Analyse des incertitudes associées aux prédictions de la teneur en argile obtenues par imagerie hyperspectrale VNIR/SWIR aéroportée (0.4-2.5µm)

Gomez C. [1], Drost A. [1,2], Roger J.-M.[3]

[1] IRD, UMR LISAH (INRA-IRD-SupAgro), F-34060 Montpellier, France
[2] Centre for Geo-Information, Wageningen University, 6708 PB Wageningen, The Netherlands
[3] IRSTEA, UMR ITAP, Montpellier, France

1 Contexte et Objectif

Un nombre croissant d'études ont montré que l'imagerie spectroscopique Visible Proche-Infrarouge (VNIR/SWIR, 0.4-2.5µm) aéroportée peut fournir une estimation géo-localisée de plusieurs propriétés de sol telles que l'Argile, la Matière Organique ou le Carbonate de Calcium. Ces estimations sont généralement réalisées par le biais de modèles de régression multi-variés construits en Cross-validation sur des bases de données d'étalonnage, puis validées sur des bases de données indépendantes. La performance globale de ces modèles, et donc des cartes des propriétés de sol estimées, est étudiée à travers l'analyse d'indices tels que l'Erreur Standard de Prédiction (SEP), le coefficient de détermination (R^2), ou encore l'Erreur quadratique moyenne (RMSEP).

Au-delà de ces indices reflétant la performance globale des modèles, l'analyse de l'erreur et de l'incertitude affectant chaque nouvelle prédiction reste un enjeu. On peut définir l'erreur comme « l'écart entre la valeur prédite et la valeur vraie », et on peut définir l'incertitude comme « la variance des prédictions » (Zeaiter et al., 2004).

Cette étude s'intéresse à l'estimation et l'interprétation de l'incertitude associée à chaque nouvelle prédiction d'argile réalisée par régression multi-variée.

2 Matériel

Les données spectrales considérées dans cette étude sont des données aéroportées acquises par le capteur AISA-DUAL (280 bandes spectrales entre 0.4-2.5µm), sur le Bassin Versant du Lebna (300km², Tunisie, Figure 1a), avec une résolution spatiale de 5m (Gomez et al., 2012).

Un masque des zones végétalisées, urbaines et des zones d'eau a été appliqué afin de ne garder que les zones de sol nu (44% de la surface totale).

Un total de 129 échantillons de sol de surface a été récolté sur la zone d'étude du Lebna. La teneur en Argile (granulométrie < 2µm) a été mesurée en laboratoire pour chacun des 129 échantillons.

L'analyse des résultats se concentre particulièrement sur le bassin versant de Kamech (rectangle noir sur la Figure 1b).

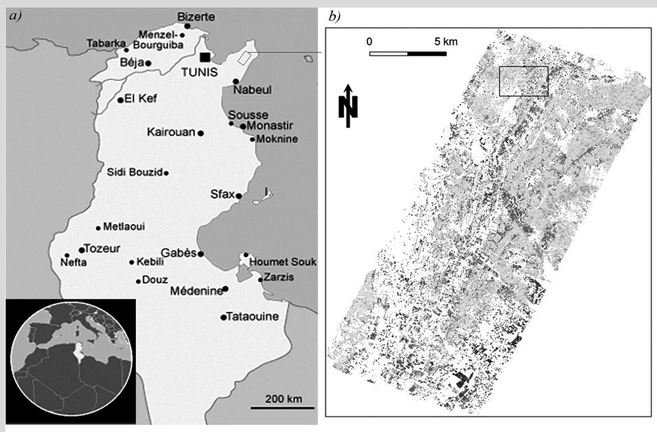

Figure 1: a) Localisation de l'image hyperspectrale (rectangle noir) en Tunisie, b) Localisation du Bassin Versant de Kamech (rectangle noir) sur l'image hyperspectral AISA-DUAL (en blanc sont représentées les zones masquées)

3 Méthodologie

Un ***modèle de régression PLS*** (Partial Least Square, Wold et al., 2001) a été construit afin d'estimer le taux d'argile $\hat{y}$ à partir d'une base de données de calibration (2/3 des 129 échantillons) et validé par une base de données indépendantes (1/3 des 129 échantillons).

Chaque estimation $\hat{y}$ correspondant à un nouvel échantillon **x** peut s'écrire :

$$\hat{y} = f(\mathbf{X}c, \mathbf{Y}c, Model, \mathbf{x})$$

où $(\mathbf{X}c, \mathbf{Y}c)$ contient les spectres de calibration et leur valeur d'argile, et *Model* représente la calibration, incluant le prétraitement et le choix des dimensions du modèle PLSR. Donc chaque estimation d'argile $\hat{y}$ est issue d'une suite d'opérations auxquelles correspondent pour chacune une source d'incertitude.

7 expressions d'incertitudes affectant l'estimation $\hat{y}$ sont calculées dans cette étude :

- La variance $\boldsymbol{var(\hat{y})_{BS}}$ de *R* prédictions issues d'un bootstrap (avec *R=999*), représente la somme de toutes les incertitudes et est considérée comme la variance "vraie" (Figure 2a et d).
- La Distance de Mahalanobis ***MD*** représente la distance entre le spectre **x** et les spectres de la BD_Calib$_0$ projetés dans un espace ACP de dimension réduite (Figure 2b).
- Le Leverage ***H*** représente la distance entre le spectre **x** et les spectres de la BD_Calib$_0$ projetés dans l'espace du modèle PLS (Figure 2b).
- L'expression $\boldsymbol{var(\hat{y})_{Ahumada}}$, proposée par Fernandez-Ahumada et al. (2012), représente une estimation de la variance des prédictions telle que :

$$\mathrm{var}(\hat{y})_{\mathbf{Ahumada}} = \underbrace{\left(1+\frac{1}{N}\right)\mathbf{b}'\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{b}}_{\boldsymbol{T1}} + \underbrace{\mathbf{z}'\boldsymbol{\Sigma}_{\mathbf{b}}\mathbf{z}}_{\boldsymbol{T2}} + \underbrace{\left(1+\frac{1}{N}\right)\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{x}}\boldsymbol{\Sigma}_{\mathbf{b}})}_{\boldsymbol{T3}}$$

Où ***T1***, ***T2*** et ***T3*** sont respectivement l'incertitude liée au spectre **x** et à ses plus proches voisins (Figure 2c), au modèle de PLS, et à l'intersection des variances dues au modèle et au spectre **x** (Figure 2d).

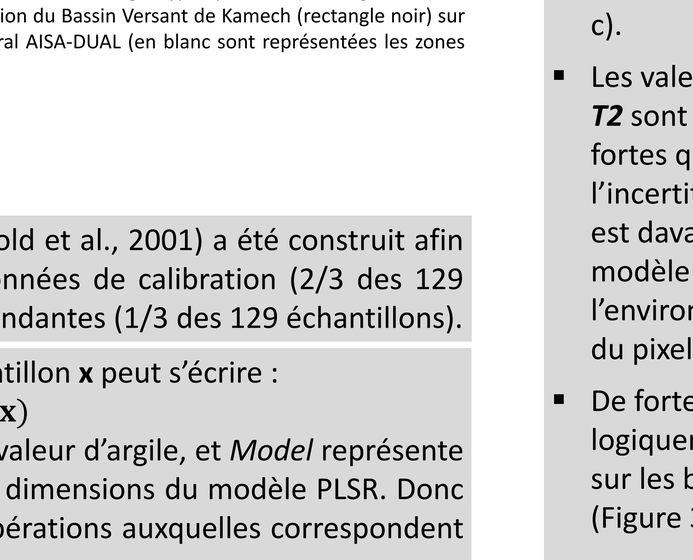

Figure 2 : Procédure de cartographie d'Argile associée au calcul d'expressions d'incertitudes.

4 Résultats

Les modèles PLSR de prédiction de taux d'Argile construits dans cette étude ont des performances acceptables : R^2_{val} = 0,7 et RPD = 1,88.

- Les prédictions d'Argile reliées à des valeurs d'incertitudes très faibles pour chacun des 7 termes (***T1, T2, T3, H, MD, …***) peuvent être considérés comme des prédictions « fiables ». C'est notamment le cas de la plupart des pixels de sol nu de la zone (Figure 3).
- Les prédictions d'Argile reliées à des valeurs d'incertitudes très fortes pour chacun des 7 termes (***T1, T2, T3, H, MD, …***) peuvent être considérés comme des prédictions « erronées ». C'est notamment le cas de certains pixels situés en zone urbaine et n'ayant pas été masqués lors des prétraitements (rectangle noir, Figure 3a).
- Les cartes de ***H*** et ***MD*** permettraient d'améliorer le masque de zones de « non-sol » (Figure 3 b et c).
- Les valeurs d'incertitudes ***T2*** sont beaucoup plus fortes que ***T1***. Donc l'incertitude de prédiction est davantage liée au modèle PLSR qu'à l'environnement spectral du pixel étudié.
- De fortes valeurs ***T1*** sont logiquement observées sur les bords de parcelle (Figure 3e).

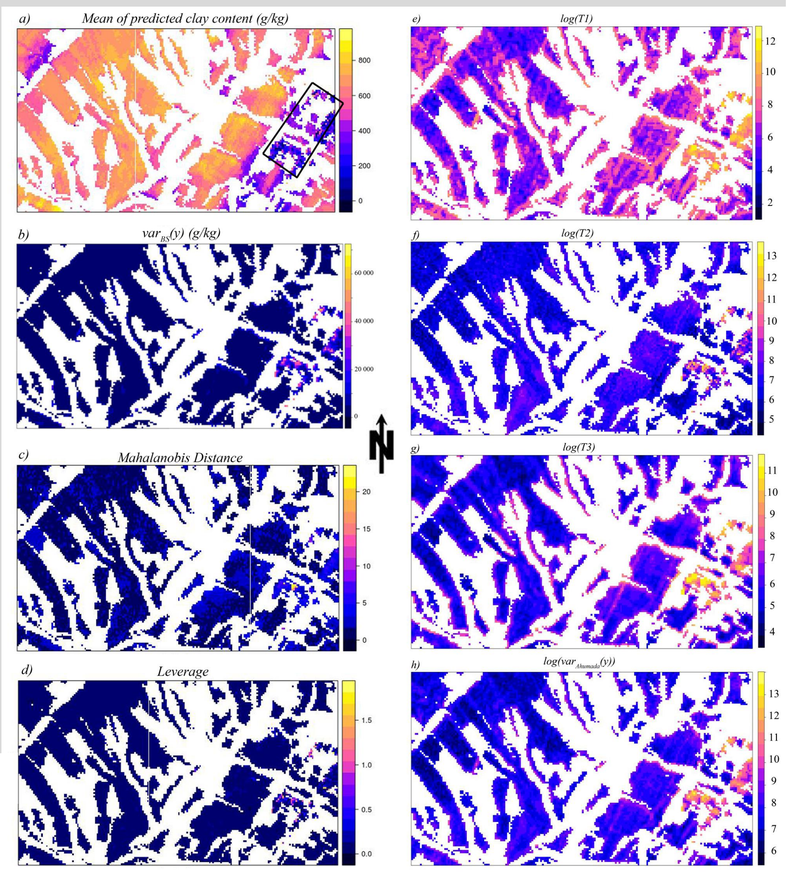

Figure 3 : Cartes a) des valeurs estimées d'argile $\hat{y}$, et des 7 expressions d'incertitude : b) "vraie" variance $\boldsymbol{var(\hat{y})_{BS}}$, c) ***MD***, d) ***H***, e) ***T1***, f) ***T2***, g) ***T3***, et h) $\boldsymbol{var(\hat{y})_{Ahumada}}$. Les zones en blanc correspondent à des pixels préalablement masqués.

5 Conclusions et perspectives

Ces travaux montrent l'intérêt d'utiliser les cartes d'incertitudes de prédiction, afin d'améliorer :

i) le masque de pixels de « non-sol »,
ii) l'échantillonnage des données de calibration
iii) l'analyse de qualité des cartes de prédiction de propriété de sol.

Afin de renforcer ces conclusions, il serait intéressant d'étendre l'étude à différentes propriétés de sol, différents modèles multi-variées et contextes pédologiques.

Références :

Gomez, C., Lagacherie P., & Bacha, S. (2012). Using an VNIR/SWIR hyperspectral image to map topsoil properties over bare soil surfaces in the Cap Bon region (Tunisia). In *"Digital Soil Assessments and Beyond"* Minasny B., Malone B.P., McBratney A.B. (Ed.). Springer, 387-392.
Fernandez-Ahumada, E., Roger, J.M., & Palagos, B. (2012). A new formulation to estimate the variance of model prediction. Application to near infrared spectroscopy calibration. *Analytica Chimica Acta*. 721:28-34.
Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of Chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58, 109-130.
Zeaiter, M., Roger, J.M., & Bellon-Maurel, V. (2004). Robustness of models developed by multivariate calibration. Part I: The assessment of robustness. *Trends in Analytical Chemistry*, 23(2), 157-170.

Ces travaux ont été financés par le projet TOSCA-CNES « HUMPER - Mission HYPXIM : Apport de la résolution spatiale de la mission HYPXIM pour l'étude des propriétés pérennes des sols et de leur humidité de surface » (2013-2014) et le projet MISTRAL 2011 Sicmed-Lebna "Biophysical and socio-economical analysis of water management within the Tunisian Cap Bon Peninsula: the Lebna study area". Les auteurs remercient l'UMR LISAH (IRD, France) et le CNCT (Centre National de Cartographie et de Télédétection, Tunisie), pour avoir fournit les données hyperspectrales AISA-Dual.

Contact : cecile.gomez@ird.fr
UMR LISAH, Montpellier