EVALUATION DE LA PREDICTION DE LA CORUBE DE RETENTION PAR MACHINE LEARNING

EVALUATING THE PREDICTION OF THE WATER RETENTION CURVE BY MACHINE LEARNING

Adel ABDALLAH [1]
[1] *Université de Lorraine, LEMTA, CNRS UMR 7563, Vandœuvre-lès-Nancy, F-54500, France*

RÉSUMÉ – Cette communication présente une évaluation de la prédiction de la courbe de rétention de l'humidité des sols non saturés par trois techniques de machine learning. Les données issues de la base de données UNSODA (Leij *et al.*, 1996) ont dû être réorganisées et nettoyées pour constituer un jeu d'apprentissage et de validation pour ces modèles. Les prédictions en chemin de séchage se sont montées meilleures qu'en humidification.

ABSTRACT – This paper presents an evaluation of the prediction of the soil water retention curve using three techniques of machine Learning. Data extracted from the UNSODA database (Leij *et al.*, 1996) needed to be reorganized and cleansed before being used for training and validating the models. The prediction models are shown to perform better on the drying path than on the wetting path.

1. Introduction

La résolution des problèmes d'écoulement dans les sols non saturés est nécessaire dans de nombreux projets géotechniques mais également en agronomie et dans les applications environnementales. Une estimation aussi précise que possible des propriétés hydrodynamiques des sols à l'état non saturé est nécessaire dans ce contexte. Ces propriétés sont la courbe de rétention de l'humidité reliant le degré de saturation ou la teneur en eau volumique à la succion d'une part ; et la courbe de conductivité hydraulique non saturée, reliant la conductivité hydraulique à la succion, d'autre part. La détermination expérimentale de ces deux courbes passe par des essais complexes et longs qui donnent bien souvent des résultats entachés d'une incertitude assez importante (notamment pour la conductivité hydraulique), qui couvrent une gamme de succion limitée imposée par la technique utilisée, et qui dépendent du chemin suivi pour les obtenir (séchage ou humidification du sol). Il est donc nécessaire de caler des modèles théoriques (d'origine empirique ou analytique) sur les résultats expérimentaux obtenus afin de donner une description complète et continue de la courbe de rétention et de la courbe de conductivité hydraulique. Une revue complète des différentes méthodes expérimentales et des différents modèles théoriques les plus souvent utilisés est donnée par (Fredlund *et al*., 2012) entre autres.

De nombreux auteurs ont proposé des fonctions appelées : « fonctions de pédo-transfert », qui permettent de décrire les propriétés hydrodynamiques des sols non saturés, en utilisant des données d'identification physique et géotechnique des sols plus faciles à obtenir comme la courbe granulométrique ou les différentes masses volumiques du sol (Aria et Paris, 1981 ; Fredlund *et al*., 2002). Ces fonctions sont en général, obtenues par des régressions statistiques effectuées sur des données locales. Elles s'expriment en fonction de paramètres qui sont le plus souvent reliés aux paramètres d'un modèle théorique décrivant une seule ou les deux courbes parmi lesquels, le modèle de

(van Genuchten, 1980 et Mualem, 1976) est le plus utilisé pour son adaptabilité à une large gamme de sols. La validité de ces équations reste toutefois restreinte au type de sols et à la gamme de succion pour lesquels, elles ont été obtenues (Patil et Singh, 2016).

Durant les dernières années, des méthodes issues de l'Intelligence Artificielle et plus particulièrement du champ du Machine Learning ont été utilisées comme des outils de régression plus puissants qui se basent sur les données disponibles pour permettre la prédiction de la courbe de rétention de l'humidité des sols non saturés (Nguyen et *al.*, 2017 ; Lamorski *et al.*, 2017 et D'Emiliano *et al*., 2018). Les jeux de données utilisés comprennent des mesures concernant de 104 à 359 sols et les auteurs ont utilisé différentes méthodes de machine learning incluant notamment : machine à support vecteur (*Support Vector Machine* SVM) et les Réseaux de Neurones Artificiels (*Artificial Neural Networks* ANN). Ils ont prédit la courbe de rétention en utilisant les teneurs en sable, en limon, en argile, la densité du sol et pour les deux premiers, la teneur en carbonates. Pour les deux premières références, les auteurs ont comparé la prédiction de la courbe de rétention point par point en incluant la succion dans les prédicteurs, et la prédiction des paramètres du modèle de (van Genuchten, 1980) et ils n'ont pas signalé des différences majeures de performance.

Dans cette communication, nous nous proposons d'évaluer la performance de différentes méthodes de machine learning (arbres de régression ou *Regression Tree* RT, SVM et ANN) pour prédire la courbe de rétention des sols point par point. Cette approche n'impose pas la forme de la courbe prédite, elle peut également prédire indifféremment les point de la courbe en séchage et en humidification, et serait plus adéquate pour s'intégrer dans un code de calcul numérique. Les données servant à l'apprentissage et à la validation des modèles sont extraites par une requête dans la base de données UNSODA (Lej *et al.,* 1996).

2. Présentation des jeux de données

UNSODA est une base de données qui a pour vocation de rassembler des données sur les propriétés hydrodynamiques des sols non saturés (courbe de rétention de l'humidité, conductivité et diffusivité hydrauliques) déterminées par des essais au laboratoire et *in situ*, à la fois en séchage et en humidification. Elle précise pour chaque point, le type d'essai et elle complète les données expérimentales par des données d'identification des sols correspondants (granulométrie, masses volumiques et porosité). Les données sont stockées dans une base Microsoft Access® et organisées en plusieurs tables. Les données de UNSODA couvrent 790 sols différents provenant des cinq continents avec une nette prédominance des sols d'Amérique du nord et d'Europe. La figure 1 présente la position de ces sols dans le diagramme ternaire de l'*United States Department of Agriculture* (USDA). Elle montre une surreprésentation claire des sols sableux et limoneux par rapport aux sols argileux.

Les données extraites pour cette étude forment deux jeux de données résultats de deux requêtes recherchant des données de courbes de rétention obtenues au laboratoire en séchage (jeu 1) et en humidification (jeu 2). Au total, les deux jeux de données représentent 65 534 points de courbe de rétention (succion ψ et teneur en eau volumique θ) ainsi que 7 paramètres d'identification complémentaires : la taille de particules, la fraction massique correspondante, la masse volumique apparente (ρ), la masse volumique des grains (ρ_s), la porosité (n), la teneur en eau volumique à la saturation (θ_{sat}) et la conductivité hydraulique saturée (k_{sat}).

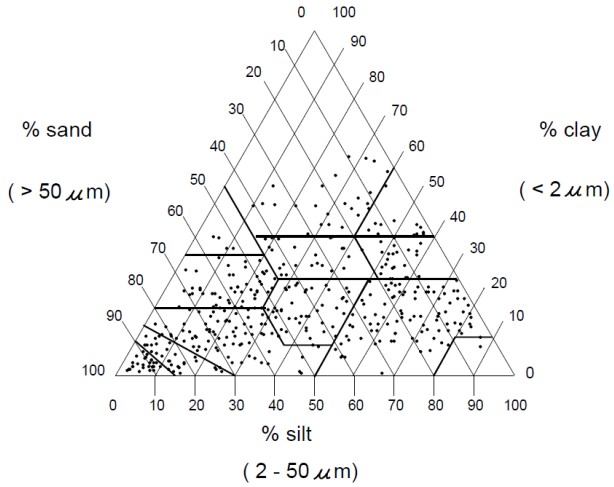

Figure 1. Diagramme ternaire de l'USDA pour les sols représentés dans la base de données UNSODA (Leij *et al.*, 1996).

2.1. Préparation des données

L'examen des données extraites de la base de données a montré que pour intégrer les données de la granulométrie, les points de la courbe de rétention sont dupliqués autant de fois qu'il y a de fractions de taille de grains dans la courbe granulométrique. Il a donc été nécessaire de réorganiser les données et de supprimer les points redondants. Un script Matlab® a été créé pour nettoyer les jeux de données. La granulométrie a été intégrée en créant trois paramètres supplémentaires correspondant aux fractions des tailles de grains inférieures à 2, 50 et 500 µm ($P_{2\mu}$, $P_{50\mu}$ et $P_{500\mu}$).

Par ailleurs, la conductivité hydraulique saturée k_{sat}, la teneur en eau volumique saturée θ_{sat}, et la masse volumique apparente ρ, n'étaient pas exploitables à la vue du pourcentage de valeurs manquantes (atteignant plus de 80%), ces paramètres ont donc été supprimés.

Il convient de noter la diminution importante de la quantité des enregistrements qui résulte de la réorganisation et du nettoyage des données. En effet, la taille des jeux de données est passée à 1 551 enregistrements pour les données de séchage et à 295 enregistrements données d'humidification.

Cette étape de restructuration et de nettoyage des jeux de données est primordiale pour assurer l'efficacité du processus d'apprentissage.

2.2. Choix des prédicteurs

Afin de sélectionner les meilleurs prédicteurs, les coefficients de corrélation entre chaque prédicteur potentiel et θ ont été calculés. La figure 2 montre le poids de corrélation normalisé de chacun de ces paramètres par rapport à θ. Le poids de la masse volumique des grains ρ_s apparaît insignifiant pour la prédiction, ce paramètre a donc été écarté.

Finalement, pour prédire θ, 5 prédicteurs semblent être pertinents : ψ, n, $P_{2\mu}$, $P_{50\mu}$ et $P_{500\mu}$. Le tableau 1 synthétise les statistiques descriptives des 6 paramètres retenus pour les deux jeux de données (séchage et humidification). Le nombre de prédicteurs retenus pour prédire la teneur en eau volumique est de 5 paramètres contre 8 paramètres présents dans le jeu initial qui étaient des prédicteurs potentiels. Ceci est dû à l'importance de la part de données manquantes pour la conductivité hydraulique saturée, la teneur en eau volumique saturée et la masse volumique apparente, et au faible poids de corrélation avec la teneur en eau volumique, pour la masse volumique des grains.

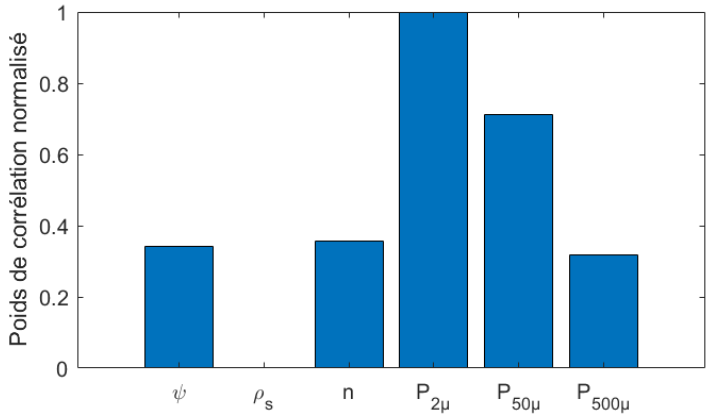

Figure 2. Poids de corrélation normalisé des différents paramètres par rapport à la teneur en eau volumique θ.

Tableau 1. Statistiques descriptives des paramètres retenus pour les jeux de données en séchage et en humidification

Paramètre (unité)	Chemin	Taille	Valeurs manquantes	Minimum	Maximum	Médiane	Moyenne	Ecart-type
ψ (kPa)	Séchage	1551	0	0	6,0 x10^3	10	107,9	400,2
	Humidification	295	0	0	1,5 x10^3	4,2	77,1	306,9
θ	Séchage	1551	0	0,02	0,58	0,29	0,27	0,12
	Humidification	295	0	0,03	0,71	0,25	0,26	0,14
n	Séchage	1551	1062	0,18	0,62	0,44	0,43	0,06
	Humidification	295	92	0,18	0,72	0,46	0,43	0,09
$P_{2µ}$	Séchage	1551	0	0	0,58	0,06	0,12	0,14
	Humidification	295	0	0	0,62	0,04	0,13	0,18
$P_{50µ}$	Séchage	1551	0	0	0,76	0,07	0,17	0,21
	Humidification	295	0	0	0,96	0	0,16	0,32
$P_{500µ}$	Séchage	1551	0	0	1	0	0,34	0,41
	Humidification	295	0	0	1	0	0,22	0,39

3. Présentation des méthodes de régression utilisées

Dans ce paragraphe, les algorithmes de machine learning utilisés sont présentés de manière très concise, Ces algorithmes sont proposés dans Matlab® et plus de détails sur les méthodes pourront être consultés notamment dans (Mathworks, 2019 ; Kotu et Deshpande, 2015).

3,1, Arbre de régression (RT)

L'arbre de régression est une méthode adaptée de celle de l'arbre de décision utilisé pour les tâches de classification. Le principe est de trouver des valeurs seuils permettant de séparer les données en sous-intervalles, d'attribuer des poids à ces valeurs seuils et d'ajuster ces poids afin de minimiser l'erreur de prédiction sur les données d'apprentissage. Chaque valeur seuil représente une branche de l'arbre de régression. Cet algorithme présente les avantages d'une convergence rapide et d'une interprétabilité facile des résultats, Il présente toutefois un risque important de surapprentissage (*overfitting*). Pour cela, les données seront divisées en 5 groupes permettant une validation croisée pendant l'apprentissage.

3.2. Machine à support vecteur (SVM)

Il s'agit également d'une méthode de classification qui a été adaptée pour la régression. Elle cherche à définir des régions dans l'espace des paramètres. Les limites entre ces régions sont appelées des hyperplans (ce sont des droites s'il n'y a que 2 paramètres et des plans s'il y en a 3). Les équations de ces hyperplans sont ajustées de manière itérative en utilisant les données. La méthode nécessite en général, un temps de calcul moyen et présente une grande flexibilité qui permet de résoudre des problèmes difficiles et de bien résister au surapprentissage. L'interprétation des résultats devient quasi-impossible si le nombre de prédicteurs est élevé et si les équations des hyperplans (on parle de fonctions noyau ou kernel) sont non linéaires. Dans cette étude, une fonction noyau quadratique a été utilisée et les données ont été divisées en 5 groupes pour permettre une validation croisée de l'apprentissage.

3.3. Réseau de neurones artificiels (ANN)

Les réseaux de neurones artificiels s'inspirent du fonctionnement des neurones biologiques. Chaque neurone est un élément avec une ou plusieurs valeurs d'entrée (*input*) et une valeur de sortie (*output*). Une fonction d'activation (sigmoïde en général) permet de calculer la valeur de sortie à partir d'une combinaison pondérée des valeurs d'entrée. Les neurones d'un réseau sont organisés en une couche d'entrée formée par autant de neurones que de prédicteurs, une couche de sortie avec un nombre de neurones égal au nombre de cibles, et une ou plusieurs couche(s) cachée(s) de neurones intermédiaires permettant de s'adapter aux problèmes difficiles. Les coefficients de pondération des différents neurones sont initialisés et puis ajustées progressivement en utilisant les données d'apprentissage par rétropropagation des gradients. Le jeu de données est généralement partitionné de manière aléatoire en 3 sous-ensembles pour l'apprentissage, la validation et le test. Cette méthode s'est montrée très performante pour les problèmes non linéaires complexes. Dans cette étude, une couche cachée avec 10 neurones a été sélectionnée, 70% des données ont été réservés à l'apprentissage, 15% à la validation et 15% aux tests.

4. Résultats

Les trois algorithmes sélectionnés ont été appliqués aux deux jeux de données (séchage et humidification), le but étant de prédire θ à partir de ψ, n, $P_{2\mu}$, $P_{50\mu}$ et $P_{500\mu}$.

La figure 3 montre les comparaisons entre les valeurs de la teneur en eau volumique prédites après entrainement, et les valeurs connues, pour la totalité des données disponibles en séchage. Les trois méthodes permettent des prédictions acceptables avec une erreur quadratique moyenne RMSE < 0,07. L'algorithme ANN donne la meilleure performance de prédiction avec RMSE ≈ 0,025.

La figure 4 montre les comparaisons entre les valeurs de la teneur en eau volumique prédites après entrainement, et les valeurs connues, pour la totalité des données disponibles en humidification. L'algorithme ANN s'avère encore ici, être le plus performant et les résultats montrent une performance moins satisfaisante des prédictions que pour les données de séchage, même si l'erreur quadratique moyenne reste comparable. Cette différence s'explique par une quantité de données moindre mais certainement aussi, par une qualité inférieure des mesures obtenues en humidification. En effet, les essais transitoires en humidification sont plus délicats à réaliser à cause notamment des problèmes de piégeage de l'air et les mesures sont alors généralement moins précises.

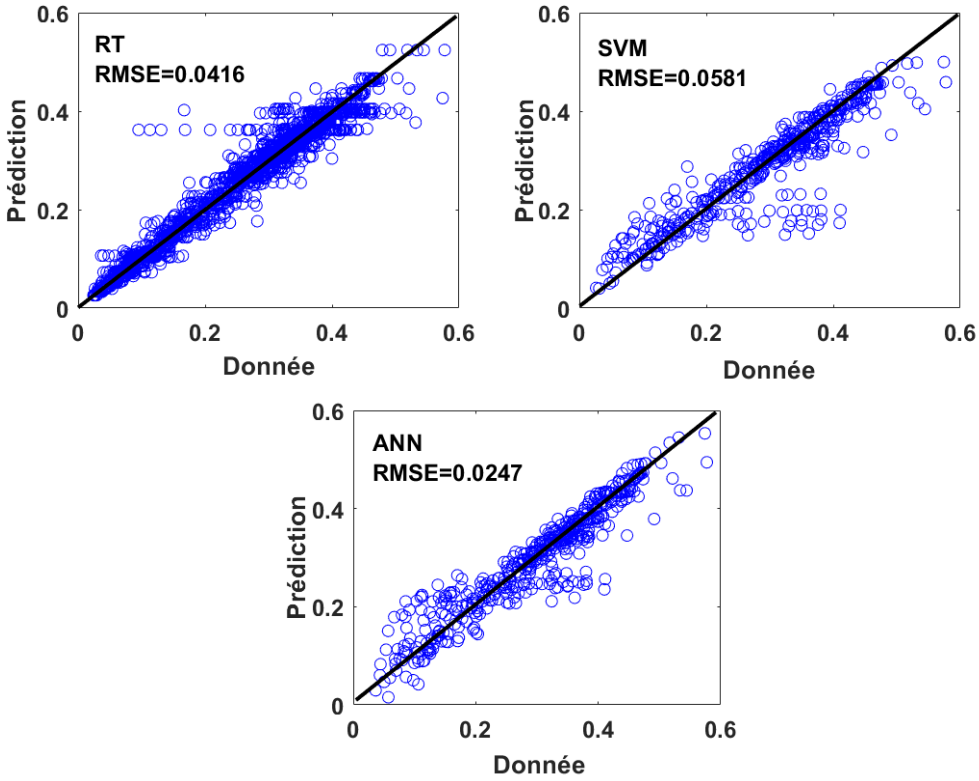

Figure 3. Comparaison entre les prédictions et les données de teneur en eau volumique en séchage avec les trois algorithmes utilisés.

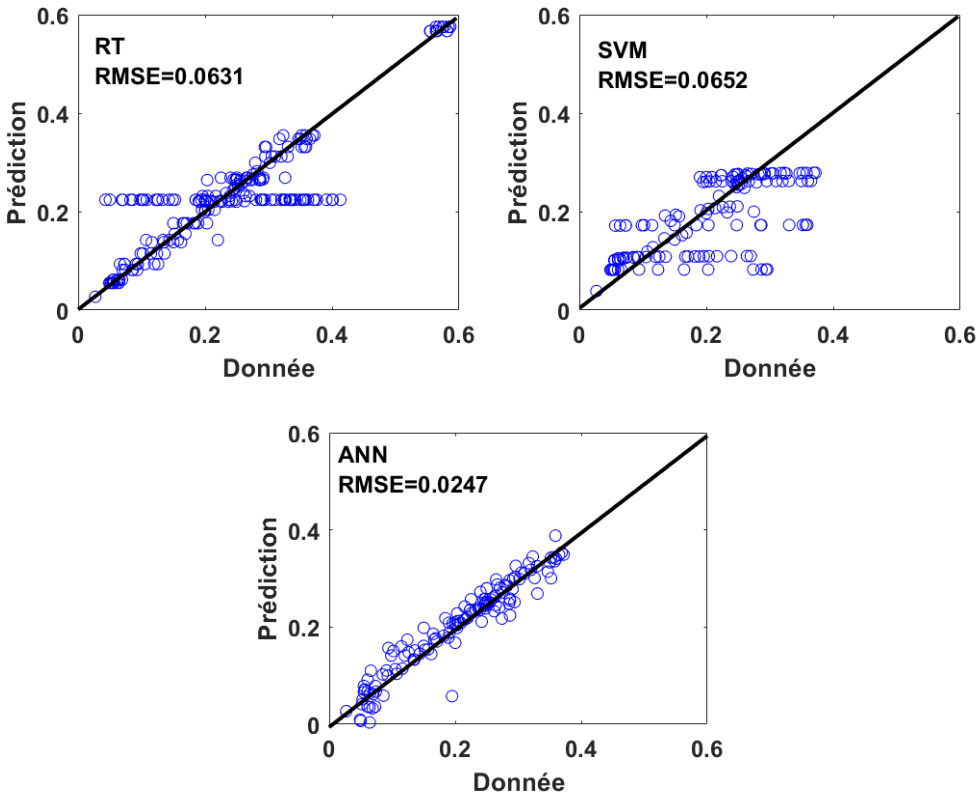

Figure 4. Comparaison entre les prédictions et les données de teneur en eau volumique en humidification avec les trois algorithmes utilisés.

5. Conclusion

Dans cette communication, deux jeux des données extraites de la base de données UNSODA (Leij *et al.*, 1996) ont été réorganisés et nettoyés, pour pouvoir être utilisés pour entrainer trois modèles de machine learning (RT, SVM et ANN) afin de prédire la courbe de rétention de l'humidité des sols. La restructuration et la suppression des données manquantes ont considérablement réduit la taille des jeux qui sont passés de plus 65 000 (avec 9 paramètres) à 1 551 et 295 enregistrements (avec 6 paramètres) respectivement pour les données des essais de séchage et de d'humidification. Cette étape est néanmoins, indispensable pour permettre des prédictions fiables.

La comparaison des prédictions de la teneur en eau volumique avec les valeurs mesurées a montré globalement des bonnes performances des régressions avec une erreur quadratique moyenne inférieure à 0,07 pour les deux jeux de données (séchage et humidification). La précision des prédictions obtenues sur les données de séchage s'est montrée toutefois globalement supérieure à celle obtenue sur les données d'humidification. Cela peut s'expliquer par la plus grande quantité mais aussi la meilleure qualité des données obtenues lors des essais en séchage. Le réseau de neurones artificiels a fourni les résultats les plus satisfaisants avec une erreur quadratique moyenne de 0,025 dans les deux cas.

Ces résultats prometteurs devront toutefois être validés en utilisant des données indépendantes de la base de données qui a servi à l'apprentissage des modèles. Les données *in situ* présentes dans la base de données, devront également être ajoutées pour vérifier les prédictions. Il serait également intéressant de tester la prédiction des courbes de rétention au moyen des paramètres de modèles théoriques.

En ce qui concerne l'utilisation pratique des techniques de machine learning, les résultats comparés des modèles entrainés sur les données de séchage et d'humidification (de taille et de qualité différentes), montrent bien l'effet important de la taille de la base de données mais également celui de la qualité des mesures disponibles. Avec ces techniques, il n'est pas possible de définir a priori une taille nécessaire de la base de données ni de présélectionner un algorithme de choix, le résultat obtenu dépend de la combinaison de la quantité et de la qualité des données d'une part, et de la complexité du phénomène physique à prédire d'autre part.

6. Références bibliographiques

Arya, L. M. et Paris J. F. (1981). A Physicoempirical model to predict the soil moisture charac-teristic from particle-size distribution and bulk density data. Soil Science Society of America Journal, 45, 1023-1030.

D'Emiliano A., Aiello R., Consoli S. et Vanella D. (2018). Artificial neural networks for predict-ing the water retention curve of Sicilian agricultural soils. Water: 10, 1431, doi:10.3390/w10101431.

Fredlund M. D., Wilson G. W. et Fredlund D. G. (2002). Use of the grain-size distribution for es-timation of the soil-water characteristic curve. Can. Getech. J. 39: 1103-1117, doi: 10.1139/T02-049.

Fredlund, D. G., Rahardjo H. et Fredlund M. D. (2012). Unsaturated soil mechanics in Engineering Practice. John Wiley & Sons. Online ISBN: 9781118280492. doi: 10.1002/9781118280492.

Kotu V. et Deshpande B. (2015). Predictive analytics and data mining. Elsevier, Imprint Morgan Kaufmann, 446 pages, ISBN: 978-0-12-8014608, https://doi.org/10.1016/C2014-0-00329-2.

Lamorski K., Simunek J., Slawunski C. and Lamorska J. (2017). An estimation of the main wetting branch of the soil water retention curve based on its main drying branch using the machine learning method. Water Ressour. Res.: 53, 1539-1552, doi:10.1002/2016WR019533.

Leij F. J., Alves W. J., van Genuchten Th. M. and Williams J. R. (1996). The Unsoda unsaturated soil hydraulic database user's manual, version 1.0. USEPA EPA/600/R-96/095.

Mathworks (2019). Statistics and machine learning toolbox™ user's guide, version R2019a.

Mualem, Y. (1976). A new model for predicting the hydraulic conductivity of unsaturated porous media, Water Resour. Res., 12(3), 513–522, doi:10.1029/WR012i003p00513.

Nguyen P. M., Haghverdi A., De Pue J., Botula Y.-D., Le K. V., Waegeman W. and Cornelis W. M. (2017). Comparison of statistical regression and data-mining techniques in estimating soil water retention of tropical delta soils. Biosystems Engineering 153: 12-27.

Patil N. G. and Singh PS. K. (2016). Pedotransfer functions for estimating soil hydraulic proper-ties: a review. Pedoshere 26, 417-430.

Van Genuchten M. Th. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. Soil Sci. Soc. Am. J., 44, 892-898.