Cartographie numérique du Réservoir Utile en eau des sols à partir de données pédologiques anciennes : application à la plaine littorale Languedocienne

Quentin Styc

THÈSE POUR OBTENIR LE GRADE DE DOCTEUR DE MONTPELLIER SUPAGRO

En Sciences du Sol

École doctorale GAIA – Biodiversité, Agriculture, Alimentation, Environnement, Terre, Eau

Portée par

Unité de recherche UMR LISAH

Cartographie numérique du Réservoir Utile en eau des sols à partir de données pédologiques anciennes : Application à la plaine littorale Languedocienne

Présentée par Quentin Styc

Le 1er Juillet 2020

Sous la direction de Philippe LAGACHERIE

Devant le jury composé de

Laura POGGIO, Ingénieur de Recherche, ISRIC – World Soil Information	Rapporteure
Christian WALTER, Professeur, UMR SAS – Agrocampus Ouest	Rapporteur
Marc VOLTZ, Directeur de Recherche, UMR LISAH	Examinateur (Président du jury)
Julie CONSTANTIN, Chargé de Recherche, UMR AGIR	Invitée
François GONTARD, Directeur d'études, BRL Exploitation	Invité
Philippe LAGACHERIE, Ingénieur de Recherche, UMR LISAH	Directeur de thèse

UNIVERSITÉ DE MONTPELLIER

l'institut Agro agriculture • alimentation • environnement | Montpellier SupAgro

BRL Exploitation

ANRT ASSOCIATION NATIONALE RECHERCHE TECHNOLOGIE

LISAH Laboratoire d'étude des Interactions Sol - Agrosystème - Hydrosystème

INRAE la science pour la vie, l'humain, la terre

REMERCIEMENTS

« Cette phrase est la première pourtant c'est la dernière que j'ai écrite »
(Disiz La Peste, ADN)

Bien que ces quelques lignes soient les premières que les lecteurs de cette thèse liront, il s'agit bien des dernières lignes rédigées qui concluent ces trois années de thèse. Cette thèse a été une véritable expérience avec des hauts, des bas, mais aussi beaucoup de connaissances et de nouvelles compétences acquises pendant ces trois années. Je tiens à commencer ces remerciements par les personnes qui ont permis à cette thèse de voir le jour. Un grand merci à Philippe Lagacherie, mon directeur de thèse. Ces trois années à travailler ensemble m'ont été très enrichissante autant sur la rigueur scientifique que demande la recherche, que sur l'ouverture d'esprit. J'ai beaucoup appris sur la manière de raisonner sur un sujet de recherche et c'est en grande partie grâce à nos échanges pendant cette thèse. Je remercie également François Gontard d'avoir participer à l'encadrement de cette thèse. Merci de m'avoir accueilli et intégré chez BRL, bien que n'étant présent que rarement ; ainsi que de m'avoir convié malgré cela à chaque évènement BRL. Merci également pour ta disponibilité et pour ton aide précieuse pour l'utilisation des données BRL qui a été très importante dans ces travaux de thèse.

Je tiens à remercier l'ensemble de l'UMR LISAH pour leur accueil depuis le début de ma thèse et même avant cela. Merci à Jêrome Molénat, Olivier Grunberger, Jean Stéphane Bailly et Frédérique Jacob de m'avoir accueilli au sein des locaux du LISAH pendant ces trois années. Je tiens aussi à dire un grand bravo et un grand merci à l'équipe du pôle administratif, Azziza, Nadia, Céline et Virginie, pour avoir assuré toutes les démarches administratives permettant une bonne installation dans les locaux du LISAH. Merci également pour avoir répondu à chacune de mes questions et de m'avoir guidé pour les différentes tâches administratives. Merci à Dominique Carrière d'avoir contribuer à ce que cette thèse se passe bien en me fournissant un équipement informatique irréprochable. Merci à Arnaud Dubreuil

d'avoir répondu à mes questions techniques sur les calculateurs entre autres, et pour sa bonne humeur et sa gentillesse. Merci à Fabrice Vinatier pour sa disponibilité, sa pédagogie et sa bonne humeur. Merci de m'avoir énormément appris sur R, j'ai sans aucun doute gagné beaucoup de temps via quelques fonctions que tu m'as transmis, c'est certain ! Je tiens également à remercier mes collègues de bureau : Laetitia, Baptiste, Bruce, Walid et Anaïs. C'était un véritable plaisir de partager le bureau 217 avec vous, bureau qui était très vivant entre parties de fléchettes, de basket, et de ping-pong, sapin de Noël, décoration de coupe du monde (⚔⚔). Merci à Gabrielle, Camille, Martin et Nicolas pour les discussions intéressantes et les bons moments pendant les pauses café. Merci également à Bruce, Manon, Pauline, Meriem, Mariem, Sebastian pour votre sympathie, bonne humeur et les multiples parties de fléchettes qui m'ont permis de mieux gérer la pression de fin de thèse. Un énorme merci à tous mes autres collègues que je n'ai pas cités (Laurent, Guillaume, Armand, Denis et tous les autres) de m'avoir permis de passer trois belles années au LISAH.

Je remercie également l'équipe de BRL Exploitation pour leur accueil à chacune de mes venues au sein des locaux. Bien que n'ayant été que peu présent, j'ai apprécié chaque passage à Nîmes.

Je tiens également à remercier Sébastien Salvador-Blanes de m'avoir fait découvrir les sciences du sol lors de mon parcours universitaire et d'avoir beaucoup contribué à mon parcours professionnel.

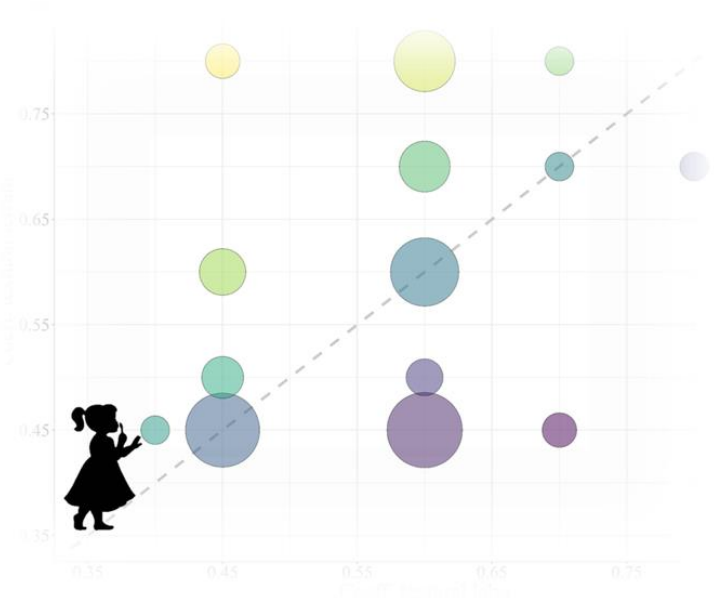
Je remercie les membres du comité de pilotage de thèse, Manuel Martin, Dino Ienco, Frédérique Seyler, Jean Stéphane Bailly et Jérôme Molénat pour m'avoir guidé dans mes travaux de thèse.

Je remercie l'ensemble des membres du jury de thèse d'avoir accepté avec plaisir de juger mon travail.

Je finirai cette partie en remerciant mes proches. Merci à mes parents pour leur soutien sans faille le long de mon parcours universitaire, mes stages, et la thèse. Merci à vous pour les valeurs que vous m'avez inculqué dès mon plus jeune âge. Vous êtes une véritable source d'inspiration et si j'ai pu aller si loin c'est en grande partie grâce à vous. A mes sœurs (Maëva, Johanne et Laurine), merci de leur soutien tout au long de la thèse et de m'avoir écouté et reboosté dans les mauvais moments. Merci également à Sylvie et Patrick de leur soutien. Merci à mes grands-parents ainsi qu'à mes tantes, oncles, cousins, cousines, enfin… à toute ma famille qui m'a également soutenu pendant cette thèse. Je remercie également mes amis proches,

Thomas, Chris, Maël, Florian, Fanny, Julie, pour tous les très bons moments malgré la distance. Merci également à Laetitia, Camille, Carole, Bruce, Martin, Nicolas, Marion G. et Marion D. (et surement d'autres) pour des bons moments passés à Montpellier. Merci aux *Swing-Nutella* également pour les lundis soir swing & barbotte et les après-midi répèt' (en principe). Merci également à mes amis de la fac Jassim, Keita, Thomas, Guillaume, Maxime et Florian, ces années universitaires ont été un vrai plaisir à vos côtés ! Je conclurai cette partie remerciement par une personne très importante qui m'accompagne et me soutient depuis un bon moment maintenant, Jenny Cesbron. Merci pour ton soutien sans faille tout au long de ces années et plus particulièrement ces dernières semaines. Merci de m'avoir écouté et reboosté dans les moments de doutes. Merci de m'avoir permis de finir cette thèse, entouré d'amour et de multiples desserts et bonbons !

Merci à tous et bonne lecture !



Le résultat d'une figure de thèse et d'un soupçon de légèreté

« Sometimes you gotta run before you can walk. »

- T.S. -

Résumé

Le réservoir utile du sol (RU) désigne la capacité des sols à stocker l'eau pour les plantes. Le RU joue un rôle majeur dans de nombreux domaines tels que la production alimentaire et la régulation des inondations. Le RU est également un élément essentiel pour prévoir et optimiser l'irrigation des sols localement. Par conséquent, il est primordial de connaitre avec précision les variations spatiales du RU. Cependant, les bases de données pédologiques actuelles ne fournissent pas une cartographie du RU qui soit à la fois exhaustive et suffisamment précise pour être utilisée à l'échelle de la parcelle. Cette limite pourrait être levée en utilisant le gisement considérable de données pédologiques anciennes non numérisées comme données d'apprentissage d'un modèle de cartographie numérique des sols. Ainsi, la société BRL Exploitation (BRLE) dispose sur son périmètre irrigué (6 636 km^2 en plaine littorale Languedocienne) de 228 000 observations de sol. L'objectif de cette thèse a été de développer une méthodologique de cartographie numérique du réservoir utile à partir de ces données pédologiques anciennes.

Le RU étant une propriété fonctionnelle de sol impliquant plusieurs propriétés de sol sur plusieurs profondeurs, les premiers travaux de cette thèse ont porté sur l'impact de la trajectoire de calcul utilisée sur les performances de prédiction du RU, une trajectoire étant définie par l'ordre selon lesquels sont effectuées les opérations de calcul du RU sur une couche de sol donné, d'agrégation des couches de sol et de spatialisation, afin d'obtenir une prédiction du RU. En prenant l'exemple du Languedoc-Roussillon, 18 trajectoires utilisant les données pédologiques disponibles dans le Référentiel Régional Pédologique pour produire une prédiction du RU ont été testées. La meilleure performance de prédiction a été obtenue par la trajectoire de calcul calculant puis spatialisant le RU pour quatre couches de sol distinctes et combinant enfin ces quatre couches.

Ensuite, les fonctionnalités du modèle de prédiction du RU ont été complétées par une prédiction des incertitudes associées, essentielle à l'aide à la décision. Ces incertitudes ont été prédites par un modèle de propagation d'erreurs utilisant les erreurs de spatialisation du RU sur les quatre couches de sol spatialisées séparément - estimées par une forêt aléatoire quantile - et prenant en compte les corrélations d'erreurs de ses composants. L'utilisation de ce modèle a montré une bonne aptitude à estimer et spatialiser l'incertitude de prédictions, dans un contexte de faibles performances de prédiction du RU.

La dernière étape a été consacrée à l'utilisation de données anciennes de BRLE pour alimenter un modèle de cartographie numérique des sols à l'échelle locale, en prenant l'exemple de la commune de Bouillargues. L'augmentation de densité spatiale par l'ajout de sondages au jeu de profils de sol, associée à l'ajout dans l'algorithme d'apprentissage de données représentant la position géographique des sondages, a considérablement amélioré la résolution spatiale, les performances de prédictions du RU et la précision des cartes d'incertitude. Cependant, les erreurs de caractérisation du RU sur ces sondages ont été constatées comme un facteur limitant les performances de prédictions du RU et de son incertitude. Une meilleure prise en compte de ces erreurs serait nécessaire pour améliorer les résultats actuels.

Les travaux de thèse ont permis de concevoir et de tester une démarche visant à valoriser l'utilisation des données pédologiques anciennes dans une approche de cartographie numérique des sols appliquée au RU. Une chaîne de traitement informatique visant à déployer la démarche sur l'ensemble du périmètre irrigué a été développée et une étude de coût/bénéfice a été réalisée. L'automatisation, au moins partielle de la saisie des sondages apparaît comme une condition nécessaire à la réalisation d'une carte du RU à l'échelle parcellaire sur tout le périmètre irrigué de BRLE.

Mots clés :

Réservoir utile ; cartographie numérique des sols ; données anciennes ; incertitude ; arbres de régression ; densité ; trajectoire de calcul ; propagation d'erreur ; échelle locale ; échelle régionale.

Abstract

Soil available water capacity (SAWC) refers to the soil capacity to store water for plants. SAWC plays an important role on many ecosystem services such as food security and flood and gas regulation. SAWC is also crucial for planning and optimizing soil irrigation at local scale. Thus, it is essential to understand the SAWC spatial distribution. However, current soil databases do not provide mapping of SAWC that is both comprehensive and accurate enough for use at plot scale. This limitation could be overcome by using the considerable repository of legacy soil data, undigitized, as learning data for a digital soil mapping model. BRL Exploitation (BRLE), thus, has 228,000 soil observations on its irrigated perimeter (6,636 km² in the Languedoc coastal plain). The aim of this thesis was to develop a useful digital mapping methodology for the SAWC based on these legacy soil data.

As the SAWC is a functional soil property involving several soil properties over several depths, the first work of this thesis focused on the impact of the inference trajectory used on the prediction performance of the SAWC, an inference trajectory being defined by the order in which the operations of computation of the SAWC on a given soil layer, aggregation of soil layers and mapping are performed, to finally obtain a prediction of the SAWC. Taking the example of Languedoc-Roussillon, 18 inference trajectories using the pedological data available in the Regional Pedological Referential to produce a prediction of the SAWC, were tested. The best prediction performance was obtained by the trajectory calculating then mapping the SAWC for four distinct soil layers and finally combining these four layers.

Then, the functionalities of the SAWC prediction model were complemented by a prediction of the associated uncertainties, which is important data for decision support. These uncertainties were predicted by an error propagation model using the mapping errors of SAWC on the four separately spatialized ground layers - each estimated using a Quantile Random Forest - and taking into account the error correlations of these components. The use of this model has shown a good ability to estimate and map the uncertainty of predictions, in a context of low overall prediction performance of the SAWC.

The last step of this thesis was devoted to the use of legacy data from BRL Exploitation to feed a digital soil mapping model at local scale, taking the example of the commune of Bouillargues. The increase in spatial density by adding auger holes to the set of soil profiles, associated with the addition in the learning algorithm of data representing the geographical

position of the auger holes, has considerably improved the spatial resolution, the prediction performance of the SAWC and the accuracy of the associated uncertainty maps. However, SAWC characterization errors on these auger holes were found to be a limiting factor in the prediction performance of the SAWC and its uncertainty. A better consideration of these errors would be necessary to improve the current results.

The thesis work has enabled the design and testing of an approach to enhance the use of legacy soil data in a digital soil mapping approach applied to the useful reservoir. A computer processing chain aimed at deploying the approach on the whole irrigated perimeter was developed and a cost/benefit study was carried out. The automation, at least partial, of the auger hole data entry operations appears to be a necessary condition for the production of a map of the SAWC on a plot scale over the whole of the BRL Exploitation irrigated perimeter.

Keywords:

Soil available water capacity; digital soil mapping; legacy soil data; uncertainty; random forest; density; inference trajectory; error propagation; local scale; regional scale.

Partenaire industriel (contrat CIFRE)

BRL Exploitation, 1105 Avenue Pierre Mendès

30000 Nîmes

Table des matières

Liste des figures

Introduction générale

Chapitre 1. État de l'art et positionnement scientifique

Chapitre 2. La zone d'étude et les données utilisées dans la thèse

Chapitre 3. Sensibilité des estimations du réservoir utile aux trajectoires de calcul

Chapitre 4. Quantification de l'incertitude liée à la spatialisation du réservoir utile par l'utilisation d'un modèle de propagation d'erreurs

Chapitre 5. Évaluation de l'intégration des données pédologiques anciennes denses et hétérogènes dans les modèles de cartographie numérique des sols appliqués au réservoir utile

Conclusion générale

Liste des tables

Chapitre 1. État de l'art et positionnement scientifique

Chapitre 2. La Zone d'étude et les données utilisées dans la thèse

Chapitre 3. Sensibilité des estimations de réservoir utile aux trajectoires de calcul

Chapitre 4. Quantification de l'incertitude liée à la spatialisation du réservoir utile par l'utilisation d'un modèle de propagation d'erreurs

Chapitre 5. Évaluation de l'intégration des données pédologiques anciennes denses et hétérogènes dans les modèles de cartographie numérique des sols appliqués au réservoir utile

Liste des abréviations et acronymes

CSMS :	Cartographie des Sols à base de Modélisation Statistique
DSM :	Digital Soil Mapping
CNS :	Cartographie Numérique des Sols
QRF :	Quantile Regression Forest
RF :	Random Forest
SIG	Système d'Information Géographique
SRTM :	Shuttle Radar Topography Mission
DEM :	Digital Elevation Model
MNT :	Modèle Numérique de Terrain
MRVBF :	Multi-Resolution Index of Valley Bottom Flatness
MRRTF :	Multi-Resolution Ridge Top Flatness
TPI :	Topographic Position Index
TWI :	Topographic Witness Index
IGN :	Institut de Géographie National
BRGM :	Bureau de Recherches Géologiques et Minères
IGCS :	Inventaire Gestion et Conservation des Sols
SAWC :	Soil Available Water Capacity
RU :	Réservoir Utile
AWC :	Available Water Capacity
$AWC_{E:}$	Elementary Available Water Content
FC :	Water content at Field Capacity
PWP :	Water content at Permanent Wilting Point
CC :	Humidité à la Capacité au Champ
PFP :	Humidité au Point de Flétrissement Permanent
ST :	Soil Thickness
SD :	Soil Depth
$SS_{MSE:}$	Mean Square Error Skill Score
RMSE :	Root Mean Square Error
PICP :	Prediction Interval Coverage Probability
SSVR :	Spatially Structured Variance Ratio

Introduction générale

La gestion de la ressource en eau est l'un des principaux enjeux agro-environnementaux qui se pose à notre société. A l'échelle de l'écosystème agricole, tout l'enjeu est de pouvoir assurer la production agricole pour une démographie croissante dans un contexte de changement climatique et d'intensification des étés secs. Au sein de cette configuration, les sols ont un rôle clé à jouer notamment par leur capacité à stocker l'eau pour la croissance des plantes.

Cette capacité de stockage de l'eau par le sol est définie par un concept bien connu qu'est le réservoir utile du sol (Veihmayer and Hendrickson, 1927). Le réservoir utile en eau du sol constitue la principale source en eau du sol mobilisable par les racines des plantes pour leur alimentation hydrique et leur transpiration sur le long terme. Si le réservoir utile en eau du sol est souvent associé à la « réserve utile en eau des sols », il s'agit de deux notions distinctes. Le réservoir utile renvoie vers une capacité de stockage quand la réserve utile informe de l'état de remplissage du réservoir utile à un instant *t*.

Le Réservoir Utile en eau du sol (RU) peut se caractériser comme une propriété fonctionnelle de sol, à mi-chemin entre une propriété de sol et une fonction du sol. Il ne peut être considéré comme propriété de sol car son estimation implique elle-même plusieurs propriétés primaires de sol. S'il n'est pas non plus considéré comme une fonction de sol, le RU permet, cependant, d'évaluer certaines fonctions de sol telles que les fonctions de stockage, filtration et transport de l'eau.

Le RU peut également être utilisé dans l'évaluation de nombreux services écosystémiques tels que : i) la régulation de la ressource en eau (qualité et quantité) et des nutriments, ii) la séquestration du carbone atmosphérique dans les sols et la régulation des gaz à effet de serre (méthane CH_4 et protoxyde d'azote N_2O), iii) les productions d'aliments, de biomasse et de fibre, iv) la régulation des inondations, v) l'évaluation du risque de désertification ainsi que la filtration des contaminants (Dominati et al., 2014). Dans un contexte d'urbanisation à outrance, le RU sert aussi à évaluer les potentialités agronomiques des sols. De ce fait, les sols à fort potentiel sont d'avantage protégés contre l'artificialisation des sols.

De ce fait, et au vu de ses larges implications dans divers domaines agro-environnementaux, le RU est largement utilisé par les modélisateurs comme indicateur agronomique et comme indicateur de l'état hydrique des sols dans des modèles tels que STICS (Brisson et al., 1998), CENTURY (Parton et al., 1987), APSIM (O'Leary et al., 2016), ou SWAT (Arnord et Williams, 1987 ; Arnold et Fohrer, 2005). Par le biais de ces modèles, le RU est un indicateur agronomique important afin de planifier et d'optimiser les cultures.

L'irrigation est également un domaine d'applicabilité du RU. Effectivement, selon la FAO (Organisation pour l'Alimentation et l'Agriculture), la part de la ressource en eau utilisée pour l'irrigation est estimée à 70 % de la ressource mondiale en eau permettant d'assurer 40 % de la production alimentaire mondiale (OCDE, 2002). Pour cela, le RU peut être utilisé comme donnée d'entrée dans des outils de pilotage d'irrigation afin de déterminer quand l'irrigation est nécessaire. Il permet également d'être utilisé comme indicateur pour identifier les zones agricoles potentiellement irrigables (Gaudin et Gary, 2012 ; Zhang et al., 2015). Il est également utilisé dans l'équation du bilan hydrique dans le but d'optimiser l'irrigation en ajustant les doses pratiques d'irrigation. De plus, l'irrigation est également au cœur d'un enjeu plus vaste socio-environnemental reposant sur l'utilisation concurrentielle de l'eau entre l'agriculture, les secteurs industriels et les villes. Il est donc primordial dans ces circonstances de pouvoir évaluer la capacité de rétention en eau verte (eau disponible pour la croissance des plantes) (Falkenmark et Rockstrom, 2006 ; Liu et al., 2009) à travers le RU, afin d'être en mesure de distribuer l'eau avec efficacité et parcimonie. Pour ce faire, des décisions doivent être prises à l'échelle locale par les décideurs et les politiques d'aménagement du territoire à partir de la connaissance des variations dans l'espace du RU. Cependant, le RU est très variable dans le paysage du fait des variations des différentes propriétés primaires de sols qui le définissent. Ainsi des cartes précises du RU doivent être établies pour étayer ces décisions.

Les possibilités de cartographier le RU sont actuellement très limitées. Les bases de données pédologiques actuellement disponibles sont renseignées à des résolutions trop grossières telles que la carte pédologique de la France au 1/1 000 000^{e} ou encore la carte du Référentiel Régional Pédologique du Languedoc Roussillon au 1/250 000^{e}. Effectivement, l'utilisation de carte pédologique n'est intéressante qu'à des échelles très détaillées permettant de mieux capturer la variabilité locale des propriétés des sols (Leenhardt et al., 1994). Or, ces types de cartes sur le territoire ne sont que peu disponibles et très souvent, non numérisées. Ces cartes sont également fastidieuses à interpréter en termes de RU notamment à cause des notices de cartes non standardisées voire incomplètes (Bornand, 1997, Voltz et al., 2018).

Pour dépasser cette carence en cartographie des sols, la communauté scientifique a développé une nouvelle approche, le « Digital Soil Mapping » (Mc Bratney, 2003), appelée en France cartographie des sols à base de modélisation statistique (CSMS) (Voltz et al, 2018). Le principe consiste à estimer en tout point de l'espace des classes de sol ou leurs propriétés au moyen d'un modèle numérique utilisant en entrée les données spatiales disponibles sur les éléments du paysage en relation avec le sol. Ce modèle est calibré sur les observations ponctuelles de sol disponibles sur la zone à étudier. La CSMS s'est révélée, ces deux dernières décennies, une solution opérationnelle de cartographie des sols, constituant ainsi une alternative à la cartographie pédologique classique. Cependant, les premiers résultats obtenus sur des bassins versants (Nussbaum et al, 2018), des régions (Vaysse et al, 2015), des pays (Mulder et al, 2016) ou à l'échelle mondiale (Hengl et al, 2017) ont montré des précisions limitées pour une majorité de propriété des sols. L'hypothèse généralement admise est que le facteur limitant premier est la faible densité spatiale des observations de sol actuellement disponibles dans les bases de données pédologiques qui sont utilisées pour calibrer les modèles de CSMS (Lagacherie, 2008)

Une voie possible pour augmenter les densités d'observation de sol disponibles (et donc les précisions des cartes obtenues par CSMS) est de mobiliser les informations géo-localisées sur les sols présents dans les rapports d'études pédologiques anciennes sous la forme de profils de sols avec ou sans analyses et de sondages pédologiques. Un tel gisement de données pédologiques non utilisé suscite un fort intérêt, décliné à différentes échelles, dans le but de compléter les bases de données actuelles via des programmes de récupération de données anciennes. A l'échelle globale, bien que le nombre de profils intégrés dans les bases de données ait considérablement augmenté (+ 1 064 % entre 2009 et 2015 selon Arrouays et al, 2017), l'acquisition des données pédologiques anciennes non saisies reste un objectif important. A l'échelle nationale, l'intégration de données anciennes a déjà été engagée dans le but d'augmenter la quantité de profils et de sondages de DoneSol (70 000 profils et 97 000 sondages répartis de façon hétérogène sur le territoire) (Voltz et al., 2018). A des échelles plus locales, de nombreux gisements de données pédologiques existent, en relation avec des prospections pédologiques souvent très détaillées. Ainsi, la compagnie d'aménagement BRL a collecté massivement des observations de sols entre 1957 et 1992 dans le cadre d'une mise en irrigation du territoire. Ces études pédologiques incluent 228 000 observations de sols réparties sur les 6 636 km^2 de la concession de BRL Exploitation, en plaine littorale Languedocienne. Cette source de données permettrait d'appréhender la variabilité spatiale des propriétés de sol et du RU à des

degrés de détail inédits. De plus, l'exploitation de cette masse de données représente une alternative d'autant plus crédible qu'elle est économiquement plus viable que la réalisation de nouveaux profils de sol. En effet, Voltz et al. (2018) ont évalué à 1 800 € le coût total de réalisation d'un nouveau profil de sol.

Dans le cadre de cette thèse, l'objectif général consistera à développer et tester une méthode de spatialisation du réservoir utile permettant d'accéder à des résolutions spatiales fines compatibles avec des prises de décisions à la parcelle tout en étant applicable sur l'ensemble du territoire géré par BRL Exploitation (Figure 0.1). La voie choisie pour atteindre cet objectif est de mettre en œuvre une approche de CSMS fondée sur l'utilisation, autant que possible et en priorité des données pédologiques anciennes disponibles.

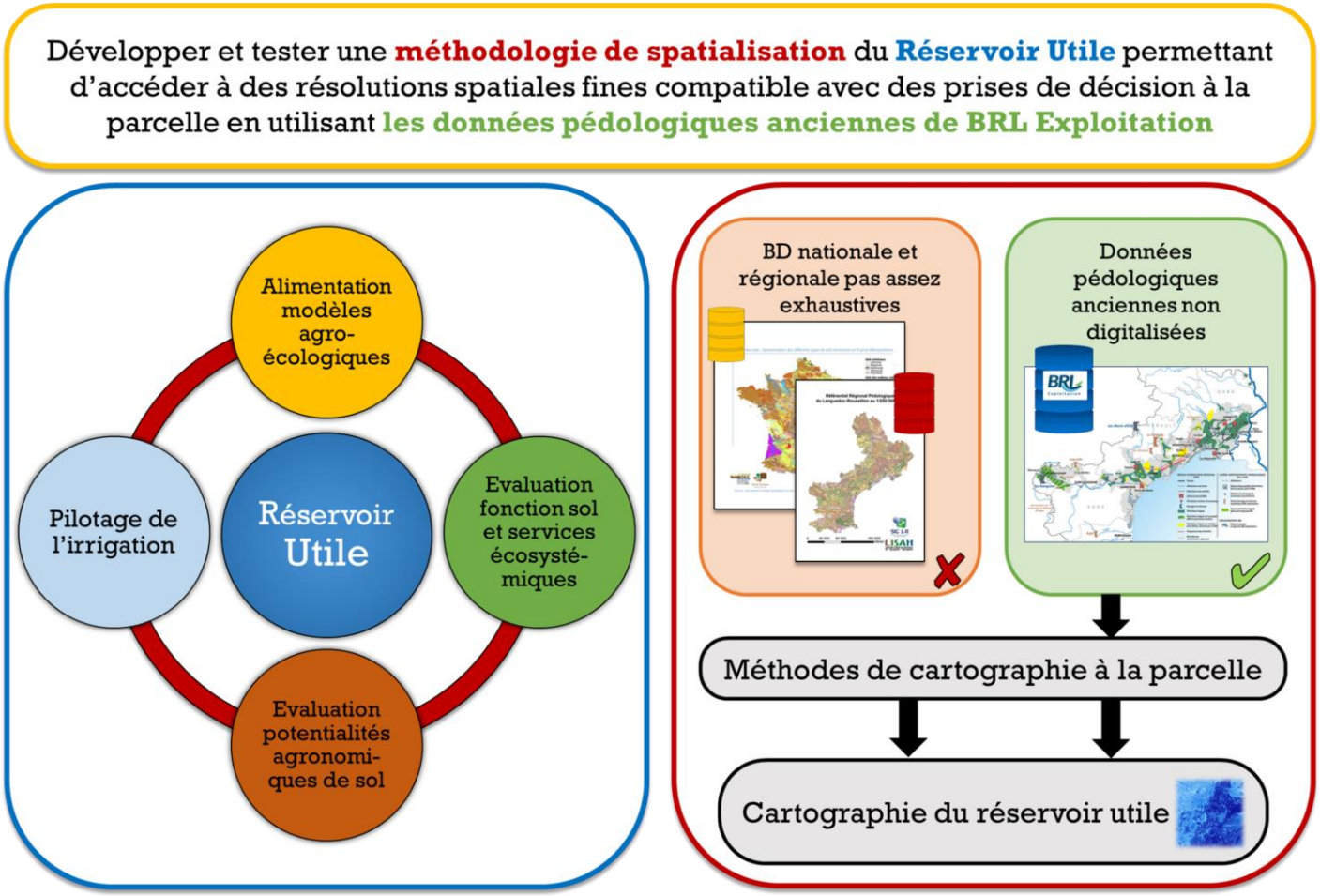


Figure 0.1. Présentation synthétique de l'objectif de thèse

CHAPITRE 1 État de l'art et positionnement scientifique

Dans ce chapitre seront présentés succinctement : i) les spécificités du réservoir utile (RU) et de son estimation locale, ii) le principe de la cartographie numérique des sols et ses applications et enfin iii) les questions méthodologiques relevées qui structureront le plan de cette thèse.

I. Estimation locale du réservoir utile

Cette section est réservée à la présentation de la méthode de calcul du RU à l'échelle d'un site. Nous présenterons la formulation utilisée afin d'estimer le RU et détaillerons chacune de ses composantes.

I.1. Calcul du réservoir utile

Le RU est une propriété fonctionnelle définie comme la capacité du sol à stocker l'eau et à la restituer à la plante. Actuellement, le RU est calculé de façon opérationnelle selon la formulation suivante (Cousin et al., 2003) :

$$RU = \sum_{n=1}^{i=Prof_{max}} dh_i * bd_i * \left(\frac{100 - st_i}{100}\right) * (\theta r_i - \theta w_i) \qquad \text{(Eq. 1.1.)}$$

Où $prof_{max}$ est la profondeur maximale (en cm) du profil, dh_i est l'épaisseur de l'horizon i (en cm), bd_i, la densité apparente (en kg.dm^{-3}), st_i la teneur volumique des éléments grossiers (en %), θr_i, l'humidité massique à la capacité au champ (en g.g^{-1}) et θw_i, l'humidité massique au point de flétrissement permanent (en g.g^{-1}).

I.2. Les propriétés de sol composantes du réservoir utile et leur mesure

I.2.1. La profondeur

Le terme de profondeur sur laquelle le RU est estimé, ne fait pas consensus dans la communauté scientifique.

En effet, historiquement, le RU était orienté comme un paramètre lié à la plante, visant à stocker et à restituer l'eau disponible à ces racines (Veihmeyer et Hendrickson, 1927). Cette définition décompose le RU en deux propriétés : i) la capacité du sol à stocker l'eau et ii) la capacité de la plante à prélever l'eau stockée dans le sol. La notion de profondeur est très dépendante de l'objectif de l'étude liée à l'utilisation du RU et de l'importance relative de ses deux propriétés. Elle peut être définie comme étant : i) **la profondeur maximale du sol**, pour les études focalisées sur la capacité de stockage en eau du sol, ou ii) **la profondeur de l'enracinement de plantes**, pour les études portées sur la capacité de prélèvement du stock en eau du sol par les plantes.

Par la bivalence de cette définition, le terme de profondeur demeure ambigu dans l'estimation du RU. Par la suite, nous allons explorer les spécificités de chaque aspect de la profondeur présentées ci-dessus.

I.2.1.1. La profondeur de sol

La profondeur de sol, telle que définie par les spécifications du programme *GlobalSoilMap* (Arrouays et al., 2014), correspond à la profondeur de la roche-mère qui se matérialise par soit : un contact net avec une roche consistante et indurée ou bien à une roche meuble et désagrégée se référant à la définition d'un contact paralithique du Soil Survey Division Staff (1993) (liseré rouge de la Figure 1.1). La présence d'un contact paralithique rejoint la notion de profondeur d'enracinement potentiel de la directive INSPIRE. Elle consiste à tenir compte des limites dues à la composition de certain horizon pédologique (horizon compacté) formant un obstacle au développement racinaire et *a contrario* à la présence de racines dans cette zone paralithique ou plus communément appelée le saprolite. Par exemple, des sols développés sur vignes ou forêts peuvent potentiellement contenir la présence de racine dans le saprolite (Graham et al., 2010 ; Brantley et al., 2017, Algayer et al, 2020).

Outre les problèmes de définition de la profondeur, la mesure de cette propriété de sol n'en est pas pour le moins aisée. En effet, les mesures de profondeur de sol répertoriées dans les bases de données sont souvent sous-estimées pour les sols profonds, quand la profondeur du contact lithique ou paralithique se situe en dessous de la profondeur maximale d'observation (Figure 1.1). Les facteurs de cette sous-estimation sont divers : limite de l'outil de prospection (exemple : tarière à main limitée à 120 cm) ou encore la présence d'obstacle (système de drainage, formation argileuse compacte, etc.) (Lacoste et al., 2016).

I.2.1.2. La profondeur d'enracinement

La profondeur d'enracinement a une forte influence sur l'estimation du RU en fonction de l'espèce de la plante présente et de son état de croissance. Malgré cela, cette profondeur n'est qu'occasionnellement mesurée *in situ* car fastidieuse à réaliser et que généralement les valeurs peuvent être déterminées à partir de résultats compilés (e.g., Allen et al., 1998). Cependant, le niveau de précision de ces données est faible avec une amplitude d'erreur atteignant 1 m selon les espèces. La profondeur d'enracinement évolue principalement en fonction de l'espèce végétale mais aussi selon les conditions pédologiques du sol (disponibilité de l'eau et des nutriments). Dans un contexte climatique de disponibilité systématique d'eau, la profondeur d'enracinement est également inférieure par rapport à un contexte climatique restrictif, pour une même espèce cultivée et un même sol.

La répartition verticale des racines dans le sol est aussi un élément important sur l'estimation du RU et la capacité des racines à prélever l'eau. Les racines sont généralement réparties de façon homogène en surface et sub-surface et deviennent progressivement éparses quand la profondeur augmente. Cette structuration du système racinaire interroge la notion de RU qui considère les racines équitablement efficaces sur l'ensemble de la profondeur. Plusieurs études ont démontré (Algayer et al., 2020 ; Tanaka et al., 2004) que l'utilisation d'environ la moitié de la profondeur de l'enracinement était nécessaire afin d'évaluer directement ou indirectement le RU. De plus, la différence de densité racinaire entre la surface et les couches profondes influent également sur la capacité des racines à prélever l'eau disponible. Des travaux menés par Cabelguenne et Debaeke (1998) mettent en évidence la capacité des racines des couches superficielles à prélever l'eau à un potentiel inférieur à -1500 kPa alors que pour les couches profondes, l'eau disponible dans le sol est faiblement prélevée due à une plus faible densité racinaire. De ce fait, il est important de distinguer l'eau disponible de l'eau accessible par les racines (Droogers et al., 1997). L'eau accessible par les racines dépend principalement de la qualité de la structure du sol, qui bien que non caractérisée dans l'estimation du RU, peut conditionner l'accessibilité de l'eau selon son degré de compacité.

Cependant, il n'existe pas à l'heure actuelle de consensus sur la mesure de la profondeur d'enracinement. Cette mesure consiste soit à mesurer la racine visible la plus profonde ou bien estimer une valeur correspondant à 95% de la distribution cumulative des racines (Combres et al., 1999) (zone verte de la Figure 1.1).

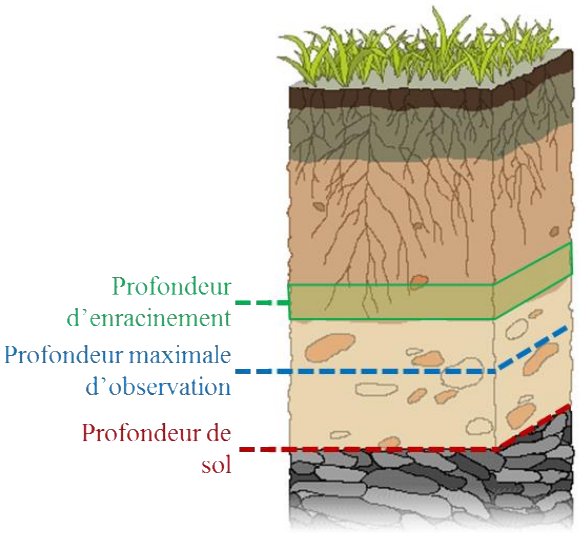


Figure 1.1. Représentation synthétique des différentes notions de profondeur

I.2.2. La densité apparente

La densité apparente de sol correspond au poids sec d'un volume de sol non remanié (structure préservée). L'intérêt de mesurer la densité apparente sur un échantillon non remanié est de pouvoir prendre en compte les fractions solides ainsi que la porosité. En raison des propriétés de gonflement ou de retrait des matériaux pédologiques, la densité apparente varie selon le degré d'humidité de l'échantillon. De ce fait, elle est déterminée à un état standard qui est l'état sec (séchage à 105°C).

Les méthodes de mesures de la densité apparente sont multiples et se répartissent en deux catégories : i) les méthodes destructives avec prélèvements et détermination du poids et du volume (non reproductible au même endroit), et ii) les méthodes non destructives sans prélèvements avec une détermination directe de la densité apparente (reproductible au même endroit) par gammamétrie. L'estimation de la densité apparente nécessite la mesure du poids de l'échantillon et de son volume apparent. Si la pesée de l'échantillon est relativement simple à réaliser, la détermination du volume apparent est plus compliquée notamment en fonction de : la nature ou l'état du sol (présence d'éléments grossiers, racines, fissures, indurations, etc.), aux perturbations liées au prélèvement (tassements, lissages), aux dimensions relatives des pores et des fissures ainsi qu'aux dimensions des volumes considérés (un volume plus important est susceptible de prendre en compte une plus grande porosité). Dans la majorité des cas, les méthodes destructives sont privilégiées car moins couteuses. Cependant, la densité apparente est souvent une propriété absente des bases de données pédologiques car son protocole de mesure est long et exigeant (plusieurs mesures parfois nécessaires pour les méthodes avec prélèvements).

I.2.3. La teneur en éléments grossiers

Les éléments grossiers correspondent aux constituants minéraux du sol de diamètre supérieur à 2 mm. Souvent négligés dans l'estimation du RU, les éléments grossiers de par leur volume d'occupation dans le sol et par leur nature lithologique, modulent plus ou moins la capacité de rétention du sol. En considérant uniquement le volume occupé par les éléments grossiers pour un horizon donné, un sol caillouteux présente un RU plus faible qu'un sol non caillouteux. Dans ce cas, ne pas prendre en compte la pierrosité dans l'estimation du RU tend à le surestimer (Cousin et al., 2003). Quand ils sont utilisés, les éléments grossiers sont caractérisés par leur abondance et comme étant des éléments inertes (Berger, 1976 ; Gras, 1994). Cependant, de nombreuses études (Tetegan et al., 2011 ; Cousin et al., 2003 ; Algayer et al., 2020) ont montré que la pétrographie des éléments grossiers joue un rôle important sur leur capacité de rétention. En effet, l'eau retenue par les éléments grossiers peuvent représenter jusqu'à 60% (pour un calcaire altéré) du RU pour un horizon donné (Tetegan et al., 2011). De plus, Algayer et al. (2020) ont également mis en évidence une meilleure corrélation entre le RU estimé et un indice de croissance des arbres quand la pierrosité est considérée dans l'estimation du RU.

Pour mesurer la teneur volumique en éléments grossiers, ces derniers sont d'abord mesurés pondéralement, après lavage et séchage, par rapport au poids de l'échantillon total séché (éléments grossiers + terre fine). Ensuite, les densités apparentes déterminées pour les éléments grossiers et la terre fine sont utilisées pour déterminer les volumes de chaque phase du sol et enfin d'estimer la proportion volumétrique des éléments grossiers.

I.2.4. Les humidités caractéristiques de sol

I.2.4.1. L'humidité à la capacité au champ

La notion d'humidité à la capacité au champ (CC) a été introduite par Briggs et MacLane (1910). Cette humidité correspond à la teneur en eau maximale qu'un sol peut stocker et restituer à la plante. Plus précisément, cette teneur en eau est obtenue après réessuyage du sol, c'est-à-dire quand l'écoulement de l'eau gravitaire faisant suite à un état de saturation du sol (forte pluie, irrigation) est terminé (Figure 1.2).

Le temps estimé du réessuyage du sol fait débat dans la communauté scientifique. Plusieurs ordres de grandeurs sont fournis dans la littérature : 2 à 3 jours selon la définition du Soil Science Society of America, Ratliff et al. (1983) augmente la durée maximale à 12 jours, Jabro et al. (2009) préconisent une durée entre 50 et 450 heures quand Assouline et Or (2014) estiment

que la valeur de la capacité au champ peut être déterminée après quelques heures ou plusieurs mois. Pour pallier le manque de consensus sur l'estimation d'un temps de réessuyage standard, il a été proposé par plusieurs auteurs d'évaluer l'humidité CC selon une pression matricielle. En 1947, Colman proposa la valeur de -33 kPa, qui demeure la valeur la plus utilisée (Nachabe, 1998), malgré quelques valeurs alternatives proposées (-20 kPa (Salter et Haworth, 1961), -10kPa (Romano et Santini, 2002) et -5 kPa (Nemes et al., 2011)).

I.2.4.2. L'humidité au point de flétrissement permanent

L'humidité au point de flétrissement permanent (PFP) correspond à la teneur en eau pour laquelle la plante est en déficit hydrique car elle n'est plus en mesure de prélever l'eau dont elle a besoin. La force de succion n'étant plus suffisante pour prélever l'eau du sol, la plante atteint progressivement le point de flétrissement permanent (Briggs et Shantz, 1912) (Figure 1.2).

Cette définition orientée biologie a été affiné au cours du temps pour estimer la valeur de l'humidité PFP selon des conditions standardisées (espèce de plante, faible évapotranspiration, température de la zone racinaire constante, etc.) (Furr and Reeve, 1945 ; Taylor et al., 1934). Cependant, cette humidité caractéristique est aussi très dépendante des profils du sol et des racines, du niveau de développement de la plante, du taux de transpiration et de l'ajustement osmotique. En tenant compte de l'ensemble des paramètres présentés, il résulte que l'humidité PFP ne peut être une valeur unique mais plutôt une gamme de valeurs (Taylor et al., 1934). Une évaluation de cette humidité selon sa définition biologique est relativement fastidieuse et exigeante. Pour cela, elle a été remplacée par une mesure de la pression matricielle, telle que réalisée pour l'humidité CC. La valeur de la pression matricielle pour l'humidité PFP communément utilisée est de -1500 kPa (Richards et al., 1949 ; Richards et Weaver, 1943, 1944 ; Sykes, 1964, Ratliff et al., 1983).

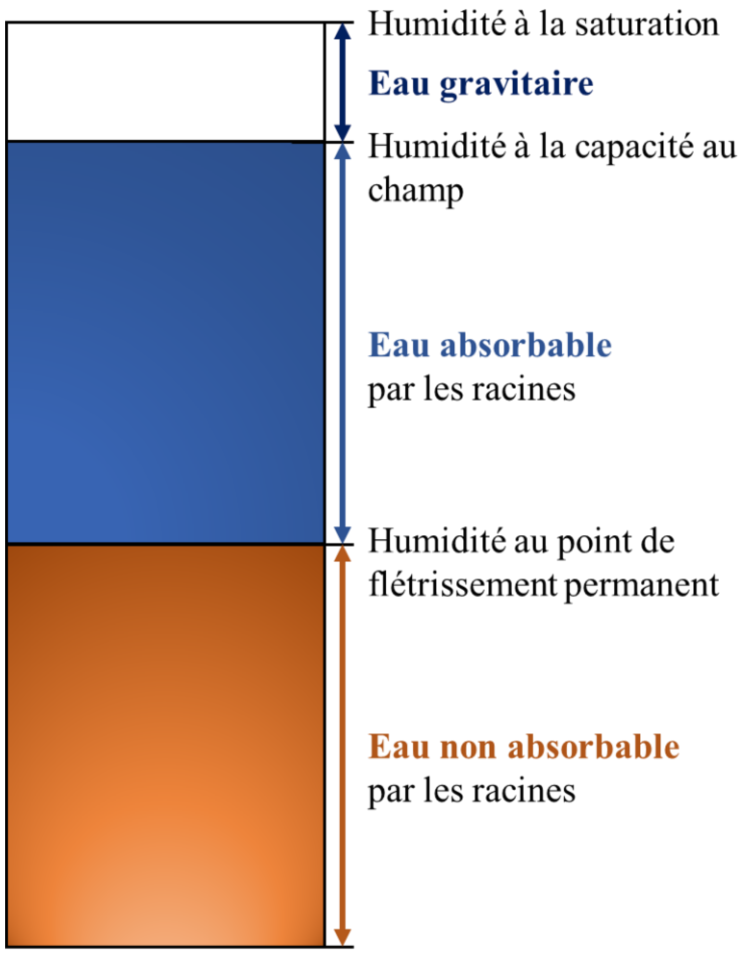


Figure 1.2. Représentation synthétique des humidités à la capacité au champ et au point de flétrissement permanent

I.2.4.3. Les mesures des humidités à la capacité au champ et au point de flétrissement permanent

Les mesures des humidités caractéristiques de sol sont généralement réalisées en laboratoire sur des échantillons remaniés ou non remaniés, prélevés pour chaque horizon du profil de sol. Ces échantillons sont ensuite placés dans le dispositif de presse à plaques de Richards et mis à l'équilibre par l'application des valeurs respectives de leur pression matricielle (-33 kPa pour l'humidité CC et -1500 kPa pour l'humidité PFP) pendant 5 à 7 jours.

En sortie de ce dispositif, les échantillons sont pesés pour estimer le poids de la teneur en eau. Ils sont ensuite séchés à 105°C pendant 2 jours avant d'être une nouvelle fois pesés pour mesurer leur poids sec, afin d'estimer la teneur en eau massique. Cependant pour estimer le RU, les humidités CC et PFP sont exprimées en volumétriques, ce qui nécessite l'estimation de la densité apparente qui peut être mesurée à partir des échantillons utilisés pour déterminer les humidités CC et PFP.

I.2.4.4. Les fonctions de pédotransfert

Les fonctions de pédotransferts (FPT) ont pour but de prédire des propriétés de sol difficiles et/ou couteuses à mesurer (ex : propriétés hydrodynamiques de sols) à partir de propriétés de sols plus accessibles (ex : fractions granulométriques) (Bouma, 1989). Les FPT sont principalement utilisées dans la prédiction des propriétés hydrodynamiques du sol (Pachepsky et Rawls, 2004 ; Wösten et al., 2001 ; Dharumarajan et al., 2019 ; Román Dobarco et al., 2019) et sont réparties en deux catégories : i) les FPT ponctuelles qui estiment des humidités de sol à une pression matricielle fixe, et ii) les FPT paramétriques où les paramètres d'un modèle de la courbe de rétention sont estimés.

La texture ou les fractions granulométriques sont très souvent considérées comme les données prédictives utilisées dans les FPT (Jamagne et al., 1977 ; Rawls et al., 1982). D'autres propriétés de sols sont également utilisées, pour améliorer la qualité de prédiction des FPT, telles que : la densité apparente (Al Majou et al., 2008 ; Vereecken et al., 1989) ou encore la teneur en carbone organique (Arrouays et Jamagne, 1993 ; Batjes, 1996).

Les FPT sont principalement élaborées par régression linéaire ou non linéaire (Romano et Palladino, 2002) inclues dans une procédure itérative permettant l'ajustement du modèle et la sélection des propriétés les plus adaptées à prédire la variable cible (Vereecken et Herbst, 2004).

Dans le but de minimiser l'incertitude introduite par une FPT, il est recommandé d'utiliser une FPT élaborée sur une région présentant des caractéristiques pédologiques et paysagère similaires (Wösten et al., 2001) et de rassembler les données selon plusieurs critères, tels que : le type d'horizon et sa position au sein du profil, la nature de la roche-mère, la classe texturale, en groupes texturales ou groupes de sols (Al Majou et al., 2008 ; Batjes, 2016 ; Bruand et al., 2004 ; Jamagne et al., 1997 ; Román Dobarco et al., 2019). Les prérequis pour sélectionner une FPT déjà publiée sont : i) que la FPT soit fournie avec son incertitude et ii) que les caractéristiques de sols de la zone étudiée soit du même ordre de grandeur que celles utilisées pour développer la FPT (Minasny et Hertemink, 2011). Un moyen de s'en assurer est l'utilisation des distances de Mahalanobis, qui permet d'évaluer l'applicabilité de la FPT (Román Dobarco et al., 2019b).

I.3. Choix d'estimation du réservoir utile et de ses composants dans la thèse

Nous détaillons ici les modalités sélectionnées dans l'estimation du RU dans cette thèse. Le RU sera estimé sur la profondeur maximale de sol, soit jusqu'au contact lithique ou paralithique. Cependant, afin de prendre en compte la profondeur racinaire des plantes, nous avons choisi de compléter cette estimation avec des RU estimés à différentes profondeurs fixes (30 cm, 60 cm et 100 cm), laissant le choix à l'utilisateur de la profondeur la plus adaptée à la culture en place. Concernant la prise en compte des éléments grossiers dans l'estimation du RU, l'abondance sera intégrée mais, ne disposant pas d'une quantité suffisante d'information permettant de tenir compte de leur nature lithologique, les éléments grossiers seront considérés comme inertes. Comme présentée en sections I.2.4.2 et I.2.4.3, les mesures en laboratoire des humidités caractéristiques de sols suivent un protocole fastidieux et long qui a pour conséquence de limiter sévèrement le nombre de sites disponibles avec cette mesure du RU. Notre objectif de spatialisation nécessitant un nombre important de sites de sols, nous utiliserons alors des estimations des humidités volumiques à la capacité et au point de flétrissement permanent issues de l'utilisation de fonction de pédotransferts.

II. La cartographie numérique des sols

Le but de cette thèse est de mobiliser des estimations locales de RU pour cartographier une zone donnée. Pour ce faire, nous plaçons ces travaux dans le contexte de la cartographie des sols à modélisation statistique (CSMS) qui se réfèrera par la suite à la cartographie numérique des sols (CNS). Par la suite, nous détaillerons les principes généraux de la cartographie numérique des sols puis les applications existantes et, enfin les principes généraux de la cartographie numérique des sols appliqués au RU.

II.1. Principes généraux de la cartographie numérique des sols

II.1.1. Le principe général de la cartographie numérique des sols

Historiquement, la cartographie des sols s'est construite sur la connaissance et la compréhension des sols acquise par une expérience collective de terrain, menée par les pédologues au fil des années. Cette expérience a mené à l'identification des facteurs environnementaux (climat, matériau parental, relief et êtres vivants) en lien avec la pédogénèse (Figure 1.3).

Figure 1.3. Schéma des facteurs environnementaux en lien avec la formation des sols

Ces relations sont rapidement théorisées, par Jenny (1941), sous la forme d'une équation conceptuelle de la formation des sols dit « CLORPT » permettant d'identifier l'impact de chaque facteur sur la formation des sols, intégré dans une dynamique temporelle :

$$S = f(c, o, r, p, t) \qquad \text{(Eq.1.2.)}$$

Où c est le climat, o les êtres vivants, r le relief, p le matériau parental et t le temps. Les premières cartographies sont alors basées sur une approche empirique et qualitative de ces paramètres selon un expert pédologue. La répartition des sols dans le paysage est définie par l'expert sur le terrain puis intégrée dans un modèle environnemental pour être spatialisée (Lagacherie et Legros, 1992).

A la fin des années 1970, les toutes premières représentations numériques des sols sont issues des travaux de digitalisation de l'information pédologique dans le but de les intégrer aux systèmes d'informations géographiques (SIG) (Webster et al., 1979). La transition des représentations numériques des sols vers la cartographie numérique des sols se fait vers la fin des années 1980 par les travaux fondateurs de cette discipline (Lagacherie et al., 1989 ; Merot et al., 1995, McKenzie et Austin, 1993 ; Bell et al., 1992). Dans le même temps, des avancées

méthodologiques importantes (méthodes d'observation et de mesures des données environnementales, augmentation de la capacité de calcul) ont permis d'amorcer la cartographie numérique des sols via des méthodes empiriques et déterministes. L'utilisation de ces modèles permet d'ajouter à l'équation conceptuelle de Jenny une dimension spatiale, la notion d'incertitude ainsi que l'utilisation de données environnementales spatiales numérisées. Ces nouvelles notions sont ensuite formalisées dans le modèle *scorpan* par McBratney et al. (2003) (Eq .1.3.), considéré depuis comme le modèle fondateur de la cartographie numérique des sols :

$$S = f(s, c, o, r, p, a, n) + \varepsilon \qquad \text{(Eq. 1.3.)}$$

Où S est une classe ou un attribut d'un sol, s les propriétés du sol, c les variables climatologiques, o les variables représentant les influences biotiques, r les indicateurs du relief, p le matériau parental, a le temps, n le position géographique et ε l'erreur du modèle.

En d'autres termes, le principe général de la cartographie numérique des sols (Figure 1.4) est de prédire des classes ou des propriétés de sol en utilisant d'une part les données pédologiques disponibles sur la zone à étudier et, d'autre part, les données spatiales représentant les éléments du paysage en relation avec la formation des sols. Pour cela des fonctions de prédictions sont utilisées. Les fonctions de prédictions sont totalement explicites, peuvent être calibrées, validées et donnent une estimation de l'incertitude de prédiction. L'intérêt de la cartographie numérique des sols est que ces modèles sont reproductibles, transférables et peuvent être constamment perfectionnés selon les avancées méthodologiques ou l'apport de nouvelles données.

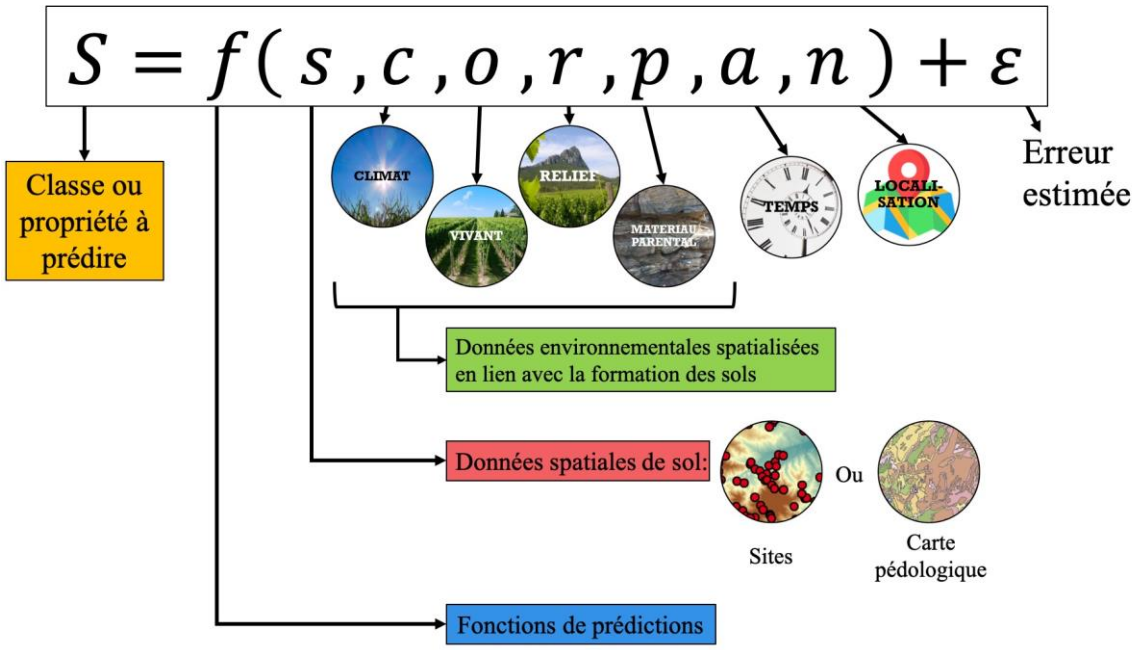


Figure 1.4. Principe générale de la cartographie numérique des sols (d'après McBratney et al., 2003)

II.1.2. Les grandes familles de modèles de la cartographie numérique des sols

La cartographie numérique des sols repose sur la construction de fonctions délivrant une prédiction de classes de sols ou de propriétés sur des sites inconnus (McBratney et al., 2003). De ce fait, la fonction de prédiction dépend également du type des données pédologiques disponibles (composant *s* du modèle *scorpan*).

Effectivement, les approches de la cartographie numérique des sols se déclinent en deux types d'approches selon le type de données pédologiques à prédire (classes vs propriété de sol) : i) une approche « *knowledge driven* » basée sur la connaissance pédologique déjà disponible et ii) une approche « *data driven* » qui utilise les données pédologiques issues d'échantillons par site (Figure 1.5).

L'approche « *knowledge driven* » est utilisée quand il existe sur la zone à étudier une expertise pédologique (issue de la connaissance d'un pédologue sur cette zone) ou d'une carte existante pédologique. Dans le premier cas, l'utilisation du savoir du pédologue pour prédire les classes de sol se fait via l'intégration de savoir sous la forme d'un arbre de décision dans des SIG (McKenzie et Gallant, 2006 ; Cole et Boettinger, 2007) ou bien de façon plus élaborée dans un modèle d'apprentissage basé sur la logique floue (Zhu, 1997 ; Cazeimer, 1999). Quand l'expertise pédologique disponible est une carte pédologique, deux types de méthodologies sont distinguées.

Le premier type de méthodologie consistant à affiner la résolution spatiale des prédictions est lui-même décliné en deux méthodes. La première consiste à dériver des estimations de propriétés de sol à partir de l'information pédologique des unités cartographiques. Pour les unités de sols simples correspondant à un seul type de sol, l'estimation se fait à partir des profils de sols représentatifs de l'unité de sol (Leenhardt et al., 1994) ou le cas échant à partir de l'information textuelle du type de sol permettant d'établir une distribution de possibilité (Cazemier et al., 2001). Pour les unités de sols complexes contenant plusieurs typologies de sol, la prédiction d'une propriété est plus fastidieuse. Les approches sont multiples : moyennes pondérées par les surfaces d'unités typologiques (Bliss et al., 1995 ; Nauman et al., 2012) ou désagrégation des unités cartographiques complexes tels que les modèles algorithmiques DSMART (Nauman et al., 2012 ; Odgers et al., 2014) ou l'area-to-point-kriging (Rawlins et al., 2009 ; Kerry et al., 2012) ou encore l'application de la logique floue aux objets pédologiques (unités cartographiques ou profils de sol) (Lagacherie et al., 1996 ; McBratney et Odey, 1997)

ou sur la détermination de classes pédologiques continues et des variables auxiliaires (Bezdek, 1974 ; Burrough et al., 1992 ; McBratney and de Gruijter, 1992).

Le second type de méthodologie s'appuie sur la formalisation de l'expertise acquise d'un secteur de référence afin de permettre son extrapolation à des superficies plus importantes présentant un contexte naturel similaire (pédologique, agriculture, géologie) (Lagacherie, 1992, Lagacherie et al., 1995 ; Lagacherie et Voltz, 2000 ; Bui et Moran, 2003 ; Salvador et al., 1997). Pour ce faire, il est essentiel de comprendre et d'établir les règles pédologiques en formalisant l'étude des relations sols-paysages et la compréhension de leur incidence sur la répartition des classes de sols (Bui et Moran, 2001 ; Mallavan et al., 2010).

Cependant, en dépit de la connaissance pédologiques des informations anciennes, la principale difficulté de l'utilisation des méthodes présentées selon une approche « *knowledge driven* » résides dans l'estimation de l'incertitude associée à la prédiction de classes dans les zones où les limites cartographiques précises des ensembles sont inconnues (Lagacherie et al., 2013). Ces méthodes sont également limitées par la rareté, passée et présente, de l'expertise pédologique disponible sous une forme ou une autre, ce qui laisse des régions potentiellement sans information à valoriser.

L'approche « *data driven* » s'appuie sur la disponibilité, sur l'étendue du cas d'étude, des données pédologiques (classe ou valeur de la propriété de sol) renseignées par sites et des données environnementales, appelées aussi covariables de sol. L'objectif de cette approche est de modéliser une variabilité régionalisée. Il s'agit d'une fonction mathématique estimant la variance d'une propriété de sol ou d'une variable environnementale selon un continuum espace/temps (Goovearts, 1999 ; Wackernagel, 2003). De nos jours, une multitude d'outils méthodologiques sont déjà disponibles et réparties en deux catégories : i) les méthodes pédo-statistiques et ii) les méthodes géostatistiques (Lagacherie et al., 2013).

Les méthodes pédo-statistiques sont principalement composées d'algorithmes de : fouilles de données (arbres de régression et de classifications) (McKenzie et Ryan, 1999 ; Breiman, 2001 ; Henderson, 2005), ou de réseau de neurones artificiels (Schaap et al., 1998), ou bien de modèles linéaires généralisés comme les régressions linéaires multiples ou les splines (Voltz et Webster, 1990 ; Bourennane et al., 2000). Le fonctionnement des algorithmes de fouilles de données, souvent mobilisés en cartographie numérique des sols, s'appuie sur la calibration de ce modèle à partir d'un ensemble d'apprentissage constitué de données pédologiques et

environnementales connues pour chaque site (Lagacherie et al., 2013). Ces modèles calibrés sont ensuite appliqués sur des sites non renseignés afin de prédire la classe ou la propriété de sol ainsi que son incertitude. Afin de permettre la spatialisation, la couverture du secteur d'étude par covariables de sol doit être intégrale (c : climat, o : interactions biotiques, r : relief, p matériau parental du modèle *scorpan,* Equation 1.3.). Par ailleurs, souvent considéré comme un algorithme non-spatial, de récents travaux (Hengl et al., 2018) ont permis d'inclure directement la dimension spatiale n du modèle *scorpan* à l'ensemble d'apprentissage.

Les méthodes géostatistiques visent à prédire une propriété de sol sur l'ensemble du secteur d'étude, en utilisant uniquement les composants sol (s), temps (a) et espace (n) du modèle *scorpan* ainsi qu'une erreur liée à l'utilisation du modèle (ε). Le principe de ces méthodes est basé sur l'analyse variographique monovariée pour une seule propriété de sol unique et multivariée pour un ensemble de propriétés (Wackernagel, 2003). L'analyse variographique est l'élaboration d'un variogramme qui représente la variance entre couples de points selon leur classe d'interdistance. Cela permet de prendre en compte la structuration spatiale de la propriété à prédire. La méthode la plus utilisée en approche monovariée est le krigeage ordinaire qui est un estimateur linéaire non-biaisé construit à partir d'un modèle appliqué au variogramme. Cet estimateur peut être ajusté par les moindres carrées ou par la méthode de maximisation de la vraisemblance REML (Laslett et McBratney, 1990 ; Lark et al., 2006). De plus, cette méthode est capable de minimiser la variance d'estimation ponctuelle en supposant que la propriété de sol à prédire est stationnaire (Cressie, 1990 ; Goovaerts, 1999). L'hypothèse de stationnarité signifie que la covariance entre les sites de la propriété régionalisée est invariante par translation et qu'elle ne dépend uniquement de la distance interpoints. Le krigeage ordinaire peut également être utilisé en complément des méthodes pédo-statistiques dans le but d'estimer spatialement les erreurs intrinsèques à l'utilisation de ces modèles, également appelée méthode régression krigeage ou krigeage par dérivé externe (Bourennane et al., 2000 ; Nussbaum et al., 2013). Pour une approche multivariée, les méthodes de krigeage disponibles sont plus complexes avec notamment un ajustement d'une matrice de co-régionalisation, permettant d'identifier la structuration spatiale entre l'échantillonnage des données environnementales et de la propriété cible. Les principales méthodes sont le co-krigeage (Wackernagel, 2003 ; Ciampalini et al., 2012) et l'analyse factorielle krigeante (Goovaerts, 1992 ; Bourennane et al., 2003).

Les deux grands ensembles d'approches présentées dans cette section présente des avantages et des inconvénients relatifs à leur utilisation. L'approche portée uniquement sur la

connaissance pédologique à l'avantage d'intégrer la non-stationnarité des relations sols-paysages afin d'élaborer des modèles cognitifs de répartition des sols dans le paysage, fondés sur le savoir du pédologue. Toutefois, l'utilisation de cette méthode ne permet pas d'expliquer la variabilité à fine échelle d'une propriété de sol, notamment en raison d'une résolution trop grossière des cartes pédologiques existante. De plus, les méthodologies utilisées lors de l'élaboration de ces cartes sont susceptibles de générer une incertitude importante en voulant désagréger ses unités cartographiques. Cette incertitude est également liée à la qualité de l'expertise du pédologue, possiblement biaisée se répercutant dans la conception de la carte pédologique. En revanche, une approche statistique des données disponibles permet une application directe des données sans besoin d'une expertise pédologique. Basée sur une analyse statistique ponctuelle, cette approche n'est ni limitée par les contours des unités cartographiques, ni par les résolutions de variabilités considérées, mais son applicabilité est fortement conditionnée par la densité d'échantillonnage et la stationnarité des variabilités des propriétés de sol. Enfin, la mise en place d'approches multi-variables se révèle plus compliquée par la nécessité d'utiliser des données disponibles sur l'ensemble du cas d'étude.

Chaque grand ensemble d'approche, de par leurs spécificités, présente des avantages et des inconvénients selon le cas d'étude et notamment les disponibilités des données environnementales et pédologiques. Cependant, ces grands ensembles d'approches n'en sont pas pour le moins complémentaires dans certaines situations. C'est pourquoi il est nécessaire d'effectuer au préalable une étude de la disponibilité des données afin de déterminer quelle est la stratégie à initier et comment associer les modèles pédo-statistiques et géostatiques pour obtenir une méthode optimale de cartographie numérique des sols.

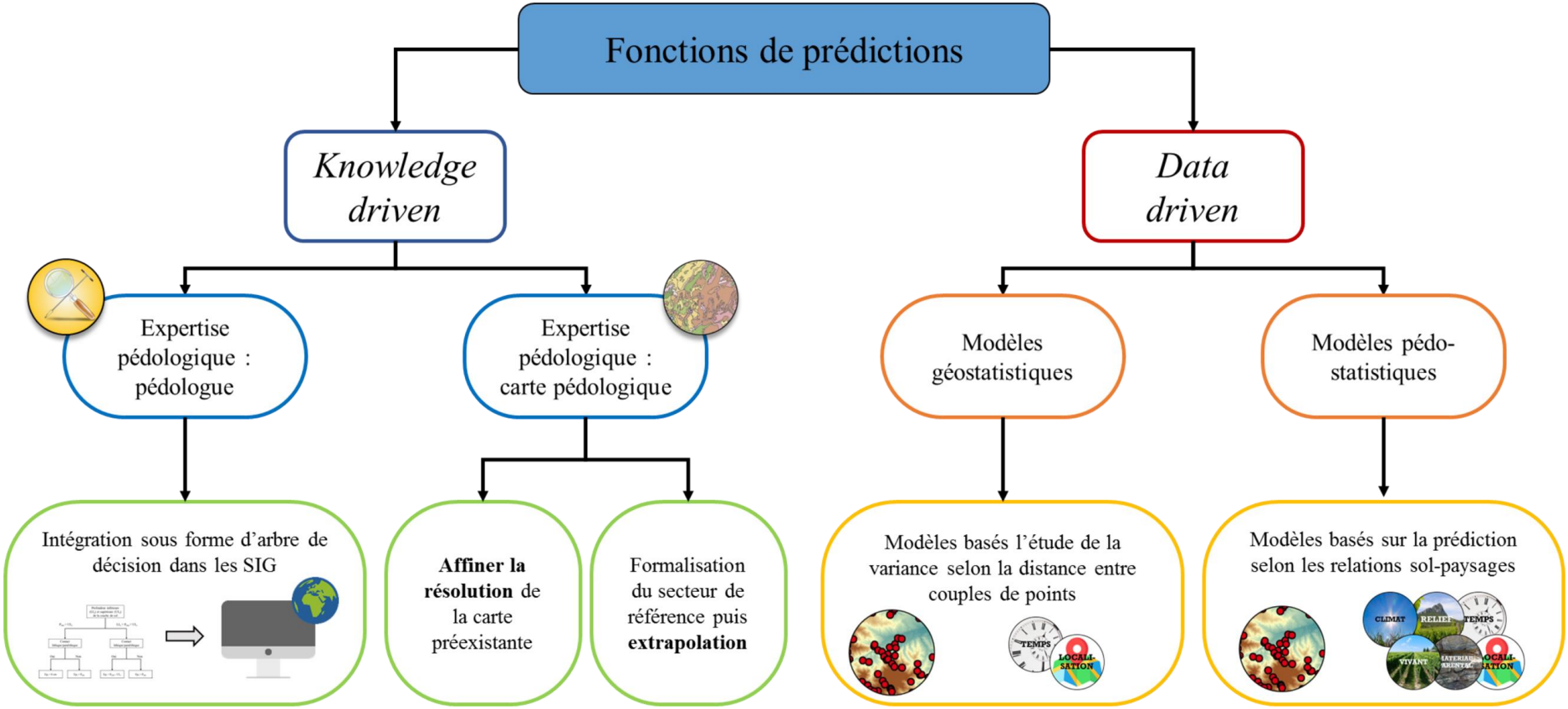


Figure 1.5. Schéma synthétique des différentes grandes familles de modèles de cartographie numérique des sols

II.1.3. Incertitudes en cartographie numérique des sols

II.1.3.1. La notion d'incertitude

Les cartographies de sols sont des représentations simplifiées de la variabilité beaucoup plus complexe des sols. Par conséquent, une prédiction d'une propriété issue de ces cartes présente une erreur importante estimée par la différence entre la valeur prédite et la vraie valeur de la propriété de sol. Cependant, n'étant pas possible de mesurer systématiquement cette erreur en tout point de l'espace, le caractère incertain quant aux vraies valeurs subsistent. L'incertitude peut se traduire par l'état d'esprit d'incertitude qu'un scientifique exprime par rapport à la précision de son modèle (Heuvelink, 2014). Avoir accès à une estimation et à une représentation spatiale de cette incertitude est tout aussi important que l'accès aux prédictions d'une propriété de sol (Wadoux et al., 2018). L'incertitude est un paramètre important pour les décideurs locaux ainsi que pour les politiques d'aménagement du territoire pouvant contribuer à des prises de décisions selon l'utilité et les limites des cartes numériques.

II.1.3.2. Estimations de l'incertitude

En cartographie numérique des sols, l'incertitude peut être plus ou moins bien définie par une analyse et une quantification des erreurs issues du modèle. Dans le cadre des spécifications *GlobalSoilMap*, il a été décidé que l'incertitude serait exprimée sous la forme d'un intervalle de confiance (voir explications section II.2). L'intervalle de confiance se définit comme un intervalle de valeurs censé inclure la vraie valeur de la variable cible (Arrouays et al., 2014).

Un modèle de cartographie numérique de sols comporte principalement trois sources d'erreurs : i) l'erreur de mesure sur les données de sol utilisées (biais de l'opérateur et/ou de l'appareil de mesure), ii) l'erreur liée une faible relation entre les variables prédictives et la variable cible, et iii) l'erreur intrinsèque à la structure du modèle (Heuvelink, 2014). L'erreur de mesure sur la variable est généralement quantifiée selon les normes du protocole de mesure. Il est possible, toutefois, qu'un biais soit généré par une dérive d'un instrument de mesure ou par une dérive engendrée des différents techniciens. L'erreur liée à une faible relation entre variables prédictives et la variable cible peut être explorée par une analyse de sensibilité afin d'estimer la propagation d'erreur liée aux données d'entrée du modèle (Cardenas et Malherbe, 2003).

II.1.3.3. Les indicateurs d'incertitude

L'erreur générée par l'utilisation du modèle est quantifiable par une validation du modèle. Pour cela, la validation s'appuie sur l'utilisation d'indicateurs statistiques pour mesurer les erreurs en sortie de modèle.

On distingue deux types d'indicateurs utilisés en validation : i) les indicateurs d'incertitude de performances voués à évaluer la capacité prédictive du modèle et ii) les indicateurs évaluant l'erreur intrinsèque du modèle.

Les indicateurs utilisés pour mesurer les performances de prédictions du modèle sont principalement : l'erreur moyenne (ME), l'erreur quadratique moyenne (MSE), la racine de l'erreur quadratique moyenne (RMSE), le pourcentage de variance expliquée (SS_{MSE}) présentant la même interprétation que le coefficient de détermination (R^2). Dans le cadre d'une étude visant à prédire des classes de sols sur des sites, la validation consistera à déterminer un taux de bon classement entre les classes prédites et les classes issues du jeu de données initial. Dans cette thèse, les indicateurs statistiques conservés sont l'erreur moyenne, la racine de l'erreur quadratique moyenne et le pourcentage de variance expliquée.

$$ME = \frac{1}{n}\sum_{i=1}^{n}(Pred_i - Obs_i) \qquad \text{(Eq. 1.4.)}$$

Où n est le nombre d'échantillons, $Pred_i$ la prédiction de la variable et Obs_i la valeur de l'échantillon.

La racine de l'erreur quadratique moyenne est un indicateur estimant la performance du modèle et permettant de comparer deux estimateurs différents, notamment dans le cas où il y a présence d'un biais par l'un des estimateurs. Cet indicateur se calcul selon l'équation suivante :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Pred_i - Obs_i)^2} \qquad \text{(Eq. 1.5.)}$$

Le SS_{MSE} permet d'estimer la part de variance expliquée par le modèle en fonction de la variance totale des données d'entrée calculée selon la formule suivante :

$$SS_{MSE} = 1 - \frac{\sum_{i=1}^{n}(Pred_i - Obs_i)^2}{\sum_{i=1}^{n}(Obs_i - \widehat{Obs_i})^2} \qquad \text{(Eq. 1.6.)}$$

Où $\widehat{Obs_i}$ est la valeur moyenne des échantillons.

Afin d'évaluer les erreurs intrinsèques à l'utilisation du modèle, l'indicateur principalement utilisé dans la littérature exprime la probabilité que tous les échantillons soient compris dans l'intervalle de confiance (PICP, Shrestha and Solomatine, 2006). Cet indicateur s'exprime selon l'équation suivante :

$$PICP = \frac{count(LPL_i \leq Obs_i \leq UPL_i)}{n} \times 100 \qquad \text{(Eq. 1.7.)}$$

Où n est le nombre d'échantillon, Obs_i la valeur de l'échantillon, LPL_i la borne inférieure de l'intervalle de confiance et UPL_i la borne supérieure de l'intervalle de confiance.

II.1.3.4. Les méthodes d'estimations des indicateurs d'incertitude

Afin de pouvoir procéder à une évaluation rigoureuse de l'incertitude, deux approches sont utilisées (Heuvelink, 2013). La première approche dite « model-based » est basée sur la capacité du modèle à fournir sa propre incertitude. L'avantage pour les modèles géostatistiques, basés sur le krigeage, est de fournir une cartographie, aussi bien ponctuelle que surfacique, de la variance de krigeage associée à la prédiction de la variable d'intérêt (Goovaerts, 2001 ; Chiles and Delfiner, 2009). Pour le krigeage ordinaire, en considérant que la distribution de la variance suit une loi normale, un intervalle de confiance peut être estimé pour chaque point d'estimation par application du théorème de la centrale limitée (Rocha et Yamamoto, 2000). Pour les modèles pédo-statistiques, essentiellement les modèles de fouilles de données, type forêt aléatoires quantiles. La méthodologie introduite par Vaysse et Lagacherie (2017) consiste à utiliser les quantiles de distributions pour estimer les bornes inférieures et supérieures de l'intervalle de confiance. La particularité de cette approche réside dans la quantification de l'incertitude aussi bien pour des variables respectant une loi normale de distribution que pour des variables ne la respectant pas. Les principaux inconvénients de cette approche sont : i) l'évaluation de l'incertitude réalisable sous certaines hypothèses (ex : stationnarité pour les modèles géostatistiques) et ii) l'utilisation en validation des données utilisées en calibration ce qui constitue une validation *a priori* susceptible d'induire une sous-estimation de l'incertitude.

La seconde approche dite « model-free » se détache de ces restrictions en évaluant l'incertitude en comparant les prédictions et la valeur observée de la variable sur des sites indépendants sélectionnés par échantillonnage aléatoire.

Selon cette approche la validation peut être réalisée selon deux modalités dépendant de la disponibilité des données de la variable à prédire : la validation croisée et la validation indépendante.

Si le nombre d'observation constituant le jeu de données est faible, la validation croisée peut être choisie. Le principe de la validation croisée consiste à retirer des individus du modèle pour ensuite tester le modèle sur ces individus. Plusieurs méthodes de validations croisées sont utilisées et répertoriées selon leur capacité d'extraction. La validation croisée « leave-one-out » partitionne l'ensemble d'individus en autant de groupes de validation. Cette méthodologie vise à tester le modèle sur l'individu soustrait de la calibration. Le test est ainsi répété sur l'ensemble des individus disponibles avant calcul des indicateurs statistiques. La validation croisée « k-fold cross-validation » partitionne l'ensemble d'individus en *k* sous-ensembles. Le test est ensuite appliqué aux individus du sous-ensemble *k* préalablement retiré de la calibration du modèle. Cette procédure est réitérée jusqu'à ce que tous les sous-ensembles soient testés. Ce type de validation ne permet pas à un échantillon de faire partie deux fois de sous-ensembles différents. La validation « out-of-bag » qui est une variation de la validation « k-fold cross-validation » consiste à retirer aléatoirement un groupe d'observations de la calibration du modèle puis de tester le modèle sur ce groupe de données. Cette étape est répétée *n* fois avec possibilité de remise des observations dans le groupe de validation. Les indicateurs statistiques sont ensuite appliqués sur ces groupes d'observations.

La validation indépendante est utilisée si un jeu de données externe au jeu de calibration du modèle est disponible. Dans ce cas, le modèle calibré est testé sur les données de validation. Les indicateurs statistiques sont ensuite calculés à partir des données prédites.

De la même façon que l'approche « model-based », l'ensemble de ces méthodes d'évaluation de l'incertitude permettent également d'estimer une incertitude à partir des données de calibration qui est une estimation de l'incertitude *a priori*. Cependant, cette approche peut être biaisée par un ajustement du modèle trop spécifique aux données de calibration (« overfitting ») et avoir tendance à sous-estimer l'incertitude en situation de prédiction. Il est donc plus rigoureux de réaliser une validation externe afin d'analyser les possibles biais.

II.2. Les produits de la cartographie numérique des sols existants

La cartographie des sols, se définissant comme une discipline auxiliaire de l'étude des sols, est désormais passée en phase opérationnelle permettant de produire des cartes numériques des propriétés pérennes de sol. Les principaux utilisateurs de ces cartes sont : la communauté scientifique ainsi que les responsables politiques et les décideurs locaux (collectivité territoriale, aménageurs du territoire, etc.). Cet objectif d'opérationnalité a été amorcé via le lancement du programme de cartographie numérique des sols international *GlobalSoilMap* (Arrouays et al., 2014) visant à produire des cartes numériques de sols selon les spécificités suivantes :

- Prédire les propriétés pérennes du sol : profondeur de la roche ; profondeur d'enracinement ; teneur en argile, limon et sable ; teneur en éléments grossiers ; teneur en matière organique ; pH des sols ; capacité d'échange cationique ; conductivité du sol ; densité apparente ; réservoir utile,
- Prédire ces propriétés selon les 6 intervalles de profondeurs : 0-5 cm, 5-15 cm, 15-30 cm, 30-60 cm, 60-100 cm et 100-200 cm,
- Prédire ces propriétés à une résolution de 3 arc-secondes (~ 90 m),
- Prédire l'incertitude de prédictions via un intervalle de confiance à 90% qui se trouve être le format correspondant aux besoins des modélisateurs qui utilisent les données pédologiques dans l'élaboration de modèles.

Le programme *GlobalSoilMap* (GSM) a été à l'origine d'une augmentation importante de cartographie numérique des sols ces dernières années. Les produits issus de ce programme se déclinent sur différentes étendues spatiales (régions, pays, continents, monde).

Le Tableau 1.1 présente un inventaire non-exhaustif des études de cartographies numériques des sols des propriétés composants le RU issus du *GlobalSoilMap*, à de multiples échelles. A l'échelle globale, les propriétés primaires de sols prédites par Hengl et al. (2014) à la résolution de 1 km x 1 km présentent une variance expliquée comprise entre 24% et 35% pour une densité d'un profil de sol tous les 1358 km². Pour les représentations cartographiques à l'échelle nationale, la densité de points augmente à 1 profil tous les 50 km² pour le Danemark (Adhikari et al., 2013) et 1 profil de sol tous les 19 km² pour la France (Mulder et al., 2016). La part de variance expliquée par les modèles varie entre 38% et 46% pour la France et entre 11% et 38% pour le Danemark. En passant à l'échelle régionale avec l'exemple des cartographies des propriétés primaires du Languedoc-Roussillon (Vaysse et Lagacherie, 2015), les performances sont comprises entre 6 % et 35 % pour une densité fluctuante entre 1 profil tous les 14 à 80 km²

selon la disponibilité des données. Enfin, à l'échelle locale, Nussbaum et al. (2018) ont obtenus des performances entre 9% et 31% de variance expliquée pour une densité comprise entre 2 et 7 profils au km², sur plusieurs terrains d'étude. Si les travaux CNS a échelle globale et nationale sont désormais acquises, le besoin d'obtenir une information plus précise à l'échelle régionale ou locale est réel (Arrouays et al., 2017). Cependant, il apparait que les faibles performances (maximum de variance expliquée à 46%) de prédictions des propriétés soient principalement dues à une faible densité des données pour alimenter les modèles CNS. De plus, la densité de données souvent insuffisante et répartie de façon hétérogène dans l'espace a été reconnue comme le principal facteur limitant les performances de prédictions.

Tableau 1.1. Tableau non exhaustif des études de cartographies numériques des sols sur les propriétés primaires de sols

Références	Propriétés prédites	Effectif des données	Densité (1/*n* km²)	Région	Echelle	Superficie (km²)	Modèle CNS	Validation	Var. expl. (%)
Hengl et al. (2014)	Argile	110 000	1358	Globale	Globale	149 400 000	RK	CV	24
	Limon	-	-	-	-	-	-		35
	Sable	-	-	-	-	-	-		24
	Elements grossiers	-	-	-	-	-	-		-
	Profondeur de la roche	-	-	-	-	-	-		-
	Densité apparente	-	-	-	-	-	-		32
Mulder et al. (2016)	Argile	28 537	19	France	Nationale	543 940	Cubist	CV	28
	Limon	28 581	19	-	-	-	-		36
	Sable	28 565	19	-	-	-	-		38
	Elements grossiers	56 179	10	-	-	-	-		22
	Profondeur de la roche	16 244	33	-	-	-	-		11
Adhikari et al. (2013)	Argile	850	50	Danemark	Nationale	42 924	Cubist	Ext	43
	Limon	850	-	-	-	-	-		45
	Sables (fins/grossiers)	850	-	-	-	-	-		38/46

Vaysse et Lagacherie (2015)	Argile	1 955	14	Languedoc-Roussillon	Régionale	27 236	RF	Ext	33
	Limon	1 894	14	-	-	-	-		27
	Sable	1 924	14	-	-	-	-		35
	Eléments grossiers	1 566	17	-	-	-	-		10
	Profondeur de roche	341	80	-	-	-	-		6
Nussbaum et al. (2018)	Argile	1 107 / 969 / -	0,24 / 0,15 / -	Greingensee/Berne/ZH forest	Locale	170 / 235 / 570	Lasso, georob, geoGAM, BRT, RF*, MA	Ext	23 / 24 / -
	Limon	1 111 / 974 / -	0,24 / 0,15 / -	-	-	-			9 / 21 / -
	Eléments grossiers	936 / 1 034 / -	0,23 / 0,18 / -	-	-	-			22 / 18 / -
	Profondeur de la roche	943 / 1 036 / -	0,23 / 0,18 / -	-	-	-	-		31 / 14 / -
	Densité apparente	- / - / 1 074	- / - / 0,47	-	-	-	-		- / - / 14

*: modèle sélectionné pour exprimer la variance expliquée ; CV : validation croisée ; Ext : validation indépendante

II.3. Cartographie numérique et réservoir utile

Bien que la cartographie numérique des sols soit passée en phase opérationnelle, les études portant sur l'estimation d'une propriété fonctionnelle telle que le réservoir utile sont proportionnellement faibles en comparaison des études portant sur ses composantes, en dépit de son importance pour les modélisateurs et décideurs locaux. D'après Grunwald (2009), seulement 6% des études CNS étaient consacrées à l'estimation du RU. Un nouveau décompte récent (Chen et al, 2020) révèle une stagnation de ce chiffre (moins de 8% des papiers publiés entre 2003 et 2019)

L'estimation du RU dépend principalement des estimations des valeurs caractéristiques de rétention en eau du sol, soit les humidités volumiques CC et PFP.

Bien que Padarian et al. (2014) ont mesuré ces deux humidités, la majorité des études les estiment par l'utilisation de FPT reflétant la difficulté de mesure de ces variables évoquées en section I.2. Román Dobarco et al. (2019) ont alimenté les FPT par des valeurs de texture pour estimer des humidités volumiques. Malone et al. (2009) ont ajouté la densité apparente aux données texturales utilisées dans les FPT. Ugbaje et Reuter ont testé plusieurs FPT pour estimer les humidités CC et PFP à partir de la texture, de la densité apparente, de la capacité d'échange cationique (CEC), du pH et de la matière organique. Ces données ont également alimenté les FPT de Leenaars et al. (2018). Hong et al. (2003) ont estimé les humidités CC et PFP massiques à partir des teneurs en sable, argile, matière organique et de la CEC. Enfin, Leenhardt et al. (1994) ont pu déterminer l'humidité PFP à partir d'une FPT calibrée sur la mesure de plusieurs potentiels matriciels (3 kPa, 30 kPa, 100 kPa et 300 kPa) dont celui de l'humidité CC (10 kPa). Les FPT peuvent aussi être utilisées pour estimer directement le RU. Poggio et al. (2010) ont estimé le RU en utilisant des classes de pédotransferts élaborées par Bibby et al. (1982) intégrant la texture, la matière organique, la profondeur de sol, la densité apparente et la teneur en éléments grossiers.

L'estimation du RU est également dépendant de la profondeur et de la teneur en éléments grossiers dont la prise en compte ne fait pas consensus. Pour la profondeur, certaines études se réfèrent aux spécifications *GlobalSoilMap* pour déterminer la profondeur d'arrêt (Malone et al., 2009 ; Padarian et al., 2014 ; Hong et al., 2013 ; Ugbaje et Reuter, 2013 ; Román Dobarco et al., 2019), ou à la profondeur de sol (contact lithique ou paralithique) (Poggio et al., 2010) ou bien à la profondeur d'enracinement (Leenaars et al., 2008). La profondeur est aussi prise en compte différemment pour spatialiser l'épaisseur de sol sur laquelle le RU est calculé. Poggio

et al. (2010) ont estimé sur les épaisseurs maximales des profils de sols quand Malone et al. (2009) ont discrétisé le RU selon les intervalles de profondeur GSM.

L'estimation du RU est également modulée par la prise en compte des éléments grossiers. Pour l'ensemble des études présentées dans le Tableau 1.2, la nature lithologique des éléments grossiers n'est pas intégrée dans l'estimation du RU. Cependant, la teneur en éléments grossiers n'est pas non plus systématiquement intégrée dans l'estimation du RU (Malone et al., 2009 ; Ugbaje et Reuter, 2013 ; Padarian et al., 2014 ; Leenhardt et al., 1994 ; Leenaars et al., 2018), ce qui tend à sous-estimer les vraies estimations de RU (voir section I.2.3).

Pour valider les modèles CNS de RU et ne pouvant être directement mesuré, deux approches de validation des prédictions de RU se distinguent : i) l'approche visant à valider uniquement les prédictions des composants du RU (38%) (humidités CC et PFP, densité apparente, teneur en éléments grossiers) (Román Dobarco et al., 2019 ; Padarian et al., 2014 ; Leenhardt et al., 1994) et/ou les données d'entrées des fonctions de pédotransferts (argile, limon, sable, densité apparente), et ii) l'approche visant à évaluer les prédictions du RU dans leur globalité (62%) (Poggio et al., 2010 ; Malone et al., 2009 ; Ugbaje et Reuter, 2013, Leenhardt et al., 1994).

Dans ces travaux de thèse, nous utiliserons un concept de RU plus générique et proche de l'attente et des besoins des décideurs selon les modalités définies en section I.3. Cependant, nous ferons quelques concessions, notamment sur : i) les humidités des éléments grossiers non prises en compte (qui n'est jamais considérée dans les études CNS du RU), ii) les vraies valeurs des humidités spécifiques de sols (prise en compte par Padarian et al., 2014) et iii) la profondeur d'enracinement, simplifiée par des profondeurs fixes (prise en compte par Leenaars et al., 2018).

Tableau 1.2. Inventaire non-exhaustif des études CNS du RU

Références	Variables prédites	Eléments grossiers	Effectif des données	Densité (1/*n*km²)	Région	Echelle	Superficie (km²)	Modèle CNS	Profondeur envisagée	Var. expl. (%)	Intervalle de confiance
Malone et al. (2009)	RU	Non	341	4	Edgeroi	Locale	1 500	NN	100 cm	8 - 29	Non
Poggio et al. (2010)	RU	Oui	2290 - 103	34 - 20	Ecosse	National - locale	78000 - 2100	GAM	Profondeur du sol	28 - 25	Oui
Ugbaje et Reuter (2013)	Composants	Non	1120	825	Nigéria	Nationale	923 768	Cubist	200 cm	25	Non
Hong et al. (2013)	RU	Oui	380	263	Corée	Nationale	100 000	Stat	100 cm	-	Non
Padarian et al. (2014)	Composants	Non	806	2171	Côte Australienne	Régionale	1 750 000	GP, SVM, Cubist	100 cm	-	Oui
Leenhardt et al. (1994)	RU	Non	109	0,12	France	Locale	13	Stat	Racinaire	64	Non
Leenaars et al. (2018)	RU	Non	58321	350	Afrique	Nationale	20 400 000	Reg	Racinaire	52	Non
Román Dobarco et al. (2019)	Composants	Oui	36381	15	France	Nationale	543 940	RK	Roche-mère	-	Oui

Ext: validation indépendante ; CV: validation croisée ; RK : régression kirgeage; QRF : forêt aléatoire quantile; CART : arbre de classification et de régression; Reg : régression linéaire; Stat : prédiction calculée par unité de sols ;NN : réseau de neurones artificiels; GP : programmation génétique; SVM : "support vector machines"; GAM : modèle généralisé additif ; Humidité CC: humidité à la capacité au champ; Humidité PFP: humidité au point de flétrissement permanent; RU : réservoir utile

III. Les questions méthodologiques en suspens

En mettant en relation l'objectif finalisé que nous nous sommes fixé (voir introduction générale) avec l'état de l'art en cartographie numérique des sols exposé précédemment, il apparaît des questions méthodologiques en suspens. Ces questions sont tout d'abord liées au caractère multivarié du RU qui induit des difficultés à appliquer une approche classique de cartographie numérique des propriétés de sol essentiellement conçue pour prédire une seule propriété à la fois. Ainsi, deux questions méthodologiques doivent être traitées :

- Quelle est la sensibilité des estimations du réservoir utile à leurs trajectoires de calcul ?
- Comment estimer et propager l'incertitude d'un indicateur fonctionnel de sol tel que le réservoir utile ?

L'objectif de ces travaux de thèse est également d'accéder à une précision fine pour pouvoir connaitre les variations spatiales du RU à l'échelle locale. La densité d'observations de sol étant reconnue comme facteur limitant aux performances de prédictions des modèles de cartographie numérique des sols, l'utilisation de données anciennes pédologiques denses et hétérogènes peut être une alternative crédible bien qu'elles ne soient pas actuellement disponibles en bases de données informatisées. Ce second constat conduit à poser la question méthodologique suivante :

- Comment valoriser l'utilisation de données pédologiques anciennes dans les modèles de cartographie numériques des sols ?

Par la suite, nous développerons ces trois questions méthodologiques.

III.1. Les trajectoires de calcul du réservoir utile

La cartographie numérique des sols a atteint la phase d'opérationnalité pour fournir des cartes numériques des propriétés primaires de sols (texture de la terre fine, teneur en carbone organique, pH du sol, etc.). Cependant, l'objectif initial n'a pour le moment pas abouti. Pour cela, la cartographie numérique des sols doit maintenant développer de nouvelles approches afin de spatialiser des propriétés fonctionnelles de sol telles que le réservoir utile.

Bien que les données d'entrées soient très souvent constituées par des propriétés primaires de sols et des fonctions de pédotransferts, l'estimation du RU et sa spatialisation ne font pas consensus dans la littérature. Les études présentées en section II.3 ont des approches relativement différentes sur la combinaison de propriétés à spatialiser (propriétés primaires vs

RU) ainsi que sur l'épaisseur de sols à prendre en compte pour calculer le RU (couches de sol du *GlobalSoilMap* vs profil de sol).

Ce constat soulève deux questions : i) quelle combinaison de propriétés de sol faut-il spatialiser ? et ii) quel niveau d'agrégation des couches de sols faut-il considérer pour spatialiser la combinaison de propriétés ?

Pour fournir des éléments de réponses, une conceptualisation de la spatialisation est nécessaire en tenant compte de ces deux questions. Nous proposons ci-dessous une représentation tridimensionnelle répertoriant les différentes trajectoires de calcul possibles pour fournir une estimation unique du RU. Les trajectoires de calcul peuvent varier selon l'ordre d'exécution de : i) la combinaison des propriétés primaires de sols (axe rouge sur la Figure 1.6), ii) l'agrégation des couches de sols (axe bleu sur la Figure 1.6) et, iii) la spatialisation (axe vert sur la Figure 1.6).

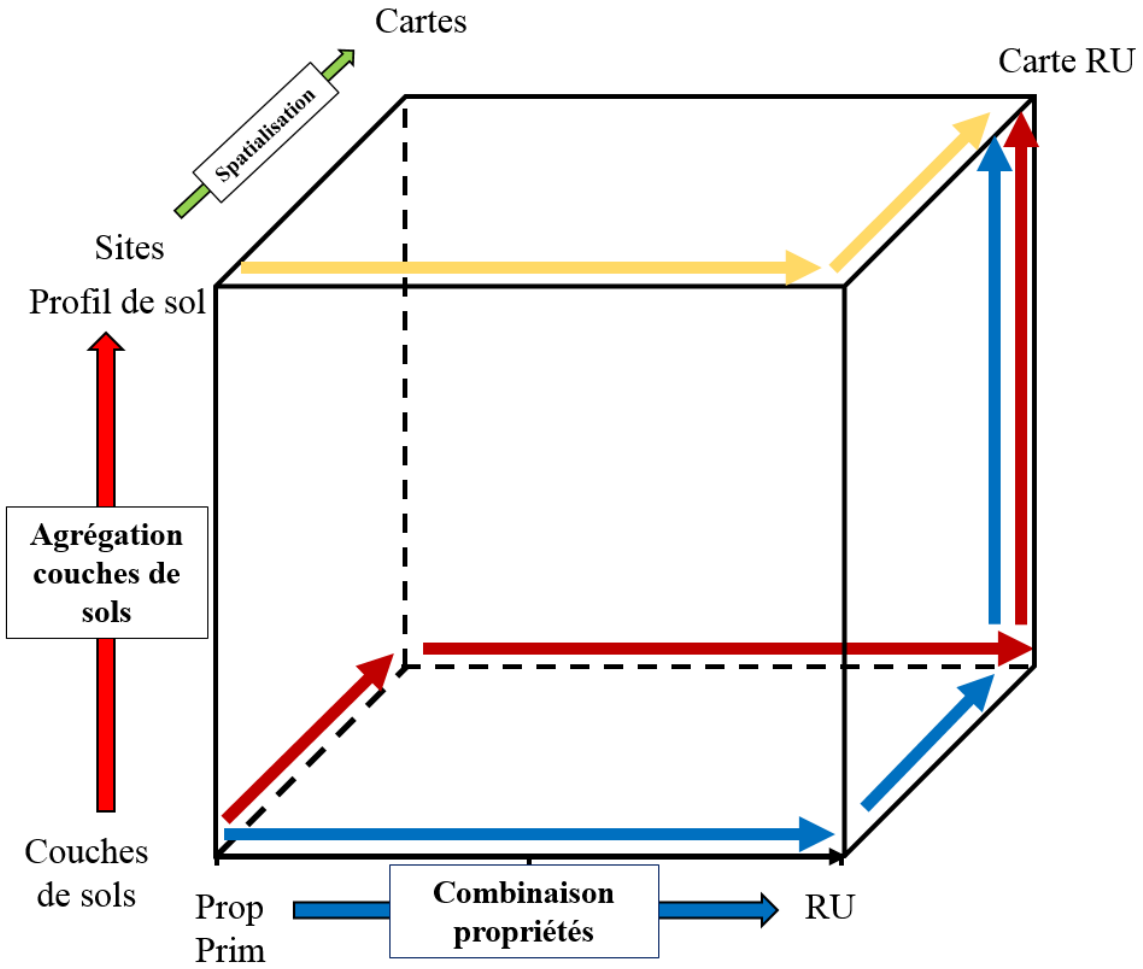


Figure 1.6. Conceptualisation de la spatialisation du réservoir utile avec des exemples de trajectoire de calcul

L'approche la plus fréquemment utilisée est la spatialisation indépendante de chaque propriété primaire de sol puis l'agrégation de couches de sol (trajectoire rouge sur la Figure 1.6) (Ugbaje et Reuter, 2013 ; Román Dobarco et al., 2019). A l'inverse, la spatialisation peut être le dernier processus à exécuter après avoir combiné les propriétés primaires et les couches de sol sur la base des données d'entrées de sol (trajectoire jaune sur la Figure 1.6) (Poggio et al., 2010). Au milieu de ces deux propositions extrêmes, une trajectoire intermédiaire consisterait à combiner les propriétés de sols pour chaque couche de sol puis spatialiser avant d'agréger (trajectoire bleue sur la Figure 1.6) (Malone et al., 2009). Ces exemples de trajectoires de calcul présentent des performances hétérogènes compris entre 25% (Poggio et al., 2010) et 44 % (Malone et al., 2009) de variance expliquée. En dehors, de ces trois exemples, il est possible d'entrevoir un grand nombre d'autres trajectoires de calcul aussi bien basées sur une combinaison partielle des propriétés de sols (application des fonctions de pédotransferts) que sur l'agrégation partielle des couches. Récemment, Laborczi et al. (2019) ont fourni quelques éléments de réponses à l'importance de la trajectoire de calcul dans la spatialisation de propriété de sols. En effet, ils ont comparé deux trajectoires de calcul pour cartographier les teneurs en argile, limon et sable à une profondeur de 30 cm. La première trajectoire de calcul consistait à agréger d'abord les couches de sols *GlobalSoilMap* (0-5 cm, 5-15 cm et 15-30 cm) puis à spatialiser, tandis que la seconde trajectoire de calcul était de spatialiser les couches de sols individuellement avant de les agréger. Les résultats obtenus sont sensiblement différents selon la trajectoire de calcul sélectionnée, la première trajectoire donnant des résultats de performances moins satisfaisants que la seconde. Bien que le sujet de cette étude couvre partiellement la problématique de spatialisation du RU (trajectoire élaborée par rapport à la considération de la profondeur), un premier élément de réponse permet de mettre en avant la sensibilité relative des estimations d'une propriété de sol quant au choix de la trajectoire de calcul utilisée.

III.2. Quantification de l'incertitude

Selon les spécifications *GlobalSoilMap*, chaque représentation cartographique doit être accompagnée d'une estimation de l'incertitude sous la forme d'un intervalle de confiance à 90%. Dans le cas d'une spatialisation d'une propriété primaire de sol, l'intervalle de confiance est souvent dérivé des erreurs liées à l'utilisation du modèle (Heuvelink et al, 2014) soit par les quantiles de distributions (Vaysse and Lagacherie, 2017) soit en utilisant le théorème de la centrale limitée à partir d'une distribution de la variance (Rocha et Yamamoto, 2000).

Cependant, la représentation de l'incertitude par un intervalle de confiance est rarement présentée dans les études CNS de RU (25%, Tableau 1.2), principalement à cause de la difficulté à estimer cette incertitude. Une incertitude du RU est une donnée importante aussi bien pour alimenter des modèles de fonctionnement des cultures et hydrologiques (STICS (Brisson et al., 1998), SWAT (Arnold and Williams, 1987 ; Arnold et Fohrer, 2005)), que pour guider les chercheurs et les décideurs politiques à atteindre les objectifs de développement durable fixés par les Nations Unis (assurer la sécurité alimentaire, promouvoir l'agriculture durable, atténuer le changement climatique et permettre une gestion durable de la ressource en eau, notamment).

L'estimation de l'incertitude d'une propriété fonctionnelle impliquant une combinaison de propriétés de sol et de profondeurs est relativement fastidieuse et nécessite de propager les sources d'erreurs selon les opérations à effectuer pour arriver à une estimation du RU répertoriées ci-après (Heuvelink et al., 1989 ; Carré et al., 2007) :

- Erreur de mesure (lié à l'instrument de mesures ou au nombre d'opérateur) ;
- Erreur liée à la structure et aux paramètres de la fonction de pédotransfert ;
- Erreur sur les limites supérieure et inférieure du RU en termes de définition des potentiels matriciels de référence (pression matricielle utilisées pour mesurer les humidités à la capacité au champ et au point de flétrissement permanent) ;
- Erreur liée à la spatialisation ;
- Erreur sur les variables environnementales utilisées dans les modèles de régression.

Poggio et al. (2010) ont déduit l'intervalle de confiance selon les bornes inférieures et supérieures de l'intervalle de confiance mais issu de la prédiction de RU et ne nécessitant pas de propagation d'erreur. Bien qu'ils aient considéré l'incertitude sur la tendance du modèle ainsi que l'incertitude locale et spatiale, ils n'ont cependant pas inclus l'incertitude liée à l'utilisation des fonctions de pédotransferts. Toutefois, les incertitudes sur les propriétés hydrodynamiques de sol (humidité CC et PFP) dues aux erreurs des FPT sont parfois faibles par rapport à l'incertitude des données d'entrées (Minasny et al., 1999).

Récemment, Román Dobarco et al. (2019) ont utilisé un modèle de propagation fondé sur le premier ordre des analyses de Taylor afin de propager les erreurs de spatialisations des propriétés et des fonctions de pédotransferts dans la prédiction finale de l'incertitude du RU. Bien que l'incertitude finale du RU ait été délivrée sans être évaluée, ces travaux ont montré que la spatialisation des propriétés composant le RU (texture et éléments grossiers) constituait

la principale source d'incertitude du RU devant l'incertitude liée aux FPT. Cependant, les erreurs liées à l'agrégation des couches de sols n'ont pas été inclus dans cette estimation. Plus tard, Algayer et al. (2020) ont démontré l'importance cruciale de l'épaisseur du sol dans l'estimation du RU, ce qui revient à supposer que ces erreurs sont indépendantes les unes des autres, mais cela n'a pas été vérifié.

Dans certains cas, l'estimation de l'incertitude est simplifiée comme étant la différence entre les limites des intervalles de confiances des humidités CC et PFP (Padarian et al., 2014). Cependant, cette estimation de l'incertitude du RU considérée comme la différence entre les humidités spécifiques de sols (non prise en compte des éléments grossiers) et ne prend alors pas en compte la corrélation entre ces variables.

Afin de délivrer une incertitude complète du RU, il est nécessaire de prendre en compte les erreurs de spatialisation de toutes les propriétés composantes du RU et leurs corrélations, ainsi que les corrélations des erreurs entre les différentes couches de sols.

III.3. Utilisation des données anciennes

Bien que la CNS soit désormais passée en phase opérationnelle (Minasny et McBratney, 2016 ; Arrouays et al., 2017), les performances des propriétés spatialisées présentent souvent une incertitude plus importante que prévue. Par exemple, les pourcentages de variances expliquées inférieurs à 50 % sont observés pour 95 % des propriétés testées à l'échelle locale (Nussbaum et al., 2018), 76 % à l'échelle régionale (Vaysse et Lagacherie, 2015), 100% à l'échelle nationale (Mulder et al., 2016) et 86% à l'échelle mondiale (Hengl et al., 2014).

L'analyse des résultats par ces auteurs a convergé vers le fait que la densité d'observations de sols utilisées pour calibrer les modèles CNS est un facteur limitant les performances de prédictions. La majeure partie de l'information pédologique utilisée dans les études CNS sont issues principalement de cartes pédologiques ou de mesures de propriétés de sols selon un plan d'échantillonnage spatial. La densité moyenne en observations de sols est relativement faible dans les études CNS opérationnelles avec 4 à 12 profils/km² à l'échelle locale (Nussbaum et al., 2018), et des densités largement inférieures à 1 profil/km² pour les échelles régionale, nationale et mondiale (Vaysse et Lagacherie, 2015 ; Mulder et al., 2016 ; Hengl et al., 2014). Cela limite les performances de prédictions car la variabilité spatiale de la variable s'exprime à une échelle beaucoup plus grande que l'espacement entre les profils de sols (Vaysse et Lagacherie ; Gomez et Coulouma, 2018). Ce constat a pu être observé et confirmé par de récentes études ayant pour but de faire varier la densité spatiale des profils de sol (Somarathna

et al., 2017 ; Wadoux et al., 2019 ; Lagacherie et al., 2020). Par conséquent, l'augmentation de la densité d'observations de sols est d'une importance primordiale dans le but d'améliorer les performances de prédictions des propriétés de sols des modèles CNS (Voltz et al., en cours de publication).

Le moyen le plus pragmatique d'augmenter la densité de données utilisées dans les modèles CNS est de collecter des données pédologiques anciennes qui n'ont pas encore été intégrées aux bases de données existantes. Ceci est en cours de réalisation, les travaux menés par Arrouays et al. (2017) ayant montré une augmentation de la quantité de profils de sols inclus dans les bases de données mondiales de 1046% et de 45% pour les bases de données nationales entre 2007 et 2015. Ces mêmes travaux ont également montré qu'un nombre important de données pédologiques anciennes demeuraient inexploitées. Par exemple, pour pallier la faible densité des données utilisées dans les modèles CNS, BRL Exploitation dispose sur l'ensemble de la plaine littorale Languedocienne de 228 000 observations de sols correspondant à une densité de 34 observations de sol / km².

Les modèles actuels de CNS utilisés de façon opérationnelle sur de larges étendues (Malone et al., 2009 ; Vaysse and Lagacherie, 2015) ont été élaborés pour être calibrés avec des données spatialement peu denses et exactes contrairement aux caractéristiques des données pédologiques anciennes. Afin de prendre en compte l'utilisation d'échantillonnages spatiaux denses, les méthodes géostatistiques sembles les plus adaptés (krigeage ordinaire, régression krigeage). Toutefois, les récents travaux de Hengl et al. (2018) ont permis d'intégrer la proximité géographique entre les points dans des modèles de forêt aléatoire, fréquemment utilisés dans les études CNS (Vaysse et Lagacherie, 2015). Enfin, la saisie d'une telle quantité de données peut être également un long processus, pour cela il est nécessaire d'estimer tout d'abord le temps de saisie nécessaire selon les spécificités des données citées ci-dessus ainsi que de réaliser des études de faisabilité et coût-bénéfice.

IV. Plan de thèse

La suite de ce mémoire de thèse est conçue afin de répondre aux trois questions méthodologiques posées avec une présentation préalable du cas d'étude. Le chapitre 2 sera donc consacré à cette présentation en développant les principales caractéristiques de la zone d'étude ainsi qu'une étude exploratoire des données pédologiques utilisées dans cette thèse. Le chapitre 3 portera sur la sensibilité des estimations du réservoir utile à sa trajectoire de calcul. Le chapitre 4 sera axé sur l'estimation de l'incertitude associée aux prédictions de réservoir utile, par utilisation d'un modèle de propagation des erreurs. Le cinquième chapitre sera dédié à l'évaluation de la valorisation de l'utilisation de données pédologiques anciennes dans les modèles de cartographie numérique des sols pour spatialiser le réservoir utile. Enfin, une conclusion générale des travaux de thèse clôturera ce manuscrit.

Les éléments importants de ce chapitre à retenir sont :

Estimation locale du réservoir utile

- Le réservoir utile en eau est estimé à l'échelle d'un site à partir de la **profondeur** (enracinement ou contact de la roche mère), des **éléments grossiers** (volume et nature lithologique), des **humidités à la capacité au champ et au point de flétrissement** et de **la densité apparente**.
- Choix d'estimation du réservoir utile dans cette thèse :
  - **Profondeur :** Considération de la profondeur à la roche mère et de la profondeur d'enracinent en estimant le réservoir utile selon trois profondeurs fixes (30 cm, 60 cm et 100 cm) laissant le choix à l'utilisateur de l'estimation la plus adaptée au type de culture,
  - **Eléments grossiers :** prise en compte uniquement de l'abondance des éléments grossiers considérés inertes car la quantité d'information de la nature lithologique est insuffisante,
  - **Humidités à la capacité au champ et point de flétrissement permanent :** utilisation de fonctions de pédotransfert alimentées par des données texturales permettant d'obtenir des humidités volumiques (prise en compte et simplification de **la densité apparente**)

Principes généraux de la cartographie numérique des sols

- Le principe de la cartographie numérique des sols consiste à prédire, à l'aide d'un modèle, une propriété ou une classe de sol à partir, d'une part des données pédologiques disponibles (cartes ou sites) et des données spatialisées des éléments du paysage en lien avec la formation des sols (relief, climat, organismes vivants, géologie) accompagnée d'une erreur d'estimation (McBratney et al., 2003).
- Les modèles de cartographie numérique des sols peuvent prédire leur propre incertitude (Heuvelink., 2013 ; Vaysse et Lagacherie, 2017). Cette prédiction d'incertitude permet d'éclairer la prise de décision. Elle est exprimée usuellement par un intervalle de confiance à 90% (Arrouays et al., 2014).

- Le principal facteur limitant de la cartographie numérique des sols est **la faible densité** d'observations de sols

Cartographie numérique des sols et réservoir utile

- Le peu d'étude portant sur la spatialisation du réservoir utile présentent **des méthodologies sensiblement différentes** (pas de consensus).
- Une quantification de l'incertitude est **rarement fournie**.

QUESTIONS METHODOLOGIQUES EN SUSPENS

1. Quelle est la sensibilité des estimations du réservoir utile à leurs trajectoires de calcul ? (Chapitre 3)
2. Comment estimer et propager l'incertitude d'un indicateur fonctionnel de sol tel que le réservoir utile ? (Chapitre 4)
3. Comment valoriser l'utilisation de données pédologiques anciennes dans les modèles de cartographie numériques des sols ? (Chapitre 5)

CHAPITRE 2 La zone d'étude et les données utilisées dans la thèse

I. Zone d'étude

Les données détenues par BRL Exploitation sont intégralement réparties sur l'ex-région Languedoc-Roussillon (Figure 2.1). De ce fait, nous présenterons, dans ce chapitre, l'ensemble des caractéristiques géologiques, climatiques, anthropiques et pédopaysagères du Languedoc-Roussillon avec systématiquement un résumé de ces caractéristiques sur la zone d'emprise des données BRL Exploitation. Par la suite, le Languedoc-Roussillon sera présenté comme région géographique.

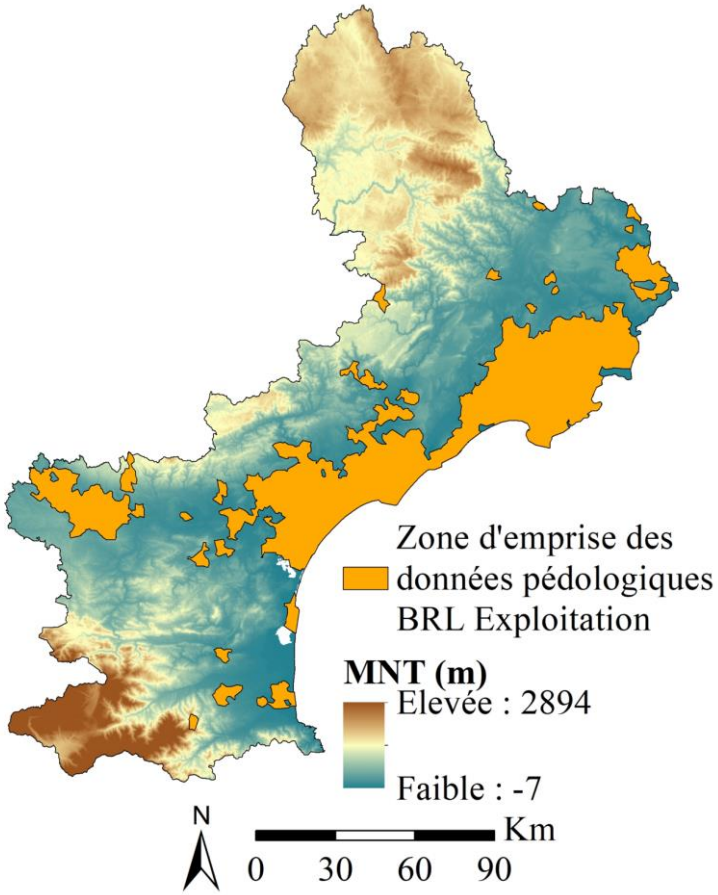


Figure 2.1. Couverture spatiale des données pédologiques anciennes de BRL sur l'ancienne région Languedoc-Roussillon

I.1. Introduction

Le Languedoc-Roussillon est une ancienne région administrative française du Sud de la France (Figure 2.2), qui compose depuis 2016, avec la région Midi-Pyrénées, la nouvelle région Occitanie. Le Languedoc-Roussillon est divisé en cinq départements : la Lozère, le Gard, l'Hérault, l'Aude et les Pyrénées Orientales. Les limites géographiques et naturelles du Languedoc-Roussillon sont les montagnes du Massif Central au Nord, la chaine Pyrénéenne à l'Ouest, la mer Méditerranée au Sud, ainsi que le fleuve du Rhône à l'Est. De par les milieux naturels de la zone d'étude, une grande diversité pédologique, géologique, paysagère, climatique et agricole est présente. Dans ce chapitre dédié à la présentation du site d'étude, chaque composante de la diversité (pédo-paysagères, géologiques, climatiques et anthropiques) du Languedoc-Roussillon sera abordée. De plus, certains passages se baseront sur les travaux de thèse de Kévin Vaysse (2015) proposant une description complète et détaillée du Languedoc-Roussillon.

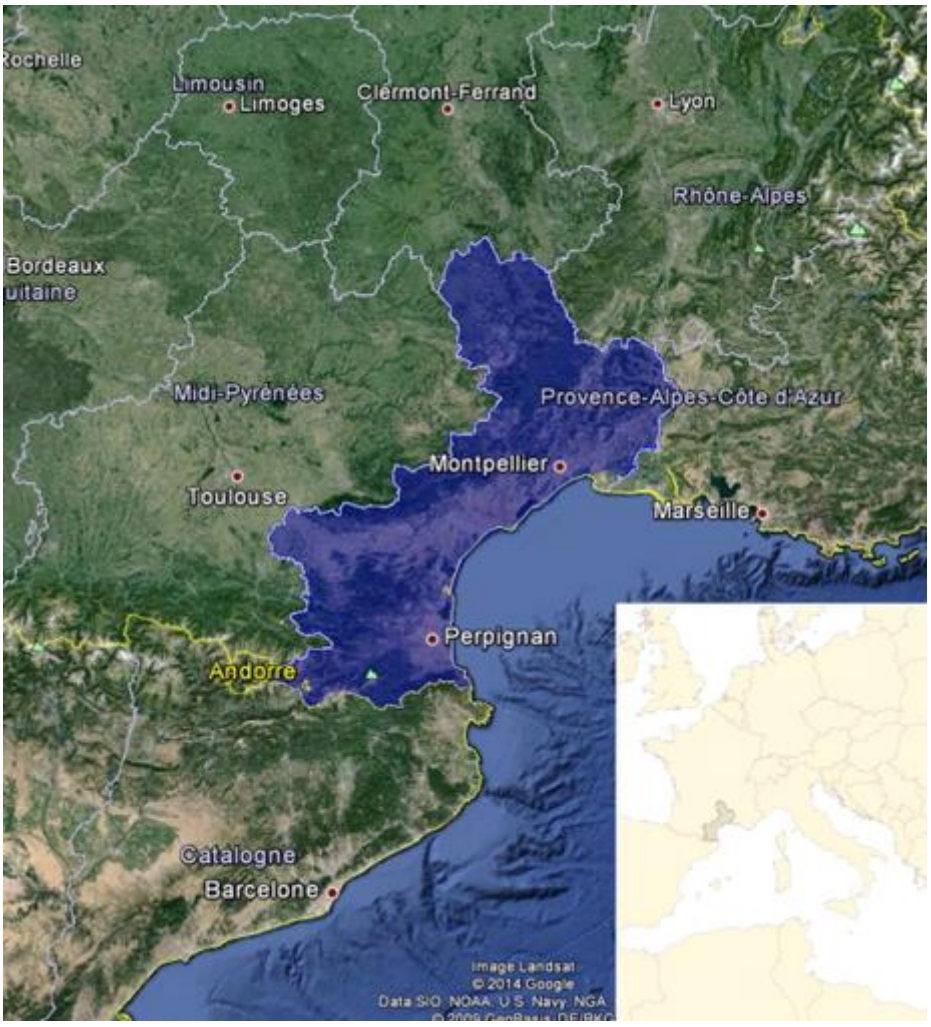


Figure 2.2. Localisation du site d'étude

I.2.Géologie (Debelmas, 1974)

Le Languedoc-Roussillon fait partie des régions les plus riches de France d'un point de vue géologique. Sa localisation, très spécifique, entre la section méridionale du Massif Central (Aubrac, Margeride, Cévennes, Causses et Montagnes Noires), l'extrémité des Pyrénées orientales, en passant par les plaines sédimentaires du Roussillon, du Bas-Languedoc et de l'Aude, résulte en une grande variété de structures et de lithologies. Cette diversité de formations géologiques découle d'un héritage géologique remarquable de 600 millions d'années, recouvrant une large gamme d'étages géologiques, du protérozoïque à nos jours (BRGM et DREAL, 2013).

La Figure 2.3 présente la répartition spatiale des grands ensembles géologiques du Languedoc-Roussillon. Plusieurs ensembles peuvent être distingués sur cette carte : le système hercynien du Massif Central (Margeride) ; la pointe orientale du massif pyrénéen et son prolongement (Corbières) ; le bassin sédimentaire de la vallée de l'Aude ; la zone de transition du couloir Rhodanien et du Bas-Languedoc englobant les plaines alluvionnaires Rhône-Hérault et la partie Sud du Massif Central (Causses).

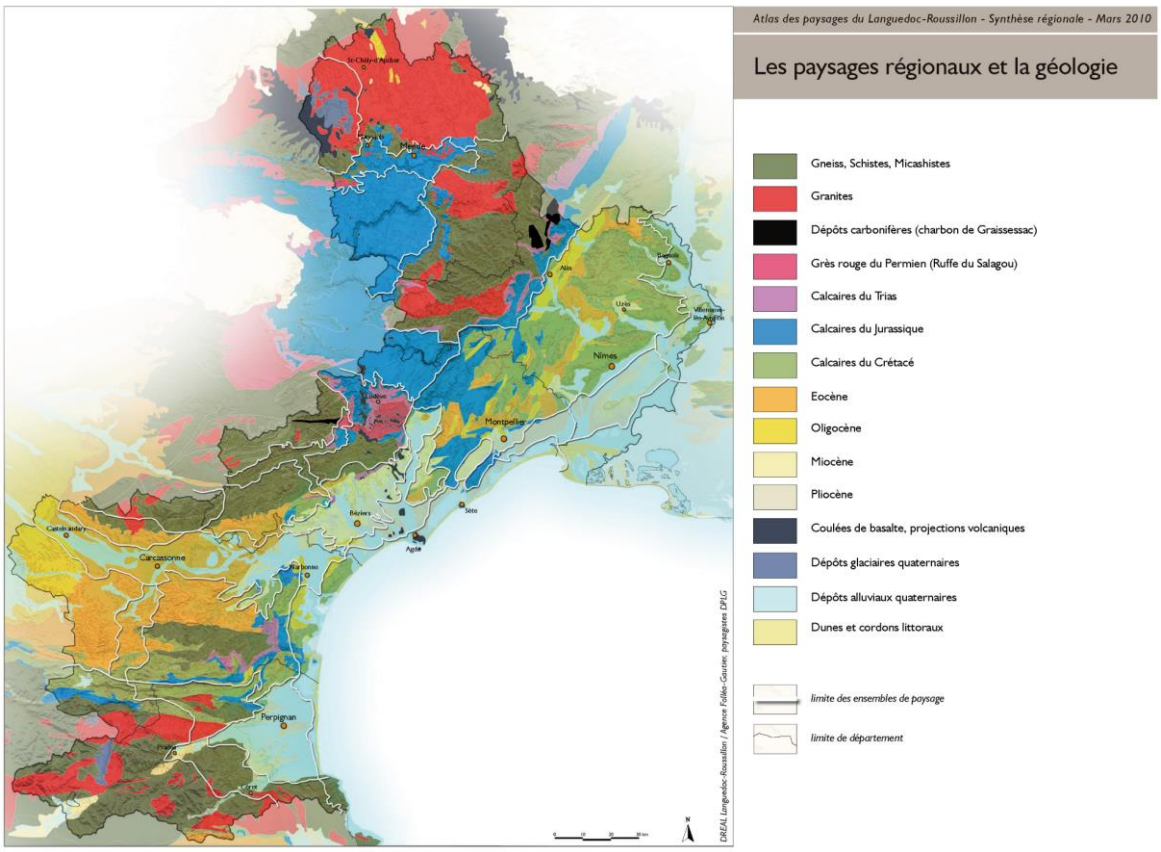


Figure 2.3. Cartographie synthétique des grands ensembles géologiques du Languedoc-Roussillon

Le couloir du Rhône et du Bas-Languedoc repose sur un socle triasique, potentiellement affleurant dans les zones remaniées en bordure de couloir et dans les Cévennes. La lithologie du Trias, dans le département de l'Hérault est principalement détritique composée de formations de grès et de Muschelkalk présentant des faciès argilo-détritiques ou carbonatés. Au nord du socle triasique, une bande jurassique composée essentiellement de calcaire marin persiste localement et les plus grandes étendues forment les Causses du Massif Central (Méjean, Sauveterre, Sud de la Margeride), où le calcaire est plus ou moins dolomitisé.

La présence de ces bandes de calcaires jurassiques est propice à la formation de plateaux karstiques et de cuestas, en migrant vers les plaines des zones de garrigues et de l'arrière-pays montpelliérains. En allant vers l'Est, dans la région Nord – Nord-Est de Nîmes, des formations calcaires d'origine marine datant du Crétacé, prédominent. Parmi les strates géologiques de cette série, c'est surtout durant l'Urgonien que les paysages de plateaux, parfois karstiques, ont été formés. La présence du Crétacé dans cette région témoigne du début d'une transgression marine.

En progressant vers les plaines, ces formations sont peu à peu enfouies par les dépôts tertiaires du Cénozoïque (Eocène, Oligocène). Durant le Néogène, les plaines du Bas-Languedoc ont connu de multiples phases de sédimentation impliquant une succession importante de transgression marine, de sédimentation fluvio-lacustres et de régression marine. Ces phases ont, ensuite, abouti à un enchevêtrement de couches soit marneuses, détritiques ou bioclastiques conduisant à des dépôts molassiques ainsi qu'à des dépôts calcaires et calcaires molassiques ou bioclastiques présentant une large gamme d'épaisseurs (5 à 130 m).

En se rapprochant du littoral, des dépôts fluviatiles composés des sédiments provenant du Rhône, de la vallée de l'Hérault et du Massif Central, constituent les principales formations géologiques de la Camargue.

Le Massif Central est principalement composé du massif granitique de la Margeride, des massifs métamorphisés de la ceinture Cévenole, ainsi que des Causses présentées ci-dessus. Le massif de la Margeride s'est formé suite à des phases de plutonismes datant du cycle hercynien, donnant principalement du granite à inclusion de feldspath, dit porphyrique, ainsi que, occasionnellement, des leucogranites (granites à biotites et muscovites). La ceinture cévenole comprend, quant à elle, de nombreuses formations métamorphiques dont la Montagne Noire et le secteur oriental de l'Albigeois et du Pays Cévenol, parmi les plus emblématiques. La montagne Noire résulte de la compression due au rapprochement du compartiment ibérique et

à la formation des Pyrénées, conduisant à une structure anticlinale orientée Est – Nord-Est. La structure anticlinale a permis de faire affleurer des formations anciennes dont les gneiss, les granites et des micaschistes, qui constituent les roches prédominantes. Ensuite, situé à l'Ouest des causses, l'Albigeois est constitué principalement de schistes. Le Pays Cévenol, présente une variabilité géologique plus importante puisque les principaux massifs métamorphiques (gneiss et micaschistes) se retrouvent entrecoupés par d'importantes intrusions magmatiques de nature granitique.

Les Pyrénées Orientales et son avant-pays demeurent des zones géologiques plus complexes du fait de la collision entre les plaque ibérique et eurasienne, ayant pour conséquence une remontée des strates sédimentaires du socle hercynien au niveau de la zone primaire axiale (extrémité Sud du Languedoc-Roussillon). Les formations géologiques sont principalement composées de gneiss, de micaschistes et de granites. « *La région des Corbières constitue la zone sous-pyrénéenne qui s'étend jusqu'aux dépôts éocènes de la vallée de l'Aude* » (Vaysse, 2015). Au sein de cette région géologique, plus précisément dans les hautes Corbières, le massif paléozoïque du Mouthoumet présente une morphologie structurée par l'orogénèse varisque survenue pendant le Carbonifère. Par ailleurs, le massif du Mouthoumet constitue avec la zone primaire axiale les limites géologiques de la nappe de Corbières, présentant des formations molassiques datant du Crétacé supérieur.

A la limite de la région des Corbières, les dépôts continentaux de l'Eocène composés de grès, poudingues, sables et calcaires lacustres, occupent principalement la vallée de l'Aude. De plus, des mouvements tectoniques de compressions ont considérablement modifié la géomorphologique de cette zone avec une rotation de l'axe de la zone de plissement, initialement identique à l'axe de la zone primaire axiale (Sud – Nord), qui a progressivement migré vers une orientation Sud-Ouest – Nord-Est dans le bassin de l'Aude (visible également dans le Bas-Languedoc).

Emprise de la zone des études BRL Exploitation :

La zone d'études BRL couvre principalement, d'Est en Ouest, les formations calcaires du Crétacé de la région de Nîmes et ses terrasses alluviennes anciennes, en passant par les dépôts fluviatiles de la Camargue, les dépôts tertiaires du Cénozoïque, ainsi que les terrasses alluviennes anciennes, pour finir sur les dépôts continentaux de l'Eocène.

I.3.Climatologie (Joly et al., 2010)

La variété et la répartition spatiale des climats présents en Languedoc-Roussillon sont essentiellement reliées à un gradient altimétrique. Les travaux de Joly et al. (2010) (Figure 2.4) ont permis d'identifier cinq climats en Languedoc-Roussillon : le climat de montagne, le climat semi-continental, le climat méditerranéen altéré, le climat du bassin du Sud et le climat méditerranéen franc.

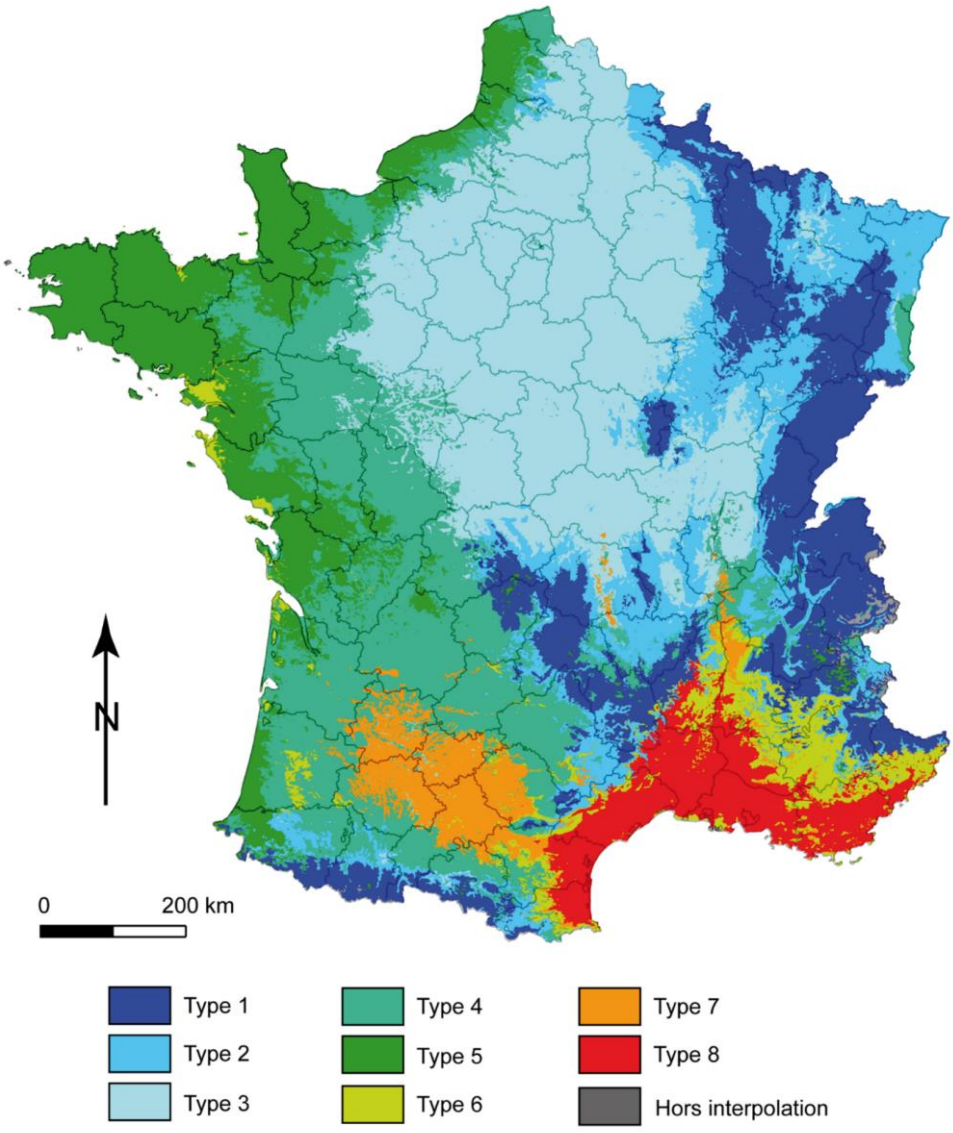


Figure 2.4. Représentation typologique des climats métropolitains (d'après Joly et al., 2010)

Les hautes altitudes du Massif Central et des Pyrénées constituent le climat montagnard (type 1) du Languedoc-Roussillon, caractérisé par un nombre de jours et un cumul élevé de précipitation, une température moyenne inférieure à 9.4 °C et corrélativement, plus de 25 jours au cours desquels la température minimale est inférieure à -5°C et moins de 4 jours avec une

température maximale supérieure à 30°C. La variabilité interannuelle des précipitations de juillet et des températures d'hiver et d'été est maximale.

Le climat semi-continental (type 2) est un climat de transition entre le climat de montagne et le climat méditerranéen franc, en Languedoc-Roussillon. Ce type de climat est souvent situé en périphéries des montagnes, présentant des températures plus douces, des précipitations plus faibles et moins fréquentes qu'en montagnes. Cependant la variabilité climatique sur la normale 1971-2000 reste tout aussi élevée.

Le climat méditerranéen altéré (type 6) est présent en liseré entre les Pyrénées Orientales et le département de l'Hérault. Il se caractérise par une température moyenne élevée et une variabilité interannuelle faible. Les étés chauds et les hivers humides sont caractéristiques de ce type de climat, conséquence directe d'une répartition temporelle des précipitations annuelles hétérogènes, bien que le cumul de précipitation annuel se situe dans la moyenne nationale.

Le climat du Bassin du Sud (type 7) se situe dans la région du Lauragais, caractérisé par une température moyenne élevée (> 13°C) et d'une faible variabilité interannuelle. Les précipitations annuelles et hivernales sont peu importantes avec une légère augmentation durant l'été (moins de 800 mm à l'année) bien qu'elles soient plus fréquentes en hiver qu'en été (9-11 jours en hiver contre < 6 jours en été).

Enfin, le climat méditerranéen franc (type 8), localisé entre les Pyrénées et le Var, est un climat beaucoup plus marqué d'une part, par l'amplitude des températures entre les périodes hivernales et estivales (> 17°C), mais également, par le rapport pluviométrique jusqu'à 6 fois plus important en automne qu'en été, due à des évènements pluvieux cévenoles. En dépit de ces évènements cévenols, le cumul de précipitation annuel reste faible (556 mm, station de Perpignan ; Météo France, 2020).

Emprise de la zone des études BRL Exploitation :

Les zones d'études BRL sont majoritairement localisées en climat méditerranéen franc et en climat Bassin du Sud pour les données situées dans région du Lauragais.

I.4. Agriculture et pression anthropique

L'occupation des sols du Languedoc-Roussillon est révélatrice des pressions anthropiques appliquées aux sols et paysages de la zone d'étude. La Figure 2.5 présente les différentes classes d'occupation du sol qui peuvent être regroupées en quatre grands types d'occupation : les espaces agricoles, les espaces forestiers, les prairies et les espaces artificialisées.

Les espaces agricoles sont principalement distribués dans les vallées (Plaines du Roussillon, de l'Aude, de l'Hérault et du Bas-Rhône) et subdivisés en trois grands types de selon leur ordre d'occupation surfacique : la viticulture, le maraichage et les grandes cultures (Barthès et al., 1999a).

Historiquement, la viticulture intensive de masse dominait l'ensemble de l'Hérault. Cependant, la surproduction vinicole et la transition des départements limitrophes (Gard et Aude) vers une diversification des cultures ont amené à l'arrachage de pieds de vigne aboutissant, selon la capacité d'irrigation, à l'artificialisation ou au changement de cultures de ces terres. Du fait d'une grande capacité d'irrigation du Gard et de l'Aude, le maraîchage et les grandes cultures prédominent dans ces deux départements.

Les prairies sont préférentiellement localisées dans des zones de hautes altitudes telles que les Causses calcaires de Massif Central (Larzac, Méjean et Sauveterre), les zones de moyennes montagnes (Aubrac, Margeride), mais aussi, ponctuellement, sur des zones basses telles que les serres cévenoles. Le développement de prairies est en lien avec des conditions climatiques favorisant la minéralisation de la matière organique ainsi que la sauvegarde du couvert végétal pendant toute l'année. L'utilisation des prairies est majoritairement vouée à l'élevage.

Les paysages forestiers sont principalement situés hors des plaines et des terrasses d'alluvions anciennes et préférentiellement dans les paysages cévenols tels que la Montagne Noire ainsi que les parcs régionaux (Parc du Haut-Languedoc et Parc National des Cévennes).

« *Les espaces forestiers sont relégués en dehors de la plaine et des terrasses alluvionnaires anciennes. Les serres et collines cévenoles ainsi que la Montagne Noire abritent une grande partie des parcs nationaux du Languedoc-Roussillon (Parc du Haut-Languedoc ; Parc National des Cévennes).* » (Vaysse, 2015)

La superficie des espaces artificialisés est en nette augmentation dans le Languedoc-Roussillon (+17 % entre 1997 et 2009) au détriment des surfaces agricoles (Balestrat et al., 2011a).

Cette politique d'artificialisation des sols fait suite aux crises successives liées à l'activité viticole couplées à une démographique et à la pression foncière qui ne cessent d'augmenter depuis les années 1960. Le Languedoc-Roussillon est également exposé à l'artificialisation de sa plaine littorale, notamment de l'Hérault, afin de soutenir et de développer le tourisme (construction stations balnéaires, ports de plaisances, etc.).

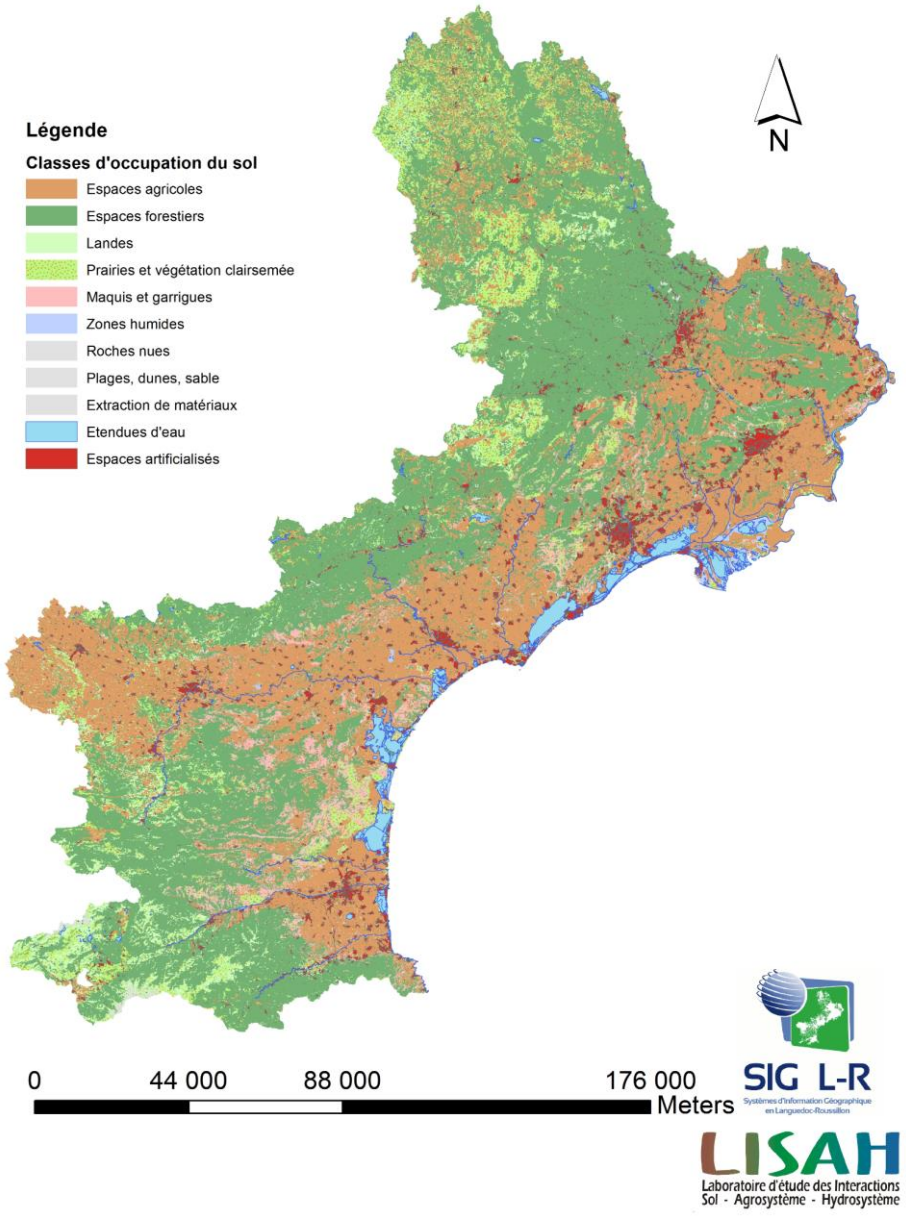


Figure 2.5. Répartition spatiale simplifiée de l'occupation des sols du Languedoc-Roussillon (d'après Vaysse, 2015)

Emprise de la zone des études BRL Exploitation :

La zone d'emprise des études pédologiques de BRL Exploitation occupe principalement les espaces agricoles avec quelques zones en domaine forestiers dans la région de Nîmes, ainsi que des espaces urbanisés.

I.5. Pédologie (Barthès et al., 1999a, 1999b, 1999c, 1999d, 1999e)

La description des sols régionaux se base sur la cartographie au 1/250 000^{e} issue du Référentiel Régional Pédologique du Languedoc-Roussillon (Arrouays et Jamagne, 1989) (Figure 2.6).

La couverture des sols de la région est très variée et contrastée, notamment due à la présence de régions naturelles très diversifiées.

« *Cette diversité pédologique implique 18 groupes de sols majeurs de la classification globale WRB représentant 56 % et 75 % de l'ensemble des groupes de sols à l'échelle mondiale et Européenne, respectivement.* » (Vaysse, 2015)

Ces sols se répartissent sur un grand nombre de types de paysages présents sur Le Languedoc-Roussillon composés de : plaines et terrasses alluvionnaires, plateaux, bassins sédimentaires, formations cévenoles, moyennes montagnes et hautes montagnes pyrénéennes.

Les paysages de plaines et de terrasses alluvionnaires constituent une unité géographique où l'on retrouve les paysages de plaines alluviales récentes, des plaines littorales ainsi que des terrasses d'alluvions anciennes.

Les plaines alluviales récentes se répartissent essentiellement le long du Rhône et fleuves côtiers cévenols pour le Gard, le long de l'Hérault, de l'Orb et de petits fleuves côtiers pour l'Hérault, le long des fleuves côtiers du Roussillon à régime torrentiel en Pyrénées Orientales et restreint au niveau des rivières lozériennes pour La Lozère. Concernant le département de l'Aude, les plaines alluviales sont essentiellement localisées dans la haute et moyenne vallée de l'Aude qui sert d'axe de drainage des eaux provenant de la Montagne Noire, du Minervois et des Corbières. Ces espaces sont caractérisés par des zones de recreusements dont l'influence dépend de l'importance de la rivière, accompagnées de remblaiements par dépôts alluviaux récents pour les principaux fleuves et des apports colluviaux fréquents dans les petites vallées fluviatiles locales. D'un point de vue pédologique, cet ensemble est caractérisé par sa distribution aléatoire de sols, une forte hétérogénéité texturale avec la présence de gradients texturaux dans les plaines alluviales les plus évoluées. Les principaux types de sols de cet ensemble sont les Fluviosols et les Colluviosols.

Les terrasses d'alluvions anciennes sont constituées par l'ensemble de dépôts mis en place depuis le Pliocène jusqu'au Quaternaire (glacis plio-quaternaires). D'origine alluviale, ces formations sont ordonnées en système de terrasses dont l'étagement fluctuent depuis les abords

des plaines alluvionnaires pour les plus jeunes, jusqu'aux altitudes les plus élevées pour les plus âgées. Leur composition et leur granulométrie, très diversifiées, varient essentiellement selon les apports alluvionnaires et colluviaux affectant les paléo-bassins de sédimentation. Bien que la majorité des sols soient des Brunisols, la diversité des apports formant ces terrasses induit l'observation de multiples types de sols (Calcolsols, Arénosols, Colluviosols).

Les plaines littorales sont principalement localisées à proximité de la côte méditerranéenne et en pourtours des étangs littoraux s'étalant en continue du Gard jusqu'aux Pyrénées Orientales. Leur faible altitude par rapport au niveau de la mer leur confère non seulement une capacité de drainage externe très médiocre mais aussi une problématique de salinisation fréquente des nappes d'eau. De ce fait, les types de sols les plus présents sont majoritairement des Sodisols, des Salisols ainsi que des Arénosols, possiblement rédoxiques.

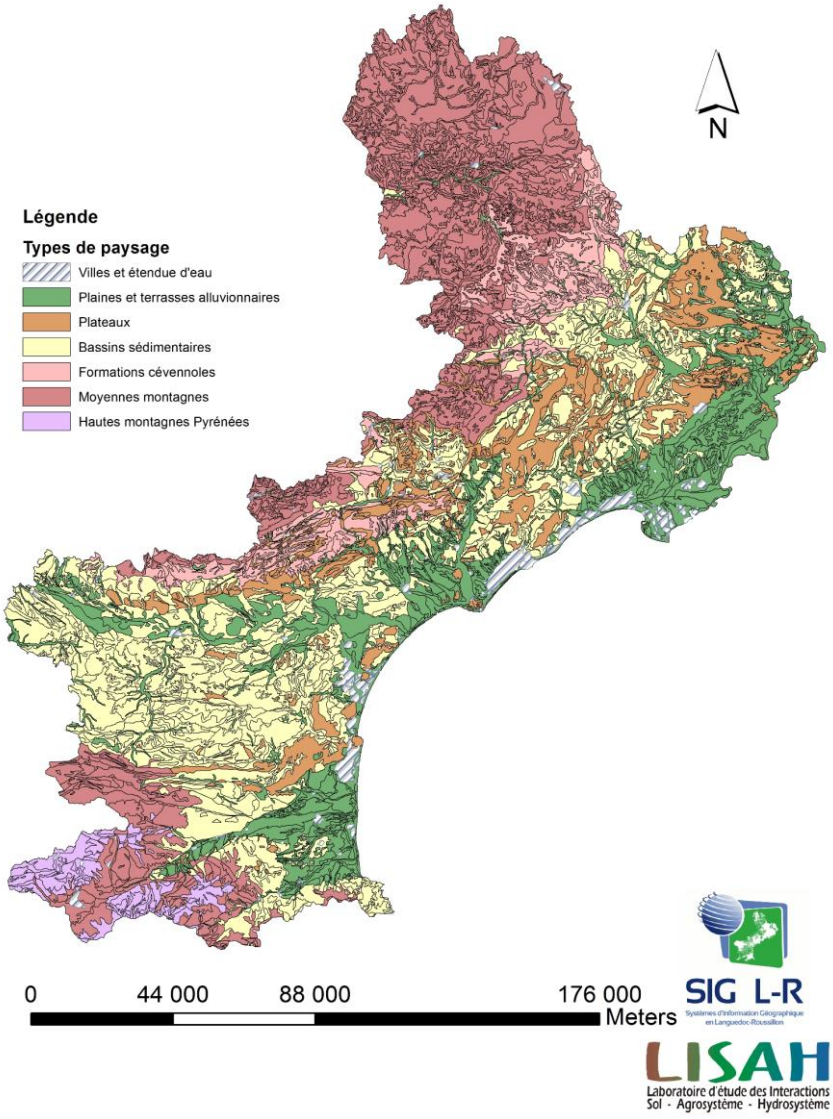


Figure 2.6. Répartition spatiale des pédo-paysages du Languedoc-Roussillon (d'apres Arrouays et Jamagne, 1989)

Les paysages de plateaux rassemblent les zones formées à la fois de plateaux mais aussi, de collines tabulaires limitées à 900 mètres d'altitude et les versants associés. Ces zones sont caractérisées par de vastes replats tabulaires liés à des formations volcaniques ou calcaires. Elles constituent des zones de transition entre les paysages de bassins sédimentaires et de plaines et terrasses alluvionnaires, et sont particulièrement présentes dans l'Hérault, le Gard et l'Aude. Les types de sols prédominants dans ces zones sont les Lithosols, les Rendosols et les Calcolsols.

Les principaux sols des paysages de plateaux sont généralement peu profonds, compris entre des affleurements du matériau parental et des profondeurs moyennes mais pouvant atteindre occasionnellement des profondeurs supérieures à 70 cm, liée principalement à une résistance à l'érosion du matériau parental moins importante.

Les paysages de bassins sédimentaires constituent un ensemble de paysages relativement hétérogènes composé de collines, versants et bassins. Ils se situent à l'interface entre les zones de moyennes montagnes et les plaines alluvionnaires. La superficie de cet ensemble représente environ un tiers de celle du Languedoc-Roussillon. Les paysages de bassins sédimentaires se caractérisent par une limite altimétrique inférieure à 900 mètres ; des pentes moyennes, des formes géomorphologiques ondulées recouvertes de végétation naturelle pour les zones collinaires et des pentes faibles, ainsi que de dépressions principalement cultivées dans les zones de bassins.

Les types de sols rencontrés sont relativement hétérogènes (exemple : Colluviosols, Brunisols, Lithosols, Fersialsols, Calcolsols, etc.) caractérisés par une profondeur moyenne à forte et une hétérogénéité texturale correspondant à la capacité du matériau parental à résister à l'érosion.

Les paysages de formations cévenoles sont localisés à la limite australe du Massif central. Les paysages de type cévenol sont formés par des ensembles collinaires sur des structures monoclinales dont les longues vallées encaissées sont dominées par des pentes ravinées et escarpées pour former les serres cévenoles.

Dans le but d'amener l'agriculture dans cet ensemble, l'homme a façonné de nombreuses terrasses afin d'y développer l'activité viticole.

L'influence de l'homme accompagnée de la nature du matériau parental ainsi que la pente des versant constituent les principaux facteurs de la répartition spatiale des sols dans ces zones. Les Brunisols, plus ou moins profonds, potentiellement composés d'horizons humifères, sont

présents en zone de faible pente, alors que les Lithosols et Rankosols sont plus propices à se développer sur des pentes fortes. Les terrasses de cultures sont, quant à elles, caractérisées par des Anthroposols.

Les paysages de moyennes montagnes regroupent l'ensemble de zones montagneuses dont l'altitude est comprise en 900 mètres et 1600 mètres pour le Massif Central et 2000 mètres pour les Pyrénées, couvrent une grande partie de la zone montagneuse. L'emprise de cet ensemble représente 25% de la superficie totale du Languedoc-Roussillon. Ces zones sont soumises à un climat contraignant (pluviométrie, faible température) favorisant une faible minéralisation des sols et, par conséquent, une concentration importante en carbone organique. Le climat est, par ailleurs, un facteur déterminant de répartition spatiale des sols et de la végétation.

Les types de sols présents sont variables avec des Lithosols et Rankosols caractérisés par une faible profondeur, les Brunisols et les Alocrisols, moyennement profonds avec des éventuelles accumulations organiques pouvant évoluer jusqu'à la formation d'horizons tourbeux (Histosols). En fonction de la végétation et de la nature du matériau parental, des Podzosols ocriques peuvent être observés.

Les paysages de hautes montagnes pyrénéennes, exclusivement localisés dans les Pyrénées, sont caractéristiques de la géomorphologie des zones montagneuses supérieures à 2 000 mètres d'altitude. La répartition des types de sols dépend principalement de la pente et du matériau parental présent mais les sols superficiels sont les plus fréquents (Lithosols).

Cependant, des unités podozolisées et humifères (Rankosols, Podozosols, Hystosols, etc.) sont également observées dans cet ensemble, en fonction de la végétation et du climat.

Emprise de la zone des études BRL Exploitation :

Les pédo-paysages où se trouvent les données BRL Exploitation sont : les plaines et terrasses alluvionnaires, les plateaux ainsi que les bassins sédimentaires. Cette diversité pédo-paysagère induit une hétérogénéité de la distribution spatiale des sols intéressante.

II. Les données environnementales

Le fonctionnement des modèles de cartographie numérique des sols repose sur le schéma conceptuel du modèle *scorpan* (McBratney et al., 2003), à savoir, l'utilisation des relations quantitatives entre la propriété de sol cible et les variables environnementales spatialisées, pour prédire la propriété cible sur la zone d'étude. La composition des variables environnementales se réfère aux composantes du modèle *scorpan*, soit : le relief, la géologie, le climat et l'occupation du sol. La sélection des données environnementales est basée sur les recommandations des travaux de thèse de Kévin Vaysse (2015).

II.1. Données relief

Les données du relief utilisées dans cette thèse sont issues de deux modèles numériques de terrain (MNT) fournissant une valeur altimétrique pour chaque nœud d'une grille de résolution fixe.

Le MNT issu de la mission Shuttle Radar Topographic Mission (SRTM), d'une résolution native de 90 m par 90 m a été utilisé pour la région Languedoc-Roussillon. A l'échelle locale, un MNT d'une résolution plus fine de 25 m par 25 m issu de la 2nde version de la BD ALTI® de l'Institut Géographique National (IGN), est utilisé.

Les MNT constituent la donnée source au développement d'indicateurs géomorphologiques permettant de caractériser au mieux le lien entre la formation des sols et le relief. L'utilisation du module « Terrain Analysis » proposé par le logiciel SAGA-GIS (SAGA, 2014) a permis de générer les indicateurs géomorphologiques suivants : la pente, les indices de courbures du plan et du profil de la pente, l'indice de position topographique (TPI), l'indice topographique d'humidité (TWI), le multi-resolution valley bottom flatness (MrVBF) et le multi-resolution ridge top flatness (MRRTF).

Ces indicateurs ont été dérivés respectivement à 90 m x 90 m et 25 m x 25 m de résolution, à l'échelle régionale et locale.

II.2. Données lithologiques

Les données lithologiques sont directement issues des travaux de thèse de Kévin Vaysse (2015) visant à produire des indicateurs lithologiques quantitatifs à partir de données qualitatives (données textuelles et non codifiées) fournies pour chaque unité cartographique de la version numérisée de carte géologie de la France (1/50 000^{e}) réalisée par le BRGM (Figure

2.7). L'élaboration de ces indicateurs lithologiques a été d'autant plus motivée par la difficulté d'exploitation de données textuelles et non codifiées en cartographie numérique des sols.

L'élaboration d'indicateurs lithologiques a été réalisée par expertise pédologique visant à associer des propriétés de sol « diagnostics », issues de profils de sol, à des unités cartographiques afin d'obtenir les indicateurs lithologiques suivant : la dureté (hardness) estimée par la profondeur de sol, la minéralogie par le pH de l'horizon le plus profond, et la texture par la teneur en sables de l'horizon le plus profond.

La dureté caractérise la sensibilité de la roche mère à l'altération selon trois modalités : faible, forte et indéterminée. La minéralogie permet d'identifier le degré d'acidité/alcalinité selon trois modalités : acide, alcaline et indéterminée. La texture détermine la granulométrie du produit d'altération de la roche mère selon trois modalités : grossière, fine et indéterminée.

Ces indicateurs vectoriels ont ensuite été rastérisées via le package R *raster* a une résolution de 90 mètres par 90 mètres.

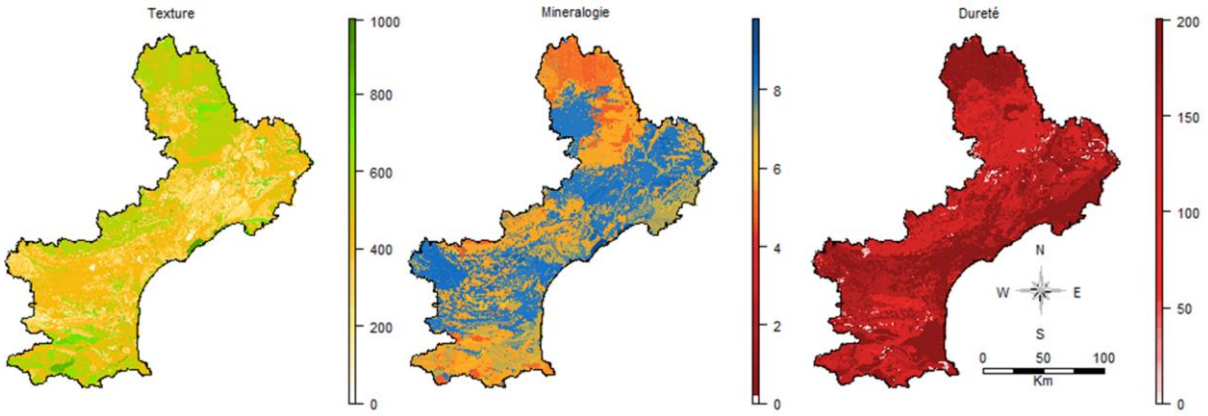


Figure 2.7. Représentation cartographique des indicateurs lithologiques (d'après Vaysse, 2015)

A l'échelle locale, ces indicateurs ont été ré-échantillonnés par interpolation bilinéaire via le package R *raster* à une résolution de 25 mètres par 25 mètres (Hijmans, 2014).

II.3. Données climatologiques

Les données climatiques utilisées proviennent d'une base de données en accès libre issue du programme WorldClim (Hijmans et al., 2005). Cette base de données compile des observations provenant de plus de 1 000 stations météorologiques réparties sur le globe (sauf l'Antarctique). Le format de cette base de données est un ensemble d'images (rasters) réunissant plusieurs indicateurs climatiques à une résolution de 1 km par 1 km. Les surfaces climatiques

ont été interpolées par l'utilisation de régressions splines via le package ANUSPLIN en utilisant les informations relevées entre 1950 et 2000.

Dans cette thèse, ont été retenues les indicateurs climatiques standards tels les précipitations totales annuelles moyennes ainsi que les températures moyennes minimales et maximales annuelles (Figure 2.8). Ces températures ont été calculées à partir des températures minimales et maximales mensuelles.

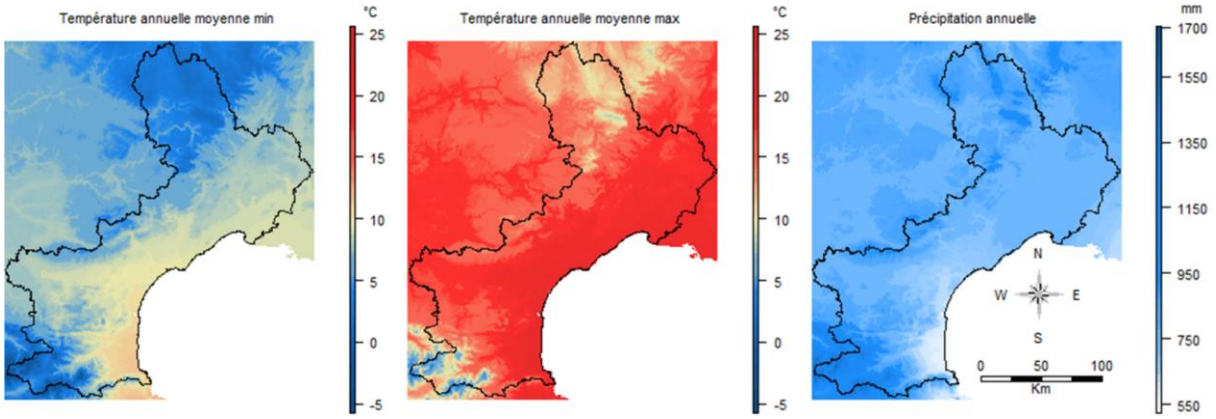


Figure 2.8. Cartographie des indices basiques climatiques : la température annuelle moyenne minimale, la température moyenne maximale et les précipitations moyennes annuelles (d'après Vaysse, 2015)

Ces indicateurs basiques sont complétés par des indicateurs plus sophistiqués : l'indice d'Emberger (Emberger, 1930) et l'indice d'aridité de De Martonne (De Martonne, 1927) (Figure 2.9). L'indice climatique d'Emberger est calculé à partir d'un rapport pluviométrique capable de caractériser différents types de climats méditerranéens, en utilisant l'équation suivante :

$$Q = P \times 100 / (T_M^2 - T_m^2) \qquad \text{(Eq.2.1.)}$$

Où P est la précipitation totale annuelle moyenne, T_M^2 est la valeur maximale des températures maximales moyennes mensuelles et T_m^2 est la valeur minimale des températures minimales moyennes mensuelles. A l'échelle locale, la majorité de la superficie est occupée par des zones subhumides définies par des valeurs de Q compris entre 40 et 80. A l'échelle régionale, cet indicateur atteint notamment des valeurs extrêmes, principalement dues au climat montagnard présent dans les Pyrénées Orientales et le Massif Central (voir section I.3).

L'indice d'aridité de De Martonne se calcule selon l'équation suivante :

$$I = P\ /(T + 10) \qquad \text{(Eq.2.2.)}$$

Où P est la précipitation totale annuelle moyenne et T, la température moyenne annuelle. Concernant l'aridité, les zones présentant un I inférieur à 30 sont définies comme semi-arides. Ce sont des zones bien représentées à l'échelle régionale et locale.

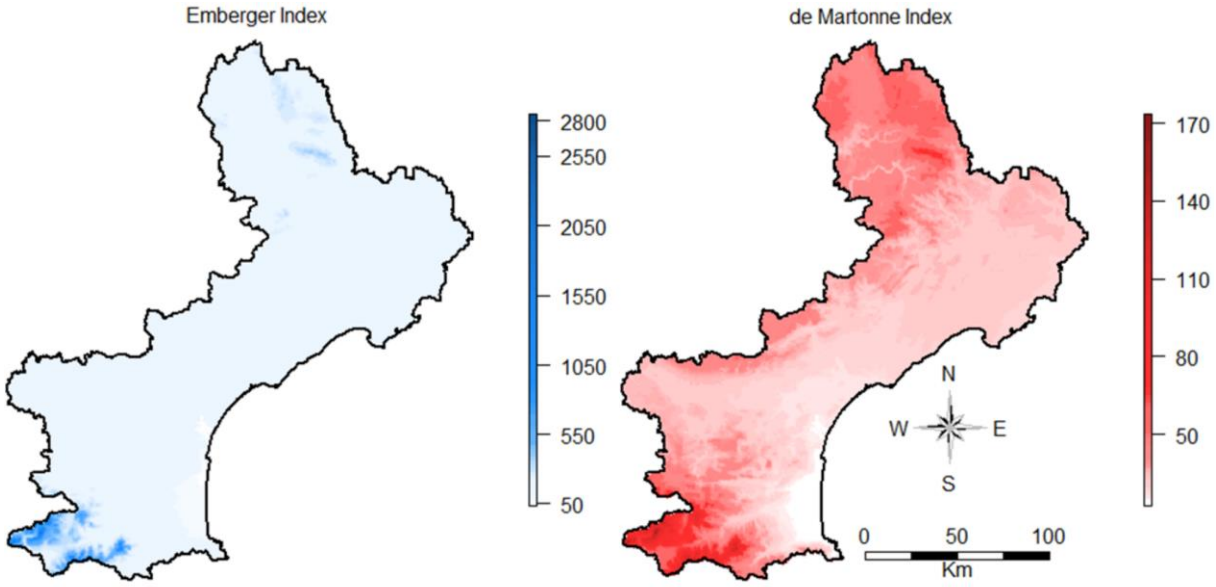


Figure 2.9. Cartographies des indices climatiques d'Emberger et de De Martonne estimés à partir des données WorldClim

Afin d'harmoniser la résolution des indicateurs climatiques à celle de l'ensemble des données environnementales, ces indicateurs ont été ré-échantillonnés par interpolation bilinéaire via le package R *raster* (Hijmans, 2014) à des résolutions de 90 m x 90 m.

En considérant la résolution spatiale native des données WorldClim (1 km x 1 km), l'exhaustivité de ces données ne sera pas assez précise pour une utilisation à l'échelle locale dans cette thèse.

II.4. Données occupation du sol

L'occupation du sol en Languedoc-Roussillon est issue du traitement d'images satellitaires Landsat 7 de 2006, réalisé et distribué par Open IG (ex-SIG L-R). La classification de l'occupation du sol a été simplifiée en 9 grands types d'occupations du sol, constitués de 43 classes distinctes. Ce choix de simplification a été motivé par la nécessité de rendre l'information plus pertinente vis-à-vis de la représentation des interactions sol-occupation du sol connues. La Figure 2.10 reprend la répartition spatiale des grands types d'occupation du sol.

Afin de faire correspondre la donnée d'occupation du sol avec le reste des données environnementales, la cartographie vectorielle a été rastérisée via le package R *raster* respectivement à des résolutions de 90 m x 90 m et de 25 m x 25 m, à l'échelle régionale et locale.

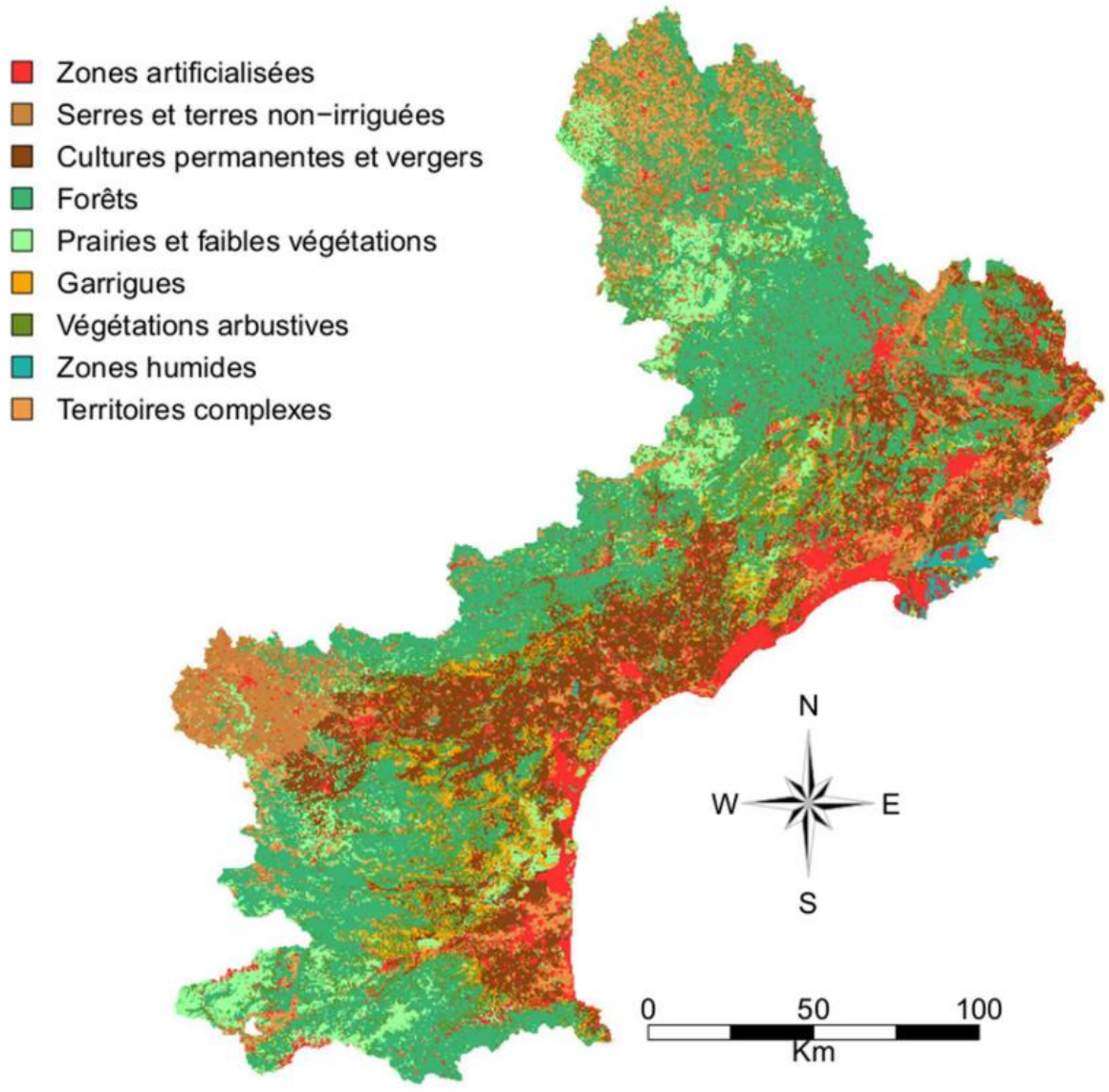


Figure 2.10. Cartographie des grands types d'occupation du sol du Languedoc-Roussillon (d'après Vaysse, 2015)

II.5. Données pédologiques

L'information liée à la distribution des sols provient de la cartographie au 1/250 000^{e} du Référentiel Régional Pédologique du Languedoc-Roussillon (RRP) (voir explications III.1). Cette cartographie vectorielle est composée de 396 Unités Cartographiques de Sols (UCS) (Figure 2.11) comprenant une codification permettant son utilisation dans la cartographie numérique des sols.

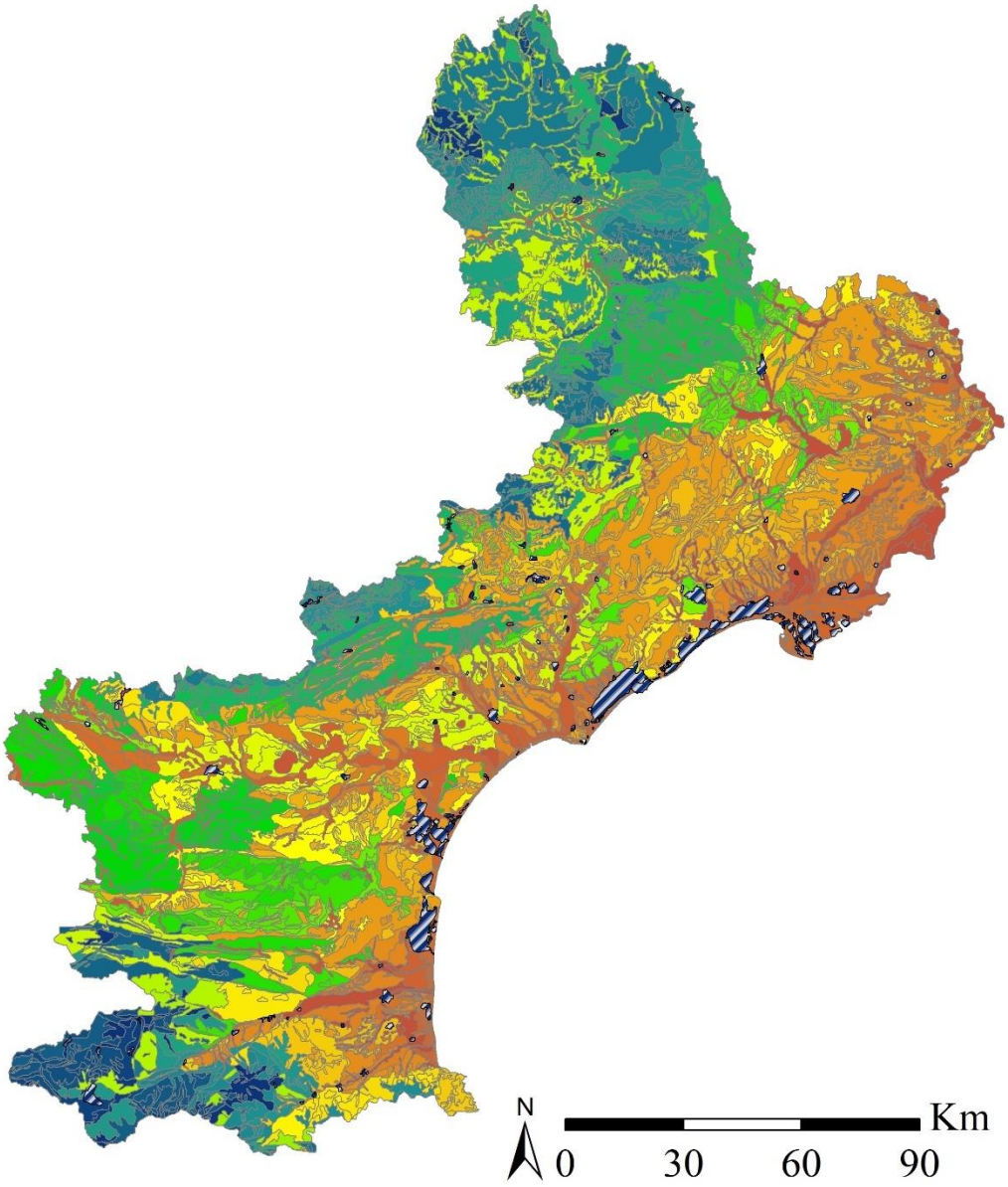


Figure 2.11. Cartographie des unités cartographiques de sol du Référentiel Régional Pédologique du Languedoc-Roussillon (d'après Arrouays et Jamagne, 1989)

Dans le but d'harmoniser les types de données environnementales utilisées dans les modèles, cette cartographie vectorielle a été rastérisée via le package R *raster* selon une grille de résolution 90 m x 90 m pour l'échelle régionale et une grille de résolution 25 m x 25 m pour l'échelle locale.

Le Tableau 2.1 reprend l'intégralité des 18 variables environnementales utilisées dans la thèse à l'échelle régionale et locale, rastérisées aux résolutions respectives.

Tableau 2.1. Tableau récapitulatif des données environnementales utilisées dans la thèse

Noms	**Résolution native**	**Source**
Relief		
MNT	90 m / 25 m	SRTM / BD ALTI
Pente	90 m / 25 m	SRTM / BD ALTI
MRVBF	90 m / 25 m	SRTM / BD ALTI
MRRTF	90 m / 25 m	SRTM / BD ALTI
TWI	90 m / 25 m	SRTM / BD ALTI
Courbure du profil	90 m / 25 m	SRTM / BD ALTI
Courbure du plan	90 m / 25 m	SRTM / BD ALTI
TPI	90 m / 25 m	SRTM / BD ALTI
Géologie		
Dureté	1 / 50 000	Carte géologique (BRGM)
Texture	1 / 50 000	Carte géologique (BRGM)
Minéralogie	1 / 50 000	Carte géologique (BRGM)
Climat		
Température minimale	1000 m	WorldClim
Température maximale	1000 m	WorldClim
Précipitation	1000 m	WorldClim
Indice de de Martonne	1000 m	WorldClim
Indice d'Emberger	1000 m	WorldClim
Organismes		
Occupation du sol	30 m	Landsat 7
Sol		
Carte pédologique	1:250 000	RRP

III. Les données pédologiques

Les données pédologiques utilisées dans cette thèse proviennent de deux sources distinctes : les données issues du RRP Languedoc-Roussillon et les données pédologiques anciennes détenues par BRL Exploitation réparties majoritairement sur la plaine littorale languedocienne (Figure 2.1). Ces données sont accompagnées de leur modalité respective d'estimation du réservoir utile utilisée pour chacune de ces sources.

III.1. Données utilisées pour une approche de cartographie numérique des sols à l'échelle régionale

III.1.1. Les données pédologiques régionales utilisées : présentation du RRP

« Issu du programme « Inventaire, Gestion et Conservation des Sols » (IGCS) visant à cartographier les sols des régions françaises à différentes échelles, le RRP a pour objectif une cartographie complète et une harmonisation de la couverture nationale des sols à l'échelle du 1/250 000. Cette harmonisation est possible par la mise en place de spécifications précises pour la description des profils » (Vaysse, 2015)

La conception du RRP du Languedoc-Roussillon (RRP LR) a été réalisée dans les années 1980 s'appuyant sur les bases méthodologiques de la carte des pédopaysages au 1/250 000^{e} élaborée par Arrouays et Jamagne (1989). Opérationnel dans les années 1990, ce fût le premier RRP à être diffusé aux utilisateurs en France. Mise au format DoneSol en 2007, le RRP LR comprend : i) une carte pédologique au 1/250 000^{e} composée de 396 Unités Cartographiques de Sols et ii) 2128 profils de sol collectés entre 1964 et 1994 (1976 profils de sol avec analyses des propriétés de sol en laboratoire), abondés par 470 profils de sol suite aux travaux de thèse de Kévin Vaysse (2015), issus de programmes de numérisations des données pédologiques anciennes, soit un total de 2598 profils de sol.

III.1.2. Estimation du réservoir utile

A l'échelle régionale, le réservoir utile a été estimé selon la formulation de Cousin et al. (2003) (section I.1., Chapitre 1).

Les fonctions de pédotransferts nationales issues des travaux de Román Dobarco et al. (2019) ont été sélectionnées afin d'estimer les humidités à la capacité au champ et au point de flétrissement permanent. Ce choix a été motivé d'une part, pour leurs applicabilités en Languedoc-Roussillon et d'autre part, car elles estiment les humidités volumiques, à la capacité au champ (Equation 2.3.) et au point de flétrissement permanent (Equation 2.4.), en intégrant

l'information de la densité apparente. Ces fonctions de pédotransfert utilisent les teneurs en argile et en sables totaux (%) en variables prédictives et sont calculées selon :

$$\hat{\theta}_{ri} = 0.278 + 2.45\ 10^{-3} * Argile - 1.35\ 10^{-3} * Sable\ tot \qquad \text{(Eq.2.3.)}$$

$$\hat{\theta}_{wi} = 0.08 + 4.01\ 10^{-3} * Argile - 2.93\ 10^{-4} * Sable\ tot \qquad \text{(Eq.2.4.)}$$

Où $\hat{\theta}_{ri}$ est l'humidité volumique à la capacité au champ et $\hat{\theta}_{wi}$ l'humidité volumique au point de flétrissement permanent.

L'utilisation de ces fonctions de pédotransfert modifie légèrement l'estimation du RU avec la simplification de la densité apparente :

$$RU = \sum_{i=1}^{n} dh_i * \left(\frac{100 - st_i}{100}\right) * \left(\hat{\theta}_{ri} - \hat{\theta}_{wi}\right) \qquad \text{(Eq.2.5.)}$$

III.2. Données issues des études pédologiques BRL Exploitation

III.2.1. Caractéristiques générales des données

III.2.1.1. Historique des données

De 1957 à 1992, la Compagnie National d'Aménagement de la Région du Bas-Rhône et du Languedoc (CNARBRL, actuellement BRL) a réalisé de nombreuses études et cartes pédologiques. Ces études et cartes pédologiques avaient pour objectif de répondre, à différents niveaux de précision, aux questions posées aux différents stades de l'aménagement et de la mise en valeur agricole, tout en apportant une connaissance générale des sols pour d'autres usages éventuels.

Les études pédologiques ont été déclinées à différentes échelles : i) études pédologiques à petite échelle, ii) les études pédologiques à moyennes échelles, et iii) les études pédologiques à grandes échelles.

Les études pédologiques à petite échelle, à savoir au 1/200 000^{e}, 1/100 000^{e} ou encore 1/50 000^{e} sont des études dîtes de « reconnaissance ». Elles ont pour but de localiser approximativement les zones cultivables, éventuellement irrigables, d'en estimer la superficie et les possibilités agricoles, et de prévoir le dimensionnement des travaux dans une perspective d'intensification des cultures via l'irrigation. Dans le cadre des études, le nombre de profils analysés est uniquement constitué par les 200 profils types issus des synthèses pédologiques existantes des cartes au 1/100 000^{e} de Montpellier et d'Arles.

Les études à moyennes échelles regroupent : les études prospectives, les études de reconnaissance, la carte départementale des terres agricoles (CDTA), ainsi que les études complètes. Les études prospectives, au 1/20 000^{e} et 1/25 000^{e} sont des cartes établies d'après des observations et l'interprétation des caractéristiques pédologiques visibles ou décelables en surface et en ne faisant qu'un nombre faible d'observations (sondages pédologiques), et sans analyses. Ces cartes sont fournies uniquement avec une légende pédogénétique et/ou de caractéristiques principales des sols. Bien que les limites des unités de sols ne soient pas précises, ces cartes améliorent, néanmoins, la connaissance des sols du secteur étudié. Les cartes de reconnaissances sont également présentes au 1/20 000^{e} et 1/25 000^{e} avec, cependant, un nombre plus ou moins réduit de profils (1 profil de sol tous les 1 à 10 km²) fournit avec un texte de commentaire ou seulement la légende. La carte départementale des terres agricoles (CDTA) a été conçue entre 1982 et 1987. Cette carte s'appuie sur une prospection de surface au 1/25 000^{e} avec un sondage à la tarière tous 0,5 km² et un profil de sol tous les 5 à 10 km². Enfin, les études pédologiques complètes constituent un inventaire complet des sols de la zone étudiée. Elles expriment les caractéristiques de chaque type de sol, leurs caractéristiques hydrodynamiques, nécessaires au calcul des réseaux d'irrigation, leurs caractéristiques agronomiques, leurs potentialités et les améliorations nécessaires ou souhaitables pour leur utilisation optimale. Ces études sont effectuées en deux phases :

- Une première cartographie sur le terrain avec photo-interprétation qui permet de délimiter les différents types de sols,
- Le creusement et la description des tranchées pédologiques (4 profils au km²) avec un prélèvement d'échantillons pour des analyses et des mesures de laboratoire avec élaboration des cartes.

Les études à grandes échelles présentées au 1/5 000^{e} et 1/2 000^{e} (quelques exceptions à 1/1 000^{e} et 1/10 000^{e}) sont des cas d'études pédologiques de détail sur le périmètre irrigué de BRL. Ce type d'étude a été réalisé jusqu'à 1986 dans le but de déterminer les caractéristiques principales des sols et de les délimiter au niveau de chaque parcelle, de servir d'assistance technique à l'agriculteur, notamment lors de la mise en eau d'une exploitation ou d'un changement de culture. Ce type d'étude était réalisée sur une zone donnée au préalable des travaux hydrauliques de mise en irrigation, pour permettre un choix judicieux des espèces et des porte-greffes, des doses d'irrigation et des travaux du sol à effectuer selon la nature des sols. La densité d'une telle étude est très importante avec 100 profils de sol au km² accompagnés pour chacun d'une fiche de description et des résultats d'analyses de chaque horizon (analyses des

propriétés de sol physiques, chimiques et hydrodynamiques). Ces études sont présentées selon deux types de documents :

- Un fascicule répertoriant les profils de sol avec un fond cadastral avec topographie, généralement au 1/2 000e, replaçant les profils de sol avec une délimitation des zones de dose pratique maximale d'arrosage homogènes
- Une carte au 1/5 000e (caractéristiques pédologiques, hydrodynamiques, etc.)

Finalement, l'ensemble de ces études s'étendent sur une surface de 6 636 km² comprenant 360 km² des études à petites échelles (1/2 000e et 1/5 000e), 5 296 km² des études à moyennes échelles (1/25 000e et 1/20 000e) et 980 km² d'étude pédologiques à grandes échelles.

III.2.1.2.Concept du réservoir utile proposé par BRL Exploitation

Historiquement, le CNARBRL a utilisé une approche différente de celle actuellement en vigueur (Cousin et al., 2003) pour exprimer le terme de rétention en eau de la terre fine ($\theta r_i - \theta w_i$) selon l'équation suivante :

$$RU = \sum_{i=1}^{n} dh_i * bd_i * \left(\frac{100 - st_i}{100}\right) * (b_i * EqW_i) \qquad \text{(Eq. 2.6.)}$$

Où EqW_i est l'humidité équivalente (en g.100g^{-1}) correspondant approximativement au terme θr_i et b_i est le coefficient de texture qui exprime l'humidité au point de flétrissement permanant en pondérant EqW_i pour prendre en compte la teneur en eau qui n'est plus disponible par la plante, c'est-à-dire, au-delà du point de flétrissement permanent. Les autres termes (épaisseur, densité apparente, pierrosité) sont conservés.

III.2.1.3. Profils de sol

Les profils de sol sont issus des descriptions de fosses pédologiques. Leur profondeur de prospection est fixée à 2,20 m (dans le cadre d'étude des potentialités viticoles) pour une profondeur effective d'observation estimée à 1,40 m. Les profils de sol sont fournis avec un plan de situation et une photographie de la fosse pédologique (utilisation pour photo-interprétation).

Pour chaque profil de sol, l'information est présentée en quatre sections : i) les informations générales du profil, ii) les caractéristiques hydrodynamiques et iii) les caractéristiques pédologiques de la description du profil et les résultats d'analyses granulométriques et iv) les résultats d'analyses chimiques.

Les informations générales (encadré noir Figure 2.12.a) du profil de sol regroupent :

- Le(s) pédologue(s) en charge de la description de la fosse pédologique,
- La date de prélèvement,
- La classification du profil de sol dans les études pédologiques ainsi que l'échelle de la carte pédologique contenant le profil,
- Les informations géographiques : nom de la commune et coordonnées longitude et latitude en Lambert III (Zone Sud),
- Les indications géomorphologiques : altitude, situation topographique (pente, sommet, bas fond, etc.), inclinaison (pente faible, moyenne ou forte), exposition (direction de la normale à la pente), relief (uni, vallonné, etc.), érosion (traces et intensité),
- L'occupation du sol (type de culture et relevée de la présence et fréquence des espèces),
- L'état de surface,
- La géologie (nature du matériau parental),
- Capacité de drainage et niveau de la nappe phréatique,
- Origine et type de sol.

Les caractéristiques hydrodynamiques (encadré vert de la Figure 2.12.a) regroupent les résultats d'analyses de :

- Les profondeurs entre lesquelles sont réalisés les prélèvements correspond à une couche homogène de sol,
- Pourcentages pondéraux (%) des éléments grossiers ayant un diamètre supérieur à 20 mm, et compris entre 2 et 20 mm et supérieur à 2 mm (= pourcentage total des éléments grossiers) par rapport à la masse de l'ensemble éléments grossiers et terre fine,
- Pourcentage pondérale de terre fine (%) estimée par la différence avec le pourcentage massique des éléments grossiers,
- Les vitesses d'infiltrations (en $cm.s^{-1}$) K_1 qui est une mesure de la vitesse d'infiltration sur le bloc Vergières lorsque l'écoulement se manifeste et K_2, qui est la mesure réalisée 3h après K_1, quand la vitesse d'infiltration diminue,
- La densité apparente, Δa (en $kg.dm^{-3}$), correspond au poids de terre fine contenu dans une unité de volume. La densité est mesurée par un prélèvement de sol non remanié par cube Vergières (Bourrier, 1965). Il est possible que cette mesure ne soit pas réalisable selon l'abondance des éléments grossiers. Dans ce cas, la densité apparente est estimée

comme $1.6 * \%\, Terre\ fine$ (où 1.6 correspond à la densité moyenne d'une terre fine sans cailloux),

- La densité réelle du sol, Δr (en $kg.dm^{-3}$), généralement avoisinant la valeur 2,6 ou bien se référant à l'indice de compacité variant entre 0 et 6 avec 0 indiquant un sol très compact et 6, un sol avec une porosité importante,
- Humidité équivalente (H.éq) ou coefficient de rétention (Cr) (en $g.100^{-1}$) selon les fiches de sol. Cette humidité correspond à l'eau retenue par un échantillon de terre fine (tamisée à 2 mm) soumis à un champ de force de 1000 fois l'accélération de la pesanteur par une centrifugeuse, équivalent à une pression de -100 kPa appliquée sur l'échantillon. L'humidité équivalente est une approximation de l'humidité à la capacité au champ, qui était anciennement mesurée à -100 kPa (pF = 3) (Baize and Jabiol, 1995),
- La dose d'irrigation (Dose) (en $m^3.ha^{-1}$ pour 1 cm de sol) est estimée comme étant égale au 2/3 du RU.

a) b)

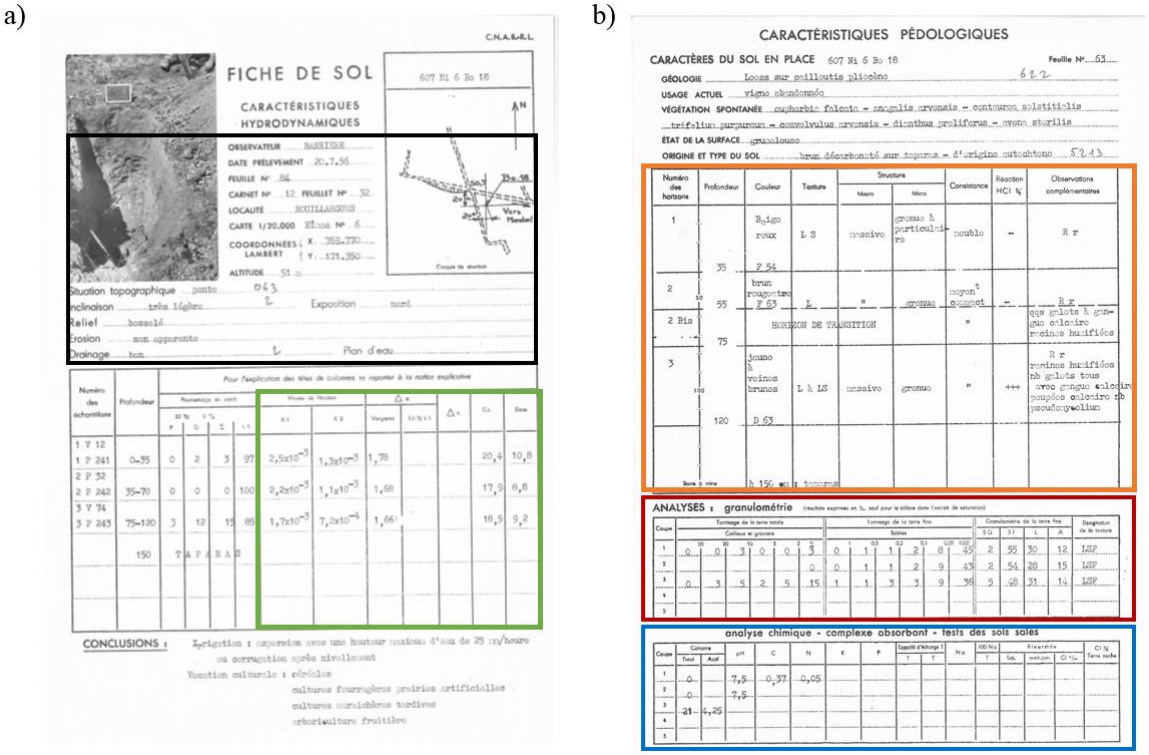


Figure 2.12. Exemple a) d'une description de fosse pédologique et b) des résultats d'analyses physiques, chimiques et hydrodynamiques effectuées en laboratoire

Les caractéristiques pédologiques (encadré orange de la Figure 2.12.b) rassemblent les résultats de descriptions et analyses granulométriques de sols (encadré rouge de la Figure 2.12.b) regroupant :

- La profondeur des horizons
- La couleur des horizons à la fois observée par expertise pédologique (information textuelle) et examinée en laboratoire sur un échantillon sec et codifiée selon le code expolaire de Cayeux et Taylor (Cailleux et Taylor, 1963).
- La macrostructure qui indique l'agencement des éléments du sol, observé sur la face du profil. La microstructure décrit la façon dont le sol se désagrège quand il est travaillé.
- La consistance des horizons qui définit la cohésion du sol (meuble, friable, cimenté, etc.)
- La réaction à l'HCl ½ qui permet d'indiquer la présence/absence de calcaire en fonction de degré d'effervescence.
- La texture de l'horizon est renseignée sous deux formes : une appréciation de la texture de l'échantillon par expertise pédologique (sur le terrain) (encadré orange de la Figure 2.12.b) ou par analyse granulométrique de la terre fine et des éléments grossiers en laboratoire (encadré rouge de la Figure 2.12.b) composée par:
    - Le pourcentage pondéral des éléments grossiers calculé à partir de la masse de terre fine et des éléments grossiers, divisé en deux fractions granulométriques : graviers (ø entre 2-20 mm) et cailloux (ø > 20 mm), mesurées par tamisage,
    - Les pourcentages pondéraux des cinq fractions granulométriques (argile (ø < 0,002 mm), limon fin (ø entre 0,002-0,02 mm), limon grossier (ø entre 0,02-0,05 mm), sable fin (ø entre 0,05-0,2 mm), sable grossier (ø entre 0,2-2 mm)) calculées à partir de la masse de terre fine. La quantification des fractions granulométriques est réalisée par la méthode à la pipette de Robinson. Il s'agit d'une méthode qui se base sur la vitesse des chutes de particules solides en suspension dans un fluide,
    - La désignation d'une classe de texture en replaçant les résultats d'analyses granulométriques de la terre fine dans le triangle de texture élaboré par BRL (Figure 2.13).

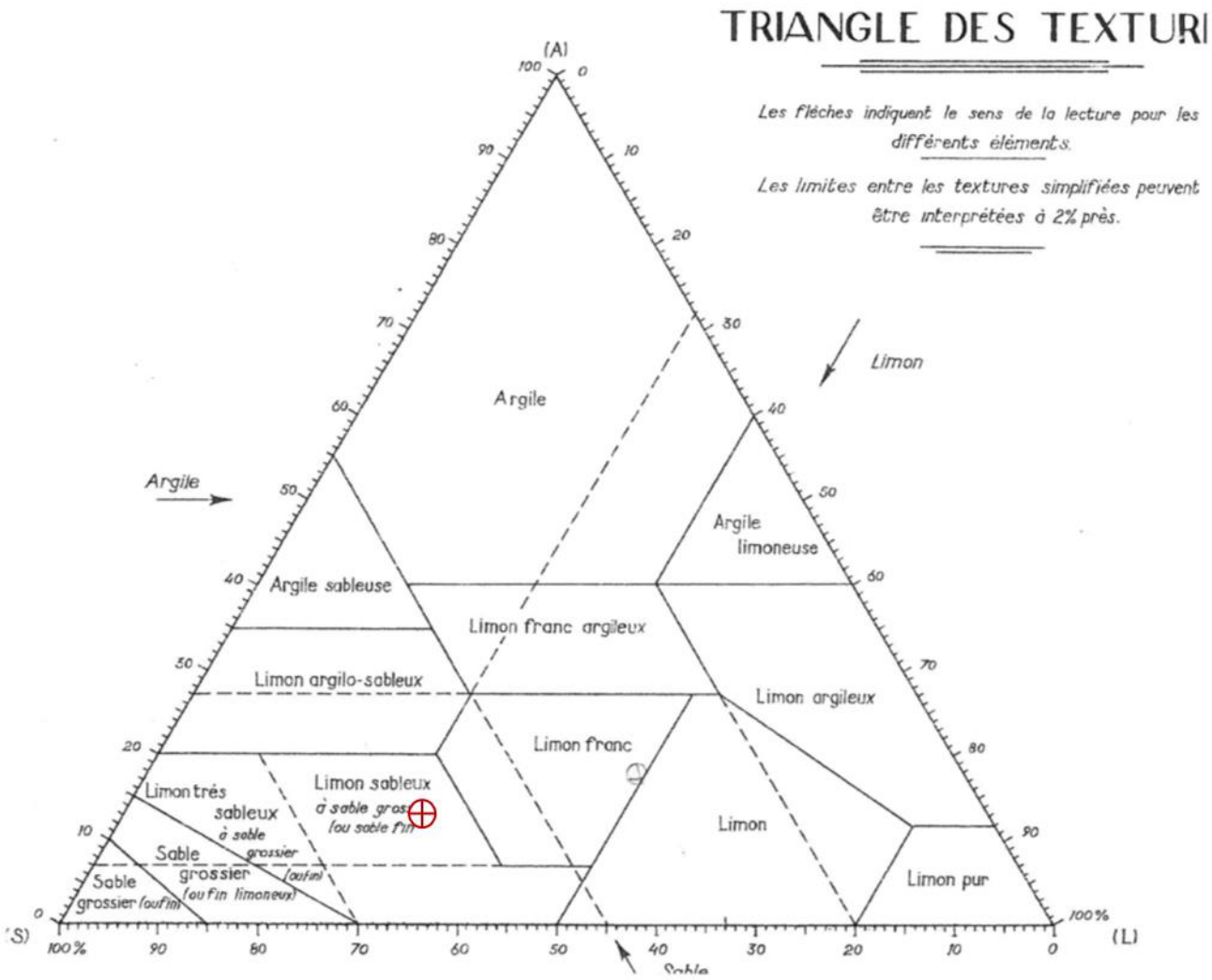


Figure 2.13. Triangle des textures élaborés par BRL avec un exemple de replacement du premier échantillon de la Figure 2.12.

Les propriétés chimiques du sol mesurées en laboratoire (encadré bleu de la Figure 2.12.b) sont :

- Le taux de calcaire actif et totaux,
- Le pH,
- Les teneurs en carbone, azote, potassium, phosphore,
- La capacité d'échange cationique (mEq/100g),
- Les bases échangeables,
- La conductivité,
- Le taux de Chlore dans la terre fine et réel,
- Le pourcentage de saturation du complexe en sodium.

Afin d'estimer la rétention d'eau de la terre fine et par conséquent le RU (Equation 2.6.), il est nécessaire d'obtenir le coefficient textural b_i. Pour ce faire, le coefficient textural est déterminé par une fonction de pédotransfert selon l'équation suivante (Equation 2.7.):

$$b = \frac{RU_{elem}}{EqW * bd} = \frac{\left(bd * \left(\frac{100 - st}{100}\right) * (0.45 * EqW)\right)}{EqW * bd} \qquad \text{(Eq. 2.7.)}$$

Où RU_{elem} est la valeur du réservoir utile élémentaire, soit le RU calculé pour un millimètre de sol (mm.mm^{-1}), EqW l'humidité équivalente, st la pierrosité, bd la densité apparente et 0.45 le coefficient utilisé pour approximer le point de flétrissement permanent comme étant équivalent à 55 % de l'humidité équivalente. Chacun de ces termes a été mesuré/calculé pour chaque horizon des profils de sol, inclus dans le périmètre BRL, par classe texturale du triangle des textures GEPPA (et non selon le triangle de texture BRL). Les valeurs du coefficient textural obtenues pour les horizons d'une même classe ont ensuite été moyennée. De ce fait, chaque classe texturale a un coefficient textural variant entre 0.8 pour le sable et 0.33 pour l'argile.

Ce traitement est aussi appliqué à l'humidité équivalente ainsi qu'à la densité apparente (Tableau 2.2).

Cependant, afin de renseigner les valeurs moyennes d'humidité spécifique, de densité apparente et de coefficient textural pour chaque classe texturale BRL, un rattachement entre les classes texturales BRL et GEPPA a été nécessaire. Ce rattachement a été effectué en sélectionnant la plus petite différence des valeurs modales de texture (argile, limon et sable) entre les classes texturales BRL et GEPPA. Il est important de noter que certaines classes texturales BRL ne sont pas rattachées aux classes texturales GEPPA en raison d'un effectif de classes moindre (18 vs 14).

Tableau 2.2. Valeurs des moyennes des humidités équivalentes, des densités apparentes et des coefficients texturaux en fonction des classes texturales

Classe texturale		Humidité équivalente ($g.100^{-1}g$)	Densité apparente ($kg.dm^{-3}$)	Coefficient textural
GEPPA	*BRL*			
S pur	-	5	1,7	0,8
SS	SF/SG/SFL	8	1,6	0,7
S	SL	12	1,6	0,6
Sl	LS	14	1,6	0,55
Ls	L	15	1,5	0,45
LL	LP	21	1,4	0,5
Sa	SA	16	1,6	0,5
Sal	-	17	1,6	0,45
LSa	L.Fr	19	1,5	0,45
L	-	23	1,4	0,45
AS	LAS	23	1,5	0,45
LAS	-	24	1,5	0,45
LA	-	25	1,4	0,45
As	AS	27	1,5	0,45
Als	L.Fr.A	28	1,5	0,45
Al	AL	32	1,4	0,45
A	A	35	1,3	0,4
AA	-	45	1,2	0,33

III.2.1.4. Sondages pédologiques

Les sondages pédologiques ont été réalisés à la tarière à main. Les profondeurs de prospection et d'observation de cet outil de mesure sont égales à 1,20 m. A la différence des profils de sol, les coordonnées spatiales des sondages pédologiques ne sont pas directement disponibles et un géoréférencement manuel est nécessaire.

Les sondages pédologiques présentent uniquement une description des caractéristiques pédologiques des horizons de sol :

- Localisation (secteur, borne, mappes ; voir explication section III.3,
- Profondeur des horizons,
- Pourcentage massique des graviers (ø entre 2-20 mm) et cailloux (ø > 20 mm) estimés visuellement,
- Pourcentage pondéral de terre fine calculé par la différence avec les pourcentages pondéraux cumulés des graviers et des cailloux,

- Appréciation de la terre fine sur terrain par expertise pédologique permettant l'obtention d'une classe texturale selon le triangle des textures BRL ou un indicateur de classe variant entre 1 et 5, où 1 = terre sableuse et 5 = terre argileuse,
- Le pH du sol,
- La densité apparente,
- L'humidité équivalente,
- La capacité de rétention (en %) correspond au volume d'eau retenu par le sol ressuyé, estimée à partir du produit entre l'humidité équivalente et la densité apparente
- La dose d'irrigation unitaire (mm/cm) et cumulée (mm).

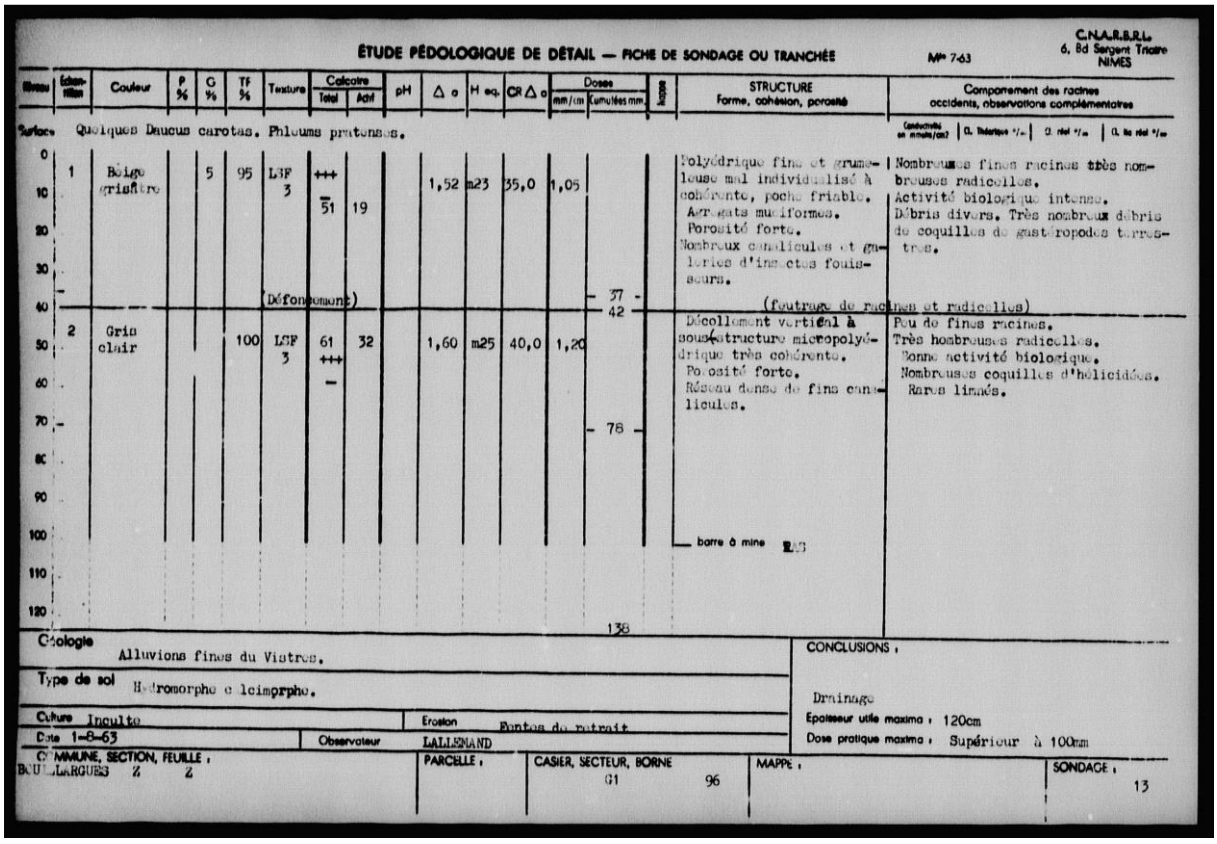


Figure 2.14. Exemple d'une description de sol d'un sondage de sol

Afin d'estimer le RU, l'humidité spécifique, la densité apparente ainsi que le coefficient textural b_i ont été déterminés en fonction de la classe texturale de l'horizon, estimée préalablement par le pédologue (Tableau 2.2).

Le Tableau 2.3 regroupe l'ensemble des propriétés présentées dans les données pédologiques de BRL Exploitation ainsi que leur méthode d'estimation et/ou de mesure.

Tableau 2.3. Tableau récapitulatif des propriétés issues des profils de sol et des sondages

Caractéristiques	Type d'observation de sol	
	Profils de sol	Sondages
Pédologique		
Profondeur	M	M
Eléments grossiers	M	E
Terre fine	M	E
Texture	M	E
Granulométrie	M	-
Densité apparente	M	E
Densité réelle	M	-
Structure	M	E
Réaction HCl 1/2	M	-
Hydrodynamiques		
Vitesses de filtrations	M	-
Humidité équivalente	M	E
Dose d'irrigation	M	-
Chimique		
Calcaire actif	M	-
pH	M	M
Teneur C/N/K/P	M	-
CEC	M	-
Bases échangeables	M	-
Conductivité	M	-
Taux de Cl	M	-
Saturation en sodium	M	-

M: mesuré; E: estimé ; - : non disponible.

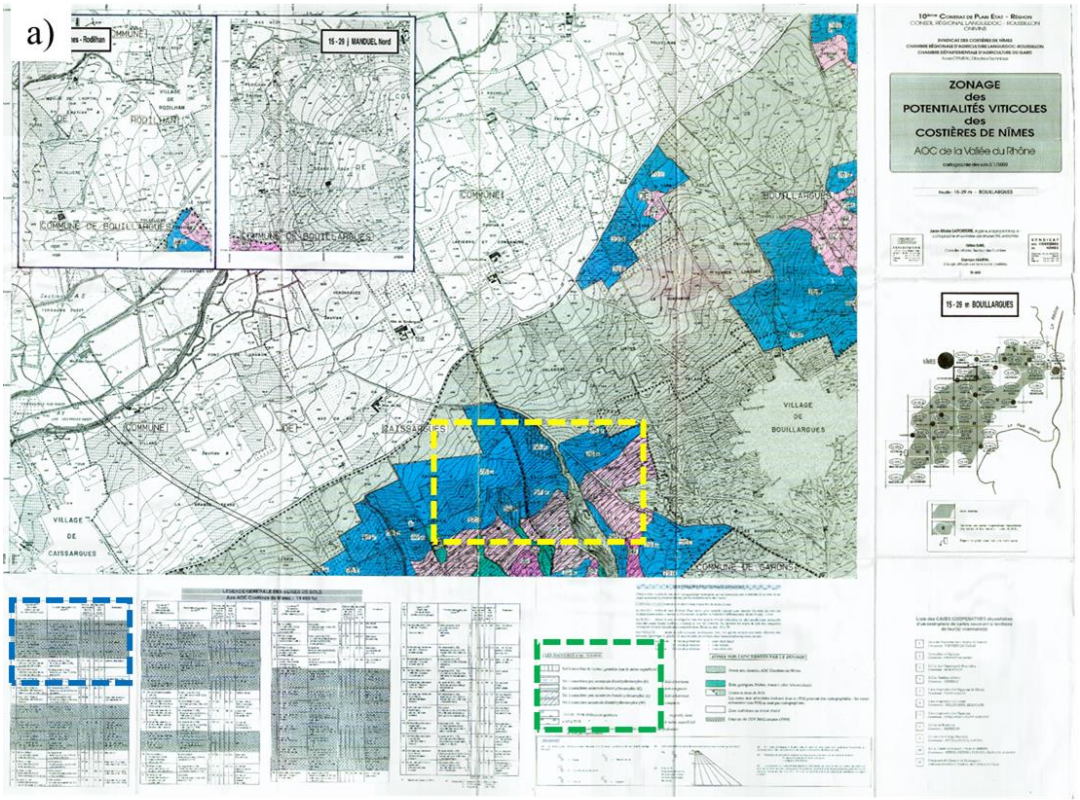
III.2.1.5. Cartes pédologiques

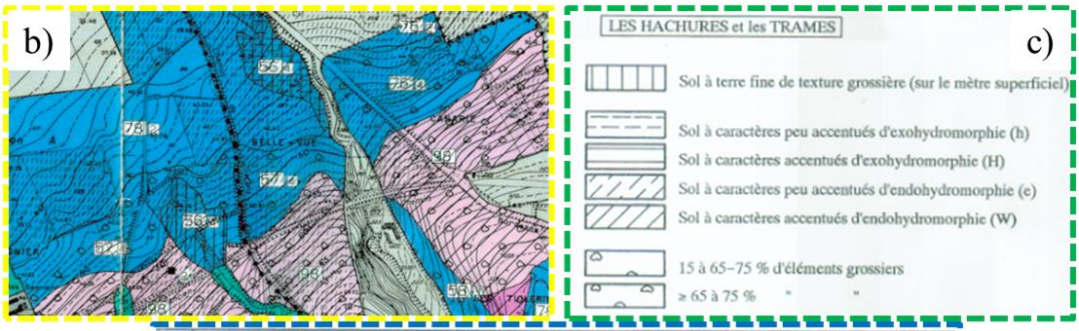
Les cartes pédologiques issues des études pédologiques sont principalement construites sur les données de profils et de sondages pédologiques, mais leur typologie varie selon la nature de leur utilisation (mise en irrigation, évaluation des potentialités, changement cultural, etc.). Le Tableau 2.4 présente la typologie des cartes en fonction de leur échelle. Les études à petites échelles sont principalement vouées à donner une description d'ensemble de caractéristiques pédologiques. Les cartes à moyennes échelles sont les plus complètes, avec des descriptions pédologiques, hydrodynamiques, pédogénétique ainsi que des cartes de mise en valeur agricole. Enfin les cartes à grandes échelles, sont utilisées pour des études de cas avec une description complète de la zone d'étude (caractéristiques pédologiques, pédogénétiques, hydrodynamiques) et des cartes de protections des sols et des cultures contre les excès d'eau basées sur une répartition de zones homogènes de dose d'irrigation pratique.

Tableau 2.4. Récapitulatif de la typologie des informations renseignées par les cartes pédologiques selon leur échelle

Echelle	Typologie des cartes et de leur légende				
	Pédologique	Hydrodynamique	Mise en valeur	Pédogénétique	Protection contre excès d'eau
Petite	☑	☒	☒	☒	☒
Moyenne	☑	☑	☑	☑	☒
Grande	☑	☑	☒	☑	☑

Bien que la quantité d'information soit importante pour les cartes à moyennes et grandes échelles, les cartes délivrées sont fréquemment poly-informatives, superposant de multiples légendes. La Figure 2.15 représente un exemple de carte pédologique évaluant les zones de potentialités viticoles. Cette carte cumule de nombreuses unités cartographiques superposées présentant (Figure 2.15.b) : les types de sol, l'hydromorphie, la charge pondérale des éléments grossiers associées à leurs légendes respectives (Figure 2.15.c). Pour l'exemple de cette carte, une estimation du RU du mètre superficiel (Figure 2.15.d) est présentée pour chaque type de sol en fonction de la charge pondérale en éléments grossiers (absence, 15-70%, > 70%) relevée sur l'unité cartographique.





d)

N° série de sol	Type de sol (1) Roche mère Géomorphologie (position dans le paysage)	Caractéristiques principales du sol (2)		Réserve utile moyen.(mm) du mètre superficiel % El. gros. RAS	15–70	> 70	Superficie aire AOC zonée Ha	%	Remarques
1	Sol peu évolué d'érosion de sable de l'Astien (versant, sommet)	Sol très profond (> 150), de texture grossière (S,SI,SS), peu calcaire (< 7 %)	P 1 C	80	40		115	0.6	Bancs de grès éventuels
4	Sol peu évolué d'érosion de sable de l'Astien (sommet, versant)	Sol profond (100 à 150), de texture grossière (SI,S,LS), peu calcaire (< 7 %)	A 1 C	80	40		20	0.1	Bancs de grès
5	Sol peu évolué d'érosion de marne du Plaisancien (sommet, versant)	Sol profond (100 à 150), de texture moyenne (LSa,LAS), calcaire (> 7 %)	A 5 T	162	80		227	1.2	Marne compacte
6	Sol peu évolué d'érosion de marne du Plaisancien (versant)	Sol profond (100 à 150), de texture fine (A,AL), calcaire	A 8 T	188			3	0.0	Marne très compacte
15	Sol peu évolué d'apport colluvial ou alluviocolluvial (versant, fond de vallon)	Sol très profond (> 150), de texture grossière (Ls,SI), peu calcaire	P 1 C	80	40		113	0.6	Sols très bien drainés
16	Sol peu évolué d'apport colluvial ou alluviocolluvial (versant, fond de vallon)	Sol très profond (> 150), de texture gros./moy. (Ls/LSa), peu calcaire	P 2 C	120	60		129	0.7	
17	Sol peu évolué d'apport colluvial ou alluviocolluvial (versant, fond de vallon)	Sol très profond (> 150), de texture moyenne (LAS,LSa), peu calcaire	P 5 C	162	80	32	347	1.8	Sols généralement bien drainés

Figure 2.15. Carte pédologique de la Commune de Bouillargues, évaluant les potentialités viticoles (a) combinant plusieurs légendes (b et c) et estimant une valeur de réservoir utile à partir de la légende des unités pédologiques (d)

III.3. Mise en forme et traitement

III.3.1. Saisie des données RRP

Les données RRP ont été extraites de la 3e version de la base de données pédologiques DONESOL. La constitution de ces données s'est faite via l'utilisation d'un script permettant l'extraction et la mise en forme des données pour un traitement statistique ou géostatistique via le logiciel de programmation R.

Le jeu de données est composé des propriétés de sols qui composent le RU (voir Equation 2.5.) : les profondeurs inférieures et supérieures de l'horizon, les teneurs en sables totaux et argile (utilisées dans les fonctions de pédotransferts, Equations 2.3 et 2.4.), ainsi que la teneur en éléments grossiers.

III.3.2. Saisie des données BRL Exploitation

La saisie des données pédologiques de BRL Exploitation a exclusivement été réalisé sur la commune de Bouillargues, en tant que « terrain pilote BRL Exploitation ». L'ensemble des types de données (profils, sondages et carte pédologique) est disponible pour cette zone.

III.3.2.1. Démarche générale

Les données pédologiques saisies sont issues des profils, sondages et d'une carte pédologique, se localisant dans les limites administratives de la commune de Bouillargues. Contrairement aux données RRP, les données pédologiques BRL Exploitation, sont sous format papier numérisé et une saisie manuelle est indispensable.

Le jeu de données est composé des propriétés de sols nécessaires pour estimer le RU (voir Equation 2.6.) : les profondeurs inférieures et supérieures d'horizon, la teneur en éléments grossiers, la densité apparente, la texture de terre fine (utilisée pour estimer le coefficient textural) et l'humidité équivalente.

III.3.2.2. Saisie des profils

Les données nécessaires pour estimer le RU sont directement accessibles sur les fiches de profils de sol comme présentées dans la section III.2.1.2. La saisie des données de localisation sous forme de coordonnées spatiales en Lambert III Zone Sud est aussi directement accessible mais nécessite une conversion pour être exprimée en Lambert 93 (projection officielle pour les cartes Françaises). L'estimation du temps de saisie d'un profil est de 0.8 min (48 s) pour la saisie des données sols et 0.2 (12 sec) pour les coordonnées spatiales.

III.3.2.3. Saisie des sondages et géoréférencement

Les sondages de sol sont disponibles sur microfiches, support de stockage qui était couramment utilisé par BRL Exploitation de 1957 à 1992. Ces microfiches contiennent des copies des descriptions de sondages (encadré bleu de la Figure 2.16) accompagnées de leur plan de localisation (encadré orange de la Figure 2.16) pour un périmètre défini. Ces informations ont été scannées au préalable par BRL Exploitation. Les sondages ne fournissent qu'une localisation relative (sans coordonnées spatiales) qui se reporte à un système de localisation basé sur le schéma de structuration du réseau hydraulique régional BRL (RHR).

Géoréférencement selon la structuration du RHR

Le plan d'échantillonnage et les descriptions des sondages contiennent un identifiant de la localisation relative de la borne de livraison en eau d'irrigation à laquelle est rattachée le périmètre défini. Cet identifiant est élaboré selon le schéma de structuration du RHR qui se scinde en trois niveaux :

- Casier : entité géographique regroupant des ensembles hydrauliques continus du RHR (Figure 2.17.a),
- Secteur : sous entité géographique situé au sein des casiers, correspondant à des unités desservies par une seule station de pompage et de mise en pression (Figure 2.17.b),
- Borne : point physique de livraison en eau d'irrigation (Figure 2.17.c).

Le format renseigné sur les microfiches est composé dans l'ordre des numéros de casier, de secteur et de la borne (ex : « 01-G1-092 » = Casier 01, Secteur G1, Borne 092) (Figure 2.16).

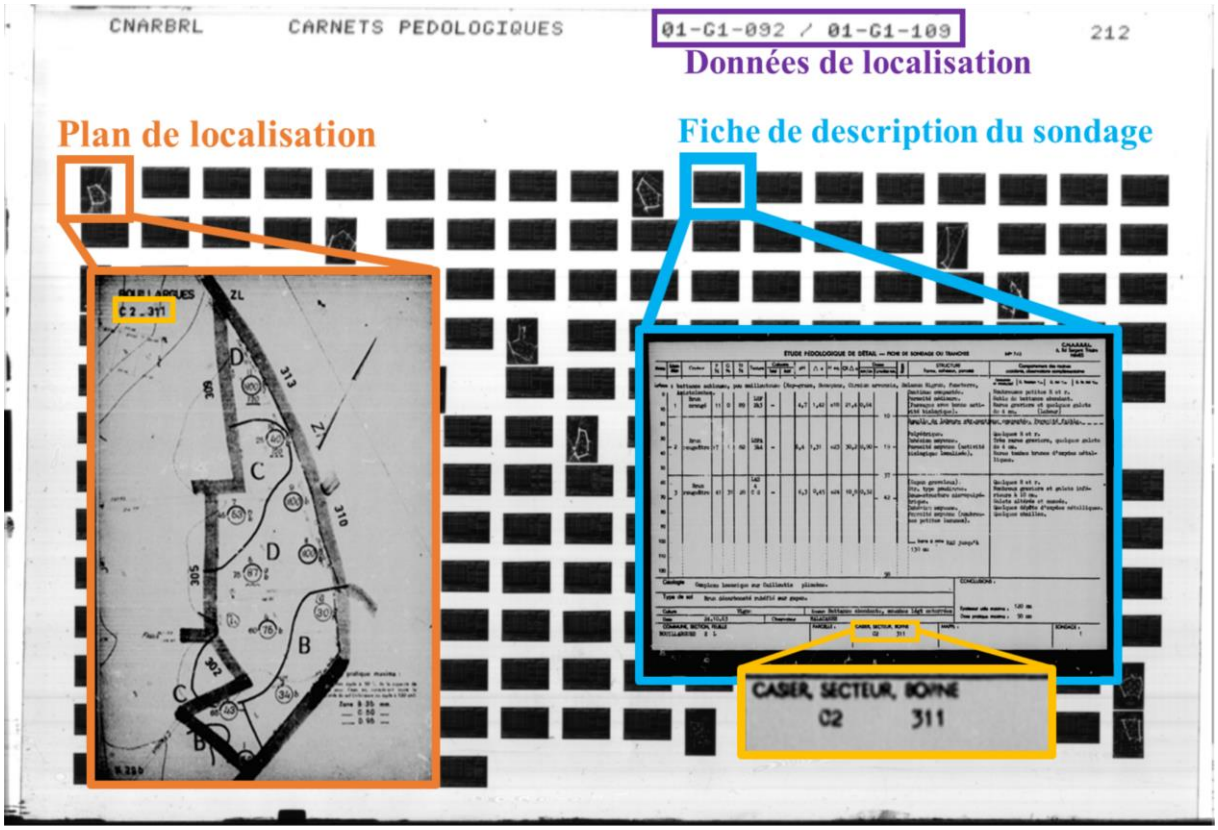


Figure 2.16. Présentation d'une microfiche répertoriant les plans de localisation (orange), les fiches de descriptions des sondages (bleu) et les données de localisation (violet)

A partir de cet identifiant, il est possible de la localiser sur l'application web-SIG de BRL Exploitation, GéEauWeb.

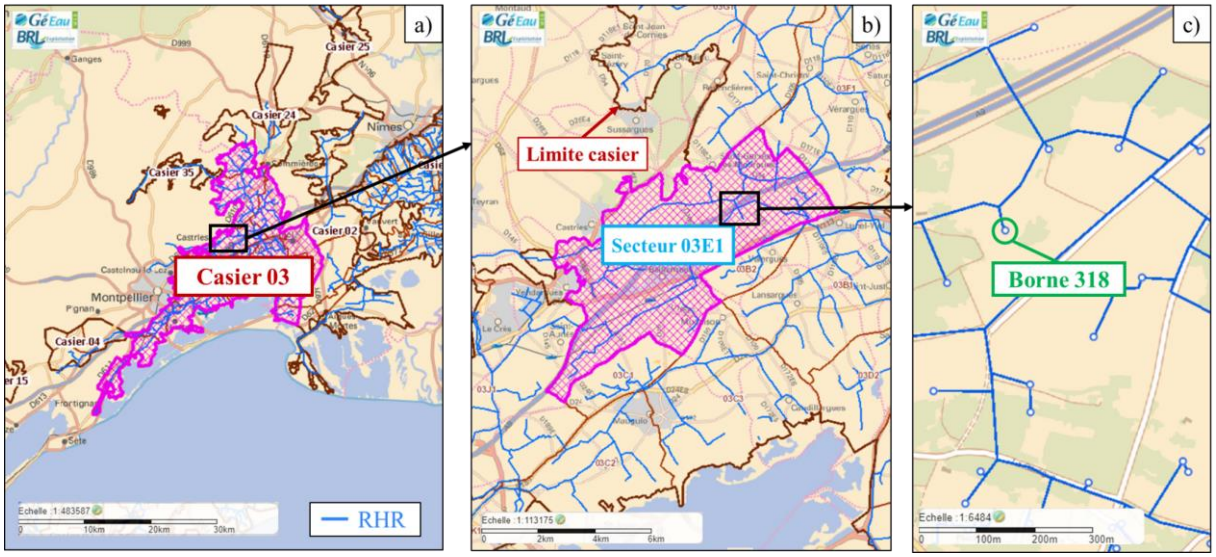


Figure 2.17. Présentation de la structuration du RHR composé de : a) les casiers, b) les secteurs et c) les bornes de livraison en eau d'irrigation

Parmi les couches géographiques disponibles sur GéEauWeb, un ancien cadastre géoréférencé (Figure 2.18.b), élaboré pendant la récolte des données pédologique BRL, permet de transposer le cadastre utilisé sur le plan de localisation des sondages sur GéEauWeb.

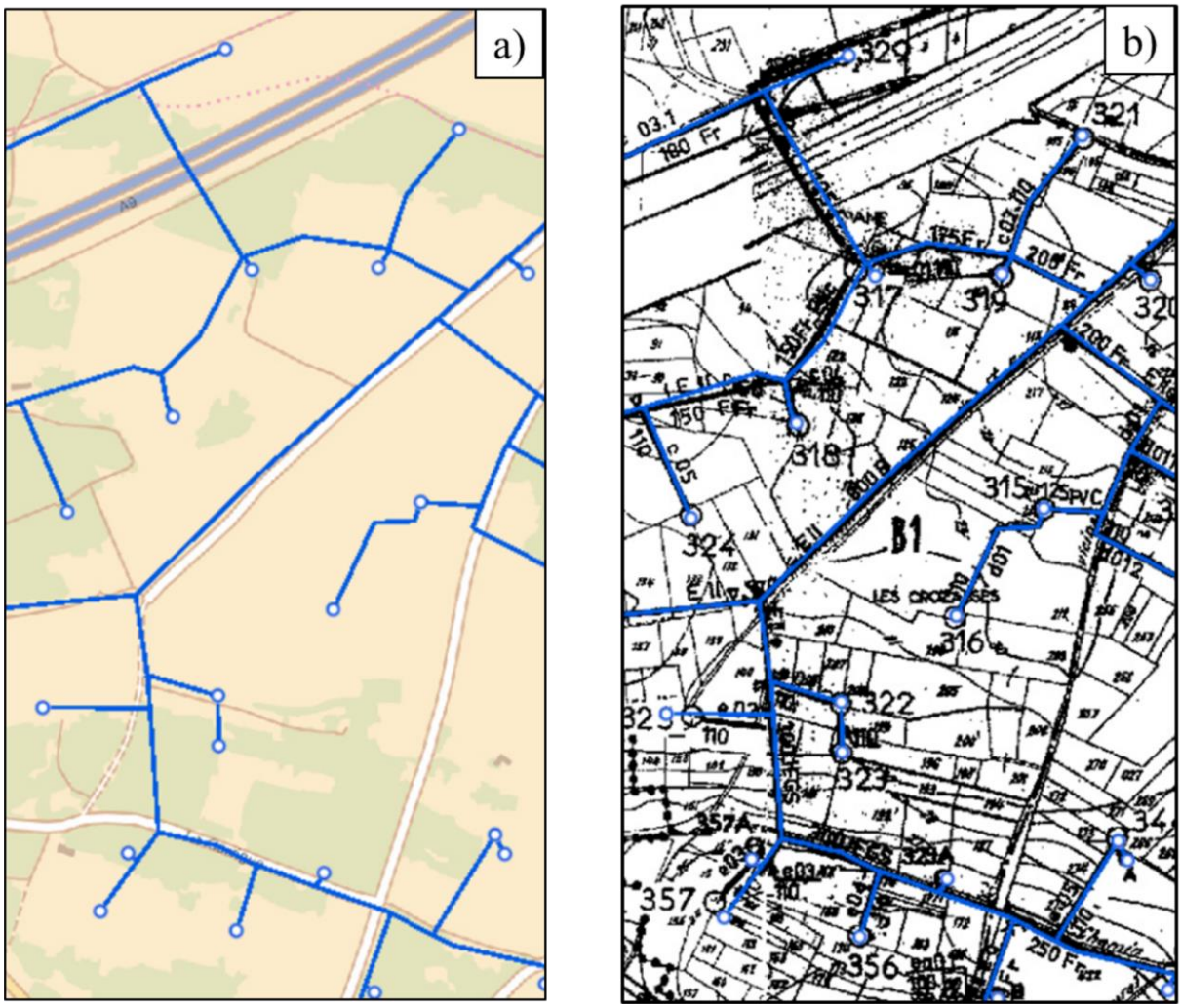


Figure 2.18. Illustration de l'ancien cadastre géoréférencé sur GéEauWeb

Ensuite, à partir d'une interprétation visuelle permettant de rattacher les limites du périmètre d'étude (Figure 2.19.a) aux voies de communications et limites de parcelles du cadastre géoréférencé (Figure 2.19.b), il est possible de mesurer les coordonnées spatiales (Lambert 93) avec l'outil « Localisation » de GéEauWeb en procédant au placement approximatif des sondages (Figure 2.19.c).

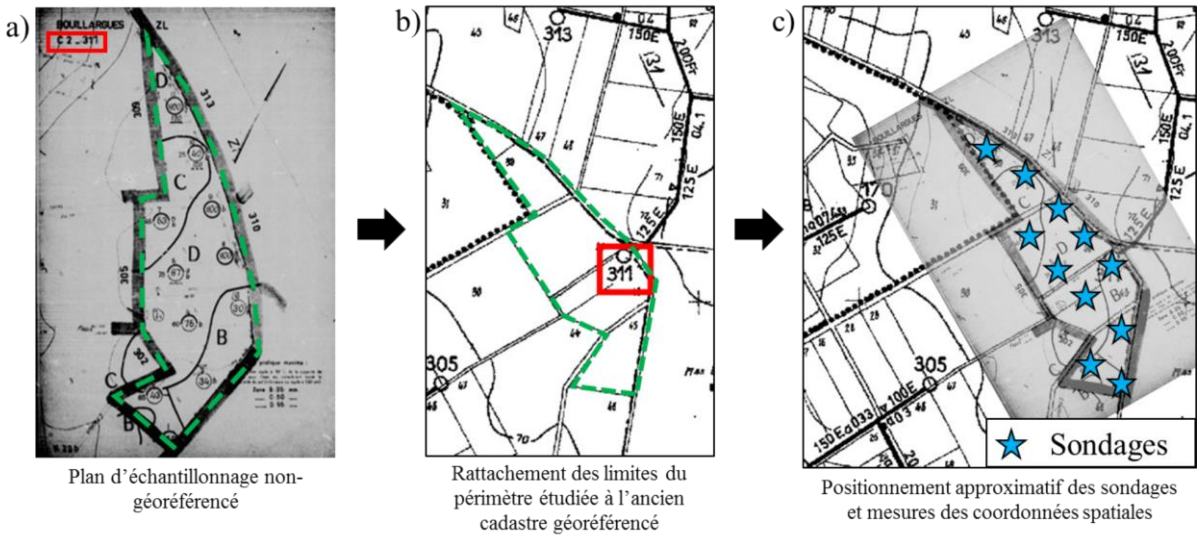


Figure 2.19. Rattachement du plan d'échantillonnage non-géoréférencé au l'ancien cadastre géoréférencé et mesure des coordonnées spatiales des sondages

Estimation du temps de saisie

La saisie des propriétés de sol intervenant dans l'estimation du RU est effectuée similairement aux profils de sol. Le temps de saisie des propriétés de sols par sondage est estimé à 0.8 min (48 s) et celui du géoréférencement à 0.5 min (30 s).

III.3.2.4. Cartes pédologiques : digitalisation des unités pédologiques

Présentées en section III.2.1.5., les cartes pédologiques par leur caractère poly-informatif, regroupant plusieurs typologies de légende (caractéristiques pédologiques, hydrodynamiques, etc.), sont des documents complexes à digitaliser. La digitalisation, réalisée sur le logiciel ArcMap de la suite SIG ArcGis, comprend : i) la digitalisation de limites des unités cartographiques définies pour chaque type de légende (teneur en éléments grossiers, hydromorphie, etc.) et ii) la codification de chaque unité cartographique selon le type de légende.

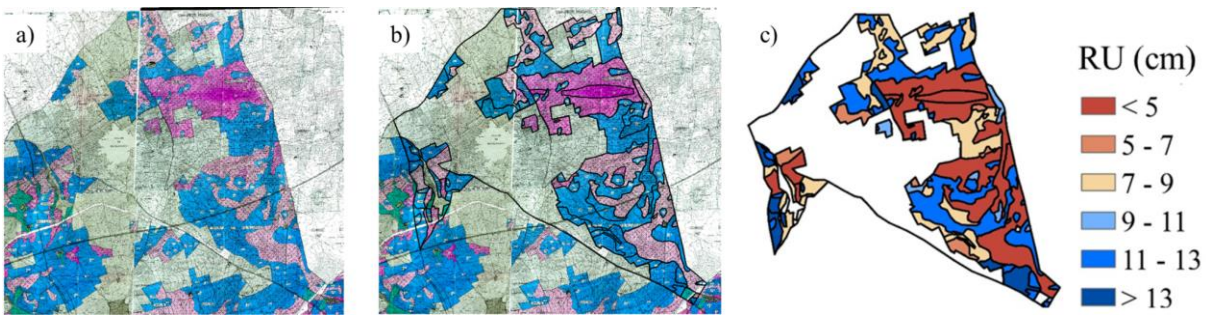


Figure 2.20. Exemple de digitalisation a) d'une carte pédologique poly-informative par b) digitalisation des unités cartographiques et c) codification des unités cartographiques

Le temps de digitalisation d'une carte pédologique est estimé à 4 heures. Ce temps estimé est à considérer avec prudence car basé sur l'unique digitalisation de carte pédologique réalisée pendant cette thèse (Figure 2.20) qui est très dépendant de la complexité de la carte pédologique.

III.3.3. Traitement des prélèvements manquants

L'estimation du RU exige un degré de complétude maximal des prélèvements des propriétés composantes du RU pour chaque horizon du profil de sol. Si un prélèvement de propriété n'a pas été relevé et renseigné pour un ou plusieurs horizons du profil, le profil n'est alors pas sélectionné pour constituer le jeu de données final. Pour cela, un script R a été développé pour appliquer ce traitement aux données pédologiques du RRP et de BRL.

III.3.4. Traitement et harmonisation de la profondeur des sols

III.3.4.1. Traitement de la profondeur des sols des données RRP

Le traitement de la profondeur des sols qui va suivre s'appuie sur les travaux de thèse de Vaysse (2015).

Selon la définition proposée par Soil Survey Division Staff (1993) et utilisée dans les spécifications GlobalSoilMap, la profondeur des sols est définie comme l'épaisseur de sol entre la surface et l'atteinte de la roche-mère. Cette limite inférieure du sol est identifiée, généralement, par l'apparition d'une roche en fond de profil (horizon noté R) soit en considérant l'apparition d'un horizon paralithique défini comme suit :

« Paralithic materials are relatively unaltered materials (do not meet the requirements for any other named diagnostic horizons or any other diagnostic soil characteristic) that have an extremely weakly cemented to moderately cemented rupture-resistance class. Cementation, bulk density, and the organization are such that roots cannot enter, except in cracks. Paralithic materials have, at their upper boundary, a paralithic contact if they have no cracks or if the spacing of cracks that roots can enter is 10 cm or more. Commonly, these materials are partially weathered bedrock or weakly consolidated bedrock, such as sandstone, siltstone, or shale » (Soil Survey Division Staff, 1993).

Cependant, le caractère ambigu de l'horizon C pouvant être un horizon paralithique ou non, nécessite une analyse plus poussée, matérialisée par un arbre de décision de type classification (Figure 2.21). « *L'arbre de décision a pour but d'interroger de la base de données DONESOL pour identifier le caractère paralithique de l'horizon. L'arbre de décision s'appuie sur quatre caractéristiques d'horizons dont l'information est largement disponible dans la base de*

données sol Languedoc-Roussillon, à savoir : le type d'horizon (classification CPCS), la structure de l'horizon, sa compacité et son niveau d'altération » (Vaysse, 2015). La structuration de l'arbre de décision est organisée selon deux niveaux d'interrogation du type du dernier horizon du profil :

- Le premier niveau interroge le degré de description du profil de sol permettant d'identifier : les profils pleinement décrits (horizon R, M, D ou horizon de transition C avec un des trois horizons précédent), les profils de type C qui n'attestent pas systématiquement d'un contact paralithique, et les descriptions de profils incomplètes par la présence de tous les autres horizons pédologiques (A, B, etc.).
- Le second niveau utilise les caractéristiques de sols supplémentaires (structure, compacité et niveau d'altération de l'horizon) uniquement lorsque le dernier horizon est de type C. Si une des trois caractéristiques indique la présence d'un contact paralithique alors le profil est considéré comme complet et sa profondeur maximale correspond à la limite inférieure de l'horizon C. Dans le cas contraire ou si la profondeur maximale n'est pas renseignée, le profil est alors considéré comme censuré à droite pour la variable profondeur, ce qui en d'autres termes consiste à dire que la profondeur maximale d'observation est inférieure à la profondeur du contact lithique ou paralithique.

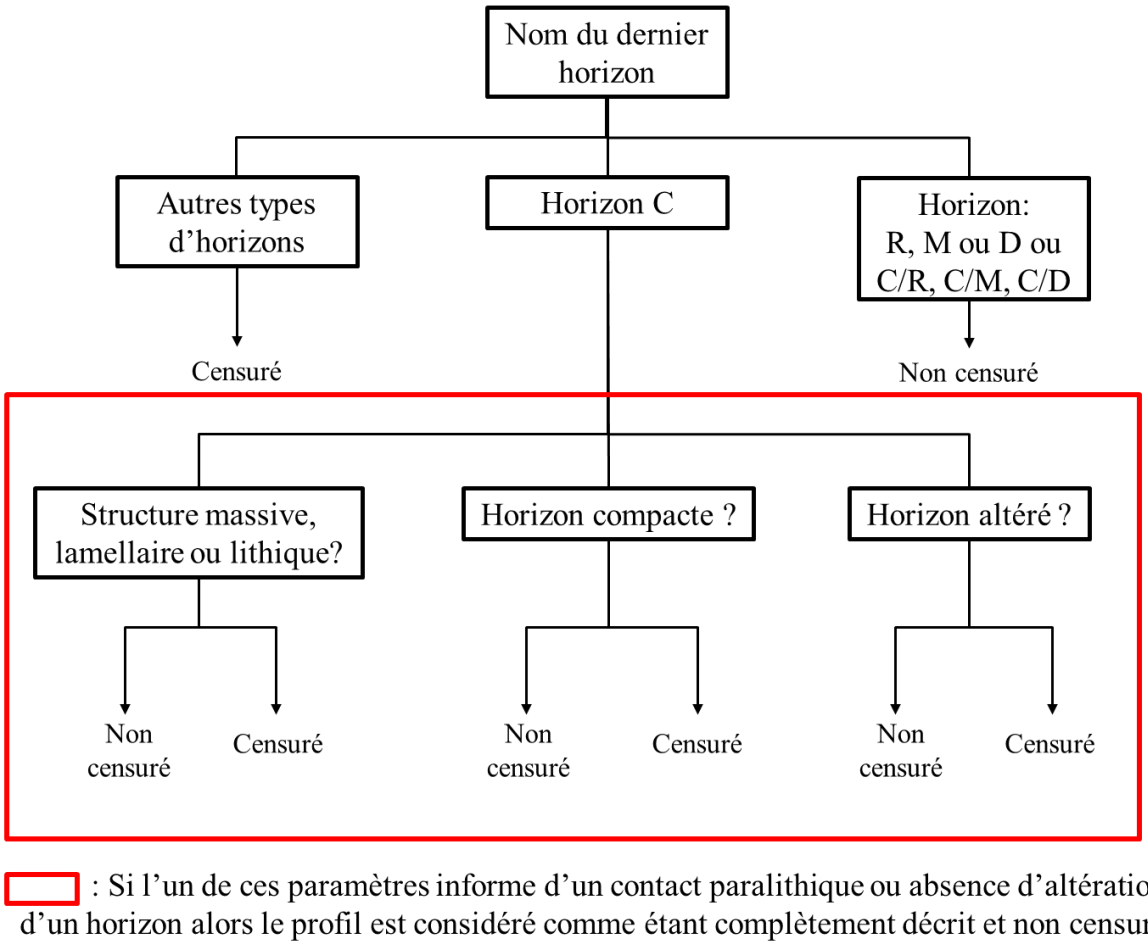


Figure 2.21. Arbre de décision de détermination du contact lithique des profils (d'après Vaysse, 2015)

Dans ces travaux de thèse, les données de sols, renseignées par horizons ont été harmonisées en six couches de sols (0-5, 5-15, 15-30, 30-60, 60-100 et 100-200 cm) selon les spécifications *GlobalSoilMap* liée aux intervalles de profondeurs (voir explication section III.3.4.2).

Par conséquent, une adaptation de la profondeur et du caractère « censuré à droite » à ses spécifications est nécessaire pour exprimer l'épaisseur effective de chaque couche.

Cette adaptation a été appliquée par un second arbre de décision tenant compte de la profondeur maximale d'observation et de la présence/absence d'un contact lithique ou paralithique.

Cet arbre de décision est également en deux niveaux :

- Le premier niveau vise à situer la profondeur maximale d'observation selon les limites de profondeurs inférieure et supérieure de la couche de sol sélectionnée : la profondeur maximale d'observation se situe soit hors des limites de profondeur de la couche de sol, soit entre ces limites de profondeur ;

- Ensuite pour chaque modalité le caractère « censuré à droite » est testé permettant de déterminer la modalité de calcul de l'épaisseur effective de la couche de sol sélectionnée.

Les profils censurés à droites sont systématiquement exclus (voir les encadrés rouge de la Figure 2.22) sauf dans le cas où la profondeur maximale d'observation est plus profonde que la couche de sol (encadré vert de la Figure 2.22). Par ailleurs, quand la profondeur du contact lithique ou paralithique est atteinte avant la couche de sol, l'épaisseur renseignée est 0 cm (encadré bleu de la Figure 2.22).

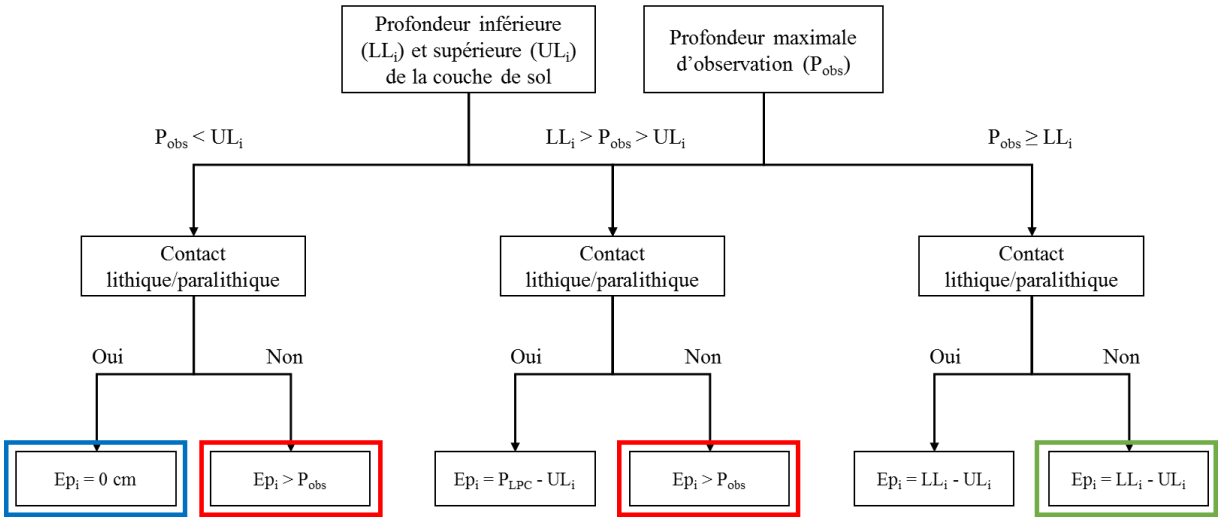


Figure 2.22. Arbre de décision pour documenter l'épaisseur de sol selon la profondeur maximale d'observation et les limites de profondeur de couche de sol (Ep_i : épaisseur effective attribuée à la couche, LL_i/UL_i : profondeur inférieure et supérieure de la couche de sol)

Ce traitement a été initié par les travaux de thèse Kévin Vaysse (2015) puis repris et adapté pour intégrer les spécificités présentées en Figure 2.22, sous forme d'une chaine de traitement contenu dans un script R.

III.3.4.2. Traitement de la profondeur des sols des données BRL

Le traitement de la profondeur des sols des données BRL s'appuie sur celui appliqué aux données pédologiques du RRP avec quelques ajustements. Les données anciennes BRL ne fournissent pas les caractéristiques de sols supplémentaires (structure, compacité et niveau d'altération du dernier horizon) ni les noms des horizons selon la classification CPCS, ce qui ne permettait pas d'estimer la présence/absence d'un contact paralithique. Dans ce cas, nous avons considéré la profondeur de la dernière description de sol la fin du sol. Cela simplifie l'arbre de décision avec uniquement la vocation à estimer l'épaisseur effective de la couche de sols à partir de ses limites de profondeurs et de la profondeur maximale d'observation (Figure 2.23).

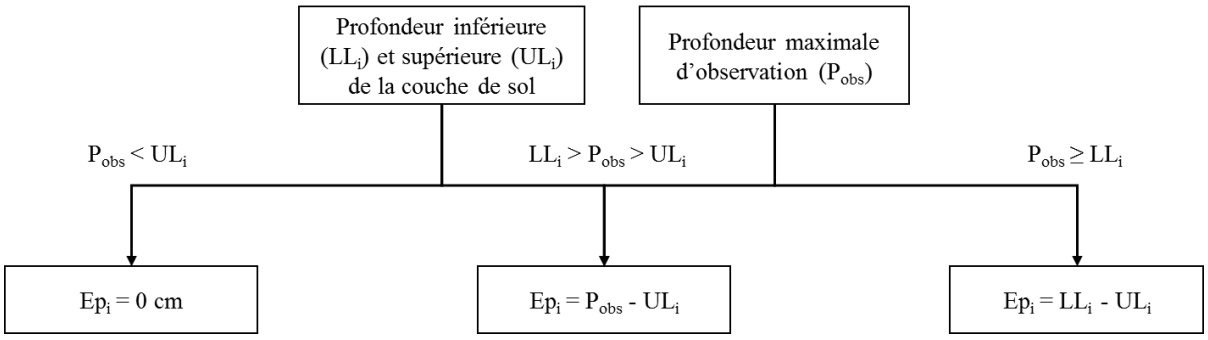


Figure 2.23. Arbre de décision de l'estimation de l'épaisseur effective des couches de sols selon la profondeur maximale observées et les limites de profondeur de la couche (Ep_i : épaisseur effective attribuée à la couche, LL_i/UL_i : profondeur inférieure et supérieure de la couche de sol)

III.3.4.3. Harmonisation des profondeurs de sols

Dans l'objectif de fournir des estimations de RU à différentes profondeurs pouvant être choisies selon les systèmes culturaux et les besoins de l'utilisateur, il est nécessaire que les données pédologiques délivrent des valeurs des propriétés de sol composantes du RU selon des intervalles de profondeurs. Pour cela, nous nous sommes référés aux intervalles de profondeur issus des spécifications *GlobalSoilMap* (0-5, 5-15, 15-30, 30-60, 60-100 et 100-200 cm). La méthodologie la plus fréquemment utilisée dans la littérature est l'utilisation de régression splines à conservation de masse (Bishop et al., 1999), caractérisé par sa capacité à limiter l'introduction d'erreur lors de l'interpolation.

D'un point de vue méthodologique, l'utilisation des splines permet de prédire sur l'ensemble d'un profil de sol, la propriété de sol pour une discrétisation de la profondeur le long du profil renseigné selon les exigences de l'utilisateur. Plus précisément, lors d'une régression spline à conservation de masse, la régression polynomiale mise en place est bornée dans les intervalles renseignées afin de faire coïncider, après régression, la valeur moyenne de l'intervalle de profondeur à la valeur introduite dans le modèle. Par exemple, si l'intervalle de profondeur de 40 à 50 cm est renseigné par une teneur en éléments grossiers de 50% en entrée du modèle, la valeur moyenne en sortie de modèle sur cet intervalle sera de 50%.

Cependant, l'erreur générée par ce type de modèle peut devenir importante dans deux cas de figures (Bishop et al., 1999) : i) la présence de très fortes discontinuités pour une propriété entre les horizons consécutifs (ex : planosol pour la teneur en argile) et ii) les effets de bords liés aux extrémités supérieures et inférieures du profil (les mesures ne sont pas systématiquement réalisées) pouvant conduire à surestimer ou sous-estimer de la propriété de sol pour l'horizon concerné. Pour atténuer au maximum ces effets, Vaysse and Lagacherie (2015) ont élaboré une démarche en deux temps :

- Si l'horizon de sub-surface présente une épaisseur supérieure à 3 cm, celui-ci est scindé en deux horizons composé d'un horizon de 0 à 3 cm et de 3 cm à la profondeur maximale de l'horizon ;
- Pour l'horizon final du profil, celui-ci est dupliqué jusqu'à 200 cm, lorsque la profondeur du contact lithique ou paralithique n'est pas renseignée.

L'application de ce modèle permet la prédiction d'une propriété de sol tout au long du profil selon une résolution centimétrique permettant d'appliquer les intervalles de profondeurs *GlobalSoilMap*. La Figure 2.24 illustre les grandes lignes de la méthodologie. L'harmonisation des profondeurs de sols a été réalisée via l'utilisation du package R *GSIF* (Hengl et al., 2013).

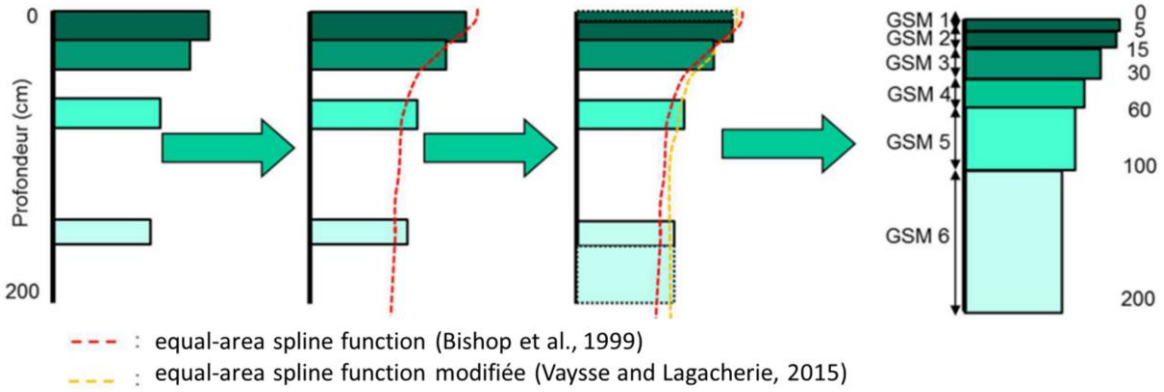


Figure 2.24. Représentation synthétique de l'harmonisation des profondeurs de sol par l'application de régression spline à conservation de masse (d'après Vaysse, 2015)

III.4.Les jeux de données constitués

III.4.1.Jeu de données RRP – Languedoc-Roussillon

Le jeu de données issus du RRP Languedoc-Roussillon est initialement composé de 2691 profils de sol réduit à 2024 profils de sol par une analyse de complétude des données pédologiques issus des travaux de thèse de Kévin Vaysse (2015) afin que chaque horizon d'un profil soit renseigné par une valeur de propriété de sol. Cependant, le jeu de données utilisé dans ces travaux de thèse est plus restreint pour deux raisons : i) chaque composant du RU doit impérativement être renseigné par horizon pédologique et ii) seuls les profils de sol présentant un contact lithique ou paralithique sont sélectionnés. 1464 profils de sol constituent ce jeu de données, ce qui correspond à une densité d'un profil tous les 19 km² (Figure 2.25). La répartition spatiale des profils est éparse avec des zones très denses (« clustérisées ») et, à l'inverse, des zones non couvertes en profils de sol.

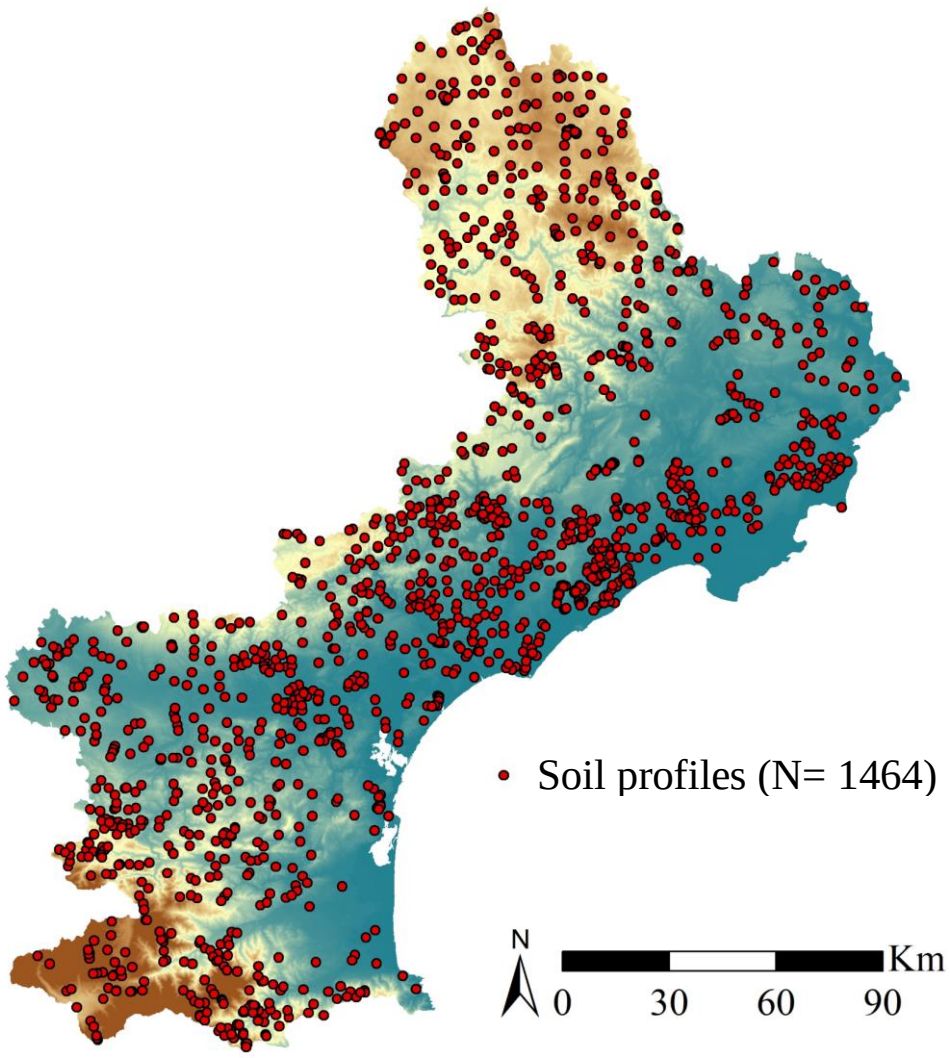


Figure 2.25. Cartographie de la répartition spatiale des profils de sol issus du RRP LR

La Figure 2.26 présente les distributions du RU et de ses composants en fonction des couches de sol définies par les intervalles de profondeur. L'argile et les sables varient peu avec la profondeur avec de légères augmentations des valeurs pour les couches plus profondes (Figure 2.26.a et b). Il en est de même de la distribution des éléments grossiers qui présente des valeurs relativement faibles pour les couches superficielles et qui augmentent pour les couches profondes (Figure 2.26.c) avec un accroissement des teneurs élevées en éléments grossiers mis en évidence par la bimodalité. Les humidités volumiques à la capacité au champ et au point de flétrissement permanent suivent la logique des distributions de leur variable prédictive (argile et sables) avec une très légère augmentation des valeurs en profondeur (Figure 2.26.d et e). Enfin le RU tend à augmenter pour les couches profondes (Figure 2.26.f), ce qui est en accord avec l'évolution de la majorité de ses composants (argile, sables, éléments, humidités spécifiques volumique et l'épaisseur ; les éléments grossiers ayant un rôle de pondérateur).

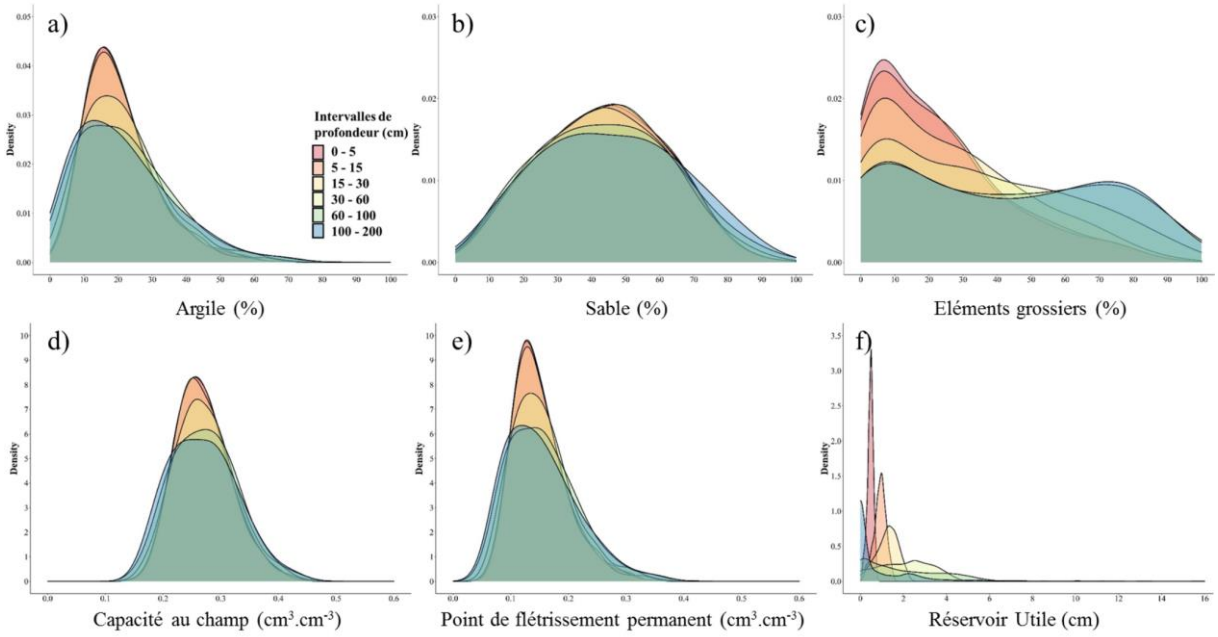


Figure 2.26. Distribution des teneurs en a) argile, b) sable, c) éléments grossiers, d) humidité à la capacité au champ, e) humidité au point de flétrissement permanent et f) de réservoir utile selon les couches de sol définies par les intervalles de profondeur des spécifications GlobalSoilMap

L'épaisseur des couches de sol est représentée en Figure 2.27 par la distribution de la profondeur de sol limitée à 200 cm selon les spécifications *GlobalSoilMap*. La majeure partie des profils de sol présente une profondeur comprise entre 5 et 120 cm. Les sols profonds sont, quant à eux, extrêmement sous-représentés.

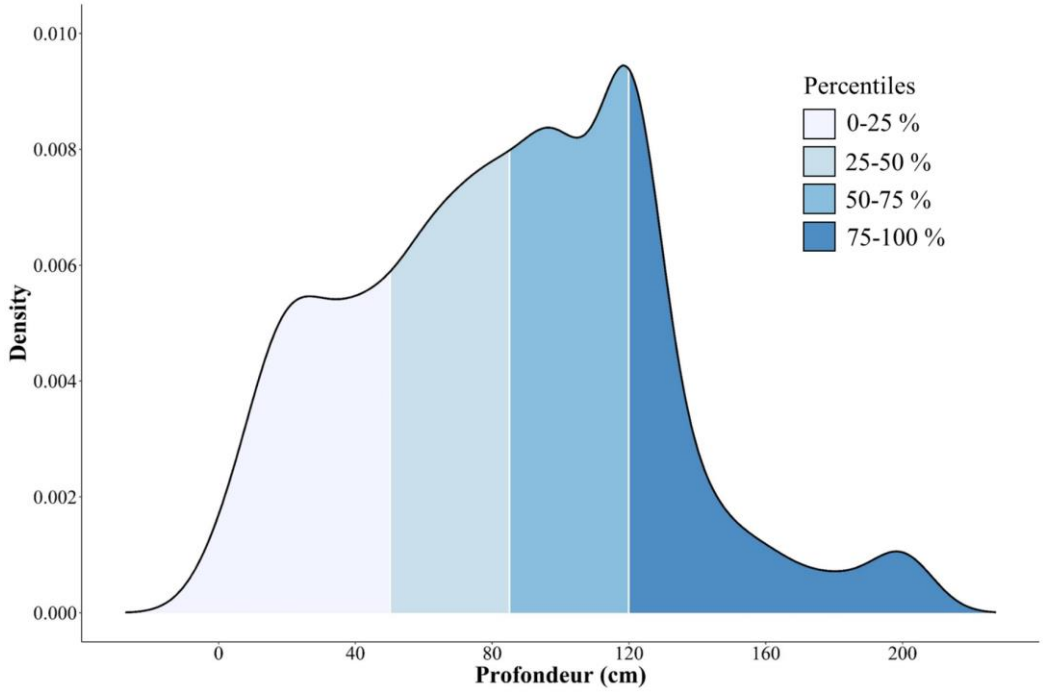


Figure 2.27. Distribution de la profondeur des sols avec une profondeur maximale bornée à 200 cm selon les spécifications GlobalSoilMap

III.4.2.Jeu de données BRL Exploitation – Bouillargues

III.4.2.1.Profils de sol et sondages

La commune de Bouillargues, localisée à l'extrémité Est de la région Occitanie, couvre une superficie de 16 km². Le jeu de données saisie est composé de 69 profils avec analyses de sol (Figure 2.28.a) et 2781 descriptions de sol sans analyses issues de sondages à la tarière (Figure 2.28.b), représentant respectivement un espacement moyen entre profils de 500 m et de 76 m entre sondages. La répartition spatiale des données profils et sondages est plutôt homogène sur le territoire avec quelques zones non couvertes en données pédologiques étant soit des zones urbanisées ou à faible potentialité agricole.

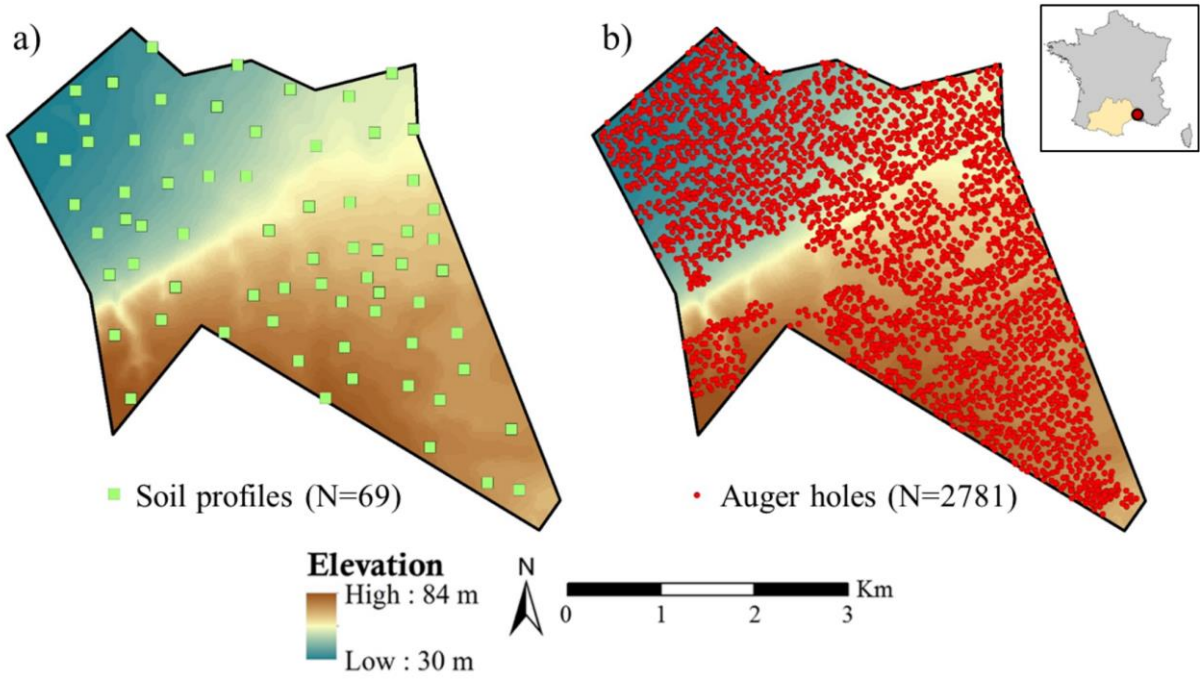


Figure 2.28. Répartition spatiale a) des profils de sol et b) des sondages à la tarière sur la commune de Bouillargues

La Figure 2.29 présente les distributions du RU et de ses composants pour chaque couche de sol. La distribution de la densité apparente (Figure 2.29.a) présente une forte représentation des valeurs élevées de densité apparente en surface qui tendent à se diversifier en profondeur. Pour les éléments grossiers (Figure 2.29.b), la surface présente une faible pierrosité et une augmentation de la teneur moyenne des éléments grossiers en profondeur est observée. Les distributions de l'humidité équivalente et le coefficient textural (Figure 2.29.c et d) suivent un comportement logique étant donné leur sens de variation antagoniste (Tableau 2.2), nous avons alors une augmentation de l'humidité équivalente moyenne et une diminution du coefficient textural moyen lorsque l'on évolue vers des couches plus profondes. Enfin, la distribution du RU (Figure 2.29.e) montre une augmentation de la valeur de RU moyenne en profondeur, avec la présence d'un pic pour des faibles valeurs de RU lié à la faible épaisseur voire à l'absence de sol sur la couche de sol 100-200 cm (Figure 2.30).

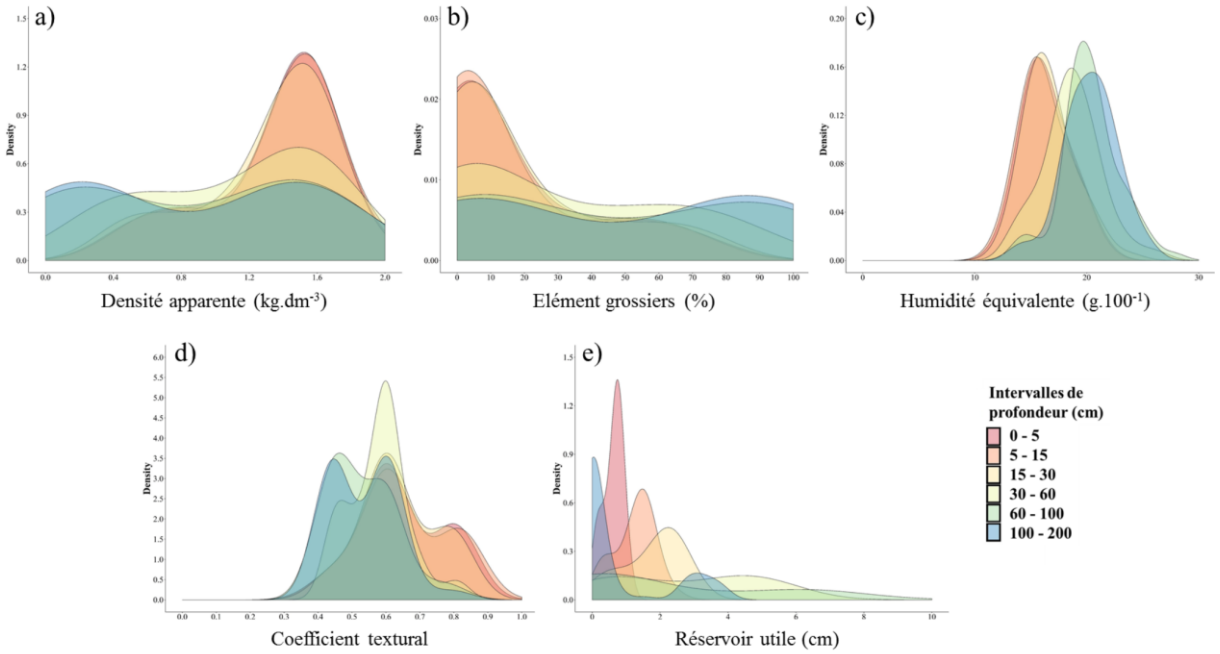


Figure 2.29. Distribution de : a) la densité apparente, b) la teneur en éléments grossiers, c) l'humidité équivalente, d) du coefficient texturale, e) du réservoir utile et f) de réservoir utile, pour les profils de sol, selon les couches de sol définies par les intervalles de profondeur des spécifications GlobalSoilMap

La Figure 2.30 présente la distribution de la profondeur de sol issu du jeu de profils. Cette distribution de la profondeur montre principalement une surreprésentation des profondeurs à 120 cm liée à la profondeur limite de prospection de la tarière à main. Des sols plus superficiels compris entre 30 cm et 90 cm sont également présents.

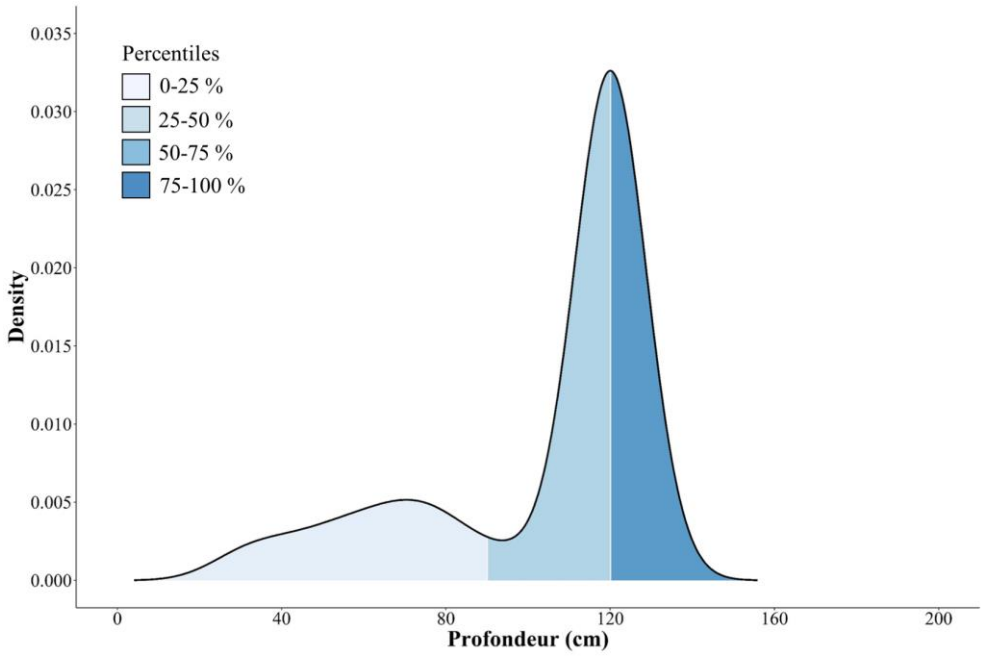


Figure 2.30. Distribution de la profondeur des sols issue des profils de sol

Les distributions du RU et de ses composantes issues des sondages (Figure 2.31) sont assez similaires à celles présentées pour les profils de sol ci-dessus, avec néanmoins des distributions bimodales plus marquées, ainsi que les distributions du coefficient textural beaucoup plus pointues. En effet, les pics de distribution pour les couches superficielles sont très marqués avec un coefficient textural avoisinant les 0.6.

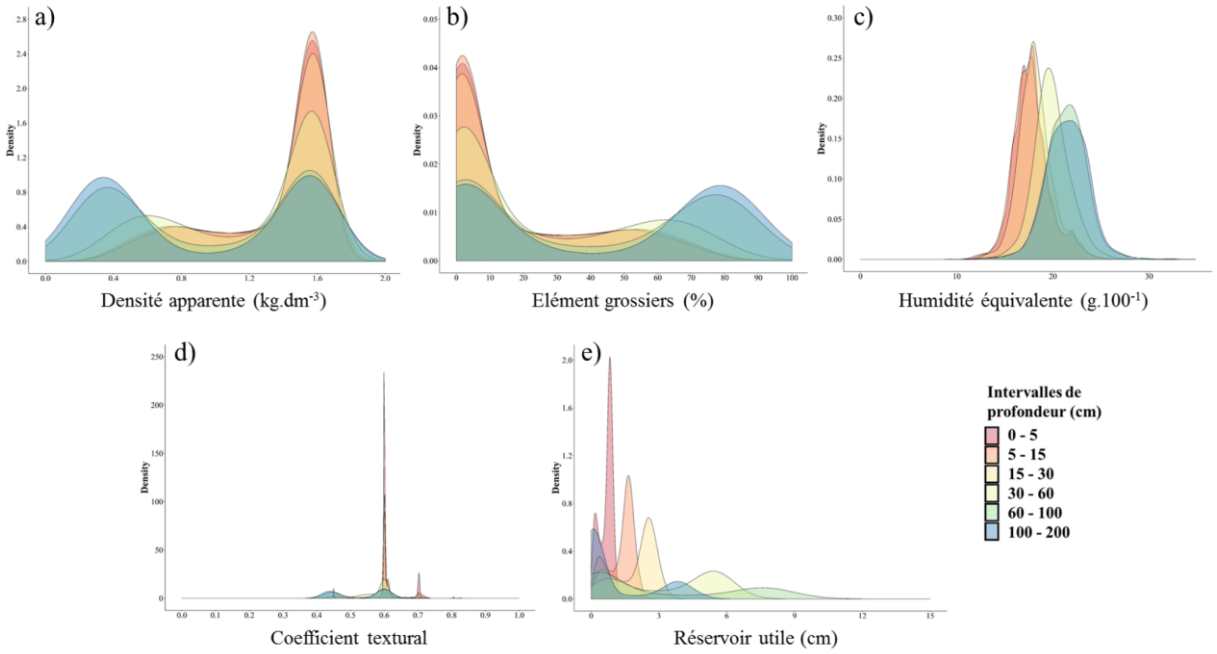


Figure 2.31. Distribution de : a) la densité apparente, b) la teneur en éléments grossiers, c) l'humidité équivalente, d) du coefficient texturale, e) du réservoir utile et f) de réservoir utile, pour les sondages de sols, selon les couches de sol définies par les intervalles de profondeur des spécifications *GlobalSoilMap*

La distribution de la profondeur de sol des sondages (Figure 2.32) est similaire à celle des profils (Figure 2.30), sans doute liée à la profondeur d'observation s'arrêtant à 120 cm également. Cependant, les valeurs de profondeurs plus faibles sont plus étalées qu'avec les profils, et présente notamment un léger pic à 100 cm.

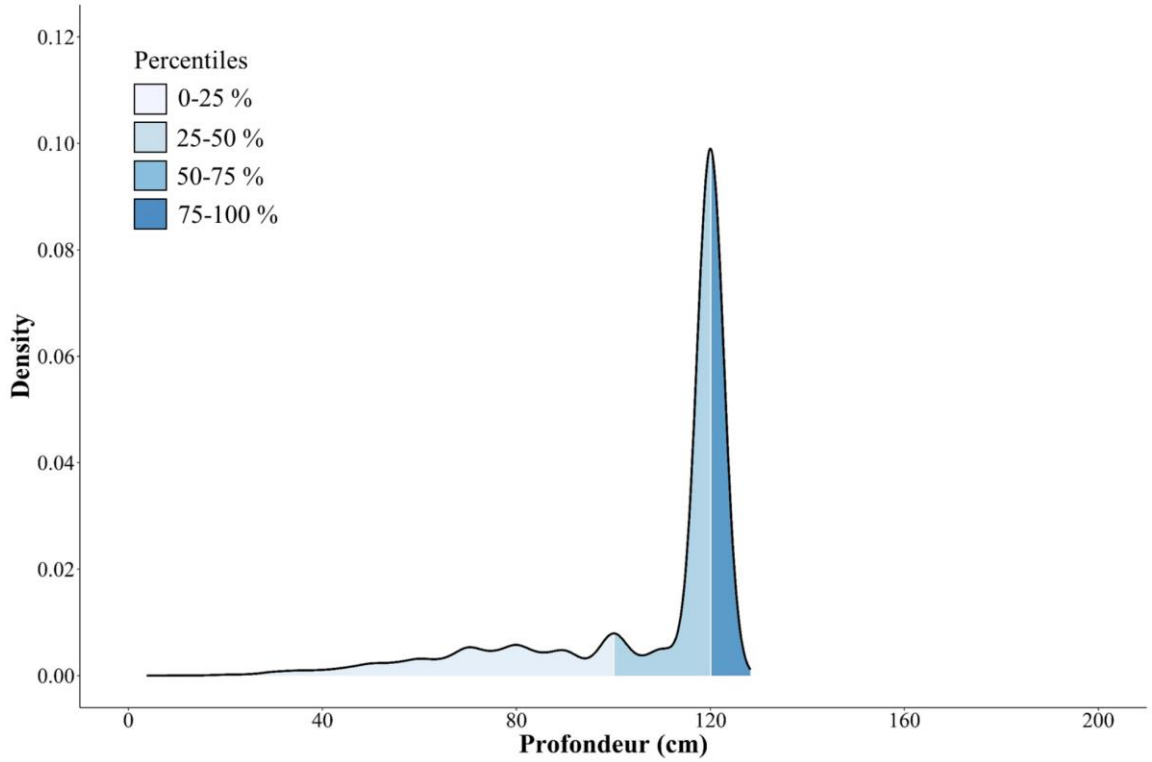


Figure 2.32. Distribution de la profondeur des sols issue des profils de sol

III.4.2.2.Carte pédologique

Le jeu de données est également constitué d'une carte de RU réalisée par digitalisation de la carte au 1/5 000^{e} des zonages de potentialités viticole de la commune de Bouillargues. Cette carte du RU (Figure 2.33) contient quelques zones non codifiées principalement dues à une non prise en compte des zones urbanisées et à faible potentialité viticole. De plus, cette carte ne couvre pas l'ensemble des limites administratives de la commune de Bouillargues.

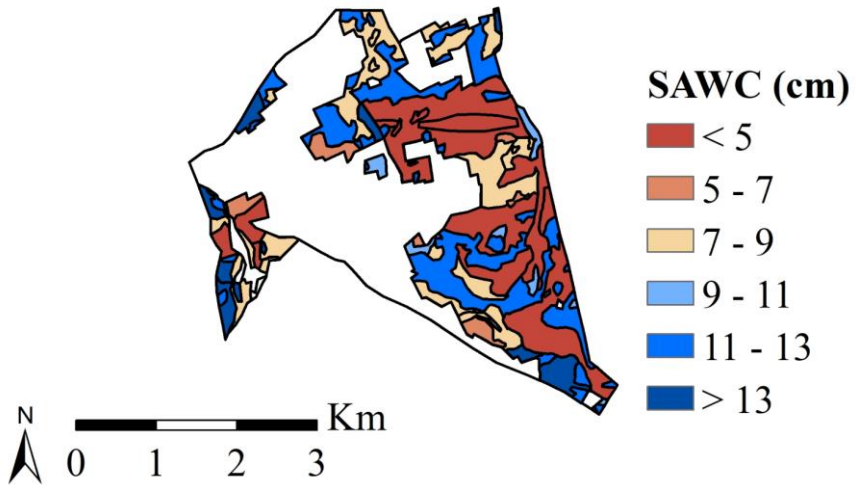


Figure 2.33. Carte du RU élaborée par digitalisation et interprétation de la carte pédologique (1/5 000^{e}) de Bouillargues

III.4.2.3. Biais d'estimation entre les données issues de profils et de sondages

Présentées en sections III.2.3.1 et III.2.3.2, les méthodes de mesures entre profils et sondages diffèrent avec la possibilité d'effectuer des analyses en laboratoire pour les profils quand les sondages se restreignent à des observations *in situ*. Cette différence peut être la source d'un biais d'estimation des propriétés composantes du RU. Les propriétés de sol susceptibles de présenter un biais d'estimation sont la texture en terre fine ainsi que la teneur en éléments grossiers.

Pour la texture en terre fine, nous présentons en Figure 2.34, la comparaison entre les valeurs du coefficient textural issues de l'analyse granulométrique des échantillons de sol et celles estimées à partir de la texture appréciée par le pédologue sur le terrain. Environ 56 % des estimations du coefficient textural par le pédologue présentent un biais par rapport aux valeurs estimées en laboratoire. Sur ces 56 %, 27% des valeurs sont surestimées, ce qui se traduit par une estimation d'une texture plus grossière que le résultat de l'analyse granulométrique, et 29 % des valeurs sont sous-estimées, ce qui est lié à la détermination de texture plus fines que celles issues de l'analyse granulométrique.

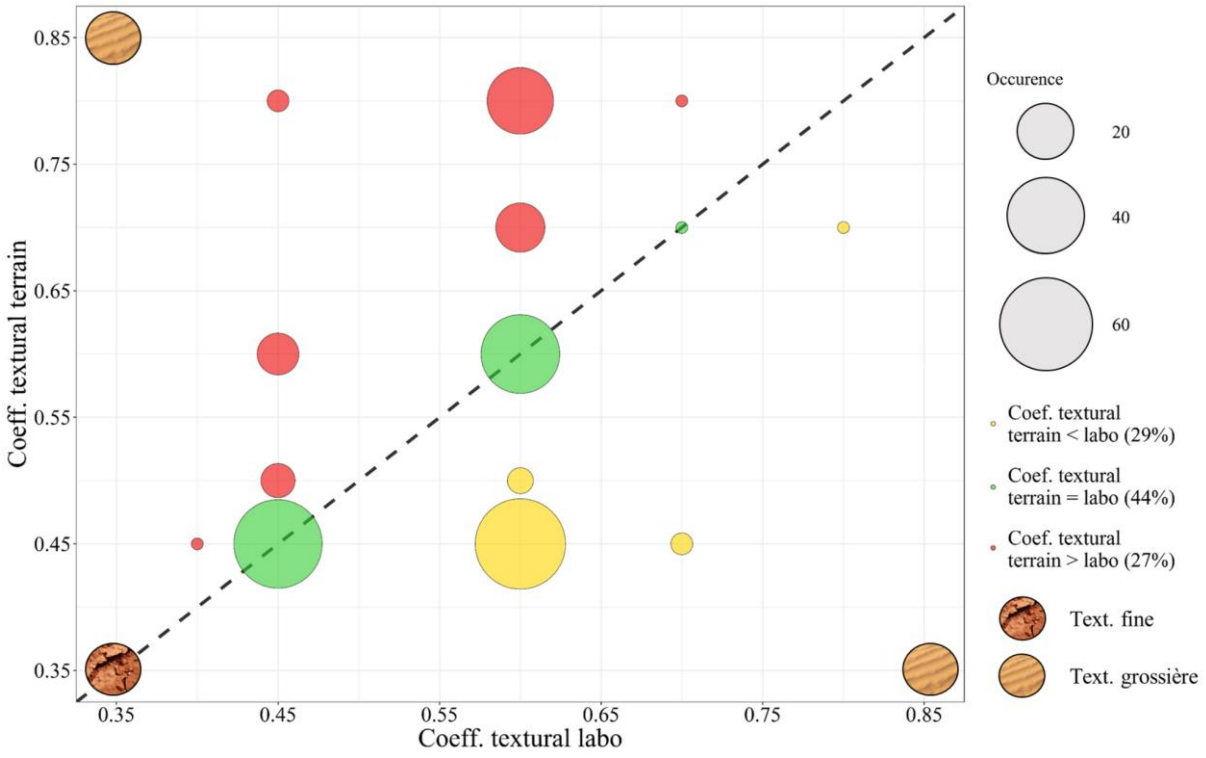


Figure 2.34. Comparatif entre les valeurs du coefficient textural déterminée par analyse granulométrique en laboratoire et par estimation de la texture sur le terrain

La Figure 2.35 représente la comparaison entre les valeurs d'humidité équivalente mesurée en laboratoire et estimée à partir les fonctions de pédotransfert locales (alimentées par la texture). Seulement 11% des valeurs de l'humidité équivalente sont correctement estimées (à +/- 1% de la différence entre l'humidité mesurée et estimée). Les valeurs des humidités estimées ont tendance à être sous-estimées par rapport aux valeurs mesurées en laboratoire (74%) ce qui se traduit vers une estimation de classes de textures plus grossières que celles issues de l'analyse granulométrique. 11% des humidités estimées sont en accord avec les résultats d'analyses et 15% des humidités estimées sont surestimées (textures plus fines relevées par le pédologue que les résultats d'analyse).

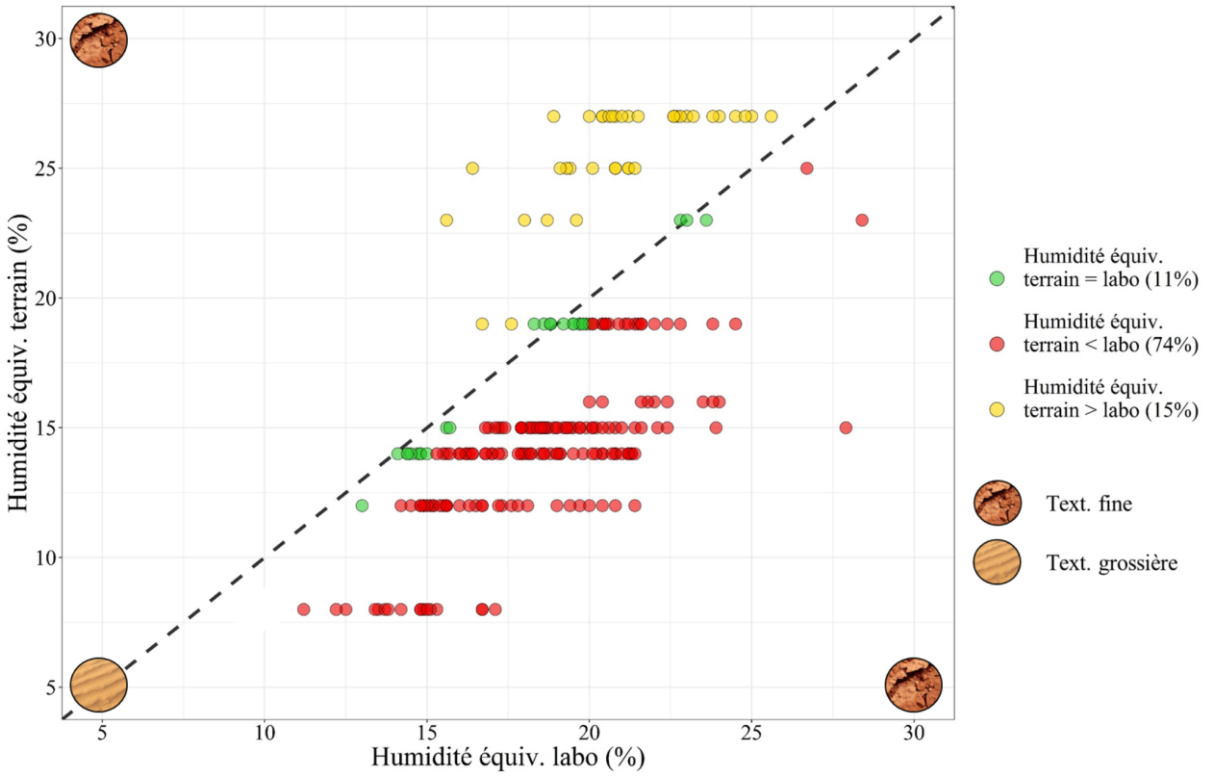


Figure 2.35. Comparaison entre les valeurs d'humidité équivalent déterminée par analyse laboratoire et par utilisation de fonction de pédotransfert via l'estimation de la texture

L'estimation de la teneur en éléments grossiers peut également être une source de biais, principalement lié à l'outil de prospection utilisé. En effet, l'estimation des éléments grossiers à la tarière à main est réalisée sur un échantillon issu d'un très petit volume comparé à la mesure faite sur le profil de sol où il est possible de mobiliser un volume beaucoup plus important. De ce fait, les mesures d'éléments grossiers effectuées à la tarière à main sont susceptibles de sous-estimer la véritable valeur et par conséquent de surestimer le RU.

Les éléments importants de ce chapitre à retenir sont :

- L'estimation de valeur du réservoir utile nécessite une complétude maximale des propriétés primaires de sols, ce qui a l'inconvénient de réduire fortement le jeu de données pédologiques issu du RRP LR à l'échelle régionale.
- L'estimation du réservoir utile selon le protocole BRL Exploitation traite la capacité de rétention de la terre fine différemment des standards actuels (Cousin et al., 2003), notamment avec l'utilisation de l'humidité spécifique et du coefficient textural pour exprimer la différence entre les humidités à la capacité au champ et au point de flétrissement permanent.
- Les données pédologiques sont denses avec **228 000 observations de sols** réparties sur 6 636 km² comprenant **25 000 profils de sol avec analyses** (espacement moyen = 515 m) et **203 000 descriptions de sol par sondages à la tarière à main** (espacement moyen = 181 m).
- Les données pédologiques anciennes BRL Exploitation sont hétérogènes au niveau de la détermination des propriétés primaires de sol (mesurées vs estimées) selon le type d'observations de sol (profils vs sondages).
- La saisie de données pédologiques anciennes est relativement simple pour les profils de sols (données directement disponibles, temps de saisie = 1 min / profil) mais est plus fastidieuse pour les sondages pédologiques (géoréférencement par l'utilisateur nécessaire, temps de saisie = 1 min 18 s / sondage) et pour les cartes pédologiques (souvent plusieurs topologies à digitaliser, temps de saisie ~ 4h)
- Selon l'étude exploratoire réalisée, des biais d'observation ont été identifiés pour la détermination de la texture en terre fine (appréciée à la main par le pédologue vs analyse granulométrique en laboratoire puis utilisation du triangle de texture) et des éléments grossiers (la mesure sur le terrain avec la tarière à main pour les sondages est beaucoup moins précise que sur le profil de sol).

CHAPITRE 3 Sensibilité des estimations du réservoir utile aux trajectoires de calcul

Les méthodes de spatialisation du réservoir utile répertoriées dans la littérature sont très hétérogènes. Le caractère multivarié du réservoir utile implique différentes opérations : « agrégation de couches de sol », « combinaison de propriétés primaires de sol » et « spatialisation » permettant de conceptualiser la cartographie numérique du réservoir utile. L'ordre d'enchaînement de ces opérations permet d'élaborer des trajectoires de calcul avec un point de départ et un point d'arrivé commun à chaque trajectoire qui sont respectivement les propriétés primaires mesurées pour des intervalles de profondeur et une cartographie du réservoir utile. Pour ce faire, ce chapitre est dédié à une étude comparative des trajectoires de calcul évaluées par des indicateurs statistiques visant à estimer la qualité des prédictions.

Ces travaux seront également accompagnés d'une étude approfondie des corrélations entre propriétés et entre couches de profondeur pour mieux comprendre les mécanismes de la spatialisation du réservoir utile et les différences de performances entre les trajectoires de calcul.

Ce chapitre se présente sous la forme d'un article publié dans la revue Soil Systems : Styc Q, Lagacherie P. (2019). What is the Best Inference Trajectory for Mapping Soil Functions: An Example of Mapping Soil Available Water Capacity over Languedoc Roussillon (France). Soil Syst. 3(34).

What is the best inference trajectory for mapping soil functions: an example of mapping soil available water capacity over Languedoc Roussillon (France)

Styc Quentin and Lagacherie Philippe

Published article in Soil Systems

Abstract

Extending digital soil mapping to the mapping of soil functions that can support end-user decisions comes to coupling a digital soil mapping procedure and a soil function assessment method. This can be done following various possible inference trajectories following the order with which "combining primary soil properties", "aggregating soil layers across depths" and "mapping" are executed to provide the targeted output.

18 inference trajectories designed for computing soil available water capacity maps in the Languedoc-Roussillon region (France) were compared with regard to their mapping performances. The best performance (SS_{MSE} = 0.37) was obtained by a trajectory that, before mapping, combined the three first *GlobalSoilMap* soil layers and computed the available water capacity of each layer. The worst (SS_{MSE} = -0.01) was observed when all the soil layers and soil properties were combined prior to mapping.

We explain the observed differences between trajectories by examining the differences in mapping errors and in error propagation between the compared trajectories, which involve both the correlations between the soil properties and between their mapping errors. This paves the way to spatial soil inference systems that could perform an ex ante selection of the best possible inference trajectory for mapping a soil function.

I. Introduction

It is increasingly recognized that soils and their functions have a part to play in the large existential challenges that have been recognized for the sustainable development of humanity and planet Earth (McBratney et al., 2014). Addressing such challenges needs to appropriately informing local and global decision making, which requires a knowledge of soils at fine resolution and global extent (Sanchez et al., 2009). Digital Soil Mapping (DSM) (McBratney et al., 2003; Lagacherie et al., 2007) has been proposed as a methodology for reaching this requirement. Various applications of DSM across the globe (Arrouays et al., 2017) demonstrated that DSM can now operationally produce sets of high resolution images representing the spatial variations of the most currently required soil properties or "primary soil properties" (e.g., soil textural fractions, soil carbon content, available water capacity, etc.).

In spite of these substantial efforts for moving to operationality, DSM has still not fully matched the initial objective. A new step is to shift from mapping primary soil properties to mapping soil functions. After the pioneering paper of Carré et al. (2007) that advocated for digital soil assessment approaches, there has been abundant literature providing conceptual advances for the description of soil functions and the related ecosystem services (Adhikari and Hartemink, 2016) and also on the valuations of soil services (Dominati et al., 2014). Greiner et al. (2017) recently proposed a set of soil function assessment (SFA) methods that cover the multiple functionalities of soils and are applicable to ecosystem service supply assessments. These SFA methods use as inputs a minimum set of primary soil properties and pedotransfer functions which makes them largely applicable provided that the spatial data on primary soil properties are made available. This carries implicitly the idea that (digital) soil mapping and soil function assessment methods are two independent processes that can be only loosely coupled through a straightforward transfer of data. This is this idea that we want to question in this paper.

Beyond their differences, all of the SFA methods share a similar data flow. They provide a single output, the valuation of the soil function, from a set of soil properties that characterize different soil layers. Figure 3.1 represents in a three-dimensional space the different possible inference trajectories that can be envisaged for producing an SFA output, with the inclusion of the mapping process in these trajectories. The inference trajectories may differ in the order with which "combining primary soil properties", "aggregating soil layers across depths" and "mapping" are executed to provide the common targeted SFA output. The commonly followed approach, i.e. mapping first then combining soil property and aggregating soil layers, is one of the possible trajectories (shown in green on Figure 3.1) but many others exist. Mapping could

be the last executed process after having combined the soil properties and the soil layers over the soil input dataset ("mapping last" in blue on Figure 3.1). Combining first the soil properties at each single soil layer, then mapping and lastly aggregating SFA outputs is an intermediate inference trajectory (shown in red on Figure 3.1) that can be envisaged too.

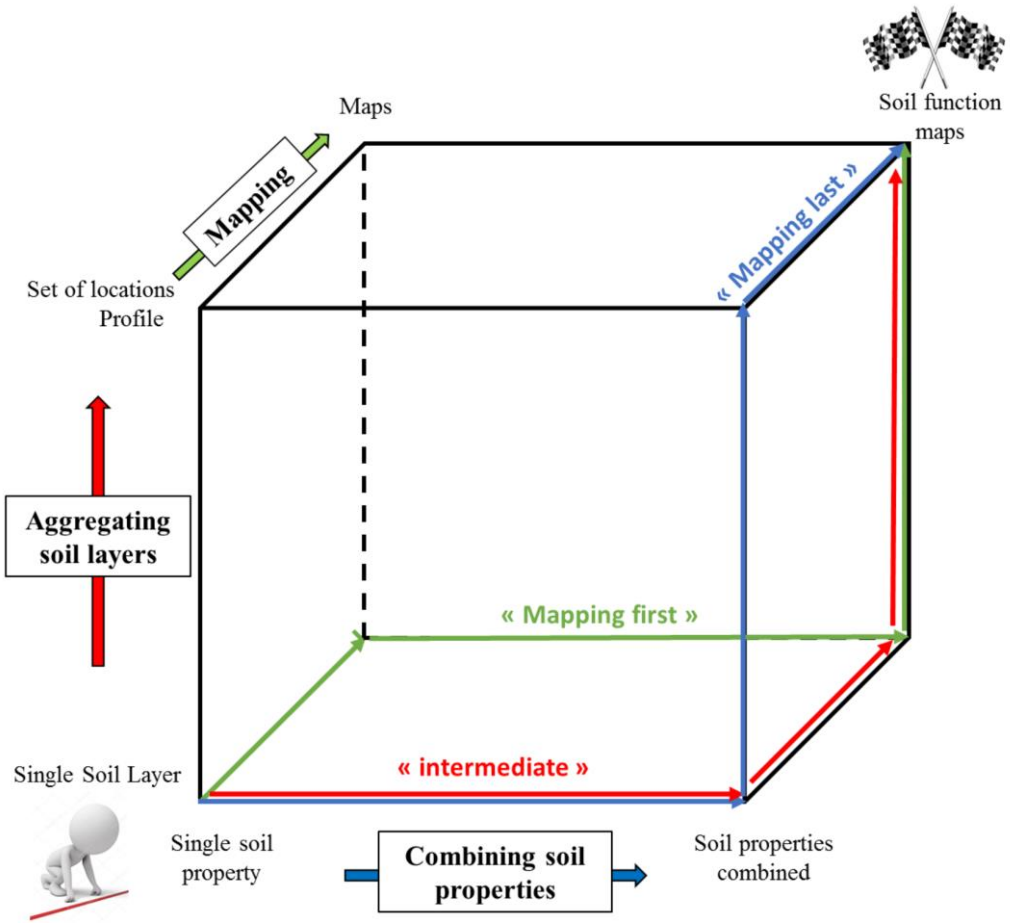


Figure 3.1. Three examples of computing trajectories for producing soil function maps

Apart from these three examples, one can imagine a large number of other inference trajectories since both partial combinations of soil properties and partial aggregations of soil layers can be envisaged.

Considering that SFA could be applied following these different mapping trajectories, the following question comes: "which inference trajectory provides the most accurate SFA output?". Recently, Laborczi et al (2019) provided a partial response to this question. They compared two inference trajectories for mapping sand, silt and clay content at 0-30 cm depth, which consisted of either first aggregating the GlobalSoilMap layers (0-5 cm, 5-15 cm and 15-30 cm) then mapping or first mapping each individual layer and then aggregating the layers. They obtained significantly different outputs between the different trajectories, the former

trajectory providing slightly worse predictions than the latter one. To our knowledge, such a comparison has not been extended yet to a more generic case study involving the whole inference trajectories for computing soil functions.

In this paper, we address the above-evoked question by taking as an example the mapping of the soil available water capacity (SAWC) of a soil profile. Although SAWC cannot be considered as a soil function per se, it shares the same issues of the choice of inference trajectories. Furthermore, SAWC is involved in the assessment of several ecosystem services (Adhikari and Hartemink, 2016), either provisioning services - food, fuel and fiber production - or regulating services - climate and gas regulation, water regulation or erosion and flood control. SAWC is mapped over the Languedoc-Roussillon region using the same inputs as earlier DSM papers dealing with the same region (Vaysse and Lagacherie, 2015, 2017).

II. The problem

SAWC is a well-known concept that has been used for a long time to express the capacity of soils to store water for plants (Veihmayer and Hendrickson, 1927); SAWC is computed using a classic expression:

$$SAWC = \sum_{i=1}^{n} dh_i * bd_i * \left(\frac{100 - st_i}{100}\right) * (\theta r_i - \theta w_i) \qquad \text{(Eq. 3.1.)}$$

where $SAWC$ is the soil available water capacity (cm), dh_i = thickness of the ith horizon (cm), bd_i = bulk density of the ith horizon, st_i = coarse fragment content of the ith horizon (% volumetric), and θr_i and θw_i are soil water contents at field capacity (FC) and at permanent wilting point (PWP) of the ith horizon ($cm^3.cm^{-3}$), respectively.

When the mapping of SAWC is targeted, some modifications to Equation 3.1 are necessary.

First, bulk density, soil water contents at FC and at PWP are expensive-to-measure soil properties that are rarely available in current soil databases, which prevents their mapping. The alternative solution is to estimate these data from more easily mappable primary soil properties (e.g., particle size fractions and organic carbon) by using pedotransfer functions (PTFs; Bouma, 1989). Many PTFs have been developed for estimating the soil properties required for computing SAWC (Wösten et al., 1999; Al Majou et al., 2007). In particular, PTFs can be used to calculate volumetric water contents at FC and PWP (e.g., Wösten et al., 1999), which embeds the bulk density information and avoids using a specific PTF to estimate bulk density.

Volumetric water contents are estimated as follows:

$$\hat{\theta}r_i = \sum_{j=1}^{n} \alpha_j PP_j + \varepsilon \quad \text{(Eq.3.2.)}$$

$$\hat{\theta}w_i = \sum_{j=1}^{n} \alpha_j PP_j + \varepsilon \quad \text{(Eq.3.3.)}$$

where $\hat{\theta}r_i$ = volumetric soil water content at FC ($cm^3.cm^{-3}$), $\hat{\theta}w_i$ = volumetric soil water content at PWP ($cm^3.cm^{-3}$), α_1 ... α_n = the coefficient of the model, $PP_1 \dots PP_n$ is the value of the selected primary soil properties as input, and ε is an estimated error.

A second modification of Equation 3.1 is the replacement of horizons whose thicknesses are variable across locations by soil layers with fixed depths. This has been introduced in digital soil mapping as a simple way for dealing with soil variability across depths (Malone et al., 2009). The general principle is to fit a continuous depth function (spline) of a given property onto the values of the property for the successive horizons, which allows further estimations of the soil properties for any possible layer defined by a user-fixed interval of depth (Bishop et al., 1999). This enables the harmonization of soil depths intervals across DSM input locations, which greatly facilitates further mapping. A discretization into 6 layers (0-5, 5-15, 15-30, 30-60, 60-100 cm) has been adopted in the specifications of the *GlobalSoilMap* project (Arrouays et al., 2014), which has made this discretization the most commonly applied.

When PTFs and fixed depth soil layers are introduced in the SAWC formula, we obtain the following formula:

$$SAWC = \sum_{i=1}^{n} SLh_i * \left(\frac{100-st_i}{100}\right) * \left[\left(\sum_{j=1}^{n} \alpha_j PP_j + \varepsilon\right) - \left(\sum_{j=1}^{n} \alpha_j PP_j + \varepsilon\right)\right] \quad \text{(Eq.3.4.)}$$

where SLh_i is the thickness of soil layers fixed by soil depth interval for $i = 1, \dots, 6$ when the specifications of the *GlobalSoilMap* project are followed.

From Equation 3.4, the variety of possible inference trajectories for mapping SAWC can be represented as an implementation of Figure 3.1 (Figure 3.2). The blue axis shows three different ways to combine soil properties (the most straightforward ones, among others): 1) considering primary property (PP) i.e., no combination, 2) considering the volumetric soil water contents at FC and PWP (WCs) i.e., application of PTFs first and 3) considering the available water capacity of a given soil layer i (AWCi) i.e., whole application of Equation 3.4 first. On the red axis, we have six possibilities for soil discretization across depths (the most straightforward, among others): from considering the six *GlobalSoilMap*-fixed soil layers, i.e., no combination,

to a full aggregation of soil layers, i.e., aggregating layers to soil profile first. Four other possibilities are considered by merging an increasing number of layers (Figure 3.2).

From Figure 3.2, it is possible to define 18 possible inference trajectories for mapping SAWC, i.e., three modalities of property combinations x six modalities of combining soil layers. For the sake of readability, only 5 out of these 18 trajectories are presented in Figure 3.2. It must be noted that the most classic inference trajectory (mapping first then aggregating properties and layers), which fully dissociate mapping and SAWC computing, was included among the 18 tested strategies.

In this paper, we investigated whether these different inference trajectories provided similar results or not, the input data and the underlying formula (Equation 3.4) being the same for all trajectories. If not, we also wanted to know which trajectory provided the best mapping performance.

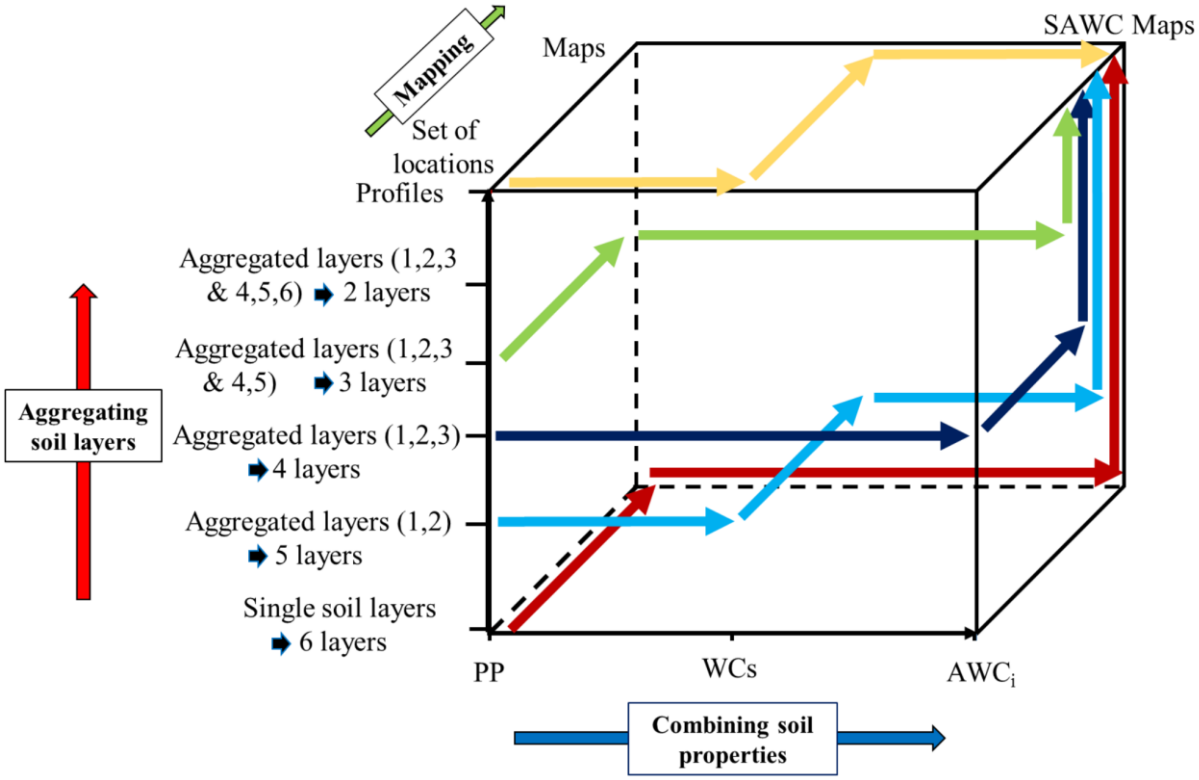


Figure 3.2. Concept of SAWC digital soil mapping with a few examples of inference trajectories

II.1. The case study

II.1.1. Study area

Languedoc-Roussillon is one of the 22 former administrative regions in France (Figure 3.3) and is now part of the country's newest region, Occitanie, which resulted from the merging of Languedoc Roussillon and Midi-Pyrénées. Located in Southern France, it covers 27,236 km² of land and stretches from the Mediterranean Sea to the Pyrenees and to the Massif Central Mountains. The region includes a wide variety of climates, parent materials and landscapes: low sedimentary plains with vineyards and/or cereals, dry limestone plateaus with scrublands and evergreen oak forests, slopes of Paleozoic mountains covered by forests, and volcanic and granitic highlands with grasslands. The soil cover of the region is consequently very diverse, including 18 WRB major soil groups that represent 56% and 75% of the total number of the soil group populations in the world and in Europe, respectively.

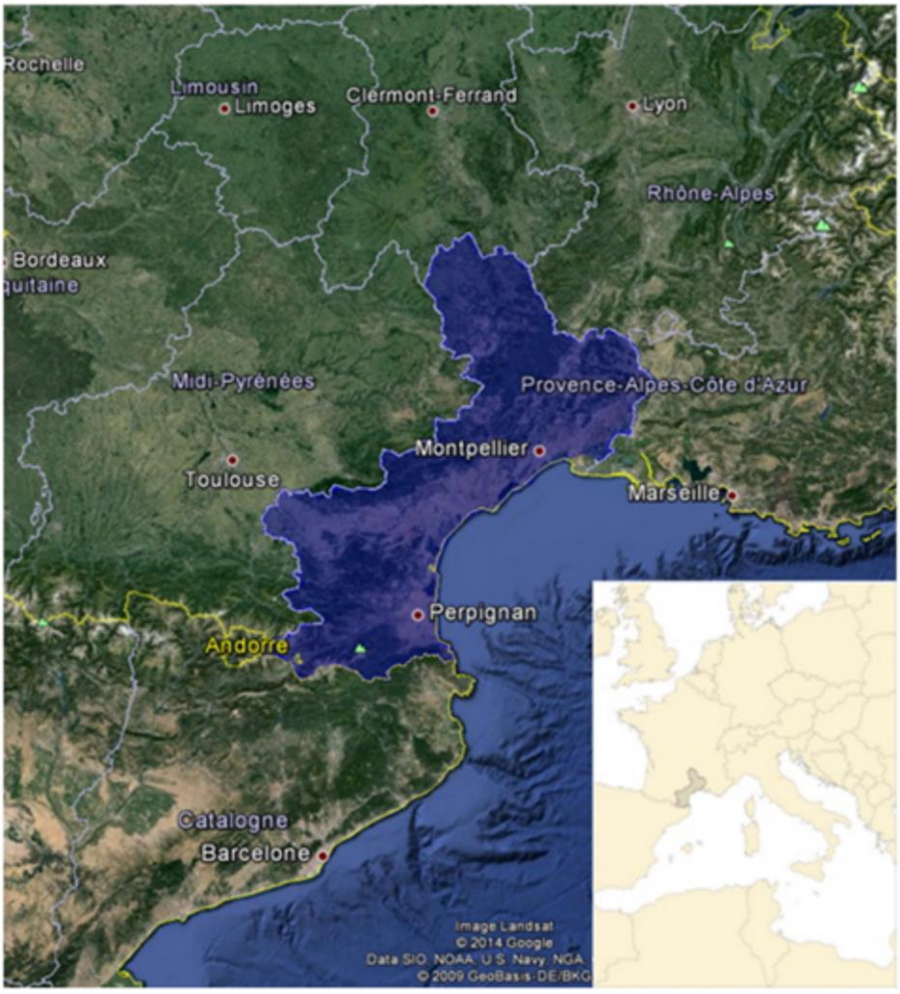


Figure 3.3. Location of the study area

II.1.2. Soil data input

In this study, we used a legacy dataset of 640 measured soil profiles selected from the 2024 used in Vaysse and Lagacherie (2015, 2017). This selection was made to ensure that the same input data was used by all of the inference trajectories. This motivated to select only the profiles that were fully documented for each soil property and for each layer. Indeed, this restrictive condition was necessary for providing non-null inputs for applying the "combine first" inference trajectory (i.e., combining all soil properties and for all soil layers before mapping). Finally, the density of this dataset of the study area was one soil profile for each 41 km².

Documenting Soil layer thickness

The soil layer thicknesses (Equation 3.4) that were initially defined through the fixed interval depth, 5, 10, 15, 30, 40, 100 cm for layers 1 to 6, respectively, needed to be updated to account for soils having a depth less than 200 cm. This was done from the following formula.

- If $SD > UL_i$, then $SLh_i = UL_i - LL_i$
- If $UL_i > SD > LL_i$, then $SLh_i = SD - LL_i$. (Eq.3.5.)
- If $SD < LL_i$, then $SLh_i = 0$.

With SD is the soil depth, SLh_i is the soil thickness of the soil layer and, UL_i and LL_i are the upper and lower limits of the considered soil layer, respectively.

Equation 3.5 requires first to document the soil depth (SD), i.e., the distance (in cm) from the soil surface to the bedrock or a paralithic contact (Soil Science Division Staff, 1993). This was done by a prior classification tree that determined whether or not the bottom horizon could be considered as bedrock or as a paralithic contact. Figure 3.4 shows the decision tree that was used for classifying the bottom horizon. The type of soil horizon was first used to identify lithic (R, M or D) vs pedological horizons. Since the horizon type, C, remained ambiguous with respect to the targeted classification, additional soil variables were used (types of structure, compaction and weathering), which allowed for the identification of paralithic horizons vs. pedological horizons.

From this classification of horizons, the soil depths were determined by applying the following rules:

- If the bottom horizon is lithic or paralithic, then the soil depth is the upper depth of the bottom horizon.
- If the bottom horizon is still a pedological one, the soil depth cannot be determined and thus the site is not selected. A total of 760 measured soil profiles were removed for this reason, which corresponded to 55% of the total removal of sites

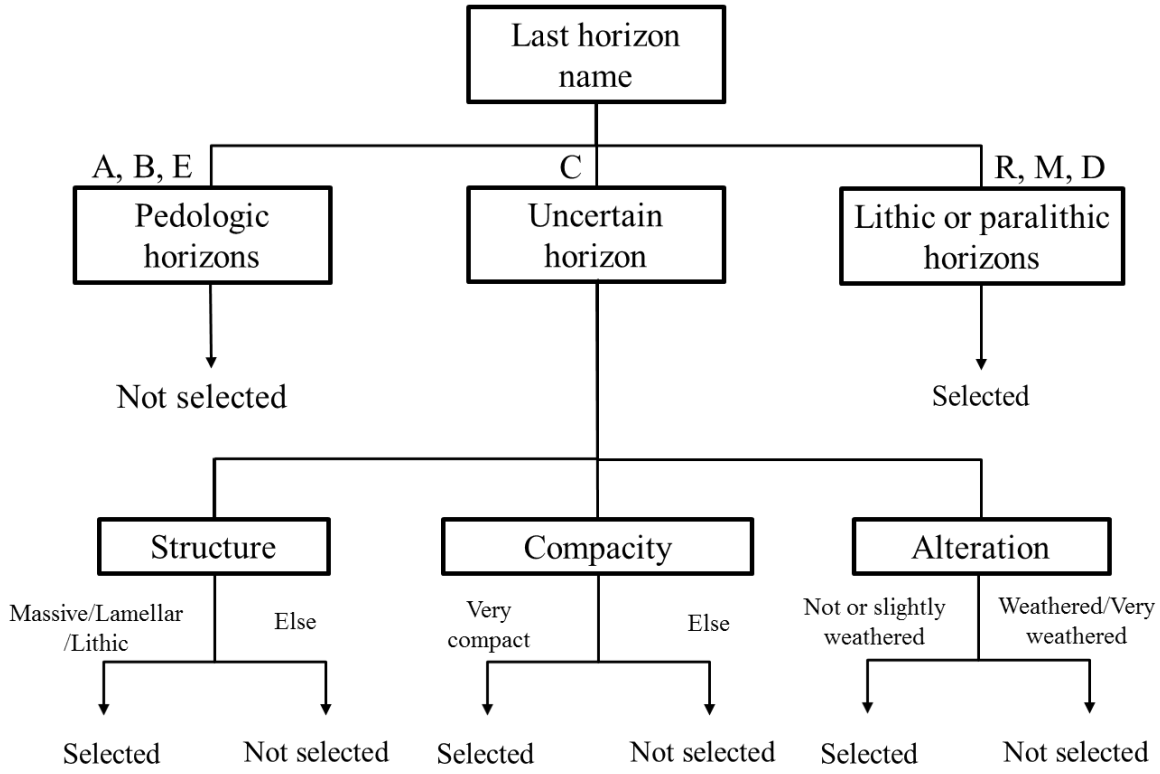


Figure 3.4. Representation of the classification tree applied to the dataset for identifying lithic or paralithic horizons and for selecting the input site

Documenting layers with the other required soil properties

As the described pedological horizons available in the soil database were defined by variable soil depths, a prior interpolation was required to document the soil layers. Spline functions work as an interpolator, respecting the average values of the target soil property, and assuming a continuous variation with depth (Bishop et al., 1999). As an outcome, spline functions deliver a set of interpolated values at specific depths which are, in our study, the depth intervals provided by *GlobalSoilMap* specifications (0-5, 5-15, 15-30, 30-60, 60-100 and 100-200 cm). Then, the mass-preserving spline functions were applied to clay, silt, sand and coarse fragment contents.

The values of the soil properties for the tested aggregated layers (see Figure 3.2) were derived from their values at the six depth intervals using a weighted mean by soil layer thickness.

Local continuous PTFs

As presented in section II, the common way to obtain soil water contents, that partially drives AWC (i.e., soil water contents at FC and PWP), is to use pedotransfer functions. The continuous PTFs used in this study are obtained from an investigation based on the analysis of a dataset of 294 pedological horizons belonging to 115 soil profiles located in the Hérault River valley in the former Languedoc Roussillon region (Leenhardt et al., 1994). The dataset contained the following variables: soil water volumetric contents at FC (10 kPa) and at PWP (1500 kPa), particle size fractions (clay, silt and sand), bulk density and organic carbon content. Before starting the investigation, bulk density and organic carbon were discarded from the dataset because not all horizons of the STIPA database reported their values, making PTFs inapplicable for local prediction. PTFs were fitted using a point estimation approach based on multiple linear regression carried out with the stats R package (R Development Core Team, 2009). Model variables were selected using the Akaike's information criterion (AIC; Akaike, 1973) applied to a backward/forward stepwise procedure. Linearity was visually analyzed on linear model graphs, then tested with the RESET test for nonlinearity (Ramsey, 1969). Linear models that fitted to the data set are reported in Table 3.1. Since particle size fractions are compositional data, stepwise selection will always remove at least one of them from the model. In this case, clay was selected for both models, sand better explained water content at field capacity predictions and silt, water content at permanent wilting point predictions. Performance assessments showed great performances for both PTFs, with a certain advantage for PWP's PTF.

Table 3.1. Outcomes of the calibrated continuous pedotransfer functions (PTFs) for calculations of water contents at field capacity (FC) and permanent wilting point (PWP)

Stratum	Water contents	B0	Sand	Silt	Clay	R^2	RMSE ($cm^3.cm^{-3}$)
All horizons (N = 294)	FC	0.4065	-0.0028		-0.0009	0.47	0.046
	PWP	-0.0064		0.0018	0.0036	0.63	0.036

II.1.3. Soil covariates

The DSM process used in this study is based on the well-known scorpan model (McBratney et al., 2003) that uses quantitative relationships between targeted soil property and environmental variables also called "covariates". The selection of the landscape covariates (Table 3.2) had been performed for a previous DSM application in the region (Vaysse and Lagacherie, 2015). It was based on two criteria: (i) the covariates could be derived from freely available geodatasets, at least at the French national level, and (ii) they have a logical and deterministic relationship to the soil properties, according to the literature. Classical geomorphometric indicators found in the DSM literature were computed from the global Shuttle Radar Topographic Mission (SRTM) digital elevation model (DEM): the elevation, slope, plan curvature, profile curvature, set of multiresolution valley bottom flatness (MRVBF), set of multiresolution ridge top flatness (MRRTF), topographic position index, and topographic wetness index.

The parent materials were characterized from the National Geological 1:50,000 map obtained from the French Geological Survey (Mansy et al., 2005). This map was translated into three parent material soil covariates, namely, the hardness, mineralogy and texture of alteration materials, following a mixed approach that involved both our pedologic knowledge and the measured legacy profiles described above (see details in Vaysse and Lagacherie, 2015). Land use was mapped across the region by a manual interpretation of Landsat 7 images from 2006. The initial classification into 43 land use types was condensed into nine types that were considered to be correlated with soil variations (e.g., artificial areas, greenhouse cultivation, permanent crops and orchards, forests, heathlands and pastures, scrublands, wetlands, complex territories composed of natural and agricultural areas, and forests in transition). The basic climate data (maximum temperature, minimum temperature, and precipitation) were extracted from the Global Climate Database at a resolution of 1 km² (Hijmans et al., 2005). Two aridity indexes were derived from these data, namely, the De Martonne index and the Emberger index (see details in Vaysse and Lagacherie, 2015). Additionally, we added to the covariate set the regional-scale soil map (1:250,000) that regroups the major landscape types across the region.

Table 3.2. Exhaustive categorical and continuous covariates

Variables	Abbreviation	Resolution/Scale	Source	Soil forming factor[1]	Type[2]
Topography					
Elevation	ELEV	90 m	SRTM	r	Q
Multi-resolution Valley Bottom Flatness	MRVBF	90 m	SRTM	r	Q
Slope	SLOPE	90 m	SRTM	r	Q
Topographic Wetness Index	TWI	90 m	SRTM	r	Q
Plan Curvature	PLANCURV	90 m	SRTM	r	Q
Profile Curvature	PROCURV	90 m	SRTM	r	Q
Multi-resolution Ridge Top Flatness	MRRTF	90 m	SRTM	r	Q
Topographic Position Index	TPI	90 m	SRTM	r	Q
Geology					
Hardness	HARDNESS	90 m	Geological map/soil profil	p	C
Texture	TEXTURE	90 m	Geological map/soil profil	p	C
Mineralogy	MINERALOGY	90 m	Geological map/soil profil	p	C
Climate					
Martonne Index	MARTONNE	90 m	WorldClim	c	C
Emberger Index	EMBERGER	90 m	WorldClim	c	C
Maximum temperature	TMAX	90 m	WorldClim	c	Q
Minimum temperature	TMIN	90 m	WorldClim	c	Q
Precipitation	PRECIPITATION	90 m	WorldClim	c	Q
Organisms					
Land use	LANDUSE	30 m	Landsat 7	o	C
Soil					
Soil map	SOILMAP	1:250 000	RRP	s	Q

[1]: SCORPAN factors (s=soil property, c = climate, o = organisms, r = relief, p=parent material)

[2]: Q = quantitative, C = categorical

SRTM = Shuttle Radar Topographic Mission ; RRP = Référentiel Régional Pédologique.

III. Methods

III.1. Mapping Model : Quantile Regression Forest

The mapping model selected for this work is the Quantile Random Forest (QRF) (Meinshausen, 2006), which is a very popular machine-learning algorithm in recent DSM operational applications (Vaysse and Lagacherie, 2015, 2017, Nussbaum et al., 2018). In this section, we first describe the random forest algorithm (Breiman, 2001) from which QRF was extended and then provide the specific features of QRF. In the following, we use excerpts from Meinshausen (2006) to present these algorithms. More details can be found in this paper. We let Y be a real-valued response variable and X be a covariate or predictor variable. A standard goal of statistical analysis is to infer the relationship between Y and X. The random forest algorithm grows a large ensemble of decision trees using n independent observations (Y_i, X_i), $i = 1, \ldots, n$. For each tree and each node, the random forest algorithm implements a random selection to determine a variable at which to split. For each tree, a bagged version of the training data is used. In addition, only a random subset of predictor variables is considered for the split point selection at each node. For regression, the prediction of a single tree of the random forest algorithm for a new data point, X = x, can be represented as the weighted average of the original observations Y_i, $i = 1, \ldots, n$:

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(x, \theta) Y_i \qquad \text{(Eq. 3.6.)}$$

where $w_i(x, \theta)$ is the weight vector given by a positive constant if the observation X_i is part of the same leaf of the tree built from the random vector of variables in which X was dropped and is otherwise 0. Using the RF algorithm, the conditional mean, $E(Y \mid X = x)$, is approximated by the averaged predictions of k individual trees, each constructed with an independent and identically distributed vector t, $t = 1, \ldots, k$. We let $w_i(x)$ be the average of $w_i(T)$ over such a collection of trees:

$$w_i(x) = k^{-1} \sum_{t=1}^{n} w_i(x, \theta_t) \qquad \text{(Eq. 3.7.)}$$

The final predictions are the average of predictions of individual trees (Breiman, 2001; Prasad et al., 2006; Biau and Scornet, 2016):

$$\hat{\mu}(x) = \sum_{i=1}^{n} w_i(x)\, Y_i \qquad \text{(Eq. 3.8.)}$$

Recently, Wright, and Ziegler (2017) developed a fast implementation of Breiman's random forest (Breiman, 2001) and Meinshausen's quantile regression forest (Meinshausen, 2006), for high-dimensional data, which is available as the *ranger* R package (Wright and Ziegler, 2017). This implementation is faster than QRF in terms of i) superior scaling of the number of samples, trees, and descriptive variables tried for splitting and ii) runtime and memory usage (see details in Wright and Ziegler, 2017). QRF was applied for mapping all possible soil properties considered in the inference trajectories. In particular, the thicknesses of soil layers (SLh_i in Equation 3.4) were mapped separately so that all trajectories could be applied. However, the thicknesses of the upper soil layers (0–5 cm, 5–15 cm, 15– 30 cm, 0–15 cm, and 0–30 cm) were not variable enough across the region for mapping (variance < 20 cm²). In that case, the predicted values at each location were fixed as the mean soil layer thickness.

III.2. Evaluation protocol

The performances of digital mapping of SAWC and of all its components (primary soil properties, hydraulic properties (soil water contents at FC and PWP for different soil layers)) were evaluated by a k-fold cross validation. This evaluation procedure involves randomly dividing the data into k subsets. Then, the holdout method is repeated k times, such that each time one of the k subsets is used as the evaluation set, the other k-1 subsets are merged to form the calibration set. Following this procedure, every data point is included in an evaluation set exactly once and is included in a calibration set k-1 times. In our case, we choose k = 10 and, to increase the robustness of the evaluation, the 10-fold cross-validation was repeated 20 times. The k-fold cross validation was performed using the *cvTools* R package (Alfons, 2012). To evaluate the prediction performances, we used classic performance indicators, e.g., mean square error skill score (SS_{MSE}) (Nussbaum et al., 2018) that has the same interpretation than the percentage of variance explained by the model, root mean square error (RMSE) and bias. The performance indicators were calculated for each of the 20 iterations and then averaged to obtain a final pooled value.

Since soil compartment differs from one AWC mapping trajectory to another, soil properties and AWC were predicted for varying soil layers. The evaluation protocol was adapted accordingly.

IV. Results

IV.1. Preliminary Results

In this section, we present basic statistics of the soil properties used for calculating as well as their correlations. We focused on correlations since they play a major role in error propagation (Heuvelink et al., 1989), which, therefore, can be of great interest for comparing the performance of the tested inference trajectories.

IV.1.1. Basic Statistics

We provide in Table 3.3 the basic statistics of the study area. Languedoc Roussillon presents soil texture with predominant sand fraction and significant amount of coarse fragment content, which varies consequently.

Table 3.3. Basic statistics of soil properties for soil horizons

Variable	Unit	Mean	CV (%)
Clay	% mass	21.46	58.18
Silt	% mass	33.99	37.12
Sand	% mass	44.45	41.76
Coarse fragment	% vol	31.63	83.27
Thickness	cm	102.41	40.40
FC	$cm^3.cm^{-3}$	0.26	16.91
PWP	$cm^3.cm^{-3}$	0.13	39.55
AWC	cm	2.64	79.35

FC: volumetric water content at field capacity; PWP: volumetric water content at permanent wilting point; AWC: available water capacity; CV: coefficient of variation.

IV.1.2. Correlation tables of input data: combining soil properties

Table 3.4 shows the averaged correlations between the properties involved in the SAWC computing (Equation 3.4) with standard deviations that show the variations in correlation across the different considered soil layers. Soil thickness was independent of the other properties, whereas the coarse fragment content was weakly, but significantly, correlated with silt and sand. Additionally, sand content was highly negatively correlated to silt and clay contents. As expected, FC and PWP showed high correlations with the input properties of their PTFs (Table 3.1) and with each other, as the result of application of PTFs that used the same properties or strongly correlated ones (Table 3.1). Low standard deviations of all these correlations revealed their low variations across the soil layers.

Table 3.4. Soil layer properties combinations averaged on soil profiles when soil layers are support SAWC computation.

	Soil properties					
	Clay	Silt	Sand	Coarse fragment	Thickness	FC
Silt	0.13 (0.016)***	-	-	-	-	-
Sand	-0.73 (0.001)***	-0.76 (0.010)***	-	-	-	-
Coarse fragment	-0.07 (0.031)^{ns}	-0.29 (0.031)***	0.24 (0.023)***	-	-	-
Thickness	0.06 (0.017)*	-0.05 (0.011)^{ns}	-0.01 (0.007)^{ns}	0.03 (0.043)**	-	-
FC	0.61 (0.003)***	0.86 (0.007)***	-0.99 (0.001)***	-0.27 (0.023)***	-0.01 (0.008)^{ns}	-
PWP	0.90 (0.002)***	0.55 (0.016)***	-0.96 (0.001)***	-0.19 (0.024)***	0.03 (0.010)^{ns}	0.90 (0.002)***

ns: not significant (p-value>0.05); *: p-value ≤0.05; **: p-value ≤0.01; ***: p-value ≤ 0.001; FC : volumetric water content at field capacity; PWP: volumetric water content at permanent wilting point.

IV.1.3.Correlation Tables of input data: aggregating soil layers

Table 3.5 shows the averaged correlations of properties between soil layers with standard deviations that show the variations in correlation across the soil properties. The three upper layers were greatly correlated while the correlations decreased for the deeper layers.

Table 3.5. Correlations of combined soil properties and available water capacity (AWC) across depths

Number of layers	**Soil properties**	**Depth intervals (cm)**					
6	All[1]		0-5	5-15	15-30	30-60	60-100
		5-15	0.99 (0.005)***	-	-	-	-
		15-30	0.94 (0.030)***	0.96 (0.015)***	-	-	-
		30-60	0.78 (0.084)***	0.79 (0.080)***	0.87 (0.044)***	-	-
		60-100	0.58 (0.141)***	0.57 (0.139)***	0.64 (0.107)***	0.85 (0.038)***	-
		100-200	0.40 (0.174)**	0.41 (0.174)**	0.46 (0.162)***	0.58 (0.137)***	0.78 (0.164)***

ns: correlation not significant (p-value>0.05); *: p-value ≤0.05; **: p-value ≤0.01; ***: p-value ≤ 0.001; [1]: clay, silt, sand, coarse fragment contents, thickness, volumetric water content at field capacity (FC), volumetric water content at permanent wilting point (PWP), available water capacity (AWC)

IV.2. Primary and hydric property mapping and performances

For the trajectories that did not map the SAWC but rather its components, Table 3.6 provides a summary of the performances obtained for each property across the different numbers of soil layers. For sake of readability, only SS_{MSE} values are shown with mean, min, and max values across the set of trajectories with different numbers of soil layers. Concerning the primary soil properties, particle size fractions mapping exhibited the best performances, while coarse fragment content and soil thickness were very badly mapped. The hydraulic properties showed slightly similar results to those of particle size fractions from which they are derived.

Table 3.6. Results of mean square error skill score (SS_{MSE}) for primary and hydraulic properties for different number of layers

Variable	Unit	SS_{MSE}		
		Min (SD)	Max (SD)	Mean (SD)
Clay	% mass	0.11 (0.050)	0.25 (0.024)	0.17 (0.027)
Silt	% mass	0.10 (0.076)	0.28 (0.028)	0.20 (0.021)
Sand	% mass	0.20 (0.083)	0.33 (0.012)	0.28 (0.024)
Coarse fragment	% vol	0.05 (0.030)	0.08 (0.016)	0.06 (0.024)
Thickness	cm	-0.02 (0.085)	0.07 (0.036)	0.03 (0.056)
FC	$cm^3.cm^{-3}$	0.19 (0.092)	0.32 (0.009)	0.28 (0.028)
PWP	$cm^3.cm^{-3}$	0.20 (0.058)	0.32 (0.010)	0.26 (0.025)

SD : standard deviation; FC : volumetric water content at field capacity; PWP : volumetric water content at permanent wilting point

It is also interesting to examine the correlations between mapping errors across properties and layers since, according to the error propagation formula (Heuvelink et al., 1989), it plays a role in the final errors on AWC. Table 3.7 shows the correlations between mapping residuals for the soil properties used in the trajectories. High correlations of residuals were observed between sand and the two other textural fractions, whereas moderate but significant correlations of residuals can be noticed between coarse fragment and two of the textural fractions (silt and sand).

Table 3.7. Residuals correlations for soil property combinations

	Soil properties					
	Clay	**Silt**	**Sand**	**Coarse fragment**	**Thickness**	**FC**
Silt	0.03 (0.095)^{ns}	-	-	-	-	-
Sand	-0.69 (0.027)***	-0.72 (0.047)***	-	-	-	-
Coarse fragment	-0.08 (0.044)^{ns}	-0.23 (0.139)***	0.22 (0.102)***	-	-	-
Thickness	0.02 (0.060)^{ns}	-0.02 (0.057)^{ns}	0.00 (0.059)^{ns}	-0.01 (0.108)^{ns}	-	-
FC	0.54 (0.043)***	0.83 (0.027)***	-0.98 (0.004)***	-0.23 (0.117)***	-0.01 (0.060)^{ns}	-
PWP	0.88 (0.010)***	0.47 (0.087)***	-0.94 (0.011)***	-0.17 (0.080)**	0.01 (0.052)^{ns}	0.86 (0.026)***

ns: not significant (p-value>0.05); *: p-value ≤0.05; **: p-value ≤0.01; ***: p-value ≤ 0.001; FC : volumetric water content at field capacity; PWP : volumetric water content at permanent wilting point.

Table 3.8 presents averaged correlations of property residuals between the six GlobalSoilMap soil layers across the set of soil properties with their associated standard deviation. The pattern of correlation of the soil properties residuals between layers is similar to the one drawn by the correlation of input data (Table 3.4), i.e., (i) the great correlation of surface soil layers and (ii) decreasing correlation when the residuals correlation coefficient between aggregated layers drops below 0.9.

Table 3.8. Residuals correlation of pooled properties across depths

Number of layers	**Soil properties**	**Depth intervals (cm)**					
6	All[1]		0-5	5-15	15-30	30-60	60-100
		5-15	0.98 (0.003)***	-	-	-	-
		15-30	0.92 (0.026)***	0.94 (0.011)***	-	-	-
		30-60	0.75 (0.056)***	0.75 (0.049)***	0.84 (0.028)***	-	-
		60-100	0.57 (0.110)***	0.48 (0.213)^{ns}	0.51 (0.220)^{ns}	0.81 (0.048)***	-
		100-200	0.39 (0.147)***	0.33 (0.184)^{ns}	0.36 (0.189)^{ns}	0.51 (0.121)***	0.65 (0.285)***

ns: not significant (p-value>0.05); *: p-value ≤0.05; **: p-value ≤0.01; ***: p-value ≤ 0.001; [1]: clay, silt, sand, coarse fragment contents, thickness, volumetric water content at field capacity (FC), volumetric water content at permanent wilting point (PWP), available water capacity (AWC)

IV.3. SAWC mapping trajectories performance comparisons

In this study, we tested 18 SAWC mapping trajectories. Table 3.9 shows the indicators of performance, SS_{MSE}, RMSE, and Bias of the 18 tested trajectories with their mean values and standard deviations across the 20 iterations of the evaluation protocol.

The results showed substantial differences in performance across the tested trajectories (SS_{MSE} between -0.02 and 0.37). These differences occurred both between trajectories that mapped different soil properties (SL.PP, SL.WC, and SL.AWC) and between trajectories that did not consider the same number of layers, although the former often exhibited differences that could be interpreted as being within the error margin of the evaluation process (Lagacherie et al., 2019). Roughly the same hierarchy of performances was found whatever the examined indicator.

Concerning the differences in performance between trajectories that mapped different soil properties, the best performances were mainly obtained by calculating first an AWC (SL.AWC) then mapping, although mapping the soil properties first (SL.PP) seems slightly better than calculating AWC first for the six-layer trajectory. However, this ranking should be considered with care because of the above noticed small differences of performances. Concerning the difference in performance between trajectories that do not consider the same number of layers, the maximum performances were obtained by considering four soil layers. The performances increased moderately from the six-layer trajectory to the four-layer trajectory, i.e., from SS_{MSE} = 0.27 to SS_{MSE} = 0.37 for SL.AWC trajectories and then dramatically decreased from the three-layer trajectory to the one-layer- trajectory, i.e., from SS_{MSE} = 0.34 to SS_{MSE} = -0.02 for SL.AWC trajectories.

According to the performance indicators, the best performance was obtained by an intermediate trajectory that considered four soil layers and mapped SAWC. It is also interesting to note that the classic trajectory, i.e., using DSM outputs for calculating a spatialized SAWC (the SL.PP trajectory with six layers) was not the best among the 18 tested.

Table 3.9. Averaged values of performance indicators SS_{MSE}, root mean square error (RMSE), and Bias, and their corresponding standard deviation (SD).

Number of layer	Depth (cm)	Trajectory	SS_{MSE}	RMSE (cm)	Bias (cm)
			Mean (SD)	*Mean (SD)*	*Mean (SD)*
6	0-200	SL.PP	0.29 (0.006)	4.21 (0.019)	0.5 (0.018)
		SL.WC	0.27 (0.006)	4.27 (0.02)	0.25 (0.018)
		SL.AWC	0.27 (0.004)	4.27 (0.012)	-0.54 (0.013)
5	0-200	SL.PP	0.30 (0.003)	4.19 (0.01)	-0.37 (0.009)
		SL.WC	0.33 (0.004)	4.08 (0.013)	-0.24 (0.008)
		SL.AWC	0.35 (0.003)	4.02 (0.01)	-0.07 (0.01)
4	0-200	SL.PP	0.34 (0.004)	4.05 (0.014)	-0.28 (0.011)
		SL.WC	0.35 (0.005)	4.02 (0.015)	-0.18 (0.009)
		SL.AWC	0.37 (0.005)	3.96 (0.017)	0 (0.008)
3	0-200	SL.PP	0.31 (0.006)	4.15 (0.018)	-0.24 (0.014)
		SL.WC	0.32 (0.005)	4.12 (0.015)	-0.15 (0.013)
		SL.AWC	0.34 (0.004)	4.07 (0.015)	0.04 (0.011)
2	0-200	SL.PP	0.16 (0.009)	4.57 (0.027)	-0.24 (0.017)
		SL.WC	0.17 (0.008)	4.55 (0.023)	-0.15 (0.018)
		SL.AWC	0.20 (0.009)	4.48 (0.027)	-0.01 (0.014)
1	0-200	SP.PP	-0.01 (0.010)	4.78 (0.026)	0.71 (0.025)
		SP.WC	0.03 (0.009)	4.7 (0.023)	0.31 (0.03)
		SP.AWC	-0.02 (0.009)	4.79 (0.024)	-0.68 (0.023)

V. Discussion

V.1. Level of performances and limitations

V.1.1. Level of performances in predicting primary soil properties

In this study, we considered inference trajectories that involved mapping of soil primary properties by using a DSM model. The partial results presented in section IV.2 can then be compared to those of previous studies in similar conditions. The performances delivered for particle size fractions were slightly worse than those provided recently by Román Dobarco et al. (2019) for all of France (R^2 = 0.27, 0.43, and 0.46 for clay, silt, and sand content, respectively) with, however, the same hierarchy between sand, silt, and clay mapping performances. Vaysse and Lagacherie (2015) obtained in the same study area better mapping performances for particle size fractions (between 0.19 and 0.36) but the same poor performances for thickness and coarse fragment, however, with another validation technique (independent validation test instead of cross validation). A comparison with results from Vaysse

and Lagacherie (2017) that used a cross validation confirmed the previous results for a given soil property (clay content for the 5–15 cm layer). The performance gap could be explained by the reduction in the size of the input dataset between Vaysse and Lagacherie (2017) applications and this study (from 1945 to 640).

V.1.2. Level of performances in predicting SAWC

Among the 18 inference trajectories, the most appropriate for predicting SAWC was the one considering four soil layers (e.g., 0–30, 30–60, 60–100, and 100–200 cm) to directly predict SAWC. This trajectory obtained much better results than any mapping of SAWC components, i.e., individual soil properties at a given depth. This can be explained by the removal of intraprofile variabilities and noise that did not play any more when SAWC is considered. In addition, this inference trajectory exhibited similar results to those obtained by Leenhardt et al. (1994) with a R^2 ranges from 0.36 to 0.45 according to the scale of the SAWC map chosen. Except for this study, evaluation of SAWC mapping has rarely been applied and this paper is, to our knowledge, among the first that has performed such an evaluation. However, this evaluation remains incomplete since we did not take into account the error of the PTFs coefficient applied. However, Román Dobarco (2019) found that PTFs error played a minor role in SAWC mapping error.

V.2. Drivers of the variability in performance between trajectories

The results exhibited substantial differences of mapping performance across the tested inference trajectories (Table 3.9). The most important differences in performance were observed when different numbers of soil layers were considered. The best performances were obtained when highly correlated soil layers (correlation > 0.9 between layers 1, 2, and 3, cf. Table 3.5) were merged before mapping whereas the worst ones were obtained when fewer correlated soil layers were merged (correlation < 0.80 between layers, Table 3.5). The former came from averaging the DSM model inputs, which may decrease the input errors of the DSM models whereas the latter forced the DSM models to cope with contradictory drivers of the combined soil layer, which could negatively impact the performance. This result is partially in accordance with Heuvelink and Pebesma (1999) who advocated a mapping first trajectory since it “enables a more efficient use of the spatial characteristics of the individual inputs”. We however brought a nuance to this statement by considering a correlation threshold beyond which combining first could be a better solution. These differences were not modified by the further combination of the soil layer maps for calculating the final SAWC since no noticeable

differences were observed in mapping error correlations (Table 3.8) that could have created differences in error propagation through this additive operation (Heuvelink et al., 1989). The differences in performance observed between trajectories that mapped different soil properties (primary soil properties, hydraulic properties, or SAWC) were more difficult to interpret. First the differences in mapping performance were much smaller than previously observed. Furthermore, in contrast to the combination of soil layers, the combination of soil properties is multiplicative (Equation 3.4). Therefore, the impacts of the correlations between soil properties and between their mapping errors could be very different. More comparisons of such inference trajectories involving contrasted mapping results and a larger range of soil function expressions would be necessary for a full understanding and an ex ante prediction of the hierarchy of performances across the trajectories. Finally, it must be noticed that the comparisons of the inference trajectories "all things being equal" required considering the same number of measured soil profiles across trajectories. This corresponded to the number of fully complete measured soil profiles that were necessary for mapping SAWC after combining all soil properties and all soil layers. However, other less restrictive inference strategies could use more locations for mapping some of the SAWC components. Indeed, mapping separately the soil properties at each soil layer would allow an increase in the amount of input data which could increase the mapping performance of some SAWC components and, in turn, might have a positive effect on the overall performances. However, this reasoning does not hold for the depth prediction that is still limited by the soil input, whatever the trajectories. The value added provided to the final mapping of SAWC should therefore be investigated in the future.

V.3. Toward soil spatial information systems

Digital mapping of soil function introduces an additional degree of complexity relative to the usually practiced monovariate digital soil mapping. Different inference trajectories could be envisaged and we showed that the decision of selecting one or another could substantially impact the performance of the soil function mapping. This can provide motivation to develop tools that would select the best possible trajectories from a prior knowledge of error propagation mechanisms and of causes of mapping errors. Such tools may refer to the ideas of soil inference systems applied to the building of pedotransfer functions (McBratney et al., 2002) that were further extended to digital soil mapping by Lagacherie and McBratney (2007). This could imply a revision of the strategies of diffusion of the digital soil mapping products since we showed that the current practice, i.e., providing a set of spatial layers of soil properties that could be further combined for obtaining soil function maps could not be the optimal one. An alternative

could be to provide to the users the soil spatial inference system of a given region so that they could produce the best possible soil function map by themselves.

VI. Conclusions

The main lessons of this study are as follows.

- A large number of inference trajectories can be envisaged for mapping soil functions. This makes the mapping of soil function much more complex than classic monovariate digital soil mapping.
- Mapping a single value per soil profile that traduces a soil function (for example, soil available water capacity) gives better results in terms of the explained variance than mapping individual soil properties for individual soil layers.
- The best trajectory is not found among the extreme ones (e.g., first map the individual properties then combine or the converse).
- The decision of combining soil layers before mapping should be made after looking at the correlations of soil properties between layers.

Les éléments importants de ce chapitre à retenir sont :

- Les estimations du réservoir utile sont sensibles aux trajectoires de calcul.
- Les estimations du réservoir utile sont plus sensibles aux différences d'agrégation **de couches de sols** (ΔSS_{MSE} = 0.39) **qu'aux différences de combinaisons des propriétés** (ΔSS_{MSE} = 0.05).
- Les trajectoires répertoriées dans la littérature ne figurent pas parmi les plus adaptées à la spatialisation du réservoir utile.
- La trajectoire de calcul ayant fourni les meilleures performances consiste à **spatialiser des valeurs de réservoir utile** estimées sur une agrégation partielle des six couches *GlobalSoilMap* en **quatre couches de sol** (0-5 cm, 5-15 cm et 15- 30 cm agrégées en 0-30 cm par moyenne pondérée des propriétés par l'épaisseur des couches) avant **agrégation de ces couches de sols**.
- La moins performante des trajectoires de calcul consistait à **spatialiser des valeurs du réservoir utile** estimées sur **une unique couche** de sol résultant de **l'agrégation totale des six couches de sol *GlobalSoilMap***.
- Le choix de la trajectoire de calcul peut être anticipé par une étude des corrélations **des propriétés primaires de sols** entre elles et entre **les couches de sols**. Les trajectoires les plus performantes ont été celles qui consistaient à agréger les couches de sols fortement corrélées avant spatialisation (r > 0.9) alors que les trajectoires les moins performantes spatialisaient des couches préalablement agrégées alors qu'elles étaient moins corrélées (r < 0.8).

CHAPITRE 4 Quantification et spatialisation de l'incertitude liée à la spatialisation du réservoir utile par propagation d'erreurs

Selon les spécifications du programme *GlobalSoilMap*, tout produit cartographique doit être accompagné d'une quantification de l'incertitude (Arrouays et al., 2014). Si ces spécifications sont communément appliquées dans le cadre d'une spatialisation d'une propriété primaire de sol, l'application au réservoir utile et à son caractère multivarié nécessitent de prendre en compte la propagation des erreurs de spatialisation de différents composants du réservoir utile. Le principe du modèle de propagation d'erreur à construire est de tenir compte d'une part des erreurs générées par l'utilisation du modèle et d'autre part, des corrélations entre erreurs de spatialisation de propriétés ou de couches de sols. Les erreurs de spatialisation sont, dans ce chapitre, estimées par un algorithme de forêts de régressions quantiles (Quantile Regression Forest, QRF) ayant la particularité de pouvoir quantifier sa propre erreur.

Ce chapitre est consacré à l'évaluation de la qualité de prédiction de l'incertitude en comparant plusieurs modalités de propagation des erreurs selon que certaines corrélations entre erreurs de spatialisation soient prises en compte ou non. Les enseignements du chapitre précédent (Chapitre 3) sont pris en compte dans ce chapitre, notamment sur le choix de la trajectoire de calcul la plus adaptée à la spatialisation du réservoir utile, avec une légère modification pour les besoins de cette étude.

Le chapitre se présente sous la forme d'un article soumis à Geoderma.

Uncertainty assessment of soil available water capacity using error propagation: a test in Languedoc-Roussillon

Styc Quentin and Lagacherie Philippe

Submitted to Geoderma

Abstract

Soil available water capacity (SAWC) is a key soil indicator that plays a major role in many ecosystem services, such as food production, irrigation management, soil drought, flood control, and climate and gas regulation. Digital soil mapping (DSM) can be used to obtain needed SAWC maps. However, SAWC differs from the usual soil properties considered in DSM in that it involves several soil properties determined at several soil layers. Therefore, a specific approach is required to obtain SAWC maps and the associated uncertainty predictions. The objective of this study was to build a SAWC mapping approach that could predict SAWC values at three maximum rooting depths (60, 100 and 200 cm) and their associated prediction uncertainties. The approach was tested in the Languedoc-Roussillon region (southern France). Elementary available water capacities of each layers (in $cm.cm^{-1}$) and soil layer thicknesses were first mapped separately at 0-30, 30-60, 60-100 and 100-200 cm and then aggregated to estimate the SAWCs at the three mentioned maximum rooting depths. SAWC uncertainty was estimated with an error propagation model that used a first-order Taylor analysis. This analysis considered the mapping errors of each involved property, which were estimated by the quantile regression forest algorithm. We tested different error propagation models that differently considered the correlations between these mapping errors: no correlation considered, correlations between soil layer thicknesses and elementary water capacities per soil layer only, correlations between soil layers only, or all correlations considered.

The performances of both SAWC predictions and their uncertainties were assessed with a 10-fold cross validation that was iterated 20 times. The SAWC predictions showed poor accuracies (percentages of explained variance ranged from 0.12 to 0.13). The uncertainties of SAWC predictions were best estimated when the correlations between the soil layer errors were considered in the error propagation model whereas the uncertainties of SAWC predictions were severely underestimated when these correlations were neglected. In spite of the poor performance in predicting SAWC at the punctual level due to the low density of soil

observations (1/19 km^2), the SAWC approach appeared promising since it produced maps that agreed with the available pedological knowledge and precisely estimated the uncertainties. Improvement of the current approach would require the collection of more soil data, such as either legacy data or new measurements or proxies of SAWC.

I. Introduction

Soil available water capacity (SAWC) refers to the capacity of soils to store water for plants (Veihmayer and Hendrickson, 1927). SAWC is a key soil indicator that plays a major role in many ecosystem services, such as food production, irrigation management, soil drought, flood control, and climate and gas regulation. It is therefore a fundamental parameter that has been used in land evaluations and recently in soil ecosystem service assessments (Dominati et al., 2014). Currently, SAWC is operationally computed in the literature as follows (Cousin et al., 2003) (Equation 4.1):

$$SAWC = \sum_{i=1}^{n} dh_i * bd_i * \left(\frac{100 - st_i}{100}\right) * (\theta r_i - \theta w_i) \qquad \text{(Eq. 4.1.)}$$

where SAWC is the soil available water capacity (cm), dh_i = the thickness of the *i*th horizon (cm), bd_i = the bulk density of the *i*th horizon (g.cm^{-3}), st_i = the coarse fragment content of the *i*th horizon (% volumetric), and θr_i and θw_i are the soil water contents at field capacity (FC) (i.e., the soil water content that remains in the soil after water has drained due to gravitational force) and at permanent wilting point (PWP) (i.e., the soil water retained so strongly that it is no longer available for plant roots, so plants wither and cannot recover their turgidity) of the *i*th horizon (g.g^{-1}), respectively.

To meet the need to map the SAWC, digital soil mapping (McBratney, 2003) can be considered an adequate approach since it provides the best solution for synergizing all data on the soils and its drivers that can be available in a given region, regardless of its size. Styc and Lagacherie (2019) proposed a modified formulation for SAWC (Equation 4.2) to make it more easily mappable from the available primary soil properties.

$$SAWC = \sum_{i=1}^{n} SLh_i * \left(\frac{100-st_i}{100}\right) * \left[\left(\sum_{j=1}^{n} \alpha_j PP_j + \varepsilon_{FC}\right) - \left(\sum_{j=1}^{n} \alpha_j PP_j + \varepsilon_{PWP}\right)\right] \qquad \text{(Eq.4.2.)}$$

where SLh_i is the thickness of the soil layers fixed by the soil depth interval, st_i is the coarse fragment content of the *i*th horizon (% volumetric), $\alpha_1 \ldots \alpha_n$ are the coefficients of the

pedotransfer functions used to calculate the volumetric water contents at the field capacity and permanent wilting point respectively, $PP_1 \dots PP_n$ are the values of the primary soil properties used as inputs for the pedotransfer functions (most often textural fractions) and, ε_{FC} and ε_{PWP} are the errors of the pedotransfer functions used to estimate field capacity and permanent wilting point, respectively. Equation 4.2 shows that the SAWC determination involved several primary soil properties at several soil layers, which created methodological questions that have not been addressed by the classical DSM framework. First, there is no consensus about the inference trajectory selected for predicting SAWC. In the DSM literature, there were i) studies that calculate first AWC at the observed sites prior using these sites for calibrating a mapping function (Vanderlinden et al., 2005; Poggio et al., 2010; Hong et al., 2013) and ii) studies that mapped the AWC components first and then combined the mapping outputs to obtain an estimate of AWC (Ugbaje and Reuter, 2013, Leenaars et al., 2018, Román Dobarco et al., 2019). However, a comparison of 18 inference trajectories that combined different AWC calculations from the primary soil properties, soil layer aggregation and mapping (Styc and Lagacherie, 2019) showed significant differences in SAWC prediction accuracies, and none of the two inference trajectories cited above were optimal. The best inference trajectory was an intermediate trajectory that, before mapping, calculated the AWC for four soil layers.

Another methodological question was the ex-ante uncertainty assessment of the SAWC mapping output. In the classical DSM framework, the different models can provide a local estimate of the uncertainty of the predicted values of the target soil properties (Heuvelink et al, 2014, Vaysse & Lagacherie, 2017). Obtaining a similar estimate for the SAWC map requires an error propagation model that combines the different errors associated with the mapping of each layer of the soil properties involved in the SAWC calculation. Román Dobarco et al. (2019) used a first-order Taylor analysis to propagate mapping errors and pedotransfer function errors to the final SAWC predictions. They showed that the mapping of the SAWC components (soil texture and coarse fragment) was the main source of SAWC mapping uncertainty. However, the soil thickness mapping errors were not considered in their analysis, although Algayer et al. (2019) demonstrated that soil thickness can be the most critical component in the SAWC estimation. Furthermore, the error propagation model proposed by Román Dobarco et al. (2019) neglected the error correlations between the SAWC components, which assumed that these errors were independent of each other, which has not been demonstrated.

The objective of this study was to build a SAWC mapping approach with the best possible inference strategy that could predict all SAWC values for three maximum rooting depths (60,

100 and 200 cm) and their associated prediction uncertainties while taking into account all SAWC component mapping errors and their correlations. The approach was tested in the former Languedoc-Roussillon region (Southern France).

II. The case study

II.1. Study area

This study was carried out in the former Languedoc-Roussillon French administrative region, which is now part of the new Occitanie region (Figure 4.1). Located in southern France, the former region covers 27,236 km² of land that stretches from the Mediterranean Sea to the Pyrenees and Massif Central mountains. The region includes a wide-ranging diversity of climates, geologies, and landscapes that lead to a large pedodiversity, with 18 WRB major soil groups, which represent 75% of all soil groups in Europe, being included in this study area. Further details can be found in Vaysse and Lagacherie (2015, 2017).

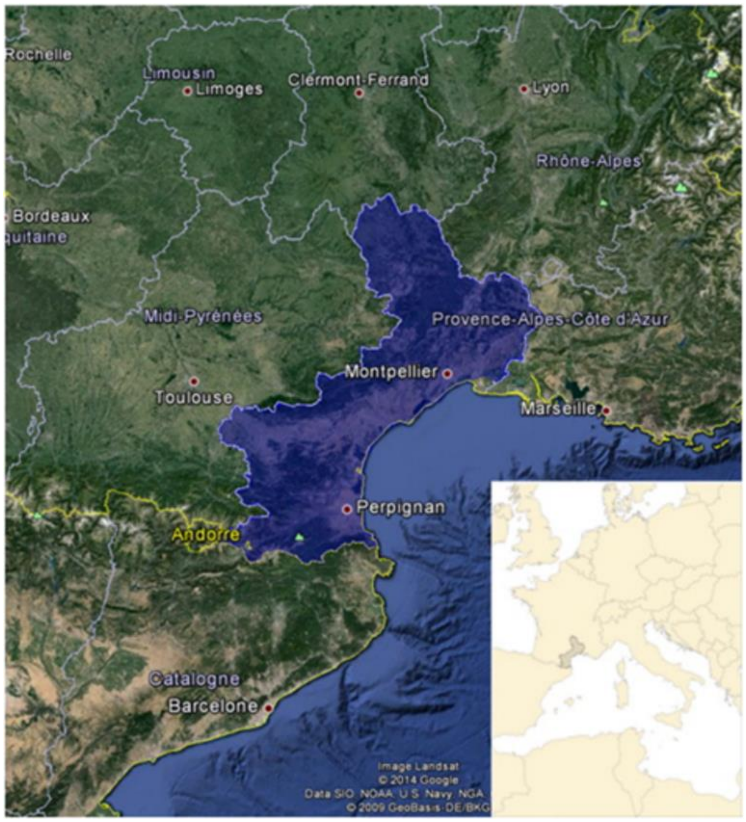


Figure 4.1. Location of the study case

II.2. Soil profiles with observed SAWC components

In this study, we used a legacy dataset of 2024 measured soil profiles from Vaysse and Lagacherie (2015, 2017). SAWC for the different soil profiles were harmonized for providing

AWC components values at the fixed soil layers that were considered for SAWC mapping (0-30 cm, 30-60 cm, 60-100 cm and 100-200 cm) (see explanations further in section II.3). For that, each AWC component was estimated at these layers from the initial soil horizons using mass conservation cubic splines (Bishop et al., 1999). For each considered soil layer, we selected the profiles at which all AWC components (soil texture, coarse fragments and soil layer thicknesses) were fully documented (see details below). This resulted in a reduction of the number of soil profiles for mapping each soil layer (Table 4.1).

Table 4.1. Number of soil profiles for each layer

Soil layer depths (cm)	Number of soil profiles
0-30	1464
30-60	1323
60-100	1064
100-200	822

Documenting soil thickness

The soil layer thicknesses (SLT) were documented by considering the following rules:

- If the lower limit of the soil layer (LL) was less than both the maximum soil observation depth (MSOD) of the soil profiles and the upper depth of a lithic or paralithic contact (UDPLC), then SLT was equal to the difference between its fixed lower and upper limits (e.g. the SLT of 30-60 cm soil layer is 30 cm)
- Else if LL was less than MSOD but greater than UDPLC, SLT was equal to the difference between UDPLC and LL (e.g. the SLT of the 30-60 cm layer with a lithic contact appearing at 50 cm is 20 cm)
- Else if LL was greater than MSOD, the SLT could not be determined, which lead to remove the soil layer from the input soil datase.
- Else if LL was greater than UDPLC, SLT was equal to 0 cm as the soil is no longer present.

II.3. Pedotransfer functions

In this study, we used the national-level pedotransfer functions (PTFs) developed by Román Dobarco et al. (2019) because our case study was in the domain of applicability of these PTFs, which ensured the best possible performances (Román Dobarco et al., 2019). The volumetric soil water contents at field capacity (Equation 4.3) and permanent wilting point

(Equation 4.4) used clay and sand contents (%) as the predictive variables, which were calculated as follows:

$$\hat{\theta}r_i = 0.278 + 2.45\ 10^{-3}\ Clay - 1.35\ 10^{-3}\ Sand \quad \text{(Eq.4.3.)}$$

$$\hat{\theta}w_i = 0.08 + 4.01\ 10^{-3}\ Clay - 2.93\ 10^{-4}\ Sand \quad \text{(Eq.4.4.)}$$

where $\hat{\theta}r_i$ and $\hat{\theta}w_i$ are the volumetric water contents at field capacity and permanent wilting point, respectively.

II.4. Soil covariates

The employed DSM process, which relies on the *scorpan* model (McBratney et al., 2003), used the quantitative relationships between the target soil properties and available spatial variables related with soil, which are also called the "soil covariates".

The soil covariates of this study area were selected by Vaysse and Lagacherie (2015) following two criteria: i) they could be derived from freely available geo-datasets for at least the French national level, and ii) they had a logical and process-based relationship with soil properties according to the literature. The soil covariates (Table 4.2) accounted for the impact of topography, climate, organisms, and parent material. The regional-scale map (1:250,000) that delineated the major landscape types across the region was also considered as a soil covariate. All the soil covariates were computed at the nodes of the 90 m x 90 m grid of the SRTM digital elevation model, which corresponded also to the resolution of the predicted SAWC map. More details can be found in the descriptions of several applications of DSM to the region (Vaysse and Lagacherie, 2015, 2017; Styc and Lagacherie, 2019).

Table 4.2. The soil covariates

Variables	Abbreviation	Resolution/Scale	Source	Soil forming factor[1]	Type[2]
Topography					
Elevation	ELEV	90 m	SRTM	r	Q
Multi-resolution Valley Bottom Flatness	MRVBF	90 m	SRTM	r	Q
Slope	SLOPE	90 m	SRTM	r	Q
Topographic Wetness Index	TWI	90 m	SRTM	r	Q
Plan Curvature	PLANCURV	90 m	SRTM	r	Q
Profile Curvature	PROCURV	90 m	SRTM	r	Q
Multi-resolution Ridge Top Flatness	MRRTF	90 m	SRTM	r	Q
Topographic Position Index	TPI	90 m	SRTM	r	Q
Geology					
Hardness	HARDNESS	90 m	Geological map/soil profil	p	C
Texture	TEXTURE	90 m	Geological map/soil profil	p	C
Mineralogy	MINERALOGY	90 m	Geological map/soil profil	p	C
Climate					
Martonne Index	MARTONNE	90 m	WorldClim	c	Q
Emberger Index	EMBERGER	90 m	WorldClim	c	Q
Maximum Tempertaure	TMAX	90 m	WorldClim	c	Q
Minimum Tempertaure	TMIN	90 m	WorldClim	c	Q
Precipitation	PRECIPITATION	90 m	WorldClim	c	Q
Organisms					
Land use	LANDUSE	90 m	Landsat 7	o	C
Soil					
Soil Map	SOILMAP	1 :250,000	RRP	s	C

[1]: SCORPAN factors (s=soil property, c = climate, o = organisms, r = relief, p=parent material)

[2]: Q = quantitative, C = categorical

SRTM = Shuttle Radar Topography Mission ; RRP = Référentiel Régional Pédologique

III. Methods

III.1. Random Forest

Random forest models (Breiman, 2001) are an ensemble learning method for both classification and regression. A forest, which is an ensemble of randomized decision trees, is built and trained based on a bootstrap approach. Individual trees are built using the principle of recursive partitioning. "*The feature space is recursively split into regions containing observations with similar response value*" (Strobl et al., 2009). The predictions of the individual trees are finally averaged to give a single prediction.

III.2. Mapping model: the quantile regression forest

In this study, we use one of the most commonly used algorithms in DSM studies, namely, the quantile regression forest algorithm (QRF; Meinshausen, 2006), which is an extension of Breiman's random forests (RF; 2001). For the regression, RF provides an ensemble prediction based on *n* regression trees. For every tree, the algorithm integrates random features by randomly selecting a subset of features to be split. While RF provides solely the conditional mean, QRF supplies the whole conditional distribution of the target variable by keeping all observations at the terminal nodes and can infer estimates for the conditional quantiles (Meinshausen, 2006). More details on QRF can be found in Meinshausen (2006).

QRF was run with the *ranger* package, which is a fast implementation of Breiman's random forest and Meinshausen's quantile regression forest (Wright and Ziegler, 2015).

III.3. Inference trajectories

Since SAWC is a soil indicator that involves several soil properties at several soil layers, it can be estimated following various possible inference following the order with which "combining primary soil properties", "aggregating soil layers across depths" and "mapping" are executed to provide the targeted output (Styc and Lagacherie, 2019). Styc and Lagacherie (2019) tested a total of 18 inference trajectories for throughout Languedoc-Roussillon that were performed to obtain the most appropriate SAWC map. From this study, we considered the best performing inference trajectory, i.e., we computed first AWC in four layers (0-30, 30-60, 60-100 and 100-200 cm) obtained by merging the three first layers defined in the GlobalSoilMap specifications (Arrouays et al., 2014), mapped them and then aggregating the maps of the four soil layers to obtain the final SAWC map. To account for the different possible rooting depths

across the different crops, these aggregations were performed over three different maximal rooting depth (60 cm, 100 cm and 200 cm).

However, we modified the inference trajectory (Figure 4.2) by mapping, the soil thickness and the elementary available water capacity (AWC_E, Equation 4.5) separately for each layer. AWC_E represents the water retention capacity for one centimeter of soil (in $cm.cm^{-1}$) and is defined as follows:

$$AWC_E = \left(\frac{100-st_i}{100}\right) * \left[\left(\sum_{j=1}^{n} \alpha_j PP_j + \varepsilon_{FC}\right) - \left(\sum_{j=1}^{n} \alpha_j PP_j + \varepsilon_{PWP}\right)\right] \quad \text{(Eq. 4.5.)}$$

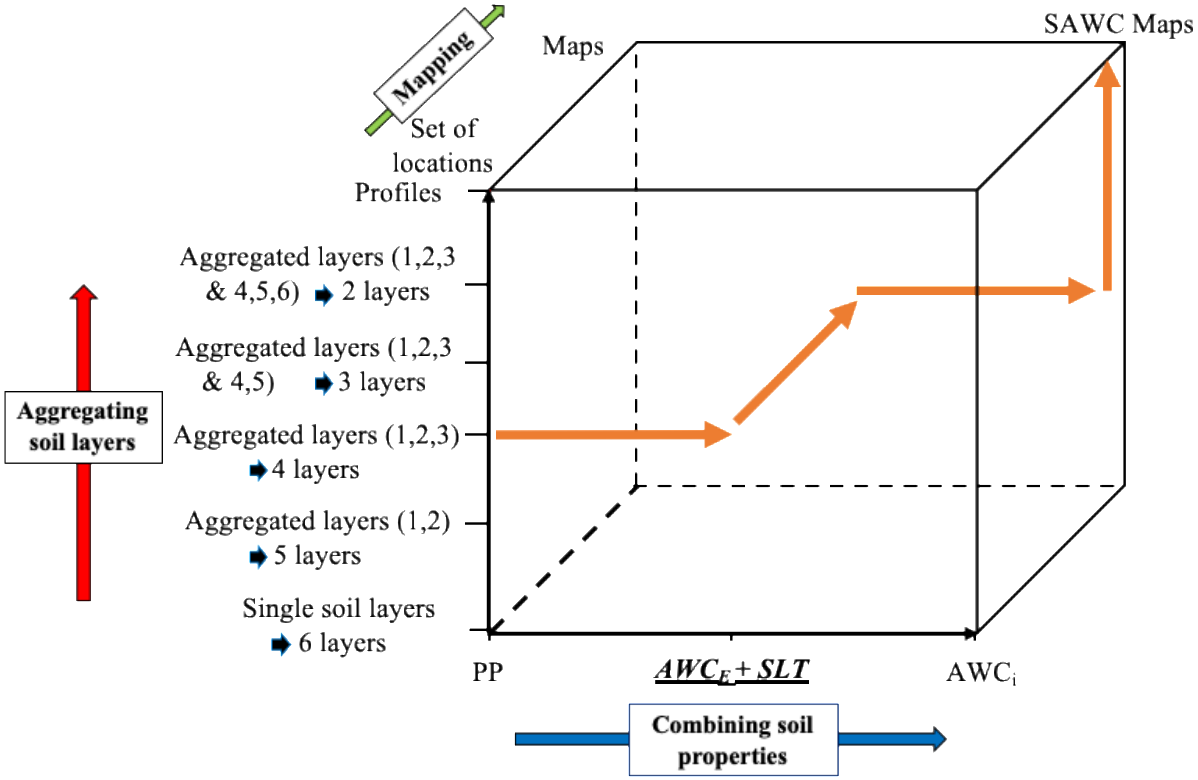


Figure 4.2. Conceptual diagram of SAWC digital soil mapping with an example of inference trajectory including the new level of soil property combination, elementary available water capacity (AWC_E) and soil layer thickness (modified from Styc and Lagacherie, 2019)

The rationale of this modification was to separately map two soil properties that exhibited very low correlations, meaning that their variations could result from different landscape drivers that could be imperfectly considered by a single mapping model.

III.4. Uncertainty analysis using error propagation modeling

Following Román Dobarco et al. (2019), the error propagation was modeled using first-order Taylor expansion to calculate the variance of the SAWC predictions. This variance was considered a proxy of the prediction uncertainty of the target variable (Heuvelink et al., 1989).

This method relies on the approximation of the estimates obtained for the soil property (i.e., the available water capacity). Let Y be an estimation of a given soil property as follows (Equation 4.6):

$$Y = f(z) \qquad \text{(Eq.4.6.)}$$

where f is a continuously differentiable function from $\mathbb{R}^n$ into $\mathbb{R}$ and z is the vector of the n input variable of f. The approximation of f uses a series centered on the mean values of the n input variables $\mu = [\mu_1, \mu_2, \ldots, \mu_n]^T$ (Heuvelink et al., 1989). The variance of $Y = f(z)$ is calculated with the following formula (Equation 4.7):

$$\sigma_y^2 \approx \sum_{i=1}^{n} \left(\frac{\delta f(\mu)}{\delta z_i}\right)^2 \sigma z_i^2 + 2 \sum_{\substack{i=1 \\ i \neq k}}^{n} \sum_{\substack{k=1 \\ k \neq i}}^{n} \left(\frac{\delta f(\mu)}{\delta z_i}\right)\left(\frac{\delta f(\mu)}{\delta z_k}\right) \sigma z_i z_k \qquad \text{(Eq.4.7.)}$$

where σ_y^2 is the variance of Y, z_i and z_k are the soil input variables, $\sigma z_i z_k$ is the covariation of the z_i and z_k errors from the i and k variables, σz_i is the standard deviation of z_i and $\frac{\delta f(\mu)}{\delta z_i}$ and $\frac{\delta f(\mu)}{\delta z_k}$ are the partial derivatives of f(z) around μ. σz_i is estimated by the standard deviations of the conditional distributions provided by QRF at each predicted location (Meinshausen, 2006). Then, the estimate of the variance can be used to compute the limits of the confidence interval. Assuming a normal distribution, the limits of the confidence interval can be computed as follows (Equation 4.8):

$$CIL_i = \hat{y}_i \pm 1.645\, \sigma_{\hat{y}_i} \qquad \text{(Eq.4.8.)}$$

where CIL_i is the interval limits of the prediction, $\hat{y}_i$ the mean of the distribution, $\sigma_{\hat{y}}$ the standard deviation and 1.645 is the Student's coefficient for a 90% confidence interval estimation.

Error propagation was performed using the *propagate* R package (Spiess, 2018).

III.5. The experiment

In this study, we considered two sources of uncertainty: i) the mapping error of SAWC components, i.e., the soil thickness and the AWC_E and ii) the error of the AWC of every soil layer.

III.5.1.The tested options of error propagation

To evaluate the importance of the correlations of the AWC components errors for the quantification of SAWC uncertainty, four options of error propagation were considered according to whether are considered: i) the error correlations between the predicted properties involved in the determination of AWC at each layer, i.e., AWC_E and soil layer thickness (denoted further SP), ii) the error correlations between the predicted AWC at different soil layers (denoted further SL), iii) both of these correlations (denoted further SP.SL) or iv) none of this correlations (denoted further NONE). To compute the error correlations, we considered the residuals calculated by the k-fold cross validation (see next section).

Additionally, we derived the SAWC predictions according to three different fixed maximum soil depth, i.e. 200 cm, 100 cm and 60 cm. The rationale was to determine if the predictions for the deepest layers (60-100 cm and 100-200 cm) played a beneficial or unbeneficial role in the SAWC predictions.

III.5.2. Evaluation protocol

The performance of the SAWC DSM was evaluated by k-fold cross validation. This evaluation procedure consisted of randomly dividing the data into k subsets. Then, the holdout method was repeated k times such that one of the k subsets was used as the validation set in each repetition, while the other k-1 subsets were merged to form the calibration set. Following this procedure, every data point was included in a calibration set k-1 times. In this study, we selected k = 10; to increase the robustness of the evaluation, the 10-fold cross validation was iterated 20 times. The k-fold cross validation was performed using *cvTools* (Alfons, 2012) that was used to define the folds.

To evaluate the prediction performances, we used classic performance indicators, e.g., the mean square error skill score (SS_{MSE}, Nussbaum et al., 2017), which has the same interpretation as the percentage of variance explained by the model, root mean square error (RMSE) and bias.

Furthermore, we evaluated the estimation of the prediction uncertainty using the prediction interval coverage probability (PICP, Equation 4.9) (Shrestha and Solomatine, 2006), which was computed as follows:

$$PICP = \frac{count(LPL_i \leq y_i \leq UPL_i)}{n} \times 100 \qquad \text{(Eq.4.9.)}$$

where n is the total number of observations in the validation set, and the numerator counts if the observation y_i fits within the prediction limits prior to estimation by the error propagation

method. For a 90% confidence level, which is usually chosen in DSM studies (Arrouays et al., 2014b), the uncertainty is optimally predicted when the PICP value is close to 90%.

In addition to PICP, we verified that the largest errors were at locations having the largest widths of estimated prediction intervals. For that, the population of validation sites was split into four quartiles of predicted interval widths and four RMSEs were computed separately for each quartile.

IV. Results

IV.1. Basic statistics

IV.1.1. Soil input distributions

In Figure 4.3, we present the distributions of the soil thickness and the elementary available water capacity across the set of soil profiles that was used as input of the mapping model.

The soil thickness ranged from 5 to 200 cm (i.e., the maximum soil observation depth was fixed at 200 cm), with the average ST at 89 cm, which was close to the median value (90 cm). The most common ST in the dataset was 120 cm soil thickness. Then, the distribution dropped dramatically, showing that deep soils were less represented than shallow soils.

The shape of the distribution was far from normal, although the skewness and kurtosis tests indicated a normal distribution. The AWC_E for the 0-30 cm and 30-60 cm soil layers were nearly normally distributed, with peaks at 0.09 and 0.10, respectively. The AWC_E for the 60-100 cm and 100-200 cm soil layers differed from the shallowest layers by the shapes of their distributions, which were bimodal (two distribution peaks located at 0.03 cm/cm and at 0.11 cm/cm). While the skewness test indicated that the distributions of the AWC_E values of every soil layer were approximatively symmetric, the excess kurtosis test showed that the distributions were less peaked and presented less extreme values than a normal distribution.

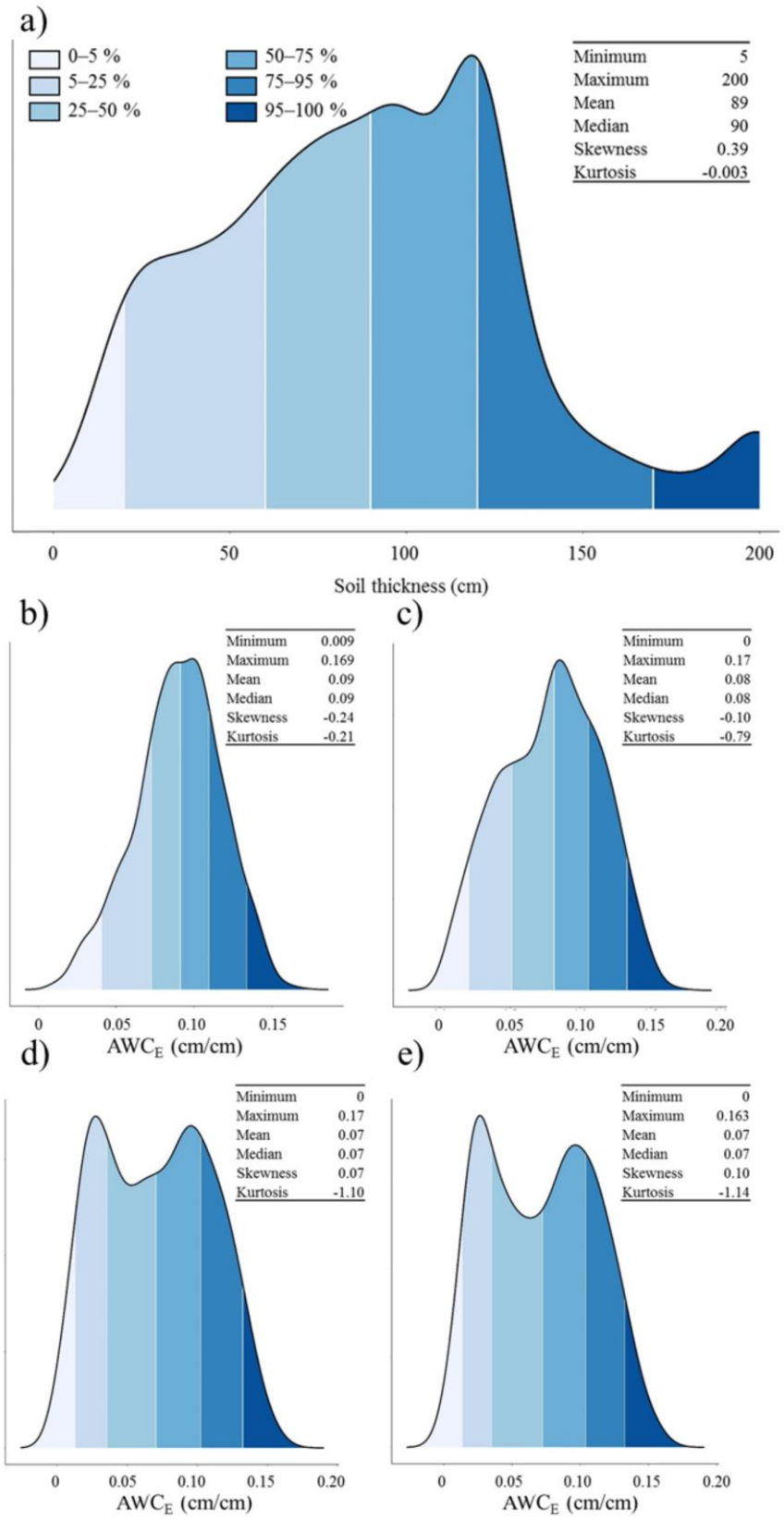


Figure 4.3. Distributions of the soil input variables; a) soil thickness, b) elementary available water capacity (AWC_E) of the 0-30 cm soil layer, c) elementary available water capacity of the 30-60 cm soil layer, d) elementary available water capacity of the 60-100 cm soil layer and e) elementary available water capacity of the 100-200 cm soil layer

IV.1.2.Correlation of error between the AWC_E and the soil thickness

The correlations between the errors from AWC_E and soil thickness (Table 4.3) were very low regardless of the considered soil layer and even nonsignificant for layer 0-30 cm and 60-100 cm. It may reveal that the mapping models of these soil properties were very different by weighing very differently the used soil covariates.

Table 4.3. Correlation of the error between the elementary available water capacity (AWC_E) and the soil layer thickness

Soil properties					
		AWC_E			
	Depth intervals (cm)	0-30	30-60	60-100	100-200
Soil layer thickness	0-30	0,01 ^{na}	-	-	-
	30-60	-	0,09 **	-	-
	60-100	-	-	0,02 ^{na}	-
	100-200	-	-	-	-0,13 *

AWC_E : elementary available water capacity; ns: not significant (p-value>0.05); *: p-value ≤0.05; **: p-value ≤0.01; ***: p-value ≤ 0.001

IV.1.3.Correlation of the errors between soil layers

Table 4.4 shows the correlation of the AWC errors between soil layers. The AWC errors of the soil layers were correlated, especially for the consecutive soil layers, and the correlations decreased for deeper layers, which may denote large similarities between mapping models of consecutive soil layers. It is worth noting that all error correlations between the soil layers were significant and generally higher than the error correlations between the primary soil properties.

Table 4.4. Correlation of the available water capacity (AWC) error between the soil layers

Soil property	**Depth intervals (cm)**			
		0-30	30-60	60-100
AWC	30-60	0,67 ***	-	-
	60-100	0,42 ***	0,63 ***	-
	100-200	0,16 **	0,24 ***	0,36 ***

AWC : available water capacity; ns: not significant (p-value>0.05); *: p-value ≤0.05; **: p-value ≤0.01; ***: p-value ≤ 0.001

IV.2. SAWC component prediction performances

In Table 4.5, we present the performances of the SAWC component predictions. While the SS_{MSE} values of both the clay and sand contents increased in depth, SS_{MSE} of both the coarse fragment content and soil thickness showed very low performances (e.g., the SS_{MSE} was close to 0%) as well low RMSE and bias values. The performances of the AWC_E indicated a weak SS_{MSE} that ranged from 6 to 14% of explained variance.

Table 4.5. Prediction performances of soil available water capacity (SAWC) components

Soil properties	Unit	Depth interval (cm)	SS_{MSE}	RMSE (cm)	Bias (cm)
Clay	(% mass)	0-30	0.16 (0.007)	10 (0)	-2 (0)
		30-60	0.2 (0.006)	11 (0)	-2 (0)
		60-100	0.26 (0.008)	12 (0)	-2 (0)
		100-200	0.27 (0.011)	12 (0)	-2 (0)
Sand	(% mass)	0-30	0.29 (0.005)	15 (0)	1 (0)
		30-60	0.29 (0.006)	18 (0)	0 (0)
		60-100	0.33 (0.008)	16 (0)	0 (0)
		100-200	0.32 (0.009)	17 (0)	-1 (0)
Coarse fragments	(% vol)	0-30	0.07 (0.008)	20 (0)	-4 (0)
		30-60	0.09 (0.01)	26 (0)	-4 (0)
		60-100	0.07 (0.013)	29 (0)	-3 (0)
		100-200	0.03 (0.013)	30 (0)	-3 (0)
Soil thickness	(cm)	0-30	-0.01 (0.005)	3.43 (0)	0.92 (0.005)
		30-60	-0.04 (0.009)	9.52 (0)	3.58 (0.023)
		60-100	0 (0.015)	17.07 (0)	3.78 (0.140)
		100-200	0.01 (0.013)	26.09 (0)	-8.82 (0.083)
AWC_E	(cm/cm)	0-30	0.12 (0.007)	0.03 (0)	0.002 (0.0001)
		30-60	0.14 (0.007)	0.03 (0)	0.003 (0.0001)
		60-100	0.11 (0.008)	0.04 (0)	0.001 (0.0001)
		100-200	0.06 (0.008)	0.04 (0)	0.001 (0.0001)

AWC_E : elementary available water capacity

We present in Table 4.6 the performances of the SAWC predictions obtained by aggregating the predicting of AWC_E and soil thickness of the four aggregated soil layers until the selected maximum rooting depths (60 cm, 100 cm or 200 cm).

SAWC was poorly predicted regardless of the maximum rooting depth considered. The variance explained by the model reached 12-13%. Slightly positive bias values (between 0.51 and 0.97 cm) denoted a slight overestimation of SAWC.

Table 4.6. Performances of the soil available water capacity (SAWC) predictions obtained by a 10-fold cross-validation that was iterated 20 times with SS_{MSE} (mean square error skill score) RMSE (root mean square error) and bias. The results are presented as the means of the 20 iterations and their standard deviations

Soil thickness (cm)	SS_{MSE}	RMSE (cm)	Bias (cm)
0-200	0.12 (0.007)	4.2 (0.018)	0.66 (0.017)
0-100	0.12 (0.007)	3.29 (0.013)	0.97 (0.013)
0-60	0.13 (0.008)	1.86 (0.008)	0.51 (0.006)

IV.3. Uncertainty in the SAWC mapping prediction performances

In Table 4.7, we present the uncertainty evaluation of the predictions averaged with their standard deviations using the PICP. The PICP values ranged from 71% to 91%; the PICP values were closer to optimal (i.e., 90%) when the correlation of the errors between the soil layers was considered regardless of whether the correlation of errors between the soil properties were considered during error propagation. It is worth noting that when the correlation of the errors between the soil layers was not accounted for, the PICP dropped dramatically to values ranging between 71 and 81%, which led to an underestimation of the SAWC uncertainty.

Table 4.7. Uncertainty prediction evaluation using PICP (prediction interval coverage probability) with a mean of the 20 iterations associated with the standard deviations with several options for error propagation: the error correlation between both soil properties and soil layers (SP.SL), solely the soil layer error correlation (SL), solely the soil property error correlation (SP) or no correlation (NONE).

PICP (%)				
Soil thickness (cm)	Error correlation			
	SP.SL	SL	SP	NONE
0-200	91 (0)	91 (0)	75 (1)	75 (1)
0-100	88 (1)	87 (1)	71 (1)	71 (1)
0-60	89 (0)	88 (0)	81 (1)	80 (0)

Table 4.8 shows the differences in the prediction performances (RMSE) for the different quartiles in the 90% confidence interval width using the error propagation model for SL (i.e., the one that considered the correlation of the error between soil layers only) for SAWC predicted at 200 cm. The RMSE calculated separately for each quartile tended to increase from the smallest width to the largest confidence interval (from 3.12 cm to 5.51 cm). Therefore, as expected, the uncertainty predicted by the model was related to the uncertainty observed through the validation protocol. Similar trends were observed for SAWC predictions at 60 cm and 100 cm.

Table 4.8. RMSEs for the quartiles of prediction interval width

Maximum rooting depth (cm)	Predicted uncertainty (cm)	RMSE (cm)
60	< 5.81	1.67
	5.81 - 6.23	1.86
	6.23 - 6.67	1.95
	> 6.67	1.96
100	<9.45	2.64
	9.45 - 10.39	3.04
	10.39 - 11.23	3.38
	> 11.23	3.96
200	< 11.6	3.12
	11.6 - 13.1	3.77
	13.1 - 15	4.02
	> 15	5.51

IV.4.Spatial distribution and uncertainly of the SAWC

IV.4.1. Spatial distribution of the SAWC

According to the previously presented results (cf. Table 4.6), there was no clear difference in performance in the SAWC predictions of 60, 100 or 200 cm, so we chose to present the SAWC map for the maximum soil thickness of 200 cm.

The SAWC map (Figure 4.4) was mainly divided in two regions of contrasting soil thickness that corresponded to different lithologies and reliefs. The low predicted values of SAWC (shown in red in Figure 4.4) were predominant in the mountainous crystalline rocks of the Pyrenees and the Massif Central mountains, which were located in the south and northwest of the region, respectively, and on the hard limestone plateaus (the Causses). The high predicted values of the SAWC (in blue color in Figure 4.4) were located in the hills and plains of the soft marine and fluviatile sediments located near the seaside and in a narrow channel in the west of the region. However, more subtle differences in the predicted SAWC could be observed within the two regions. In the sedimentary area, a gradient was observed from high predicted SAWCs in the alluvium valleys, which had deep soil with low coarse fragment contents, to low predicted SAWCs in the stony soils of the old alluvial terraces (e.g., Nîmes Costières, which is in a red circle in Figure 4.4); the soils on tertiary sediment ("molasse") hillsides showed intermediate values. The mountainous crystalline rock areas and the Causses also showed identifiable differences in the predicted SAWCs (dark blue circle in Figure 4.4) that could be explained by the soil map and DEM derivative covariates (e.g., MRVBF, MRRTF and slope). Similar situations were encountered in the Causses (light blue circles in Figure 4.4).

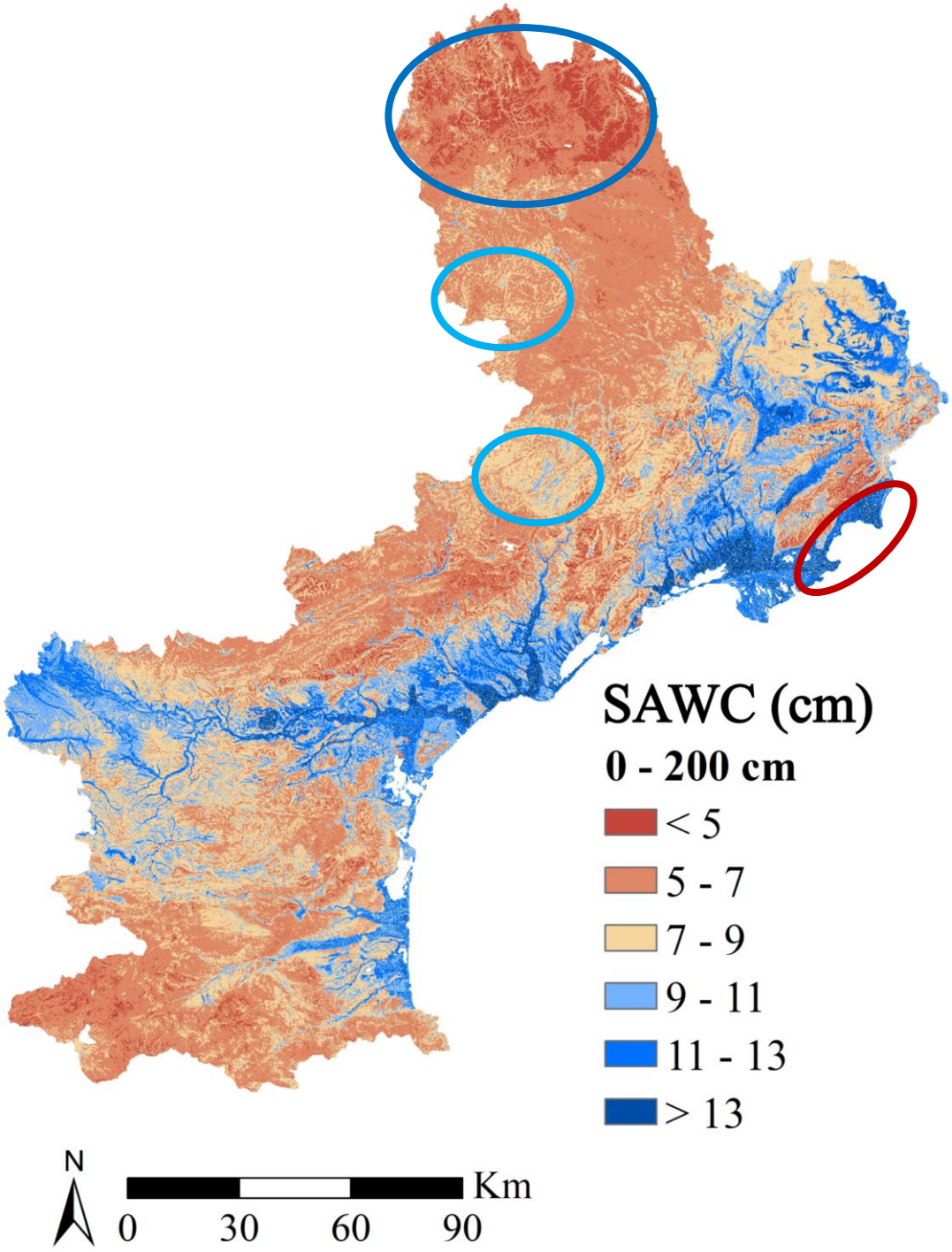


Figure 4.4. Predicted SAWC map of Languedoc Roussillon

IV.4.2. Spatial distribution of the SAWC prediction uncertainty

In Figure 4.5, we provide an uncertainty map estimated as the width of the 90% confident prediction interval represented according to the quartile classes. The high uncertainty predictions were mainly located in the alluvium valley in the littoral region. The moderated uncertainty predictions were located in a large portion of the alluvium valley and in the littoral region, while the low uncertainty predictions occupied the rest of the study area (i.e., the

plateaus and mountain regions). It is worth noting that the amount of predicted uncertainty seemed to be closely related to the values of the predicted SAWC, the largest uncertainty being related with the predictions of the largest SAWC values.

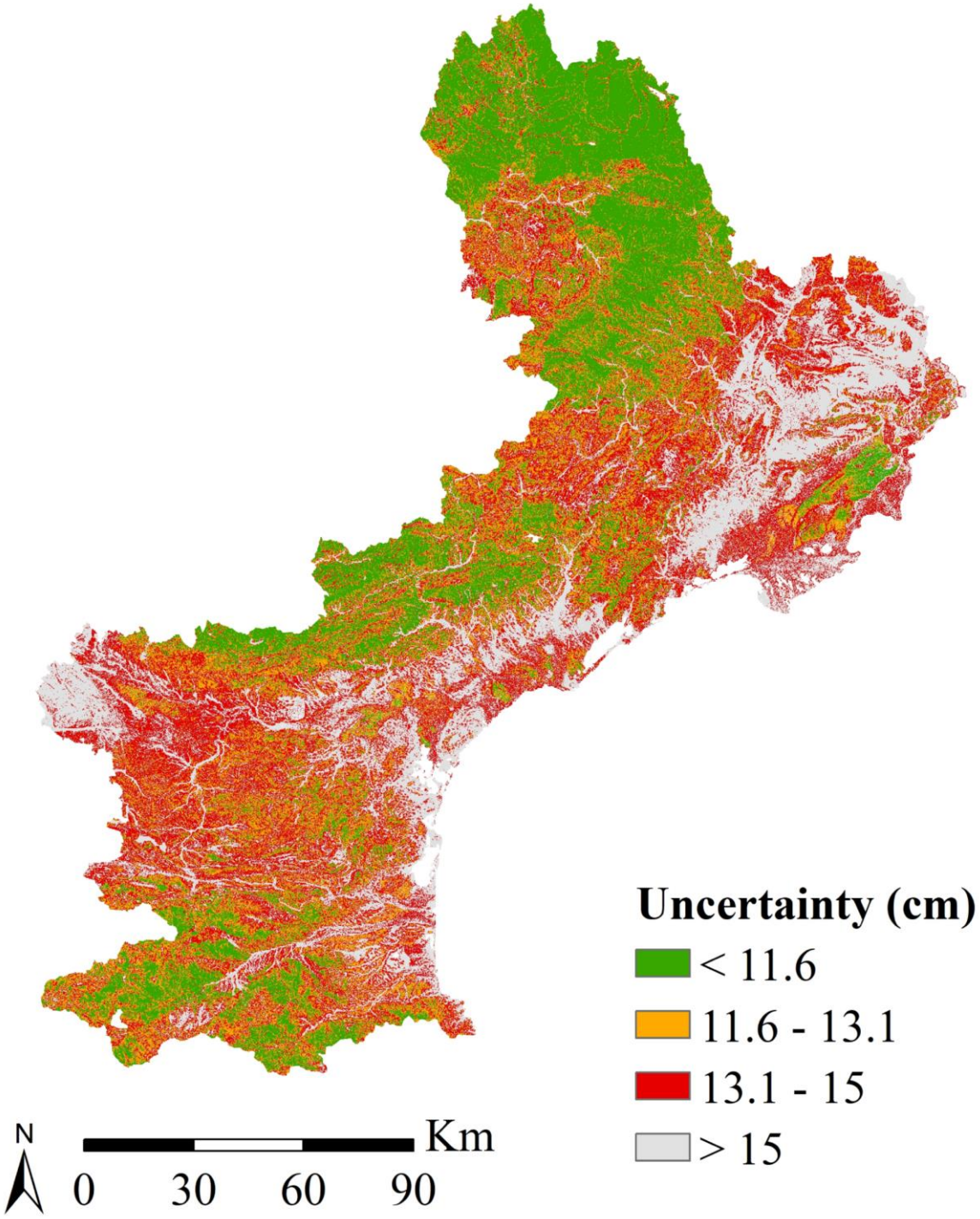


Figure 4.5. Uncertainty map presented by the classes estimated from the quartiles of the validation distribution

V. Discussion

V.1. Evaluation protocol

The evaluation protocol that was applied in this study consisted of a 10-fold cross validation approach with 20 repetitions, with the reference data being the soil profiles at punctual sites with the observed properties (textural fractions, coarse fragments and soil thickness) used to produce a local estimation of the SAWC using the pedotransfer functions. This protocol ensured both an evaluation of the SAWC predictions from independent samples and a comprehensive consideration of the mapping errors of all SAWC components. To our knowledge, this is the first time that these two conditions were fulfilled in an evaluation protocol dealing with the SAWC, which makes difficult the comparisons with previous papers dealing with SAWC (e.g., Poggio et al., 2010; Hong et al., 2013; Leenaars et al., 2018) difficult. However, our evaluation protocol had two main limitations. First, the evaluation sites were not characterized by the real SAWC measurements, and the fine earth water retention was estimated with the pedotransfer functions. This did not allow us to account for the PTF errors, as Román Dobarco et al. (2019) did in their study. However, these authors showed that PTF errors played a minor role in comparison with the mapping errors of the SAWC components. This should be even more true with the addition of the ST mapping errors in the evaluation protocol in comparison with the protocol of Román Dobarco et al (2019). Furthermore, the introduction of the PTF error in our error propagation models following the technique proposed by Román Dobarco et al. (2019) did not modify the ex-ante uncertainty evaluation of the SAWC paper (result not shown in this paper). However, a full evaluation of the SAWC mapping would be preferable, which means investing in costly field and laboratory measurements or finding accurate and inexpensive proxies for the SAWC (Coulouma et al, 2020).

A second limitation of our evaluation protocol was that SAWC mapping was evaluated at the site level, which was not representative of the decision-making units of the end-user and represents the worst case scenario for assessing the soil property prediction quality. Ideally, the evaluation should be performed for areal units (Bishop et al, 2015), which would produce more realistic results that would be in accordance with the visual evaluation of the map (see further). However, evaluating the SAWC from the areal units would require data collection that cannot be reasonably envisaged. Consistency checking involving data available at larger spatial support and closely related with AWC could be an alternative (Vanderlinden, 2005)

V.2. Error propagation model

Following Román Dobarco et al. (2019), an error propagation model using a first-order Taylor expansion was developed for the ex-ante estimation of SAWC mapping uncertainty. This model was, to some extent, more complete than the one developed by Román Dobarco et al. (2019) in that it considered the error in the soil thickness maps and the correlations between the error in the SAWC component maps that were not considered in Román Dobarco et al. (2019). This model was obtained by selecting an inference trajectory that included separate mapping of the soil layer thicknesses, which allowed easy application of the first-order Taylor expansion. The results revealed that the consideration of the error correlations had an impact on the final result if they reached a given level, which was the case for the error correlations between the soil layer maps (Table 4.4). The results showed that the ex-ante estimation of uncertainty was only slightly biased (differences with nominal values of 90% less than 1 for two out of three SAWC maps), which corresponded to much smaller uncertainty estimation biases than those obtained by Román Dobarco et al. (2019). We also verified that the RMSE obtained from the validation protocol was closely related to the predicted uncertainty (Table 4.8), which, to our knowledge, has not yet been verified. We note, however, that the error propagation model built in this study did not consider the PTF errors. This error could be easily added by following the procedure proposed by Román Dobarco et al. (2019).

V.3. General performances of the SAWC predictions

Although the SAWC map of the Languedoc-Roussillon region exhibited expected and pedologically sound soil patterns, poor results were obtained from the evaluation protocol (Table 4.6). This was likely related to the difficulty of mapping the two most critical components of the SAWC, namely, the coarse fragments and soil layer thicknesses (Table 4.5). As observed by Vaysse and Lagacherie (2015), the soil thickness and coarse fragments were characterized by a pre-eminence of short-scale variations that could not be captured by a DSM model using a so sparse soil spatial sampling according to the soil layer depths (Table 4.1). A denser spatial sampling is, therefore, necessary in this situation. Furthermore, the coarse fragment data were obtained from visual observations of the soil profile, which carry a greater uncertainty than the ones of other soil properties that are measured in a laboratory. More accurate field protocols for measuring the proportion of coarse fragments (Algayer et al, 2019) are required to improve this situation.

It is also important to notice that biases were important components of the SAWC prediction error (between 15 and 30% on Table 4.6), which generated an overall SAWC overestimation. This overestimation can be related with the important positive biases observed for the predictions of the thicknesses of the three first layers (Table 4.5). Such biases should be caused by the difficulties of the random forest algorithm to deal with the important subset of location having null soil layer thicknesses. Dealing with zero inflated input datasets of regression models in a well-known problem in ecology (Martin et al., 2005). Specific regression approach adapted to zero inflated datasets (Savage et al., 2015) should be applied to mitigate this problem.

Conversely, the last soil layer exhibited a negative bias (Table 4.5) that can be related with the unbalanced distribution of soil thicknesses in the set of sampling locations (Figure 4.3). Indeed, a very small proportion of deep soils were sampled because of the low maximum observation depth (only 15% of the soil profiles had maximum observation depths greater than 120 cm). This too small proportion generated the underestimation of the deepest soil layers thicknesses since, the random forest algorithm is known to behave like an interpolator and to smooth the outliers (Song et al., 2015). A more balanced sampling of the ST across the study region is therefore necessary.

VI. Conclusions

- We developed a DSM model that mapped the SAWC values and provided an ex ante local estimation of the prediction uncertainty. For the first time, this uncertainty model took into account all SAWC component mapping errors for all soil layers.
- The results showed weak performances of the SAWC predictions, although the final map exhibited pedologically sound spatial patterns of predicted SAWC. This paradoxical result could be caused by the inadequate spatial support at which the evaluations were conducted (punctual one).
- Improvement of the current model would require the collection of more data, such as either legacy data or new measurements or proxies of SAWC.

Les éléments importants de ce chapitre à retenir sont :

- Un modèle de cartographie numérique des sols a été développé, permettant de fournir **une spatialisation du réservoir utile** mais également une **quantification spatialisée de l'incertitude de prédiction**.
- Pour la première fois, les erreurs de spatialisation de **l'ensemble des composantes du RU** (réservoir utile élémentaire et épaisseur de sol) sur **toutes les couches de sols** ont été considérées dans **un modèle de propagation d'erreur** pour l'estimation locale de l'incertitude.
- Les corrélations des erreurs liées à **l'agrégation des couches de sol** ont une **plus grande importance** dans la **propagation d'erreur** que les corrélations des combinaisons de **propriétés constituant le réservoir utile**, cela pour **estimer correctement l'incertitude**.
- Le taux d'inclusion des sites de validation dans les intervalles de confiances prédits est proche de la valeur nominale (90%) et une bonne correspondance entre les incertitude prédites (largeurs des intervalles de confiance) et les erreurs mesurées par validation montre que le modèle de propagation d'erreur estime sans biais et spatialise correctement les erreurs.
- Bien que les performances de prédictions du réservoir utile soient faibles, la distribution spatiale des valeurs prédites semble conforme à **une connaissance pédologique** (ex : réservoir utile faible en montagne (sols peu développés) et important sur la plaine littorale), soulevant **les limites de validation par sites**. Une alternative serait de valider sur des unités territoriales mais cela implique l'acquisition de nouvelles données (non envisageable).
- La carte de l'incertitude locale fournie dans cette étude pourra être utilisée pour des prises de décisions à l'échelle régionale.

CHAPITRE 5 Évaluation de l'intégration des données pédologiques anciennes denses et hétérogènes dans les modèles de cartographie numérique des sols appliquée au réservoir utile

Il a été reconnu que la densité d'observations de sol est le principal facteur limitant les performances des modèles de cartographie numérique des sols. Les travaux qui vont suivre visent à utiliser les données pédologiques anciennes de BRL inclues dans la commune de Bouillargues afin d'étudier la sensibilité des performances des modèles de cartographie numérique des sols à la forte densité de ces données. Cette sensibilité sera analysée selon deux axes : i) un ajout progressif d'observations de sols permettant d'augmenter la densité ainsi que ii) l'utilisation d'un modèle de forêt de régressions quantiles (Quantile Regression Forest, QRF) spatiale (QRF_{dist}) permettant de prendre en compte la structuration spatiale des données comparé à l'utilisation standard du QRF qui ne la considère pas. Une étude coût-bénéfice est également réalisée en fonction de la densité afin d'estimer le coût du gain de performances de la prédiction du RU par rapport à l'augmentation de la densité.

La méthodologie utilisée ci-après prend en compte les conclusions des chapitres 3 et 4, à savoir, le choix de la trajectoire de calcul pour spatialiser le RU et l'utilisation d'un modèle de propagation d'erreurs pour quantifier et spatialiser l'incertitude de prédiction du RU.

Ce chapitre se présente sous la forme d'un article soumis à Geoderma Regional.

How far harvested legacy soil data can improve the Digital Soil Mapping of Available Water Capacity over a Mediterranean irrigation perimeter?

Styc Quentin, François Gontard, and Lagacherie Philippe

Submitted to Geoderma Regional

Abstract

The density of the soil observations that are used as the input of digital soil mapping (DSM) models has been recognized as a strong limiting factor. Although considerable work has been conducted in recent decades to build global and national soil databases, the legacy data from some former soil survey campaigns still remain unused. The objective of this study was to determine the interest in harvesting legacy data for mapping the soil available water capacities (SAWCs) at different rooting depths (30 cm, 60 cm, 100 cm) and to the maximal observation depth, over the commune of Bouillargues (16 km^2, Occitanie region, southern France). An increasing number of available auger hole observations with SAWC estimations – from 0 to 2781 – were added to the existing soil profiles to calibrate quantile regression forests (QRFs) using the geographical locations of the sites as soil covariates. The SAWC was first mapped separately for different soil layers, and the mapping outputs were pooled to estimate the required SAWC. The uncertainty of the SAWC prediction was estimated from the estimated mapping uncertainties of the individual soil layers by an error propagation model using a first-order Taylor analysis. The performances of the SAWC predictions and their uncertainties were evaluated with a 10-fold cross validation that was iterated 20 times. The results showed that the use of a quantile random forest that was fed with auger hole observations and that used the geographical locations as soil covariates considerably augmented the performances of the SAWC predictions (percentages of explained variance from 0.39 to 0.70) compared to the performance of a classical DSM approach, i.e., a QRF that solely used soil profiles and only soil landscape covariates (percentages of explained variance from 0.04 to 0.51). The analysis of the results revealed that the performances were also dependent on the spatial patterns of the

different examined SAWCs and was limited by the observational uncertainties of the SAWCs determined from auger holes. The best performance tended to also provide the best view of the uncertainty patterns with an overestimation of uncertainty. Despite these gains in performance, the cost-efficiency analysis showed that the augmentation of soil observations was not cost efficient because of the highly time-consuming manual data harvesting protocol. However, this result did not account for the observed gain in map details. Furthermore, the cost efficiency could be further improved by automation.

I.Introduction

Digital soil mapping (DSM) has been recognized as the appropriate solution to provide spatial soil information for land users, scientist communities and policy and decision makers in agriculture and the environment (McBratney et al., 2003; Sanchez et al., 2009). The principle of DSM is to predict a soil property and its associated uncertainty by determining the quantitative relationships between the punctual soil information available and the spatial data reflecting the state factors of soil formation (e.g., soil covariates). DSM has now moved from a largely academic movement toward an operational activity (Minasny & McBratney, 2016, Arrouays et al, 2017).

However, the performances of DSM predictions of soil properties often exhibit more uncertainty than initially expected. For example, the percentages of explained variances of less than 0.5 were observed for 95%, 76%, 100% and 86% of the tested soil properties for DSM applications at the catchment scale (Nussbaum et al., 2018), at the regional scale (Vaysse and Lagacherie, 2015), at the national scale (Mulder et al., 2016), and at the global scale (Hengl et al., 2014), respectively.

These authors converged toward the conclusion that the density of soil observations used for calibrating the DSM models was the main factor that limited the DSM performances. Most of the soil information used as input in DSM applications has been either soil maps or the spatial sampling of sites with soil property measurements. The average densities used in most operational DSM applications have been low, e.g., 4-12 sites/km² (several study areas in Nussbaum et al., 2018), 0.07 sites/km² (Vaysse and Lagacherie, 2015), 0.03 sites/km² (Mulder et al., 2016), and 0.001 sites/km² (Hengl et al., 2014), which limits the performances of soil prediction, especially when the pattern of variation in the soil property is largely below the spacing of soil profiles (Vaysse and Lagacherie, 2015; Gomez and Coulouma, 2018). In addition, further experiments that consisted of varying the spatial density of soil input

confirmed this analysis (Somarathna et al. 2017, Wadoux et al. 2019, Lagacherie et al, 2020). Consequently, it is of paramount importance to increase the density of soil inputs to improve the performance of DSM models in predicting soil properties (Voltz et al., 2020).

The most straightforward way to increase the density of DSM model soil inputs involves harvesting the legacy soil data that have not yet been stored in the existing soil databases. Arrouays et al. (2017) showed that during the period 2009-2015, the numbers of legacy soil profiles stored in global and national soil databases increased by 1,046% and 45%, respectively. However, they estimated that a large amount of soil legacy data can still be harvested. This is even more true in some areas across the world where soil surveying has been particularly active in the past. For example, in southern France, the BRL irrigation company conducted detailed soil surveys over its irrigation perimeter between 1957 and 1992, which resulted in detailed soil maps, 25,000 soil profiles (5/km²) and 203,000 auger hole observations (31/km²). At this stage, such soil data have not yet been harvested and therefore cannot be used as input for DSM applications. However, this data has great potential for improving DSM performance and should be thoroughly examined.

In this paper, we present a cost-benefit analysis of soil data harvesting for improving the performances of DSM models in mapping soil available water capacities for different rooting depths (0-30 cm, 0-60 cm, 0-100 cm) and at maximum observation depth, and the associated uncertainties. The study is conducted in the commune of Bouillargues, which is one of the communes included in the BRL irrigation perimeter. The cost of harvesting the available soil profiles and auger hole data has been evaluated and compared with the increase in the performance of a DSM model that addressed the high spatial density of the harvested data.

II. The case study

II.1. The study area

This study took place in the administrative commune of Bouillargues in the Occitanie administrative French region (Figure 5.1). Located in the Southern France, Bouillargues covers 16 km² and is mainly devoted to vineyards, agricultural lands, forests, and scrublands.

The study area is topographically split in two sub-regions with the large flat valley of the Vistrenque in the northern part and old fluviatile alluvium terraces belonging to the Nîmes "Costière" in the southern part. The two sub-regions have contrasted parent materials with i) loess and loamy clay deposition in the Vistrenque valley and ii) old alluvium in Nîmes Costière

part, covered by some loess deposits. The contrast of parent materials induces variations of soils with i) fluviosol and calcisol developed in loess and loamy clay deposition, characterized by an absence of coarse fragment and a loamy texture and ii) chromic luvisols developed in old alluvium terraces characterized by important stone contents and compacted clay accumulations (Figure 5.1).

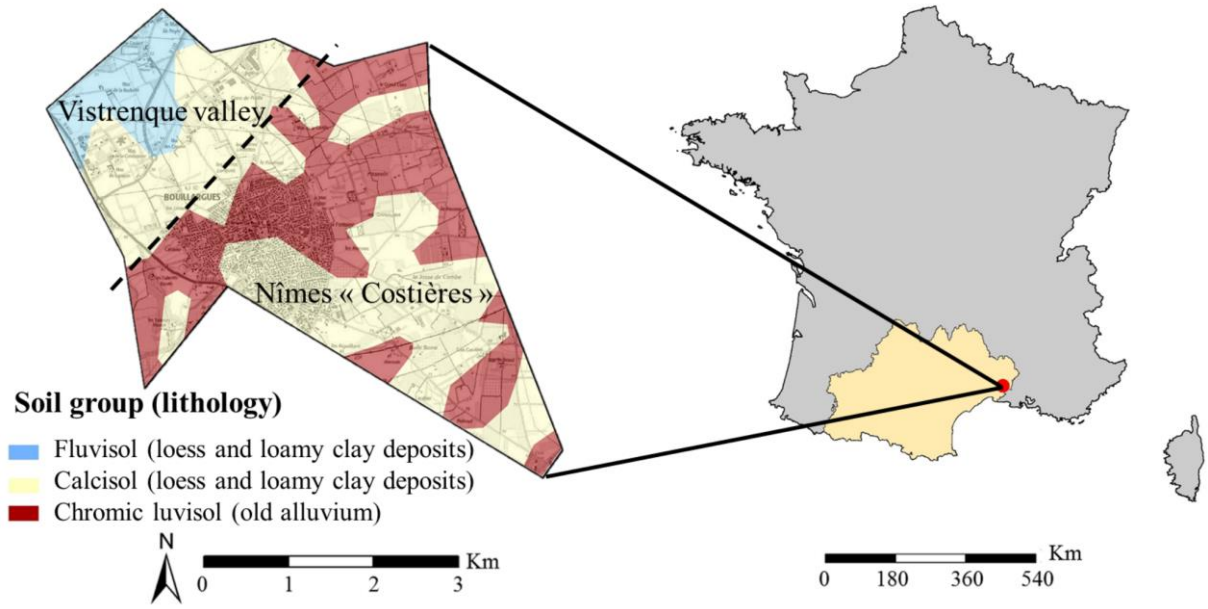


Figure 5.1. Location of the study case

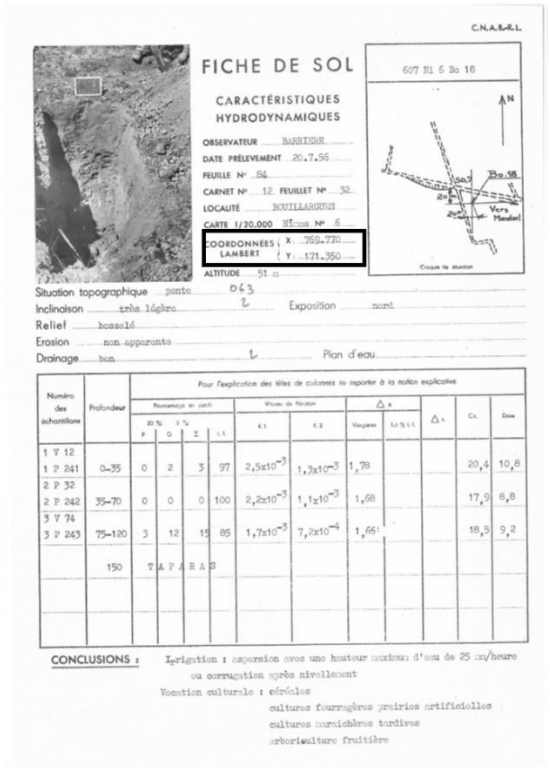
II.2. Soil data

II.2.1. History and content of BRL soil database

The soil data of this study is a part of the soil survey led by the "Compagnie Nationale d'Aménagement de la Région du Bas-Rhône et du Languedoc" (CNARBRL) between 1957 and 1992 over the irrigated perimeter of this irrigation company, which covers 6,636 km². The objectives of this survey were to provide the suitable soil information for: i) improving the development master plan of the irrigation perimeter and estimating the surface area of arable and potentially irrigable lands and for ii) supporting the cultural intensification made possible by irrigation, assessing the irrigation supply, and setting a technical assistance for landholders to start irrigation and crop conversion.

The compilation of those studies resulted in a database of 228,000 soil observations with 25,000 soil profiles description and laboratory analysis (Figure 5.2) and 203,000 auger holes (Figure 5.3), which correspond to an average spacing of 515 m and 181 m for soil profiles and auger holes respectively.

a)



FICHE DE SOL

CARACTÉRISTIQUES HYDRODYNAMIQUES

OBSERVATEUR BARTHES

DATE PRÉLÈVEMENT 20.7.56

FEUILLE N° 64

CARNET N° 12 FEUILLET N° 32

LOCALITÉ BOUILLARGUES

CARTE 1/20.000 Nîmes N° 6

COORDONNÉES LAMBERT X : 769.770 Y : 171.350

ALTITUDE 51 m

607 N1 6 Bo 18

Situation topographique ponte 063

Inclinaison très légère 2 Exposition nord

Relief bosselé

Erosion non apparente

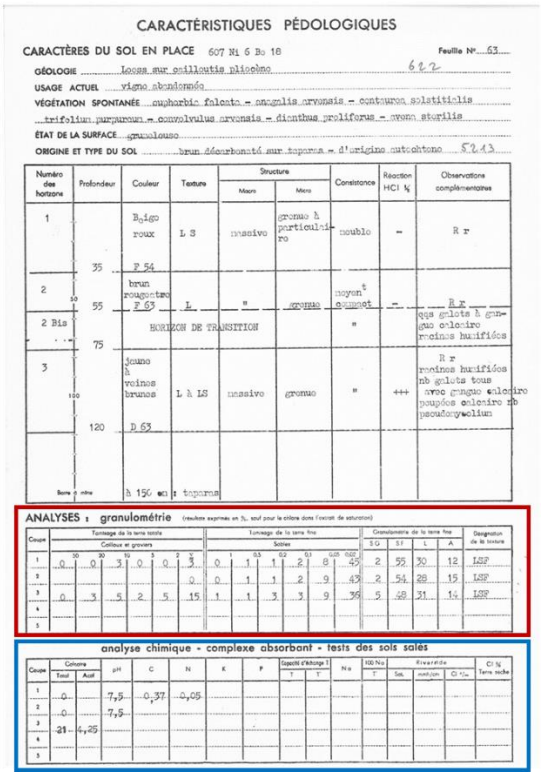
Drainage bon 2 Plan d'eau

Numéro des échantillons	Profondeur	Pourcentage en poids 2 mm %				Valeurs de filtration K1	Valeurs de filtration K2	Δ e Valeurs	Δ s	Cc	Dsw
		P	G	Σ							
1 Y 12											
1 P 241	0-35	0	2	3	97	$2,5x10^{-3}$	$1,3x10^{-3}$	1,78		20,4	10,8
2 P 32											
2 P 242	35-70	0	0	0	100	$2,2x10^{-3}$	$1,1x10^{-3}$	1,66		17,9	6,8
3 Y 74											
3 P 243	75-120	3	12	15	85	$1,7x10^{-3}$	$7,2x10^{-4}$	1,66		18,5	9,2
	150	T A P A R A S									

CONCLUSIONS : Irrigation : aspersion avec une hauteur maximum d'eau de 25 mm/heure ou corrugation après nivellement

Vocation culturale : céréales, cultures fourragères prairies artificielles, cultures maraîchères tardives, arboriculture fruitière

b)



CARACTÉRISTIQUES PÉDOLOGIQUES

CARACTÈRES DU SOL EN PLACE 607 N1 6 Bo 18 Feuille N° 63

GÉOLOGIE Lœss sur cailloutis pliocène 622

USAGE ACTUEL vigne abandonnée

VÉGÉTATION SPONTANÉE euphorbia falcata - onopalis arvensis - centaurea solstitialis - trifolium purpureum - convolvulus arvensis - dianthus prolifer us - avena sterilis

ÉTAT DE LA SURFACE granuleuse

ORIGINE ET TYPE DU SOL brun décarbonaté sur taparas - d'origine autochtone 5243

Numéro des horizons	Profondeur	Couleur	Texture	Structure Macro	Structure Micro	Consistance	Réaction HCl ½	Observations complémentaires
1	35	B_2 4go roux P 54	L S	massive	grenue à particulaire	meuble	-	R r
2	55	brun rougeâtre P 63	L	"	grenue	moyennt compact	-	R r
2 Bis	75	HORIZON DE TRANSITION				"		qqs galets à gangue calcaire racines humifères
3	120	jaune à veines brunes P 63	L à LS	massive	grenue	"	+++	R r racines humifères nb galets tous avec gangue calcaire poupées calcaire nb pseudomycelium
	Barre à mine	à 150 cm : taparas						

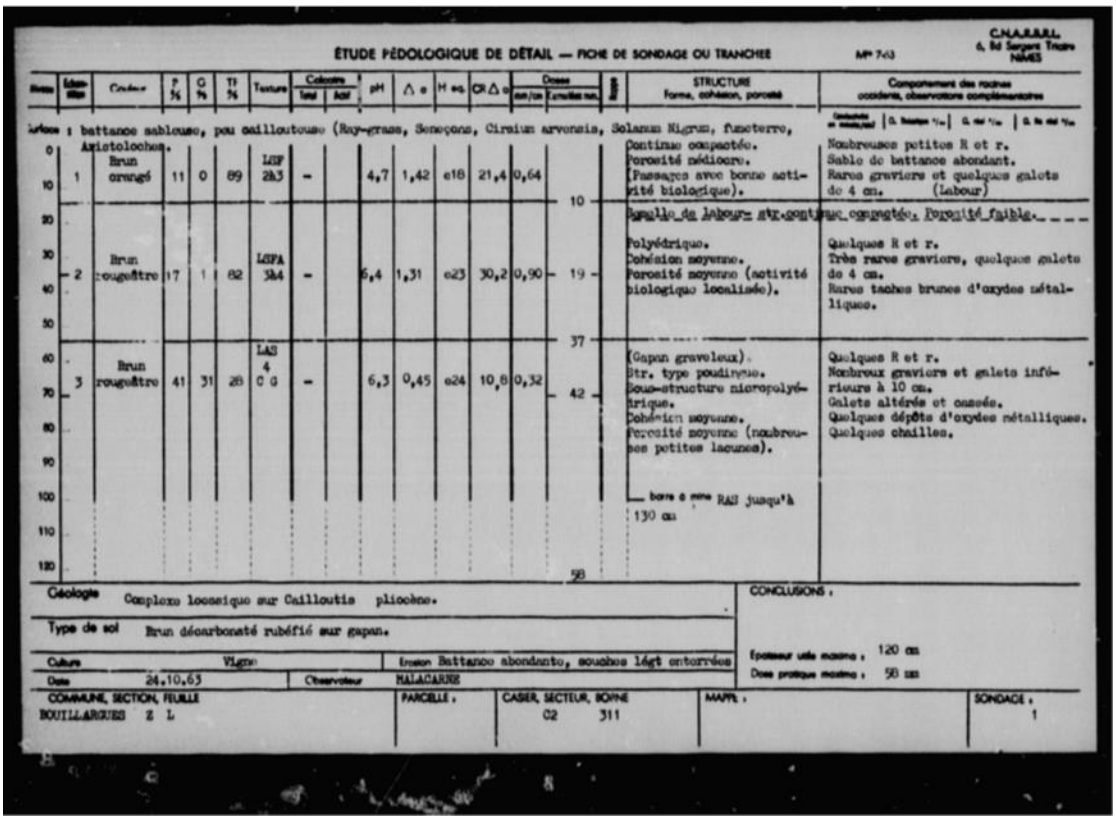
ANALYSES : granulométrie (résultats exprimés en %, sauf pour le calcaire dans l'extrait de saturation)

Coupe	Tamisage de la terre totale Cailloux et graviers							Tamisage de la terre fine Sables							Granulométrie de la terre fine SG	SF	L	A	Désignation de la texture
	50	20	10	5	2	1	Σ	1	0,5	0,2	0,1	0,05	0,02						
1	0	0	3	0	0	3	3	0	1	1	2	8	45		2	55	30	12	LSF
2							0	0	1	1	2	9	43		2	54	28	15	LSF
3	0	3	5	2	5	15		1	1	3	3	9	36		5	48	31	14	LSF
4																			
5																			

analyse chimique - complexe absorbant - tests des sols salés

Coupe	Calcaire Total	Calcaire Actif	pH	C	N	K	P	Capacité d'échange T	Capacité d'échange T	Na	Rapport 100 Na/T	Rivacidité Sat.	Rivacidité matière	Cl ‰	Cl ‰ Terre sèche
1	0		7,5	0,37	0,05										
2	0		7,9												
3	21		6,25												
4															
5															

Figure 5.2. Soil profile) a) horizon descriptions with geographical coordinates (black box) and b) laboratory analysis results, physical analysis (red box) and chemical analysis (blue box)



ÉTUDE PÉDOLOGIQUE DE DÉTAIL — FICHE DE SONDAGE OU TRANCHEE N° 743

C.N.A.R.R.R.L. 4, Bd Sergent Triaire NIMES

Horizon	Echantillon	Couleur	P %	G %	T %	Texture	Calcaire Total	Calcaire Actif	pH	Δ e	H eq.	CE Δ e	Dosse dm/cm	Dosse Cailloux sup.	Racines	STRUCTURE Forme, cohésion, porosité	Comportement des racines accidents, observations complémentaires
0		Aristoloches.														Horizon : battance sableuse, peu caillouteuse (Ray-grass, Senecons, Cirsium arvensis, Solanum Nigrum, fumeterre,	
10	1	Brun orangé	11	0	89	LSF 2b3	-		4,7	1,42	e18	21,4	0,64	10		Continu compacté. Porosité médiocre. (Passages avec borne activité biologique).	Nombreuses petites R et r. Sable de battance abondant. Rares graviers et quelques galets de 4 cm. (Labour)
20																Semelle de labour str.continue compactée. Porosité faible.	
30	2	Brun rougeâtre	17	1	82	LSFA 3b4	-		6,4	1,31	e23	30,2	0,90	19		Polyédrique. Cohésion moyenne. Porosité moyenne (activité biologique localisée).	Quelques R et r. Très rares graviers, quelques galets de 4 cm. Rares taches brunes d'oxydes métalliques.
40																	
50														37			
60						LAS 4										(Gapon graveleux). Str. type poudreuse. Sous-structure micropolyédrique. Cohésion moyenne. Porosité moyenne (nombreuses petites lacunes).	Quelques R et r. Nombreux graviers et galets inférieure à 10 cm. Galets altérés et cassés. Quelques dépôts d'oxydes métalliques. Quelques chailles.
70	3	Brun rougeâtre	41	31	28	C G	-		6,3	0,45	e24	10,8	0,32	42			
80																	
90																	
100																barre à mine RAS jusqu'à 130 cm	
110																	
120														58			

Géologie Complexe loessique sur Cailloutis pliocène.

Type de sol Brun décarbonaté rubéfié sur gapon.

Culture Vigne — Horizon Battance abondante, souches légt enterrées

Date 24.10.63 — Observateur MALACARNE

CONCLUSIONS :

Epaisseur utile maxima : 120 cm

Dose pratique maxima : 50 mm

COMMUNE, SECTION, FEUILLE BOUILLARGUES Z L — PARCELLE : — CASIER, SECTEUR, BOYNE C2 311 — MAPPE : — SONDAGE : 1

Figure 5.3. Auger hole descriptions

II.2.2. Spatial sampling and georeferencing in the study area

Focusing on the commune of Bouillargues, the harvested dataset is composed of 2850 punctual sites that include 2781 auger holes and 69 soil profiles, which correspond to average spacings of 76 m and 500 m, respectively (Figure 5.4). Both the soil profiles and the auger hole observations were fairly evenly distributed over the study area, however, some gaps corresponded to urbanized areas or lands that were not expected to have any agricultural potential.

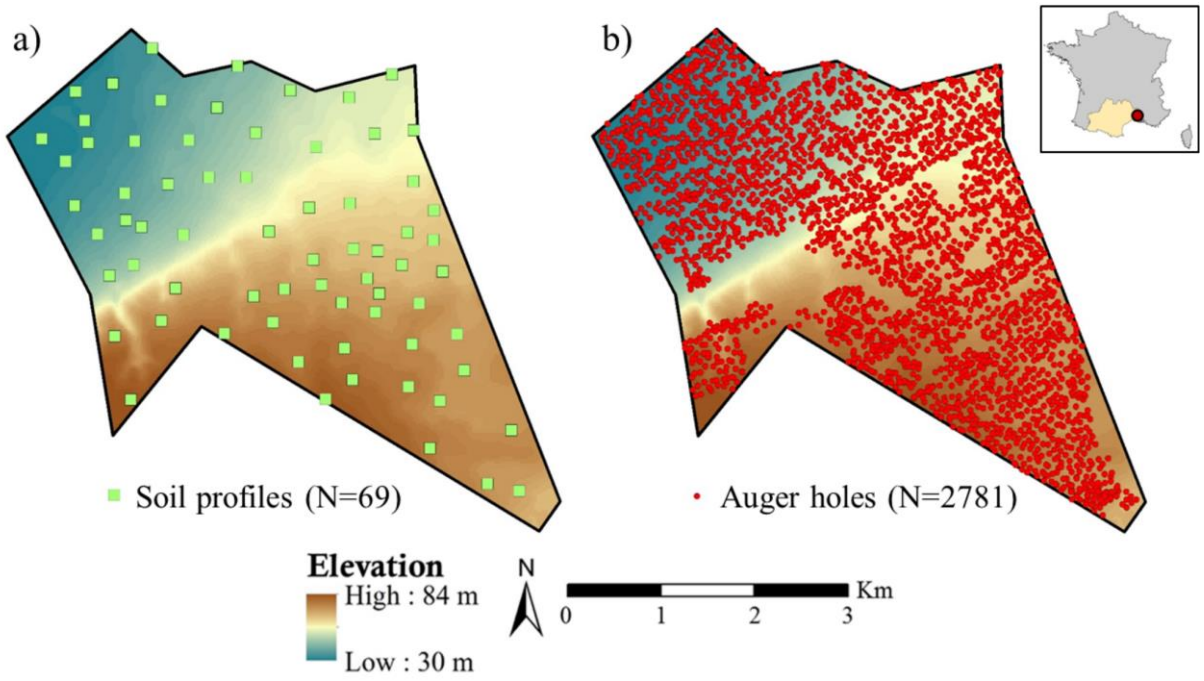


Figure 5.4. Spatial distribution of a) soil profiles and b) auger holes over the commune of Bouillargues

The soil profiles data records included geographical coordinates (Lambert III, black box in Figure 5.2.a), whereas a manual preprocessing was necessary for georeferencing the auger holes.

The auger holes were initially located through a non-georeferenced map representing the local sampling scheme (Figure 5.5.a). Each sampling scheme corresponded to an area of water distribution supplied by an irrigation water access point of the BRL irrigation network. This access point was georeferenced and could be positioned onto a georeferenced former cadastre (red box on Figure 5.5.b). To acquire the coordinates of the auger holes, the sampling scheme was first located in the georeferenced cadastre using the coordinates of the irrigation water

access point. Its boundaries were then positioned (green dashed perimeter on Figure 5.5.b) using the geometry of the parcels and communication paths. Finally, each auger hole was manually positioned onto the georeferenced cadastre (blue stars on Figure 5.5.c) using the sampling scheme (Figure 5.5.a) and the coordinates of the auger holes were obtained using the coordinates acquisition tool of the BRL's web-GIS (Figure 5.5.c).

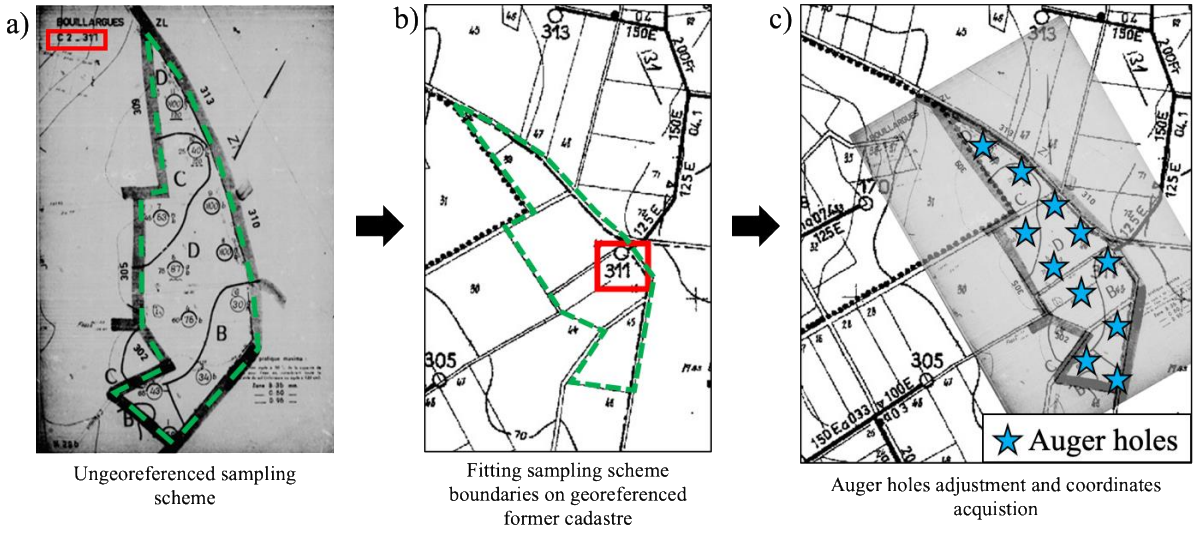


Figure 5.5. Fitting the ungeoreferenced sampling scheme of auger holes in the georeferenced former cadastre

II.3. Soil Available Water Capacity determinations at punctual sites

This study took the mapping of soil available water capacity (SAWC) as an example of applying DSM. SAWC refers to the capacity of the soil to store water for plant growth (Veihmayer and Hendrickson, 1927). This functional property plays a key role in many ecosystem services, such as food production, soil drought or climate and gas regulation. Consequently, it is a crucial parameter used in land evaluations and recently in ecosystem services assessments (Dominati et al., 2014). Information about the SAWC distribution in space is essential for planning and management in agriculture and for ecological modeling. In the present example, SAWC was required for fulfilling the irrigation objectives evoked above (section II.2.1). Currently, SAWC is computed in the literature as follows (Cousin et al., 2003):

$$SAWC = \sum_{i=1}^{n} dh_i * bd_i * \left(\frac{100 - st_i}{100}\right) * (\theta r_i - \theta w_i) \qquad \text{(Eq. 5.1.)}$$

where SAWC is the soil available water capacity (cm), dh_i = the thickness of the ith horizon (cm), bd_i = the bulk density of the ith horizon, st_i = the coarse fragment content of the ith horizon (% volumetric), and θr_i and θw_i are the gravimetric soil water contents at field capacity

(i.e., the soil water content that remains in the soil after water has drained due to gravitational force) and at permanent wilting point (i.e., the soil water retained so strongly that it is no longer available for plant roots, so plants wither and cannot recover their turgidity) of the ith horizon ($g.g^{-1}$), respectively.

Historically, the CNARBRL had a different approach for expressing the water retention term of the fine earth, i.e. $(\theta r_i - \theta w_i)$, which leads to the following equation:

$$SAWC = \sum_{i=1}^{n} dh_i * bd_i * \left(\frac{100 - st_i}{100}\right) * (b_i * EqW_i) \qquad \text{(Eq. 5.2.)}$$

The equivalent water content (EqW_i) corresponds to θr_i of Equation 5.1 and the textural coefficient b_i is an expression of the water content at permanent wilting point that weights EqW_i to account for the water content that is not available for plant (i.e. beyond the wilting point defined as θw_i in Equation 5.1).

The values of bd_i and EqW_i were measured at each soil profiles. bd_i was determined in the field following the Vergières protocol (Bourrier, 1965) but was estimated as 1.6 times of the mass fraction of fine earth from the ensemble coarse fragment and fine earth, when the coarse fragment phase of the soil sample was too important to perform the Vergières protocol (Legros, 1996).

The EqW_i of sieved samples was determined in the laboratory using a centrifuge apparatus set at 100 kPa (pF = 3.0), a reference pressure that was considered, at the time of the CNARBRL soil survey, as yielding the best approximation of the water content at the field capacity (see section II.2.1) (Baize and Jabiol, 1995). EqW_i were estimated on auger hole observations by local pedotransfer functions using the field estimated textural classes.

The b_i coefficient was determined both on soil profiles and on auger hole observations by a local pedotransfer function using the textural classes determined from granulometric analyses and field estimation, respectively for soil profile and auger hole observations.

The stone content and the horizon thicknesses of Equation. 5.2 were retrieved from the descriptions physical analyses and descriptions of the soil profiles and of the auger hole observations, respectively (Figure 5.2 and Figure 5.3). Different total soil thicknesses (i.e. $\sum_{i=1}^{n} dh_i$) were considered to determine the different rooting depths related with the different possible crops of the study area (from market gardening to vineyard passing by annual crops). In addition to the maximum soil thicknesses given by the soil observations that were considered

for calculating the maximum Soil Available Water Capacity ($SAWC_{max}$), restricted thicknesses of 30 cm, 60 cm and 100 cm were then considered, leading to different restricted SAWC denoted further $SAWC_{30}$, $SAWC_{60}$, $SAWC_{100}$.

It must be noted that both the profiles and auger holes had limited observation depths of 140 and 120 meters, respectively, which may cause underestimations of $SAWC_{max}$.

II.4. Environmental covariates

The DSM process used, lied on the scorpan model (McBratney et al., 2003), quantitative relationships between target soil property and environmental variables, which were also known as "covariates".

The selection of soil covariates depends on two criteria: i) they could be derived from geodatasets freely available at least at the French national level, and ii) they have a logical and process-based relationship with soil properties according to the literature. Following these criteria, we derived covariates related to the scorpan model component, i.e., topography, organisms, and parent material, that regroups the major landscape types across the study area. Climate data were not considered in this study since we did not find any climate data at a spatial resolution fine enough to represent the climate variations over such a small area. The relief component was described by a set of geomorphometric indicators currently considered in DSM studies: elevation, slope, aspect, multiresolution valley bottom flatness (MRVBF), multiresolution ridge top flatness (MRRTF), topographic wetness index (TWI), topographic position index, plan curvature and profile curvature. These indicators were derived from the French altimetry database (BD ALTI, 25 m resolution) digital elevation model (DEM). Organisms and parent materials were derived from the Landsat 7 imagery and geological map, respectively, and were both resampled at the native resolution of the DEM (i.e., 25 m). Additionally, parent material covariates were developed by Vaysse and Lagacherie (2015) from the geological map (1:50,000) qualitative descriptions to quantitative indicators describing the hardness, mineralogy and texture of alteration materials (Table 5.1).

Table 5.1. Exhaustive categorical and continuous covariates

Variables	Abbreviation	Resolution/Scale	Source	Soil forming factor[1]	Type[2]
Topography					
Elevation	ELEV	25 m	BD ALTI	r	Q
Multi-resolution Valley Bottom Flatness	MRVBF	25 m	BD ALTI	r	Q
Slope	SLOPE	25 m	BD ALTI	r	Q
Topographic Wetness Index	TWI	25 m	BD ALTI	r	Q
Plan Curvature	PLANCURV	25 m	BD ALTI	r	Q
Profile Curvature	PROCURV	25 m	BD ALTI	r	Q
Multi-resolution Ridge Top Flatness	MRRTF	25 m	BD ALTI	r	Q
Topographic Position Index	TPI	25 m	BD ALTI	r	Q
Geology					
Hardness	HARDNESS	25 m	Geological map/soil profil	p	C
Texture	TEXTURE	25 m	Geological map/soil profil	p	C
Mineralogy	MINERALOGY	25 m	Geological map/soil profil	p	C
Organisms					
Land use	LANDUSE	25 m	Landsat 7	o	C

[1]: SCORPAN factors (s=soil property, c = climate, o = organisms, r = relief, p=parent material)

[2]: Q = quantitative, C = categorical

II.5. Acquisition process and cost assessment

In section II.2, we presented the main difference in using soil profiles and auger holes in a DSM application, i.e., the accessibility of the data. While soil profiles acquisition is quite straightforward, i.e., recording soil data and locations, auger hole acquisition is more complicated as the locations are not directly available and a manual georeferencing is required,

the acquisition process is longer. In Table 5.2, we provided the main information about the acquisition process for soil profiles and auger holes. As the number of auger hole observation is substantially larger than the number of soil profiles and take longer to record, we provided, in addition, an assessment of the cost of soil data acquisition.

Table 5.2. Information to assess the cost of the acquisition process

	Auger holes	Soil profiles
Record time of soil properties (min/observation)	0.8	0.8
Recording time of locations (min/observation)	0.5	0.2
Number of observations	2781	69

To compute the cost of the acquisition process, we applied the following formula using information in Table 5.2:

$$Cost = \left(\frac{Number\ of\ observation * recording\ time}{Daytime}\right) * Salary \qquad \text{(Eq.5.3)}$$

III. Methods

III.1. Digital Soil Mapping models for soil profiles

In this study, we used several mapping models derived from the Random Forest algorithm. Hereafter, we provide the general description of Random Forest and its derivatives used in this study.

III.1.1. Random Forest

Random forest models (Breiman, 2001) are an ensemble learning method for both classification and regression. A forest, i.e., an ensemble of randomized decision trees, is built and trained based on a bootstrap approach. Individual trees are built using the principle of recursive partitioning. "*The feature space is recursively split into regions containing observations with similar response value*" (Strobl et al., 2009). The predictions of the individual trees are finally averaged to obtain a single prediction.

III.1.2. Quantile regression forest

The quantile regression forest algorithm (Meinshausen, 2006) is an extension of random forests that has become one of the most commonly used algorithms in DSM studies (Hengl et al., 2015; Ugbaje and Reuter, 2013; Vaysse and Lagacherie, 2017). As RF, QRF provides an ensemble prediction based on *n* regression trees. However, while RF provides solely the conditional mean, QRF supplies the whole conditional distribution of the target variable by keeping all observations at the terminal nodes. This allows to infer estimates for the conditional quantiles (Meinshausen, 2006). More details on QRF can be found in Meinshausen (2006).

QRF was performed with the *ranger* package, which is a fast implementation of Breiman's random forest and Meinshausen's quantile regression forest for big data (Wright and Ziegler, 2017).

III.2. Mapping models for dense spatial sampling

The usual applications of RF and its derivative to DSM only exploit the relationships between the soil properties to be predicted with landscape elements characterized by a set of covariates derived from the available spatial data. However, they do not consider the spatial relationships between sites, or spatial autocorrelation, which allows the spatial interpolations of a given soil property between sites. This can lead to suboptimal predictions and possibly systematic over- and underestimation of predictions, especially if the target variable is spatially autocorrelated and if point patterns show clear sampling bias (Hengl et al., 2018). In case of dense sampling, such spatial interpolation can be of great interest to overcome the limitations of the landscape covariates for predicting the soil properties (Lagacherie et al, in press).

To correct the non-spatial approach of RF and its derivative, Hengl et al. (2018) proposed adding new covariates that consider the locations of the sites. These covariates are defined as the Euclidian buffer distances from the observation sites. To limit the number of covariates and the computing time in case of large dataset (> 1,000 sites), these distances to the nearest points were not calculated for each individual observation site but for *n* equal classes (from low to high AWC values). As RF is sensitive to the number of classes (Hengl et al., 2018), we performed a trial and error process, which conducted to choose different classes according to the maximal soil thickness considered and to the density scenario (number of classes varying between 6 and 15). For each targeted SAWC, a map was generated. In this DSM model, we considered soil profile and auger hole observations indifferently as soil inputs, omitting their possible differences of uncertainty on the SAWC determinations. This model will be denoted

further QRF_{dist}. The Euclidian buffer distances mapping was performed using the *GSIF* package (Hengl, 2019).

III.3. Inference trajectories

Since we aimed to map SAWC, which is a soil indicator involving several soil properties and several soil depths, it could be estimated following various possible inferences following the order with which "combining primary soil properties", "aggregating soil layers across depths" and "mapping" were performed to provide the SAWC (Styc and Lagacherie, 2019). Styc and Lagacherie (2019) experienced a total of 18 inference trajectories for throughout Languedoc-Roussillon that were performed to obtain the most appropriate SAWC map. From this study, we considered the best performing inference trajectory, i.e., we mapped first AWC of four separate layers (0-30, 30-60, 60-100 and 100-200 cm), and then aggregating the maps of the four soil layers to obtain the final SAWC map.

III.4. Uncertainty analysis using error propagation

In this section, we provide the main details of uncertainty assessment using propagation error. More detailed of the procedure can be found in Román Dobarco et al. (2019) and Styc and Lagacherie (submitted).

The selected inference trajectory, i.e., SAWC estimated as the aggregation of AWC predicted at four depth soil layers, required an error propagation to estimate the variance of SAWC, considered as a proxy of the uncertainty prediction of the target variable (Heuvelink et al., 1989). In this study, we used a first-order Taylor expansion to calculate the error variance of SAWC that results from the error variances of its components (here, the different mapped AWC for the four considered soil layers). This calculation involved i) the error variances of AWC for each soil layer obtained from the conditional distributions provided by QRF for each predicted location (Meinshausen, 2006) and ii) the correlation coefficients between the errors at each soil layer provided by the mapping residuals. Then, the estimate of the SAWC variances were translated into a 90% prediction interval, assuming a normal distribution, by:

$$CIL_i = \widehat{y_i} \pm 1.645\ \sigma_{\widehat{y_i}} \qquad \text{(Eq.5.4.)}$$

where CIL_i is the interval limits of the prediction, $\widehat{y_i}$ the mean of the distribution, $\sigma_{\hat{y}}$ the standard deviation and 1.645 is the Student's coefficient for a 90% confidence interval estimation.

Error propagation was performed using the *propagate* R package (Spiess, 2018).

III.5. The experiment

The goal of the experiment was two-fold: i) to evaluate the efficiency of the DSM model proposed for dealing with dense spatial sampling of auger holes (QRF_{dist}) and ii) to evaluate the cost-efficiency ratio of using auger hole observations with increasing densities.

For that, QRF_{dist} was applied to different soil input scenarios with increasing numbers of auger holes. The performances of the QRF_{dist} was compared with those of a baseline QRF application that did not consider any spatial relation between the sites, as practiced in most DSM applications. The four SAWCs presented in section II.3 were considered. In the following, we provide some details about the sampling strategy for selecting auger holes, the evaluation protocol and the cost-benefit analysis.

III.5.1. The sampling procedure of auger holes

Different data scenarios were considered, all of which included all the available soil profiles. An increasing number of auger holes were sampled from the available set and added to the soil profiles in the soil input datasets (from 10% to 100% of auger hole observations each 10%, e.g. average spacing of 278 m, 556 m, 834 m, 1112 m, 1391 m, 1669 m, 1947 m, 2225 m, 2503 m and 2781 m).

At each step, the auger holes were selected using a stratified random sampling technique using compact geographical strata (Walvoort et al., 2010), as recommended by Brus et al. (2011). Thirty-three geographical strata of 0.5 km^2 were considered. The spatial stratification sampling was performed using the *spcosa* R package (Walvoort et al., 2018).

III.5.2. Evaluation protocol

The performance of the SAWC DSM models was evaluated by k-fold cross validation. This evaluation procedure consisted of randomly dividing the data into k subsets. Then, the holdout method was repeated k times such that one of the k subsets was used as the validation set in each repetition, while the other k-1 subsets were combined to form the calibration set. Following this procedure, every data point was included in a calibration set k-1 times. In this study, we selected k = 10 and to increase the robustness of the evaluation, the 10-fold cross validation was iterated 20 times. The k-fold cross validation was performed using *cvTools* (Alfons, 2012).

To avoid uncertain estimations of the model performances due to the inherent uncertainty of SAWC estimations from the auger hole observations, the evaluation protocol presented hereafter, was solely applied to the soil profiles.

To evaluate the prediction performances, we used classic performance indicators, e.g., the mean square error skill score (Nussbaum et al., 2018), which has the same interpretation as the percentage of variance explained by the model, the root mean square error (RMSE) and the bias.

Furthermore, we evaluated the estimation of the prediction uncertainty using the prediction interval coverage probability (PICP; Shrestha and Solomatine, 2006) and error-predicted-uncertainty estimations. PICP was computed as follows:

$$PICP = \frac{count(LPL_i \leq y_i \leq UPL_i)}{n} \times 100 \qquad \text{(Eq.5.4.)}$$

where n is the total number of observations in the validation set, and the numerator counts if the observation y_i fits within the prediction limits prior to estimation by the error propagation method. For a 90% confidence level, which is usually chosen in DSM studies (Arrouays et al., 2014b), the uncertainty is optimally predicted when the PICP value is close to 90%.

The PICP provides an assessment of the overall uncertainty prediction bias (underestimation or overestimation) but does not tell anything about the ability to map differences in uncertainty across the study area. The PICP was therefore completed by error-predicted-uncertainty estimations that materialized the evolution of the cross validation RMSE with the widths of the predicted confidence intervals. To remove noise, the RMSEs were averaged per quartile of prediction interval widths denoted "low/fairly low/fairly high/high predicted uncertainty". It was expected that the RMSE would increase from low to high predicted uncertainty.

III.5.3. The cost-efficiency of SAWC Digital Mapping

Soil data need to be recorded, but this process can be time consuming, therefore, costly. In order to answer to the question, "Is all the data necessary to reach quality predictions?" we set two indicators to assess: i) the cost of a unit of gained RMSE and ii) the relative cost-efficiency, which were both calculated for each percentage of auger holes added to soil profiles.

The cost of a unit of RMSE was evaluated using the following equation (Eq 5.5):

$$Err_{cost} = \frac{cost_i}{RMSE} \quad \text{(Eq.5.5.)}$$

where Err_{cost} is the cost of a unit of RMSE (in €/cm), $RMSE_i$ the root mean square error of the combination of i% of auger holes and soil profiles datasets.

The relative cost-efficiency was assessed following the recommendation of Kish (1965) used by Viscarra Rossel and Brus (2018) (Equation 5.6):

$$CE_r = \frac{cost_{ref} * RMSE_{ref}}{cost_i * RMSE_i} \quad \text{(Eq.5.6.)}$$

Where CE_r is the relative cost-efficiency ratio, $cost_{ref}$ and $RMSE_{ref}$ are the cost and the error of a reference design, respectively, here using solely soil profiles in SAWC DSM, and $cost_i$ and $RMSE_i$ the cost and the error, respectively, of the combination of i% of auger hole observation and soil profiles. A CE_r larger than one reveals a more efficient sampling than the reference (Viscarra Rossel and Brus, 2018).

IV. Results

IV.1. Preliminary results

In Figure 5.6, we present the distributions of $SAWC_{30}$, $SAWC_{60}$, SAW_{100} and $SAWC_{max}$ for soil profiles (left panel of Figure 5.6) and auger holes (right panel of Figure 5.6). We first observed that the distributions of SAWC regardless of the considered soil depth were bimodal for both soil profiles and auger holes, with i) a higher peak for higher values of $SAWC_{30}$ and $SAWC_{60}$ and with ii) a higher peak for lower values of $SAWC_{100}$ and $SAWC_{max}$. Additionally, it is worth noting that both of the SAWC ranges and means of the auger holes, were systematically greater than soil profiles. This could be explained by i) possible underestimations of coarse fragment by visual determinations on very small volumes using auger holes, compared to real measurements of coarse fragments on larger volumes using soil profiles ii) possible biases of field determination of textural class on auger holes compared with laboratory analyses performed on soil profiles.

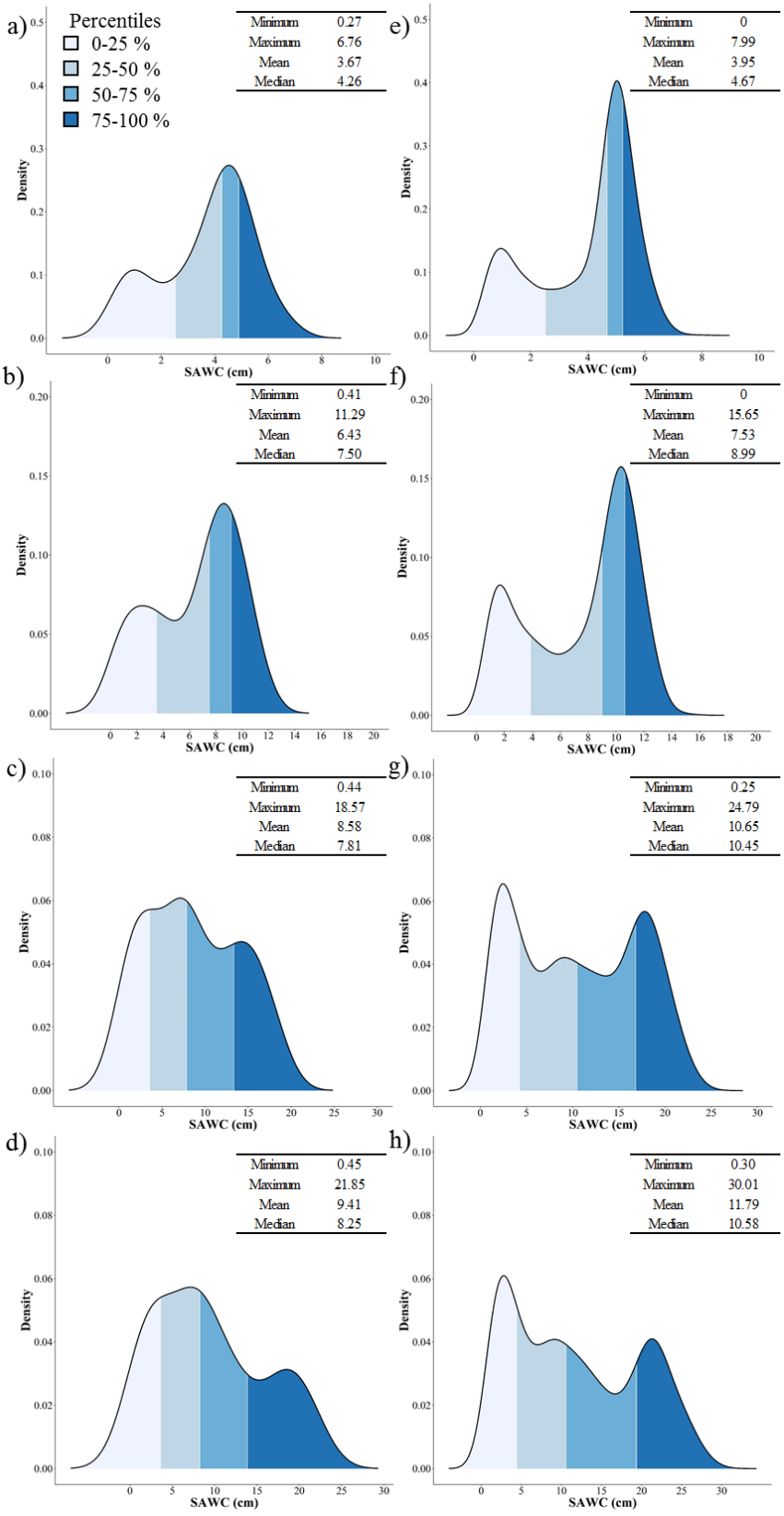


Figure 5.6. Distributions of the soil available water capacity of soil profiles at a) 0-30 cm, b) 0-60 cm, c) 0-100 cm and d) 0-depth$_{max}$; and of auger holes at e) 0-30 cm, f) 0-60 cm, g) 0-100 cm and h) 0-depth$_{max}$

In addition, empirical variograms and their fitted models were computed using the *gstat* package (Pebesma, 2004) both from the soil profiles data (Figure 5.7, left panel) and from the auger hole data (Figure 5.7, right panel), and for the different considered SAWC (lines of Figure 5.7). The Spatially Structured Variance Ratio (SSVR, Equation 5.7) which estimated the portion of the variance that was spatially structured, was computed from the variograms as follows:

$$SSVR = 1 - \left({}^{nugget}/_{variance}\right) x\ 100 \qquad \text{(Eq.5.7.)}$$

First, we noted that the variogram of the SAWC determined from auger hole observations exhibited clear spatial structures regardless of the maximal depth (SSVR ranging from 66% to 76%). The variograms showed a mix of short-range spatial structures (fitted ranges between 332 and 341 m) and large-range structures (fitted ranges exceeding 30 km). Conversely, the variograms of $SAWC_{30}$ and $SAWC_{60}$ determined from the soil profile empirical variogram exhibited less clear spatial structures (SSVR of 25% and 33%), whereas a clear structure appeared for $SAWC_{100}$ and $SAWC_{max}$ (SSVR of 82% and 89%). Because of their larger spacing, the soil profiles did not allow us to see the short-range spatial structures revealed by the auger hole observations. Additionally, significant decreases in nuggets were observed from the variograms of $SAWC_{30}$ and $SAWC_{60}$ processed from profiles to those processed from auger holes. This decrease can be interpreted as the result of increasing sampling densities that better captured the short-range spatially structured variance that was otherwise included in the profile variogram nuggets. It is interesting to note that the converse occurred for $SAWC_{100}$ and $SAWC_{max}$. The probable increase in the uncertainty of observations with depth due to the difficulties in observing deep horizons from auger holes yielded a nugget increase that largely counterbalanced the effect of the sampling density evoked previously.

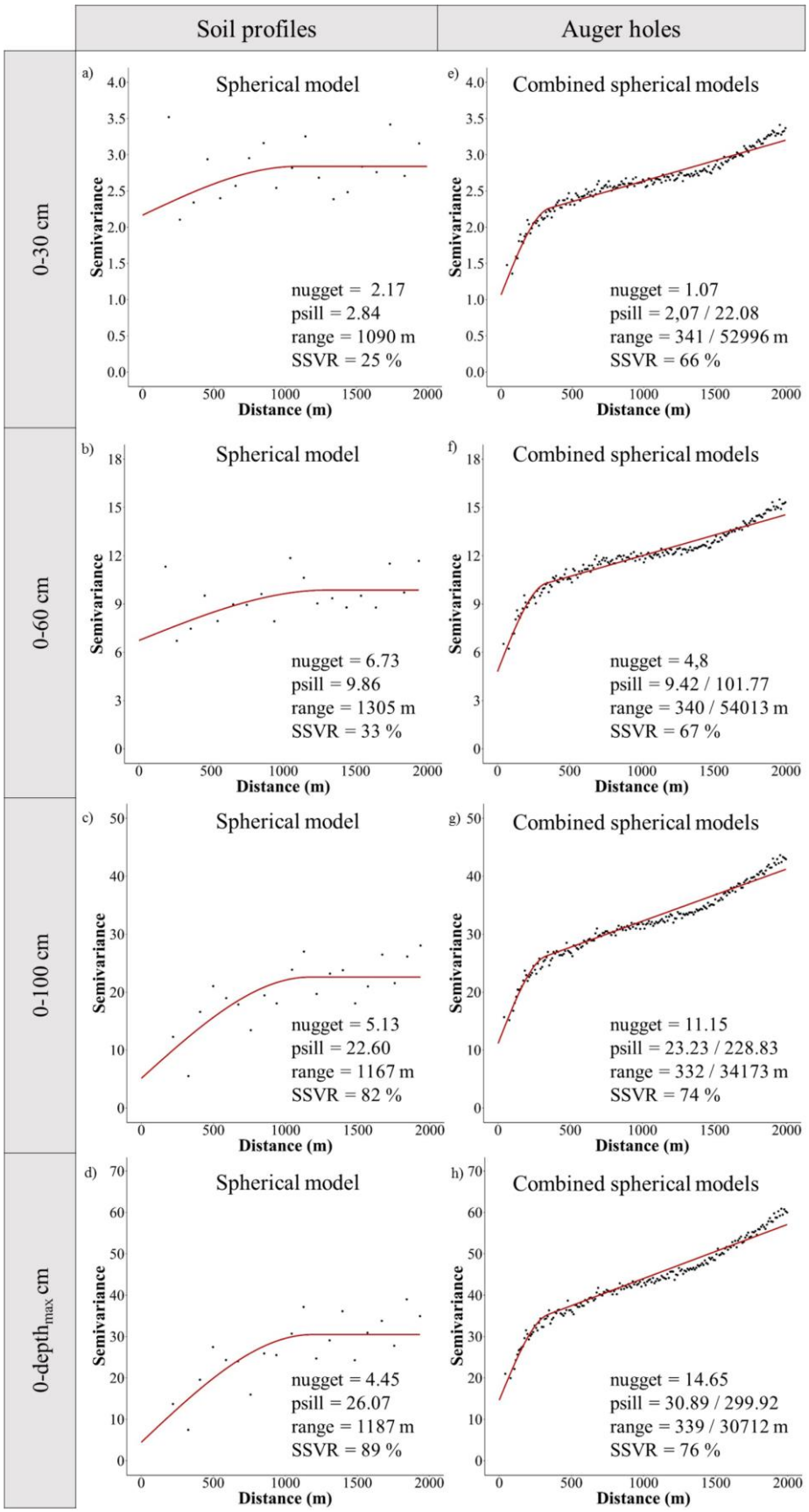


Figure 5.7. Empirical variograms computed for SAWC using 69 soil profiles at a) 30 cm, b) 60 cm , c) 100 cm and d) depth$_{max}$, and, using 2781 auger hole observations at e) 30 cm ,f) 60 cm , g) 100 cm and h) depth$_{max}$, and their theoretical variograms.

IV.2. Comparing DSM models prediction and uncertainty prediction performances

Table 5.3 shows the prediction and the uncertainty prediction performances of the two considered DSM models in predicting the SAWC at four different depths. Only the extreme data scenario, i.e. no auger hole vs the whole set of auger holes, is shown. First, better performances of SAWC predictions were generally obtained by adding the auger hole observations, with the noticeable exceptions of the predictions of $SAWC_{60}$, $SAWC_{100}$ and $SAWC_{max}$ using a classical (non-spatial) QRF. When using QRF_{dist}, the performance increases by adding auger hole observations tended to decrease as the considered depth increased. Additionally, using QRF_{dist} that included geographical information led to better prediction performances regardless of the SAWC only when the auger hole observations were added to soil profiles. Otherwise (i.e. when only the soil profiles were used for calibrating the model), using QRF yielded equal or slightly better prediction performances. Concerning the ability of the models to provide unbiased estimates of prediction uncertainty, as measured by the PICP, larger PICP values were obtained with QRF_{dist} than with QRF, except for the PICP for $SAWC_{100}$ with only soil profiles. Furthermore, the effects of including auger holes in QRFs calibration were different according to the selected model: the PICP decreased when QRF was selected whereas the PICP increased when the QRF_{dist} model was selected. As far as the closeness to the nominal value of 90% is concerned, better results were generally obtained when the auger hole observations were not used, with the noticeable exception of $SAWC_{30}$ predictions using QRF. Furthermore, QRF_{dist} had more PICP values close to the 90% nominal value (< 2%) than did QRF (4 out of 8 vs 1 out 8).

Table 5.3. Prediction and uncertainty prediction performances of SAWC using multiple DSM models

DSM models		QRF				QRF_{dist}			
SAWCs	Auger holes portion (%)	SS_{MSE}	RMSE (cm)	Bias (cm)	PICP (%)	SS_{MSE}	RMSE (cm)	Bias (cm)	PICP (%)
$SAWC_{30}$	0	0,04	1,66	0,17	86	-0.02	1,71	0,32	85
	100	0,38	1,34	0,49	86	0,49	1,22	0,37	90
$SAWC_{60}$	0	0,33	2.74	1,08	87	0,3	2,79	0,35	89
	100	0,32	2,76	1,28	83	0,54	2.26	0,82	93
$SAWC_{100}$	0	0,55	3,73	-0,47	92	0,46	3.97	0.22	90
	100	0,43	4,06	1,82	85	0,63	3.27	1.09	95
$SAWC_{max}$	0	0.61	4.01	-0.68	90	0.53	4.41	-0.56	91
	100	0.54	4.37	1.88	85	0.7	3.54	0.18	96

As expected, the averaged RMSE tended to increase with the widths of the confidence intervals predicted by QRF_{dist} (Table 5.4), which demonstrated the overall validity of the uncertainty predictions. However, non-monotonous increases were observed for the SAWC predictions at small depths that also exhibited the weakest performances (Table 5.3). This non-monotonousness was clearer when the auger hole observations were added. Similar trends were observed for the confidence interval widths predicted by QRF (results not shown).

Table 5.4. Error-predicted uncertainty results of QRF_{dist} using only soil profiles and using soil profiles and auger hole observations for predicting SAWC at multiple depths

SAWCs	Uncertainty	RMSE (cm)	
		Soil profiles	Soil profiles and auger holes
$SAWC_{30}$	Low	1.09	1.31
	Fairly low	1.25	0.79
	Fairly high	2.75	1.10
	High	1.9	1.59
$SAWC_{60}$	Low	2.31	1.25
	Fairly low	2.24	2.02
	Fairly high	2.81	3.25
	High	3.46	2.08
$SAWC_{100}$	Low	2.81	1.52
	Fairly low	2.82	2.81
	Fairly high	3.49	3.69
	High	5.71	4.32
$SAWC_{max}$	Low	3.07	2.24
	Fairly low	2.88	2.82
	Fairly high	4.55	4.20
	High	6.09	4.37

IV.3. Spatial distribution of the SAWC and its associated uncertainty

All the predicted maps of SAWC (Figure 5.8) exhibited spatial patterns of variation that were globally in accordance with the lithological variations shown on Figure 5.1. The highest values of SAWC were predicted in the northeastern section of the study area with fluvisols developed on loess. The smallest values corresponded to chromic luvisols developed on the old stony alluvial deposits.

The spatial pattern became increasingly clear and contrasted as the considered soil depth for calculating SAWC increased (from the top to the bottom of Figure 5.8). The incorporation of auger holes (from left to right column of Figure 5.8) led to: i) an increase in the predicted variabilities of the SAWC, leading to more contrasted patterns regardless of the predicted SAWC ii) an increase in the spatial resolution of the SAWC pattern delineations, showing very fine details of variation , iii) removing some obvious artefacts of the map of $SAWC_{100}$ obtained from the soil profile (Figure 5.8.c) and iv) adding some artefacts (isolated pixels) in $SAWC_{30}$ and $SAWC_{60}$ maps (Figure 5.8.e and f).

The uncertainty maps of SAWC predictions (Figure 5.9) obtained from the QRF_{dist} model exhibited spatial patterns that were both complex and very contrasted across predicted SAWCs and soil inputs. When examining the variations between quartiles of predicted uncertainty that looked significant according to the error-predicted-uncertainty results (Table 5.4), some of the maps revealed strong spatial pattern similarities with those of some uncertainty drivers, i.e. $SAWC_{30}$ uncertainty map using soil profiles (Figure 5.9.a) with the lithology map (Figure 5.1), $SAWC_{100}$ map using soil profiles (Figure 5.9.c) with the spatial density of soil profiles that is observable on the map of soil profiles (Figure 5.4.a), $SAWC_{30}$ uncertainty map using auger hole observations (Figure 5.9.e) with the spatial density of auger hole observations that is observable on the map of auger hole observations (Figure 5.4.b), $SAWC_{max}$ uncertainty map using auger hole observations (Figure 5.9.h) with the predicted map of $SAWC_{max}$. The other uncertainty maps (Figure 5.9.b, d and f) showed less interpretable patterns, with probably mixed impacts of the above evoked drivers.

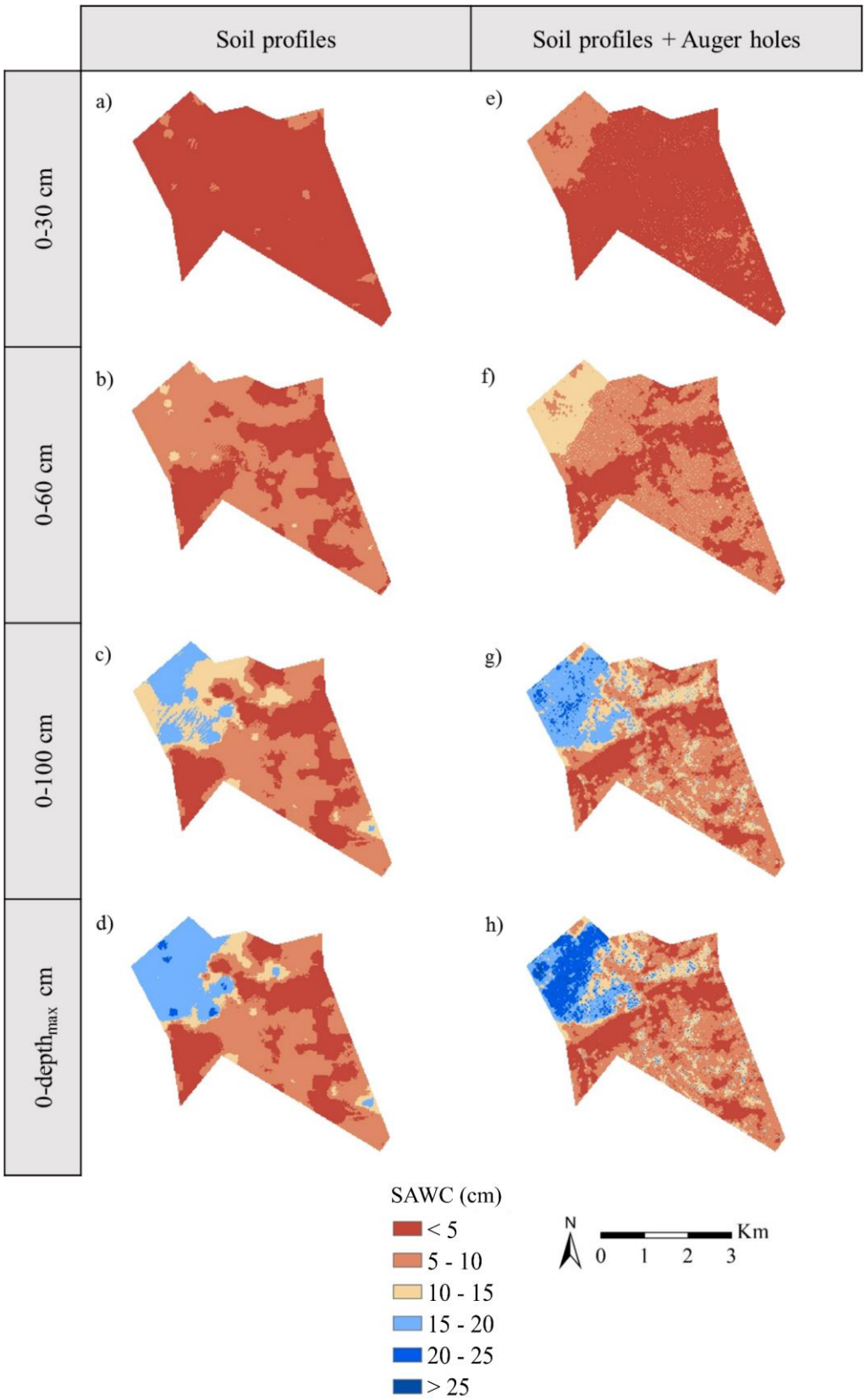


Figure 5.8. Predicted maps of SAWC over Bouillargues using QRF_{dist} with soil profiles for predicting a) $SAWC_{30}$, b) $SAWC_{60}$, c) $SAWC_{100}$, d) $SAWC_{max}$ and using QRF_{dist} with soil profiles and auger hole observations for predicting e) $SAWC_{30}$, f) $SAWC_{60}$, g) $SAWC_{100}$, f) $SAWC_{max}$

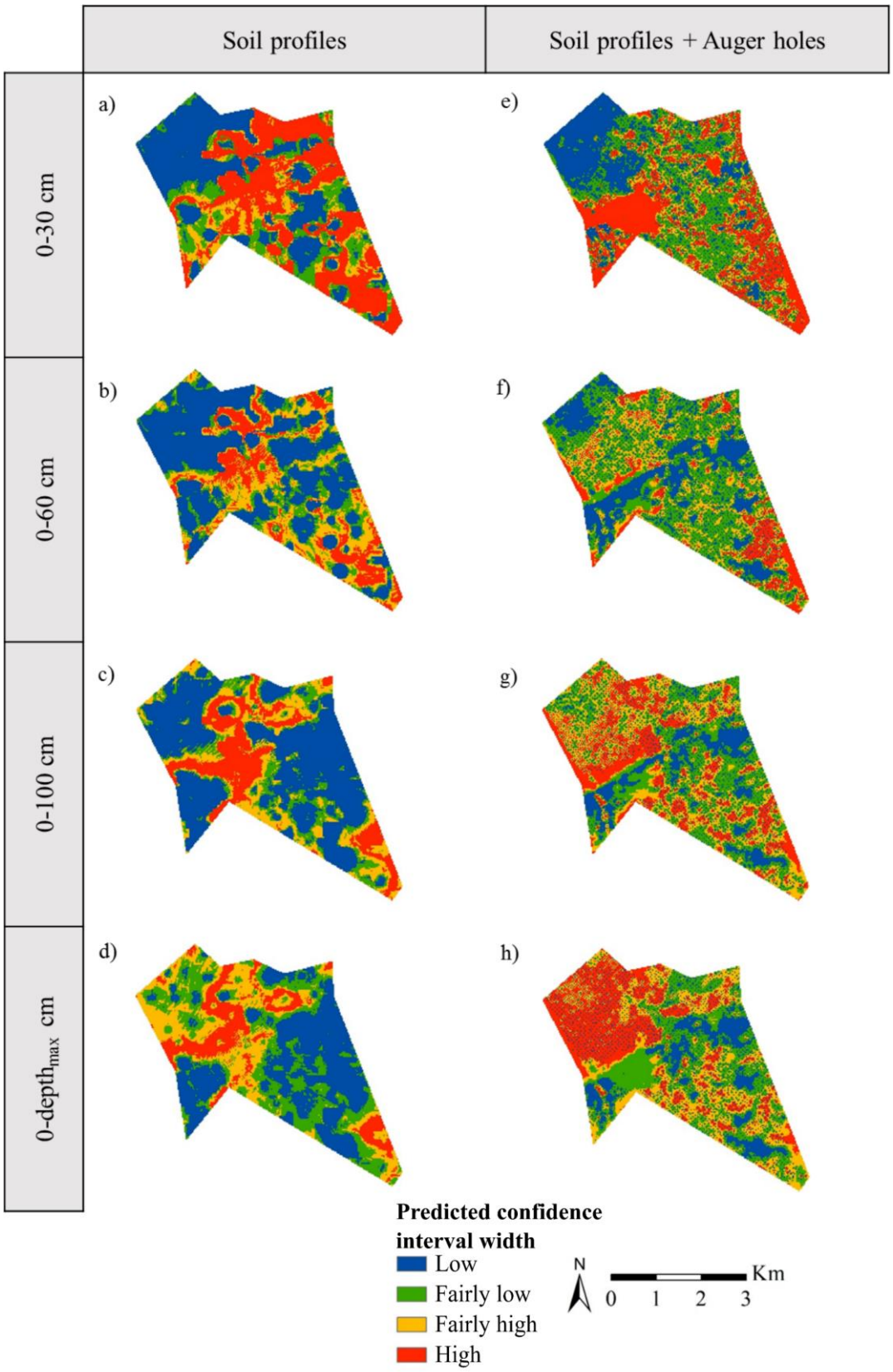


Figure 5.9. Predicted uncertainty maps of SAWC predictions over Bouillargues presented by the classes estimated from the quartiles of the validation distribution using: QRF_{dist} with soil profiles for predicting SAWC at a) 30 cm, b) 60 cm and c) 100 cm d) $depth_{max}$; QRF_{dist} with soil profiles and the whole set of auger hole observations in covariates set for predicting SAWC e) 30 cm, f) 60 cm g) 100, cm and f) $depth_{max}$

IV.4. Comparing spatial densities of auger hole observations

In Figure 5.10, we present the evolution of the SS_{MSE} with the increasing number of auger hole observations in the calibration process. The density in number of observations/km^2 is also expressed as the average spacing between observation sites, which means that the density increases as the average spacing decreases. The average spacing between observation sites was estimated as follows:

$$Average\ spacing = \sqrt{\frac{total\ area}{size}} \quad \text{(Eq.5.8)}$$

As already observed from Table 5.3, the general trend was an increase in performance as the number of the auger hole observations increased regardless of the maximal depth at which the SAWC were calculated. However, some punctual decreases in performances were observed, e.g. on $SAWC_{60}$ and $SAWC_{100}$ predictions when adding 10% of auger holes or on $SAWC_{100}$ and $SAWC_{max}$ predictions when passing from 20 to 30% of auger holes. Conversely, the addition from 10% to 20% auger holes and from 60% to 70% seemed beneficial for all predictions of the SAWC.

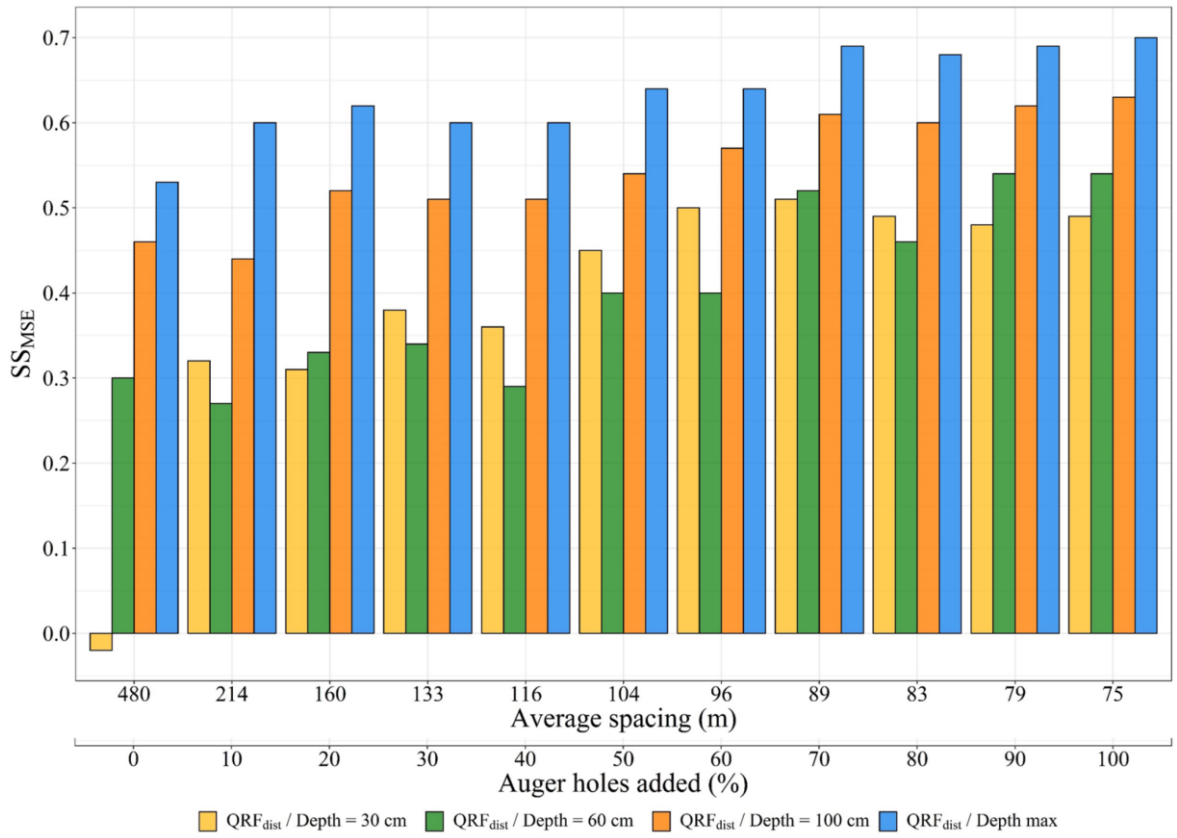


Figure 5.10. Evolutionary SS_{MSE} according to the number of auger hole observations added to the inputs for the four SAWCs

When considering the costs of adding new auger hole observations according to the two cost-efficiency indicators described in section III.5.3, it appeared that the cost of gaining one unit of RMSE (the error cost, Err_{cost}) was important until the first addition of the auger holes and further linearly increased as new auger holes were added (Figure 5.11). This is traduced by the relative cost-efficiency ratio (CE_r) by a dramatic decrease under the 1:1 ratio when adding the first auger hole observations and then a slow decrease for further additions.

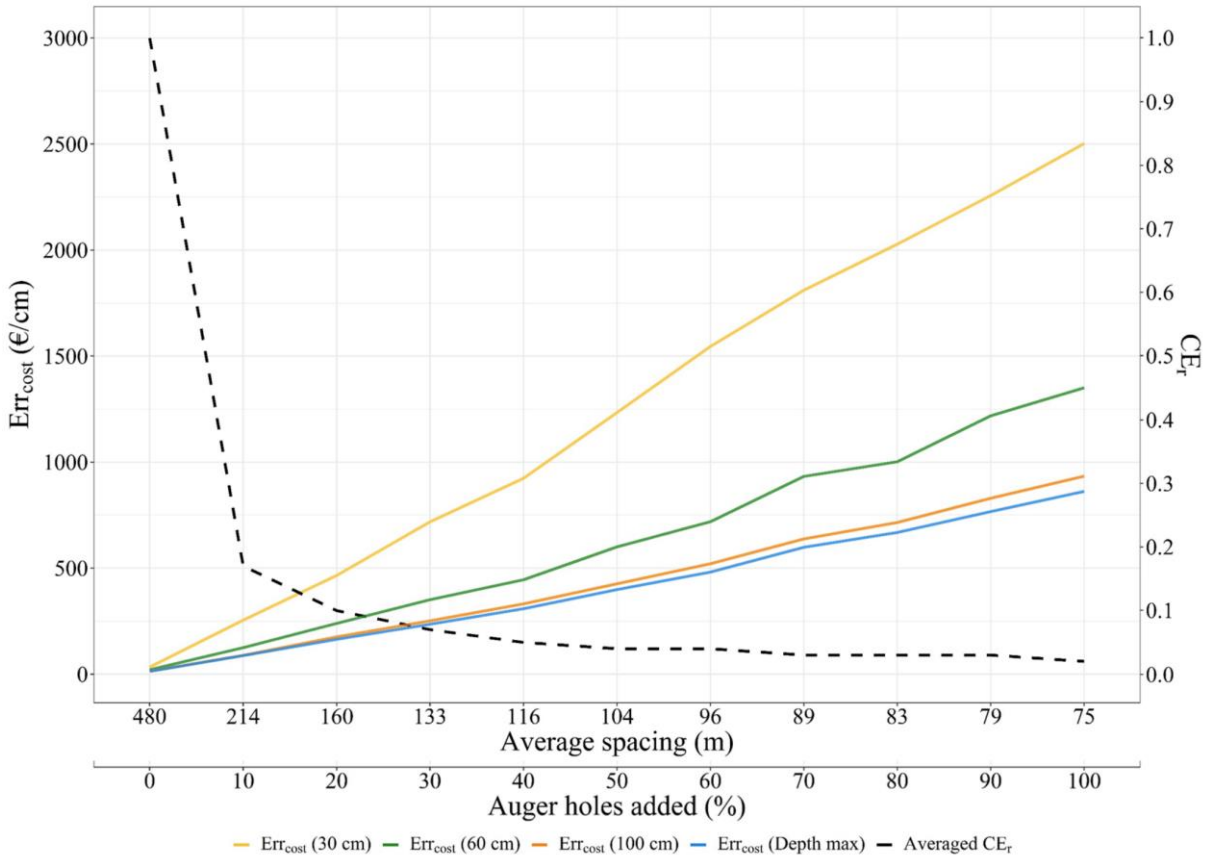


Figure 5.11. Cost-efficiency ratios according to the average spacing related to the amount of auger holes

V. Discussion

V.1. Soil Available Water Capacity

The selected case study considered the soil available water capacity, which is among the most highly demanded properties of the end-users, as the targeted soil property (Richer de Forges et al, 2019). This paper completes the small set of papers that were devoted to the digital mapping of SAWC (Hong et al., 2013; Malone et al., 2009; Padarian et al., 2014; Poggio et al., 2010; Román Dobarco et al., 2019; Ugbaje and Reuter, 2013, Amirian-Chackan et al., 2019) and the even smaller set of papers that addressed all the SAWC components as defined by the

original definition reported by Cousin et al (2003) (Equation 5.1) (Leenaars, 2018; Romàn Dobarco et al., 2019; Styc and Lagacherie, 2019, submitted).

However, as in many DSM applications, the SAWC was determined at local sites without the full measurements of its components. Visual estimations of coarse fragment content and of the soil depth generated observational uncertainties and, for the latter, right-censored estimations due to the limitation in observation depths. Furthermore, the water retention capacity of each horizon was not fully measured although it is worth noting that some components of this retention that are usually not measured (bulk density, field capacity) were here measured on the soil profiles. To overcome the measurement limitations, pedotransfer functions were used (see section II.3). It is worth noting that these pedotransfer functions were highly case specific both regarding their input (textural classes + field capacity measurements) and their target (the b coefficient). The addition of all these peculiar uncertainties should result in a significant overall uncertainty of the soil inputs that is well reported by the nuggets of the variograms of the densest datasets (Figure 5.7 right panel). This uncertainty may greatly explain the limitation of performances that was observed, even for the densest datasets.

V.2. The interest of "spatial RFs"

Our results showed that the SAWC prediction performances were nearly systematically increased by adding some geographical information, i.e. the n of "scorpan" in McBratney et al.'s (2003) formula, to the set of candidate covariates used in a random forest. This confirmed the results obtained by Hengl et al (2018) from various case studies. This, however, enriched these results by showing that the gains in performances provided by the addition of geographical covariates depend on the density of the sampling. Indeed, these gains were only effective when the dense sampling of auger hole observations were used (76 m spacing), whereas the low density of soil profiles did not provide clear improvements (Table 5.3). At high density levels, the classical landscape covariates were not sufficient to account for the variability shown in the dataset of soil inputs as represented by the variograms of Figure 5.7 (right panel) whereas the proximity effects brought by the geographical covariates allowed us to overcome this limitation.

In digital soil mapping, proximity effects have been traditionally addressed by using regression kriging (Hengl et al., 2004; Malone et al., 2009; Vaysse and Lagacherie, 2015). However, spatial QRF was demonstrated as having similar performances (Hengl et al, 2018) while having some decisive advantages in the context of our case study. It does not require any rigid statistical assumptions about the distribution and the stationarity of the target variable, which allows us

to handle the bimodal distributions of SAWCs (Figure 5.6). It also does not require any geostatistician expertise for the manual fitting of variograms, which opens the possibility to fully automate the procedure so that non pedometrician such as BRL staff, could use it for the other communes of the irrigation perimeter.

V.3. The interest of adding auger hole observations

The addition of dense spatial sets of auger hole observations in the modeling process significantly increased the level of performance when considering the best model (QRF_{dist}), which is in accordance with several previous experiments studying the impact of soil sampling densities (Somarathna et al. 2017, Wadoux et al. 2019 and Lagacherie et al, in press). The performances observed in this case study were better than most of the published DSM applications dealing with SAWC (Ugbaje and Reunter, 2013; Styc and Lagacherie, 2019, submitted) which was the result of a much greater spatial density of the soil input (from 6/km^2 to 26/km^2) than these previous applications (from 0.01/km^2 to 0.05/km^2).

However, strong limitations in the SAWC prediction performances were still observed, even when using the most dense set of auger hole observations. These limitations increased as the maximum depth at which the SAWC was calculated decreased (Table 5.3). This means that significant proportions of the SAWC variabilities were not mapped despite of the large densities of the auger hole observations used as input. To explain this fact, it is first interesting to note that for both soil profiles and soil profiles plus auger holes inputs, the performances and the spatially structured variance ratios of the input soil datasets were ranked similarly across SAWCs and spatial densities (Figure 5.7), which was already observed in the same region for different soil properties and study extent by Vaysse and Lagacherie (2015). Concerning the results using solely the soil profiles, this revealed that a part of the short-range variability shown by the variograms built from auger holes (Figure 5.7, left panel) was not captured by the soil dataset because of a limitation of spacing. However, this limitation decreased as the considered depth of SAWC calculation increased, which explained the observed increase in performance from $SAWC_{30}$ to $SAWC_{max}$. Concerning the results using the auger hole observations, a similar trend was observed since the local uncertainty as revealed by the variogram nuggets (Figure 5.7, right panel) remained important due to observational uncertainty (see section IV.1.), which may induce noise that may perturb the calibration of the QRF model.

Finally, it should be recalled that these performances were calculated for punctual predictions of the SAWC, whereas SAWC is required for field or in-field management zones for most of

the decision making. It could be expected that these performances would increase when the SAWC prediction will be spatially aggregated (Vaysse et al, 2017).

V.4. Uncertainty predictions

Since SAWC is a soil functional property composed by several primary soil properties, uncertainty predictions were provided by a specific error model previously proposed by Román Dobarco et al. (2019) and further refined by Styc and Lagacherie (submitted). The uncertainty predictions were classically evaluated with regard to their unbiasness (PICP, Table 5.3). They were also evaluated for their ability to identify contrasted uncertainty areas (comparisons between residuals and predicted uncertainty, Table 5.4), which, to our knowledge, has never been done in the DSM literature before Styc and Lagacherie (submitted). The results were highly variable across models and spatial densities. However, the most accurate models tended to also provide the best pictures of the uncertainty patterns (Figure 5.9) with an overestimation of uncertainty (QRF_{dist} on Table 5.3). This overestimation was already observed by Lagacherie et al (2020) and was assumed to be due to the inclusion of outliers as the average spacing decreased, which probably disturbs the limit estimations of the confidence interval. More attention must be paid in the future to uncertainty predictions in view of identifying the possible causes of these uncertainty mispredictions.

It is interesting to note that some of the produced uncertainty maps showed strong similarities with possible drivers (see comments of Figure 5.9), which can be interpreted from our common sense pedological knowledge. The largest uncertainties were estimated i) in chromic Luvisols (Figure 5.9.a) because of the large rates of coarse fragment content that are known to be difficult to quantify in the field, ii) in areas of lower densities of soil observations (Figure 5.9.c and e) because of difficulties of model calibration at these locations and iii) for the largest predicted values of SAWCs with the best models (Figure 5.9.h) because the estimates of relative uncertainty reached an unsurmountable floor that is likely related with the observational uncertainty. All these observations reinforce the credibility of the presented uncertainty maps.

V.5. The level of performances obtained and cost.

The use of auger hole observations as complementary soil input to soil profiles led to a substantial increase in performance, but the harvesting process was very time consuming, which result in high costs (see section IV.4). Figure 5.10 curves show that the performance gains were obtained by increasing costs as the density of auger hole increased. A compromise should then be found which can be formulated as "the number of auger hole observations that reach an

acceptable level of performance while keeping an acceptable cost level". The cost indicator curves of Figure 5.11 did not reveal a clear compromise. However, such curves could be used with a prior definition of what performance and costs are acceptable. Furthermore, such cost curves could be improved if either more sophisticated sampling are used (e.g. van Groningen et al, 1998) or if the harvesting costs could be reduced by a partial automation of digitizing procedures (Yang and Yang, 2017)

Finally, it should be stressed that the quantitative evaluation of prediction performance that served as a basis for building the curve costs should be completed by a qualitative examination of the maps. As revealed by the spatial patterns of the predicted SAWC maps, considerable gains in spatial resolution were obtained by adding auger holes, which may enable field-level decision making. This may constitute a more decisive added value than the moderate gain in precision quantitatively evaluated by the cost indicators.

VI. Conclusion

In this study, the main lessons were as follows:

- A QRF using covariates that represent the locations of the site to be predicted outperformed a classical QRF approach in predicting SAWC with a dense set of profiles and auger holes,
- The addition of a dense spatial sampling of auger hole observations dramatically increased the performance in predicting SAWCs and increased the spatial resolutions of the predicted maps, but there were limitations due to the uncertainty of the auger hole observations,
- The performances in predicting both SAWC values and the prediction uncertainty varied following a set of drivers that could be identified: the SAWC variation patterns, average spacing of sites, type of observations (soil profiles vs. auger holes),
- The cost-efficiency analysis did not reveal a clear compromise in terms of limiting the costly harvesting of auger hole data. Rather, this compromise should be user specific and should be updated as soon as partial automation is possible.

Les éléments importants de ce chapitre à retenir sont :

- L'intégration d'un jeu de données spatialement dense a **considérablement augmenté** les performances des modèles de cartographie numérique des sols.
- L'utilisation d'un QRF permettant d'intégrer la structuration spatiale des données sous forme de covariables (QRF_{dist}) a fourni de meilleurs résultats que l'utilisation standard du QRF.
- L'association d'un jeu de données dense à un QRF spatial a permis d'obtenir des cartes de RU et de son incertitude associée avec **un grain de détail inédit**.
- Les performances obtenues dans la **spatialisation du réservoir utile et de son incertitude** ont également varié selon plusieurs facteurs : la variabilité spatiale du réservoir utile, l'espacement moyen entre les sites et le type de données utilisées (profil de sols vs sondages pédologiques).
- Les performances restent toutefois **limitées** par une incertitude liée à un biais d'observation de certaines propriétés du RU sur les sondages pédologiques.
- L'étude coût-bénéfice n'a pas permis d'identifier de compromis clair sur une limite financière à la saisie des sondages pédologiques. Le compromis devrait plutôt être décidé selon **les besoins et les exigences de l'utilisateur**. De plus, le coût de la saisie pourrait être diminué par la mise en place d'une saisie automatisée des observations de sol.

Conclusion générale

Autour de l'usage concurrentielle de l'eau, les différents acteurs se mobilisent pour atteindre une utilisation raisonnée de la ressource en eau. Dans le cadre d'une gestion de l'eau pour la production agricole, viticole et fruitière telle qu'assurée par BRL Exploitation sur la plaine littorale Languedocienne, l'objectif est de garantir une distribution de l'eau d'irrigation avec efficacité et parcimonie. Pour cela, ces travaux de thèse se sont inscrits dans une démarche visant à élaborer une méthodologie de cartographie de la capacité du sol à stocker l'eau utilisable par les plantes (le « Réservoir Utile » du sol), à haute résolution spatiale et en mobilisant les données pédologiques anciennes fournies par BRL Exploitation. Afin de répondre à cet objectif et en relation avec les travaux déjà réalisés sur le sujet, plusieurs verrous méthodologiques ont été identifiés :

Spatialisation du réservoir utile par l'utilisation de trajectoires de calcul

La spatialisation d'une propriété fonctionnelle telle que le réservoir utile des sols est un défi tant le nombre de trajectoires de calcul est important. La trajectoire de calcul se définit comme la succession d'opérations permettant d'obtenir une cartographie du réservoir utile à partir de propriétés primaires de sols, selon l'enchaînement des opérations suivantes : « combiner les propriétés primaires de sols », « agréger les couches de sols » et « spatialiser ». Bien que la spatialisation du réservoir utile soit peu abordée dans la littérature, les quelques applications utilisent des trajectoires de calcul sensiblement variées. Dans le but d'étudier l'influence du choix de la trajectoire sur la qualité des estimations de RU, il a été important d'effectuer une étude comparative entre les trajectoires référencées dans la littérature mais également d'introduire des trajectoires de calcul intermédiaires non encore essayées.

Quantification et spatialisation de l'incertitude par l'application d'un modèle de propagation d'erreurs

La quantification et la spatialisation de l'incertitude de prédictions du réservoir utile est une donnée de premier ordre pour guider les prises de décisions par les différents acteurs de la ressource en eau. Selon les spécifications du programme *GlobalSoilMap*, la cartographie d'une propriété de sol doit être systématiquement délivrée avec une carte de son incertitude sous la

forme d'un intervalle à 90%. Si cette demande est assez bien maîtrisée dans la spatialisation d'une seule propriété de sol, l'estimation de l'incertitude de spatialisation du RU, propriété fonctionnelle à caractère multivarié, nécessite une propagation des différentes sources d'erreurs selon la trajectoire de calcul choisie et l'ordre des opérations appliquées (combinaison des propriétés et agrégation des couches de sol).

Utilisation des données pédologiques anciennes dans les modèles de cartographie numérique des sols.

Les performances des modèles de cartographie numérique des sols actuelles sont principalement limitées par une densité trop faible de données pédologiques ponctuelles. Cette faible densité ne permet pas aux modèles de capturer la variabilité de la propriété à prédire qui se situe généralement à une échelle plus fine. Malgré plusieurs programmes de récupération de données anciennes, les bases de données pédologiques actuelles (nationale et régionale) ne sont pas assez riches pour réaliser une cartographie détaillée à haute résolution. L'utilisation de données pédologiques anciennes disponibles localement pour augmenter la densité des données d'entrée des modèles de CSMS se présente comme une alternative à la réalisation de nouveaux profils de sols, à la fois crédible et moins coûteuse.

Par la suite, nous résumerons les principaux résultats et conclusions de ces travaux de thèse selon i) un point de vue méthodologique dans l'intérêt d'un approfondissement des connaissances de la cartographie numérique de sols appliquée au réservoir utile pour la communauté scientifique et ii) un point de vue plus technique relevant d'un caractère opérationnel de la méthodologie présentée dans cette thèse.

Tout d'abord, nous présentons en quelques points les principales avancées méthodologiques des méthodes de cartographie numérique des sols appliquées au RU :

- L'évaluation de performances des prédictions pour les 18 trajectoires de calcul testées a montré une sensibilité des estimations de réservoir utile aux trajectoires de calcul (Chapitre 3). La trajectoire de calcul la plus adaptée à l'échelle du Languedoc-Roussillon consiste à spatialiser des estimations du réservoir utile par couches de sols partiellement agrégées (calcul d'une moyenne pondérée par l'épaisseur des trois couches de sols les plus superficielles (0-5 cm, 5-15 cm et 15-30 cm)). En revanche, l'agrégation de la totalité des couches a fourni les résultats les moins satisfaisants. Par

ailleurs, les estimations du réservoir utile sont davantage sensibles au degré d'agrégation des couches (ΔSS_{MSE} = 0,39) de sols que celui des combinaisons de propriétés (ΔSS_{MSE} = 0,05). Aussi, il a été mis en avant qu'une étude des corrélations entre propriétés de sol et entre couches de sol permettait également d'anticiper la trajectoire de calcul la plus adaptée. En effet, les trajectoires de calcul aux meilleures performances ont été obtenues lorsque les couches de sols très corrélées (coefficient de corrélation > 0.9) ont été agrégées avant spatialisation. Inversement, les trajectoires de calcul aux faibles performances consistaient à agréger des couches de sols moins corrélées (coefficient corrélation < 0.8).

- L'estimation de l'incertitude de prédiction a pu être fournie via l'utilisation de modèles de propagation d'erreurs alimentés par les erreurs de spatialisation de l'épaisseur de sol et du réservoir utile élémentaire (Chapitre 4). La prise en compte, dans ces modèles, de multiples combinaisons des corrélations entre les propriétés et entre les couches a été testée sans toutefois prendre en compte les erreurs liées à la structure des FPT jugées négligeables dans notre cas d'étude, conforté par les résultats précédents de Román Dobarco et al. (2019). Les résultats ont mis en avant l'importance de la prise en compte des corrélations d'erreurs entre couches de sol par rapport à celles des propriétés composant le RU, dans l'évaluation de l'incertitude de prédiction selon plusieurs profondeurs de sol. En effet, la largeur de l'intervalle de confiance était sous-estimée lorsque les corrélations des erreurs entre couches de RU n'étaient pas considérées dans le modèle de propagation d'erreur. Il a été également montré, en vérifiant l'adéquation entre largeurs estimées d'intervalles de confiance et erreurs mesurées par validation, que le modèle de propagation d'erreur spatialisait correctement les erreurs. Une carte de l'incertitude de prédiction du RU selon ces quartiles de distribution de l'intervalle de confiance a également été délivrée permettant d'identifier les zones hautement incertaines généralement localisées dans les zones de fortes valeurs de RU prédites, tandis que les zones faiblement incertaines sont localisées dans les zones de faibles RU prédites.
- L'utilisation des données anciennes a permis une augmentation nette du niveau de performance de prédiction du RU pour les différentes profondeurs (Chapitre 5). De plus, les récentes avancées méthodologiques de Hengl et al. (2018) ont permis d'introduire dans l'algorithme de forêt aléatoire quantile, la prise en compte de la structuration spatiale d'un jeu de données. La combinaison de l'utilisation d'un modèle spatial (QRF_{dist}) ainsi que d'un jeu de données denses (espacement moyen jusqu'à 75 m) a

permis d'exploiter la structuration spatiale des données, donnant des niveaux de performances rarement observés dans la littérature. De plus, les cartes issues de l'utilisation de ces modèles ont affiché un grain de détail également inédit. Les modèles les plus performants ont eu toutefois tendance à surestimer l'incertitude, bien que délivrant des cartes très détaillées également.

D'un point de vue plus technique, cette thèse a pu répondre à divers points :

- L'élaboration d'une approche de cartographie numérique des sols ne nécessitant pas d'expertise pédologique et géostatistique avec l'utilisation d'un modèle d'apprentissage non-paramétrique, forêt aléatoire quantile. De ce fait, les opérations nécessaires dans le cadre d'un modèle géostatistique tel que les transformations de données (ex : transformation logarithmique) afin de respecter l'hypothèse de stationnarité, ainsi que l'ajustement d'un modèle de variogramme nécessairement réalisé par l'utilisateur, ne sont pas requises. Dans le cadre d'une démarche opérationnelle reproductible sur l'ensemble du territoire détenu par BRL Exploitation, le compromis entre performances de la méthodologie et facilité d'utilisation semble acquis. Si possible, une vérification qualitative des résultats par un pédologue serait néanmoins utile pour éviter certains artefacts.
- Une étude coût bénéfice a également été réalisée (Chapitre 5) afin d'estimer le coût d'utilisation et de saisie des données pédologiques anciennes par rapport aux performances de prédictions du RU. Pour cela, le scénario de référence a été l'utilisation seule des profils facilement et rapidement numérisables. Selon les indicateurs proposés, le coût lié à l'augmentation des performances par l'intégration des sondages pédologiques aux profils de sols est élevé, du fait d'un protocole de saisie très fastidieux et long à exécuter. Cependant, ces indicateurs se sont révélés incomplets du fait qu'il ne soit pas encore possible de mesurer le gain de résolution spatiale visible à l'examen comparatif des cartes de RU élaborées avec les profils et les sondages pédologiques. Si un compromis entre performances de prédictions du RU et coût de saisie ne peut être clairement défini, c'est également parce que les exigences de l'utilisateur (niveau de performances de prédictions de RU et de son incertitude) ne sont pas intégrées à ces indicateurs.
- Cependant, d'un point de vue économique, la récupération des profils de sols issus des données pédologiques anciennes a démontré son intérêt par un coût de saisie estimé à 1€ par profil de sol contre 1 800 € pour la réalisation d'un nouveau profil de sol.

En association avec les résultats de ces travaux de thèse exposés ci-dessus, nous avons pu identifier certains verrous et limites afin d'améliorer la cartographie numérique appliquée au réservoir utile.

Estimation du réservoir utile sur les sites observés

Pour une estimation du réservoir utile à l'échelle régionale, nous nous sommes appuyés sur une formulation opérationnelle de l'estimation du RU selon Cousin et al. (2003). Bien que nous ayons intégré la profondeur d'enracinement des cultures en dérivant des cartes du RU à différentes profondeurs, cette démarche est simplificatrice. La profondeur d'enracinement étant éminemment variable selon le type de culture, les pratiques culturales et les conditions climatiques (Cousin et al, soumis), les valeurs du RU prédites dans certaines zones peuvent être surestimées ou sous-estimées. Ensuite, pour les éléments grossiers, bien que leur teneur volumique ait été pris en compte afin d'éviter une surestimation du RU, la nature lithologique des éléments grossiers n'a pas été considéré ce qui peut potentiellement correspondre à sous-estimer le RU (Tetegan et al, 2011). Enfin, l'utilisation de fonction de pédotransferts calibrées à l'échelle nationale par Román Dobarco et al. (2019) peuvent aussi être une source d'erreurs d'estimations, bien qu'il ait été démontré que le Languedoc-Roussillon soit inclus dans les zones d'applicabilités des FPT.

A l'échelle locale du périmètre BRL, la formulation utilisée historiquement par BRL Exploitation pour calculer le RU présente également des limites. Bien que cette formulation soit adaptée aux données mesurées durant la période 1957-1992, une mise à jour vers une estimation plus générique telle que proposée par Cousin et al. (2003) serait judicieuse pour plusieurs raisons. Tout d'abord, les paramètres permettant de déterminer la capacité de rétention de la terre fine, l'humidité équivalente et le coefficient de filtration tendent à surestimer cette capacité. Pour les profils, les valeurs d'humidités équivalentes ont été estimées par l'application d'une force centrifuge équivalant à une pression de -10 kPa. Ces valeurs ont été considérées comme une approximation des valeurs de l'humidité à la capacité au champ. Or, la pression de référence appliquée pour mesurer l'humidité à la capacité au champ est de -33 kPa (Nachabe, 1998). De plus, le coefficient textural est également une source d'erreur car son estimation est issue de fonction de pédotransfert utilisant les valeurs d'humidité équivalente mesurées. Plusieurs alternatives pourraient être envisagées. D'une part, les FPT développées à l'échelle nationale par Román Dobarco pourraient être utilisées pour calculer les humidités

caractéristiques de sol à partir de la teneur en sable et argile ainsi que la densité apparente, données disponibles sur les profils de sol de BRL. D'autre part, des FPT locales plus précises car élaborées sur un territoire plus conforme au contexte pédologique du périmètre irrigué de BRL (Bastet et al, 1999) pourraient également être envisagées. Par ailleurs, il serait également possible d'améliorer l'estimation de la capacité de rétention en eau et du point de flétrissement sur les sondages en corrigeant les biais d'estimation observés dans l'appréciation qualitative des classes de texture. Pour cela une nouvelle fonction de pédotransfert corrigeant les valeurs médianes de granulométrie pour chaque classe de texture pourrait être calibrée à partir des profils de sol où on dispose à la fois de l'appréciation texturale qualitative et de l'analyse granulométrique. Enfin, la teneur en éléments grossiers étant massique dans l'estimation du RU pourrait être également convertie en volumique afin d'atteindre l'ensemble des conditions requises pour appliquer une formulation plus générique de l'estimation du RU. Notons enfin que la difficulté de prendre en compte le RU des éléments grossiers se heurte au fait que la nature des cailloux n'a pas été enregistrée sur les observations de BRL alors qu'elle conditionne fortement ce RU (Algayer et al, 2020).

Application des trajectoires de calcul par zones homogènes

Comme présenté ci-dessus ainsi que dans le Chapitre 3, les estimations du RU sont sensibles au choix des trajectoires de calcul. Bien que les performances ont montré une sensibilité vis-à-vis du degré d'agrégation des couches de sol, avec la moins performante des trajectoires visant à agréger l'ensemble des couches de sol, il serait possible que ces mauvaises performances soient due à une application trop générale de la trajectoire. En effet, il est possible d'imaginer que cette trajectoire puisse être utilisée pour des sols relativement homogènes en profondeur. Afin de tenir compte du degré de variabilité d'un profil de sol, une étude préliminaire au choix de la trajectoire pourrait être réalisée en veillant à déterminer un indice de variabilité des propriétés du RU en profondeur. Cet indice serait ensuite comparé à un seuil fixé à dire d'expert ou par analyse de sensibilité. Si l'indice est inférieur au seuil, le sol est alors considéré comme homogène en profondeur (Figure 6.1.a) ou hétérogène en profondeur le cas échéant (Figure 6.1.b). Cela permettrait d'identifier les profils homogènes où une trajectoire de calcul sur une seule couche de sol pourrait potentiellement être plus adaptée à délivrer une meilleure prédiction que l'application d'une trajectoire de calcul global. Ce procédé serait également appliqué aux profils de sols hétérogènes nécessitant une discrétisation détaillée de la profondeur. Il s'agirait alors d'élaborer une « côte mal taillée » en appliquant des trajectoires

par zone de sols homogènes/hétérogènes (Figure 6.1.c). Toutefois, cette méthode présenterait quelques inconvénients tels que i) la nécessité d'avoir une densité d'observations de sol importante telle que celle présentée avec l'utilisation des données BRL Exploitation, ii) la segmentation de l'ensemble d'apprentissage en sous-ensembles d'apprentissages moins importants et potentiellement moins efficaces pour l'apprentissage et iii) la création de discontinuité de prédictions à l'interface de deux zones.

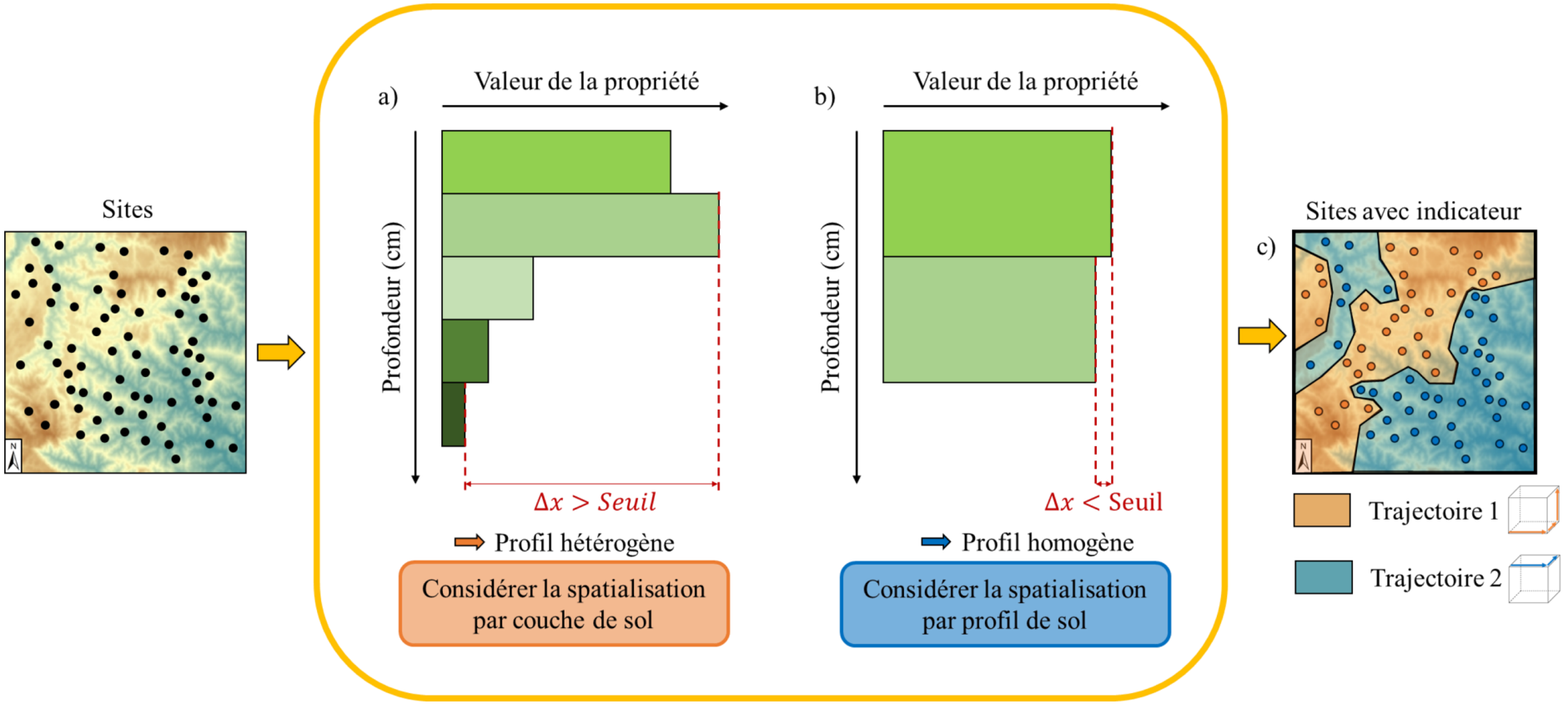


Figure 6.1. Représentation d'une application zonale des trajectoires de calcul pour la spatialisation du RU

Protocole d'évaluation de l'incertitude appliqué au réservoir utile

Le principe de l'évaluation en CNS consiste à comparer les prédictions avec les valeurs mesurées d'une propriété de sol. Dans le cas d'une approche CNS du RU, nous avons vu que la mesure directe du RU n'est pas faisable. Néanmoins, nous avons constitué un jeu de données semi-indépendant, en estimant un RU par profil équivalent à la somme des RU estimés par horizon selon la formulation du Cousin et al. (2003) (Chapitre 3) pour valider des prédictions du RU estimées par couches puis agrégées. Toutefois, cette validation ne prend pas en compte les vraies valeurs du RU mais des valeurs estimées de la capacité de rétention de la terre fine par utilisation des FPT. Une validation plus rigoureuse des prédictions du RU pourrait être réalisée à partir de sites avec des mesures d'humidités spécifiques en laboratoire. Ce type de mesure étant très lourde et coûteuse, cette option ne paraît pas réaliste en pratique. L'alternative serait de tester la concordance avec des données de substitution du RU (« proxies » en anglais), notamment mesurées sur la plante, à l'image de l'indice de fertilité des arbres utilisé par Algayer et al (2020) pour tester différentes estimations du RU. Dans cette perspective, le rapport isotopique du carbone $\delta^{13}C$, un indicateur de stress hydrique de la plante, pourrait constituer un bon proxy du RU (Coulouma et al., 2020).

Perfectionnement des modèles de cartographie numérique des sols

L'un des principaux avantages de la cartographie numérique des sols est de pouvoir constamment perfectionner les méthodes utilisées par l'ajout de nouvelles données ou d'algorithmes plus sophistiqués (Lagacherie et al., 2013). Si les modèles de fouilles de données sont parmi les plus utilisées dans les études CSMS (Vaysse et Lagacherie, 2015, 2017), depuis quelques années les modèles d'apprentissage profond (deep-learning) émergent peu à peu dans les études CSMS (Wadoux et al., 2019). Ces modèles ont vocation de mieux caractériser les relations sols-paysages souvent matérialisées par une régression linéaire entre la valeur d'une propriété de sol pour un site et la valeur de la covariable extraite à ce même point. Cependant, les relations sols-paysages sont beaucoup plus complexes et une contextualisation de la covariable en considérant, non pas seulement un point, mais une zone incluant la localisation de ce point est intéressante à envisager (Robbez-Masson, 1994). Par exemple, les sols sur une pente faible peuvent éventuellement présenter une accumulation importante en matière organique, accumulation toutefois conditionnée par les pentes environnantes (Wadoux et al., 2019). La contextualisation des covariables intégrée dans les modèles spécifiques de deep

learning (Convolutional Neural Networks) a montré une qualité de prédictions supérieure (R^2 = 0.55, Wadoux et al., 2019) à l'utilisation classiques des modèles d'apprentissages type forêt aléatoire (R^2 = 0.35, Wadoux et al., 2019). Les modèles d'apprentissage profonds utilisés en CSMS sont principalement des modèles basés sur une famille de réseaux de neurones artificiels appelées « Convolutional Neural Networks » (Behrens et al., 2018 ; Wadoux et al., 2019 ; Padarian et al., 2019). Il resterait à vérifier que les améliorations observées par ces nouveaux modèles soient toujours effectives dans notre contexte d'application, ces modèles ayant été testés pour l'instant dans les conditions d'utilisation classiques de la CSMS, c'est à dire avec des données de sol peu denses.

Correction des biais liée aux données pédologiques anciennes

Ce dernier verrou a un caractère plus technique. L'utilisation des données anciennes dans les modèles CSMS a permis de s'affranchir du facteur limitant principal des performances de ces modèles : la densité spatiale des observations de sol. Si les données pédologiques sont effectivement denses, elles sont également hétérogènes (Chapitre 5) et possiblement biaisées (Chapitre 2). L'hétérogénéité des données s'applique aussi bien à l'estimation des propriétés de sols (analyses de sol vs observations de terrain), au géoréférencement (positionnement « manuel » vs relevés des coordonnées géographiques) et de couverture spatiale (répartition homogène vs sous forme de clusters sur le territoire). L'hétérogénéité liée aux données a été identifiée comme un facteur limitant les performances de modèles CNS (Chapitre 5). En effet, l'ajout de sondages pédologiques permet d'augmenter la densité et de pouvoir capter la variabilité locale de la RU mais les données issues des sondages comportent des biais d'observation liés à plusieurs causes : limitations de l'outil tarière (ex pierrosité, profondeur), subjectivité des observateurs multiples (appréciations de classes texturales), aux dates de prélèvement induisant des différences de conditions d'observation (ex : sols trop secs pour être pénétrés) ou des changements de protocole analytique. Toutefois, il existe dans la littérature des méthodes de corrections de biais (Baume et al 2011, Ciampalini et al, 2013) qui pourrait être appliquées à ces données.

En conclusion, si les résultats et avancées majeurs de cette thèse ont permis de répondre à la problématique posée, ils ont également contribué à exposer de nouveaux verrous méthodologiques pouvant alimenter des prochains travaux. L'ensemble de ces travaux de thèse sera prochainement compilé en une chaîne de traitement opérationnel qui sera livrée à BRL Exploitation afin de déployer cette méthodologie de cartographie numérique du réservoir en eau du sol sur l'ensemble du périmètre irrigué. Notons que le déploiement de cette méthodologie est actuellement conditionné à la capacité de récupérer les données pédologiques disponibles à un coût raisonnable. Bien que notre étude ait montré que cette récupération avait des coûts sans commune mesure avec le coût d'un nouveau profil de sol, une automatisation, au moins partielle, de cette récupération serait nécessaire. Des procédures de dématérialisation des données (Yang & Yang, 2017), actuellement pratiquées dans le monde industriel pourraient être intéressantes à mettre en œuvre.

Annexe

Annexe 1 : Notice explicative de la saisie des données pédologiques anciennes

NOTICE EXPLICATIVE DE LA SAISIE DES DONNEES PEDOLOGIQUES ANCIENNES

PAR QUENTIN STYC

I. Présentation et estimation du réservoir utile

La réserve utile est un indicateur dépendant de plusieurs propriétés primaires de sols et intégrant plusieurs profondeurs (i.e., horizons pédologiques). Son estimation dépend :

- Teneur en éléments grossiers : plus un sol sera riche en éléments grossiers et plus la portion de terre fine susceptible de retenir l'eau sera faible et donc une réserve utile faible
- Composition de la terre fine : les fines particules (argile/limon) sont capables de mieux retenir l'eau que les grosses particules (sable)
- Profondeur : un sol profond est capable de stocker plus d'eau qu'un sol superficiel (à composition de terre fine et teneur en éléments grossiers équivalent).

Le protocole mis en place pour estimer le RU utilise la formule suivante :

$$RU = \sum_{i=1}^{n} dh_i * bd_i * \left(\frac{100 - st_i}{100}\right) * (b_i * EqW_i)$$

Où EqW_i est l'humidité équivalente correspondant approximativement au terme θr_i and b_i est le coefficient de texture qui exprime l'humidité au point de flétrissement permanant en pondérant EqW_i pour prendre en compte la teneur en eau qui n'est plus disponible par la plante, c'est-à-dire, au-delà du point de flétrissement permanent, st_i les éléments grossiers et dh_i l'épaisseur de l'horizon et bd_i la densité apparente.

II. Présentation des données pédologiques anciennes

II.1. Profil de sol

Les profils de sol sont issus des descriptions de fosses pédologiques. Leur profondeur de prospection est fixée à 2,20 m (dans le cadre d'étude des potentialités viticoles) pour une profondeur effective d'observation estimée à 1,40 m. Les profils de sol sont fournis avec un plan de situation et une photographie de la fosse pédologique (utilisation pour photo-interprétation).

Les données d'intérêts pour pouvoir estimer le RU sont :

- **Les informations géographiques : nom de la commune et coordonnées longitude et latitude en Lambert III (Zone Sud),**
- **Les profondeurs entre lesquelles sont réalisés les prélèvements correspond à une couche homogène de sol**
- **Pourcentages pondéraux (%) des éléments grossiers ayant un diamètre supérieur à 20 mm, et compris entre 2 et 20 mm et supérieur à 2 mm (= pourcentage total des éléments grossiers) par rapport à la masse de l'ensemble éléments grossiers et terre fine.**
- **Pourcentage pondérale de terre fine (%) estimée par la différence avec le pourcentage massique des éléments grossiers.**
- **La densité apparente, Δa (en kg.dm^{-3}), correspond au poids de terre fine contenu dans une unité de volume. La densité est mesurée par un prélèvement de sol non remanié par cube Vergières (Bourrier, 1965). Il est possible que cette mesure ne soit pas réalisable selon l'abondance des éléments grossiers. Dans ce cas, la densité apparente est estimée comme $1.6 * \% Terre\ fine$ (où 1.6 correspond à la densité moyenne d'une terre fine sans cailloux).**
- **Humidité équivalente (H.éq) ou coefficient de rétention (Cr) (en g.100^{-1}) selon les fiches de sol. Cette humidité correspond à l'eau retenue par un échantillon de terre fine (tamisée à 2 mm) soumis à un champ de force de 1000 fois l'accélération de la pesanteur par une centrifugeuse, équivalent à une pression de -100 kPa appliquée sur l'échantillon. L'humidité équivalente est une approximation de l'humidité à la capacité au champ, qui était anciennement mesurée à -100 kPa (pF = 3) (Baize and Jabiol, 1995).**
- **Classe texturale déterminée à partir des résultats d'analyse granulométrique replacés dans le triangle des textures**
- Le coefficient textural déterminée selon la classe texturale de l'horizon (Table 1)

Table 1. Valeurs moyennes des humidités équivalentes, des densités apparentes et des coefficients texturaux à partir des classes texturales

Classe texturale		Humidité équivalente ($g.100^{-1}g$)	Densité apparente ($kg.dm^{-3}$)	Coefficient textural
GEPPA	*BRL*			
S pur	-	5	1,7	0,8
SS	SF/SG/SFL	8	1,6	0,7
S	SL	12	1,6	0,6
Sl	LS	14	1,6	0,55
Ls	L	15	1,5	0,45
LL	LP	21	1,4	0,5
Sa	SA	16	1,6	0,5
Sal	-	17	1,6	0,45
LSa	L.Fr	19	1,5	0,45
L	-	23	1,4	0,45
AS	LAS	23	1,5	0,45
LAS	-	24	1,5	0,45
LA	-	25	1,4	0,45
As	AS	27	1,5	0,45
Als	L.Fr.A	28	1,5	0,45
Al	AL	32	1,4	0,45
A	A	35	1,3	0,4
AA	-	45	1,2	0,33

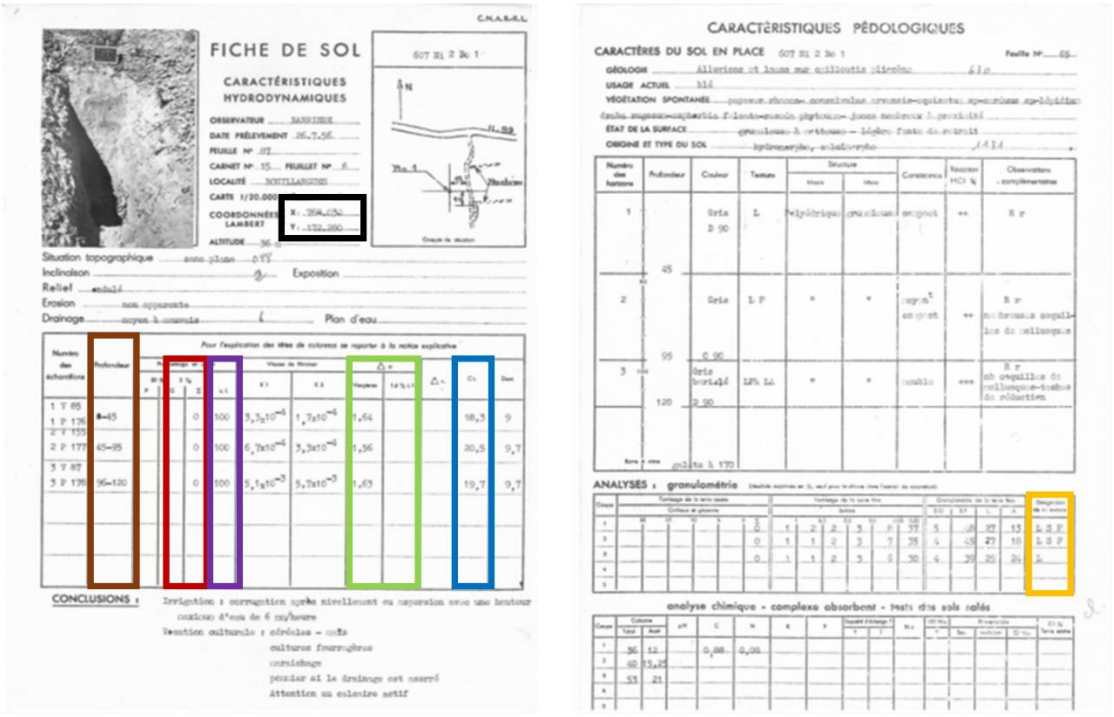


Figure 1. Exemple d'un profil de sol de la commune de Bouillargues

II.2. Sondages pédologiques

Les sondages pédologiques ont été réalisés à la tarière à main. Les profondeurs de prospection et d'observation de cet outil de mesure sont égales à 1,20 m. A la différence des profils de sol, les coordonnées spatiales des sondages pédologiques ne sont pas directement disponibles et un géoréférencement manuel est nécessaire.

Les sondages pédologiques présentent uniquement une description des caractéristiques pédologiques des horizons de sol (Figure 2):

- **Profondeur des horizons**
- **Pourcentage massique des graviers (ø entre 2-20 mm) et cailloux (ø > 20 mm) estimés visuellement**
- **Pourcentage pondéral de terre fine calculé par la différence avec les pourcentages pondéraux cumulés des graviers et des cailloux**
- **Appréciation de la terre fine sur terrain par expertise pédologique permettant l'obtention d'une classe texturale selon le triangle des textures BRL ou un indicateur de classe variant entre 1 et 5, où 1 = terre sableuse et 5 = terre argileuse,**
- **La densité apparente,**
- **L'humidité équivalente**

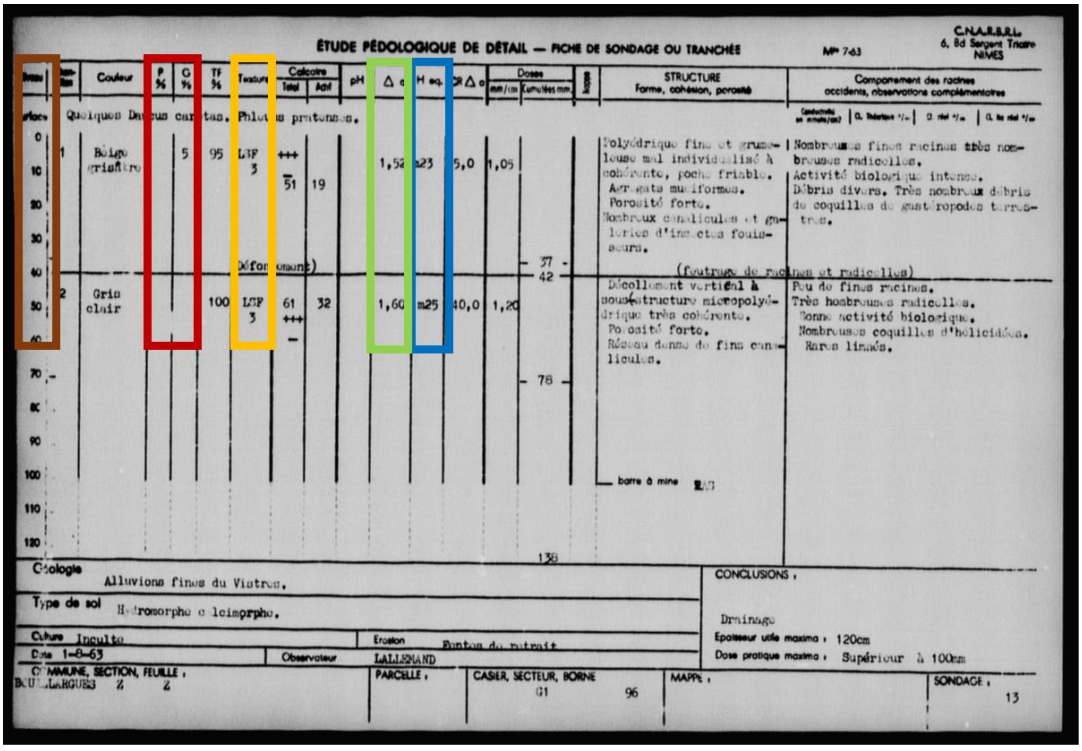


Figure 2. Exemple d'une description d'un sondage pédologique

Afin d'estimer le RU, l'humidité spécifique, la densité apparente ainsi que le coefficient textural ont été déterminés en fonction de la classe texturale de l'horizon (Table 1).

III. Géoréférencement

III.1. Profils de sol

La saisie des données de localisation sous forme de coordonnées spatiales en Lambert III Zone Sud est directement accessible mais nécessite une conversion pour être exprimée en Lambert 93 (projection officielle pour les cartes Françaises). L'estimation du temps de saisie d'un profil est de 0.8 min (48 s) pour la saisie des données sols et 0.2 (12 sec) pour les coordonnées spatiales.

III.2. Sondages pédologiques

Les sondages de sol sont disponibles sur microfiches, support de stockage qui était couramment utilisé par BRL Exploitation de 1957 à 1992. Ces microfiches contiennent des copies des descriptions de sondages (encadré bleu de la Figure 3) accompagnées de leur plan de localisation (encadré orange de la Figure 3) pour un périmètre défini. Ces informations ont été scannées au préalable par BRL Exploitation. Les sondages ne fournissent qu'une localisation relative (sans coordonnées spatiales) qui se reporte à un système de localisation basé sur le schéma de structuration du réseau hydraulique régional BRL (RHR).

Géoréférencement selon la structuration du RHR

Le plan d'échantillonnage et les descriptions des sondages contiennent un identifiant de la localisation relative de la borne de livraison en eau d'irrigation à laquelle est rattachée le périmètre défini. Cet identifiant est élaboré selon le schéma de structuration du RHR qui se scinde en trois niveaux :

- Casier : entité géographique regroupant des ensembles hydrauliques continus du RHR (Figure 4.a),
- Secteur : sous entité géographique situé au sein des casiers, correspondant à des unités desservies par une seule station de pompage et de mise en pression (Figure 4.b),
- Borne : point physique de livraison en eau d'irrigation (Figure 4.c).

Le format renseigné sur les microfiches est composé dans l'ordre des numéros de casier, de secteur et de la borne (ex : « 01-G1-092 » = Casier 01, Secteur G1, Borne 092) (Figure 3).

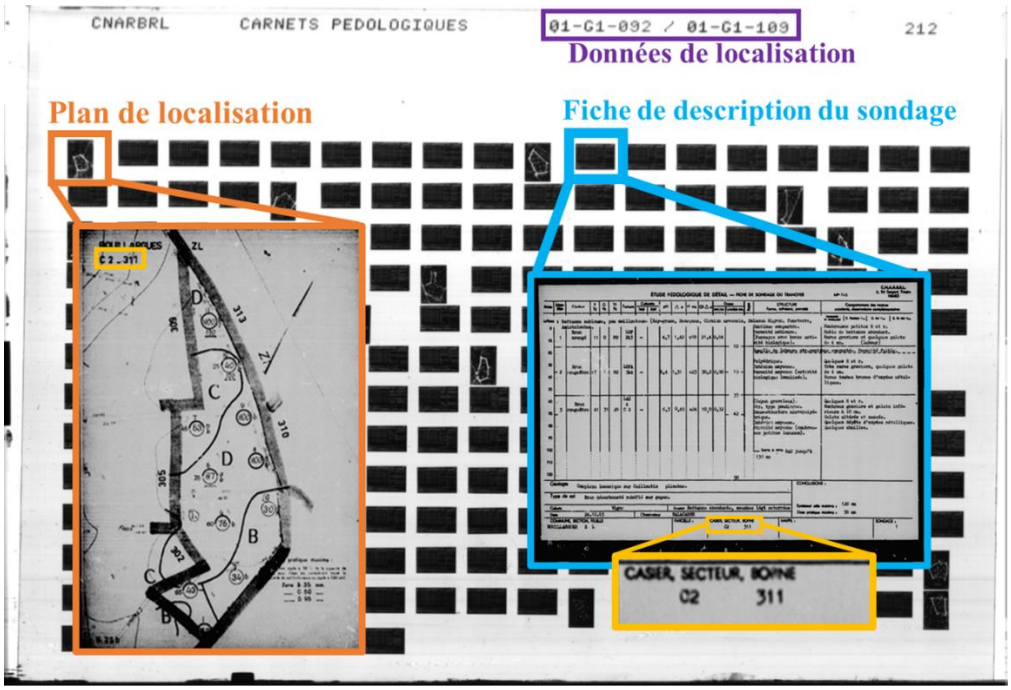


Figure 3. Présentation d'une microfiche répertoriant les plans de localisation (orange), les fiches de descriptions des sondages (bleu) et les données de localisation (violet)

A partir de cet identifiant, il est possible de la localiser sur l'application web-SIG de BRL Exploitation, GéEauWeb.

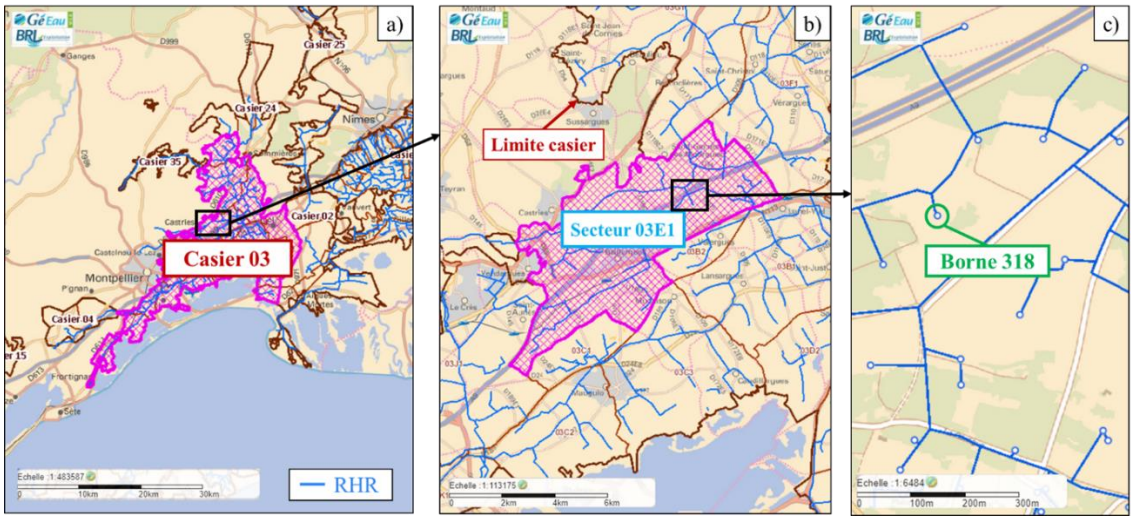


Figure 4. Présentation de la structuration du RHR composé de : a) les casiers, b) les secteurs et c) les bornes de livraison en eau d'irrigation

Parmi les couches géographiques disponibles sur GéEauWeb (Figure 6), un ancien cadastre géoréférencé (Figure 5.b), élaboré pendant la récolte des données pédologique BRL, permet de transposer le cadastre utilisé sur le plan de localisation des sondages sur GéEauWeb.

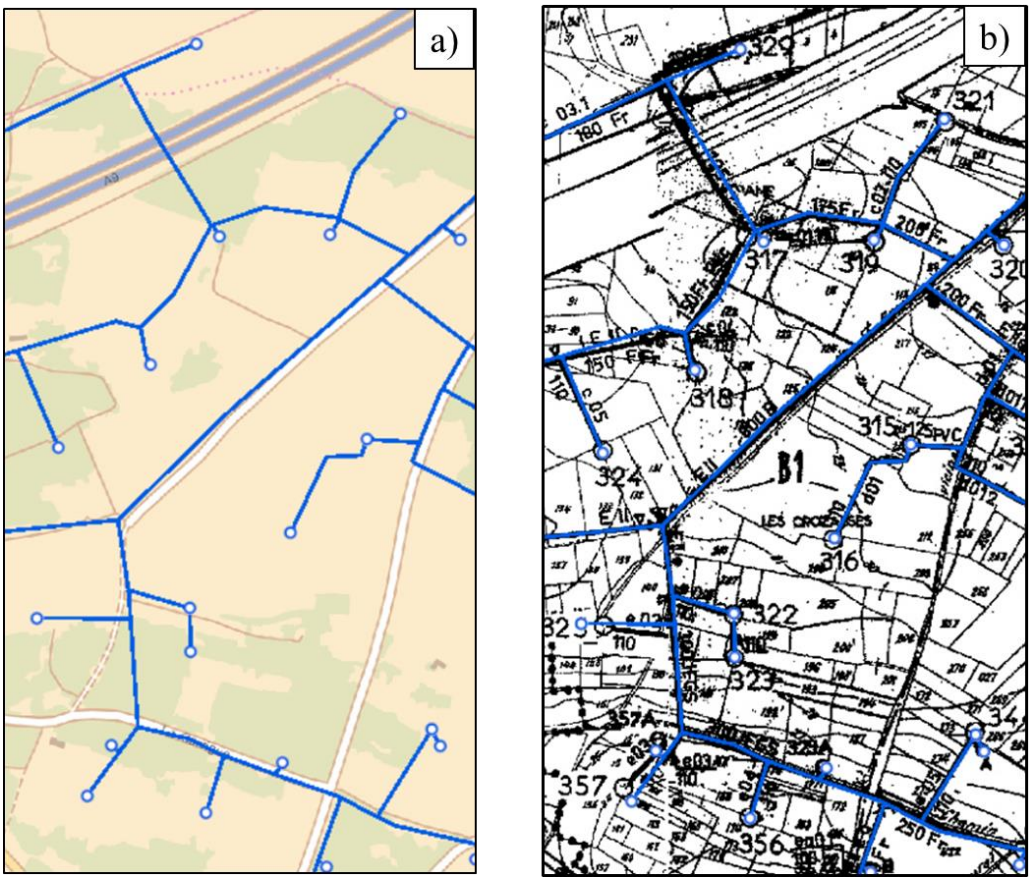


Figure 5. Illustration de l'ancien cadastre géoréférencé sur GéEauWeb

Ensuite, à partir d'une interprétation visuelle permettant de rattacher les limites du périmètre d'étude (Figure 7.a) aux voies de communications et limites de parcelles du cadastre géoréférencé (Figure 7.b), il est possible de mesurer les coordonnées spatiales (Lambert 93) avec l'outil « Localisation » de GéEauWeb (Figure 8) en procédant au placement approximatif des sondages (Figure 7.c).

Figure 6. Couche SIG du cadastre ancien géoréférencé

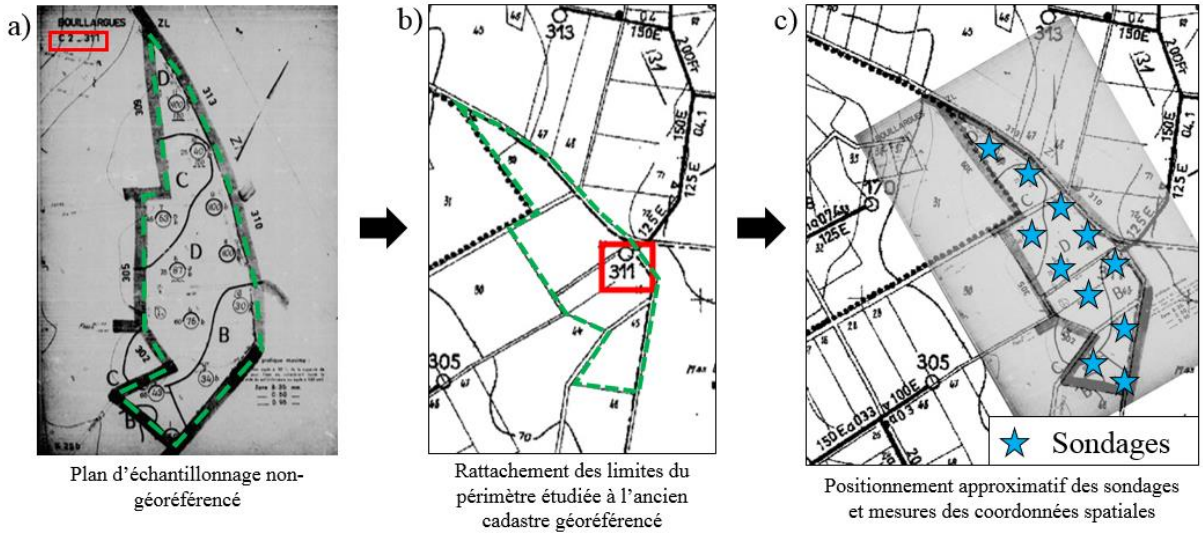


Figure 7. Rattachement du plan d'échantillonnage non-géoréférencé au l'ancien cadastre géoréférencé et mesure des coordonnées spatiales des sondages

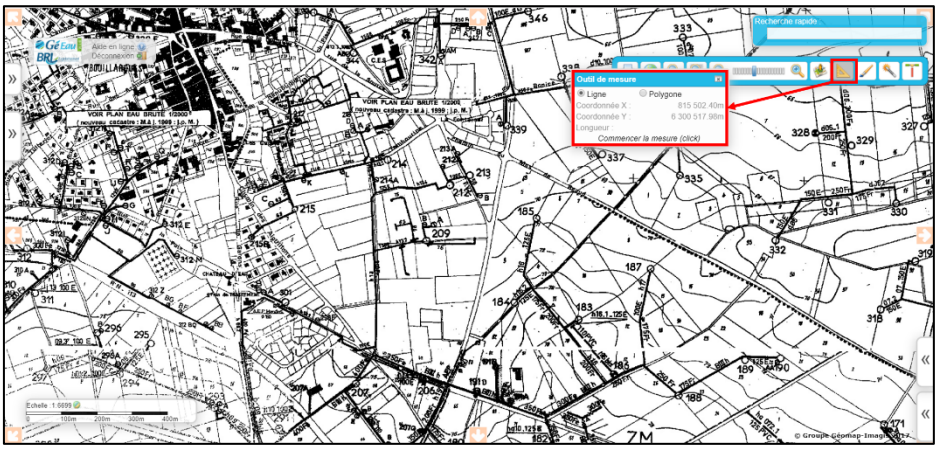


Figure 8. Outil de localisation (Lambert 93)

Estimation du temps de saisie

La saisie des propriétés de sol intervenant dans l'estimation du RU est effectuée similairement aux profils de sol. Le temps de saisie des propriétés de sols par sondage est estimé à 0.2 min (12 s) et celui du géoréférencement à 0.3 min (18 s).

Géoréférencement selon le système mappillon

Les données pédologiques de BRL Exploitation peuvent aussi être localiser selon un système de rangement des mappes, unité surfacique régulière (Figure 9).

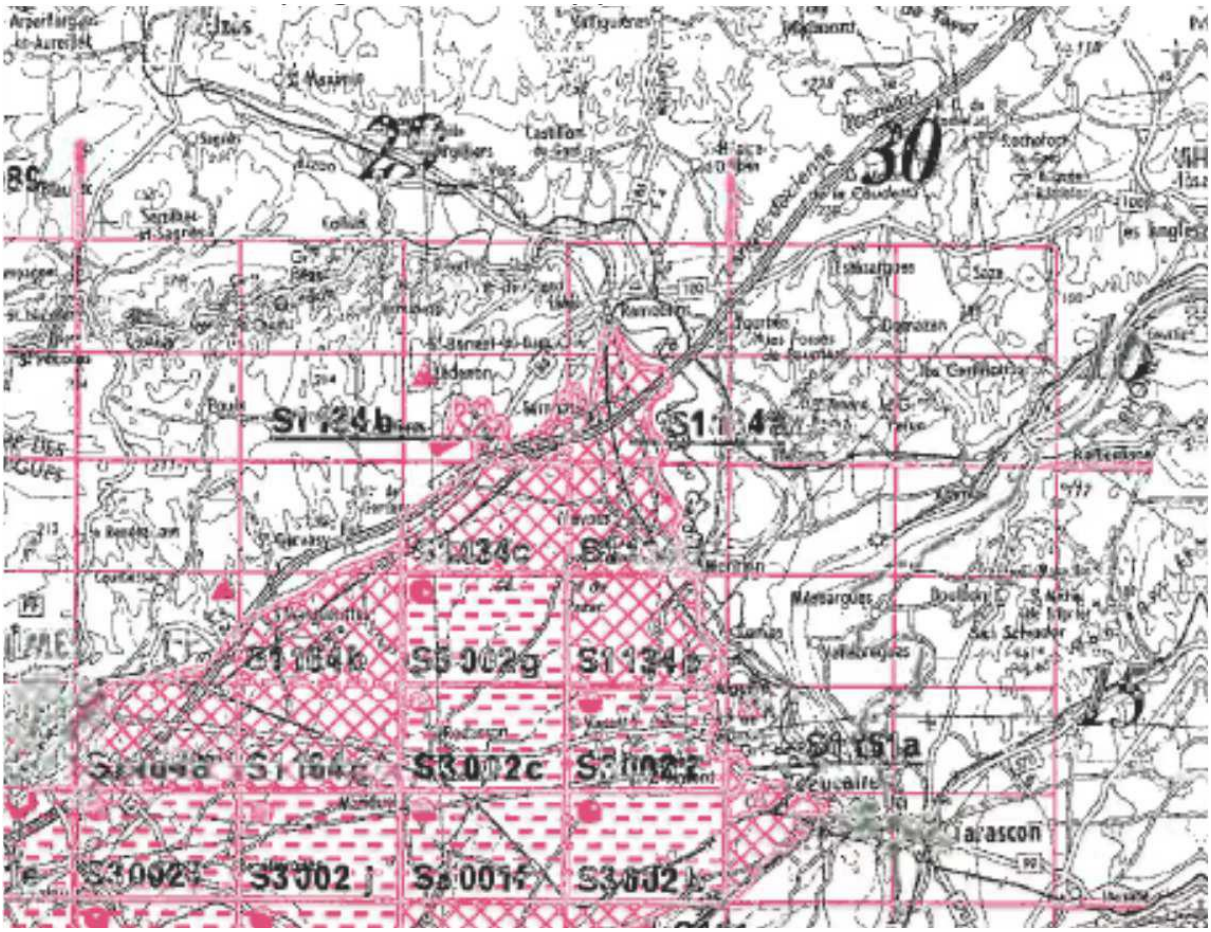


Figure 9. Représentation d'une mappe (unité rose)

Les mappes sont divisées par un découpage régulier identifier par un identifiant longitudinale et un identifiant latitudinale ainsi qu'un identifiant désignant l'emplacement dans la mappe (Figure 10).

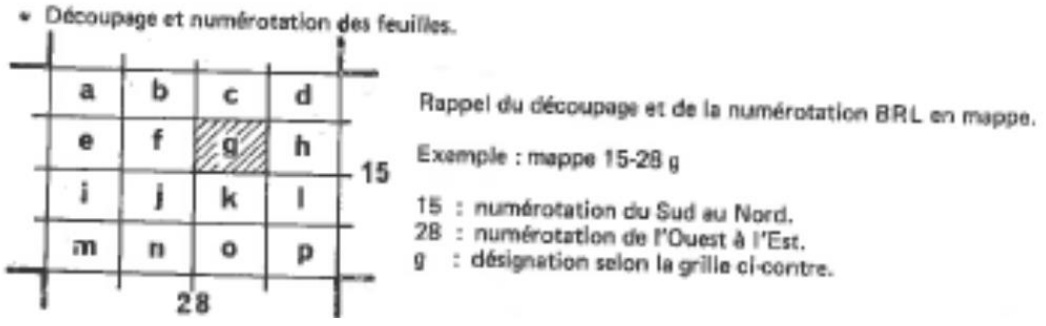


Figure 10. Composition d'une mappe

A partir des données disponibles sur la microfiche, il est possible de déterminer l'emplacement de la zone selon la mappe (Figure 11, Figure 12.a).

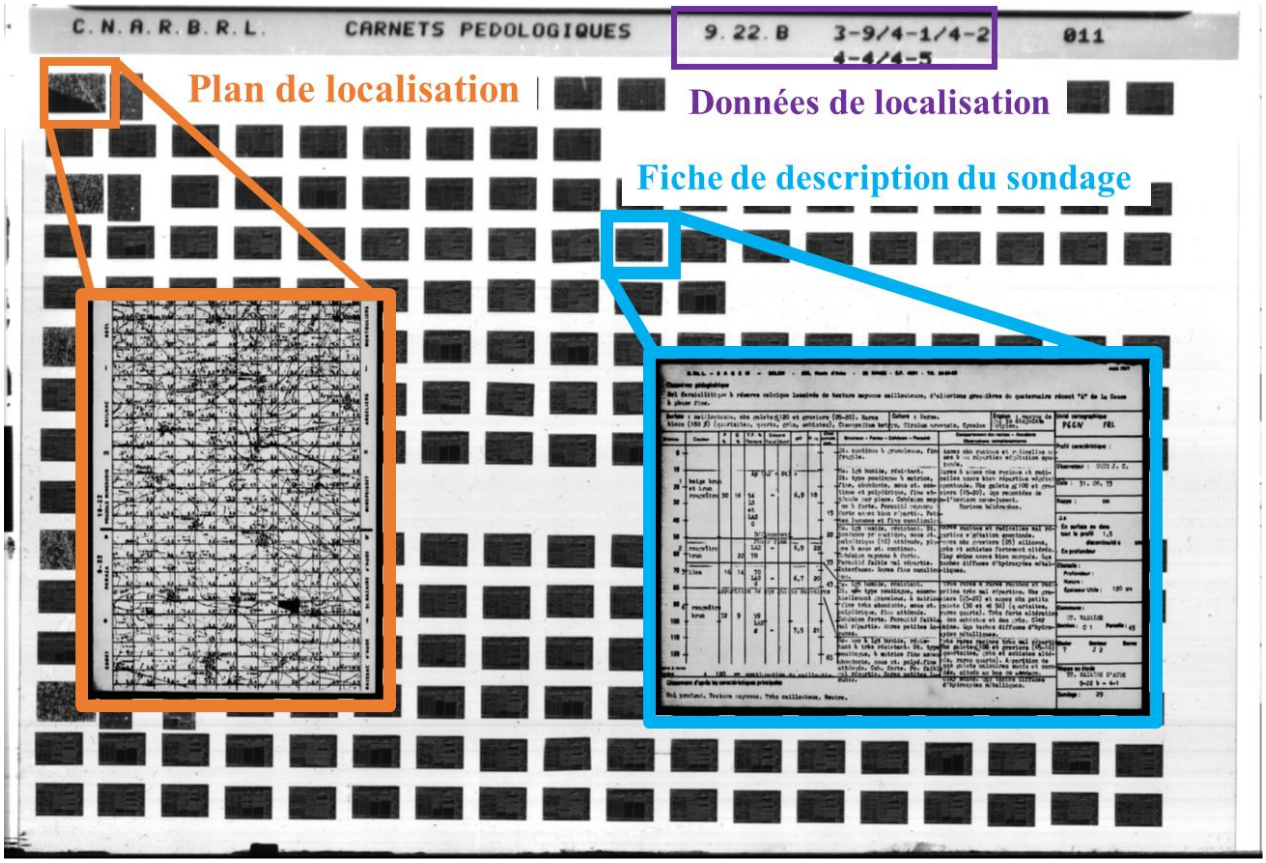


Figure 11. Présentation d'une microfiche répertoriant les plans de localisation (orange, mappe), les fiches de descriptions des sondages (bleu) et les données de localisation (violet)

En sélectionnant la subdivision de la mappe 9.22 B les identifiants latitude et longitude ainsi que la désignation de l'emplacement dans la mappe (encadré violet transparent, Figure 12b). Cette sous-division de la mappe est elle-même divisée en 36 sous unités appelées « mappillons » (6 x 6). A partir de l'identifiant du mappillon fourni (encadré vert, Figure 12a), il est possible de repérer visuellement la localisation du mappillon qui correspond à la zone d'étude (Figure 12.b et c).

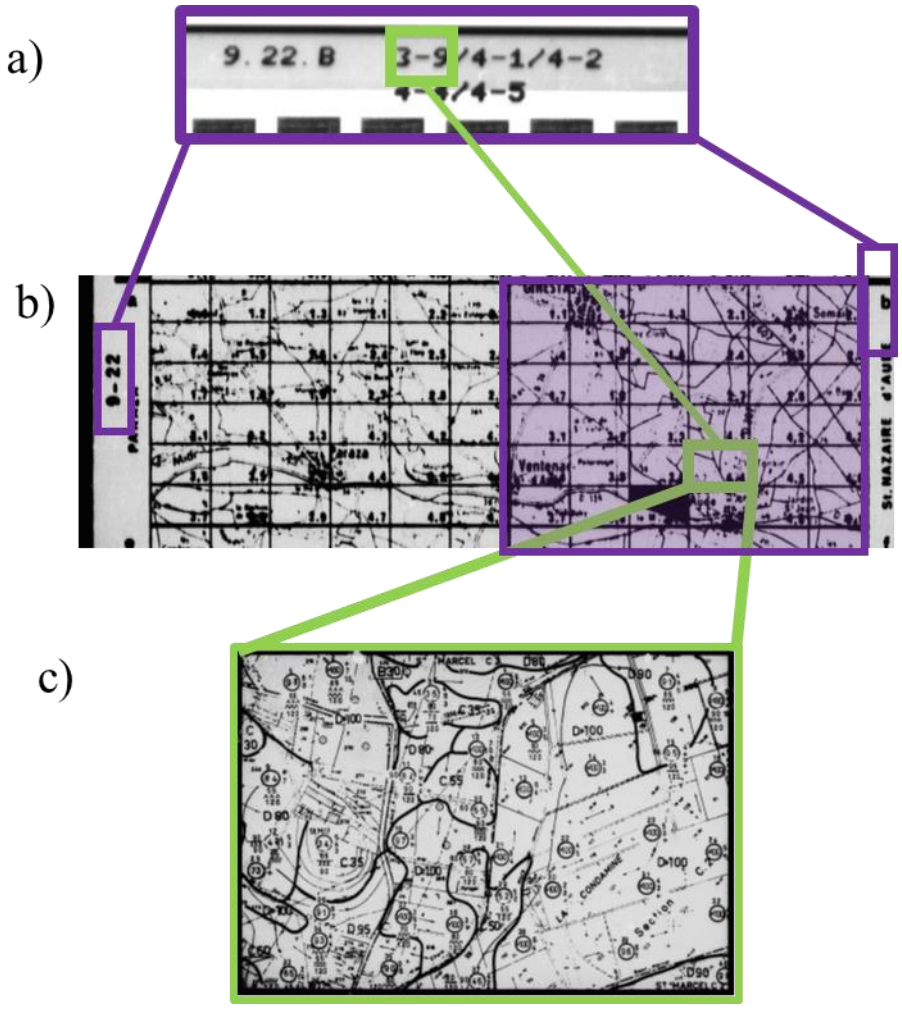


Figure 12. a) récupération de la localisation de la zone d'étude selon le système de localisation de la mappe, b) replacement de ses informations dans la mappe afin de retrouver les mappillons où est localisé la zone d'étude et c) replacement de la zone dans la mappe

Une fois ces informations obtenues il est possible de replacer la localisation suivant la même procédure que pour la localisation du réseau RHR en localisant par tâtonnement selon la localisation générale fournie par les mappes et le plan de localisation des mappes.

Références

Adhikari, K., Hartemink, A.E., (2016). Linking soils to ecosystem services - A global review. Geoderma 262, 101–111.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B., Csaki, F. (Eds.), Second International Symposium on Information Theory. Akademia Kiado, Budapest, Hungary, pp. 267–281

Al Majou, H., Bruand, A., Duval, O. 2008a. Use of in situ volumetric water content at field capacity to improve prediction of soil water retention properties. Can. J. Soil Sci., 88(4), 533–541.

Al Majou, H., Bruand, A., Duval, O., Le Bas, C. and Vautier, A., 2008. Prediction of soil water retention properties after stratification by combining texture, bulk density and the type of horizon. Soil Use and Management, 24(4): 383-391

Al Majou, H., Bruand, A., Duval, O., Cousin, I. (2007). Comparaison de fonctions de pédotransfert nationales et européennes pour prédire les propriétés de rétention en eau des sols. Étud. Gest. Sols, 14 (2) (2007), pp. 103-116.

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

Algayer, B., Lagacherie, P., Lemaire, J., 2020. Adapting the available water capacity indicator to forest soils : An example from the Haut-Languedoc (France). Geoderma 357, 113962.

Allen, R.G., Pereira, L.S., Raes, D. and Smith, M., 1998. Crop evapotranspiration: guidelines for computing crop water requirements. FAO Irrigation and Drainage Paper (56): xxvi + 300 pp.-xxvi + 300 pp

Amirian-Chakan, Alireza, Budiman Minasny, Ruhollah Taghizadeh-Mehrjardi, Rokhsar Akbarifazli, Zahra Darvishpasand, and Saheb Khordehbin, 2019. Some Practical Aspects of Predicting Texture Data in Digital Soil Mapping. Soil and Tillage Research 194:104289.

Arnold, J. G., and J. R. Williams. 1987. Validation of SWRRB: Simulator for water re- sources in rural basins. J. Water Resour. Plan. Manage. ASCE 113(2): 243–256

Arnold, J. G., and N. Fohrer. 2005. SWAT2000: current capabilities and research op- portunities in applied watershed modeling. Hydrol. Process. 19(3): 563–572.

Arrouays, D. and Jamagne, M., 1993. Sur la possibilité d'estimer les propriétés de rétention en eau de sols limoneux lessivés hydromorphes du sud-ouest de la France à partir de leurs caractéristiques de constitution. Compte Rendu de l'Académie d'Agriculture de France, 79(1): 111-121

Arrouays, D., Grundy, M.G., Hartemink, A.E., Hempel, J.W., Heuvelink, G.B., Hong, S.Y., Lagacherie, P., Lelyk, G., McBratney, A.B., McKenzie, N.J. and dL Mendonca-

Santos, M., 2014. GlobalSoilMap: Toward a fine-resolution global grid of soil properties. In Advances in agronomy (Vol. 125, pp. 93-134). Academic Press.

Arrouays, D., Lagacherie, P., Hartemink, A.E., 2017. Digital soil mapping across the globe. Geoderma Reg. 9.

Arrouays, D., Leenaars, J. G. B., Richer-de-Forges, A. C., Adhikari, K., Ballabio, C., Greve, M., … Rodriguez, D. 2017. Soil legacy data rescue via GlobalSoilMap and other international and national initiatives. GeoResJ.

Arrouays, D., McBratney, A.B., Minasny, B., Hempel, J.W., Heuvelink, G.B.M., MacMillan, R.A., Hartemink, A.E., Lagacherie, P., McKenzie, N.J., 2014b. The GlobalSoilMap project specifications, in: Arrouays, D., McKenzie, N.J., Hempel, J.W., Richer-de-Forges, A.C., McBratney, A.B. (Eds.), globalSoilMap: Basis of the Global Spatial Soil Information System. CRC Press, Taylor&Francis, Boca Raton, USA. pp. 9–12.

Assouline, S. and Or, D., 2014. The concept of field capacity revisited: Defining intrinsic static and dynamic criteria for soil internal drainage dynamics. Water Resources Research, 50(6): 4787-4802

B.P. Malone, A.B. McBratney, B. Minasny, G.M. Laslett, (2009). Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma, 154 (1–2), pp. 138-152

Baize, D., & Jabiol, B. (1995). Guide des sols, Quae.

Barthès, J.P., Bornand, M., Falipou, P., 1999a. Référentiel Pédologique de la France: Pédopaysages de l'Hérault. Echelle au 1/250000ème. INRA - ENSA. Science du Sol Montpellier, Montpellier.

Barthès, J.P., Bornand, M., Falipou, P., 1999b. Référentiel Pédologique de la France: Pédopaysages de la Lozère. Echelle au 1/250000ème. INRA - ENSA. Science du Sol Montpellier, Montpellier.

Barthès, J.P., Bornand, M., Falipou, P., 1999c. Référentiel Pédologique de la France: Pédopaysages du Gard. Echelle au 1/250000ème. INRA - ENSA. Science du Sol Montpellier, Montpellier.

Barthès, J.P., Bornand, M., Falipou, P., 1999d. Référentiel Pédologique de la France: Pédopaysages des Pyrénées Orientales. Echelle au 1/250000ème. INRA - ENSA. Science du Sol Montpellier, Montpellier

Barthès, J.P., Bornand, M., Falipou, P., 1999e. Référentiel Pédologique de la France : Pédopaysages de l'Aude. Echelle au 1/250000ème. INRA - ENSA. Science du Sol Montpellier, Montpellier

Batjes, N.H., 1996. Development of a world data set of soil water retention properties using pedotransfer rules. Geoderma, 71(1-2): 31-52.

Batjes, N.H., 2016. Harmonized soil property values for broad-scale modelling (WISE30sec) with estimates of global soil carbon stocks. Geoderma, 269: 61-68.

Baume, O., Skøien, J.O., Heuvelink, G.B.M., Pebesma, E.J., 2011. A geostatistical approach to data harmonization — application to radioactivity exposure data. International Journal of Applied Earth Observation and Geoinformation 13 (3), 409–419.

Bell J.C., Cunningham R.L., and Havens M.W., 1992 - Calibration and Validation of a Soil-landscape Model For Predicting Soil Drainage Class. Soil Science Society of America Journal, 56, 1860-1866.

Berger, E., 1976. Partitioning the parameters of stony soils, important in moisture determinations in to their constituents. Plant and Soil, 44:201-207.

Bezdek, D.J.C., 1974. Numerical taxonomy with fuzzy sets. J. Math. Biol. 1, 57–71.

Biau, G. Scornet, E. TEST. 2016. 25: 197. https://doi.org/10.1007/s11749-016-0481-7

Bishop, T.F.A., Horta, A., Karunaratne, S.B., 2015. Validation of digital soil maps at different spatial supports. Geoderma 241–242, 238 – 249.

Bliss, N.B., Waltman, S.W., Petersen, G.W., 1995. Preparing and soil carbon inventory for the United States using geographical information systems., in: Soil and Global Change. Lal, R. et al, Boca Raton, pp. 275 – 295

Bornand M., 1997 - Connaissance et suivi de la qualité des sols en France. Etat des lieux, enjeux, besoin en données, proposition pour une gestion rai- sonnée de la ressource en sol. Rapport d'expertise du ministère de l'agri- culture, du ministère de l'environnement et de l'Inra. Montpellier. 176p.

Bouma, J., 1989. Using soil survey data for quantitative land evaluation. Adv. Soil Sci. 9, 177–213. doi:https://doi.org/10.1007/978-1-4612-3532-3_4.

Bourennane, H., King, D., Couturier, A., 2000. Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities. Geoderma 97, 255–271

Bourennane, H., Salvador-Blanes, S., Cornu, S., King, D., 2003. Scale of spatial dependence between chemical properties of topsoil and subsoil over a geologically contrasted area (Massif central, France). Geoderma, Pedometrics 2001 112, 235–251

Brantley, S.L., Eissenstat, D.M., Marshall, J.A., Godsey, S.E., Balogh-Brunstad, Z., Karwan, D.L., Papuga, S.A., Roering, J., Dawson, T.E., Evaristo, J., Chadwick, O., McDonnell, J.J., Weathers, K.C., 2017. Reviews and syntheses: on the roles trees play in building and plumbing the critical zone. Biogeosciences 14, 5115–5142. https://doi.org/10.5194/bg-14-5115-2017.

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

BRGM, DREAL, 2013. L'inventaire du patrimoine géologique du Languedoc-Roussillon.

Briggs, L.J. and MacLane, J.W., 1910. Moisture equivalent determinations and their application. American Society of Agronomy Proceedings, 2: 138-147.

Briggs, L.J. and Shantz, H.L., 1912. The Wilting Coefficient and Its Indirect Determination. Botanical Gazette, 53: 20-37

Brisson, N., Mary, B., Ripoche, D., Jeuffroy, M. H., Ruget, F., Nicoullaud, B., et al., 1998. STICS: a generic model for the simulation of crops and their water and nitrogen balances. I. Theory and parameterization applied to wheat and corn. Agron., 18(5–6), 311–346

Bruand, A., duval, O. and Cousin, I., 2004. Estimation des propriétés de rétention en eau des sols à partir de la base de données SOLHYDRO: Une première proposition combinant le type d'horizon, sa texture et sa densité apparente. Etude et Gestion des Sols, 11(3): 323-334.

Bruand, A., Pérez Fernández, P., Duval, O., 2003. Use of class pedotransfer functions based on texture and bulk density of clods to generate water retention curves. Soil Use Manage. 19, 232–242.

Brus, D. J., Kempen, B., & Heuvelink, G. B. M., 2011. Sampling for validation of digital soil maps. European Journal of Soil Science, 62(3), 394–407. https://doi.org/10.1111/j.1365-2389.2011.01364.x

Bui, E.N., Moran, C.J., 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. Estim. Uncertain. Soil Models 103, 79–94

Burrough, P.A., Macmillan, R.A., van Deursen, W., 1992. Fuzzy classification methods for determining land suitability from soil profile observations and topography. J. Soil Sci. 43, 193–210

Cabelguenne, M. and Debaeke, P., 1998. Experimental determination and modelling of the soil water extraction capacities of crops of maize, sunflower, soya bean, sorghum and wheat. Plant and Soil, 202(2): 175-192

Cailleux, A., Taylor, G. 1963. Notice sur le code expolaire. Editions N. Boubee.

Cardenas, G., Malherbe, L., 2003. Evaluation des incertitudes associées aux méthodes géostatistiques

Carré, F., McBratney, A.B., Mayr, T. & Montanarella, L., (2007). Digital Soil Assessment: beyond DSM. Geoderma 142. 69-79.

Cazemier, D., 1999. Utilisation de l'information incertaine dérivée d'une base de données sols : Application à la cartographie des propriétés hydriques à l'échelle régionale. Montpellier, ENSA

Chen, S., Arrouays, D., Mulder, T. Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannan, J., Meersmans, J. and Walter, C., Digital Mapping of Soil Properties at broad scale. Soumis à Earth-Science Reviews

Chiles, J.-P., Delfiner, P., 2009. Geostatistics: modeling spatial uncertainty. John Wiley & Sons

Ciampalini, R., Lagacherie, P., Gomez, C., Grünberger, O., Hamrouni, M. H., Mekki Insaf, I., & Richard, A. 2013. Detecting, correcting and interpreting the biases of measured soil profile data: A case study in the cap bon region (Tunisia). Geoderma, 192(1), 68–76.

Ciampalini, R., Lagacherie, P., Hamrouni, H., 2012a. Documenting GlobalSoilMap.net grid cells from legacy measured soil profile and global available covariates in Northern Tunisia, in: Digital Soil Assessments and Beyond. CRC Press, pp. 439–444.

Ciampalini, R., Lagacherie, P., Monestiez, P., Walker, E., Gomez, C., 2012b. Co-kriging of soil properties with Vis-NIR hyperspectral covariates in the Cap Bon region (Tunisia), in: Digital Soil Assessments and Beyond. CRC Press, pp. 393–398

Cole, N.J., Boettinger, J.L., 2006. Chapter 27 Pedogenic Understanding raster classification methodology for mapping soils, powder river basin, Wyoming, USA, in: P. Lagacherie, A.B.M. and M.V. (Ed.), Developments in Soil Science, Digital Soil Mapping: An Introductory Perspective. Elsevier, pp. 377–619

Colman, E., 1947. A laboratory procedure for determining the field capacity of soils. Soil Science, 63(4): 277-283.

Combres, J. C., Le Mezo, L., Mete, M., & Bourjon, B. (1999). Réserve utile et mesures d'humidité. Difficulté de calage des modèles de bilan hydrique

Coulouma, G., Prevot, L., & Lagacherie, P. 2020. Carbon isotope discrimination as a surrogate for soil available water capacity in rainfed areas: A study in the Languedoc vineyard plain. Geoderma, 362, 114-121.

Cousin, I., Buis, S., Lagacherie, P., Doussan, C., Le Bas, C. The Available Water Capacity: a multidisciplinary and multiscale point of view. Soumis à Geoderma.

Cousin, I., Nicoullaud, B. and Coutadeur, C., 2003. Influence of rock fragments on the water retention and water percolation in a calcareous soil. Catena, 53(2): 97-114.

Cressie, N., 1990. The origins of kriging. Math. Geol. 22, 239–252. Cressie, N., Kang, E.L., 2010. High-resolution digital soil mapping: Kriging for very large datasets, in: Proximal Soil Sensing. Springer, pp. 49–63

Debelmas, J., 1974. Géologie de la France. Doin.

Dharumarajan, S., Hegde, R., Lalitha, M., Kalaiselvi, B. and Singh, S.K., 2019. Pedotransfer functions for predicting soil hydraulic properties in semi-arid regions of Karnataka Plateau, India. Current Science, 116(7): 1237-1246

Dominati,E., Mackay, A., Green, S., Patterson, M. 2014.A soil change-based methodology for the quantification and valuation of ecosystem services from agro-ecosystems: a case study of pastoral agriculture in New Zealand. Ecological Economics. Econ., 100, pp. 119-129

Droogers, P., vanderMeer, F.B.W. and Bouma, J., 1997. Water accessibility to plant roots in different soil structures occurring in the same soil type. Plant and Soil, 188(1): 83-91

Falkenmark, M., Rockström, J. 2006. The new blue and green water paradigm: Breaking new ground for water resources planning and management. Journal of Water Resources Planning and Management, 132(3), 129–132

Furr, J.R. and Reeve, J.O., 1945. Range of soil-moisture percentages through which plants undergo permanent wilting in some soils from semi-arid irrigated areas. Jour. Agr. Res., 71: 149.

Gaudin, R. and Gary, C., 2012. Model-based evaluation of irrigation needs in Mediterranean vineyards. Irrigation Science, 30(5): 449-459

Gomez, C., & Coulouma, G. (2018). Importance of the spatial extent for using soil properties estimated by laboratory VNIR/SWIR spectroscopy: Examples of the clay and calcium carbonate content. Geoderma. https://doi.org/10.1016/j.geoderma.2018.06.006

Goovaerts, P., 1992. Factorial kriging analysis: a useful tool for exploring the structure of multivariate spatial soil information. J. Soil Sci. 43, 597–619

Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89, 1– 45

Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma 103, 3 – 26

Graham, R.C., Rossi, A.M., Hubbert, K.R., 2010. Rock to regolith conversion: producing hospitable substrates for terrestrial ecosystems. GSA Today 20, 4–9.

Gras. R., 1994. Sols caillouteux et production végétale, Paris, 175 pp.

Greiner, L., Keller, A., Grêt-Regamey, A., Papritz, A., (2017). Soil function assessment: review of methods for quantifying the contributions of soils to ecosystem services. Land use policy 69, 224–237.

Grunwald, S., 2009. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma 152, 195–207

Hengl, T. (2019). GSIF : Global Soil Information Facilities. R package version 0.5-5. https://CRAN.R-project.org/package=GSIF.

Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B.M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Gonzalez, M.R., 2014. SoilGrids1km — Global Soil Information based on automated mapping. PLoS ONE 9, e105992

Hengl, T., Heuvelink, G. B., Kempen, B., Leenaars, J. G., Shepherd, K., Sila, A., … Tondoh, J. E. (2015). Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions S1 Regression-kriging in R using the Meuse data set. PlosOne, 10(6), e0125814. https://doi.org/https://doi.org/10.1371/journal.pone.0125814

Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ, 2018(8).

Heuvelink, G.B., 2013. Uncertainty quantification of GlobalSoilMap products. Glob. Basis Glob. Spat. Soil Inf. Syst. 327–332.

Heuvelink, G.B.M., 2014. Uncertainty quantification of GlobalSoilMap products. In: Arrouays, D., McKenzie, N.J., Hempel, J.W., Richer-de-Forges, A.C., McBratney, A.B. (Eds.), GlobalSoilMap: Basis of the Global Spatial Soil Information System. CRC press & Taylor & Francis group, Boca Raton, USA, pp. 327–332.

Heuvelink, G.B.M., Burrough, P.A., Stein, A., 1989. Propagation of errors in spatial modelling with GIS Int. J. Geogr. Inf. Syst. 3, 303–322.

Heuvelink, G.B.M., Pebesma, E.J., 1999. Spatial aggregation and soil process modelling. Geoderma 89, 47–65.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978.

Hong, S. Y., Minasny, B., Han, K. H., Kim, Y., Lee, K., 2013. Predicting and mapping soil available water capacity in Korea. PeerJ, 1:e71

J.H.M. Wösten, A. Lilly, A. Nemes, C.Le Bas., 1999. Development and use of a database of hydraulic properties of European soils. Geoderma, 90, pp. 169-185

Jabro, J.D., Evans, R.G., Kim, Y. and Iversen, W.M., 2009. Estimating in situ soil-water retention and field water capacity in two contrasting soil textures. Irrigation Science, 27(3): 223- 229

Jamagne, M., Betremieux, R., Bégon, J.C. and Mori, A., 1977. Quelques données sur la variabilité dans le milieu naturel de la Réserve en Eau Utile des sols. Bulletin Technique d'Information, 324-325: 627-641

Jenny, H., 1941. Factors of Soil Formation. . McGraw-Hill, New York

Joly, D., Brossard, T., Cardot, H., Cavailhes, J., Hilal, M., Wavresky, P., 2010. Les types de climats en France, une construction spatiale. Eur. J. Geogr.

Joly, S., 2015. Dossier de spécifications fonctionnelles et techniques de l'Infrastructure de données géographiques du Languedoc-Roussillon.

Kerry, R., Goovaerts, P., Rawlins, B.G., Marchant, B.P., 2012. Disaggregation of legacy soil data using area to point kriging for mapping soil organic carbon at the regional scale. Geoderma 170, 347– 358

Kish, L. 1965. Survey sampling. New York, London: John Wiley

Laborczi, A., Szatmári, G., Dezs, A., Pásztor, L., in press. Comparison of soil texture maps synthetized from standard depth layers with directly compiled products. Geoderma.

Lacoste, M., Mulder, V.L., Richer-de-Forges, A.C., Martin, M.P. and Arrouays, D., 2016. Evaluating large-extent spatial modeling approaches: A case study for soil depth for France. Geoderma Regional, 7(2): 137-152.

Lagacherie P., and Voltz M., 2000 - Predicting soil properties over a region using sample information from a mapped reference area and digital elevation data : a conditional probability approach. Geoderma 97 (3-4), pp. 187-208.

Lagacherie, P, 1992. Formalisation des lois de distribution des sols pour automatiser la cartographie pédologique à partir d'un secteur pris comme référence : cas de la petite région naturelle moyenne vallée de l'Hérault. Thèse de doctorat de l'université des sciences et techniques du Languedoc, Montpellier II. 227 pages.

Lagacherie, P. and McBratney, A.B., 2007. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. in Digital Soil Mapping: an introductory perspective. P. Lagacherie, A. McBratney and M. Voltz. Amsterdam, Elsevier: 3-24.

Lagacherie, P., 2008. Digital soil mapping: A state of the art, in: Hartemink, A.E., McBratney, A.B., Mendonca-Santos, M.D. (Eds.), Digital Soil Mapping with Limited Data. Springer, pp. 3–14.

Lagacherie, P., Andrieux, P., Bouzigues, R., 1996. Fuzziness and uncertainty of soil boundaries: from reality to coding inGIS. Geogr. Objects Indeterminate Boundaries 2, 275

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., & Nkuba-Kasanda, L. Analysing the impact of soil spatial sampling on the performances of Digital Soil Mapping models and their evaluation: a numerical experiment using clay contents obtained from Vis-NIR-SWIR hyperspectral imagery. In press Geoderma.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M., Saby, N. (2019). How far can the uncertainty on a Digital Soil Map be known?: A numerical experiment using pseudo values of clay content obtained from Vis-SWIR hyperspectral imagery. Geoderma.

Lagacherie, P., Arrouays, D., Walter, 2013. Cartographie numérique des sols : principe, mise en œuvre et potentialités. Etude Gest. Sols, Etude et Gestion des Sols 20, 83 – 98

Lagacherie, P., Depraetere, C., 1989. Analyse des relations sol-paysage au sein d'un secteur de référence en vue d'un zonage pédologique semi-automatisé d'une petite région naturelle

Lagacherie, P., Legros, J.P., 1992. Formalisation des lois de distribution des sols pour automatiser la cartographie pédologique à partir d'un secteur pris comme référence = Automatization of soil mapping using laws governing the distribution of soils, established from a reference area [WWW Document]. URL http://cat.inist.fr/?aModele=afficheN&cpsidt=154826 (accessed 8.13.15)

Lagacherie, P., Legros, J.P., Burfough, P.A., 1995. A soil survey procedure using the knowledge of soil pattern established on a previously mapped reference area. Geoderma 65, 283–301

Lagacherie, P., McBratney, A.B., Voltz.,M. (2007) Digital Soil Mapping :an introductory perspective. Developments in Soil Science, vol. 31. Elsevier Amsterdam 400 pages Amsterdam, Elsevier.

Lagacherie, P., Sneep, A.-R., Gomez, C., Bacha, S., Coulouma, G., Hamrouni, M.H., Mekki, I., 2013. Combining Vis–NIR hyperspectral imagery and legacy measured soil profiles to map subsurface soil properties in a Mediterranean area (Cap-Bon, Tunisia). Geoderma 209–210, 168–176

Lark, R.M., Cullis, B.R., Welham, S.J., 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. Eur. J. Soil Sci. 57, 787–799.

Laslett, G.M., McBratney, A.B., 1990. Further Comparison of Spatial Methods for Predicting Soil pH. Soil Sci. Soc. Am. J. 54, 1553

Leenaars, J.G.B., Claessens, L., Heuvelink, G.B.M., Hengl, T., Ruiperez González, M., van Bussel, L.G.J., Guilpart, N., Yang, H., Cassman, K.G., 2018. Mapping rootable depth and root zone plant-available water holding capacity of the soil of sub-Saharan Africa. Geoderma 324, 18–36.

Leenhardt, D. , Voltz, M. , Bornand, M. and Webster, R., (1994). Evaluating soil maps for prediction of soil water properties. European Journal of Soil Science, 45: 293-301. doi:10.1111/j.1365-2389.1994.tb00512.x

Leenhardt, D., Voltz, M., Bornand, M., Webster, R., 1994. Evaluating soil maps for prediction of soil water properties. Eur. J. Soil Sci. 45, 293–301.

Legros, J. P. 1996. Cartographies des sols: de l'analyse spatiale à la gestion des territoires. In Collection Gérer l'environnement. Retrieved from https://books.google.fr/books?id=MiONzDc-jnQC

Liu, J., Zehnder, A. J. B., & Yang, H. (2009). Global consumptive water use for crop production: The importance of green water and virtual water. Water Resources Research, 45(5), 1–15

Mallavan, B., Minasny, B., McBratney, A., 2010. Homosoil, a methodology for quantitative extrapolation of soil information across the globe, in: Digital Soil Mapping. Springer, pp. 137– 150.

Malone, B.P., McBratney, A.B., Minasny, B., Laslett, G.M., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154, 138–152

Mansy, J.-L., Guennoc, P., Robaszynski, F., Amédro, F., Auffret, J.-P., Vidier, J.-P., Lamarche, J., Lefevre, D., Somme, J., Brice, D., Mistiaen, B., Prud'Homme, A., Rohart, J.-C., Vachard, D., (2008). Notice explicative de la carte géologique de la France (1/50000)., BRGM. ed.

Martonne, E. de, 1927. Regions of Interior-Basin Drainage. Geogr. Rev. 17, 397–414.

Martin, T.G., Wintle, A., Rhodes, J.R., Field, A., Low-choy, S.J., 2005. Zero tolerance ecology : improving ecological inference by modelling the source of zero observations. Ecol. Lett. 1235–1246.

McBratney A.B., Mendonca Santos M.L., and Minasny B., 2003 - On digital soil mapping. Geoderma, 117, pp. 3-52.

Mcbratney, A., Field, D.J., Koch, A., 2014. Geoderma. The dimensions of soil security. Geoderma 213, 203–213.

McBratney, A.B., de Gruijter, J.J., 1992. A continuum approach to soil classification by modified fuzzy k-means with extragrades. J. Soil Sci. 43, 159–175.

McBratney, A.B., Minasny, B., Cattle, S.R., Vervoort, R.W. (2002). From pedotransfer functions to soil inference systems. Geoderma, 109, pp. 41-73, 10.1016/S0016-7061(02)00139-8

McBratney, A.B., Odeh, I.O.A., 1997. Application of fuzzy sets in soil science: fuzzy logic, fuzzy measurements and fuzzy decisions. Geoderma 77, 85–113

McKenzie N.J., and Austin M.P., 1993 - A quantitative Australian approach to medium and small scale survey based on soil stratigraphy and environmental correlation. Geoderma, 57, pp. 329-355.

McKenzie, N.J., Gallant, J.C., 2006. Chapter 24 Digital Soil Mapping with Improved Environmental Predictors and Models of Pedogenesis, in: P. Lagacherie, A.B.M. and M.V. (Ed.), Developments in Soil Science, Digital Soil Mapping: An Introductory Perspective. Elsevier, pp. 327–349.

McKenzie, N.J., Ryan, P.J., 1999. Spatial prediction of soil properties using environmental correlation. Geoderma 89, 67–94

Meinshausen, N. 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.

Merot P., Ezzahar B., Walter C., Aurousseau P., 1995 - Mapping waterlogging of soils using Digital Terrain Models. Hydrological Processes, 9(1), pp. 27-34

Météo France, 2020. CLIMAT FRANCE par Météo-France - Normales et relevés sur la France métropolitaine

Minasny, B. and Hartemink, A.E., 2011. Predicting soil properties in the tropics. Earth-Science Reviews, 106(1-2): 52-62

Minasny, B., & McBratney, A. B., 2016. Digital soil mapping: A brief history and some lessons. Geoderma, 264, 301–311.

Minasny, B., McBratney, A.B., 2015. Digital soil mapping: A brief history and some lessons. Geoderma

Mulder, V. L., Lacoste, M., Richer-de-Forges, A. C., Arrouays, D. 2016. GlobalSoilMap France: High-resolution spatial modelling the soils of France up to two meter depth.

Science of the Total Environment, 573, 1352–1369. https://doi.org/10.1016/j.scitotenv.2016.07.066

Nachabe, M.H., 1998. Refining the definition of field capacity in the literature. Journal of Irrigation and Drainage Engineering-Asce, 124(4): 230-232.

Nemes, A., Pachepsky, Y.A. and Timlin, D.J., 2011. Toward Improving Global Estimates of Field Soil Water Capacity. Soil Science Society of America Journal, 75(3): 807-812.

Nussbaum, M., Papritz, A., Baltensweiler, A., Walthert, L., 2013. Estimating soil organic carbon stocks of Swiss forest soils by robust external-drift kriging. Geosci. Model Dev. Discuss. 6, 7077–7116

Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., … Papritz, A. 2018. Evaluation of digital soil mapping approaches with large sets of environmental covariates. Soil, 4(1), 1–22. https://doi.org/10.5194/soil-4-1-2018

OCDE. 2002. Rapport annuel de l'OCDE 2002, Éditions OCDE, Paris.

Odgers, N.P., Libohova, Z., Thompson, J.A., 2012. Equal-area spline functions applied to a legacy soil database to create weighted-means maps of soil organic carbon at a continental scale. Geoderma 189–190, 153–163

Odgers, N.P., Sun, W., McBratney, A.B., Minasny, B., Clifford, D., 2014. Disaggregating and harmonising soil map units through resampled classification trees. Geoderma 214–215, 91–100.

O'Leary, G.J., Li Liu, D., Ma, Y., Li, F.Y., McCaskill, M., Conyers, M., et al., 2016. Modelling soil organic carbon 1. Performance of APSIM crop and pasture modules against long-term experimental data. Geoderma, 264, 227–237.

Pachepsky, Y. and Rawls, W.J., 2004. Preface: status of pedotransfer functions. In: Y. Pachepsky and W.J. Rawls (Editors), Development of pedotranfer functions in soil hydrology. Development in soil science. Elsevier, pp. vii-xvi.

Padarian, J., Minasny, B., McBratney, A. B., & Dalgliesh, N. (2014). Predicting and mapping the soil available water capacity of Australian wheatbelt. Geoderma Regional, 2–3(C), 110–118.

Padarian, J., Minasny, B., and McBratney, A. B.: Using deep learning for digital soil mapping, SOIL, 5, 79–89.

Parton, W.J., Schimel, D.S., Cole, C.V., Ojima, D.S., 1987. Analysis of factors controlling soil organic matter levels in Great Plains Grasslands 1. Soil Sci. Soc. Am. J., 51(5), 1173–1179.

Pebesma, E. J. 2004. Multivariable geostatistics in S: The gstat package. Computers and Geosciences, 30(7), 683–691.

Poggio, L., Gimona, A., Brown, I., Castellazzi, M., 2010. Soil available water capacity interpolation and spatial uncertainty modelling at multiple geographical extents. Geoderma, 160(2), 175–188.

Prasad, A., Iverson, L., and Liaw, A. 2006. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems, 9(2):181–199.

R Development Core Team 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ramsey, J.B. 1969. Tests for specification errors in classical linear least squares regression analysis. J. Roy. Statist. Soc. B., 31(2):350–371.

Ratliff, L.F., Ritchie, J.T. and Cassel, D.K., 1983. Field-measured limits of soil-water availability as related to laboratory-measured properties. Soil Science Society of America Journal, 47(4): 770-775

Rawlins, B.G., Marchant, B.P., Smyth, D., Scheib, C., Lark, R.M., Jordan, C., 2009. Airborne radiometric survey data and a DTM as covariates for regional scale mapping of soil organic carbon across Northern Ireland. Eur. J. Soil Sci. 60, 44–54

Rawls, W.J., Brakensiek, D.L. and Saxton, K.E., 1982. Estimation of soil water properties. Transactions of the ASAE, 25(5): 1316-1320

Richards, L.A. and Weaver, L.R., 1943. Fifteen-atmosphere percentage as related to the permanent wilting percentage. Soil Science, 56: 331-339

Richards, L.A., Campbell, R.B. and Healton, L.H., 1949. Some Freezing Point Depression Measurements on Cores of Soil in Which Cotton and Sunflower Plants Were Wilted. Proceedings. Soil Science Society of America Proceedings, 14: 47-50

Richer-de-forges, A.C., Arrouays, D., Bardy, M., Bispo, A., Lagacherie, P., Laroche, B., Lemercier, B., Sauter, J., Voltz, M., (2019). Mapping of Soils and Land-Related Environmental Attributes in France : Analysis of End-Users ' Needs. Sustainability 11, 1–15.

Robbez Masson, J.M. 1994. Reconnaissance et délimitation de motifs d'organisation spatiale : application à la cartographie des pédopaysages. Thèse de Docteur Ingénieur de l'Ecole Nationale Supérieure Agronomique de Montpellier.

Rocha, M.M. da, Yamamoto, J.K., 2000. comparison between kriging variance and interpolation variance as uncertainty measurements in the Capanema Iron Mine, State of Minas Gerais— Brazil. Nat. Resour. Res. 9, 223–235

Román Dobarco, M., Bourennane, H., Arrouays, D., Saby, N. P. A., Cousin, I., & Martin, M. P. (2019). Uncertainty assessment of GlobalSoilMap soil available water capacity products: A French case study. Geoderma, 344(February), 14–30.

Román Dobarco, M., Cousin, I., Le Bas, C., Martin, M.P., 2019. Pedotransfer functions for predicting available water capacity in French soils, their applicability domain and associated uncertainty. Geoderma, 336, 81–95.

Román Dobarco, Mercedes & Bourennane, Hocine & Arrouays, Dominique & Saby, Nicolas & Cousin, Isabelle & Martin, Manuel. (2019). Uncertainty assessment of

GlobalSoilMap soil available water capacity products: A French case study. Geoderma. 344. 14-30. 10.1016/j.geoderma.2019.02.036.

Romano, N. and Palladino, M., 2002. Prediction of soil water retention using soil physical data and terrain attributes. Journal of Hydrology, 265(1-4): 56-75

Romano, N. and Santini, A., 2002. Field water capacity. In: J.H.D.a.G.C.T. (Ed.) (Editor), Methods of Soil Analysis: Part 4. Physical Methods. Soil Science Society of America, Madison, Wis, pp. 722-738.

Salter, P.J. and Haworth, M., 1961. The available-water capacity of a sandy loam soil. 1. A critical comparison of methods of determining the moisture content of soil at field capacity and at the permanent wilting percentage. Journal of Soil Science, 12.

Salvador S., Lagacherie P., and Morlat R., 1997 - Zonage prédictif des terroirs viticoles à partir de secteurs pris comme référence. Etude et Gestion des Sols, 4, pp. 175 –190

Sanchez, P.A., Ahamed, S., Carre, F., Hartemink, A.E., Hempel, J., Huising, J., Lagacherie, P., McBratney, A.B., McKenzie, N.J., Mendonca-Santos, M. de L., Minasny, B., Montanarella, L., Okoth, P., Palm, C.A., Sachs, J.D., Shepherd, K.D., Vagen, T.-G., Vanlauwe, B., Walsh, M.G., Winowiecki, L.A., Zhang, G.-L., 2009. Digital Soil Map of the World. SCIENCE 325, 680–681

Savage, S.L., Lawrence, R.L., Squires, J.R., 2015. Predicting relative species composition within mixed conifer forest pixels using zero-inflated models and Landsat imagery. Remote Sens. Environ. 171, 326–336.

Schaap, M.G., Leij, F.J., van Genuchten, M.T., 1998. Neural network analysis for hierarchical prediction of soil hydraulic properties. Soil Sci. Soc. Am. J. 62, 847

Shrestha, D.L., Solomatine, D.P., 2006. Machine learning approaches for estimation of prediction interval for the model output. Neural Networks 19 (2), 225–235

Soil Science Division Staff. 2017. Soil survey manual. C. Ditzler, K. Scheffe, and H.C. Monger (eds.). USDA Handbook 18. Government Printing Office, Washington, D.C.

SOIL SCIENCE SOCIETY OF AMERICA - SSSA. Glossary of soil science terms. Madison, 1984. 38p.

Soil Survey Division Staff, 1993. Soil survey manual. Soil Conservation Service: U.S. Department of Agriculture Handbook, 18

Somarathna, P. D. S. N., Minasny, B., & Malone, B. P. (2017). More Data or a Better Model? Figuring Out What Matters Most for the Spatial Prediction of Soil Carbon. Soil Science Society of America Journal, 81(6), 1413–1426. https://doi.org/10.2136/sssaj2016.11.0376

Song, J., 2015. Bias correction for Random Forest in regression using residual rotation. Journal of the Korean Statistical Society, 44, 321-326.

Spiess, A-N., 2018. Propagate: Propagation of Uncertainty. R package version 1.0-6. Available on : https://CRAN.R-project.org/package-propagate

Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol. Methods 14, 323–348.

Styc, Q., Lagacherie, P. (2019). What is the Best Inference Trajectory for Mapping Soil Functions: An Example of Mapping Soil Available Water Capacity over Languedoc Roussillon (France). Soil Systems, 3(34), 17. https://doi.org/10.3390/soilsystems3020034

Styc, Q., Lagacherie, P. Uncertainty assessment of soil available water capacity using error propagation: a test in Languedoc Roussillon. Submitted on Geoderma since 24/01/2020.

Sykes, D.J., 1964. The availability of soil moisture plants. Retrospective Theses and Dissertations. Paper 3011

Bishop, T.F.A., McBratney, A.B., Laslett, G.M. 1999. Modelling soil attribute depth functions with equal-area quadratic smoothing splines. Geoderma, 91 (1–2), pp. 27-45

Tanaka, K., Takizawa, H., Kume, T., Xu, J.Q., Tantasirin, C. and Suzuki, M., 2004. Impact of rooting depth and soil hydraulic properties on the transpiration peak of an evergreen forest in northern Thailand in the late dry season. Journal of Geophysical Research-Atmospheres, 109(D23): 10

Taylor, C.A., Blaney, H.F. and McLaughlin, W.W., 1934. The wilting-range in certain soils and the ultimate wilting-point. Transactions of the American Geophysical Union, 15: 436- 444

Tetegan, M., Nicoullaud, B., Baize, D., Bouthier, A., Cousin, I., 2011. The contribution of rock fragments to the available water content of stony soils: proposition of new pedotransfer functions. Geoderma 165, 40–49.

TW Nauman, JA Thompson, NP Odgers, Z Libohova, 2012. Fuzzy disaggregation of conventional soil maps using database knowledge extraction to produce soil property maps, in: Digital Soil Assessments and Beyond. CRC Press, pp. 203–207

Ugbaje, S.U., Reuter, H.I., 2013. Functional digital soil mapping for the prediction of available water capacity in Nigeria using legacy data. Vadose Zone J., 12(4).

Van Groenigen, J. W., Stein, A. 1998. Constrained optimization of spatial sampling using continuous simulated annealing. Journal of Environmental Quality 27 (5), 1078–1086.

Vanderlinden, K., Giráldez, J. V., Van Meirvenne, M., 2005. Soil water-holding capacity assessment in terms of the average annual water balance in southern Spain. Vadose Zone J., 4, 317–328.

Vaysse, K. & Lagacherie, P. 2015. Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). Geoderma Regional, 4, pp.20–30.

Vaysse, K., Arrouays, D., McKenzie, N.J., Coste, S., Lagacherie, P., 2014. Estimation of GlobalSoilMap.net grids cells from legacy soil data at the regional scale in Southern

France, in: GlobalSoilMap : Basis of the Global Spatial Soil Information System. CRC Press, pp. 133–138.

Vaysse, K., Lagacherie, P. 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64.

Vaysse K. Application et développement de méthodes de cartographie numérique des propriétés des sols à l'échelle régionale : cas du Languedoc-Roussillon [Internet]. 2015. Available from: http://www.theses.fr/2015NSAM0036/document

Veihmeyer, F.J., Hendrickson, A.H. 1927. The relation of soil moisture to cultivation and plant growth. Soil Sci. 3, 498–513.

Vereecken, H. and Herbst, M., 2004. Statistical regression. In: Y. Pachepsky and W.J. Rawls (Editors), Development of pedotranfer functions in soil hydrology. Development in soil science. Elsevier, pp. 3-19

Vereecken, H., Maes, J., Feyen, J. and Darius, P., 1989. Estimating the soil moisture retention characteristic from texture, bulk density, and carbon content. Soil Science, 148(6): 389- 403

Viscarra Rossel, R. A., Brus, D. J. 2018. The cost-efficiency and reliability of two methods for soil organic C accounting. Land Degradation and Development, 29(3), 506–520. https://doi.org/10.1002/ldr.2887

Voltz, M., Arrouays, D., Bispo, A., Lagacherie, P., Laroche, B., Lemercier, B., Richer de Forges, A., Sauter, J., Schnebelen, N. 2018. La cartographie des sols en France : Etat des lieux et perspectives. INRA, France, 114 pages

Voltz, M., Arrouays, D., Bispo, A., Lagacherie, P., Laroche, B., Lemercier, B., Richier-de-Forges, A., Sauter, J., Schnebelen, N. (in press). Disseminating Digital Soil Mapping in national soil mapping programmes: a prospective analysis in France. Submitted in Geoderma Regional.

Voltz, M., Webster, R., 1990. A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. J. Soil Sci. 41, 473–490.

Wackernagel, H., 2003. Multivariate geostatistics. Springer Science & Business Media

Wadoux, A. M. J. C., Brus, D. J., Heuvelink, G. B. M. 2019. Sampling design optimization for soil mapping with random forest. Geoderma. https://doi.org/10.1016/j.geoderma.2019.

Wadoux, A.M.J.-C., Brus, D.J., Heuvelink, G.B.M., 2018. Accounting for non-stationary variance in geostatistical mapping of soil properties. Geoderma 324, 138–147.

Walvoort, D. J. J., Brus, D. J., de Gruijter, J. J. 2010. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Computers and Geosciences, 36(10), 1261–1267. https://doi.org/10.1016/j.cageo.2010.04.005

Walvoort, D. J. J., Brus, D. J., de Gruijter, J. J. 2018. Spatial Coverage Sampling and Random Sampling from Compact Geographical Strata. R package version 0.3-8. https://CRAN.R-project.org/package=spcosa

Webster, R., Harrod, T.R., Staines, S.J., Hogan, D.V., 1979. Grid sampling and computer mapping of the ivybridge area, DEVON

Wösten, J.H.M., Pachepsky, Y.A. and Rawls, W.J., 2001. Pedotransfer functions: bridging the gap between available basic soil data and missing soil hydraulic characteristics. Journal of hydrology, 251: 123-150

Wright, M., & Ziegler, A., (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. Journal of Statistical Software, 77(1), 1 - 17.

Yang, C. S., & Yang, Y. H. (2017). Improved local binary pattern for real scene optical character recognition. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2017.08.005

Zhang, X.Y., Zhang, X.Y., Liu, X.W., Shao, L.W., Sun, H.Y. and Chen, S.Y., 2015. Incorporating root distribution factor to evaluate soil water status for winter wheat. Agricultural Water Management, 153: 32-41

Zhu, A.-X., Band, L., Vertessy, R., Dutton, B., 1997. Derivation of Soil Properties Using a Soil Land Inference Model (SoLIM). Soil Sci. Soc. Am. J. 61, 523

Résumé

Le réservoir utile du sol (RU) désigne la capacité des sols à stocker l'eau pour les plantes. Le RU joue un rôle majeur dans de nombreux domaines tels que la production alimentaire et la régulation des inondations. Le RU est également un élément essentiel pour prévoir et optimiser l'irrigation des sols localement. Par conséquent, il est primordial de connaitre avec précision les variations spatiales du RU. Cependant, les bases de données pédologiques actuelles ne fournissent pas une cartographie du RU qui soit à la fois exhaustive et suffisamment précise pour être utilisée à l'échelle de la parcelle. Cette limite pourrait être levée en utilisant le gisement considérable de données pédologiques anciennes non numérisées comme données d'apprentissage d'un modèle de cartographie numérique des sols. Ainsi, la société BRL Exploitation (BRLE) dispose sur son périmètre irrigué (6 636 km2 en plaine littorale Languedocienne) de 228 000 observations de sol. L'objectif de cette thèse a été de développer une méthodologique de cartographie numérique du réservoir utile à partir de ces données pédologiques anciennes. Le RU étant une propriété fonctionnelle de sol impliquant plusieurs propriétés de sol sur plusieurs profondeurs, les premiers travaux de cette thèse ont porté sur l'impact de la trajectoire de calcul utilisée sur les performances de prédiction du RU, une trajectoire étant définie par l'ordre selon lesquels sont effectuées les opérations de calcul du RU sur une couche de sol donné, d'agrégation des couches de sol et de spatialisation, afin d'obtenir une prédiction du RU. En prenant l'exemple du Languedoc-Roussillon, 18 trajectoires utilisant les données pédologiques disponibles dans le Référentiel Régional Pédologique pour produire une prédiction du RU ont été testées. La meilleure performance de prédiction a été obtenue par la trajectoire de calcul calculant puis spatialisant les RU pour quatre couches de sol distinctes et combinant enfin ces quatre couches. Ensuite, les fonctionnalités du modèle de prédiction du RU ont été complétées par une prédiction des incertitudes associées, essentielle à l'aide à la décision. Ces incertitudes ont été prédites par un modèle de propagation d'erreurs utilisant les erreurs de spatialisation du RU sur les quatre couches de sol spatialisées séparément - estimées par une forêt aléatoire quantile - et prenant en compte les corrélations d'erreurs de ses composants. L'utilisation de ce modèle a montré une bonne aptitude à estimer et spatialiser l'incertitude de prédictions, dans un contexte de faibles performances de prédiction du RU.

La dernière étape a été consacrée à l'utilisation de données anciennes de BRLE pour alimenter un modèle de cartographie numérique des sols à l'échelle locale, en prenant l'exemple de la commune de Bouillargues. L'augmentation de densité spatiale par l'ajout de sondages au jeu de profils de sol, associée à l'ajout dans l'algorithme d'apprentissage de données représentant la position géographique des sondages, a considérablement amélioré la résolution spatiale, les performances de prédictions du RU et la précision des cartes d'incertitude. Cependant, les erreurs de caractérisation du RU sur ces sondages ont été constatées comme un facteur limitant les performances de prédictions du RU et de son incertitude. Une meilleure prise en compte de ces erreurs serait nécessaire pour améliorer les résultats actuels.

Les travaux de thèse ont permis de concevoir et de tester une démarche visant à valoriser l'utilisation des données pédologiques anciennes dans une approche de cartographie numérique des sols appliquée au RU. Une chaîne de traitement informatique visant à déployer la démarche sur l'ensemble du périmètre irrigué a été développée et une étude de coût/bénéfice a été réalisée. L'automatisation, au moins partielle de la saisie des sondages apparaît une condition nécessaire à la réalisation d'une carte du RU à l'échelle parcellaire sur tout le périmètre irrigué de BRLE.

Mots clés :

Réservoir utile ; cartographie numérique des sols ; données anciennes ; incertitude ; arbres de régression ; densité ; trajectoire de calcul ; propagation d'erreur ; échelle locale ; échelle régionale.

Abstract

Soil available water capacity (SAWC) refers to the soil capacity to store water for plants. SAWC plays an important role on many ecosystem services such as food security and flood and gas regulation. SAWC is also crucial for planning and optimizing soil irrigation at local scale. Thus, it is essential to understand the SAWC spatial distribution. However, current soil databases do not provide mapping of SAWC that is both comprehensive and accurate enough for use at plot scale. This limitation could be overcome by using the considerable repository of legacy soil data, undigitized, as learning data for a digital soil mapping model. BRL Exploitation (BRLE), thus, has 228,000 soil observations on its irrigated perimeter (6,636 km² in the Languedoc coastal plain). The aim of this thesis was to develop a useful digital mapping methodology for the SAWC based on these legacy soil data.

As the SAWC is a functional soil property involving several soil properties over several depths, the first work of this thesis focused on the impact of the inference trajectory used on the prediction performance of the SAWC, an inference trajectory being defined by the order in which the operations of computation of the SAWC on a given soil layer, aggregation of soil layers and mapping are performed, to finally obtain a prediction of the SAWC. Taking the example of Languedoc-Roussillon, 18 inference trajectories using the pedological data available in the Regional Pedological Referential to produce a prediction of the SAWC, were tested. The best prediction performance was obtained by the trajectory calculating then mapping the SAWC for four distinct soil layers and finally combining these four layers.

Then, the functionalities of the SAWC prediction model were complemented by a prediction of the associated uncertainties, which is important data for decision support. These uncertainties were predicted by an error propagation model using the spatialization errors of SAWC on the four separately spatialized ground layers - each estimated using a Quantile Random Forest - and taking into account the error correlations of these components. The use of this model has shown a good ability to estimate and map the uncertainty of predictions, in a context of low overall prediction performance of the SAWC.

The last step of this thesis was devoted to the use of legacy data from BRL Exploitation to feed a digital soil mapping model at local scale, taking the example of the commune of Bouillargues. The increase in spatial density by adding auger holes to the set of soil profiles, associated with the addition in the learning algorithm of data representing the geographical position of the auger holes, has considerably improved the spatial resolution, the prediction performance of the SAWC and the accuracy of the associated uncertainty maps. However, SAWC characterization errors on these auger holes were found to be a limiting factor in the prediction performance of the SAWC and its uncertainty. A better consideration of these errors would be necessary to improve the current results.

The thesis work has enabled the design and testing of an approach to enhance the use of legacy soil data in a digital soil mapping approach applied to the useful reservoir. A computer processing chain aimed at deploying the approach on the whole irrigated perimeter was developed and a cost/benefit study was carried out. The automation, at least partial, of the auger hole data entry operations appears to be a necessary condition for the production of a map of the SAWC on a plot scale over the whole of the BRL Exploitation irrigated perimeter.

Keywords:

Soil available water capacity; digital soil mapping; legacy soil data; uncertainty; random forest; density; inference trajectory; error propagation; local scale; regional scale.