Estimation des incertitudes liées à la prédiction ponctuelle de variables pédologiques à partir de bases de données géographiques sur les sols.

Violaine Murciano, Jean-Baptiste J.-B. Paroissien, Nicolas N. Saby, Anne C Richer-De-Forges, Manuel Pascal Martin, R. Emilion, Dominique D. Arrouays

Estimation des incertitudes liées à la prédiction ponctuelle de variables pédologiques à partir de bases de données géographiques sur les sols

Exemple de l'utilisation des strates issues du programme français Inventaire, Gestion et Conservation des Sols (IGCS)

V. Murciano(1,2), J-B. Paroissien(1), N.P.A. Saby(1), A.C. Richer de Forges(1), M.P. Martin(1), R. Emilion(2) et D. Arrouays(1)*

1) INRA, Unité InfoSol, US1106, F-45075 Orléans, France
2) Université d'Orléans et CNRS, UMR 7349, France

* : Auteur correspondant : Dominique.Arrouays@orleans.inra.fr

RÉSUMÉ

Le programme *GlobalSoilMap* a pour objectif de produire une base de données à haute résolution spatiale des propriétés des sols du monde, assorties de leurs incertitudes. Parmi les méthodes possibles pour atteindre cet objectif, l'une d'elles consiste en une prédiction de ces propriétés à partir de moyennes pondérées et d'estimations des intervalles de confiance issues de bases de données cartographiques (unités cartographiques et typologiques de sol). En France, ces bases décrivent en particulier les « strates », qui sont des horizons conceptuels caractérisés à partir des valeurs modales et extrêmes d'un certain nombre de propriétés. Cependant, dans de nombreux cas, ces valeurs extrêmes n'ont pas été renseignées dans les bases de données. Notre objectif est donc de tester la possibilité d'estimer ces valeurs extrêmes à partir des valeurs modales et d'autres variables environnementales en utilisant des techniques d'apprentissage automatique. Les données utilisées proviennent de l'extraction de DoneSol des strates contenant les valeurs modales et les valeurs extrêmes de certaines propriétés. Nous illustrons ici la démarche par des résultats primaires concernant 7 variables pédologiques requises pour les « produits » *GlobalSoilMap* (carbone organique, pH_{eau}, teneurs en argile, limon, sable, éléments grossiers, et capacité d'échange cationique) puis nous détaillons deux exemples portant sur le carbone organique et le pH. Les qualités de prédiction les plus satisfaisantes concernent le carbone. L'intérêt principal de ce type d'approche est de pouvoir dériver des valeurs par défaut de ces indicateurs de dispersion lorsqu'elles sont manquantes dans les bases de données. On peut toutefois penser que les types de modèles que nous avons utilisés pourraient parfois conduire à un « sur-ajustement » qui donne une fausse idée de leur performance. Pour vérifier cela, il faudrait disposer d'une validation externe entièrement indépendante.

Mots clés

GlobalSoilMap, cartographie numérique des sols, incertitude, apprentissage automatique, fouille de données, France.

Reçu : août 2014 ; Accepté : octobre 2014

SUMMARY

ESTIMATING UNCERTAINTIES LINKED TO POINT PREDICTION OF SOIL VARIATES USING SOIL GEOGRAPHICAL DATABASES. An example on the French soil inventory and mapping programme.

The GlobalSoilMap *project aims to produce a digital soil map of the world. The ultimate objective of the project is to build a free high resolution (100x100-m) downloadable database of key soil properties at multiple depths, mostly using existing soil information and environmental covariates. The soil properties will be delivered as predictions with uncertainty at specified depths. Among the possible methods, one is to use data from geographical databases, and to apply weighed averages on the predicted properties. However, in numerous cases, only mean or modal values are available and therefore it is impossible to estimate the range of possible values within a 100x100-m area. Our aim is to test the feasibility of estimating the range of possible values using this information, additional ancillary co-variates and various regression and machine learning tools. We show that for some properties (i.e. carbon content) the quality of the prediction is quite good. The main interest of these methods is to provide default values of the range of soil properties when they are missing in the databases. However, to check if there is no over-fitting, further studies should be conducted.*

Key-words

GlobalSoilMap*, digital soil mapping, uncertainty, machine learning, data mining, France.*

RESUMEN

ESTIMACIÓN DE LAS INCERTIDUMBRES LIGADAS A LA PREDICCIÓN PUNTUAL DE VARIABLES PEDOLÓGICAS A PARTIR DE BASES DE DATOS GEOGRÁFICAS SOBRE LOS SUELOS.

Ejemplo del uso de los estratos derivados del programa francés Inventario, Gestión y Conservación de los suelos (IGCS)

El programa GlobalSoilMap tiene por objetivo producir una base de datos de alta resolución espacial de las propiedades de los suelos del mundo, acompañadas de sus incertidumbres. Entre los métodos posibles para alcanzar este objetivo, uno consiste en una predicción de estas propiedades a partir de medias ponderadas y de estimación de los intervalos de confianza derivadas de bases de datos cartográficos (unidades cartográficas y tipológicas de suelos). En Francia, estas bases describen en particular los "estratos", que son horizontes conceptuales caracterizados a partir de valores modales y extremas de algunos números de propiedades. Sin embargo, en numerosos casos, estos valores extremos no fueron indicados en las bases de datos. Nuestro objetivo es por lo tanto probar la posibilidad de estimar estos valores extremos a partir de los valores modales y de otras variables medioambientales usando técnicas de aprendizaje automático. Los datos utilizados provienen de la extracción de DoneSol de los estratos que contiene los valores modales y los valores extremos de ciertas propiedades. Ilustramos aquí el enfoque por resultados primarios que concierne 7 variables pedológicas requeridas para los "productos" GlobalSoilMap (carbono orgánico, pHagua , contenidos en arcilla, limo, arena, elementos gruesos, y capacidad de intercambio catiónico) luego detallamos dos ejemplos sobre el carbono orgánico y el pH. Las calidades de predicción las más satisfactorias conciernen el carbono. El interés principal de este tipo de enfoque es poder derivar valores por defecto de estos indicadores de dispersión cuando faltan en las bases de datos. Se puedo, no obstante pensar que los tipos de modelos que utilizamos podrían a veces llegar a un "sobre – ajuste" que da una falsa idea de su eficiencia. Para averiguarlo, sería necesario disponer de una validación externa enteramente independiente.

Palabras clave

Cartografía numérica de suelos, incertidumbre, aprendizaje automático, búsqueda de datos, Francia.

La connaissance et la protection des sols sont reconnues comme étant des piliers majeurs pour répondre aux grands enjeux planétaires tels que la sécurité alimentaire, le changement climatique, l'accaparement des terres, l'urbanisation et la gestion de l'eau. Il est donc indispensable de mettre en place des outils permettant de prendre en compte les propriétés des sols à l'échelle mondiale. Toutefois, de nombreuses régions du monde manquent encore cruellement de données sur les sols. Dans ce contexte, la communauté scientifique mondiale propose de mettre en place le programme *GlobalSoilMap* dont l'objectif est de produire une base de données à haute résolution spatiale des propriétés des sols du monde, assorties de leurs incertitudes (Sanchez *et al*., 2009 ; Hempel *et al*., 2013 ; Arrouays *et al*., 2014a et b).

Plusieurs méthodes ont été proposées pour atteindre cet objectif (Minasny et McBratney, 2010). Parmi ces méthodes, l'une d'elles consiste en une prédiction de ces propriétés à partir de moyennes pondérées et d'estimations des intervalles de confiances issues de données cartographiques (unités cartographiques et typologiques de sol). Cette solution pourrait être envisagée pour la France, à partir des bases de données géographiques du programme IGCS (Inventaire Gestion et Conservation des Sols ; Laroche *et al*., 2014) stockées dans la base de données nationale des informations pédologiques (DoneSol) (Grolleau *et al*., 2004). Ces bases décrivent en particulier les « strates » (voir par exemple, Richer de Forges *et al*., 2014), qui sont des horizons conceptuels décrits à partir des valeurs modales et extrêmes d'un certain nombre de propriétés. Ces valeurs extrêmes peuvent être rapprochées de celles d'un intervalle de confiance à 95 % comme récemment montré par Nauman et Thompson (2014) et Libohova *et al.,* (2014) aux Etats-Unis.

Il reste que dans de nombreux cas, ces valeurs extrêmes n'ont pas été renseignées dans les bases de données. Notre objectif est donc de tester la possibilité d'estimer ces valeurs extrêmes à partir des valeurs modales et d'autres variables environnementales.

MATÉRIEL ET MÉTHODES

Données

Les données utilisées proviennent de l'extraction de DoneSol des strates contenant les valeurs modales et les valeurs extrêmes de certaines propriétés. Nous illustrons ici la démarche par des résultats primaires concernant 7 variables pédologiques requises pour les « produits » *GlobalSoilMap* (carbone organique, pH_{eau}, teneurs en argile, limon, sable, éléments grossiers, et capacité d'échange cationique) puis nous détaillons deux exemples portant sur le carbone organique et le pH. Le nombre de strates est d'environ 13000 pour le carbone, 30000 pour la CEC et compris entre 37000 et 42000 pour les autres variables. La couverture du territoire métropolitain est relativement clusterisée, liée à l'état d'avancement des programmes (Laroche *et al*., 2014 ; Richer de Forges *et al*., 2014). Elle couvre néanmoins assez bien la diversité des situations pédologiques (Laroche *et al*., 2014 ; Richer de Forges *et al*., 2014).

Nous avons renseigné les strates par les variables environnementales suivantes :
- altitude maximale (m),
- altitude minimale (m),
- superficie de l'UCS (ha),
- catégorie de matériel parental regroupé en 13 catégories,
- forme morphologique regroupée en 20 catégories,
- profondeur minimale d'apparition (cm),
- profondeur moyenne d'apparition (cm),
- profondeur maximale d'apparition (cm),
- épaisseur minimale (cm),
- épaisseur moyenne (cm),
- épaisseur maximale (cm),
- valeur modale (dans l'unité de la variable pédologique considérée),
- altitude moyenne (m),
- altitude maximale moins altitude minimale (m),
- profondeur totale du sol (cm),
- superficie de l'UTS (ha).

Traitement

Pour chaque strate, nous avons calculé les indicateurs de dispersion suivants :
- valeur minimale (val_min),
- valeur maximale (val_max),
- étendue (val_max-val_min).

Les modèles de fouille de données

Nous avons ensuite calibré des modèles de prédiction de ces indicateurs à partir des valeurs modales et des variables environnementales. Nous avons utilisé 3 méthodes dérivées de techniques de régression et d'apprentissage automatique :

Random Forest (RF) :

Cet algorithme, Random Forest (ou Forêts aléatoires), est une amélioration de l'algorithme bagging, proposé par Breiman (2001). Cette amélioration est l'ajout d'une « randomisation ». Il a pour objectif de rendre plus indépendants les arbres de l'agrégation en ajoutant du hasard dans le choix des variables qui interviennent dans les modèles.

Cette méthode semble plus efficace lorsque le nombre de variables explicatives p est très important. En effet, la variance de la moyenne de N variables indépendantes et identiquement distribuées, chacune de variance σ^2, est σ^2/N**.** Si les variables sont identiquement distribuées mais

avec une corrélation ρ des variables prises deux à deux, la variance de la moyenne devient : (1)

$$\rho\sigma^2 + \frac{1-\rho}{N}\sigma^2 .$$

Comme pour le premier cas, le deuxième terme décroît avec N**.** Par défaut, c'est le nombre minimum d'observation par nœuds qui limite la taille de l'arbre, il est fixé à 5 par défaut. Ce sont donc des arbres plutôt complets qui sont considérés, chacun de faible biais mais de variance importante.

La sélection aléatoire d'un nombre réduit de m prédicteurs potentiels à chaque étape de construction d'un arbre, augmente significativement la variabilité en mettant en avant nécessairement d'autres variables. Chaque modèle de base est évidemment moins performant, mais la réunion de tous conduit finalement à de bons résultats.

L'évaluation itérative de l'erreur *out-of-bag* permet de contrôler de nombre N d'arbres de la forêt et éventuellement d'optimiser le choix de m.

GBM (Generalized Boosted Regression Models)

L'algorithme GBM a été mis en œuvre par Friedman (2002) tout d'abord sous le nom Multiple Additive Regression Trees (MART) puis sous celui de GBM. Ce modèle fait partie d'une famille d'algorithme basés sur une fonction perte supposée différentiable et notée Q. Le principe de cet algorithme est de construire une séquence de modèles afin qu'à chaque étape, chaque modèle ajouté à la combinaison apparaisse comme un pas vers une meilleure solution. L'innovation apportée concerne le pas effectué à chaque étape, qui est franchi dans la direction du gradient de la fonction de perte lui-même approché par un arbre de régression.

Cet algorithme permet de réduire la variance ainsi que le biais. Il donne généralement de meilleurs résultats.

Cubist

Cubist est un modèle de régression axé prédiction qui combine les idées de modèle à base de règles, décrites dans Quinlan (1992) avec des corrélations supplémentaires basées sur les voisins les plus proches dans l'ensemble d'apprentissage (voir Quinlan (1993) pour plus de détails).

Initialement, une arborescence est créée, mais chaque chemin de l'arbre termine par une règle. Un modèle de régression est adapté pour chaque règle basée sur un sous-ensemble de données défini par les règles. L'ensemble des règles est élagué ou éventuellement combiné. Les variables candidates pour les modèles de régression linéaire sont les prédicteurs qui ont été utilisés dans la partie de la règle qui a été élaguée.

Cubist généralise ce modèle pour ajouter du « boosting » (lorsque le paramètre committees est supérieur à 1) et des corrélations à base d'instances. Le nombre d'instances est fixé au temps de la prédiction par l'utilisateur et n'est pas nécessaire pour la construction du modèle.

L'ensemble de ces algorithmes a été mis en œuvre avec le logiciel R (R Core Team, 2014) et les paquets R suivants : randomForest (Liaw et Wiener., 2002), gbm (Ridgeway *et al*., 2013) et Cubist (Kuhn *et al*., 2014)

Construction des modèles et évaluation de leur qualité prédictive

On effectue la construction des modèles en réalisant une validation croisée avec la méthode k-fold. On fait une validation croisée pour avoir une gamme de valeurs la plus homogène possible.

La méthode k-fold consiste à diviser le jeu de données en k échantillons, on sélectionne ensuite un des k échantillons comme ensemble de validation et les k-1 autres échantillons constituent l'ensemble d'apprentissage. On répète l'opération k fois de façon à ce que chaque échantillon serve une fois pour la validation.

L'algorithme cvFolds du paquet R cvTools (Alfons., 2012) est utilisé pour la création des k échantillons. Le pourcentage p du jeu de données servant comme échantillon de validation est fixé à 0.2 et permet de calculer le nombre de blocs nb (d'échantillons) à partir de p :

$$nb = (\text{taille du jeu de données})/\text{taille du jeu de données} \times p) \quad (2)$$

On construit ensuite les k blocs de façon aléatoire.

Afin de pouvoir juger si un modèle est performant ou bien de pouvoir comparer des modèles entre eux, on utilise des indicateurs de qualité. Chacun de ces indicateurs se calcule en fonction des données observées x et des valeurs prédites y.

Nous avons calculé les indicateurs suivants :

- Le coefficient de détermination R^2 est compris entre 0 et 1. Plus il est proche de 1, plus le modèle est de bonne qualité. Il se calcule de la manière suivante :

$$R^2(x,y) = cor(x,y)^2 \quad (3)$$

- L'erreur moyenne de prédiction (EMP, *en anglais: MPE*), mesure le biais de prédiction et doit être proche de 0. Il se calcule de la manière suivante :

$$EMP = \frac{1}{n}\sum_{i=1}^{n}(x-y)^2 \quad (4)$$

- La racine carrée de l'erreur quadratique moyenne (REQM, *en anglais : RMSE*) est utilisée pour mesurer la précision du modèle, plus il est faible, plus le modèle est de bonne qualité. Il se calcule de la manière suivante :

$$REQM = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x-y)^2} \quad (5)$$

En utilisant ces trois modèles, notre objectif principal n'est toutefois pas de comparer leurs performances, mais plutôt de vérifier la cohérence et la robustesse de nos résultats.

Pour chacun des modèles, l'importance des covariables utilisées pour la construction de celui-ci a été calculée. Cette importance des variables est disponible en sortie des modèles mais elle est calculée de manière différente pour les modèles Random Forest, GBM ou Cubist. Elle est estimée de la manière suivante :

• Pour *Random Forest*, si un prédicteur est important dans le modèle actuel, alors en attribuant d'autres valeurs à ce prédicteur au hasard mais de manière réaliste (par exemple en permutant les valeurs du prédicteur sur l'ensemble du jeu de données, cela devrait avoir une influence négative sur la prédiction. On calcule donc le MSE (= $RMSE^2$) avec l'ensemble de données d'origine puis avec l'ensemble de données permuté et on fait la différence. Ainsi plus la différence est élevée, plus le prédicteur est important.

• Pour *GBM*, pour chaque arbre de décision T, on mesure la pertinence pour chaque variable prédictive X_i de la manière suivante :

$$(6)\quad I_l^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 I(v(t) = l)$$

La somme porte sur les *J-1* nœuds internes de l'arbre. A chaque nœud t, l'une des variables d'entrée $X_{v(t)}$ est utilisée pour partitionner la région associée à ce nœud en deux sous-régions ; à l'intérieur de chacune une constante est ajustée pour séparer les valeurs de réponse. La variable particulière choisie est celle qui donne l'amélioration estimée maximale i^2_t dans le risque de l'erreur quadratique (MSE). L'importance relative quadratique de la variable X_l est la somme de ces améliorations au carré sur tous les nœuds internes pour lequel il a été choisi comme variable de séparation.

Cette mesure de l'importance est facilement généralisable à une combinaison d'arbres ; c'est tout simplement la moyenne sur les arbres :

$$(7)\quad I_l^2 = \frac{1}{M}\sum_{m=1}^{M} I_l^2(T_m)$$

Les importances réelles sont les racines carrées des importances obtenues. Étant donné que ces mesures sont relatives, il est d'usage d'attribuer au plus important une valeur de 100, puis mettre à l'échelle les autres en conséquence.

• Pour *Cubist*, le pourcentage de fois où chaque variable a été utilisée est retourné.

Les quatre variables les plus importantes ont été étudiées et leur influence sur le modèle final a été interprétée et analysée graphiquement à travers la mise en relation entre les données observées et les données prédites par la covariable. Pour produire ces graphiques on crée tout d'abord une grille contenant toutes les combinaisons des valeurs (déciles s'il s'agit d'une variable quantitative) des quatre variables les plus importantes. On ajoute ensuite à la grille les autres variables en leur attribuant la valeur médiane. On effectue une prédiction en fonction de cette grille puis on calcule la valeur médiane pour les prédictions. Enfin, on trace les graphiques des valeurs prédites en fonction de chacune des quatre variables.

RÉSULTATS

Le *tableau 1* présente les résultats pour les 3 modèles et les 7 variables.

On constate pour chaque indicateur de dispersion (val_min, val_max et étendue) des variables une cohérence entre les 3 méthodes. On remarque également que les R^2 sont très élevés et que les valeurs des RMSE apparaissent très faibles. Les indicateurs les mieux prédits sont les valeurs minimales et les valeurs maximales. La variable la mieux prédite est le carbone.

A titre d'exemple nous montrons ici les graphiques valeurs prédites, valeurs observées pour deux variables prédites et 3 modèles.

Pour le carbone, on observe une relation linéaire et sans biais. Pour l'étendue du pH, la relation est moins satisfaisante et on observe un biais systématique, avec une sous-estimation des valeurs fortes et une surestimation des valeurs faibles.

L'importance relative des variables est présentée dans le *tableau 2*.

L'ordre d'importance des variables est globalement cohérent entre les 3 méthodes. La variable la plus importante est dans presque tous les cas la valeur modale du paramètre ; on constate deux exceptions, pour l'étendue de l'abondance des éléments grossiers avec la méthode GBM, et pour l'étendue de la teneur en limons avec cette même méthode. Toutefois, dans ces deux cas, la valeur modale intervient en deuxième position. Les autres variables intervenant le plus souvent concernent la classe de matériau parental, la forme morphologique et l'épaisseur maximale de la strate.

Nous présentons ici deux résultats pour illustrer l'analyse de l'effet des variables explicatives.

L'effet de la valeur modale de la teneur en carbone des strates sur leur valeur minimale est monotone et croissant et assez proche d'une relation linéaire. En d'autres termes, il semble qu'il y ait une relation de proportionnalité entre la valeur modale et la valeur minimale, proportionnalité qui peut être modulée en fonction d'autres co-variables, qui ont cependant un poids beaucoup plus faible.

En revanche, l'effet de la valeur modale du pH des strates montre une courbe en cloche, ce qui signifie que plus les pH sont proches de valeurs extrêmes de la gamme rencontrée, moins leur variabilité au sein de la strate est forte.

Tableau 1 - Indicateurs de performance de prédiction des variables cible pour les 3 modèles. Les chiffres en gras représentent le meilleur indicateur.

Table 1 - *Indicators of performance for the prediction of the variates of interest and for the three models. The best indicators are in bold.*

		Val_min			Val_max			Etendue		
		R^2	MPE	RMSE	R^2	MPE	RMSE	R^2	MPE	RMSE
Carbone g/kg	RF	0,9939	0,0919	3,09	0,9903	**-0,0227**	**7,97**	**0,9581**	**-0,0420**	**9,36**
	GBM	**0,9973**	-0,0200	**2,35**	**0,9907**	-0,1738	8,12	0,9558	-0,2691	9,56
	Cubist	0,9971	**0,0117**	2,53	0,9886	-0,0332	8,90	0,9548	-0,0723	9,60
pH	RF	**0,9684**	0,0155	**0,24**	**0,9636**	0,0033	**0,21**	**0,8260**	0,0020	0,36
	GBM	0,9655	0,0012	0,25	0,9631	-0,0039	0,21	0,8254	-0,0054	**0,35**
	Cubist	0,9656	**-0,0008**	0,25	0,9625	**0,0006**	0,21	0,8184	**-0,0005**	0,36
Argile g/kg	RF	**0,9492**	1,753	**24,80**	**0,9478**	**0,1631**	**34,28**	**0,8079**	0,4719	47,89
	GBM	0,9443	**-0,1868**	25,76	0,9458	-0,5981	34,89	0,8065	-0,7481	**47,61**
	Cubist	0,9448	-0,2141	25,64	0,9436	0,5002	35,58	0,7949	**0,3341**	49,06
Limons g/kg	RF	0,9410	2,883	38,47	0,9542	0,6118	39,90	0,8304	-0,3052	62,59
	GBM	**0,9413**	**0,06711**	**38,09**	**0,9560**	-0,7256	**38,83**	**0,8443**	-1,1850	**59,68**
	Cubist	0,9377	-0,3839	39,24	0,9499	**0,3632**	41,39	0,8140	**0,1955**	65,19
Sable g/kg	RF	0,9598	2,909	41,57	**0,9602**	**0,0084**	**45,54**	0,8261	**-0,0890**	71,29
	GBM	**0,9609**	-0,1287	**40,74**	0,9595	-0,8220	45,87	**0,8385**	-0,9295	**68,03**
	Cubist	0,9568	**-0,0733**	42,77	0,9565	-0,0685	47,53	0,8154	-0,6928	72,72
Abondance EG %	RF	0,9724	0,1442	2,89	**0,9633**	**0,0177**	**5,13**	0,9133	**0,0442**	6,62
	GBM	0,9745	**-0,0324**	2,76	0,9560	-0,1614	5,59	0,9057	-0,1566	6,86
	Cubist	**0,9758**	-0,0417	**2,69**	0,9609	-0,0911	5,28	**0,9138**	-0,1410	**6,56**
CEC cmol/kg	RF	**0,9610**	0,0745	**1,23**	**0,9591**	**0,0023**	**2,02**	0,8410	**0,0102**	2,75
	GBM	0,9590	-0,0048	1,25	0,9451	-0,0189	2,33	0,8026	-0,0242	3,06
	Cubist	0,9572	**-2,028e-03**	1,28	0,9565	-0,0054	2,08	**0,8414**	-0,0256	**2,75**

DISCUSSION

Nous discutons ici successivement de la convergence et de la divergence des modèles, de leur interprétation et de leur performance et intérêt.

Convergence et divergence des modèles

Les modèles donnent des résultats qui sont globalement convergents. En particulier le fait que la valeur modale soit toujours la première en termes d'importance est cohérent et *a priori* logique. En revanche on constate quelques différences dans l'importance des variables présentes dans les positions immédiatement suivantes. Pour l'étendue du pH, un effet des matériaux parentaux était attendu, ce qui fut le cas.

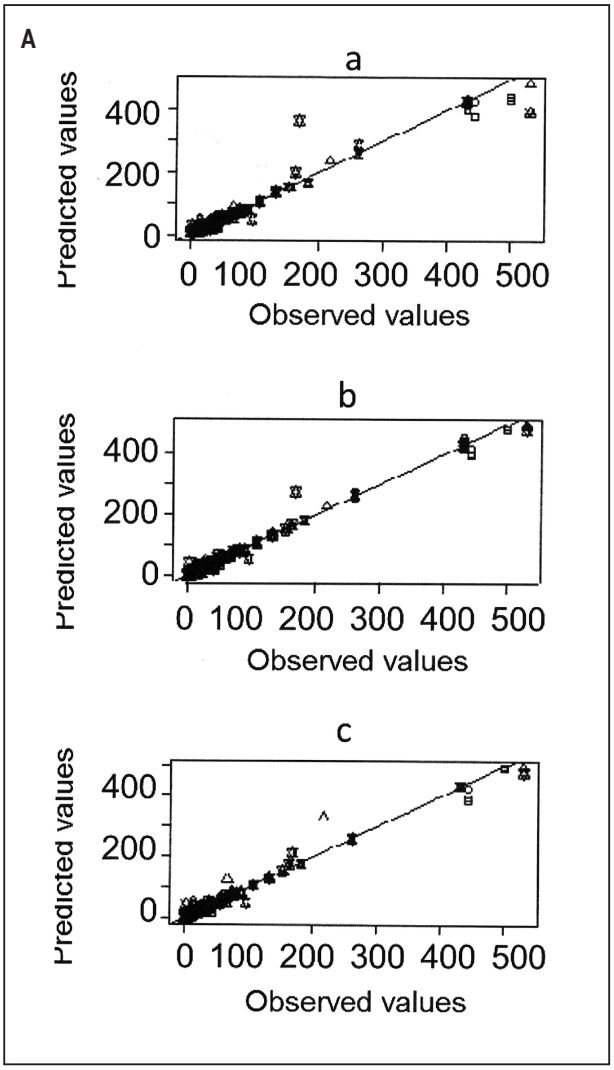
Interprétations

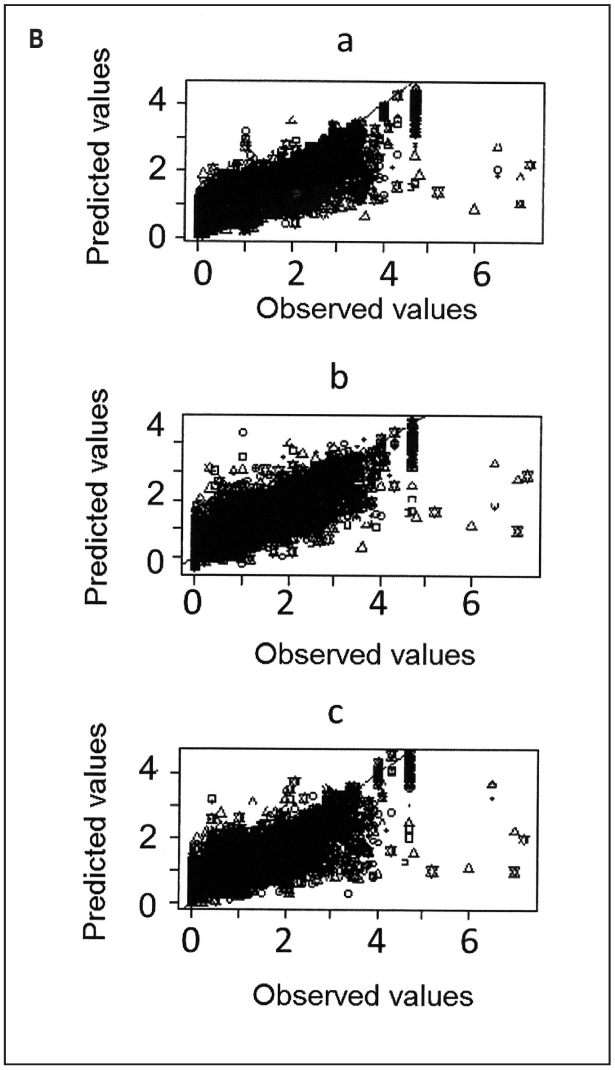
Pour le carbone, la très forte dépendance des valeurs minimales et maximales aux valeurs modales était attendue. Il a été montré de nombreuses fois que ce paramètre suivait une loi proche d'une distribution log-normale, et ce à des échelles allant de la parcelle au monde entier (voir par exemple : Arrouays *et al.*, 1995, 2003 ; Saby *et al.*, 2008 ; Jones *et al.*, 2005 ; Batjes, 1996). Le fait que les valeurs de carbone soient liées à la profondeur est également logique.

Pour le pH, l'effet du matériau est logique. Les sols sur matériaux calcaires ont généralement - à de notables exceptions près - des pH élevés et dans une gamme de variation réduite. La relation entre la valeur modale et l'étendue du pH est également intéressante et logique. En milieu basique (calcaire ou sodique) le pH est élevé mais ses variations sont faibles car il est toujours

Figure 1 - A : Valeurs observées et prédites pour la valeur minimale de la teneur en carbone organique d'une strate ; a) RF, b) BRT, c) Cubist. B. Etendue observée et prédite pour la valeur de pH_{eau} d'une strate ; a) RF, b) BRT, c) Cubist. Les figurés différents correspondent à différentes réalisations de la partition du jeu de données.

Figure 1 - *A: Observed vs predicted lower values of SOC; a) RF, b) BRT, c) Cubist. B. Observed vs predicted range of pH; a) RF, b) BRT, c) Cubist. Points shapes correspond to different k samples.*





tamponné par les cations Ca^{++} ou Na^{+}. Inversement, en milieu très acide le pH est très bas, mais ne varie pas non plus car il est contrôlé par l'ion Al^{+++}. C'est donc dans les gammes de pH «neutres» que l'on observe le plus de variations car dans ces milieux, le pH n'est pas stabilisé naturellement et dépend essentiellement des apports agricoles (engrais, chaulage…) qui sont très fortement variables dans le temps et dans l'espace.

Nous attendions un effet de la surface des UTS sur l'étendue observée au sein des strates. En d'autres termes, nous pensions que plus une UTS serait petite, plus les strates la décrivant seraient homogènes. Or on observe une faible dépendance des valeurs prédites à la superficie des unités. Ceci peut s'expliquer par le fait que de très vastes UTS peuvent être parfaitement homogènes pour certains caractères (comme par exemple la granulométrie dans les Landes de Gascogne) ou par le fait que de petites UTS peuvent être très variables et bien décrites dans la base de données.

Qualité de la prédiction

Les qualités de prédiction sont assez remarquables pour certaines variables, comme le carbone. En revanche, elles sont

Tableau 2 - Ordre d'importance relative des variables prédictives. A : altitude maximale, B : altitude minimale, C : superficie de l'UCS, D : catégorie de matériel parental regroupé en 13 catégories, E : forme morphologique regroupée en 20 catégories, F : profondeur minimale d'apparition, G : profondeur moyenne d'apparition, H : profondeur maximale d'apparition, I : épaisseur minimale, J : épaisseur moyenne, K : épaisseur maximale, L : valeur modale, M : altitude moyenne, N : altitude maximale moins altitude minimale, O : profondeur totale du sol, P : superficie de l'UTS.

***Table 2** - Order of importance of the co-variates. A : highest elevation, B : lowest elevation, C : soil mapping unit area, D : parent material class (13 classes), E : landform (20 classes), F : minimal upper depth, G : mean upper depth, H : maximal upper depth, I : smallest thickness, J : mean thickness, K : largest thickness, L : modal value, M : mean elevation, N : highest elevation minus lowest elevation, O : total soil depth, P : soil typological unit area.*

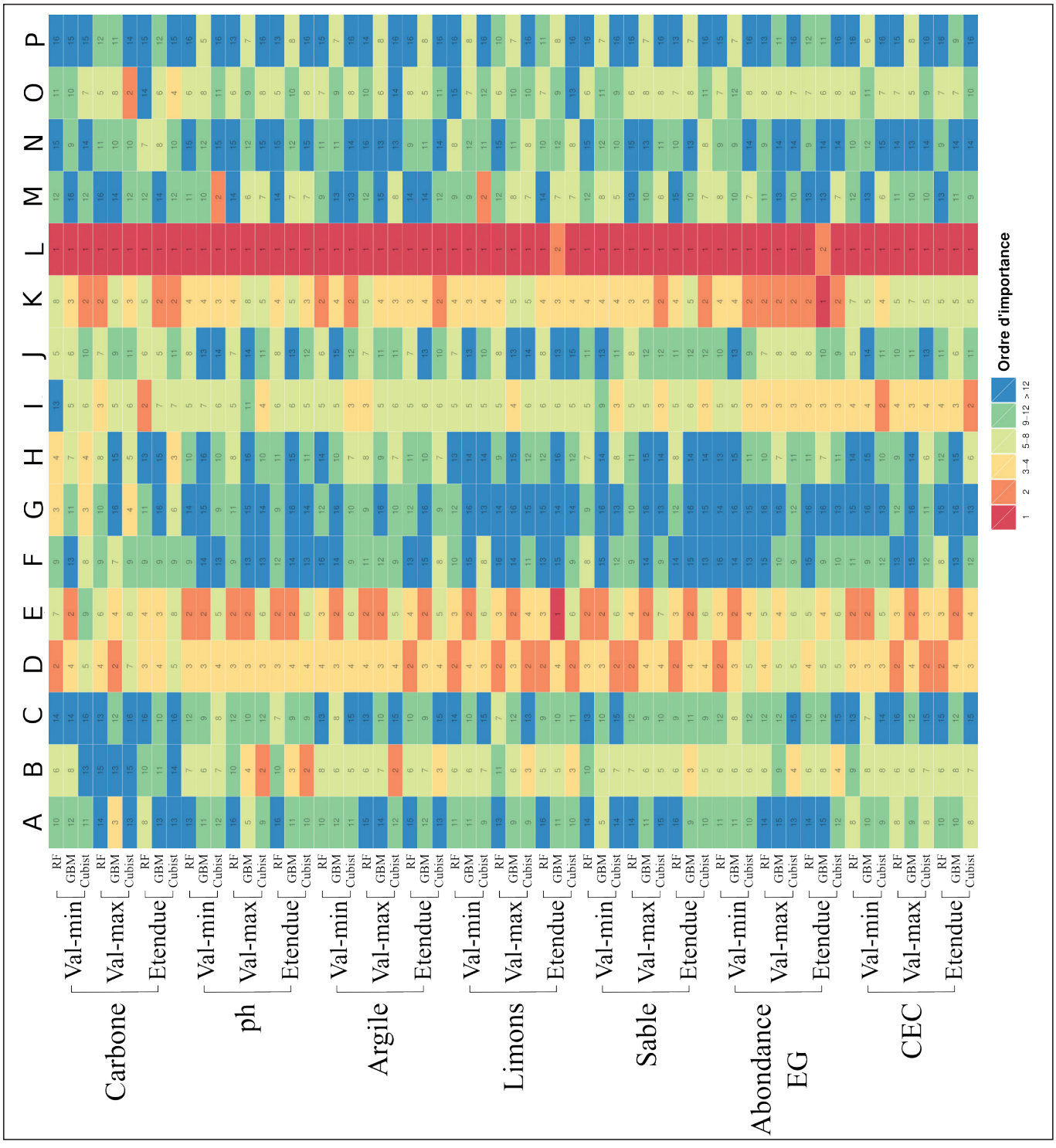
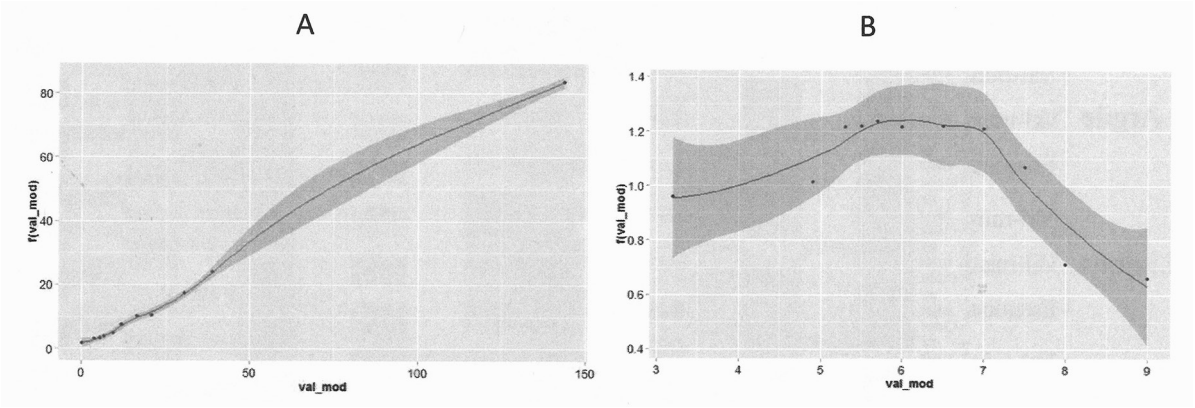


Figure 2 - A : Effet de la valeur modale du carbone (abscisse) sur la valeur minimale de la strate (ordonnée). Modèle GBM. B : Effet de la valeur modale du pH (abscisse) sur l'étendue de la strate (ordonnée). Modèle RF.

Figure 2 - *A: Effect of the modal value of carbon content (x axis) on the minimal value of carbon content (y axis). GBM model. B: Effect of the modal value of pH (x axis) on pH range (y axis). RF model.*



assez médiocres pour d'autres, comme le pH. Ceci était attendu du fait de la variabilité temporelle du pH qui dépend fortement des apports locaux et dont on sait qu'elle présente également une dynamique saisonnière forte (voir des ordres de grandeur de cette variabilité dans Baize (1988)).

Indépendance des prédictions

Une question que l'on peut se poser est : « vaut-il mieux prédire séparément valeurs minimales et valeurs maximales » ou directement l'étendue en supposant une symétrie ? Pour éclairer ce point, nous avons fait un calcul de propagation d'erreur sur l'opération « valmax-valmin » (moyenne quadratique des erreurs individuelles en supposant les erreurs indépendantes, ce qui sur-estime sans doute un peu le calcul d'erreur). Ce test a montré de très faibles différences entre les deux approches.

Intérêt et précautions d'emploi

Le fait que nous soyons parvenus à de bonnes qualités de prédiction pour certaines variables ne signifie aucunement que l'on doit se passer de l'expertise du pédologue pour fournir ces informations sur leur variabilité intra-strate. Le type de modèle que nous avons développé est principalement destiné à compenser un manque de données dans le cas de certaines études anciennes où seule la valeur modale est renseignée et où l'on ne dispose ni de suffisamment de données pour calculer ces indicateurs de dispersion, ni de financements pour compléter les études en question. En d'autres termes, le recours à ces modèles doit être considéré comme une mesure palliative et les méthodes d'estimation de cette variabilité, ainsi que les incertitudes qui y sont associées, doivent être précisées dans les bases de données.

CONCLUSION

L'intérêt principal de ce type d'approche est de pouvoir dériver des valeurs par défaut de ces indicateurs de dispersion lorsqu'elles sont manquantes dans les bases de données. On peut toutefois penser que les types de modèles que nous avons utilisés pourraient parfois conduire à un « sur-ajustement » qui donne une fausse idée de leur performance. Pour vérifier cela, il faudrait disposer d'une validation externe entièrement indépendante.

Trois solutions pourraient être envisageables:

- Une validation purement externe par un échantillonnage indépendant, qui pourrait par exemple consister en un échantillonnage aléatoire stratifié des strates où l'on ne connaît que la valeur modale et les variables environnementales et à un retour sur le terrain pour y réaliser de nombreuses analyses afin d'estimer les indicateurs de dispersion et de les comparer aux valeurs prédites par les modèles. Cette solution, qui est théoriquement la plus satisfaisante, pourrait toutefois être d'un coût prohibitif.
- Une autre solution pourrait être le recours à des avis d'experts (par exemple les auteurs des bases de données en cours de constitution) en leur soumettant les valeurs prédites et en leur demandant de les expertiser.
- Une troisième solution plus globale, serait de vérifier si, en utilisant ces estimations d'incertitude, on ne dégrade pas la prédiction du pourcentage de sites tombant en dehors de l'intervalle de confiance par rapport à la prédiction déduite des

valeurs minimale et maximale renseignées par les pédologues dans les bases de données. On pourrait ainsi utiliser les profils DoneSol comme données de validation (même si c'est en théorie biaisé car ils ont le plus souvent aussi servi à construire les UTS). Nous pourrions aussi faire une validation globale du même type en utilisant les profils du RMQS.

REMERCIEMENTS

Le programme IGCS est financé par le ministère en charge de l'agriculture dans le cadre du Groupement d'intérêt scientifique Sol). Ce travail n'aurait pas pu être mené sans la contribution de tous les pédologues qui alimentent la base de données DoneSol, nous les remercions collectivement. Ce travail s'inscrit également dans le cadre des travaux du RMT Sols et Territoires et du programme mondial « *GlobalSoilMap* ». Nous remercions Philippe Lagacherie et un lecteur anonyme pour leurs critiques constructives sur cet article.

BIBLIOGRAPHIE

Alfons A., 2012 - cvTools: Cross-validation tools for regression models. R package version 0.3.2.

Arrouays D., Vion I. et Kicin J.L., 1995 - Spatial analysis and modeling of topsoil carbon storage in forest humic loamy soils of France. *Soil Science*, 159, 191-198.

Arrouays D., Feller C., Jolivet C., Saby N.P.A., Andreux F., Bernoux M., Cerri C.E.P. , 2003 - Estimation de stocks de carbone organique des sols à différentes échelles d'espace et de temps. *Etude et Gestion des Sols*, 10(4) : 347-354.

Arrouays D., Grundy M. G., Hartemink A. E., Hempel J. W., Heuvelink G.B.M., Hong S.Y., Lagacherie P., Lelyk, G., McBratney A. B., McKenzie, N. J. Mendonça-Santos M.D., Minasny B., Montanarella L., Odeh, I. O. A., Sanchez P. A., Thompson J. A., Zhang G. L., 2014a - *GlobalSoilMap*: towards a fine-resolution global grid of soil properties. *Advances in Agronomy*, 125, 93-134.

Arrouays D., McKenzie N.J., Hempel J., Richer de Forges A.C., McBratney A.B. (eds), 2014b - GlobalSoilMap. Basis of the global spatial soil information system. CRC Press, Taylor&Francis, 478 p.

Baize D., 1988. Guide des analyses courantes en pédologie. INRA, Paris, 172 p.

Batjes N.H., 1996 - Total carbon and nitrogen in the soils of the world. *Eur. J. Soil Sci.* 47: 151-163.

Breiman, L., 2001 - Random Forests. *Machine Learning*, 45(1), 5-32.

Friedman J-H., 2002 - Stochastic Gradient Boosting. *Computational Statistics and Data Analysis*, 38(4), 367-378.

Grolleau E., Bargeot L., Chafchafi A., Hardy R., Doux J., Beaudou A., Le Martret H., Lacassin J.-Cl., Fort J.-L., Falipou P. et Arrouays D., 2004 -. Le système d'information national sur les sols: DoneSol et les outils associés. *Etude et Gestion des Sols,* 11(3), 255-269..

Hempel J.W., McBratney A.B., McKenzie N.J., Hartemink A.E., McMillan R., Lagacherie P., Arrouays D., 2013 - Vers une cartographie numérique des propriétés des sols du monde : Le programme *GlobalSoilMap*. *Etude et Gestion des Sols*, 20(1), 7-14.

Jones R.J.A., Hiederer R., Rusco E., Montanarella L., 2005 - Estimating organic carbon in the soils of Europe for policy support. *Eur. J. Soil Sci.* 56: 655-671.

Kuhn M., Weston S., Keefer C., Coulter N., 2014 - C code for Cubist by Ross Quinlan (2014). Cubist: Rule- and Instance-Based Regression Modeling. R package version 0.0.15.

Laroche B., Richer de Forges A.C., Leménager S., Arrouays D., Schnebelen N., Eimberck M., Toutain B., Lehmann S., Tientcheu E., Héliès F., Chenu J-P., Parot S., Desbourdes S., Girot G., Voltz M., Bardy M. 2014 - Le programme Inventaire Gestion Conservation des Sols de France : volet Référentiel Régional Pédologique. *Etude et Gestion des Sols*, 21, 125-139.

Liaw A., Wiener M., 2002 - Classification and Regression by randomForest. *R News* 2(3), 18--22.

Libohova Z., Wills S. et Odgers N.P., 2014 - Legacy data quality and uncertainty estimates for United States *GlobalSoilMap* products. In: Arrouays D., McKenzie N.J., Hempel J., Richer de Forges A.C., McBratney A.B.. (eds). GlobalSoilMap. Basis of the global spatial soil information system. CRC Press, Taylor&Francis, pp. 63-68.

Minasny B et McBratney A.B., 2010 - Methodologies for Global Soil Mapping. In: J.L. Boettinger *et al*., (eds). Digital Soil mapping. Bridging Research, Environmental Aplication and Operation. Logan, Utah. Springer.

Nauman T.W., Thompson J.A., 2014 - Semi-automated disaggregation of conventional soil maps using knowledge driven data mining and classification trees. *Geoderma*, 213, 385-399.

Richer de Forges A.C., Baffet M., Berger C., Coste S., Courbe C., Jalabert S., Lacassin J.-C., Maillant S., Michel F., Moulin J., Party J.-P., Renouard C., Sauter J., Scheurer O., Verbèque B., Desbourdes S., Héliès F., Lehmann S., Saby N.P.A., Tientcheu E., Jamagne M., Laroche B., Bardy M., Voltz M., 2014 - La cartographie des sols à moyennes échelles en France métropolitaine. *Etude et Gestion des Sols*, 21, 25-36.

Quinlan J.R., 1992 - Learning with continuous classes. Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, pp. 343-348.

Quinlan J.R., 1993 - Combining instance-based and model-based learning. Proceedings of the 10th International Conference on Machine Learning, pp. 236-243.

Ridgeway G with contributions from others, 2013 - gbm: Generalized Boosted Regression Models. R package version 2.1.

Saby N.P.A., Arrouays D., Antoni V., Foucaud-lemercier B., Follain S., Walter C., Schvartz C. 2008. Changes in soil organic carbon content in a French mountainous region, 1990-2004. *Soil Use and Management*, 24, 254-262.

Sanchez, P. A., S. Ahamed, F. Carre, A. E. Hartemink, J. Hempel, J. Huising, P. Lagacherie, A. B. McBratney, N. J. McKenzie, M L. de Mendonca-Santos, *et al*., 2009 - "Digital Soil Map of the World", *Science*, 325(5941), 680-681.