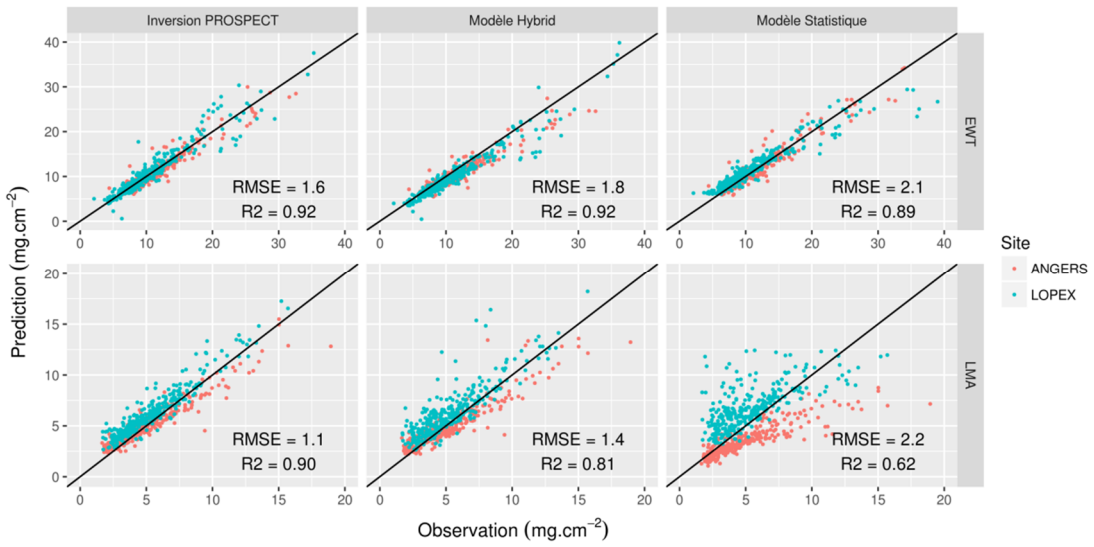

Figure 2: Comparaison des différentes modèles d'estimation de EWT et LMA. Seuls les meilleurs modèles selon l'intervalle spectrale, l'algorithme d'apprentissage et le niveau de bruit à l'apprentissage, sont utilisés ici.

Les résultats de régression, présentés dans la Figure 2, montrent les performances des meilleurs modèles de chaque type (inversion du modèle physique, modèle hybride, modèle statistique). On peut noter que les modèles d'apprentissage machine entraines à partir de données expérimentales sont clairement moins performants que les approches physiques et hybrides, notamment pour l'estimation du LMA. Bien que légèrement moins bonnes que l'inversion PROSPECT par optimisation, les performances des modèles hybrides restent comparables, même pour des valeurs extrêmes.

5 Discussion & Conclusions

Cette étude s'intéresse au potentiel des méthodes d'apprentissage machine pour produire des modèles génériques de propriétés biophyiques foliaires à partir de propriétés optiques simulées par le modèle physique PROSPECT. Le principal écueil de ce type d'approche dite hybride est le risque d'apprentissage de caractéristiques du modèle physique non généralisable à la réalité (défauts du modèles, sensibilité au bruit de mesure, etc.). Afin d'éviter ce type de problèmes, deux stratégies ont été testées dans cette étude.

La première stratégie consiste à restreindre l'intervalle de longueurs d'ondes aux domaines les plus caractéristiques, i.e. [900, 2400] nm pour EWT et [1800, 2400] nm pour LMA. Cette stratégie montre une nette amélioration des performances notamment pour l'estimation du LMA, en éliminant les longueurs d'ondes pour lesquelles le modèle PROSPECT est peu précis.

La seconde stratégie consiste à ajouter du bruit sur les résultats de simulations, dans le but concentrer l'apprentissage sur les longueurs d'ondes les mieux simulée par le modèle physique. Les résultats montrent que l'ajout de bruit permet une légère amélioration des résultats de régression.

Contrairement aux modèles statistiques basés uniquement sur des données expérimentales, les modèles hybrides testés ici montrent une meilleure robustesse à la généralisation sur plusieurs sites, et de meilleures performances aux valeurs extrêmes. En comparaison de l'inversion de PROSPECT, ils permettent d'atteindre des performances similaires.

Ces résultats contribuent à mieux définir les conditions d'application des méthodes d'apprentissage machine à l'échelle de la canopée, dont l'utilisation est bien plus développée qu'à l'échelle de la feuille. Ils ouvrent également la voie à d'autres expériences, avec des méthodes d'apprentissage profond nécessitant un nombre de données inaccessibles par mesures terrain. Des expériences sont en cours avec des Réseaux de Neurones Convolutifs [14] pour identifier les domaines spectraux pertinents, améliorer l'estimation de EWT et LMA, et élargir à d'autres propriétés biophysiques.

Références

[1] C. Violle *et al.*, "Let the concept of trait be functional!," *Oikos*, vol. 116, no. 5, pp. 882–892, May 2007.

[2] I. J. Wright *et al.*, "The worldwide leaf economics spectrum," *Nature*, vol. 428, no. 6985, pp. 821–827, Apr. 2004.

[3] G. P. Asner *et al.*, "Taxonomy and remote sensing of leaf mass per area (LMA) in humid tropical forests," *Ecol. Appl.*, vol. 21, no. 1, pp. 85–98, 2011.

[4] G. P. Asner, R. E. Martin, A. J. Ford, D. J. Metcalfe, and M. J. Liddell, "Leaf chemical and spectral diversity in Australian tropical forests," *Ecol. Appl.*, vol. 19, no. 1, pp. 236–253, Jan. 2009.

[5] J.-B. Féret, A. A. Gitelson, S. D. Noble, and S. Jacquemoud, "PROSPECT-D: Towards modeling leaf optical properties through a complete lifecycle," *Remote Sens. Environ.*, vol. 193, pp. 204–215, May 2017.

[6] R. Colombo *et al.*, "Estimation of leaf and canopy water content in poplar plantations by means of hyperspectral indices and inverse modeling," *Remote Sens. Environ.*, vol. 112, no. 4, pp. 1820–1834, Apr. 2008.

[7] J. Verrelst, J. P. Rivera, M. Mardashova, and J. Moreno, "ARTMO's Global Sensitivity Analysis (GSA) toolbox to quantify driving variables of leaf and canopy radiative transfer models." EARSeL eProceedings, 2015.

[8] J.-B. Féret *et al.*, "Optimizing spectral indices and chemometric analysis of leaf chemical properties using radiative transfer modeling," *Remote Sens. Environ.*, vol. 115, no. 10, pp. 2742–2750, 2011.

[9] J.-B. Féret *et al.*, "PROSPECT-4 and 5: Advances in the leaf optical properties model separating photosynthetic pigments," *Remote Sens. Environ.*, vol. 112, no. 6, pp. 3030–3043, Jun. 2008.

[10] B. Hosgood, S. Jacquemoud, G. Andreoli, J. Verdebout, A. Pedrini, and G. Schmuck, "Leaf Optical Properties Experiment 93 (LOPEX93)," Joint Research Centre, Institute for Remote Sensing Applications, European Commission EUR 16095 EN, 1994.

[11] J.-B. Feret, G. Le Maire, and S. Jay, "Estimating leaf mass per area and equivalent water thickness based on leaf optical properties: potential and limitations of physical modeling and machine learning," in prep.

[12] Z. Wang, T. Chu, L. A. Choate, and C. G. Danko, "Rgtsvm: Support Vector Machines on a GPU in R," *ArXiv170605544 Cs Stat*, Jun. 2017.

[13] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *J. Mach. Learn. Res.*, vol. 12, p. 2825−2830, Oct. 2011.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *ArXiv151203385 Cs*, Dec. 2015.