Modélisations spatio-temporelles d'indicateurs bio-physiques des sols et de fonctions fournies par les écosystèmes

Alexandre M.J.-C Wadoux

Mémoire déposé en vue de l'obtention de l'Habilitation à Diriger des Recherches

Modélisations spatio-temporelles d'indicateurs bio-physiques des sols et de fonctions fournies par les écosystèmes

Présenté par Alexandre Wadoux
le 5 juillet 2023 à 9h30

Devant le jury composé de

Mr Hocine Bourennane, Ingénieur de recherche, INRAE, Info&Sols	**Examinateur**
Mme Isabelle Cousin, Directrice de recherche, INRAE, Info&Sols	**Rapporteuse**
Mme Chantal de Fouquet, Directrice de recherche, Mines Paris - PSL	**Examinatrice**
Mr Philippe Lagacherie, Ingénieur de recherche, INRAE, UMR LISAH	**Examinateur**
Mme Madlene Nussbaum, collaboratrice scientifique, Haute école spécialisée bernoise	**Examinatrice**
Mr Christian Walter, Professeur, Institut Agro, UMR SAS	**Rapporteur**
Mr Bas van Wesemael, Professeur, Earth and Life Institute, UCLouvain	**Rapporteur**

UNIVERSITÉ DE MONTPELLIER

Je déclare avoir respecté, dans la conception et la rédaction de ce mémoire d'HDR, les valeurs et principes d'intégrité scientifique destinés à garantir le caractère honnête et scientifiquement rigoureux de tout travail de recherche, visés à l'article L.211-2 du Code de la recherche et énoncés par la Charte nationale de déontologie des métiers de la recherche et la Charte d'intégrité scientifique de l'Université de Montpellier. Je m'engage à les promouvoir dans le cadre de mes activités futures d'encadrement de recherche.

Alexandre Wadoux
Modélisations spatio-temporelles d'indicateurs bio-physiques des sols et de fonctions fournies par les écosystèmes

Habilitation à Diriger des Recherches, Université de Montpellier, Montpellier, France (2023)
71 pages sans annexes.

Table des matières

Tirés à part

CHAPITRE 2

Mémoire des travaux scientifiques

2.1 Introduction

Comprendre comment les **propriétés cibles du paysage** évoluent avec les changements de l'écosystème a récemment attiré beaucoup d'attention, non seulement à des fins scientifiques mais également pour l'élaboration de politiques publiques et le développement d'incitations financières pour leur conservation. La caractérisation spatiale et temporelle de ces propriétés permet d'acquérir des connaissances sur les mécanismes physiques, biologiques et chimiques qui déterminent le fonctionnement de l'écosystème. **Les cartes mondiales, continentales et régionales des variables biophysiques fournissent des données de base pour évaluer la façon dont les écosystèmes réagissent aux perturbations humaines et au réchauffement climatique.** Celles-ci sont utilisées par les modélisateurs ainsi que par les décideurs politiques (Schmidt-Traub, 2021). Elles appuient également la recherche sur le changement climatique. Par exemple, les cartes mondiales des stocks et des flux de carbone aériens et souterrains sont essentielles pour évaluer s'il existe une émission terrestre positive ou négative nette de carbone dans l'atmosphère. Les décideurs s'intéressent aussi davantage au rôle que certaines de ces variables, tel que le sol, pourrait jouer dans la séquestration du carbone et l'atténuation du changement climatique (Rumpel et al., 2018). Les stocks de carbone organique du sol sont par exemple l'un des trois indicateurs de neutralité en matière de dégradation des terres développés par la Convention des Nations Unies sur la lutte contre la désertification (UNCCD). Pour ces raisons, de nombreuses études ont tenté d'**estimer comment les concentrations, stocks et flux de variables biophysiques varient dans l'espace, dans le paysage ou pour de vastes surfaces continentales**, et de comprendre comment celles-ci réagissent aux changements de l'environnement.

La modélisation spatiale de variables biophysiques peut se faire à l'aide de modèles statistiques qui interpolent les données mesurées et utilisent un ensemble de covariables environnementales dont des cartes sont disponibles, telles que des images de télédétection et des attributs de terrain. Le développement des techniques géostatistiques dans les années 1960 par Matheron (1963) et l'apparition de la méthode de prédiction géostatistique de base connue sous le nom de krigeage (Krige, 1951), ont conduit les scientifiques à développer une large gamme de variantes de krigeage adaptée aux données sur lesquelles elles ont été appliquées. Des exemples de ces modèles sont le krigeage avec dérive externe qui assouplit l'hypothèse de stationnarité du premier ordre ou le krigeage indicateur pour l'utilisation sur des réponses binaires. Les modèles basés sur les techniques géostatistiques et de prédiction via le krigeage sont couramment utilisés, tels qu'en science des sols, en météorologie, épidémiologie et pour l'évaluation des ressources minérales. Un exemple récent d'étude utilisant cette approche est Kempen et al. (2019) pour la cartographie des stocks de carbone organique dans les sols de Tanzanie à l'aide d'un modèle géostatistique avec une moyenne variant dans l'espace. Depuis les années 2000, des outils complexes, statistiques et algorithmiques du domaine de l'apprentissage automatique sont devenus populaires pour la modélisation spatiale,

notamment parce qu'ils sont généralement plus précis que les modèles statistiques simples et évitent les hypothèses relatives au système de krigeage. Ils sont apparus comme un outil précieux pour faire des prédictions spatiales « sans hypothèses » à partir de grands ensembles de covariables environnementales. L'inconvénient de ces modèles est la complexité d'obtention d'informations sur leur structure. Quelques exemples récents d'utilisation de ces modèles pour la modélisation spatiale sont Delgado-Baquerizo et al. (2018) pour la spatialisation des bactéries du sol, ou Baccini et al. (2012) pour la cartographie de la biomasse aérienne. Une alternative aux modélisations spatiales empiriques est la modélisation dynamique qui a principalement été réalisée à l'aide de modèles mécanistes et semi-mécanistes tels que DNDC, CENTURY ou RothC en science des sols pour la modélisation des stocks de carbone (voir Lugato et al. (2014) et Martin et al. (2021), par exemple) ou Biome-BGC en écologie pour estimer le stockage et le flux de carbone, d'azote et d'eau dans la végétation (voir aussi l'application aux forêts européennes dans Pietsch et Hasenauer, 2002). Les modèles semi-mécanistes sont attrayants car ils rendent justice aux processus sous-jacents et permettent l'intégration des connaissances existantes dans la modélisation. Ils sont également particulièrement adaptés lorsque l'objectif n'est pas seulement de prédire, mais aussi de comprendre les facteurs à l'origine d'un résultat. Il reste cependant plusieurs défis pour l'application de modèles mécanistes et semi-mécanistes sur de grandes aires continentales (par exemple, la paramétrisation du modèle et les forçages), qui n'ont été que partiellement résolus dans la littérature.

Mes travaux s'inscrivent dans cette logique de modélisation spatiale à travers **le développement de cadres conceptuels et méthodologiques pour la caractérisation spatio-temporelle d'indicateurs biophysiques des sols et de certaines fonctions fournies par les écosystèmes** (érosion hydrique des sols, carbone des sols et biomasse). Les méthodes permettent la cartographie numérique de propriétés cibles du paysage peu ou pas directement observables (sols, précipitations) mais en relation avec d'autres qui, elles, sont plus observables (p.ex. relief) ou déjà produites par des experts (cartes géologiques). J'utilise principalement des **méthodes empiriques de géostatistique et d'apprentissage automatique ou profond** que j'adapte et améliore pour s'adapter aux données particulières auxquelles j'ai été confronté, à la fois spatiales, rares et incertaines. Les approches scientifiques que je développe s'appliquent à de nombreuses disciplines et c'est donc logiquement que des cas d'études en science des sols, hydrologie et écologie sont intégrés dans ce mémoire. Je m'intéresse néanmoins particulièrement aux sols pour lesquels les techniques statistiques et les connaissances pédologiques me permettent de déduire les processus à l'oeuvre pour comprendre comment le sol se forme, varie dans l'espace géographique et change dans le temps.

Mes activités de recherches ont été structurées en trois périodes dans lesquels j'ai continuellement élargie mes champs d'actions disciplinaires. La première période est d'abord de 2012 à 2015 à l'Université de Tübingen en Allemagne au sein du projet sino-allemand Yangtze GEO qui visait à modéliser et à comprendre les mécanismes et facteurs de contrôle de l'érosion hydrique des sols dans l'écosystème en amont du barrage des Trois Gorges en Chine centrale. J'ai participé à l'ensemble des travaux de recherche de la partie sol, d'abord au travers de larges campagnes de terrain en Chine et ensuite via la spatialisation des propriétés de sols et la modélisation semi-mécaniste de l'érosion. Dans une seconde période de 2015 à 2019, j'ai réalisé mon doctorat à l'Université de Wageningen au Pays-Bas au sein du réseau de recherche Européen Marie-Curie QUICS qui visait à quantifier les incertitudes dans les études intégrées de bassin versant. Mes travaux ont d'abord été centrés sur la modélisation hydrologique au sein de bassin versant. Dans ce cadre, j'ai réalisé quatre courts séjours de travail dans des départements étrangers auprès de chercheurs spécialistes de la modélisation hydrologique et statistique (voir Partie 1). J'ai ensuite orienté mes travaux vers la modélisation spatiale de propriétés de sols et vers les techniques d'échantillonnages spatiales qui accompagnent cette modélisation. Enfin, dans une troisième période de 2019 à ce jour j'ai travaillé en mi-temps pour cartographier numériquement des propriétés dynamiques du sols et de biodiversité à l'échelle de l'Australie, et en mi-temps sur des sujets de recherche choisis au sein du groupe de science des sols

de l'Université de Sydney en Australie. Dans ce cadre j'ai bénéficié d'une grande liberté sur mes sujets de recherches, que j'ai utilisé pour poursuivre une ligne de recherche intégrant de multiples aspects de la cartographie numérique.

Il serait en effet inexact de considérer la modélisation spatiale comme une simple interpolation de données biophysiques. Le plan d'échantillonnage des données utilisées pour l'interpolation a un effet important sur la spatialisation. De même, l'erreur de ces données peut être substantielle et impacter de manière significative la qualité de la prédiction. Il s'agit aussi de considérer une modélisation spatiale en changement, passant d'une ère avec peu de données à une ère avec plus de données (numériques) mais venant de nouveaux acteurs (p.ex. à travers les sciences participatives) et de techniques variées (télédétection ou détection proche) utilisées principalement dans des disciplines scientifiques annexes (p.ex. la chimiométrie). En outre, les cartes contiennent de multiples sources d'erreurs qui affectent leur qualité, de sorte qu'il est courant de rapporter une estimation de l'incertitude de la carte. Cette incertitude est estimée différemment en fonction du type de modèle de spatialisation (empirique basé sur l'apprentissage automatique, semi-mécaniste ou basé sur les géostatistiques). Enfin, la structure de ces modèles doit être compris pour garantir que les prédictions soient faites pour les bonnes raisons, et les estimations spatiales de propriétés biophysiques doivent être évaluées à l'aide techniques statistiques. La synergie entre ces multiples aspects de la modélisation spatiale est au coeur de mes recherches, que j'ai complété par une réflexion sur les ressorts épistémologiques de l'utilisation de ces techniques pour la caractérisation des sols et le développement de nouvelles connaissances.

Cette réflexion a émané d'interrogations personnelles sur l'utilisation croissante des techniques empiriques et basée sur les données pour la caractérisation et la cartographie numérique des sols. Mon cheminement sur ce sujet épistémologique a commencé durant la fin de ma seconde période d'activité de recherche en 2019, alors que je terminais un mémoire de Master sur les aspects épistémologiques de la science des sols à la fin du XIX*e* siècle. Je me suis d'abord interrogé sur les risques d'utilisations erroné des modèles entièrement basés sur les données, pour enfin mener un réflexion sur leur bénéfice potentiel pour générer des hypothèses de travail à travers un raisonnement abductif (voir la Partie 2.3.4). J'ai approfondi ce sujet dans une longue réflexion sur les aspects épistémologiques d'une partie de la science des sols où la méthode scientifique est principalement basée sur les données. L'intérêt de ma démarche se trouve dans les aller-retours entre réflexion épistémologique sur une discipline et applications sur la méthodologie et sur la façon de faire la recherche dans une « fertilisation croisée » des deux approches. Je me suis par exemple opposé à une démarche communément admise qui consiste à choisir un modèle parcimonieux pour la cartographie numérique de propriétés de sol, en soutenant que la question d'un choix entre un modèle parcimonieux (c.a.d dont l'interprétation est simple) ou complexe était mal posée : d'aucun ne devrait pas limiter la complexité des modèles empiriques si ils sont plus précis, car ils représentent mieux la complexité de la variation des sols. À la place, il s'agit de développer des outils statistiques pour obtenir des informations à partir de ces modèles complexes. Dans ce registre, plusieurs de mes travaux sont basés sur un constat dressé après une réflexion des limites et manquements de ma discipline.

J'aborde ce mémoire en sept parties qui incluent une introduction et une synthèse. La structure suit ma vision des différentes étapes de la modélisation spatiale et des nouveaux enjeux que j'ai traités : échantillonnage spatial, modélisation, quantification et propagation de l'incertitude et enfin interprétation et validation des modèles et prédiction. Certains travaux de recherches publiés y sont traités de manière transversales dans plusieurs sous-parties. Dans la Partie 2.6, j'aborde plusieurs sujets élargissant le spectre de la modélisation spatiale des variables biophysiques et discute des sujets annexes qui irriguent les autres parties, soit car ils permettent une meilleure collection de données, l'inclusion de nouveaux acteurs, une meilleure compréhension des concepts de ce mémoire ou enfin parce qu'ils permettent une vision à long terme des défis en rapport au numérique. Dans la synthèse et conclusion, je montre en quoi mes travaux passés sont une base solide pour le projet de recherche que je propose ensuite.

2.2 Optimisation des méthodes d'échantillonnages

La surveillance et la prévision spatio-temporelle des variables environnementales (p.ex. les propriétés pérennes de sol) demandent d'acquérir des observations qu'il est difficile d'obtenir partout et à tout moment. Nous ne pouvons ainsi collecter qu'un fragment, un échantillon de la variable cible dans l'espace et dans le temps, dans le but d'utiliser cet échantillon pour déduire et cartographier cette variable aux endroits et moments non visités. L'échantillonnage peut représenter une tâche coûteuse et fastidieuse, et il est judicieux de disposer de stratégies efficaces pour produire un plan d'échantillonnage optimal pour la cartographie. La plupart des recherches menées dans ce domaine se cantonnent au cas de la cartographie prédictive utilisant la géostatistique. Au cours des dernières années, les modèles géostatistiques et les techniques de cartographie associées ont largement évolué, nécessitant une adaptation des plans d'échantillonnage disponibles. **Mes recherches sur l'optimisation des méthodes d'échantillonnage avaient pour objectif principal d'aborder le plan d'échantillonnage de récentes avancées en analyse spatiale.**

2.2.1 Optimisation pour un modèle géostatistique de variance non stationnaire

Les méthodes conventionnelles de géostatistique font la prédiction d'une variable à des endroits non-visités en prenant en compte les mesures de cette variable aux points d'échantillonnages. Dans la plupart des cas, les prédictions sont meilleures si le modèle prend en compte la relation entre la variable prédite et une série de covariables environnementales disponibles en tous points sous forme de cartes, telles que les modèles numériques de terrain ou les images satellites. Cela nous amène à des techniques de krigeage avec dérive externe dont la moyenne est localement estimée par une fonction linéaire des covariables environnementales. Dans ce modèle, la variance est présumée stationnaire, c'est à dire constante au sein de l'aire d'étude. Cette hypothèse n'est pas plausible dans certaines situations où la variation spatiale des résidus est différente dans certaines parties de l'aire d'étude. Plusieurs études géostatistiques nous ont montrés ce problème. Par exemple, l'étude de Voltz et Webster (1990) a trouvé des différences significatives entre les variogrammes calculés sur des échantillons d'argiles obtenus sur deux types de sédiments Jurassique. Dans une étude qui sera décrite plus en détail dans le Chapitre 2.3, je développe un modèle géostatistique non stationnaire (voir Wadoux et al., 2018). Dans l'étude décrite ci-dessous (Wadoux et al., 2017), j'explore **l'optimisation du plan d'échantillonnage pour le modèle géostatistique de variance non stationnaire** défini dans Wadoux et al. (2018).

La prise en compte de cette non-stationnarité dans la variance des propriétés environnementales des paysages complexes permet de mieux quantifier l'incertitude associée aux cartographies. Cette méthode est appliquée dans une étude de cas de cartographie des précipitations journalières dans le nord de l'Angleterre et d'optimisation spatiale des pluviomètres. Je montre dans ce chapitre que la prévision spatiale et temporelle des précipitations est meilleure avec un modèle qui inclut la non-stationnarité dans la moyenne et la variance, comme le montrent les statistiques de vraisemblance et de critère d'information d'Akaike. L'optimisation du réseau pluviométrique est obtenue par un recuit simulé spatial, une technique d'optimisation développée et décrite dans Van Groenigen et Stein (1998).

Le réseau de pluviomètres ainsi optimisé améliore légèrement la précision de la cartographie des précipitations. Le gain de précision reste limité, car j'ai utilisé un échantillon identique pour tous les pas de temps, tandis que les zones avec une plus grande incertitude de prévision varient de jour en jour. Le plan d'échantillonnage optimisé montre également une configuration spécifique, avec une distribution spatiale assez uniforme (Figure 2.1) mais une densité accrue dans les zones où la variance résiduelle est grande. Je teste ensuite un plan d'échantillonnage optimisé en utilisant une réduction de 10% du nombre total de pluviomètres. Je montre alors **une amélioration significative par rapport au plan d'échantillonnage**

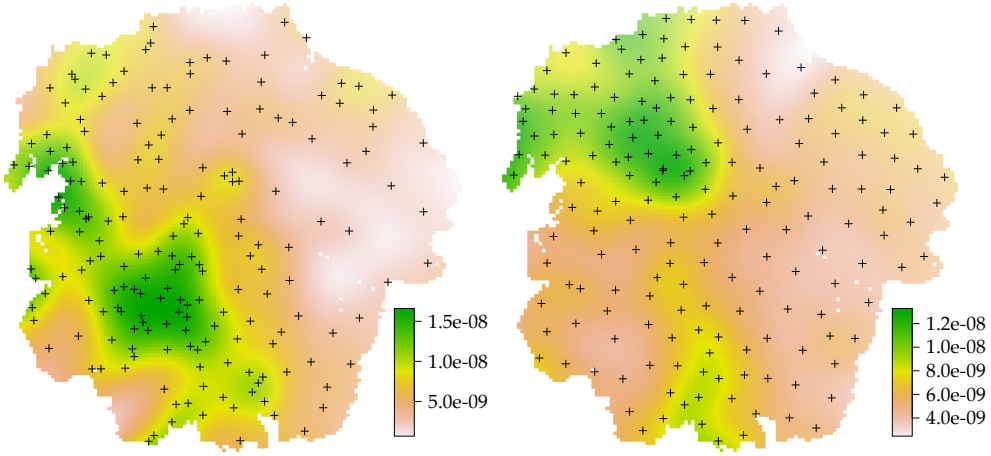

Figure 2.1 *– Plan d'échantillonnage initial (à gauche) et optimisé (à droite) du réseau de pluviomètre dans une aire d'étude en Angleterre. La couleur des cartes représente la densité spatiale des pluviomètres, et est exprimée en pluviomètre par carré de 500 m × 500 m. D'après Wadoux et al. (2017).*

original utilisant tous les pluviomètres disponibles. Je conclus que 10% des pluviomètres peuvent être supprimés (par exemple pour des réductions de coûts) sans perte de précision pour la cartographie, à la condition que les pluviomètres soient placés de manière optimale.

2.2.2 Optimisation pour la variance d'erreur de prévision et l'incertitude des paramètres

Les plans d'échantillonnage optimaux pour les modèles géostatistiques et la prédiction avec le krigeage sont connus depuis les années 90 (p.ex. avec les études de Van Groenigen et Stein, 1998). Les plans d'échantillonnage optimaux sont obtenus avec des techniques d'optimisation, par exemple le recuit simulé, ayant pour objectif la minimisation d'un critère. Le critère est le plus souvent la moyenne spatiale de la variance de prédiction obtenue par krigeage. Le plan d'échantillonnage obtenu de cette manière répartie de manière presque homogène les points d'échantillons dans l'espace géographique. Plutôt que d'utiliser des techniques d'optimisation, plusieurs études ont défini des plans d'échantillonnage simples, basés sur des critères géométriques. Par exemple, Royle et Nychka (1998) ont proposé un critère basé sur la couverture spatiale de l'aire d'étude. Plus tard, Brus et al. (1999) ont proposé la moyenne de la plus courte distance au carré comme critère, pour que le plan soit optimisé de manière rapide avec l'algorithme *k*-means. Bien que simples et adéquates pour la prédiction géostatistique, ces techniques d'échantillonnages sont sous-optimales pour l'estimation du variogramme. Le variogramme demande quelques échantillons à courte distance les uns des autres, de telle manière que la corrélation à courte distance soit correctement estimée. Il est possible d'estimer l'erreur totale, qui prend en compte la variance d'erreur de prévision et l'incertitude dans l'estimation des paramètres du variogramme (Marchant et Lark, 2004). Dans ce cadre, une récente étude (Lark et Marchant, 2018) nous a montré qu'un plan d'échantillonnage simple pour ce critère serait une répartition homogènes des points dans l'espace, avec une proportion d'environ 10% de ces points qui seraient pris à une courte distance des points existants.

Dans mon étude (Wadoux et al., 2019b), **j'examine l'utilisation de stratégies d'échantillonnage simples pour prendre en compte un critère englobant à la fois la variance d'erreur de prévision et l'incertitude des paramètres du variogramme dans la cartographie géostatistique des propriétés du sol**. J'examine de manière empirique si des plans d'échantillonnage optimisés pour ce critère sont substantiellement mieux que des plans d'échantillonnage très simples décris dans Lark et Marchant (2018)

Pour cela, je teste deux plans d'échantillonnage : le premier correspond à une répartition homogène des points dans l'espace (couverture spatiale quasi-homogène) et le suivant correspondant à la couverture spatiale complétée par un sous-ensemble d'unités proches. Je compare ces plans à un plan optimisé pour le critère d'erreur totale.

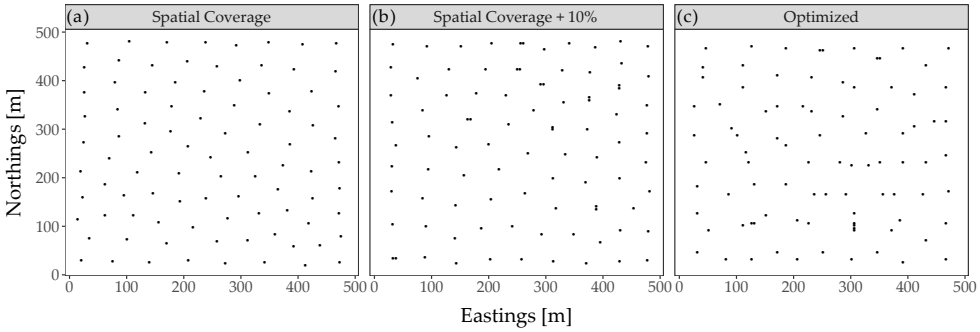

***Figure 2.2** – Exemples de plans d'échantillonnage a) de couverture spatiale, b) de couverture spatiale avec un sous-ensemble pris à courte distance des points existants, et c) optimisé pour 90 points avec un critère englobant l'erreur totale. D'après Wadoux et al. (2019b).*

Je montre qu'un plan d'échantillonnage de couverture spatiale donne de piètres résultats pour la cartographie utilisant le krigeage ordinaire en raison du manque d'informations à courte distance pour estimer les paramètres du variogramme. Ceci est valable pour des séries de paramètres estimés sur la base d'un variogramme de Matèrn. Utiliser un échantillonnage optimisé (Figure 2.2) fonctionne toujours légèrement mieux, mais présente plusieurs inconvénients dont notamment celui de devoir connaître les paramètres du variogramme. Cela implique également de définir une fonction objective caractérisant l'erreur totale et de la minimiser à l'aide d'algorithmes d'optimisations. En revanche, **un échantillonnage dit de couverture spatiale complété par un sous-ensemble d'unités proches offre des résultats précis pour la plupart des variogrammes testés**. Je recommande donc d'utiliser cette dernière méthode d'échantillonnage pour la conception d'un échantillonnage géostatistique, à moins que le variogramme ne soit préalablement connu (par exemple, si l'on dispose d'un variogramme moyen). Si un variogramme moyen est disponible pour la propriété d'intérêt, il peut être utilisé pour optimiser l'échantillonnage.

Finalement, j'ai aussi testé le nombre minimum d'unités nécessaires pour estimer le variogramme d'une étude géostatistique et montre que cela dépend fortement du degré de corrélation spatiale de la variable étudiée. Pour les grandes valeurs de portée effective du variogramme et petits ratios de pépite sur palier, il est montré que seulement 15 unités suffisent pour rendre l'analyse géostatistique intéressante, c'est-à-dire plus précise qu'une estimation fondée sur un échantillonnage non probabiliste.

2.2.3 Optimisation spatiale pour un modèle de forêt d'arbres décisionnels

La cartographie n'est pas toujours effectuée à l'aide de méthodes géostatistiques. Il existe un intérêt croissant pour la cartographie et la spatialisation de données environnementales utilisant des techniques d'apprentissage automatique non linéaires basées sur les données. Les techniques d'apprentissage automatique sont très populaires, car elles ne sont pas basées sur des présupposés statistiques rigides telles que la distribution normale des données d'entrée, et peuvent traiter des données qui ont une relation non-linéaire avec les covariables environnementales. Pour la spatialisation des données environnementales, de nombreuses études récentes ont utilisé des techniques d'apprentissage automatique, par exemple l'étude de Cook-Patton et al. (2020) a utilisé une forêt d'arbres décisionnels pour faire une carte de l'accumulation potentiel de carbone suite à la repousse de forêts naturelles à l'échelle mondiale. Ces techniques d'ap-

prentissage automatique sont basées entièrement sur les données, et sont donc particulièrement sensibles au plan d'échantillonnage spatiale. Pour autant, très peu d'études existent pour définir ce qui constitue un plan d'échantillonnage optimal pour une prédiction spatiale basée sur des techniques d'apprentissage automatique. Dans la littérature scientifique, des plans d'échantillonnage communs sont utilisés pour collecter des données d'entrée aux modèles d'apprentissage automatique, par exemple un plan basé sur la couverture spatiale homogène. Certaines études (p.ex. Brus, 2019) ont supposé qu'un plan basé sur la couverture homogène des covariables environnementales serait à privilégier, sans pour autant le démontrer empiriquement.

L'objectif principale de l'étude présentée dans Wadoux et al. (2019a) était d'**étendre nos connaissances sur l'optimisation de l'échantillonnage pour la cartographie à l'aide de forêt d'arbres décisionnels et de la comparer aux plans d'échantillonnage conventionnels**. La méthode d'apprentissage automatique de forêt d'arbres décisionnels est l'une des plus courante en analyse spatiale. J'ai testé des plans d'échantillonnage tels que : aléatoire simple, de couverture spatiale, de couverture des covariables, par hypercube latin conditionné, et je les ai comparés à un plan d'échantillonnage optimisé basé sur la minimisation de l'erreur quadratique moyenne (EQM). J'ai ensuite développé une procédure pour obtenir une distribution de l'EQM à partir des différentes méthodes d'échantillonnages, et pour différentes tailles d'échantillons. Cette méthodologie est testée dans le cadre de scénarios d'applications potentiels, en cartographiant le carbone organique de la couche superficielle du sol à l'échelle européenne en considérant les données LUCAS comme population d'intérêt.

Je démontre qu'un échantillonnage optimisé est toujours plus précis que d'autres plans d'échantillonnage couramment utilisés. Cependant, cette approche n'est possible que dans un cas restreint où on procède à un sous-échantillonnage d'un jeu de données existant avec des valeurs connues de la propriété du sol. En comparant l'EQM des cartes obtenues par un échantillonnage optimisé à celles obtenues par des plans échantillonnages communs, il est montré que l'optimisation d'un échantillonnage en termes d'EQM n'est pas toujours intéressant. Lorsque la taille de l'échantillon augmente, la précision des cartes produites par les différents types d'échantillons convergent vers des valeurs similaires. Dans une étude de cas sur la cartographie du carbone organique du sol à grande échelle, une densité d'échantillonnage supérieure à 1 unité d'échantillonnage par 4000 km^2 réduit considérablement la différence en terme de EQM moyen entre les types d'échantillons. Un échantillon optimisé pour la distance normalisée quadratique moyenne la plus courte dans l'espace des covariables correspond le mieux à un échantillon optimisé en termes d'EQM. En analysant l'emplacement des échantillons dans l'espace géographique et dans l'espace des covariables, je montre également qu'**un échantillon optimisé n'est pas distribué uniformément dans l'espace géographique, mais semble être réparti de manière assez uniforme dans l'espace des covariables**, et en particulier en considérant les variables les plus importantes pour le modèle d'apprentissage automatique. Il est toutefois difficile de tirer des conclusions supplémentaires en raison de la dispersion complexe des unités dans l'espace des covariables. Des recherches complémentaires sont nécessaires dans cette direction. Certains des aspects de ce travail on été étudiés par un étudiant de Master que j'ai supervisé. Cet étudiant a notamment regardé la façon dont les conclusions que j'ai tirées pour une forêt d'arbres décisionnels sont applicables à d'autres techniques d'apprentissage automatique.

2.2.4 Optimisation pour des données d'entrée de modèle

Les plans d'échantillonnage étudiés et optimisés dans les trois parties précédentes ont en commun leur objectif : être optimal pour la spatialisation de données environnementales. Ces données environnementales sous forme de cartes sont utilisées pour de nombreuses applications telles que l'aide à la prise de décision, l'agriculture de précision, ou comme entrées dans les modèles conceptuels, empiriques ou à

base physique. Les modèles hydrologiques de pluie-débit prennent en entrée des séries temporelles de précipitation, de température, et sont étalonnés avec des séries temporelles de débit (sortie du modèle). Suivant la représentation de l'espace des variables hydrologiques, ces modèles prennent en entrées des données globales dans un bassin versant (une moyenne, la variation spatiale est ignorée), des données semi-distribuées (plusieurs entités spatiales au sein du bassin versant), ou distribuées (maillage régulier). Dans chacun des cas, l'incertitude des cartes des données d'entrée (p.ex. les précipitations) dépend fortement du plan d'échantillonnage utilisé pour leur estimation, et se propage dans la prédiction du modèle pluie-débit (Refsgaard et al., 2007). Une réduction du nombre de pluviomètres, par exemple dans le cadre d'une réduction des dépenses dans un bassin versant, se répercute sur la qualité des prédictions du débit. Inversement, une augmentation du nombre de pluviomètres dans un bassin versant a tendance à réduire l'erreur de prédiction d'une carte de précipitation, et donc de fournir une incertitude de prédiction du modèle pluie-débit plus étroite. Plusieurs études ont tenté de comprendre le lien entre plan d'échantillonnage des données d'entrée et la prédiction d'un modèle (p.ex. Bárdossy et Das, 2008). Cependant, l'optimisation d'un plan d'échantillonnage est complexe lorsque le but ultime est de fournir une carte utilisée comme entrée pour un modèle dont la prévision est le principal objectif. La qualité de prédiction du modèle n'est pas seulement subordonnée au plan d'échantillonnage des données d'entrée, mais aussi à un ensemble d'incertitudes (p.ex. des paramètres, de la structure du modèle) qui peuvent se quantifier et qui se propagent dans la prédiction (Renard et al., 2011).

J'aborde cette question dans Wadoux et al. (2020a) avec un cas d'étude où **la géostatistique est utilisée pour la cartographie des précipitations et la calibration bayésienne d'un modèle hydrologique pour la prévision des débits**. La calibration bayésienne permet de capturer les incertitudes d'entrée, d'état initial, de paramètre et de structure du modèle, tout en tenant compte des incertitudes des mesures de sortie. Je teste différents scénarios de réduction des pluviomètres. Pour chaque scénario, le nombre de pluviomètre est fixe et leur emplacement au sein du bassin versant est optimisé en fonction de l'erreur de prédiction du modèle de krigeage de bloc. La Figure 2.3 nous montre les 6 scénarios testés et l'emplacement optimisé des pluviomètres. Le modèle est ensuite ré-étalonné pour tous les scénarios et une comparaison est faite en fonction de l'intervalle de prédiction des estimations du débit. J'ai testé cette méthodologie dans un bassin versant du nord-ouest de la Suisse où une riche base de donnée était disponible.

L'une des principales conclusions de cette étude est la suivante : **dans un cas de prévision du débit d'une rivière à l'aide d'un modèle pluie-débit et de cartes des précipitations, un seul pluviomètre peut suffire pour obtenir un étalonnage précis des paramètres du modèle et des prévisions du débit. L'ajout de cinq pluviomètres améliore cependant la prévision du modèle**. En ajouter davantage ne produit qu'une amélioration marginale de la précision des prévisions. Le calibrage de la série chronologique des précipitations en tant que paramètres supplémentaires permet d'obtenir des performances de modèle plus précises que dans le cas où l'incertitude des précipitations n'est pas actualisée à l'aide des mesures de débit. En outre, il est démontré pour cette étude de cas que l'incertitude des paramètres du modèle est le principal facteur d'incertitude de la loi postérieure du débit et que l'incertitude des entrés a une contribution relativement faible. Cependant, l'étude montre également que l'étalonnage Bayésien de la pluviométrie présente de graves inconvénients de calcul. En particulier, la calibration dans un temps raisonnable d'un grand nombre de paramètres d'entrées de pluie reste un défi majeur.

2.2.5 Comment comparer les méthodes d'échantillonnages spatiales ?

Les études énumérées précédemment nous montrent qu'il existe une multitude de plans d'échantillonnage afin de collecter des échantillons pour étalonner un modèle statistique ou d'apprentissage automatique. Ces plans sont, par exemple, le hypercube latin conditionné, répartition homogène dans l'espace, un

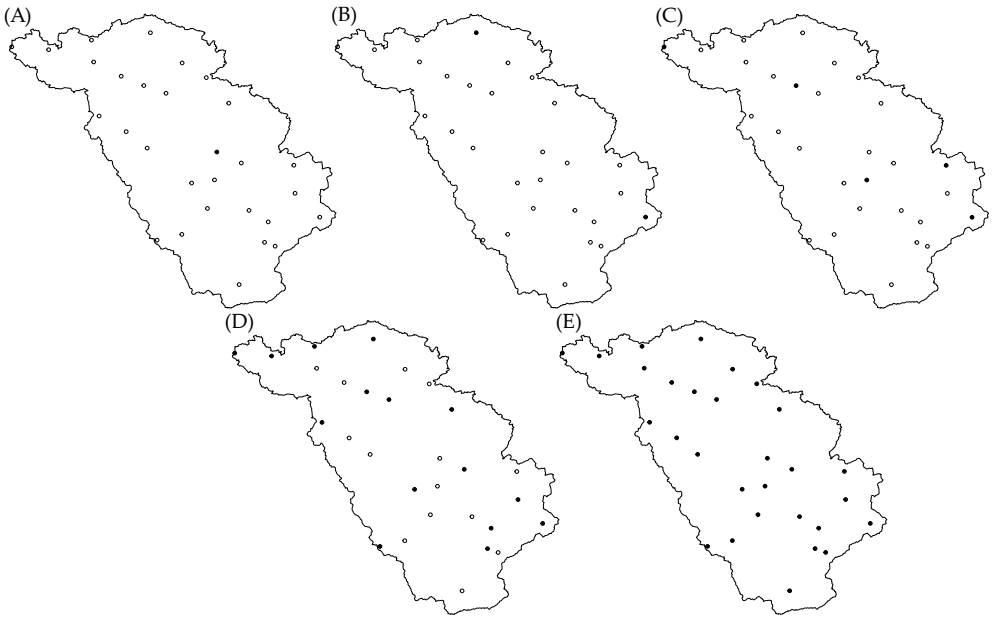

Figure 2.3 *– Emplacement optimisé des pluviomètres au sein du bassin versant pour 6 scénarios comprenant (A) un, (B) deux, (C) cinq, (D) quinze et (E) tous (29) les pluviomètres. Le point noir représente un emplacement choisi pendant l'optimisation et le point vide représente un emplacement candidat pendant l'optimisation mais non sélectionné. D'après Wadoux et al. (2020a).*

échantillon aléatoire simple ou stratifié, ou encore une couverture homogène de l'espace des covariables. L'utilisation de ces plans d'échantillonnage pour la spatialisation de données environnementales a été revue par De Gruijter et al. (2006) et Brus (2022). La multiplicité des plans d'échantillonnage disponibles a naturellement conduit à des études comparant l'effet des différents types de plans d'échantillonnage sur la qualité des spatialisations faites avec les échantillons collectés. Par exemple, l'étude de Schmidt et al. (2014) a évalué trois types de plans d'échantillonnage pour leur précision dans l'estimation de prédiction de cinq propriétés de sol à l'échelle d'un champ. Pour les trois plans d'échantillonnage, un seul plan d'échantillonnage est collecté et est utilisé pour étalonner un modèle. La prédiction du modèle pour un plan d'échantillonnage est évaluée en utilisant les échantillons des autres plans.

Cette étude, avec d'autres, illustre la difficulté de comparer des plans d'échantillonnage : en répétant la sélection du plan d'échantillonnage plusieurs fois on obtiendrait différent plans, et donc différentes estimations de la qualité des spatialisations. C'est le cas de tous les plans d'échantillonnage qui ont une part d'aléatoire, c'est à dire les plans d'échantillonnage énoncés ci-dessus, mais aussi les plans optimisés. En effet, dans les plans optimisés la configuration initiale des points est obtenue aléatoirement et l'optimisation peut converger vers des résultats différents. Afin de comparer les plans d'échantillonnage, il faut prendre en compte l'aléatoire de leur configuration et l'impact que cela produit sur le critère d'évaluation de la qualité des cartes produites.

C'est une question que j'aborde dans Wadoux et Brus (2021) en développant **une méthode pour la comparaison de plans d'échantillonnage dans les études sur des données réelles et sur des populations simulées**. Je défend l'idée qu'un plan d'échantillonnage utilisé pour la spatialisation doit être évalué sur la distribution des statistiques de validation des cartes obtenues sur un ensemble de répétitions du plan d'échantillonnage, pas sur une seule répétition. De la même manière, les plans d'échantillonnage probabilistes sont toujours évalués sur la base d'une distribution, par exemple dans l'estimation de la moyenne d'une population. Je montre l'importance de considérer la distribution des statistiques de validation dans deux cas d'étude, l'un en rapport avec une propriété sol continue, et l'autre pour la classification du type

d'occupation des sols (variable catégorielle). Dans les deux cas, deux méthodes communes d'échantillonnages sont comparées sur la base de leur distribution : hypercube latin conditionné (voir le papier de Minasny et McBratney (2006), cLHS dans la Figure 2.4) et aléatoire simple (SRS).

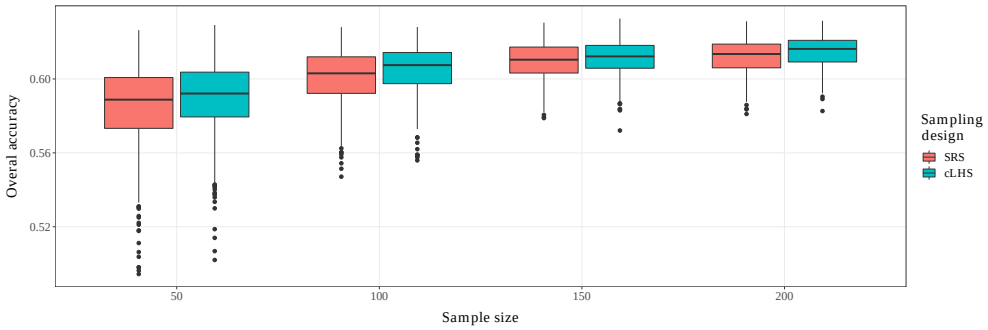

***Figure 2.4** – Distribution de l'estimation de l'indice de précision général de la population pour la classification de l'occupation des sols, pour les deux types de plans d'échantillonnage (cLHS et SRS) et plusieurs tailles d'échantillon. D'après Wadoux et Brus (2021).*

La figure 2.4 nous montre un des résultats de cette étude : **il est important de répéter la sélection des échantillons quand le but est de comparer des plans d'échantillonnage**. Les distributions obtenues dans la Figure 2.4 sont larges et se chevauchent pour une taille spécifique de l'échantillon. Dans ce cas les différences entre plans sont à peine visibles. En ne prenant qu'un seul échantillon, nous avons jusqu'à 50% de chance pour qu'un plan soit considéré meilleur que l'autre, et vice-versa. Il y a un risque d'obtenir des résultats trompeurs et de conclure sur la qualité d'un plan en comparaison d'un autre, alors qu'en prenant en compte la moyenne de la distribution la conclusion serait différente. Ce type d'étude n'est possible que dans des populations connues, car répéter la sélection des échantillons n'est pas possible en pratique. Cette étude, néanmoins, met en lumière la nécessité de tester les plans d'échantillonnage dans des situations différentes : i) des cas d'études avec des données biophysiques, ii) des cas d'études utilisant des échantillons très larges qui sont considérés comme une population cible, iii) des cas d'études dans lesquels une carte est considérée sans erreur et peut être échantillonnée plusieurs fois, et enfin iv) des cas d'études avec des simulations géostatistiques.

2.3 Modélisation spatiale

L'étude de la variabilité spatiale des variables environnementales est utilisée dans un nombre croissant d'applications, telles que dans l'aide à la prise de décision, l'optimisation de l'allocation des ressources ou l'estimation de la dégradation de certaines propriétés cibles du paysage. La cartographie numérique permet aussi des approches de suivis dans le temps, par exemple pour suivre l'évolution des stocks de carbone dans les sols agricoles. Les techniques de cartographie sont nombreuses et ont en commun **la spatialisation de propriétés du paysage ou propriétés environnementales qui sont rares ou coûteuse à mesurer, mais qui peuvent être estimées spatialement avec l'utilisation de modèles et de covariables déjà produites** (p.ex. relief ou images satellites). Les modèles peuvent être empiriques et statistiques (modèles géostatistiques ou d'apprentissage automatique) ou à base physique. J'ai abordé la cartographie numérique de variables biophysiques (p.ex. sols et précipitations) sous plusieurs aspects : prioritairement en développant des méthodes géostatistiques et d'apprentissage automatique, mais aussi par l'utilisation d'un modèle à base physique. Ces études constituent un ensemble méthodologique, mais contiennent aussi des réflexions sur l'intérêt scientifique des cartes et des techniques empiriques de production de ces cartes dans une visée d'amélioration de nos connaissances et de notre compréhension des

processus environnementaux.

2.3.1 Modélisation à base physique

Le maintien de la fertilité des sols par la régulation de l'érosion fait partie des services dit de régulation assurés par les écosystèmes. Les bénéficiaires directs de ce service sont les agriculteurs grâce au maintien du potentiel agronomique de leur sol. La diminution de la provision de ce service, au contraire, menace l'aptitude des sols à remplir leur fonctions (infiltration, ruissellement). En amont du barrage des Trois-Gorges en Chine centrale, la pression démographique sur des terres agricoles aux sols minces, en pentes ou en terrasses, combinée à un climat subtropical de mousson et des géo-risques accrus, entraînent des changements rapides de l'écosystème qui se traduisent par une vulnérabilité des sols à l'érosion. L'intérêt de ce sujet de recherche était de **caractériser des propriétés au service de l'étude du processus d'érosion hydrique des sols dans un agroécosystème de montagne aux dynamiques marquées et fortement anthropisé**, afin d'obtenir une estimation détaillée et spatialement explicite du déplacement des sédiments dans le bassin versant, tout en prenant en compte le peu de données disponibles. Bien que plusieurs recherches aient déjà quantifié spatialement l'érosion et la déposition des sédiments à l'échelle du bassin versant, peu d'études étaient disponibles pour des bassins avec cultures en terrasses et peu de données, malgré une forte demande des autorités locales dans le réservoir des Trois-Gorges.

Nous avons développé une démarche combinant la cartographie numérique des sols et l'utilisation d'un modèle à base physique spatialement explicite d'érosion/déposition. Le choix d'échantillonnage a fait l'objet de deux études utilisant des méthodes statistiques d'échantillonnage spatiale (hypercube latin, couverture spatiale) se basant sur les connaissances actuelles de la représentation du processus d'érosion. Celles-ci ont eu pour but de i) déterminer préalablement à la première campagne de terrain les lieux d'échantillonnages, de ii) prendre en compte les lieux inaccessibles, et finalement iii) d'améliorer, par la suite, les cartes des sols produites avec les échantillons de la première campagne. Ces recherches sur l'échantillonnage spatial ont fait l'objet de deux publications (Stumpf et al., 2016, 2017b). Nous avons par la suite utilisé des approches classiques de cartographie numérique pour produire des carte de propriétés de sol (pH, argile, matière organique) de l'horizon de surface, en mobilisant des données satellites (SPOT 5, Hyperion) et des approches d'apprentissage automatique. Ces cartes du sol sont utilisées comme variables d'entrées dans le modèle à base physique d'érosion (Erosion 3D, voir Schindewolf et Schmidt, 2012) et permettent de palier le peu de données disponibles. Le modèle Erosion 3D simule le ruissellement de surface, l'érosion, le dépôt et le détachement de particules de sol pour des périodes de précipitation. L'incorporation mathématique de ces processus repose sur deux sous-programmes, traitant du ruissellement et plus explicitement de l'érosion, modifié pour la prise en compte des terrasses et des cultures en rotation. Nous nous sommes concentrés sur 14 évènements pluvieux majeurs de 2013 et avons étalonné le modèle à l'aide de nos propres relevés pluviométriques et limnimétriques.

Les résultats de la modélisation ont été cartographiés en quatre classes (sévère, haute, modéré et faible) pour l'érosion et la déposition des sédiments, spatialement et pour chaque type d'utilisation des terres (Stumpf et al., 2017a). Le peu de données disponibles est partiellement résolu par l'utilisation des données satellites, par l'utilisation des techniques de cartographie numérique, et par la définition de plans d'échantillonnages optimaux. Les terres cultivées en zones de basse altitude sont les plus sujettes à un risque sévère d'érosion hydrique. Les zones de déposition des sédiments se concentrent dans les dépressions topographiques, les zones de petites prairies et d'habitations. Cette représentation spatiale de l'érosion et des dépositions nous permet de vérifier certains facteurs connus de l'érosion hydrique dans les bassins versants : i) la majorité des sédiments est mobilisée en l'espace de quelques épisodes pluvieux, ii) les terres en culture de rotations sont particulièrement vulnérables à cause du manque

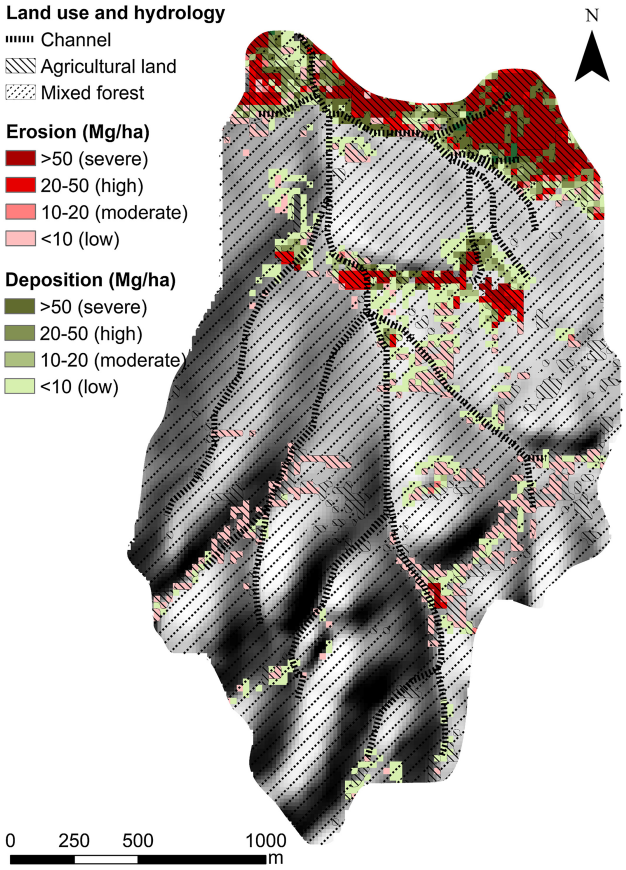

Figure 2.5 *– Source et déposition des sédiments pour les évènements érosifs avec un budget supérieur à 1 t. D'après Stumpf et al. (2017a).*

de couverture végétale pendant une partie de l'année. La visualisation de la connectivité des sédiments obtenu par l'analyse spatiale nous a permis de proposer des méthodes simples et locales pour atténuer le risque érosion, comme la mise en place de nouvelles terrasses ou l'augmentation de la couverture végétale pendant les périodes de fortes précipitations.

Ces travaux me servent de base pour une étude actuellement en cours avec une doctorante que je supervise et avec qui nous étudions la spatialisation d'un modèle mécaniste pour la cartographie des récoltes agricoles en Afrique de l'ouest. Dans cette étude, nous étudions particulièrement l'influence des données spatialisées de sol dans la prédiction des récoltes.

2.3.2 Modélisation géostatistique et krigeage

Les statistiques sont couramment utilisées pour la spatialisation des variables environnementales : pédologiques, écologiques et biologiques avec l'utilisation, entres autres, des statistiques Bayésiennes, de la régression logistique spatiale ou encore des copules. La variation spatiale des variables environnementales est le plus souvent représentée par un modèle stochastique ce qui permet de faire une inférence avec les géostatistiques. C'est la modélisation géostatistique déjà mentionnée plusieurs fois dans ce mémoire. Pour la modélisation statistique, nous utilisons les mesures comme des échantillons d'un champ aléatoire. La corrélation spatiale est ensuite représentée à l'aide d'un corrélogramme dont les paramètres sont soit connus, soit doivent être estimés. Finalement, la prédiction spatiale est faite avec des techniques dites de krigeage qui exploitent la distance entre les échantillons et les points à prédire. Les méthodes géostatistiques et de prédictions avec le krigeage sont populaires car elles possèdent plusieurs variantes, telles que

le krigeage avec dérive externe pour avoir une moyenne variante, ou le krigeage indicateur qui permet de cartographier des variables binaires. Dans ma carrière **trois techniques géostatistiques ou de krigeage ont été particulièrement étudiées et développées**, elles sont décrites ci-après.

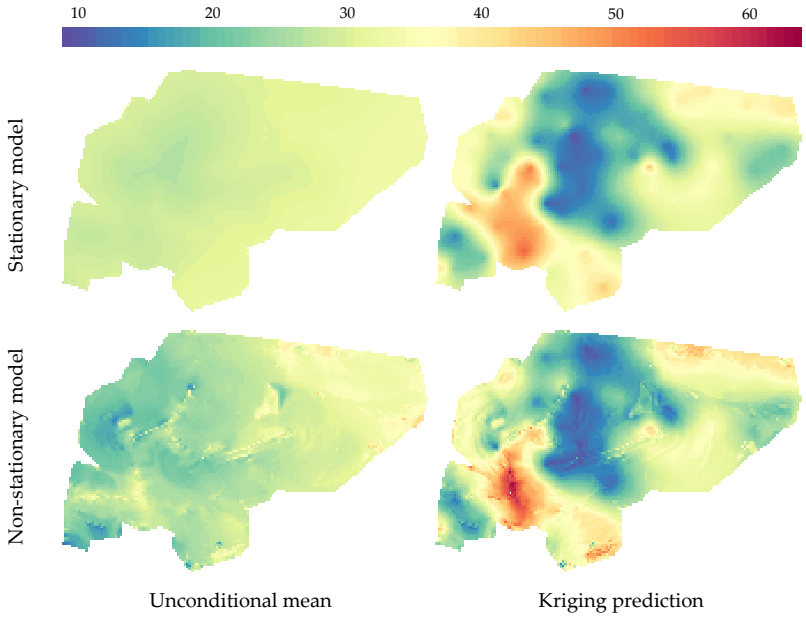

***Figure 2.6** – Comparaison des prédiction de Potassium pour le modèle stationnaire et non stationnaire. D'après Wadoux et al. (2018).*

Krigeage avec variance non stationnaire Je me suis appuyé sur un contexte de cartographie numérique des sols et des précipitations en prenant en compte la **variabilité spatiale des variables environnementales complexes présentant, par exemple, une non-stationnarité dans la variance**. Pour résoudre ce problème, j'ai étendu un modèle géostatistique de type krigeage en modélisant l'écart type du modèle comme un régression linéaire de covariables environnementales, de telle manière que les résidus soient standardisés. Cette extension a posé des difficultés dans l'estimation des paramètres, résolues par l'utilisation de l'estimateur du maximum de vraisemblance réduite (Lark, 2000). Deux cas d'étude, pour la cartographie du Potassium obtenu par spectrométrie gamma, et pour la cartographie des précipitations journalières en Angleterre pour l'année 2010, ont aidé à affiner la méthode. La Figure 2.6 illustre les résultats de prédiction pour le cas d'étude de la cartographie du Potassium, pour les deux modèles comparés : le modèle stationnaire et le non stationnaire. En analysant l'incertitude des prédictions, une des conclusions est **l'amélioration de la quantification de l'incertitude par le modèle non stationnaire**. Ce travail a aussi été approfondi par deux étudiants travaillant sur deux sujets distincts en rapport avec la modélisation géostatistique. Les deux ont été réalisés pour la cartographie des précipitations dans le cadre d'un mémoire de licence.

Géostatistique robuste Un grand nombre d'études a utilisé le krigeage avec dérive externe mentionné précédemment. L'estimation des paramètres du variogramme (ou du corrélogramme) demande des données qui suivent approximativement une loi normale. De plus, les données sont souvent contaminées par des données aberrantes (des *outliers*). Dans certains cas, ces données ne suivent pas la disposition de la majorité des observations, mais ne sont pas en tant que telles des données aberrantes. Une situation fréquente est d'exclure les données aberrantes de l'étude. Une autre solution, plus recommandée, est d'utiliser des statistiques robustes qui ne sont pas sensibles aux données aberrantes. Le but de l'étude reportée dans Ramirez-Lopez et al. (2019) était d'étudier un **modèle géostatistique dont l'estimation des paramètres de variogramme est robuste et dont les données d'étalonnage suivent une loi log-normale**. Dans

un cas d'étude de cartographie de propriétés pérennes du sol, nous avons utilisé un couplage de données de laboratoire et de données inférées par un modèle de spectroscopie infrarouge. Les données aberrantes surviennent de plusieurs façons dans ce cas de figure, et sont prises en compte par l'utilisation de l'estimateur du maximum de vraisemblance réduite robuste développée par Künsch et al. (2013). Nous trouvons que plusieurs données ont reçu un poids inférieur à 0.8 (entre 0 et 1) et que les données inférées par la modèle de spectroscopie infrarouge avaient plus de chance d'être considérées comme aberrantes. Nous avons aussi trouvé en comparant la méthodologie basée sur les statistiques robustes que **la qualité de la prédiction était meilleure que dans un cas d'utilisation des statistiques conventionnelles**.

Krigeage de bloc Les méthodes de krigeage évoquées précédemment ont en commun une prédiction ponctuelle, c'est à dire une prédiction sur un volume ou un support similaire à celui des données utilisées pour l'étalonnage du modèle. Par exemple, un échantillon de sol a un volume de l'ordre de 1 dm^3 bien que dans certains cas des échantillons composites aient un support de l'ordre d'une surface de bloc de taille 5 m $\times$ 5 m ou 10 m $\times$ 10 m. En écologie la biomasse aérienne est souvent mesurée sur des parcelles de l'ordre d'une dizaine de mètres carrés ou sur une aire circulaire de quelques mètres de rayon. Le krigeage de bloc peut être utilisé lorsque l'objectif est de **prédire une variable à un support plus large de celui sur lequel elle est observée**. L'avantage de cette méthode géostatistique est la prise en compte de la corrélation spatiale lors de l'estimation de l'incertitude associée à la prédiction. Sans cette prise en compte l'incertitude tendrait vers zéro pour une agrégation vers un support plus large. Ce point est discuté plus en détail dans la Partie 2.4. La Figure 2.7 nous montre un résultat d'agrégation spatiale pour le carbone organique des sols de France métropolitaine, avec des prédictions faites à des supports différents : à point, pour les départements, les régions et enfin à l'échelle du pays. La prédiction est faite avec le krigeage de bloc à partir d'échantillons de sols ayant un support à points. Les résultats nous montrent un phénomène bien connu des géostatisticiens, à savoir, **l'incertitude diminue quand la taille du support sur laquelle la prédiction est faite augmente**.

2.3.3 Modélisation avec l'apprentissage automatique et profond

La convergence de multiples facteurs tels qu'une grosse demande pour les données spatialisées, l'accumulation de plusieurs bases de données et de covariables environnementales et le développement des méthodes numériques et algorithmiques couplés avec des ressources pour le traitement de données, ont favorisé l'émergence des méthodes d'apprentissage automatique et profond pour la modélisation spatiale. Conventionnellement, les méthodes statistiques et géostatistiques ont été largement appliquées pour l'évaluation spatiale des ressources naturelles, et ce depuis les années 80. Les méthodes géostatistiques ont plusieurs avantages : i) un modèle statistique valide est utilisé, ii) la corrélation spatiale est explicitement modélisée, et iii) une estimation de l'incertitude est fournie avec les prédictions, ce qui rend leur utilisation pratique dans deux nombreuses applications comme la création de données spatiales pour l'aide à la prise de décision ou l'agriculture de précision. Ces mêmes méthodes ont cependant certaines limitations qui n'ont été résolues que partiellement dans la littérature scientifique. Parmi ces difficultés, plusieurs hypothèses entourent la distribution que doit prendre les résidus du modèle et leur variation spatiale (stationnaire et isotropic). L'utilisation de beaucoup de covariables est aussi un défi ainsi que la modélisation de relations non-linéaires entre la propriété cible et les covariables. De plus, la variation spatiale qui contient des changements abrupts et graduels (par exemple, une faille rocheuse) est difficile à appréhender. Finalement, les modèles géostatistiques sont aussi très lents en calcul si le nombre de points pour l'étalonnage ou la prédiction est très important. Pour ces raisons, et la convergence de facteurs évoqués précédemment, **des techniques d'apprentissage automatiques et profondes sont devenues populaires pour la spatialisation des données environnementales**. Pour la cartographie numérique des sols, plusieurs techniques ont été développées dès les années 1990. L'utilisation de ces techniques est en-

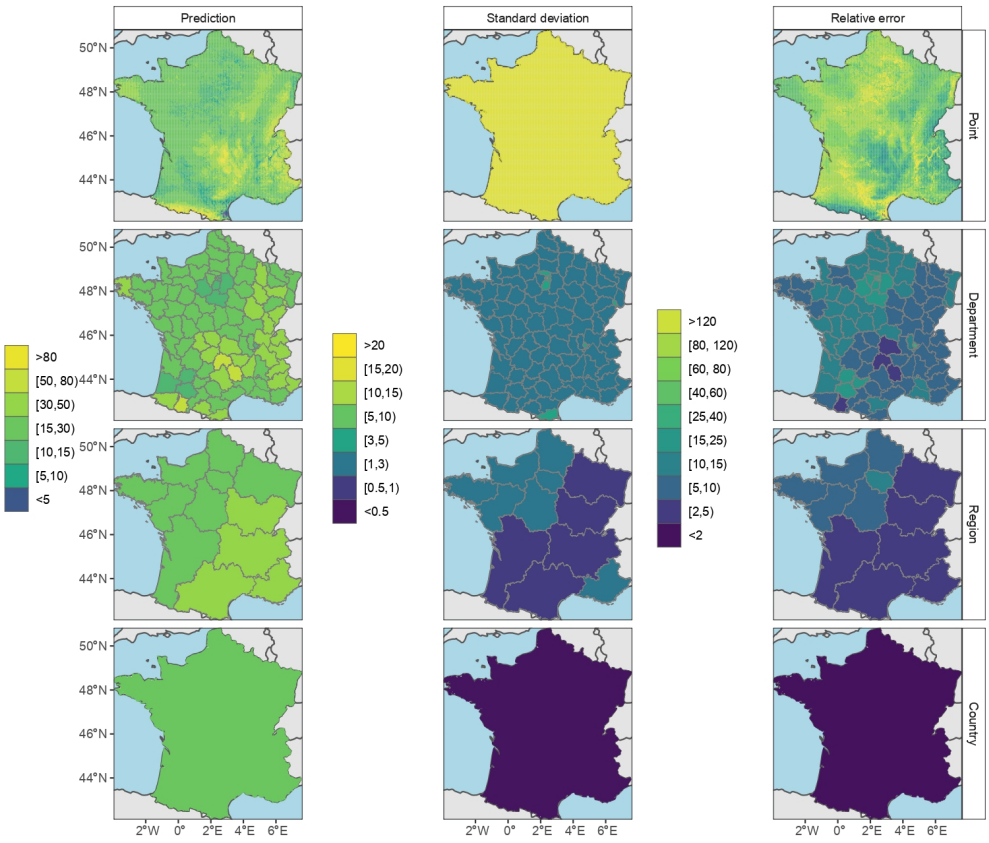

***Figure 2.7** – Cartes des prédictions, écart type de prédiction, et erreur relative pour la prédiction du carbone organique de l'horizon supérieur du sol (en g kg^{-1}) et pour des supports de tailles différentes.*

core à un stade de développement dans les sciences environnementales, et certaines dimensions, telle que l'obtention de l'incertitude des prédictions, sont encore à un stade de balbutiement. J'ai participé à ces développements à travers plusieurs études, autant sur l'apprentissage automatique que sur l'apprentissage profond, et certains dans le cadre d'une supervision d'une étudiante de doctorat (Nenkam et al., 2022).

Apprentissage automatique Dans une optique de cartographie numérique des sols, la technique de régression quantile de forêt (Meinshausen, 2006) a été rendue populaire par le papier de Vaysse et Lagacherie (2017). La régression quantile de forêt est simplement une forêt d'arbres décisionnels, mais en lieu de rapporter une seule valeur (c.a.d la prédiction à travers la moyenne des valeurs de la dernière branche), nous rapportons l'ensemble de la distribution conditionnelle de ces branches. Le principal intérêt de la régression quantile de forêt est la possibilité d'obtenir une estimation de l'incertitude de prédiction. J'ai abondamment utilisé ces avantages dans mes études, par exemple pour l'optimisation des méthodes d'échantillonnages (voir Wadoux et al., 2020a) ou pour la cartographie de la biomasse aérienne Wadoux et al. (2021a). Une contribution majeure a été la génération des nouvelles cartes de carbone organique des sols de l'Australie à très haute résolution spatiale. **Dans Wadoux et al. (2023), je bénéficie de la rapidité et de l'estimation d'incertitude de la régression quantile de forêt pour faire une cartographie numérique des sols d'Australie avec une estimation de l'incertitude**. Ces cartes, disponibles à deux résolution fine (30 m × 30 m et 90 m × 90 m) sont les nouvelles cartes de référence qui peuvent être utilisées par les décideurs et les utilisateurs des sols pour appuyer l'aide à la prise de décision. Pour une aire d'étude régionale, j'ai supervisé une étudiante qui a travaillé sur la variation spatiale des polluants du sol à l'aide de techniques d'apprentissage automatique.

L'utilisation grandissante des techniques d'apprentissage automatique pour la cartographie numérique

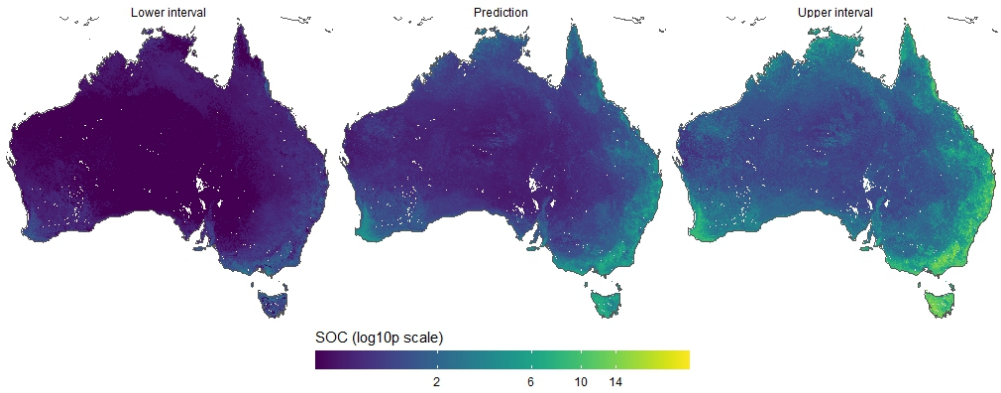

Figure 2.8 *– Cartes du carbone organique des sols pour l'horizon 0-5 cm (au centre) obtenue avec la régression quantile de forêt, avec l'incertitude de prédiction comprenant l'intervalle de prédiction pour le percentile 0.05 (à gauche) et 0.95 (à droite). D'après Wadoux et al. (2023).*

des sols, telle que la régression quantile de forêt, mais encore les réseaux neuronaux ou les machines à vecteurs de support, m'a amené à faire une revue de ces développements. Dans Wadoux et al. (2020b) je fais une **revue des études existantes en apprentissage automatique appliquées à la cartographie numérique des sols**, et définis une série de défis et de brèches dans la littérature scientifique. Je donne aussi des solutions potentielles qui sont tirées de la littérature scientifique des sciences naturelles. Une des conclusions de cette étude est la prédominance des recherches visant à faire une prédiction, au détriment sûrement d'utiliser les modèles empiriques d'apprentissage automatique pour tenter d'obtenir une meilleure compréhension des processus pédologiques. Dans beaucoup de situations, l'obtention d'une prédiction précise est importante, par exemple lorsque l'objectif est la production d'une information spatiale. Cependant, lorsque l'objectif est de comprendre les processus du sol, d'autres considérations sont à prendre en compte, comme l'inclusion de connaissance pédologique pour contraindre le modèle, ou encore l'interprétation de la structure du modèle d'apprentissage automatique qui est souvent très complexe. Je donne trois recommandations conceptuelles pour aider à une meilleure utilisation de ces modèles en cartographie numérique des sols : **dans les développements futurs, l'apprentissage automatique pour les sols pourrait incorporer trois éléments fondamentaux : la *plausibilité*, l'*interprétabilité* et l'*explicabilité***, ce qui devrait inciter les pédologues à coupler la prédiction du modèle avec l'explication pédologique et la compréhension des processus sous-jacents du sol.

Apprentissage profond Les modèles d'apprentissage profond utilisant les réseaux de neurones artificiels ont été récemment adaptés pour la spatialisation des variables environnementales, avec des succès certains (Heuvelink et Webster, 2022). En cartographie numérique des sols, j'ai fait deux études qui chacune ont participé au développement de ces techniques avec l'utilisation des réseaux neuronaux convolutifs. En plus des avantages des techniques d'apprentissage automatiques pour l'analyse spatiale décrite ci-avant (rapidité d'exécution, non-linéarité), les réseaux neuronaux convolutifs ont d'autres avantages qui en font un candidat intéressant à l'analyse spatiale des données sol. Le premier avantage est l'étalonnage du modèle qui s'effectue à l'aide d'une fonction objective. Cela permet de la flexibilité dans l'étalonnage du modèle par la définition d'une fonction spécifique à l'objectif de l'étude, par exemple à travers une prédiction multiple avec un seul modèle. Le deuxième avantage, spécifique aux réseaux neuronaux convolutif, est l'utilisation d'une image en donnée d'entrée, au lieu d'un vecteur. L'utilisation d'images en entrée signifie que les covariables ne sont plus seulement représentées à l'échelle du point, mais que le modèle inclut une multitude d'informations contextuelles locales dans le voisinage du point d'observation. Ces deux avantages certains des réseaux neuronaux convolutifs ont été exploités dans deux études distinctes :

Wadoux (2019) et Wadoux et al. (2019c). Ces deux études ont en commun **l'utilisation de réseaux neuronaux convolutifs pour la prédiction de multiples propriétés de sol** (Figure 2.9). Dans Wadoux et al. (2019c), un seul modèle est étalonné pour prédire le carbone organique des sols à deux profondeurs, alors que dans Wadoux (2019) un seul modèle prédit six propriétés de sol pour l'horizon de surface. Dans les deux cas aussi, l'interrelation entre les différentes propriétés est exploitée par l'utilisation de couches communes dans le modèle. Il est admis dans ces études que **cette utilisation bénéficie à la qualité des prédictions, en particulier pour les profondeurs de sol généralement difficiles à prédire**. Cette « prédiction multivariée » est aussi intéressante lorsque l'aire d'étude est large, comme dans Wadoux (2019) où étalonner **un seul modèle est suffisant pour prédire plusieurs propriétés de sols à l'échelle de la France**. Finalement, dans les deux études, il est trouvé que la taille de la fenêtre des données d'entrée a une importance significative sur les résultats. En testant différentes tailles, j'ai obtenu une taille optimale, qui est similaire à la distance de corrélation spatiale obtenue par l'étalonnage d'un variogramme.

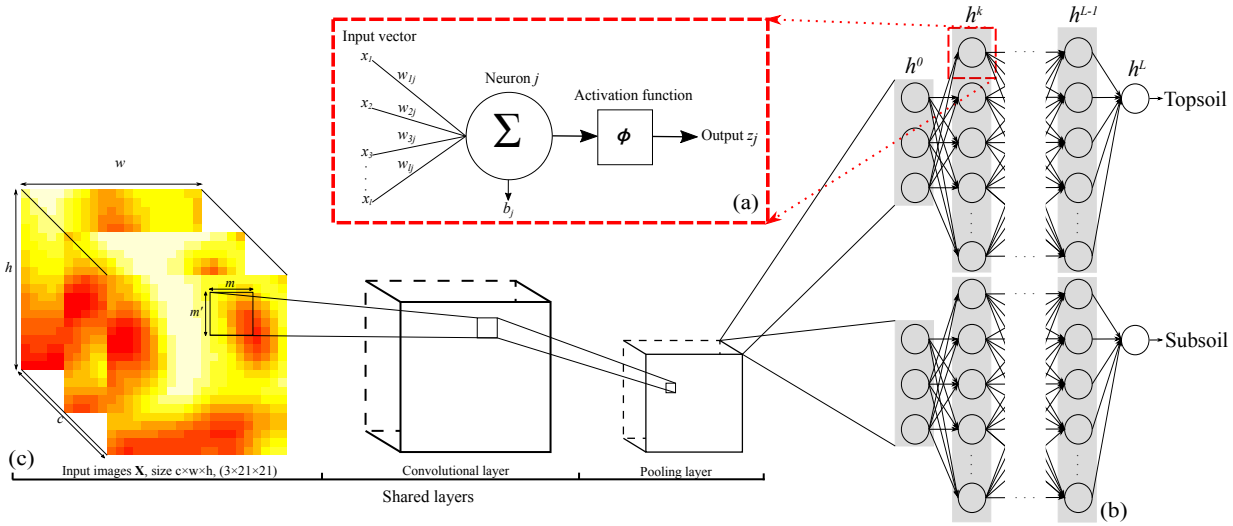

***Figure 2.9** – Représentation d'un réseau neuronal convolutif pour une utilisation dans un contexte spatial et plusieurs sorties (dans ce cas, deux profondeurs de sol), avec la visualisation de la structure interne : a) un neurone, (b) l'architecture composée de réseaux de neurones artificiels (c) les couches de convolutions. La partie c) est la structure partagée pour les différentes sorties et celle-ci se sépare ensuite en deux branches dans la partie b). D'après Wadoux et al. (2019c).*

2.3.4 Utilité et risque de la modélisation spatiale empirique

L'utilisation des techniques empiriques d'apprentissage automatique en cartographie numérique des sols engendre un série de questions quant à l'intérêt de ces techniques pour la compréhension des sols. Certaines de ces questions ont été relevées dans mon article de revue (Wadoux et al., 2020b), à savoir, **comment utiliser ces modèles empiriques pour tenter d'obtenir une meilleure compréhension des processus pédologiques ?** J'approfondis le raisonnement dans deux notes dans lesquels je discute i) l'utilité des méthodes d'apprentissage automatique pour la découverte d'un arrangement spécifique dans les données qui peuvent fournir des hypothèses au pédologue, et ii) le risque de former une interprétation causale sur les corrélations trouvées dans les données.

Utilité Il est communément accepté que dans une recherche scientifique dite conventionnelle ou normale une hypothèse est formulée en début de recherche et est la motivation pour une expérience ou une collection de données. En fonction des résultats de cette expérience, l'hypothèse est soit réfutée soit corroborée. Cela se fait dans une démarche inductive ou déductive. La déduction se forme à partir d'une théorie ou

d'une supposition, dans laquelle le scientifique fait une hypothèse qui est ensuite testée et confrontée à une expérience ou des données. La démarche inductive, au contraire, commence avec les données avec lesquelles sont inférées des principes plus généraux - du particulier au général. Dans la pratique, les deux approches opèrent et interagissent. Dans une note (Wadoux et McBratney, 2021b), je défends l'idée que **les objectifs des études utilisant le forage de données et les méthodes d'apprentissage automatique pour la cartographie numérique des sols ne sont pas clairement définis, car des hypothèses ne sont ni testées, ni développées.** Je tente de caractériser les objectifs de ces études (Figure 2.10) qui peuvent être la production d'une information spatiale précise pour une application, ou bien la production de connaissances scientifiques. Dans la premier cas, la carte peut être une fin en soi. Dans le second cas, la carte n'est pas suffisante et des hypothèses devraient être testées ou développées. Je plaide en faveur d'un raisonnement abductif dans lequel les hypothèses sont formulées d'une manière à trouver une explication *potentielle* à des motifs trouvés par l'algorithme. Dans ce cadre, la carte n'est pas une fin en soi, mais le début d'une analyse qui se base sur les questions générées par les résultats de l'algorithme.

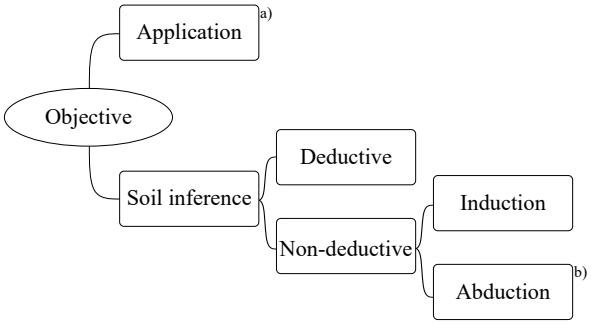

***Figure 2.10** – Illustration du parcours suivi par le scientifique utilisant un algorithme d'apprentissage automatique pour la cartographie numérique des sols. Le scientifique doit décider si l'objectif de l'étude est a) une application, ou b) à des fins scientifiques dans un raisonnement abductif. Dans a) la carte est destinée à être appliquée dans un contexte différent de la science des sols tandis que dans b) l'objectif est le développement d'hypothèses plausibles à partir du modèle prédit et des corrélations trouvées dans les données par le modèle d'apprentissage automatique. D'après Wadoux et McBratney (2021b).*

Risque Le risque grandissant, avec l'utilisation des méthodes d'apprentissage automatique pour la cartographie numérique, est l'interprétation erronée de l'importance des variables environnementales aidant à la prédiction (climat, végétation, variables de terrain). Dans ce contexte, mes collaborateurs et moi-même avons développé un cas d'étude tentant de prouver par l'utilisation de variables d'entrées dénuées de sens que **des prédictions qui semblent correctes peuvent être, en réalité, de pures chimères dans un contexte pédologique** (Wadoux et al., 2020c). Dans un cas d'étude hypothétique, la concentration de carbone de la surface de sol est prédite spatialement avec un modèle d'arbres de forêts décisionnelles et 41 covariables. Les covariables sont des photos de pédologues récupérées librement sur internet. Dans ce cas d'étude, je montre que ces covariables peuvent nous fournir une prédiction précise du carbone car la méthode d'apprentissage automatique a été capable d'identifier des corrélations dans les covariables. Cela montre qu'il faut beaucoup de prudence et ne pas prendre les corrélations pour des phénomènes de cause à effet.

2.4 Quantification et propagation de l'incertitude

Les utilisateurs des sorties de modèle ne sont pas seulement intéressés par l'estimation d'une variable, mais aussi par l'incertitude associée à la prédiction de celle-ci. Les cartes produites par des techniques

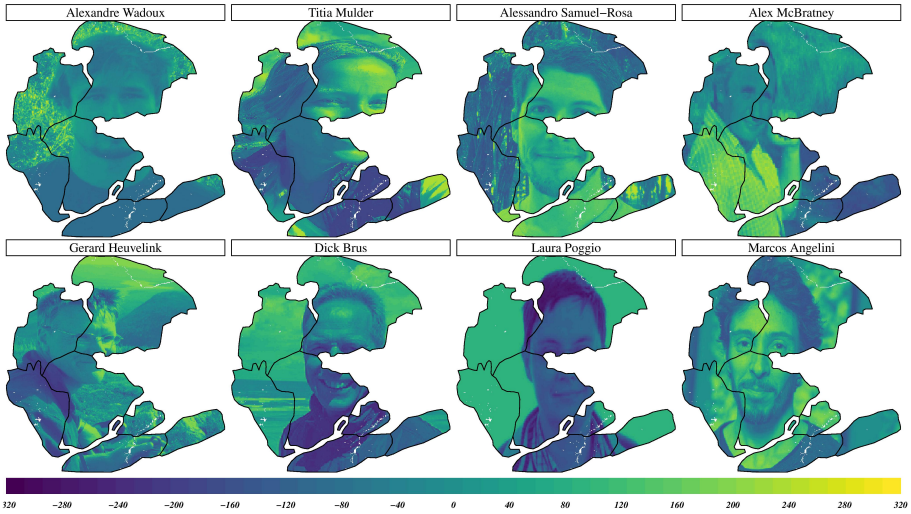

Figure 2.11 *– Des exemples de photos de pédologues utilisées comme covariables dans le modèle d'apprentissage automatique pour la prédiction du carbone organique du sol. D'après Wadoux et al. (2020c).*

d'interpolation spatiale (c.à.d. la modélisation géostatistique et le krigeage) et d'apprentissage automatique, ne sont pas sans erreurs. Ces cartes sont une représentation de la réalité et contiennent plusieurs types d'erreurs, les plus importantes étant les erreurs des observations d'entrée utilisées pour étalonner le modèle et l'erreur d'interpolation spatiale. L'erreur d'interpolation peut être estimée avec des modèles géostatistiques qui fournissent une estimation de l'incertitude. Quantifier l'erreur des données d'entrée représente un défi car dans la plupart des cas aucune information sur l'erreur de ces données n'est reportée. Il est alors compliqué de quantifier et de propager cette erreur dans la modélisation. La quantification et propagation de l'incertitude est plus complexe quand l'erreur d'entrée n'est plus le type d'erreur principal, mais se confronte à d'autres sources d'erreurs qui affectent la prédiction. C'est le cas des modèles conceptuels, tels que les modèles de pluie-débit. Ceux-ci sont des approximations de la réalité et des erreurs affectent la structure du modèle. Les paramètres qu'il faut estimer contiennent aussi des erreurs, et ceux-ci sont estimés avec des données d'étalonnages qui ne sont pas non plus sans erreur. Bien que toutes ces erreurs affectent la modélisation, elles ne sont pas toujours prises en compte. Il devient en conséquence difficile d'évaluer si une prédiction est suffisamment précise pour en tirer des conclusions. Une partie de mes recherches s'est orientée vers la **prise en compte de plusieurs types d'incertitudes dans la modélisation et, dans certains cas, leur propagation dans l'analyse spatiale.**

2.4.1 Erreur des données mesurées dans la cartographie numérique

Les données mesurées et utilisées en entrée dans les modèles sont sujettes à de multiples sources d'erreurs. L'erreur ici est définie comme la différence entre la valeur mesurée d'une variable et sa véritable valeur. À cause de ces multiples sources d'erreurs, une mesure d'une variable n'est jamais qu'une approximation de sa véritable valeur. L'erreur de mesure peut être décomposée en deux composantes, l'une étant systématique et l'autre aléatoire. L'erreur systématique est la même pour une répétition de la mesure avec le même instrument, méthode ou individu. L'erreur aléatoire varie entre différentes mesures. La quantification de ces deux sources d'erreurs et la propagation de cette erreur dans la modélisation est un défi car l'estimation de l'erreur en elle-même est rarement faite. Obtenir une estimation de l'erreur de mesure peut être coûteuse, par exemple dans le cas des échantillons de sol, il s'agit de répéter l'analyse plusieurs fois. C'est rarement le cas et l'erreur de mesure, si reportée, est souvent représentée par un terme d'erreur globale (p.ex. la moyenne des erreurs au carré dans Lagacherie et al., 2019), ce qui ne permet pas d'en

séparer la contribution des deux composantes, mais permet toutefois une propagation de l'erreur dans la modélisation qui utilise des données mesurées en entrées. Dans la cartographie numérique des sols, la plupart des données d'entrée sont soit mesurées au laboratoire soit prédites avec un modèle de spectroscopie à partir de données infrarouges. Les modèles de cartographie numérique sont généralement à base de géostatistique ou d'apprentissage automatique (voir Partie 2.3), mais **la prise en compte de l'erreur des données d'entrée dans la cartographie numérique n'est que très rarement effectuée**. J'ai tenté de d'inclure et de propager l'erreur des données d'entrée dans deux études, l'une avec un modèle d'apprentissage profond et la seconde avec un modèle géostatistique. Dans les deux cas, j'ai estimé l'erreur des mesures grâce à l'utilisation de la spectroscopie infrarouge.

Prendre en compte l'erreur des données de mesures pour la cartographie à l'aide des techniques géostatistiques et le krigeage est possible. Le krigeage fournit une interpolation exacte, c'est à dire que sur les points échantillonnés, la valeur sera préservée dans la prédiction. Il est néanmoins possible de modifier légèrement le système d'équations afin d'inclure un terme de variance supplémentaire (Delhomme, 1978) dans la diagonale de la matrice de covariance. Dans Ramirez-Lopez et al. (2019) nous avons propagé l'erreur des mesures dans l'analyse spatiale. L'erreur des mesures est obtenue à l'aide d'un modèle de spectroscopie. Nous avons défini l'erreur des mesures comme la variance des résidus obtenus entre la valeur mesurée en laboratoire et les valeurs prédite à l'aide de données infrarouges et d'un modèle statistique. Cette erreur est propagée dans la modélisation spatiale. La Figure 2.12 nous montre la prédiction spatiale de l'argile en utilisant i) les données de laboratoire uniquement et ii) les données de laboratoire avec en complément les données prédites par la spectroscopie. Dans ce dernier cas, l'erreur est propagée. Le résultat nous montre des cartes des propriétés du sol avec un motif beaucoup plus lisse. Cela était attendu dans la mesure où la prédiction à un emplacement échantillonné n'est pas l'observation elle-même. Le lissage augmente avec la taille de la variance de pépite du variogramme. La précision des cartes produites avec les données infrarouges était légèrement inférieure à celle des mesures en laboratoire, mais cet effet pourrait être atténué en tenant compte de l'erreur des mesures en laboratoire des propriétés du sol ; ce qui n'a pas été pris en compte dans notre étude de cas mais qui peut se révéler important (van Leeuwen et al., 2022).

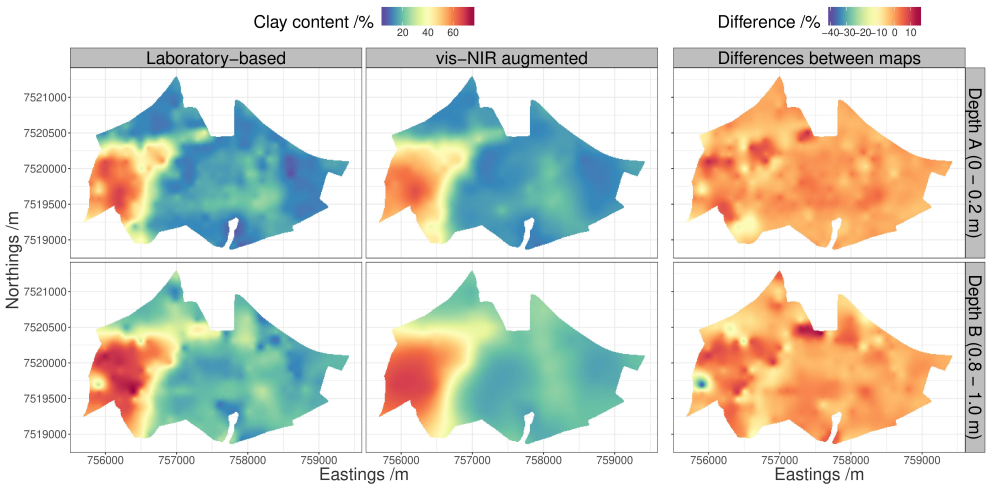

***Figure 2.12** – Cartes du pourcentage d'argile obtenu avec des données d'argile mesurées en laboratoire et avec des données de spectroscopie infrarouge. Les cartes à droite montrent la différences entre les deux. Notez que pour les cartes à base de données infrarouge l'incertitude des données d'entrée est propagée dans la prédiction. D'après Ramirez-Lopez et al. (2019).*

L'étude précédente illustre la propagation de l'erreur des mesures dans la prédiction géostatistique. Dans les méthodes d'apprentissage automatique, cependant, la prise en compte de cette erreur n'a été que peu

étudiée. Dans Wadoux (2019) je propose d'inclure l'erreur des observations dans l'étalonnage d'un modèle d'apprentissage profond de réseau de neurones convolutif. Une solution consiste à attribuer un poids d'erreur de mesure à chaque valeur de la propriété du sol, en fonction de son erreur relative par rapport à une « vraie » mesure de la propriété du sol au même endroit. Ici vrai est relatif, car nous considérons que les valeurs obtenues en laboratoire contiennent peu d'erreurs. Nous savons que cette approximation est à balancer en fonction du type de propriété étudié. Nous fournissons des poids différents en fonction de l'origine des prédictions (p.ex. un spectre proche ou moyen infrarouge) pour la prédiction du carbone organique du sol. Le modèle d'apprentissage profond est étalonné avec la minimisation d'une fonction objectif. Celle-ci est communément la moyenne des erreurs au carré pour les variables continues. Dans notre cas, nous avons modifié cette fonction pour y inclure le poids obtenu à l'étape précédente. Nous avons trouvé que le poids donné aux mesures obtenues pas des prédictions à partir de spectres proche infrarouge était moindre que celui donné aux données venant de spectre moyen infrarouge.

2.4.2 Incertitudes d'un modèle conceptuel

La régulation du cycle de l'eau est un service majeur assuré par les écosystèmes. Celui-ci s'évalue communément à partir de plusieurs fonctions, parmi lesquels le ruissellement, l'infiltration et la rétention de l'eau. Dans l'Union Européenne, la directive-cadre sur l'eau oblige tous les états membres à prendre des mesures pour garantir le bon état des masses d'eaux souterraines et de surface d'un point de vue chimique et écologique, et de les restaurer à une date cible définie. Les conséquences économiques et écologiques de toute intervention conçue pour améliorer la qualité des eaux de surface doivent être préalablement quantifiées. Ces quantifications sont généralement prévues à l'aide de modèles statistiques et informatiques qui simulent les processus hydrologiques dans l'ensemble du bassin versant (par exemple, ruissellement des précipitations, écoulement de l'eau et des nutriments dans les systèmes de drainage urbains et les rivières). Ces modèles sont affectés par de multiples sources d'erreurs, et il s'agit de développer des méthodes pour prendre en compte l'incertitude dans les observations en entrée, les paramètres et la structure des modèles utilisés. L'intérêt d'une telle démarche était de **développer des outils de quantification de l'incertitude basés sur les statistiques** pour aider à la prise de décision et d'optimiser les ressources nécessaires pour l'amélioration de ces modèles.

J'ai considéré un modèle conceptuel de pluie-débit qui s'applique à l'échelle du bassin versant pour l'estimation des débits d'une rivière en utilisant en entrée des séries temporelles de précipitation, température et d'évapotranspiration potentielle. **J'ai considéré les incertitudes liées aux données d'entrées (précipitations), à l'estimation des paramètres du modèle (p.ex. le coefficient de percolation), de leur état initial avant calibration, de la structure du modèle et finalement de l'incertitude dans les données de débit utilisées pour l'étalonnage des autres sources d'incertitudes**. Ces incertitudes sont intégrées individuellement dans un étalonnage utilisant des statistiques bayésiennes, c'est-à-dire utilisant des distributions de probabilité pour chaque source d'erreur. La corrélation temporelle de l'erreur est aussi prise en compte avec l'utilisation du filtre de Kalman. Finalement, les variations de la prédiction du modèle pluie-débit sont testées sous différents scénarios de prise en compte des différents types d'incertitudes (Figure 2.13) et pour trois approches. Les trois approches sont étudiées pour différent types de prise en compte de l'incertitude des précipitations. L'incertitude des précipitations est difficile à étudier dans des approches de prédiction car les distributions de probabilité postérieures ne peuvent pas être connus. Dans Wadoux et al. (2020a) j'ai testé i) une approche dans laquelle une correction linéaire est faite sur la distribution antérieure, ii) une approche dans laquelle la distribution postérieure est considérée comme étant la même que la distribution antérieure, et iii) une approche dans laquelle l'incertitude des précipitations est directement propagée dans la prédiction. Ces approches sont testées pour différentes densités de pluviomètres dans l'aire d'étude, c'est à dire pour différents niveaux d'incertitudes dans les données de

précipitation.

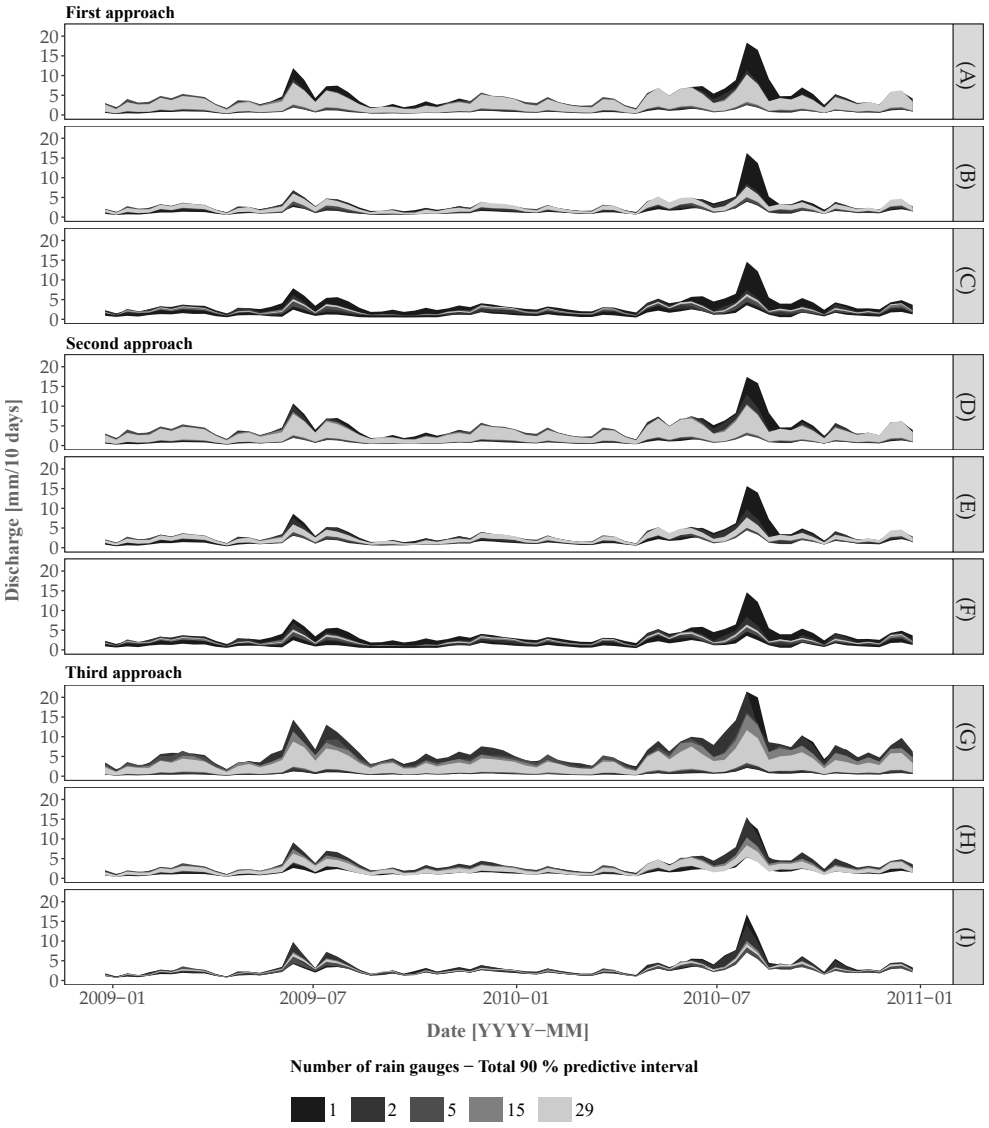

Figure 2.13 *– Prédiction du débit d'une rivière en utilisant trois approches pour les cas où (A, D, G) toutes les sources d'incertitudes sont prisent en compte, (B, E, H) l'incertitude de la structure du modèle est ignorée, (C, F, I) l'incertitude de la structure du modèle et des paramètres sont ignorés. Les trois approches sont différentes dans leur manière de prendre en compte l'incertitude des données d'entrée. D'après Wadoux et al. (2020a).*

Les résultats nous montrent que la contribution de l'incertitude des précipitations est assez petite et que l'incertitude des prédictions est dominée par l'incertitude des paramètres et de la structure du modèle. En conséquence, **la densité des pluviomètres nécessaire n'influe que de manière légère la qualité des prédictions : l'utilisation d'un seul pluviomètre se révèle suffisant pour la prédiction du débit de ce bassin versant de montagne**. Dans les approches testées, la structure du modèle avait un impact majeur dans l'incertitude totale de prédiction. Nous avons aussi conclu qu'en termes pratiques, nous ne ferions pas une grave erreur en prenant l'approche 3 et en n'étalonnant pas l'incertitude des précipitations comme paramètres supplémentaires dans l'analyse d'incertitude bayésienne. Cette conclusion est particulièrement valable lorsque le nombre de pluviomètres est supérieur à cinq dans notre étude de cas, mais cela pourrait ne pas être valable pour tous les cas. Dans toutes les approches, l'incertitude a été quantifiée avec précision, comme montrée par les figures de précisions (en anglais *accuracy plots*, pour plus d'informations sur ces figures, voir Wadoux et al. (2020a) ou Wadoux et al. (2018)). Cela signifie que malgré

des performances en prédiction plus faibles, l'approche 3 fournie toujours des informations précieuses car l'incertitude de prévision est assez faible. Dans la plupart des cas, les praticiens sont intéressés par une prédiction fiable du débit, mais compte tenu de la complexité de calcul des approches 1 et 2, les modélisateurs pourraient légitimement opter pour l'approche 3 avec des conséquences négligeables en termes de quantification de l'incertitude et une diminution relativement faible des performances de prédiction.

2.4.3 Incertitude de prédiction

Les techniques d'interpolation spatiale basées sur la géostatistique et le krigeage fournissent une quantification de l'incertitude de prédiction avec une mesure de la variance de l'erreur d'interpolation. Cette mesure peut être couplée avec une propagation de l'erreur des données d'entrée (voir aussi Partie 2.4.1). Cependant, **la quantification de l'incertitude de prédiction dans les méthodes d'apprentissage automatique reste un défi**. En cartographie numérique des sols, par exemple, des techniques basées sur la régression quantile de forêts ont été testées par Vaysse et Lagacherie (2017), et sont utilisées à de nombreuses reprises dans mes recherches (p.ex. Wadoux et al., 2023). Néanmoins, cette technique n'est applicable qu'aux forêts d'arbres décisionnels alors que de nombreuses autres techniques d'apprentissage automatique existent. Dans une première étude, j'ai utilisé les réseaux de neurones pour la cartographie numérique (Wadoux et al., 2019c). Les réseaux de neurones sont appréciés, car ils fournissent généralement des prédictions très précises mais ce sont des modèles complexes : ils comprennent potentiellement des milliers de paramètres pour une architecture de modèle qui a seulement quelques couches et neurones. Les méthodes développées dans la littérature scientifique pour quantifier l'incertitude des réseaux de neurones classent les techniques d'incertitude dans trois catégories : les techniques basées sur le bootstrap avec en complément un terme d'erreur des données, les méthodes bayésiennes et les méthodes basées sur la règle delta.

Dans Wadoux (2019), **j'ai adapté une technique de bootstrap avec un terme d'erreur des données pour obtenir l'intervalle de prédiction d'une cartographie numérique des propriétés du sol**. Les deux composantes de l'incertitude sont modélisées séparément, par la combinaison de méthodes de bootstrap et d'estimation par maximum de vraisemblance. Pour le bootstrap, cent modèles sont étalonnés pour la minimisation de l'erreur carré de prédiction. La variance est prise sur les 100 prédictions. Le second terme est obtenu par l'étalonnage d'un modèle supplémentaire pour la minimisation d'une fonction logarithme de maximum de vraisemblance. Les deux termes sont ensuite additionnés pour obtenir un intervalle de prédiction. La méthodologie est testée dans un scénario d'application potentielle pour la cartographie de l'argile de la couche arable, du limon et sable, du carbone organique, de l'azote total et du pH en France métropolitaine. Un seul modèle est utilisé fournissant une prédiction spatiale pour les six propriétés avec une quantification de l'incertitude, et en incluant des contraintes pour la prédiction des propriétés du sol corrélées (argile, limon et sable). J'ai ensuite validé les cartes des sols ainsi que l'incertitude. Un aperçu de l'architecture du modèle de neurones est fourni en Figure 2.14.

Les résultats de l'étude nous montrent que **l'incertitude de prédiction a été correctement estimée bien que légèrement sous-estimée**. C'est un résultat quelque peu attendu étant donné que la méthode se base sur le bootstrap et que certains sous-échantillons peuvent être biaisés. Néanmoins, la comparaison des résultats de cette étude avec d'autres études existantes utilisant la même base de données montre que l'incertitude est bien mieux estimée avec la méthode proposée. Un autre avantage est la prédiction simultanée de plusieurs propriétés de sol, ce qui en fait une approche particulièrement intéressante pour la prédiction des propriétés corrélées ou plus difficiles à prédire. Les statistiques de validation des prédictions nous montrent qu'ils sont aussi légèrement mieux qu'un modèle de forêts d'arbres décisionnels étalonné sur chaque propriété individuellement. Finalement, une évaluation visuelle des cartes nous confirme que les prédictions du modèle sont similaires aux cartes existantes pour ces même propriétés.

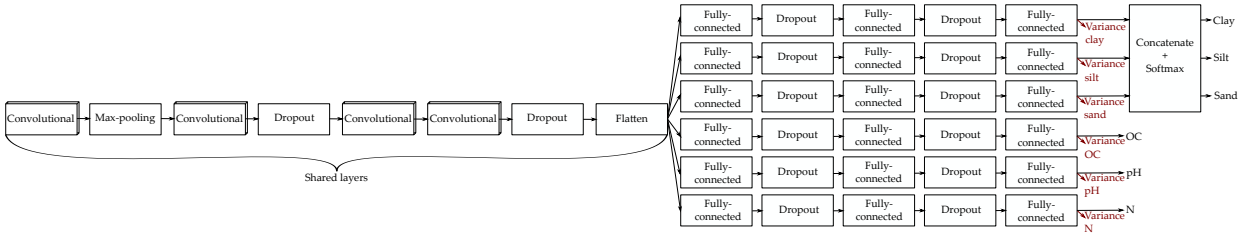

***Figure 2.14** – Représentation graphique du modèle de réseau de neurones développé dans Wadoux (2019). La partie représentée en noire montre la partie du modèle utilisé pour le bootstrap alors que l'ensemble (qui comprend la partie en rouge) est utilisée pour l'estimation du terme de variance.*

2.4.4 Incertitude dans l'agrégation spatiale

Plusieurs études reportent une estimation d'une agrégation spatiale, soit une moyenne ou un total, par exemple pour un bassin versant ou des espaces géographiques (p.ex. états, espaces bio-géographiques). Baccini et al. (2012) par exemple, a rapporté le carbone total stocké dans la végétation ligneuse vivante aérienne par pays et par région. Pour ce faire, ils ont calculé la somme de toutes les valeurs de pixel des stocks de carbone dans la zone et ont comparé leurs estimations avec les produits existants. En règle générale, l'agrégation spatiale par sommation ou moyenne est simple et effectuée en additionnant ou en faisant la moyenne des prédictions spatiales sur la zone d'intérêt. Cependant, la question de l'obtention de l'incertitude d'une agrégation spatiale est plus complexe. Il est incorrect d'estimer l'incertitude d'une moyenne spatiale en faisant la moyenne des incertitudes à tous les points sur lesquels la moyenne est calculée. Cela sous-estime fortement l'incertitude de l'agrégat spatial, car cela ignore le fait que les erreurs cartographiques s'annulent partiellement. Il s'avère que **l'incertitude d'un agrégat spatial dépend fortement du degré d'autocorrélation spatiale des erreurs cartographiques et qu'il faut donc en tenir compte lorsque l'incertitude d'une moyenne ou d'un total spatial est calculée**. Cela est bien connu des géostatisticiens avec le recourt à des techniques de krigeage de bloc (voir aussi Partie 2.3), mais cela est ignoré par une grande partie des cartographes en sol et écologie. Plusieurs raisons en sont la cause, mais la principale est la difficulté à utiliser des techniques de krigeage de bloc pour des grandes aires d'études, et l'utilisation renforcée ces dernières années des techniques d'apprentissage automatique pour la cartographie à large échelle spatiale. Ignorer la corrélation spatiale dans le calcul des agrégats peut avoir d'importantes conséquences, telle qu'une sous-estimation prononcée de l'incertitude des agrégats. En d'autres termes, les utilisateurs des cartes peuvent être trop confiants sur la qualité des estimations et cela peut se propager dans la prise de décision.

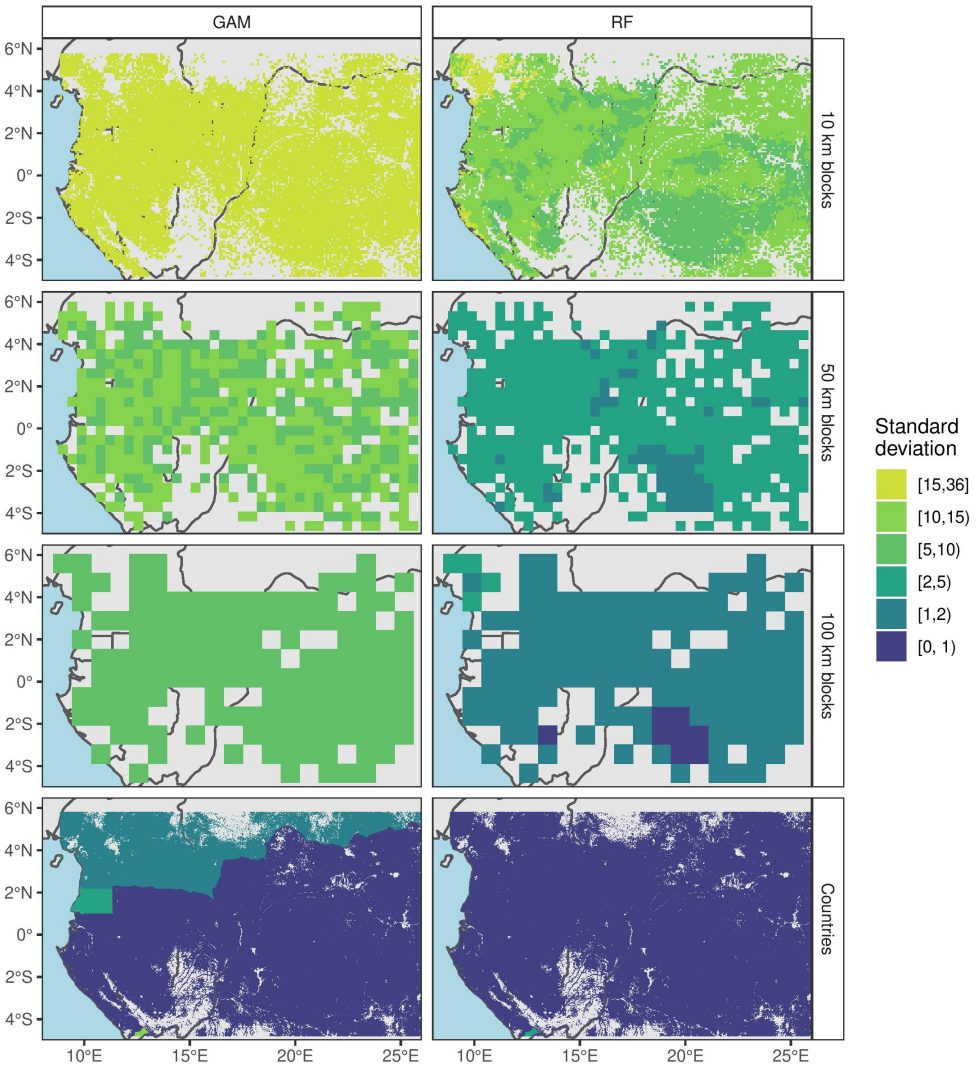

***Figure 2.15** – Cartes de l'écart type pour différents niveaux d'agrégation spatiale de la biomasse aérienne et obtenus par l'intégration Monte Carlo de la variance de l'erreur de prédiction de l'agrégat. D'après Wadoux et Heuvelink (2023).*

Je propose donc une méthode qui tienne compte de l'autocorrélation spatiale des erreurs cartographiques pour quantifier l'incertitude dans les moyennes et les totaux spatiaux, qui évite la complexité numérique du krigeage de bloc et qui est réalisable pour des études d'apprentissage automatique à grandes échelles. Dans Wadoux et Heuvelink (2023), je développe la méthodologie et la teste pour la cartographie de la biomasse aérienne en Afrique de l'ouest. La méthode suppose que les erreurs de prédiction de support ponctuel soient quantifiées, tiennent compte d'une variance non stationnaire des erreurs de prédiction mais suppose que leur corrélation spatiale est stationnaire. La fonction de corrélation est estimée à l'aide d'un variogramme standardisé et la variance de l'erreur de prédiction d'une somme ou totale est ensuite estimée à l'aide d'une intégration de Monte Carlo. Je rapporte l'écart-type de la moyenne du bloc pour différents niveaux d'agrégations et pour les deux modèles couramment utilisés en analyse spatiale (modèle additif généralisé - GAM et forêts d'arbres décisionnels - RF).

Nous avons testé l'agrégation sur des blocs de 10 km × 10 km, 50 km × 50 km, 100 km × 100 km et pour les six pays de la zone. Les cartes dans la Figure 2.15 montrent que les valeurs d'écart type de GAM sont supérieures à celles de RF. Dans tous les cas, l'incertitude diminue lorsqu'elle est agrégée pour des blocs plus grands. L'avantage de cette méthode est d'**éviter la complexité numérique de l'approche**

standard impliquant le krigeage de blocs. Celle-ci est réalisable pour des applications à grande échelle.

2.5 Interprétation et validation statistique des modèles

Dans une large proportion, nos modèles et nos cartes de prédiction de variables biophysiques peuvent être utilisés directement : pour les politiques publiques par des décideurs ou plus indirectement à travers l'information que ces modèles fournissent pour aider à la prise de décision sur des questions, par exemple, de pollution des sols ou de subventions agricoles. **L'évaluation des modèles et des cartes est une étape avant leur utilisation potentielle. Il s'agit d'établir la validité du modèle ou de la carte, ce qui signifie comprendre si le modèle est adapté à l'usage**. Cela peut être fait de plusieurs manières. La plus courante est l'utilisation de statistiques de validation qui comparent la prédiction avec des observations indépendantes. Un autre type de validation est de comprendre comment le modèle est arrivé à une prédiction ; le modèle prédit-il pour les bonnes raisons ? Tous les processus ou facteurs pertinents sont-ils inclus dans le modèle ? J'ai abordé ces questions avec plusieurs points de vue : à travers le développement de techniques statistiques pour l'interprétation de modèles complexes, à travers le développement d'un diagramme pour la représentation intégrée de plusieurs statistiques de validation, et finalement à travers une discussion et une proposition de méthodes pour prendre en compte la corrélation spatiale dans l'estimation des statistiques de validation.

2.5.1 Interprétation statistique des modèles complexes

Depuis le début des années 80 et l'essor des techniques de géostatistique pour la cartographie numérique des sols, l'attention s'est progressivement déplacée vers des outils statistiques et algorithmiques plus complexes du domaine de l'apprentissage automatique et profond. J'ai participé à ces développement à travers plusieurs études (voir la Partie 2.3). La précision de ces modèles pour la prédiction de propriétés du sol est souvent supérieure à celle des modèles classiques utilisés en modélisation spatiale. Ils sont également particulièrement utiles dans les situations où la relation entre la propriété du sol et les covariables environnementales est trop complexe pour être modélisée de manière mécaniste ou avec des modèles statistiques simples. Cependant, l'essor des modèles complexes de variation du sol s'est faite au détriment de la compréhension des mécanismes de variation du sol et de ses propriétés : **il est difficile d'obtenir un aperçu du fonctionnement et de la structure interne de ces modèles, de sorte que ceux-ci sont souvent appelés des « boîtes noires ».** Des exemples de ces modèles sont les forêts d'arbres décisionnels, les machines à vecteurs de support et les réseaux de neurones artificiels. En science des sols, plusieurs tentatives ont été faites pour obtenir des informations sur la structure de ces modèles. La plus commune est de fournir une estimation de l'importance des covariables dans la prédiction. D'autres techniques existent, tel que l'algorithme de Garson. Bien qu'utiles pour obtenir des informations sur des modèles complexes de variation du sol, ces méthodes sont spécifiques à un modèle, c'est-à-dire qu'elles empêchent la comparaison entre des modèles différents (p.ex. forêts d'arbres décisionnels et machine à support de vecteurs). Plusieurs nouvelles méthodes dites « indépendantes », c'est à dire pouvant être appliquées à tout modèle, ont récemment été développées dans la littérature statistique.

Dans un cadre de cartographie numérique des sols, deux types d'interprétations peuvent être distinguées : locale et globale. Une interprétation locale est appropriée lorsque l'objectif est d'évaluer comment la prédiction à un seul emplacement spatial est faite. Il est en effet raisonnable de supposer que l'importance de certains facteurs environnementaux varie d'un endroit à l'autre et d'un paysage à l'autre. Une interprétation globale, à l'inverse, donne un aperçu du fonctionnement global du modèle. Les méthodes globales

exposent l'importance de chaque facteur de variation spatiale, leur interaction, ainsi que la forme fonctionnelle de l'association entre la covariable environnementale et la propriété du sol. En pratique, des méthodes globales et locales sont utilisées conjointement pour interpréter et visualiser des aspects différenciés du modèle.

Synergies entre méthodes C'est dans un cadre que j'ai fait une étude sur **l'utilité des méthodes ainsi que les avantages et inconvénients entre méthodes d'interprétation locales et globales à utiliser dans les études cartographiques** (Wadoux et Molnar, 2022). De telles méthodes peuvent être appliquées à n'importe quel modèle (c'est-à-dire qu'elles sont indépendantes du modèle), bien qu'en pratique, il ne soit pas toujours judicieux de les appliquer sur des modèles simples dont la structure est facilement compréhensible (p.ex., la régression linéaire). J'illustre les méthodes d'interprétation de deux modèles dans une étude de cas sur la cartographie du carbone organique de la couche arable en France. Je présente huit méthodes : la permutation, l'espérance conditionnelle individuelle, les figures de dépendance partiel, l'effet local accumulé, l'interaction avec le statistique H, la modélisation de substitut, et enfin les valeurs de Shapley.

Le cas d'étude et l'utilisation des méthodes d'interprétation ont fourni des informations précieuses sur les facteurs de variation du carbone organique en France, leur interaction, ainsi que sur la forme fonctionnelle de l'association entre les covariables environnementales et le carbone organique. Ces informations ont été obtenues soit pour un emplacement géographique unique, soit globalement à partir du modèle dans son ensemble. Par exemple, Figure 2.16 nous montre le profil de dépendance du carbone organique à l'altitude, pour trois techniques d'interprétation différentes. Bien que l'interprétation de ces valeurs diffèrent entre les techniques, elles nous montre un même profil dans les trois cas.

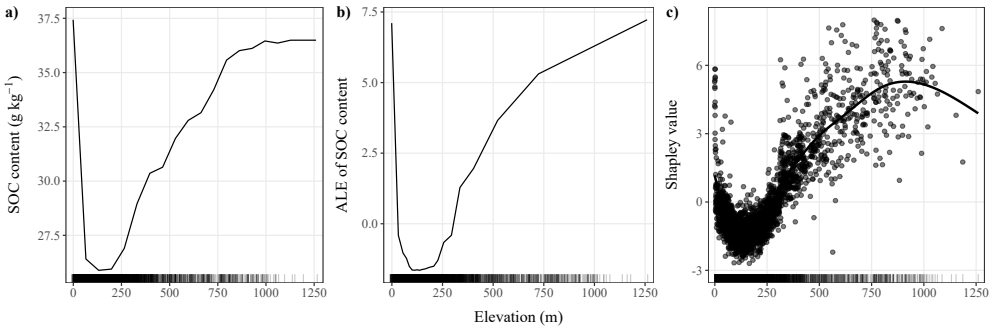

***Figure 2.16** – Effet de l'altitude sur le carbone organique des sols en France et estimé avec a) la dépendance partielle, b) l'effet local cumulé et c) les valeurs de Shapley. L'axe des abscisses montre la distribution marginale de l'altitude dans l'ensemble des données d'étalonnage. En c), les points noirs représentent les valeurs de Shapley individuelles et la courbe noire est une ligne lissée obtenue sur les valeurs de Shapley avec une fonction moyenne conditionnelle. A noter que ces résultats ont été obtenus avec le modèle de forêts d'arbres décisionnels. D'après Wadoux et Molnar (2022).*

Les valeurs de Shapley Une des conclusions de l'étude reportée dans Wadoux et Molnar (2022) est que l'utilisation des valeurs de Shapley pour interpréter des modèles complexes de variation du sol est une future ligne de recherche prometteuse. Les valeurs de Shapley sont polyvalentes : elles permettent une interprétation à la fois locale et globale, elles sont faciles à interpréter et ont une théorie sous-jacente. Dans Wadoux et al. (2022a), j'approfondis donc le sujet et cherche à comprendre **comment les valeurs de Shapley peuvent aider à expliquer la relation trouvée par un modèle d'apprentissage automatique complexe entre une propriété du sol et des covariables environnementales pour une grande aire d'étude**. Dans une étude de cas en France métropolitaine, j'utilise des valeurs de Shapley sur un modèle de forêt d'arbres décisionnels avec un grand nombre d'arbres, et je décris comment les valeurs sont utilisées pour i) comprendre l'importance moyenne des variables dans la prédiction des stocks

de carbone organique des sols, ii) obtenir des informations sur la variation spatiale de l'importance de chaque variable, c'est-à-dire comment l'importance varie localement, dans l'espace et par zones carbone-paysage, et iii) déduire la forme fonctionnelle de l'association entre le stock de carbone organique des sols et les facteurs environnementaux. Les valeurs de Shapley sont additives, c'est à dire qu'elles peuvent être additionnées par groupe de variable ou géographiquement. J'en tire profit dans cette étude et propose une interprétation de l'importance des variables par zone carbone-paysage et par groupes de covariables classées par catégories (c.a.d. conditions climatique moyenne, saisonnalité, conditions climatiques extrêmes, topographie, sol et organismes/végétation).

Un exemple de représentation spatiale des valeurs de Shapley est fournie en Figure 2.17. Cette figure nous montre à gauche la covariable la plus importante pour la prédiction du stock de carbone organique des sols, et à droite la proportion de cette même covariable dans le total de la prédiction. Les valeurs de Shapley ont révélé dans cette étude que la contribution des covariables à la prédiction des stocks de carbone organique des sols variait considérablement entre les emplacements géographiques et entre les zones carbone-paysage. **Les résultats suggèrent des relations entre les covariables environnementales et les stocks de carbone organique des sols qui ont été abondamment documentées dans la littérature**. Par exemple, dans un test de prédiction des stocks de carbone organique des sols à deux emplacements géographiques avec des valeurs de stock similaires mais des environnements très différents, nous avons obtenu des valeurs de Shapley qui montrent une contribution individuelle des covariables à la prédiction qui est cohérente avec nos connaissances pédologiques sur les processus de stockage du carbone. **D'autres relations, au contraire, mettent en évidence les limites de la modélisation empirique** pour la prédiction des stocks de carbone organique des sols.

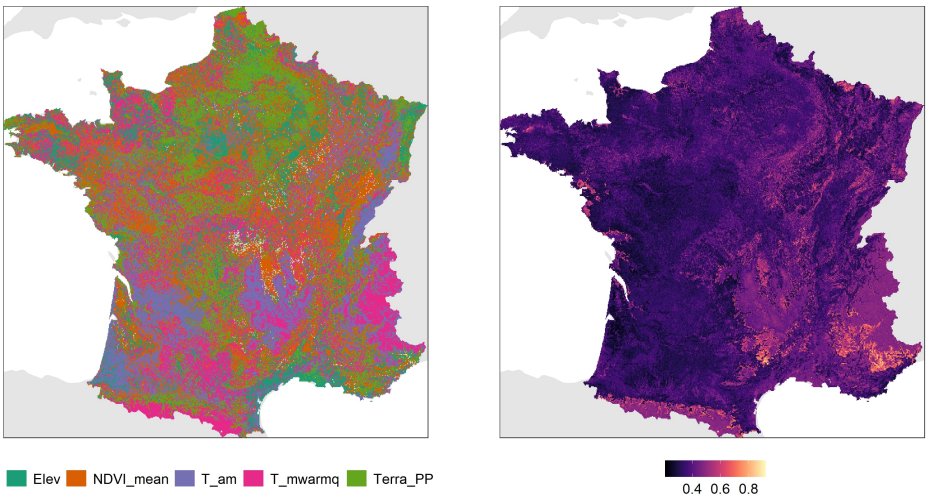

***Figure 2.17** – Carte de la covariable la plus importante contribuant à la prévision du stock de carbone organique des sols (à gauche) pour cinq covariables (sur 24), et proportion de cette covariable par rapport à la prévision du stock de carbone organique totale (à droite). D'après Wadoux et al. (2022a).*

2.5.2 Évaluation des cartographies numériques

La détermination de la qualité des cartes pédologiques est un domaine de recherche depuis de nombreuses années. Les premières études visaient à évaluer la qualité des cartes de sols conventionnelles (discrètes ou thématiques) avec des statistiques de précision, par exemple avec des indices de qualité du point de vue de l'utilisateur par rapport au producteur ou avec la théorie de l'information. Plus récemment, des statistiques ont été développées pour évaluer la variance de krigeage qui combine la prédiction erreur et la variance de l'erreur de prédiction. En cartographie numérique des sols, la qualité de la carte est généralement évaluée

avec des indices statistiques globaux de l'erreur calculés par la comparaison de paires des prédictions et d'observations de la propriété cible. Ces indices sont globaux car ils résument la qualité de l'ensemble de la carte en un seul indice statistique, c'est-à-dire qu'ils quantifient la précision globale de tous les emplacements de la zone de cartographie pris ensemble. Souvent, ces indices de qualité de carte sont basés sur des statistiques de corrélation (par exemple, le coefficient de corrélation de Pearson ou le coefficient de corrélation de concordance) ou des indices d'erreur moyens (par exemple, l'erreur moyenne ou l'erreur quadratique moyenne). Ils sont valable et utiles pour l'évaluation d'aspects spécifiques de la qualité des cartes, mais un certain nombre d'outils développés en dehors de la science des sols sont disponibles pour permettre d'évaluer simultanément plusieurs indices statistiques avec des diagrammes récapitulatifs. J'ai développé un de ces outils, qui fournit une approche intégrée à l'évaluation des cartes quantitatives des sols.

Dans Wadoux et al. (2022b) je me suis basé sur le diagramme de Taylor, largement utilisé en climatologie et hydrologie pour l'évaluation des prédictions. Le diagramme de Taylor reconnaît une relation entre l'indice de corrélation, l'écart type de l'erreur et l'écart type des valeurs observées et prédites à travers la loi des cosinus. Cela permet de représenter ensemble ces indices de manière graphique dans un système de coordonnées polaires. L'inconvénient de cette représentation est l'absence de l'erreur moyenne, qui fournit une estimation du biais de la carte. Dans ce cadre, j'ai développé **une représentation graphique alternative qui inclue l'écart type de l'erreur, l'erreur moyenne et l'erreur quadratique moyenne, qui sont visualisés ensemble grâce à la décomposition de l'erreur quadratique moyenne en deux éléments**. C'est la base de la Figure 2.18 : l'axe des abscisses nous montre le biais et l'axe des ordonnées nous montre l'écart type de l'erreur. Un point sur la figure est à une distance de l'origine exprimée en terme d'erreur quadratique moyenne. Enfin, la relation statistique de ces éléments avec deux autres indices : la corrélation et le coefficient d'efficacité de la modélisation, nous permet de les placer sur la figure. La figure 2.18 nous montre cinq indices statistiques de manière simultanée, chacun représentant un aspect spécifique de la qualité de la carte.

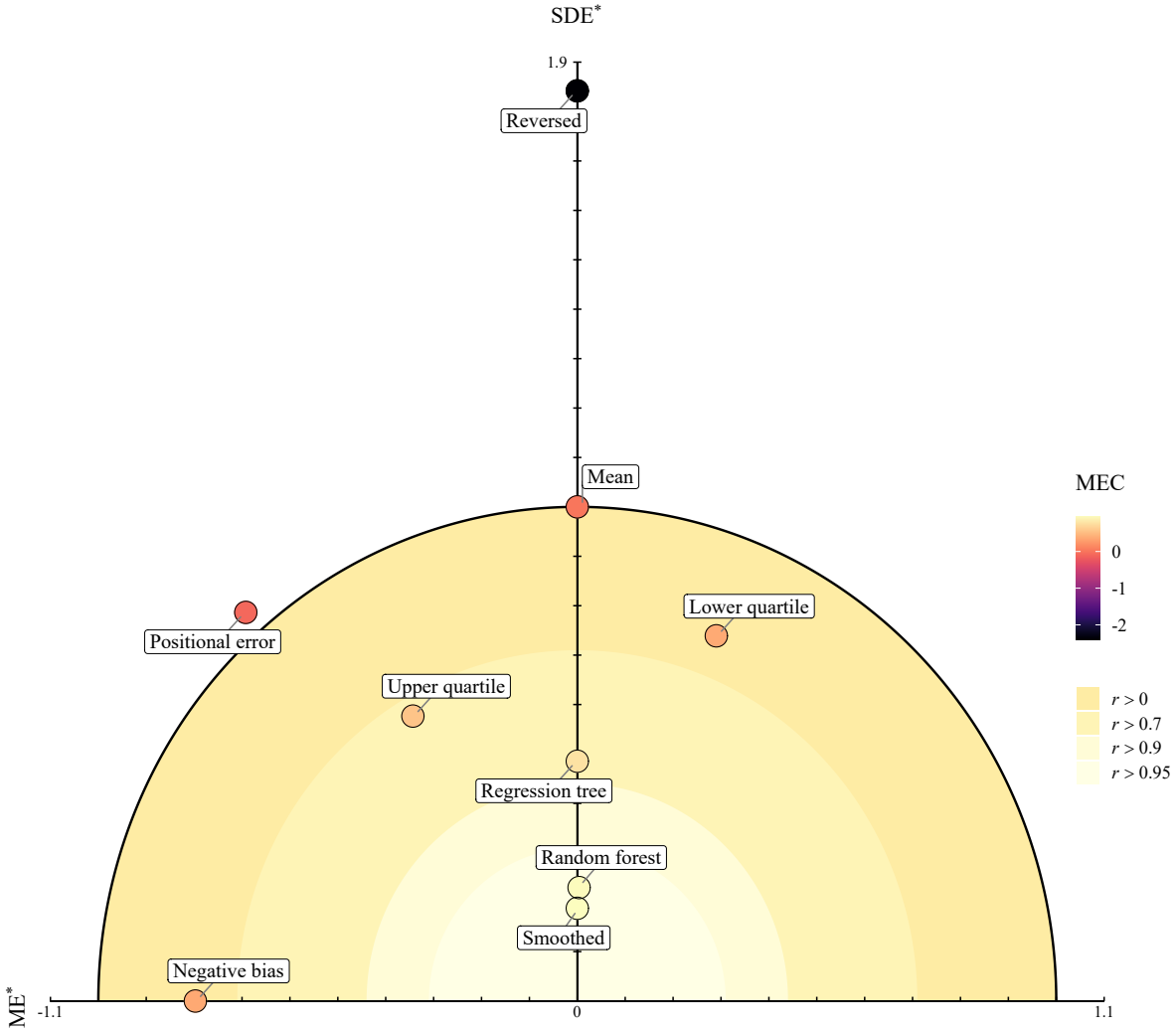

Figure 2.18 *– Diagramme solaire normalisé rendant des cartes simulées et la carte de référence. Les points à l'intérieur de la ligne noire épaisse à RMSE* = 1 indiquent le seuil pour lequel la carte est meilleure que la moyenne de la carte de référence prise lors de la prédiction. Le cercle extérieur et les cercles intérieurs délimitent les zones où la corrélation r est supérieur à 0, 0,7, 0,9 et 0,95. L'échelle de couleur sur les points représente le coefficient d'efficacité de modélisation (MEC) entre la carte de référence et la carte simulée. D'après Wadoux et al. (2022b).*

La figure 2.18 nous montre des cartes numériques des sols simulées et comparées à une carte de référence. Par exemple, la carte « negative bias » contient un biais et se trouve donc entièrement sur l'axe des abscisses : l'erreur quadratique moyenne de la carte est entièrement représentée par son biais. Cette carte a une valeur de corrélation au moins positive et une valeur de coefficient d'efficacité de modélisation elle aussi positive. Ce diagramme est particulièrement **utile pour comparer plusieurs cartes ou des prédictions d'un modèle lors d'un étalonnage. Il permet aussi de voir le mérite relatif de plusieurs cartes et de faire un choix raisonné sur la qualité de chacune d'elles.**

2.5.3 Critique de la validation sans échantillon probabiliste

De nombreuses études utilisent des méthodes empiriques, (géo)statistiques et d'apprentissage automatique pour spatialiser des variables environnementales. Dans ces études, la précision de la carte est généralement évaluée à l'aide de mesures statistiques qui évaluent la proximité des prédictions avec la réalité pour un ensemble d'emplacements dans la zone d'intérêt. Ces mesures statistiques sont par exemple la moyenne des erreurs au carré (voir plusieurs statistiques de validation utilisés dans Partie 2.5.2). Ces indices sont calculés soit en collectant un nouvel ensemble de données, soit en utilisant l'ensemble de données existantes pour l'étalonnage et la validation du modèle. La collecte d'un nouvel ensemble de données se fait idéalement par échantillonnage probabiliste. Les données d'échantillons sont utilisées pour estimer les indices de précision de la carte par une inférence statistique basée sur le plan qui vient de la théorie classique de l'échantillonnage. Cette méthode est solide sur le plan statistique et a été décrite en détail dans la littérature statistique et environnementales Une alternative est d'évaluer la précision de la carte par validation avec les données utilisées pour étalonner le modèle de cartographie sous-jacent. Dans ce cas, l'ensemble de données existant est divisé en deux sous-ensembles appelés plis d'étalonnage et de validation. Le pli d'étalonnage est utilisé pour étalonner un modèle de cartographie et faire des prédictions tandis que le pli de validation est utilisé pour estimer les indices de précision de la carte. Cette procédure peut être répétée plusieurs fois, comme dans le bootstrap lorsque plusieurs échantillons bootstrap avec remplacement sont utilisés pour l'étalonnage et la prédiction, ou comme dans la validation croisée.

Plusieurs études récentes, cependant, ont soutenu que la validation statistique des cartes devrait tenir compte de l'autocorrélation spatiale entre les points d'échantillons. Les données recueillies à des points géographiquement proches sont généralement plus similaires qu'à des points géographiquement éloignés. En conséquence, ces études affirment que les indices de précision des cartes tels que dérivés à l'aide de la validation croisée standard sont biaisés car les points d'étalonnage ne sont pas statistiquement indépendants des points de validation. Cette conception de la validation du modèle a conduit au développement récent de techniques de validation croisée qui évitent l'autocorrélation spatiale, telles que la validation croisée spatiale K-fold et la validation croisée bufferisée excluant chaque point un par un. **J'ai voulu montrer que cela donnait lieu à une conception erronée de la validation statistique des cartes dans un contexte spatial**. Dans Wadoux et al. (2021a) j'explique pourquoi nous ne devrions pas utiliser les techniques de validation croisée spatiale pour estimer les indices de précision des cartes et adhérer à la place à des méthodes de validation statistiquement rigoureuses via l'échantillonnage probabiliste et l'inférence basée sur le plan.

Je le démontre dans un cas d'étude simple dans lequel je teste plusieurs méthodes de validation pour l'évaluation d'une carte de biomasse aérienne en Amazonie réalisé à l'aide d'un modèle de forêt d'arbres décisionnels. Je teste trois bases de données spatiales d'étalonnages : (i) échantillonnage aléatoire systématique (ii) échantillonnage aléatoire simple et (iii) un échantillonnage aléatoire en grappes à deux étapes. Pour chaque base, je réalise ensuite une validation avec plusieurs techniques : validation croisée aléatoire, spatiale et spatiale avec l'exclusion d'un point à la fois, et je les compare à une validation par échantillon probabiliste et inférence basée sur le plan. Les résultats sont présentés pour l'estimation de l'erreur moyenne au carré dans la Figure 2.19. La figure montre que l'estimation basée sur le plan de l'erreur moyenne au carré n'a pas de biais et que les estimations varient peu. Ce sont des propriétés attrayantes qui montrent la supériorité de la validation basée sur le plan pour évaluer les performances de prédiction du modèle de cartographie et estimer la précision de la carte, mais comme indiqué précédemment, cela nécessite un échantillonnage probabiliste de la population. La validation croisée à l'aide d'une validation croisée aléatoire standard est presque sans biais pour les plans d'échantillonnage aléatoires systématiques et simples, mais trop optimiste dans le cas de l'échantillonnage en grappes. Les deux méthodes de validation croisée spatiale sont trop pessimistes.

Cela montre que **les études de validation croisée spatiale propagent une idée fausse très répandue sur la validation statistique des cartes**. Idéalement, la précision de la carte devrait être estimée avec un échantillonnage probabiliste et une inférence statistique basée sur le plan. Une telle méthodologie et inférence est valide sans qu'il soit nécessaire d'ajuster l'autocorrélation spatiale dans les données. Les scientifiques et les utilisateurs peuvent procéder en toute confiance en sachant que la précision de la carte obtenue par les méthodes standard basées sur la théorie de l'échantillonnage est valide.

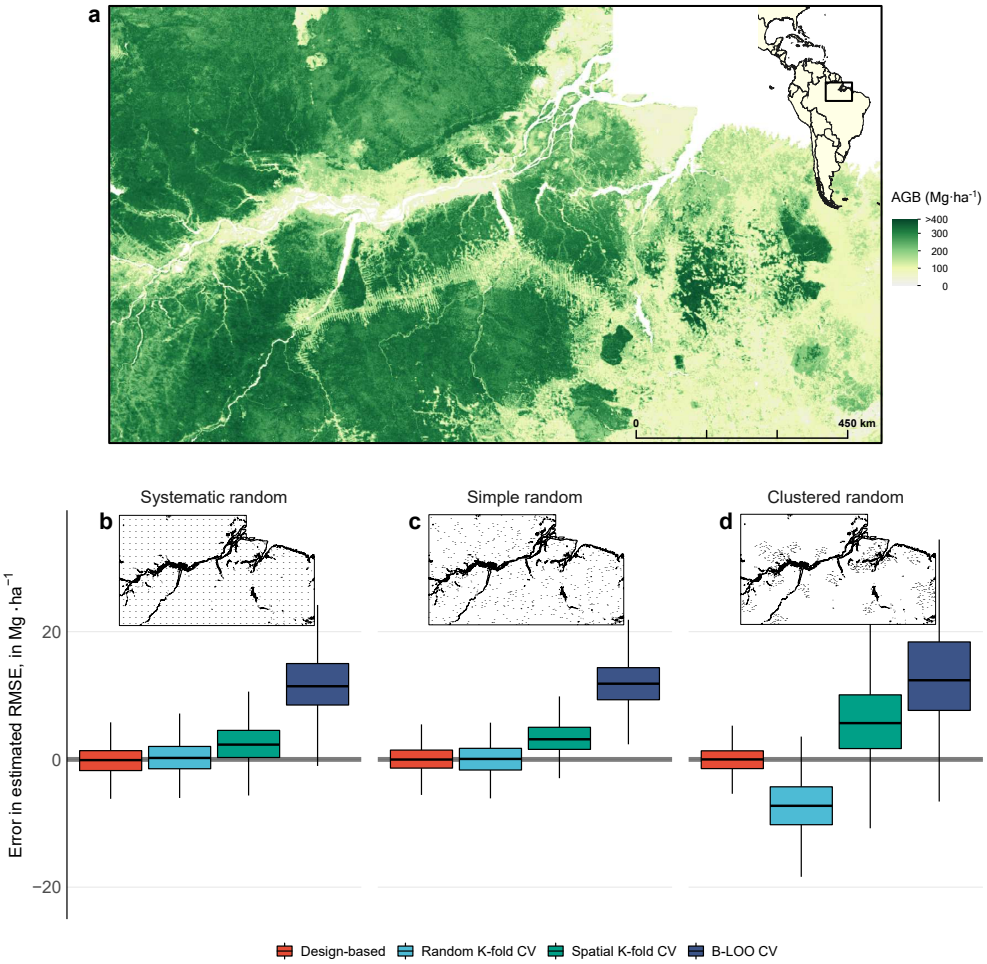

Figure 2.19 *– Aperçu de la zone d'étude et résultats de l'évaluation des stratégies de validation.* ***a*** *Zone d'étude dans le bassin amazonien avec les valeurs de la biomasse aérienne, selon la carte de Baccini (Baccini et al., 2012).* ***b-d*** *Erreur dans les estimations de du RMSE de la population (en Mg·ha^{-1}) pour les échantillons de calibrage collectés par échantillonnage aléatoire systématique (**b**), aléatoire simple (**c**) et aléatoire en grappes en deux étapes (**d**). Notez que la ligne grise horizontale à 0 dans **b-d** fait effectivement référence à au RMSE de la population, car les écarts par rapport au RMSE de la population sont tracés. Les emplacements d'échantillonnages indiqués sur les cartes dans **b-d** sont une réalisation sur 500.*

2.5.4 Estimation des statistiques de validation avec un échantillonnage en grappes

La critique de la validation des cartes sans échantillon probabiliste de la Partie 2.5.3 ne résout pas la question de l'impact de la corrélation spatiale dans l'estimation des statistiques de validation. J'ai montré que la validation croisée spatiale n'est pas la solution car elle commence sur un postulat de départ faux (c.a.d. les échantillons de validation proches géographiquement des échantillons d'étalonnage doivent être exclus de l'analyse) et qu'elle fournit des statistiques de validation largement pessimistes. Dans le cas

d'une estimation des statistiques de validation fondée sur le plan à partir d'un échantillon probabiliste, les estimateurs et leur variance ne sont pas biaisés, quelle que soit l'ampleur de la corrélation spatiale. Si les points de données sont assez uniformément répartis dans l'espace, la validation croisée conventionnelle k-fold produit des résultats raisonnables, mais les estimations des statistiques de validation de la carte peuvent être biaisées et aucun intervalle de confiance ne peut être dérivé. Les points fortement regroupés en grappes peuvent ne pas être représentatifs de l'ensemble de la zone d'étude car ils sur-représentent certaines régions tout en sous-représentant ou même en manquant d'autres. Cela implique que les modèles d'apprentissage automatique sont étalonnés intensément sur les zones densément échantillonnées qui ont également le plus grand impact sur la précision estimée de la carte. Les estimations conventionnelles de la précision des cartes de validation croisée basées sur de tels échantillons préférentiels ont tendance à être trop optimistes et des méthodes sont nécessaires pour corriger cela.

Dans la littérature scientifique, peu de méthodes ont été proposées en dehors de la validation croisée spatiale. **Il s'agit de s'attaquer au biais dans les estimations des statistiques de validation des cartes à partir de la validation croisée** en équilibrant l'impact des résidus dans les régions avec différentes intensités d'échantillonnage. Nous présentons donc et évaluons des approches alternatives de validation croisée pour évaluer la précision de la carte thématique lorsque les données de l'échantillon sont regroupées (De Bruin et al., 2022). La première méthode proposée est une approche de quasi-randomisation utilisant une pondération inverse de l'intensité d'échantillonnage pour corriger le biais de sélection en donnant plus de poids aux observations dans les zones peu échantillonnées et moins de poids aux observations dans les zones densément échantillonnées. Les deux autres approches sont des méthodes basées sur des modèles ancrées dans la géostatistique. Ceux-ci tiennent compte des informations redondantes des résidus spatialement regroupés en utilisant des fonctions de corrélation spatiale (variogrammes). La première variante suppose l'homoscédasticité des résidus, tandis que la seconde tient compte de l'hétéroscédasticité des résidus, toujours en fonction de l'intensité d'échantillonnage. Nous expliquons comment ces méthodes fonctionnent et comparons leurs estimations de précision de la carte et celles de la validation croisée conventionnelle k-fold et de la validation croisée spatiale bloquée par rapport aux mesures de précision de la carte de référence. Une véritable démonstration de notre approche nécessiterait que les vraies valeurs de la variable cible soient connues partout, ce qui est irréalisable dans la réalité. Pour imiter une situation de données de référence connues en tout point, les variables environnementales cibles ont été échantillonnées à partir de cartes existantes de la biomasse aérienne (AGB) et du stock de carbone organique du sol (OCS) et les échantillons acquis ont été utilisés pour l'ajustement et la prédiction avec des modèles de forêts d'arbres décisionnels.

Les résultats sont présentés dans la Figure 2.20. Le biais dans les statistiques de validation de la carte évalué sur plusieurs réalisations des plans d'échantillonnage par validation croisée pondérée était beaucoup plus faible que celui de la validation croisée spatiale bloquée pour les échantillons non regroupés à modérément regroupés. Pour l'échantillonnage en grappes où de grandes parties des cartes étaient prédites par extrapolation, la validation croisée spatiale bloquée était la plus proche des mesures de précision de la carte de référence. Cependant, la validation croisée spatiale bloquée peut toujours produire des estimations biaisées des statistiques de validation de la carte, car il est impossible de déterminer la taille des blocs pour s'adapter à la détérioration des statistiques de validation de la carte causée par l'extrapolation. Au contraire, l'extrapolation doit être évitée par un échantillonnage supplémentaire ou une limitation de la zone de prédiction.**Nous recommandons une validation croisée aléatoire conventionnelle pour les échantillons avec une couverture spatiale modérée et une validation croisée pondérée pour les échantillons modérément en grappes.**

Ces conclusions ont par ailleurs été utilisées pas la doctorante que je supervise dans la validation de cartographie numérique des sols pour laquelle des échantillons en grappes étaient disponibles. Les résultats sont publiés dans Nenkam et al. (2022).

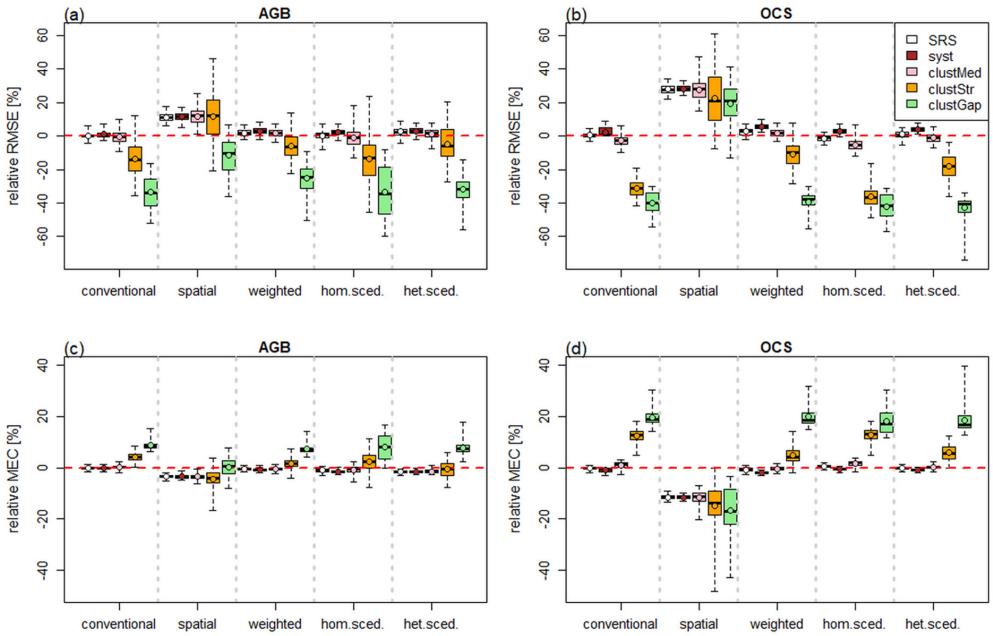

Figure 2.20 *– Écart relatif de (a, b) et (c, d) par rapport à leurs métriques de référence pour l'AGB (a, c) et l'OCS (b, d) pour les modèles entraînés sur 100 réalisations des plans d'échantillonnage explorés, selon la validation croisée. Méthodes répertoriées le long de l'axe des abscisses : (hom.sced = basé sur un modèle homoscédastique ; het.sced = basé sur un modèle hétéroscédastique). D'après De Bruin et al. (2022).*

2.6 Une perspective pluridisciplinaire et épistémologique

Des techniques novatrices accompagnent les développements en modélisation spatiale. Pour la collection de données, les techniques de chimiométrie faisant appel à la spectroscopie infrarouge sont désormais conventionnelles en science des sols. Ces techniques participent à satisfaire la demande croissante pour de nouvelles données, moins précises mais plus abondantes et rapide à obtenir. De nouveaux acteurs apparaissent également pour la collection de données, notamment à travers les approches participatives. Les approches participatives on été considérées récemment pour la collection de données dans les projets environnementaux. Elles reconnaissent les intérêts croissants des non-spécialistes dans les sciences, mais aident aussi à démocratiser les connaissances scientifiques. **Ces nouvelles techniques et ces nouveaux acteurs sont à analyser conjointement avec les approches existantes ou émergentes en modélisation spatiale. Ils forment des nouveaux enjeux qui sont tous permis et amplifié par la convergence numérique.** Dans les parties qui suivent, je décris mes recherches sur ces nouveaux enjeux, aussi bien d'un point de vue méthodologique d'épistémologique, dans un but de définir une vision moyen/long-terme des défis du numérique.

2.6.1 L'apport des données spectrales

Les données spectrales peuvent provenir de différents capteurs et longueurs d'onde : rayons gamma, rayons X et infrarouges, entre autres, et de scans effectués soit à proximité du sol - sur le terrain ou en laboratoire, soit à distance, par exemple, lorsque le capteur est embarqué dans un avion ou un satellite. L'utilisation de données spectrales pour caractériser les propriétés chimiques, minéralogiques, biologiques

et physiques du sol a gagné en popularité dans la recherche pédologique. Les progrès des capteurs et des logiciels se produisent à un rythme rapide. La détection du sol, en particulier l'utilisation de la spectroscopie du sol, est maintenant largement disponible en utilisant une gamme de modalités et de longueurs d'onde à travers le spectre électromagnétique. La signature spectrale concerne les caractéristiques du sol telles que les composants organiques et minéraux. Les mesures spectroscopiques sont rapides, économiques et non destructives et peuvent être effectuées à la fois en laboratoire et in situ sur le terrain. La composition et les caractéristiques du sol sont dans le spectre à des longueurs d'onde spécifiques du spectre électromagnétique. Par exemple, les spectres infrarouge moyen ont des informations sur la minéralogie du sol ou la composition de la matière organique du sol, qui peuvent être évaluées quantitativement ou qualitativement en utilisant l'absorption ou la réflectance à des longueurs d'onde spécifiques. La gamme visible et infrarouge du spectre électromagnétique a suscité beaucoup d'intérêt en science des sols. **La mesure du spectre infrarouge d'échantillons de sol permet la quantification de plusieurs propriétés du sol à partir de leur réponse spectrale de manière plus rapide et moins chère que par les méthodes conventionnelles d'analyses de sol**. De plus, l'enregistrement d'un spectre infrarouge n'utilise aucun réactif chimique et peut être réalisée aussi bien en laboratoire que pour des analyses de sol sur le terrain. Les spectres infrarouges sont sensibles aux matériaux organiques et inorganiques du sol, ce qui fait de la spectroscopie un excellent outil pour l'évaluation quantitative du sol. La gamme infrarouge moyen (MIR) du spectre, en particulier, contient plus d'informations et d'informations directes sur les composants organiques et minéraux du sol que la gamme visible et proche infrarouge (vis-NIR). Par exemple, divers composants de la matière organique du sol ont une signature spectrale très distincte dans le domaine de l'infrarouge moyen. La raison en est que les vibrations moléculaires fondamentales se produisent dans la gamme de l'infrarouge moyen, tandis que les harmoniques et les combinaisons se produisent dans le vis-NIR. En pratique, cela signifie que les caractéristiques d'absorption détectées dans le vis-NIR sont moins nombreuses, plus larges et plus complexes que celles enregistrées dans l'infrarouge moyen.

Alors que la spectroscopie est utilisée en science des sols depuis les années 1950, les deux dernières décennies ont vu une augmentation de son utilisation, en particulier la spectroscopie vis-NIR et MIR, pour remplacer et compléter les analyses de sol. Cette augmentation a été soutenue par le développement de la chimiométrie (l'application de méthodes mathématiques et statistiques à l'analyse des données chimiques), l'analyse statistique multivariée et l'augmentation des moyens informatiques. Les propriétés du sol ont des schémas d'absorption complexes. Les bandes spectrales infrarouges sont en grande partie non spécifiques (c'est-à-dire qu'elles ne sont pas linéairement liées à une seule propriété du sol) et se chevauchent entre les propriétés. Ceci est particulièrement important dans la gamme vis-NIR des spectres. Pour obtenir des estimations quantitatives d'une propriété du sol, les pédologues ont utilisés des fonctions de transfert mathématiques pour corréler les longueurs d'onde spectrales aux propriétés du sol. La fonction de transfert est calibrée en utilisant les longueurs d'onde spectrales comme variables indépendantes et les valeurs mesurées en laboratoire des propriétés du sol comme variable dépendante. Une fois calibrée sur les spectres, la propriété du sol peut être prédite en utilisant uniquement les informations spectrales. C'est dans ce cadre que j'ai écrit **un « livre de cuisine » pour aborder la spectroscopie du sol et résumer les récents développements de ces dernières années** (Wadoux et al., 2021b). Ce livre aborde toutes les étapes quantitatives dans la spectroscopie du sol : choix des techniques de pré-traitement des spectres, échantillonnage pour décider quelles échantillons envoyer au laboratoire, choix et étalonnage de modèles pour relier les spectres avec des mesures de laboratoire, évaluation statistique et contrôle de la qualité des modèles, développement d'une bibliothèque spectrale numérique, et, finalement, transfert des modèles statistiques sur des spectres obtenus par d'autres instruments. Toutes les étapes sont agrémentées de codes dans la langue de programmation R, ce qui en fait un livre presque entièrement dédié à l'aspect pratique de l'analyse de données spectrales infrarouge sur les sols.

Dans la continuations du livre j'ai repris le développement d'un **progiciel R « R package » sur la modé-**

lisation de données spectrales complexes (Ramirez-Lopez et al., 2022). La vulgarisation des techniques d'analyse spectrales sur les sols a poussé le développement d'un grand nombre de bases de données spectrales infrarouge sur les sols. Les échantillons de sols viennent d'aires d'études différentes, pour des aires géographiques à l'échelle du champ, de la région mais aussi de l'état et du continent. La taille et la diversité de telles librairies spectrales augmente de fait leur complexité, ce qui demande une adaptation des techniques d'analyses statistiques et algorithmiques pour en extraire des informations. C'est dans ce cadre que le progiciel R resemble fournit plusieurs outils pour extraire efficacement et avec précision des informations quantitatives à partir de bases de données spectrales vastes et complexes. Les fonctionnalités de base du progiciel incluent i) la réduction de la dimensionnalité, ii) le calcul des mesures de dissimilarité, iii) l'évaluation des matrices de dissimilarité, iv) la recherche de voisins spectraux, v) l'ajustement et la prédiction de modèles spectroscopiques locaux. Un exemple de fonctionnalité du progiciel est illustré dans la Figure 2.21 avec une modélisation locale. Dans ce type de modélisation l'espace multidimensionnel des composants principaux n'est pas entièrement utilisé pour étalonner un modèle, mais seulement une petite partie de cette espace qui est la plus similaire au point qui doit être modélisé. À cet égard, le progiciel inclut des fonctions pour réduire la dimension et projeter les données dans l'espace des composants principaux, puis d'analyse des distances entre points et enfin de la modélisation locale utilisant plusieurs types de modèles.

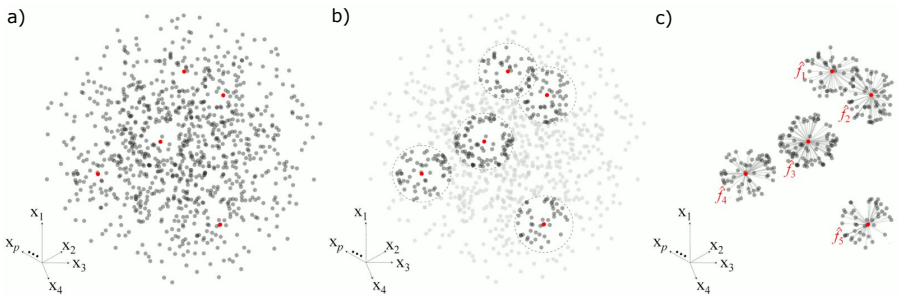

***Figure 2.21** – Les trois étapes d'une modélisation locale dans l'espace des composant principaux : a) les points en rouge sont les points pour laquelle la propriété cible doit être estimée en utilisant les données spectrales, b) les points les plus proche dans l'espace sont sélectionnés, et c) un modèle local est étalonné pour chaque point en utilisant les points les plus similaires. D'après le progiciel resemble (Ramirez-Lopez et al., 2022).*

2.6.2 Les sciences participatives

Les approches participatives ont attiré beaucoup d'attention récemment, non seulement pour permettre des développements scientifiques et techniques, mais aussi pour obtenir des résultats sociaux et éducatifs. Il est en effet reconnu que de nombreuses questions de durabilité environnementale ont un haut niveau de complexité et ne peuvent être traitées isolément les unes des autres. Ces problèmes environnementaux appelés problème « épineux » (en anglais, « wicked », voir Rittel et Webber, 1973) n'ont pas de meilleure solution (Bouma et McBratney, 2013), mais plutôt une série de résultats possibles équilibrant les besoins et les intérêts des différentes parties prenantes. **En tant qu'efforts de collaboration entre experts, scientifiques et autres participants, les approches participatives qui équilibrent les intérêts, les attentes et les connaissances sont de plus en plus utilisées pour résoudre les problèmes à l'interface entre la science et la société.** Outre l'avantage pour les projets scientifiques à grande échelle (par exemple en termes de collecte de données), la participation est censée construire une vision partagée entre les participants, réduire les conflits et, par conséquent, augmenter les chances de succès d'un projet. Ci-après participatif est basé sur une définition de Von Korff et al. (2012) et fait référence à l'implication non seulement de professionnels formés (par exemple, des scientifiques, des participants quasi-spécialistes)

mais également de toutes les autres parties intéressées, y compris, par exemple, des profanes, des amateurs, une communauté locale de citoyen intéressée, des étudiants ou encore des écoliers.

La participation de non-experts au développement scientifique n'est pas nouvelle. Il existe de nombreux exemples dans l'histoire des sciences naturelles d'interactions entre non-expert et scientifiques, par exemple celui des ornithologues amateurs du XIXe siècle participant à l'observation des oiseaux en France ou les sociétés et clubs amateurs décrivant et classant des objets naturels dans le Yorkshire de la fin de l'époque victorienne. Un regain d'intérêt actuel pour la recherche participative a émergé d'une variété de sous-disciplines à l'interface entre la durabilité et la société, allant de la gestion de l'eau à l'adaptation au changement climatique ou à la prise de décision socio-environnementale. De nombreuses études sont disponibles documentant l'utilisation croissante des approches participatives dans le processus scientifique, montrant en même temps une compréhension de l'**utilité de ces approches pour les politiques publiques, notamment comme outil de communication scientifique**.

Il existe dans les faits plusieurs définitions, parfois utilisées à tort, de la participation. Souvent, la participation non experte à la science est qualifiée de science citoyenne. Le terme de science citoyenne nous vient de l'étude d'Irwin (1995) qui a défini au sens large le terme comme une science pratiquée « par » et « pour » le peuple (Strasser et al., 2019), en d'autres termes, la science devrait se concentrer sur les préoccupations des citoyens et le processus scientifique devrait inclure leurs connaissances et expériences contextuelles locales qui ne sont pas disponibles dans les institutions académiques formelles. Une autre signification de la science citoyenne vient de Bonney et al. (2009) et est peut-être la plus populaire aujourd'hui. Dans Bonney et al. (2009), les sciences citoyennes sont définis comme des non-scientifiques contribuant à la collection de données scientifiques et un outil d'éducation du public à la science. Les deux concepts pointent en quelque sorte dans une direction opposée et sont parfois appelés science contributive ou science citoyenne démocratisée. Une variété d'autres concepts sont couramment utilisés, ils se chevauchent en partie mais diffèrent dans l'aspect de la participation qu'ils représentent. Notamment, les typologies de participation existantes décrivent plutôt la participation comme un spectre composé de la qualité de la participation, du stade et du degré d'implication, ou encore des pratiques épistémologiques.

C'est sur cette constatation que dans Wadoux et McBratney (2023) je me concentre sur la participation dans la recherche et la gestion des sols. La recherche documentaire est basée sur une procédure en deux étapes combinant une recherche documentaire systématique et une recherche de littérature grise dans un moteur de recherche standard. Je classe la littérature en fonction de la qualité de la participation non experte aux projets scientifiques. La classification utilisée pour discriminer le corpus en trois classes représente trois phases de projet auxquelles les participants sont généralement associés. Je distingue les ***approches contributives*** qui sont « conçues par des scientifiques et pour lesquelles les membres du public apportent principalement des données », des ***approches collaboratives*** où les participants « apportent des données » mais aussi « aident à affiner concevoir des projets, analyser des données et/ou diffuser des résultats » et finalement les ***approches co-créées*** où les scientifiques et le public « travaillent ensemble » et où au moins certains des participants sont « impliqués dans la plupart des ou toutes les étapes de la démarche scientifique », c'est-à-dire y compris pour définir la question d'étude et pour discuter des résultats et poser de nouvelles questions (Bonney et al., 2009, p. 17). Ces trois grandes phases sont représentées dans la Figure 2.22.

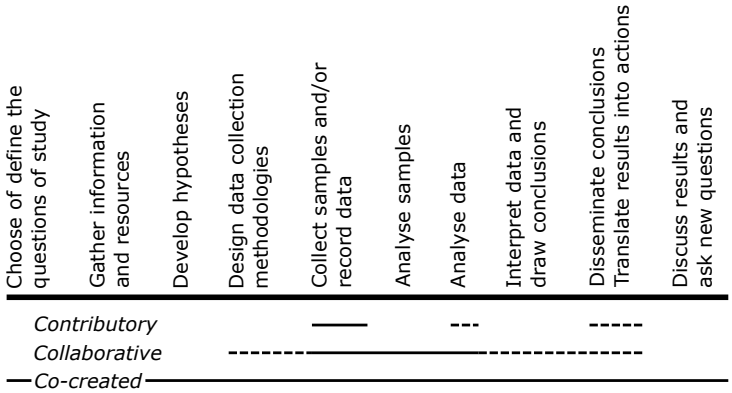

Figure 2.22 *– Les différentes phases de la recherche auxquelles les membres du public pourraient participer. La ligne pleine indique que le public participe et la ligne noire pointillée indique que le public pourrait participer à cette phase, pour chacune des trois grandes catégories. Adapté de Bonney et al. (2009).*

Pour chaque phase de participation, je décris ensuite la littérature en fonction de cinq éléments caractérisant la qualité de la participation : entrées, activités, sorties, résultats et impacts, à partir du cadre conceptuel proposé dans Shirk et al. (2012). Ce cadre est un modèle logique axé sur les résultats qui décrit le résultat de la participation comme un équilibre entre l'intérêt public et scientifique. Les cinq éléments du cadre (c'est-à-dire *entrées, activités, sorties, résultats* et *impacts*) sont décrits dans le Tableau 2.1. Outre la qualité de la participation, le cadre permet de distinguer les résultats des approches participatives pour les participants (par exemple, les compétences, les connaissances acquises), la science (par exemple, les publications et les nouvelles découvertes) et les systèmes socio-écologiques (par exemple, la législation, l'amélioration de la prise de décision, conservation des agroécosystèmes).

Table 2.1 *– Les cinq éléments caractérisant la qualité de la participation et leur description, d'après Shirk et al. (2012) et adapté dans Wadoux et McBratney (2023).*

Élément	Décrit...
Entrées	...quels sont les intérêts et les attentes des participants et des scientifiques professionnels (par exemple, connaissances, conservation des sols, éducation).
Activités	...quel est le gros du travail à effectuer dans le projet, les tâches. Cela peut inclure la conception des stratégies d'échantillonnage et la gestion de la conception et de la mise en œuvre du projet, ainsi que la communication entre les participants et la formation..
Sorties	...quels sont les résultats des activités (par exemple, nouvelles données collectées). Ceci est généralement facile à quantifier.
Résultats	...quels sont les résultats qui résultent des sorties. Cela peut inclure des ensembles de compétences et une sensibilisation accrue des participants. Les résultats identifiés pour la science sont une meilleure compréhension scientifique ou des techniques innovantes. Pour les systèmes socio-écologiques, cela comprend l'amélioration des relations entre les agences environnementales et les utilisateurs des terres, ou une meilleure politique de gestion des ressources en sols.
Impacts	...quels sont les impacts à long terme souhaités et mesurés qui favorisent une meilleure gestion des sols, le bien-être humain ou le développement des connaissances scientifiques.

Les résultats de cette revue me font constater que **presque tous les projets ont été lancés à partir d'intérêts scientifiques plutôt que d'intérêts publics**. Dans la plupart des projets contributifs, par exemple, l'objectif est de collecter des données pouvant produire un résultat scientifique. L'intérêt public est assuré pour maintenir l'exactitude des données et une bonne couverture spatiale. J'ai trouvé des situations similaires dans les études de sol collaboratives et co-créées qui ont été examinées. Une explication

pourrait être que ma revue de la littérature est principalement basée sur la littérature scientifique, et j'ai peut-être ignoré les petits projets initiés dans l'intérêt du public. Bien que nous ayons constaté dans cette revue que les projets participatifs étaient dirigés et initiés par des scientifiques, le lancement du projet était, à l'inverse, motivé par l'idée que le public et les non-experts pouvaient améliorer la mise en œuvre. Dans certains cas, la participation a été initiée, notamment pour assurer la prise en compte des savoirs locaux et pour renforcer la pertinence des solutions locales.

L'approche de classification adoptée dans cette revue a révélé que **les approches contributives (« science citoyenne » ou « crowdsourcing de données scientifiques ») étaient principalement appliquées et documentées dans le contexte des pays développés, alors que les projets qui suggèrent une plus grande implication des participants (c'est-à-dire collaboratifs et approches co-créées) ont été principalement formulés et appliqués dans les pays en développement**. L'examen a dans les fait mis en évidence deux corps de littérature distincts. Le premier concerne un grand nombre de projets contributifs initiés par des scientifiques et menés dans le but de contribuer à la science (c'est-à-dire en collectant des données pour des programmes de surveillance à grande échelle). Ces projets ont généralement un faible potentiel d'amélioration des connaissances scientifiques des participants, bien que certaines activités puissent cibler l'éducation. Il existe un deuxième corpus de littérature où la participation est appliquée dans les pays en développement dans le contexte de l'agriculture, de la durabilité et du développement en général. De tels projets sont plus alignés sur le changement socio-écologique que sur la contribution à la science. Outre les différents résultats pour les types de projets participatifs, nous supposons que les coûts pour les chercheurs et les individus et les communautés locales expliquent pourquoi la participation est appliquée différemment dans les pays développés et en développement. Les projets contributifs sont coûteux à mettre en place et à maintenir pour les scientifiques. Ils sont également coûteux pour les particuliers car cela nécessite du temps et de l'engagement pour la collecte de données. Les projets collaboratifs et co-créés sont également coûteux à mettre en place, mais ce coût est supporté par les scientifiques alors que les coûts de mise en œuvre et de maintenance supportés par les communautés locales sont faibles.

Les résultats rapportés de ces projets correspondent généralement aux attentes des participants et à leur degré d'implication. **Les projets contributifs aboutissent à un meilleur processus scientifique grâce à la collecte de données, tandis que les projets co-créés sont adaptés au changement social et à la promotion de la durabilité des agro-écosystèmes**. Nous avons cependant constaté un manque d'informations sur l'impact à long terme, à la fois sur les résultats d'apprentissage des participants ou sur les effets sur les systèmes socio-écologiques. Les impacts à long terme sont difficiles à mesurer car ils peuvent se produire plus d'une décennie après le lancement du projet. Le fait de ne pas rapporter les résultats d'apprentissage à long terme des participants n'est pas non plus spécifique aux études de recherche et de gestion des sols, car cette question a également été abondamment rapportée dans la littérature en écologie (e.g. Shirk et al., 2012). Récemment, plusieurs travaux ont contribué à une meilleure compréhension de l'impact durable à long terme des projets participatifs et ont proposé un cadre d'évaluation de l'impact. Souvent, ceux-ci sont souvent basés sur une conception inversée qui détermine d'abord les résultats souhaités, puis le type de participation qui peut atteindre ces résultats. L'évaluation des impacts apporterait certainement une contribution précieuse à la compréhension du rôle du projet participatif sur les sols dans l'amélioration du bien-être humain et la durabilité des sols.

Je conclus avec cette revue que **la surveillance des sols bénéficie de la participation**. Des réseaux de non-experts offrent la possibilité d'améliorer notre connaissance de la couverture du sol en fournissant les observations dont nous avons besoin avec une densité raisonnable. Cela a également été discuté dans Rossiter et al. (2015) : les non-spécialistes agissent en tant qu'observateurs ou expérimentateurs. Cela profite d'abord aux scientifiques et aux cartographes numériques des sols qui peuvent utiliser ces informations pour produire ou améliorer les cartes des propriétés des sols. Les pédologues ont également la possibilité

d'améliorer leur connaissance des sols en tenant compte des connaissances locales. Les utilisateurs du sol, tels que les agriculteurs et les communautés rurales, ont des connaissances tacites ou une expérience accumulée grâce à la pratique, car ils peuvent reconnaître des caractéristiques du sol qui ne correspondent pas aux légendes des cartes. Souvent ces connaissances ou observations sont définies dans une terminologie locale dont le vocabulaire n'est pas facilement accessible aux pédologues. Le défi pour accéder à ces informations est de définir un langage commun pour transférer les connaissances de l'utilisateur du sol au pédologue. On parle souvent d'ethnopédologie (Barrera-Bassols et Zinck, 2003). Enfin, les projets co-créés ont l'avantage de **transmettre aux pédologues des informations sur des phénomènes d'intérêt qui préoccupent les communautés locales**, mais peuvent ne pas être bien corrélés avec des processus clairement définis et compris par les pédologues. Comprendre le désaccord entre la connaissance du sol et les phénomènes locaux déclenchera indéniablement de nouvelles hypothèses qui pourront ensuite être testées. La participation dans ce sens est un outil utile en tant que fournisseur d'informations sur les phénomènes pédologiques locaux.

2.6.3 L'exploration de données en science des sols

La dernière décennie a vu une augmentation considérable de l'information numérique électronique et des technologies de l'information disponibles pour la recherche universitaire. Cette augmentation interroge la façon dont les sciences sont abordées d'un point de vue méthodologique. Des critiques sont formulées à l'égard de la recherche fondée sur des données et les outils informatiques, relançant ainsi les débats sur la méthode scientifique et les pratiques scientifiques dans de nombreux domaines. La science des sols ne fait pas exception. Malgré le consensus selon lequel la science des sols repose fondamentalement sur des experts et une connaissance approfondie du domaine, **le développement de techniques de détection, de méthodes analytiques d'analyse des sols et la facilité de stockage et de traitement de ces données modifient la pratique de la science des sols**. Une grande attention a donc été récemment accordée à la recherche à forte intensité de données ou axée sur les données dans la littérature scientifique ou populaire sur les sciences du sol. La recherche à forte intensité de données adopte une approche où le progrès est imposé par les données, par opposition aux approches « axées sur les connaissances » ou « centrées sur l'expert » dans lesquelles une hypothèse est développée ou corroborée par des données. La recherche à forte intensité de données émerge grâce à la combinaison de plusieurs facteurs opportuns, qui sont 1. la facilité de génération, de traitement et de stockage des données, 2. le développement de l'informatique, de la puissance de calcul et des ressources logicielles, et 3. la vulgarisation d'outils statistiques et algorithmiques complexes qui font de plus en plus appel à des calculs d'apprentissage automatique, pour explorer ces référentiels de données.

L'utilisation de statistiques pour explorer des bases de données n'est pas nouvelle en science des sols. Dernièrement cependant, les pédologues ont assisté à une augmentation de l'utilisation de modèles flexibles basés sur les données et des stratégies algorithmiques, en particulier d'apprentissage automatique, pour s'attaquer à cet environnement riche en données. Les réseaux de neurones ou les forêts d'arbres décisionnels sont des exemples de ces modèles. La recherche axée sur les données a suscité beaucoup d'enthousiasme en science des sols, en particulier dans le sous-domaine de la pédométrie, où la recherche s'est toujours appuyée davantage sur des données de terrain et d'observation que sur des expériences manipulées. **L'abondance des données et leur utilisation comme principal moteur de connaissances dans certaines sous-disciplines de la science des sols a plusieurs implications méthodologiques et épistémologiques qui n'ont été que peu documentées.**

Dans Wadoux et al. (2021c) je soulève plusieurs questions ; La recherche sur les sols basée sur les données est-elle nouvelle ? Tous les types de données sont-ils valables pour une utilisation dans une science axée

sur les données ? Quels sont les risques, les défis et les revendications extrêmes de la recherche sur les sols basée sur les données ? Si la connaissance se trouve dans les données, devrions-nous investir tous nos efforts pour générer plus de données ? Je tente de fournir **certains contextes et une perspective introductive aux défis conceptuels liés aux stratégies de recherche axées sur les données plutôt que sur des hypothèses**. Je traite des questions de recherche fondées sur la connaissance versus une recherche axée sur les données à travers ma suggestion que la recherche basée sur les données n'est pas nouvelle en science des sols. Je continue avec un questionnement sur l'objectif d'une recherche basée sur les données ainsi que sur le type de données nécessaire pour y faire face. Je discute ensuite trois propositions émanant du nominalisme dans un contexte de science des sols : 1. l'analyse des données est exempte de toute théorie, 2. les données parlent d'elles-mêmes, exemptes de préjugés humains, et 3. aucune connaissance du domaine n'est nécessaire. En reconnaissant que la science des sols axée sur les données s'inscrit avec la vision nominaliste de longue date de la science des sols et que ses pièges sont enracinés dans l'empirisme, la question qui suit logiquement est de savoir comment les scientifiques du sol obtiennent-ils une explication scientifique à partir des données ? Pour répondre cette question, je discute l'utilisation de techniques de modélisation (c'est-à-dire d'outils et de modèles statistiques et mathématiques, d'exploration de données) pour détecter des formes dans des bases de données multivariées et complexes, ou pour fournir une analyse valide et un représentation généralisable de la réalité, comme base pour faire des prédictions.

J'ai soutenu que même si cela peut sembler révolutionnaire pour certains, **la science des sols a une longue histoire d'efforts intensifs en données et exploratoires pour générer des connaissances à partir de données sur le sol**. C'est précisément parce que la science des sols, comme les autres sciences naturelles, a commencé par un inventaire des différentes propriétés représentatives de la diversité et de la complexité du sol. Travailler dans un environnement riche en données n'est donc pas nouveau pour les pédologues, mais peut-être encore moins dans les sous-disciplines des sciences du sol (par exemple la pédométrie) avec une longue histoire de données collectées à partir d'une observation de l'environnement non contrôlée plutôt que des expériences contrôlées et manipulées. Tout comme les pédologues stockaient et classaient les données sur les sols sur des morceaux de papier et dans des archives physiques dans le passé, les outils disponibles pour le stockage des données sur les sols d'aujourd'hui sont les ordinateurs et les bases de données électroniques. Ces bases de données électroniques sont analysées simultanément et à distance par plusieurs utilisateurs, mais pour autant la logique de stockage des données à un seul endroit n'a pas été bouleversée. Les bases de données sont encore centralisées géographiquement dans de grands instituts de recherche alliant puissance informatique, stockage et grands projets apportant un flux constant de nouvelles données. Peut-être que le paysage social montre un signe de changement. Alors que les citoyens ont beaucoup contribué à la collecte de données sur les sols au cours des dernières années, le développement de bases de données sur les sols accessibles au public et l'accessibilité des logiciels fournis par les ordinateurs personnels encouragent de plus en plus les citoyens à analyser les données et à participer à la production de connaissances.

Ce qui est généralement considéré comme révolutionnaire dans la science contemporaine axée sur les données est donc dans une large mesure un changement d'ampleur dans la quantité de données collectées et la capacité de les analyser avec une puissance de calcul. Les données sur les sols sont actuellement générées rapidement, en grande quantité et à partir de sources multiples, ce qui soulève des inquiétudes quant à leur capacité à être combinées efficacement. Les méthodes d'analyse de ces données proviennent en grande partie de l'utilisation de la puissance des ordinateurs et de solutions statistiques et algorithmiques complexes. Dans cette « nouvelle » science axée sur les données, les experts en informatique jouent un rôle dans la production de connaissances sur les sols par l'analyse des données. Ce n'est pas un hasard si corollaire à cette augmentation, des discours dans d'autres domaines, comme en science des données, se font en science des sols. Ces prétentions, qui ne tiennent pas longtemps en sciences du sol, s'enracinent

dans une forme radicale d'empirisme exprimant que l'analyse des données du sol peut s'affranchir de toute théorie, hypothèse ou connaissance pédologique.

En substance, je conclus donc qu'**il ne semble pas y avoir eu de changement majeur récent dans la manière dont les pédologues obtiennent une explication scientifique à partir des données**. Trouver des corrélations dans les données de sol est au mieux un point de départ, utile dans une phase exploratoire pour générer des hypothèses et être utilisé comme heuristique pour développer des modèles mécanistes plus réalistes basés sur la causalité lorsque les connaissances augmentent. Dans la quête d'explications sur les processus du sol, les pédologues sélectionnent ainsi des modèles simples au détriment de la précision. Peut-être qu'en posant le problème différemment, peut-on être tenté d'utiliser des modèles complexes, souvent plus précis que des modèles simples. Les modèles complexes sont plus précis et peuvent donc fournir une meilleure représentation du système naturel étudié. Il s'agit d'une caractéristique souhaitable lorsque la technique d'analyse des données vise à fournir une explication statistique inductive d'un processus de sol.

En analysant les enjeux épistémologiques de la recherche scientifique axée sur les données à la lumière de la littérature historique, nous avons constaté qu'il existe **une continuité des pratiques**, certaines étant certes amplifiées par les évolutions technologiques récentes, mais que **le cœur des méthodes d'enquête scientifique à partir des données, c'est-à-dire les méthodes scientifiques de production de connaissances, restent largement inchangé**.

2.6.4 La convergence numérique comme outil de progrès

Dans une certaine mesure, toutes nos recherches et tous nos enseignements ont été rendus possible, améliorés et élargis par la **convergence numérique**. Les capteurs fixes et mobiles et les systèmes de communication sans fil assistés par Internet produisent des flux bon marché et abondants de données numériques sur le sol qui peuvent être facilement utilisées pour la modélisation et la génération d'informations. Les pédologues ont largement adopté l'environnement numérique, composé de uns et de zéros binaires au lieu de données analogiques nécessitant une interprétation humaine. Il est possible de créer des cartes numériques à l'échelle globale à l'aide de grandes bases de données numériques électroniques sur les sols (c.a.d avec plus > 10 Go de données) et de centaines de milliers de profils de sols. Ces données de sol sont acquises rapidement par des capteurs et des instruments numériques. L'analyse des données est rendue possible par les ordinateurs, l'imagerie numérique et le *cloud computing*. L'innovation majeure de ces dernières années est venue de la conversion analogique-numérique. L'imagerie, les cartes, la lecture des spectromètres et des capteurs et les dessins ont dû être transformés en une représentation numérique qui, avec les ressources émergentes de l'informatique, permettent la comparaison, l'analyse, le traitement et le partage des données. Outre la conversion analogique-numérique, les capteurs et instruments générant des données numériques ont progressivement remplacé les instruments analogiques qui nécessitent une lecture humaine. Ainsi, l'une des principales caractéristiques de cette conversion analogique-numérique est la possibilité de produire plus de données à moindre coût, facilement utilisables par les ordinateurs et les technologies de l'information. La convergence numérique s'est poursuivie avec Internet, et plus récemment avec le développement de nouveaux outils pour gérer et traiter de grandes quantités de données électroniques. La science numérique du sol est, par glissement sémantique, l'étude du sol aidée par les outils de la convergence numérique.

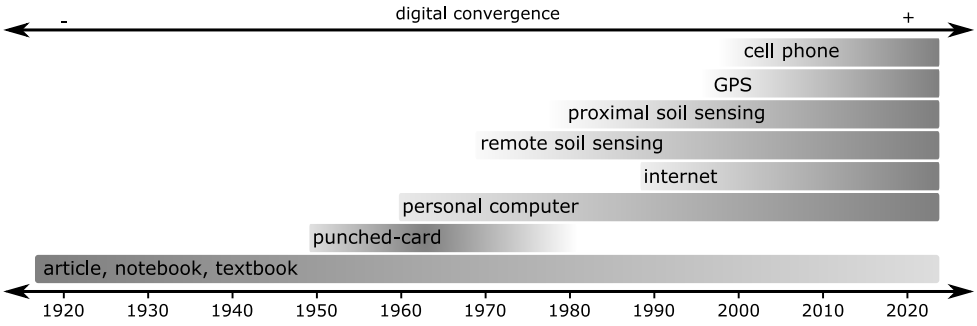

Figure 2.23 *– Chronologie simplifiée des évènements majeurs de la convergence numérique en science des sols. D'après Wadoux et McBratney (2021a).*

Les événements majeurs de la convergence numérique en science des sols, à savoir la convergence des données numériques, les bases de données électroniques stockées dans les ordinateurs, Internet et les nouveaux outils pour gérer et traiter de grandes quantités de données électroniques, ont suivi les développements de la révolution numérique dans d'autres sciences et dans la société. Nous décrivons une brève chronologie des principaux événements dans la Figure 2.23. Les données numériques existaient bien avant l'ère informatique sous une forme numérique pré-électronique stockée dans des armoires ou des instituts de recherche sous forme de nombres et de texte dans des articles, des cahiers ou des manuels. Le système de cartes perforées a été adopté en science des sols à partir des années 1950, mais c'est l'ordinateur personnel et la conversion de bases de données numériques pré-électroniques sur les sols en bases de données numériques qui ont amené un nouvel outil pour le pédologue dans les années 1960. Avec l'ordinateur, l'expansion spectaculaire de la science numérique du sol a commencé, aidée par Internet dans les années 1990 et l'utilisation généralisée de capteurs et d'instruments numériques à distance, aéroportés (années 1980) ou proximaux (années 1990). L'un des premiers instruments proposant un signal numérique fut les capteurs géophysiques EM31 dans les années 1970. Les systèmes de positionnement global (GPS) sont devenus largement accessibles aux civils dans les années 2000. Enfin, les téléphones portables et l'internet haut débit sont devenus accessibles à un large public à partir des années 2000.

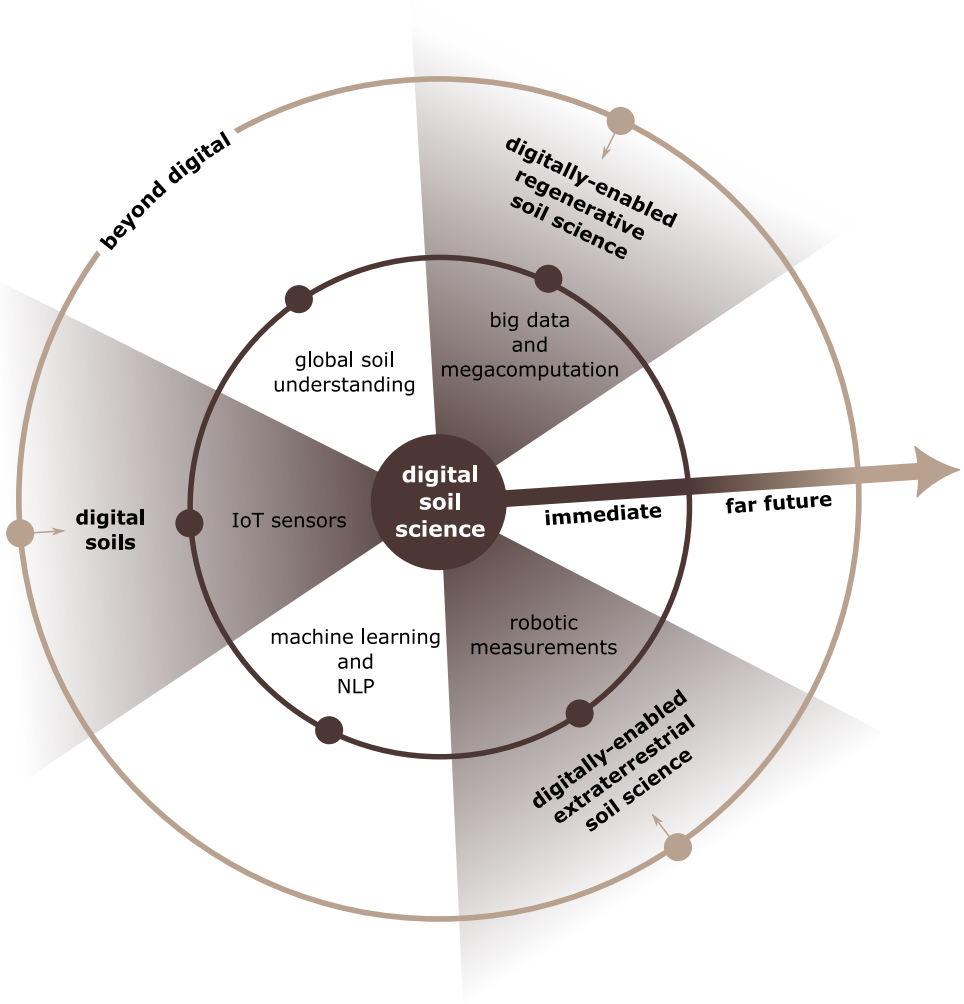

Figure 2.24 *– Schéma récapitulatif des futurs aspects de la science des sols permis par le numérique. L'avenir immédiat de la science numérique du sol soutient le développement des technologies et des techniques numériques appliquées à l'étude du sol, ainsi qu'une compréhension globale croissante des sols. Ces nouvelles techniques et technologies peuvent à leur tour permettre d'aller au-delà du numérique dans un avenir lointain, de soutenir la création de sols numériques (c'est-à-dire un jumeau numérique du sol) et d'aider la science des sols régénérative et extraterrestre. D'après Wadoux et McBratney (2021a).*

Dans Wadoux et McBratney (2021a) j'explique comment **la compréhension de tous les aspects du « sol » a été renforcée et intensifiée par les données, les technologies et les approches numériques**. Je décris comment la science des sols a changé à l'aide d'illustrations des développements intellectuels et techniques permis par la convergence numérique. Le numérique a facilité a grand nombre de développement rapide : télédétection du sol, détection proximale du sol, cartographie numérique, la caractérisation des microbes avec la bioinformatique que l'analyse des données numériques produites par les techniques de séquençage a rendu possible, mais aussi les applications pour téléphones portables. Il y a eu des échecs notables tels que le manque relatif de fédération et de partage de données, de description véritablement numérique des sols sur le terrain et de développement de systèmes taxonomiques numériques à l'échelle mondiale. De même, la convergence numérique présente des pièges, notamment le manque de nouvelles théories sur les sols, le manque de connaissances de base en sciences du sol par les chercheurs et le fait d'essayer d'en faire trop avec trop peu d'informations ou de données. Je tente finalement de définir le future proche et plus lointain de la science des sols aidée par le numérique (Figure 2.24) : la science numérique du sol sera dominée par l'apprentissage automatique, l'IoT, la mesure robotique et le *Big Data*,

tous conduisant à une meilleure compréhension globale du sol. Dans un avenir lointain, la science numérique du sol pourrait permettre de régénérer le sol ici sur terre, de créer des sols multifonctionnels sur d'autres planètes et potentiellement de créer des sols intelligents, auto-organisés et orientés vers des objectifs spécifiques.

2.7 Synthèse et conclusion

Des travaux scientifiques ; ils se sont concentrés autour du développement d'un ensemble méthodologique pour la caractérisation spatiale d'indicateurs biophysiques des sols. Les développements méthodologiques m'ont permis d'élargir mes cas d'études à la spatialisation de certaines fonctions fournies par les écosystèmes notamment en hydrologie ou pour la cartographie de la biomasse aérienne. Les méthodes utilisées sont statistiques, empiriques et *data-driven*, avec une attention particulière sur des nouvelles méthodes d'intelligence artificielle et d'apprentissage profond. Mon positionnement récent se concentre sur un double objectif : l'évaluation statistique des modèles spatiaux et l'amélioration de ces modèles, avec un positionnement plus cognitif de recherche sur les sciences numériques et techniques pédométriques appliquées aux sciences du sol. Ce positionnement me permet une **fertilisation entre épistémologie des approches basées sur les données et le numérique et développements méthodologiques pour la caractérisation spatiale d'indicateurs biophysiques.**

Un parcours de recherche ; il est riche de séjours dans des institutions diverses. Bien que mes trois grandes périodes de recherche aient été effectuées dans des groupes liés directement à la science des sols, et que les sols soient mes objets d'études privilégiés, j'ai effectué plusieurs périodes dans des groupes de recherches en statistique appliqué, hydrologie ou ingénierie civile. Ces séjours m'ont fait développer une vision plus large de la spatialisation. À coté des institutions de recherche, plusieurs méthodes ont été développées à travers des collaborations avec des acteurs du Nord comme du Sud. Certaines actions ont aussi visé le transfert des innovations vers les usagers des sols (agriculteurs), le secteur privé (bureaux d'études), ou plus récemment vers des services et institutions publiques (agence de l'eau, TERN - *Australia's Land Ecosystem Observatory* en Australie).

Des encadrements d'étudiants ; l'encadrement d'étudiants en stage de Master et de doctorant donne l'opportunité d'explorer des nouvelles méthodes et applications de techniques. Plusieurs des travaux présentés dans ce mémoire ont été soit co-réalisés soit approfondis par des étudiants que j'ai encadré. Certains ont été publiés, d'autres sont restés au stade de mémoire et n'ont pas été révélés à une audience plus large.

Une démarche ; elle se caractérise par une volonté conjointe de faire de la recherche mais aussi de tenter de comprendre notre manière de pratiquer la science. Sur certains sujets, notamment sur l'application des techniques basées sur les données, de nombreuses questions se posent et nos pratiques habituelles venant souvent des *data scientists* me paraissent inadaptées à l'étude des sols. Il s'agit de comprendre et d'étudier ce qui apparaît à première vue comme un renouveau méthodologique mais dont les prémices sont présents depuis plusieurs décennies.

Une base pour un projet de recherche ; l'évaluation des indicateurs biophysiques et des fonctions dans l'ensemble de mes recherches est faite presque exclusivement de manière monofonctionnelle avec toutefois l'utilisation de plusieurs indicateurs en entrées dans des modèles empiriques ou à base physique. Mon projet de recherche ci-après vise à proposer des méthodes pour spatialiser les fonctions du sols, et quantifier la multifonctionnalité de ces sols. Cette quantification a recourt aux techniques de spatialisation décrite dans mes travaux passés, pour lesquels de développements supplémentaires sont toutefois nécessaires.

CHAPITRE 3

Perspectives de recherche

3.1 Introduction et cadre conceptuel

Les écosystèmes remplissent une multitude de fonctions conditionnant leur contribution à des services essentiels au bien-être et à l'activité économique des hommes. Le concept de « service écosystémique », développé dès les années 70 (Westman, 1977; Randall, 1988), met en exergue les avantages que l'homme tire des écosystèmes et fournit un cadre conceptuel pour leur préservation. Les sols assurent des fonctions essentielles au sein de ces écosystèmes. L'évaluation et la valorisation des services rendus par les fonctions des sols, par exemple la production de biomasse ou la régulation du climat, constituent le fondement de nombreux programmes de recherche de ces dernières années (Walter et al., 2015).

Néanmoins, les modes d'évaluations et de quantification des services écosystémiques font encore défaut. **Il n'existe pas de méthode consensuelle d'évaluation quantitative des services écosystémiques fournis par les sols**. Les modèles se présentant en « cascade » (Potschin et Haines-Young, 2011; Tibi et Therond, 2017) permettent de relier de manière intégrée des jeux d'indicateurs (p.ex. la quantité de matière organique d'un sol), qui correspondent à des propriétés biologiques, chimiques et physiques du sol, à un ensemble de fonctions ou services. L'évaluation d'un service met généralement en jeu plusieurs fonctions, elles-mêmes définies par une série d'indicateurs, suivant le triptyque indicateur-fonction-service (Bünemann et al., 2018). De ce fait, les différentes fonctions du sol sont souvent interdépendantes puisqu'elles s'appuient sur des propriétés du sol semblables. Un changement dans les pratiques agricoles pour améliorer une fonction (p.ex. production) peut simultanément affecter positivement ou négativement d'autres fonctions (p.ex. l'habitat des micro-organismes). À cause de ces synergies et antagonismes entre fonctions, un sol ne peut jamais mobiliser à son plein potentiel chacune des fonctions qu'il peut supporter (Zwetsloot et al., 2021; Vrebos et al., 2021). De plus, ces synergies et antagonismes varient avec les dynamiques du paysage, à des échelles spatiales et temporelles emboîtées. **La quantification et la modélisation spatiale des fonctions, prenant en compte leurs antagonismes et synergies au sein du paysage, devient primordiale afin, à terme, de pouvoir mobiliser les services écosystémiques qui en découlent dans des processus de décision**. Il s'agit aussi d'une condition préalable pour quantifier l'effet sur les sols de perturbations externes telles que l'intensification des pratiques agricoles ou le changement climatique global.

Il s'agit aussi de prendre en compte qu'une partie du bien-être et à l'activité économique des populations passe par l'éventail des services écologiques et écosystémiques gratuits fournis par les sols (Robinson et al., 2012). Dans les pays ayant des problématiques de développement, notamment, il s'agit de concilier un besoin d'augmentation de la production agricole pour soutenir la croissance de la population avec une nécessaire conservation des agrosystèmes et de l'environnement (Lal, 2000). La dégradation des fonctions du sol entraîne une cascade de nuisances sur les populations telles que des pertes de biodiversité et des

détériorations de la qualité de l'eau. De nombreux pays du Nord, *a contrario*, font face à une transition des agroécosystèmes pour un changement vers des principes d'agriculture durables, tel que l'agroécologie en France. Dans ces deux problématiques se superpose le changement climatique global qui, sur certaines zones géographiques, exacerbe les problèmes de production agricole et modifie les interactions biologiques, physicochimiques et hydriques qui prennent place au sein des sols. **Une meilleure quantification et modélisation de la distribution spatiale des fonctions du sol et de leur interaction doit permettre de s'appuyer sur les processus écologiques pour produire plus et durablement**.

Par ailleurs, la relative rareté de certaines données pédologiques en comparaison avec les données disponibles dans d'autres disciplines rend inopérante certaines solutions méthodologiques connues. Il est plus que jamais nécessaire de produire des données sol alternatives à l'analyse de sol géolocalisée, et de développer de modèles capables de valoriser ces données en utilisant des moyens numériques. Dans ce contexte, **ce projet de recherche entend se focaliser sur un verrou scientifique conséquent mais peu étudié en lui-même : la modélisation et la spatialisation des fonctionnalités des sols – tant, dans les études sur les services écosystémiques fournis par les sols, les questions de quantification et de modélisation sont souvent évoquées mais rarement traitées**. Pour traiter les questions de quantification et de modélisation spatiale des fonctions du sols, plusieurs enjeux scientifiques et méthodologiques doivent être questionnés :

1. Comment prendre en compte la corrélation entre les indicateurs (c.à.d. passer d'une modélisation mono-variée à multivariée) ?
2. Peut-on utiliser les données hétérogènes de source et d'incertitude différentes pour la modélisation des fonctions ?
3. Comment spatialiser les modèles mécanistes d'estimation des fonctions ? Il s'agit de prendre en compte la variation spatiale d'un grand nombre de paramètres ainsi que leur sensibilité et incertitude, tout en évitant la complexité beaucoup plus grande que cela induit.
4. Comment agréger des indicateurs avec différentes unités et gammes de valeurs ?
5. Comment quantifier la multifonctionnalité des sols, en prenant en compte les synergies et antagonismes entre fonctions et en cherchant un compromis entre complexité des modèles de simulation et leur compréhension par les usagers ?

Axes de recherche Le projet s'articule autour de trois axes de recherche principaux, correspondant à des types de modélisations complémentaires, et de deux axes transversaux méthodologiques qui répondent à des problématiques communes aux trois premiers axes et dont les techniques soit existent et assistent à la réalisation des axes principaux, soit nécessiteront des développements méthodologiques spécifiques.

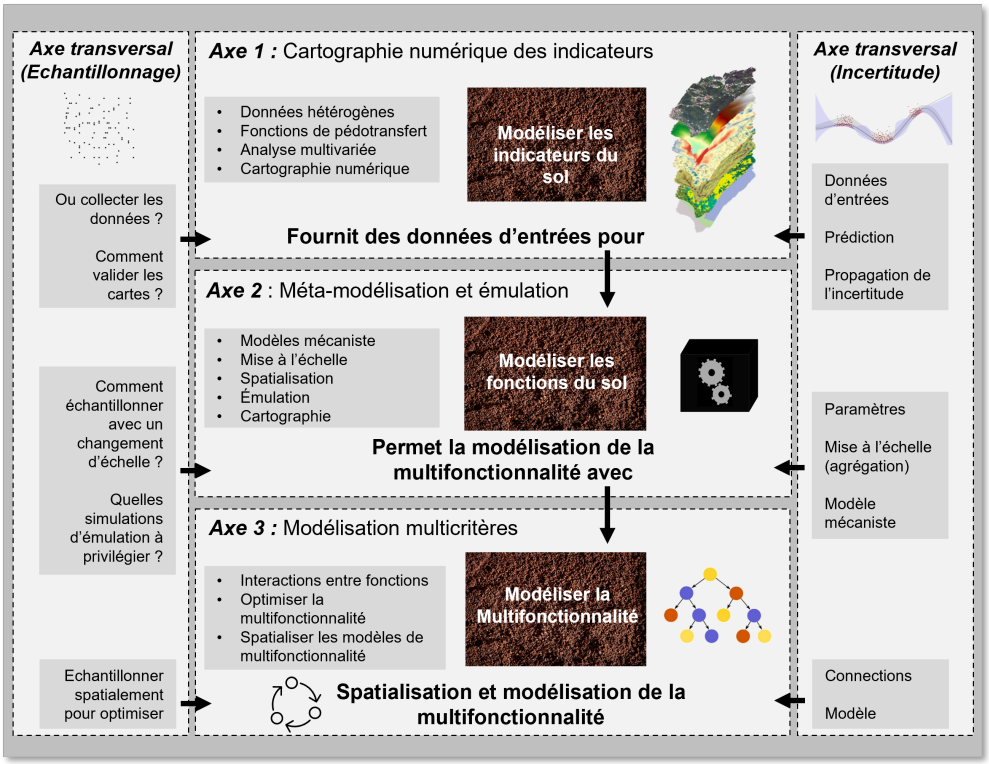

Figure 3.1 *– Représentation schématique des trois axes de recherche et des deux axes transversaux pour le projet de recherche.*

3.2 Axe 1 : Cartographie numérique des indicateurs

Pour spatialiser les fonctions du sol, il faut spatialiser les indicateurs. Ceux-ci doivent être spatialement explicites, mais aussi à résolution spatiale fine (p.ex. parcelle de taille réduite en contexte Sud) afin de pouvoir les combiner pour développer une modélisation des fonctions. Cet axe est crucial pour la réalisation des Axes 2 et 3, dans la mesure où la quantification des indicateurs est souvent un facteur limitant dans la modélisation de la multifonctionnalité des sols. Cet axe est aussi une poursuite de certains de mes travaux passés. Bien que les approches de cartographie numérique des sols monovariées aient donné lieu à de nombreuses publications ces dernières années et ont aussi fait l'objet de plusieurs de mes recherches passées, certains développements méthodologiques sur les approches multivariées restent à approfondir pour pourvoir mobiliser ces techniques dans l'évaluation des fonctions du sol.

Dans ce contexte, cet axe visera à

(i) définir et développer des approches statistiques multivariées qui prennent en compte les interactions entre indicateurs (par exemple, argile et limon ou carbone et pH),

(ii) développer des méthodes statistiques pour combiner les nouvelles technologies de mesures (p.ex. spectroscopie) et fonctions de pédotransfert à l'aide de bases de données locales et nationales pour permettre, suivant un jeux de d'indicateurs simples, d'estimer les indicateurs difficiles à obtenir (propriétés hydrauliques, stabilité structurale) ou coûteux (minéraux argileux),

(iii) spatialiser les indicateurs à résolution spatiale fine avec une quantification de l'incertitude.

Les méthodes mobilisées veilleront à explorer des méthodes statistiques et celles faisant appel aux techniques d'apprentissage automatique. Mes recherches passées m'orientent vers l'utilisation des méthodes

d'apprentissages automatiques non-linéaires. Je m'intéresserai à la façon dont la dépendance entre indicateurs peut être conservée dans la prédiction de ces indicateurs, à travers

(i) la modélisation explicite de la corrélation avec l'étalonnage conjointe d'un modèle autorégressif entre indicateurs,

(ii) la modification de la fonction de perte pour inclure la différence de corrélation entre les indicateurs et leur prédiction,

(iii) la création d'un modèle unique avec une architecture partagée entre les indicateurs.

Ces tests sont possibles pour tous les modèles qui incluent une fonction de perte (par exemple, les réseaux de neurones), mais d'autres modèles seront considérés, telles que les forêts d'arbres décisionnels multivariés. Pour les deux autres points évoqués dans cet axe, les techniques devront permettre de prendre en compte les données de source hétérogène et d'incertitude variables. Ces techniques seront décrites dans l'Axe transversal (Incertitude). Des avancées récentes dans les technologies (p.ex. spectroscopie) doivent être utilisées conjointement avec les méthodes statistiques de chimiométrie pour développer des fonctions de pédotransfert qui prennent en compte les nouveaux acteurs (p.ex. les données qui viennent des approches participatives). Finalement, la spatialisation des indicateurs devra étendre les modèles statistiques et d'apprentissage automatique précédemment évoqués avec les techniques de géostatistiques qui prennent en compte la corrélation spatiale. Je m'intéresserai particulièrement aux modèles dit « hydrides » qui permettent de concilier prise en compte de la corrélation spatiale des indicateurs et modélisation très précise à l'aide de modèles non-linéaires.

3.3 Axe 2 : Méta-modélisation et émulation

Une alternative à la modélisation des fonctions avec une suite d'indicateurs est l'utilisation des modèles basés sur les processus du sol pour extraire les fonctions. Cet axe vise à articuler la compréhension des processus du sol au service de la quantification des fonctions du sol. Certaines fonctions prioritaires telles que la dynamique du carbone organique du sol ou l'érosion pourront être spatialisées avec des modèles basés sur des processus (process-based model). Ceux-ci permettent d'estimer les réserves et les flux de carbone organique et de nutriments du sol pour les agroécosystèmes avec un degré de complexité variable et peu de données d'entrées requises (p.ex., Century ou LPJ-GUESS). Cependant, très peu d'études se sont intéressées à l'utilisation de ces modèles mécanistes pour la modélisation des fonctions, principalement à cause des difficultés à spatialiser ces modèles complexes développés pour une application à l'échelle d'une parcelle ou d'un profil de sol. Notons toutefois quelques tentatives récentes (p.ex Choquet et al., 2021) sur lesquelles il faudra s'appuyer.

Cet axe visera à résoudre certains verrous méthodologiques en

(i) contribuant au développement de techniques de compréhension des échelles spatiales pour soutenir l'utilisation de modèles mécanistes (développés à une échelle fine) à l'échelle du paysage, de la région et du pays,

(ii) ajoutant une dimension spatiale et temporelle aux modèles mécanistes,

(iii) palliant aux déficiences des modèles mécanistes (temps de calcul long) pour l'utilisation sur micro-ordinateur et avec une application sur une large échelle (émulation).

La spatialisation des modèles basés sur les processus se fera en testant des stratégies diverses :

(i) spatialisation les paramètres d'entrée du modèle,

(ii) spatialisation des paramètres du modèle

(iii) spatialisation des sorties du modèle mécaniste.

Les comparaisons entres stratégies seront faites à partir d'une modélisation de l'incertitude (voir aussi l'Axe transversale). Finalement, pour pallier le temps de calcul long des modèles mécanistes, des techniques d'émulations seront développées. À ma connaissance, elles n'ont jamais été développées pour des modèles sols. L'émulation consistera à développer un modèle de régression qui est étalonné pour relier les jeux de paramètres d'entrée et les sorties des modèles mécanistes. Le modèle de régression pourrait s'appuyer par exemple sur des techniques d'apprentissage automatique ou sur des modèles géostatistiques telles que le surrogate kriging. Enfin, j'utiliserai les théories de mise à l'échelle avec des techniques de krigeage de bloc qu'il faudra adapter si des techniques de machine learning sont employées et que l'incertitude doit être estimée. En effet, l'incertitude d'une agrégation d'une variable spatiale diminue avec une augmentation de la taille de l'aire. De manière alternative, des techniques d'échantillonnage composées peuvent être développées.

3.4 Axe 3 : Modélisation multicritères

La plupart des études existantes ont estimé des fonctions du sol individuellement, et de manière indépendante, sans prendre en compte les synergies et antagonismes entre fonctions. Il existe actuellement des connaissances limitées sur la façon dont la capacité d'un sol à fournir des fonctions et l'interaction entre elles varient dans l'espace géographique. Des études récentes (p.ex. Rabot et al., 2022; Zwetsloot et al., 2021) ont suggéré la possibilité de développer des outils statistiques et empiriques pour cartographier la multifonctionnalité. Elles fournissent une preuve qu'une telle approche est faisable. Dans cet axe, je propose la tâche ambitieuse de développer des méthodes de cartographie de la multifonctionnalité. La façon dont les synergies et antagonismes entre les fonctions varient avec le climat, le type de sol et l'occupation des sols sera explorée pour la première fois. Leurs potentiels effets non-additifs sont imprévisibles à partir des études modélisant les fonctions individuellement. Cet axe s'appuie sur l'estimation des indicateurs de l'Axe 1 et des fonctions de l'Axe 2.

Cet axe visera à développer des modèles statistiques existants et à innover pour construire des nouvelles techniques afin de

(i) prendre en compte les interactions (synergies et antagonismes) entre fonctions du sol,

(ii) comprendre comment ces interactions varient dans le paysage,

(iii) proposer des techniques pour optimiser la provision de la multifonctionnalité pour différent types de sol et d'occupation des terres,

(iv) spatialiser les modèles de multifonctionnalité.

Je m'appuierai sur les modèles développés et appliqués principalement en sociologie et écologie telles que le modèle d'équations structurelles, les réseaux de cooccurrence, et les réseaux de probabilités Bayésienne. Ce dernier, en particulier, modèle explicitement la corrélation et permettrait de combiner modélisation statistique et expertise sur les sols. Par exemple, les connections entre fonctions peuvent être examinées par des experts et rejetées si elles ne sont pas liées à un processus de sol existant. Je m'intéresserai à formuler un réseau Bayésien spécifique à chaque type d'occupation des sols, en utilisant les indicateurs de sols (Axe 1) comme variables explicatives et les fonctions (Axe 2) comme variables dépendantes. Les fonctions devront être standardisées dans une unité commune (par exemple, entre 0 et 1) pour permettre leur combinaison. J'explorerai ensuite l'étalonnage de trois catégories de réseaux bayésiens,

(i) l'étalonnage de deux réseaux distincts pour discriminer entre le potentiel intrinsèque du sol à remplir diverses fonctions -sa capacité, et l'état actuel de ce dernier -sa condition,

(ii) étalonnage d'un seul réseau pour capacité et condition,

(iii) étalonnage d'un seul réseau pour la capacité, la condition et pour les différents types d'utilisation des terres.

La dernière approche sera testée en utilisant le type d'occupation des sols comme variable explicative. Finalement, la spatialisation d'un réseau Bayésien est un défi en soi, mais plusieurs pistes sont envisagées : i) inclure la corrélation spatiale explicitement à l'aide de techniques géostatistiques de type variographie, ii) étalonner un modèle et l'appliquer sur les cartes d'indicateurs, cela serait une approche simple et efficace sur le plan des calculs, et enfin iii) des cartes d'indicateurs de sol sont utilisées pour estimer les cartes des scores individuels des fonctions du sol (entre 0 et 1, en utilisant la même méthodologie que celle en i) sur lequel un nouveau modèle est étalonné. C'est une méthode conceptuellement simple mais qui requiert une puissance de calculs phénoménale. Pour pallier cette difficulté, des méthodes d'échantillonnages spatiales devront être mise en place (voir Axe transversal : échantillonnage).

3.5 Axe transversal : Collection des données et échantillonnage spatial

L'axe transversal sur la collection des données et échantillonnage spatial représente un ensemble de techniques qu'il faudra mobiliser pour optimiser et pallier les problèmes de ressources (en données, et en puissance de calcul) pour obtenir des résultats aux trois axes principaux. Je mobiliserai mes connaissances sur mes recherches passées pour développer des nouvelles méthodes d'échantillonnage et optimiser l'utilisation des données existantes. Ces méthodes permettront de répondre à des verrous méthodologiques des trois axes présentés plus haut :

- Axe 1 : Où collecter les données en prenant en compte les données existantes ? Les données de sols peuvent être chères à collecter en grand nombre. Il faudra développer des techniques d'échantillonnage spatiale qui prennent en compte les données existantes, les nouvelles technologies (télédétection), et le coût d'accès pour obtenir de nouveaux échantillons. J'utiliserai des techniques d'optimisation basées sur le recuit simulé spatial.
- Axe 1 : Comment valider les cartes d'indicateurs ? Nous savons qu'un échantillon probabiliste est optimal pour valider des cartes d'indicateurs. Il faudra développer des méthodes d'échantillonnage probabilistes qui utilisent des données existantes.
- Axe 2 : Comment échantillonner avec un changement d'échelle ? Développer des approches d'échantillonnage multi-échelle est un défi, mais des pistes de recherche sur la théorie de valeur de l'information pourraient être développées en combinaison avec des approches de spectroscopie.
- Axe 2 : Quelles simulations sont à privilégier pour l'émulation ? Le recourt à l'émulation admet que le modèle a un temps de calcul très long. Il faut choisir où (dans le temps et dans les espaces géographiques) les simulations doivent être réalisées pour optimiser le temps de calcul. J'utiliserai des méthodes d'échantillonnage qui prennent en compte la variabilité environnementale de l'aire d'étude (p.ex. Latin hypercube sampling).
- Axe 3 : Comment choisir un sous-échantillon pour étalonner le modèle ? L'Axe 3 demande d'étalonner un modèle en utilisant un nombre très important de données. Il faudra utiliser des techniques d'échantillonnage spatiale optimisées pour réduire le temps de calcul sans compromettre la qualité de prédiction.

3.6 Axe transversal : Estimation et propagation de l'incertitude

Cet axe transversal sur l'estimation et la propagation de l'incertitude vise à utiliser et développer des techniques qui viendront compléter les modélisations des trois axes principaux. Je mobiliserai des compétences

sur les statistiques Bayésiennes et la quantification et propagation de l'incertitude. Les questions spécifiques envisagées constituant des verrous méthodologiques pour les trois axes précédemment présentés sont comme suit.

- Axe 1 : Incertitudes dans les données d'entrées. Nous savons que différentes sources d'erreurs affectent les données sol. De plus, j'utiliserai aussi des données issues des approches participatives. Ces erreurs sont rarement quantifiées mais peuvent être substantielles. J'utiliserai des approches qui prennent en compte l'erreur des observations sols, en incluant possiblement la corrélation croisée spatiale. Cela peut être fait en utilisant des méthodes Monte-Carlo.
- Axe 1 : Propagation de l'incertitude. L'erreur des observations affecte la prédiction. Il est important de propager cette erreur dans la modélisation spatiale. C'est possible en utilisant des techniques géostatistiques mais reste un défi pour des méthodes d'apprentissage automatique qui ne prennent pas en compte la structure spatiale de l'erreur. Des approches basées sur une pondération des observations ont été testées dans mes recherches passées, mais l'utilisation de simulations stochastiques me semblent une piste de recherche prometteuse. De manière alternatif, des techniques utilisant des séries de Taylor peuvent aussi être employées.
- Axe 2 : Incertitude des paramètres et du modèle mécaniste. Le modèle mécaniste souffre de plusieurs sources d'erreurs, principalement celles des paramètres et du modèle en lui-même. D'autres types d'erreurs existent et pourraient aussi être pris en compte (p.ex. erreurs des données d'étalonnage et de validation). Des approches basées sur les statistiques Bayésiennes peuvent être utilisées, bien que cela n'a pas encore été développé pour des modèles de ce type en sciences du sol. Mes résultats sur des modèles hydrologiques me rendent confiant de la possibilité d'appliquer ces techniques Bayésiennes aux modèles mécanistes sol.
- Axe 2 : Prise en compte de l'incertitude dans le changement d'échelle spatiale. L'agrégation spatiale a un effet direct sur l'incertitude de la prédiction, car les erreurs positives et négatives s'annulent lorsque la moyenne est prise. Des techniques de krigeage de bloc existent mais celles-ci n'ont pas encore été étendues pour une utilisation hybride avec de l'apprentissage automatique.
- Axe 3 : Incertitude des connections et du modèle. Le modèle de neurones Bayésiens est complexe mais permet de quantifier l'incertitude des connections et du modèle. Il faudra comprendre comment cette incertitude se propage dans la prédiction et peut être communiquée à l'usager.

3.7 Vers une quantification de la sécurité des sols ?

Dans une vision long terme de ce projet de recherche, la quantification et la spatialisation des fonctions et des services fournis par les sols peuvent être pensés dans un cadre plus global, tel que celui de **sécurité des sols** énoncé récemment (*soil security*, voir McBratney et al. (2014)). La concept de sécurité des sols comprend cinq dimensions : la condition, la capacité, la codification, la connectivité et le capital. Chacune de ces dimensions reconnaît un aspect d'interaction entre les hommes et les sols, ancrant ce concept dans les défis sociétaux auxquels l'humanité est confrontée. Une myriade de concepts existent cependant déjà et reconnaissent que la dégradation des sols peut avoir un impact sur la productivité agricole et les services écosystémiques : conservation des sols, qualité des sols, santé des sols ou encore protection des sols. Ils tentent tous de placer le sol au centre et de rendre justice à la nécessité de maintenir et de gérer la condition du sol. La Figure 3.2 replace ces concepts dans le cadre de la sécurité des sols et prend en compte pour chaque dimension les fonctions et services énoncés précédemment dans ce projet de recherche, complété par la prise en compte des menaces sur les sols.

Dans Evangelista et al. (2023) nous avons récemment fait une proposition d'évaluation de la sécurité des sols. Chaque dimension peut être évaluée par rapport aux fonctions, services et menaces sur les sols.

Cela ouvre la voie à une quantification de la sécurité des sols car des indicateurs peuvent être utilisés pour mesurer et quantifier chaque dimension. Dans le projet décrit ci-avant, je propose de quantifier les fonctions fournies par les sols en différenciant la condition actuelle du sol et son état potentiel (c.a.d. sa capacité). De nouvelles perspectives de recherches, parfois pluridisciplinaires, s'ouvrent en considérant le cadre conceptuel de la Figure 3.2. Il s'agira de quantifier dans un premier temps les menaces pesant sur la capacité et la condition actuelle du sol. Ces menaces sont, par exemple, l'érosion, la salinisation ou la perte de carbone. Cette dernière peut être évaluée avec des indicateurs telles que le contenu de carbone actuel et celui des genosols, pour quantifier respectivement les menaces sur la condition et la capacité.

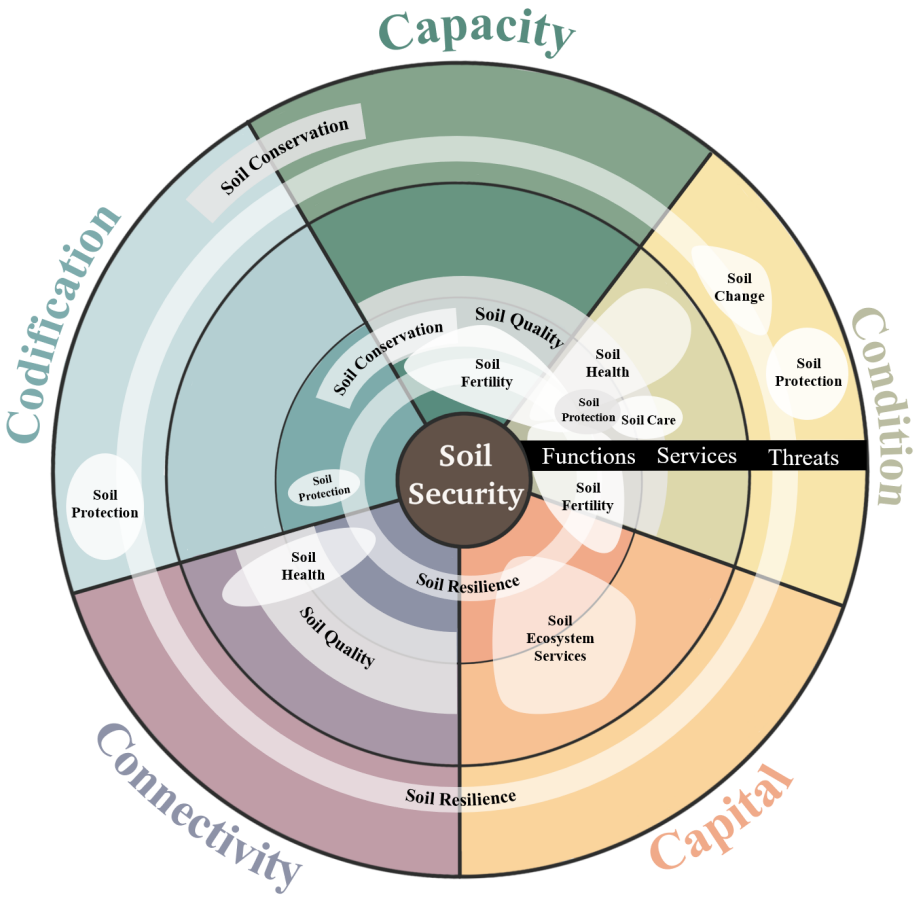

***Figure 3.2** – Les cinq dimensions de la sécurité des sols en rapport avec les fonctions, services et menaces sur les sols. Les concepts existant sont replacés dans ce cadre conceptuel. D'après Evangelista et al. (2023).*

La quantification des autres dimensions de la sécurité des sols : le capital, la connectivité et la codification, pourra se faire avec des collaborations en socio-économie et le recours à des approches participatives. La fonction stockage de carbone, par exemple, pour la dimension de capital pourra se quantifier avec des indicateurs sur la présence d'un marché de crédit de carbone, ou le coût de mise en place de technologies pour la séquestration du carbone, Pour la dimension de connectivité, cela se fera sûrement avec enquêtes participatives dans lesquels sont évalués la perception de l'importance de maintenir les stocks de carbones de sol, ou le niveau de connaissance sur les pratiques de stockages de carbone. La codification enfin, s'évaluera avec des indicateurs d'absence ou de présence de directives locales ou internationales sur la réglementation du stockage de carbone ou de politiques d'incitations visant à stocker du carbone dans les sols agricoles.

3.8 Conclusion

Je mets le développement de méthodes de quantification et de modélisation spatiale des fonctions du sol au cœur de ce projet de recherche. Mon inscription au sein d'un groupe pluridisciplinaire permettra une **fertilisation de ces développements méthodologiques avec des spécialistes des processus pédologiques au sein du paysage**. Les approches que je développerai pourront être étendues à la modélisation d'un ensemble de processus du paysage et de services écosystémiques, par exemple sur l'érosion des sols, les contaminants, la retenue dans le paysage agricoles, ainsi qu'à un ensemble de processus hydrologiques de surface. Là encore, **la multiplicité des compétences nécessaires pour évaluer la multifonctionnalité du paysage se joindra à mes compétences sur la modélisation et l'analyse spatiale**.

Ce projet n'est réalisable qu'avec **une ambition forte en terme de financement et d'encadrement**. La mise en avant des sujets de services écosystémiques fournis par les sol au niveau national (p.ex. sur l'artificialisation des sols) mais surtout européen (p.ex. *EU Mission : A Soil Deal for Europe*) me permettra de chercher des financements pour réaliser le projet présenté. De plus, mon réseau européen me facilitera le montage et la candidature à des projets européens de grandes envergures qui regroupent plusieurs institutions. J'envisage un rythme d'encadrement avec au moins un post-doctorant, deux doctorants et deux étudiants en stage de Master/ingénieur par an. Certaines perspectives de recherches seront cependant bien mieux entreprises avec des financements à l'échelle locale et nationale. Je considérerai donc aussi la candidature à des projets nationaux via l'agence nationale de la recherche et les agences régionales. Plus ponctuellement, des fonctions spécifiques peuvent intéresser le secteur privé avec lequel des partenariats pour l'encadrement de thésards sont à prévoir, par exemple à travers le dispositif des conventions industrielles de formation par la recherche (Cifre). Un rythme d'encadrement de jeunes chercheurs soutenu et une recherche de financements réussie sur plusieurs projets me donnent une vision tangible de ma progression en tant que directeur de recherche dans un horizon de 5 à 10 ans. Dans le même élan, l'**animation de la recherche** locale, nationale et internationale sera un point clef pour permettre cette progression. À court terme, mon intégration permanente au sein d'un groupe de recherche me permettra de prétendre à l'animation de la recherche, par exemple l'animation de l'équipe sol au sein du LISAH. Sur le long terme, je compléterai ces animations avec des responsabilités supplémentaires, telles que la direction d'UMR ou ma participation à des comités de pilotage nationaux.

Un glissement de la quantification des fonctions du sol vers la sécurité des sols sur le long terme ne pourra se faire sans reconnaître qu'un élargissement du champ disciplinaire est nécessaire. **Plusieurs des thématiques abordées dans la sécurité des sols ne pourront être traitées avec les connaissances immédiatement mobilisables au sein d'une seule UMR**. Je mettrai en place des collaborations pour répondre à ces questions. Les questions de connectivité, de codification et d'évaluation monétaire pourraient être étudiées en parallèle avec des approches de sociologie et d'économétrie spatiale. Ces aspects sont à ma connaissance rarement traités dans les unités sols, mais des partenariats seront à créer avec des unités spécialistes en France. Des montages de projet ainsi que des doctorants chevauchant plusieurs UMR sont à prévoir. Une partie en rapport avec l'évaluation monétaire des services écosystémiques fournis par les sols pourraient servir à des initiatives de collaborations avec des entreprises ou des institutions publiques désireuses de connaître le prix d'un sol, de sa dépollution, de sa conservation ou du coût potentiel de son artificialisation.

References

Baccini, A., Goetz, S.J., Walker, W.S., Laporte, N.T., Sun, M., Sulla-Menashe, D., Hackler, J., Beck, P.S.A., Dubayah, R., Friedl, M.A., et al., 2012. Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. Nature Climate Change 2, 182–185.

Bárdossy, A., Das, T., 2008. Influence of rainfall observation network on model calibration and application. Hydrology and Earth System Sciences 12, 77–89.

Barrera-Bassols, N., Zinck, J.A., 2003. Ethnopedology : a worldwide view on the soil knowledge of local people. Geoderma 111, 171–195.

Bonney, R., Ballard, H., Jordan, R., McCallie, E., Phillips, T., Shirk, J., Wilderman, C.C., 2009. Public Participation in Scientific Research : Defining the Field and Assessing Its Potential for Informal Science Education. A CAISE Inquiry Group Report. Washington, D.C. : Center for Advancement of Informal Science Education (CAISE).

Bouma, J., McBratney, A.B., 2013. Framing soils as an actor when dealing with wicked environmental problems. Geoderma 200, 130–139.

Brus, D.J., 2019. Sampling for digital soil mapping : A tutorial supported by R scripts. Geoderma 338, 464–480.

Brus, D.J., 2022. Spatial Sampling with R. CRC Press.

Brus, D.J., Spätjens, L.E.E.M., De Gruijter, J.J., 1999. A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. Geoderma 89, 129–148.

Bünemann, E.K., Bongiorno, G., Bai, Z., Creamer, R.E., De Deyn, G., de Goede, R., Fleskens, L., Geissen, V., Kuyper, T.W., Mäder, P., et al., 2018. Soil quality–A critical review. Soil Biology and Biochemistry 120, 105–125.

Choquet, P., Gabrielle, B., Chalhoub, M., Michelin, J., Sauzet, O., Scammacca, O., Garnier, P., Baveye, P.C., Montagne, D., 2021. Comparison of empirical and process-based modelling to quantify soil-supported ecosystem services on the saclay plateau (France). Ecosystem Services 50, 101332.

Cook-Patton, S.C., Leavitt, S.M., Gibbs, D., Harris, N.L., Lister, K., Anderson-Teixeira, K.J., Briggs, R.D., Chazdon, R.L., Crowther, T.W., Ellis, P.W., et al., 2020. Mapping carbon accumulation potential from global natural forest regrowth. Nature 585, 545–550.

De Bruin, S., Brus, D.J., Heuvelink, G.B.M., van Ebbenhorst Tengbergen, T., Wadoux, A.M.J.C., 2022. Dealing with clustered samples for assessing map accuracy by cross-validation. Ecological Informatics 69, 101665.

De Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for natural resource monitoring. Springer Science & Business Media, Dordrecht, NL.

Delgado-Baquerizo, M., Oliverio, A.M., Brewer, T.E., Benavent-González, A., Eldridge, D.J., Bardgett, R.D., Maestre, F.T., Singh, B.K., Fierer, N., 2018. A global atlas of the dominant bacteria found in soil. Science 359, 320–325.

Delhomme, J.P., 1978. Kriging in the hydrosciences. Advances in Water Resources 1, 251–266.

Evangelista, S.J., Field, D.J., McBratney, A.B., Minasny, B., Ng, W., Padarian, J., Román Dobarco, M., J.-C., W.A.M., 2023. A proposal for the assessment of soil security : soil functions, soil services and threats to soil. Under Review .

Heuvelink, G.B.M., Webster, R., 2022. Spatial statistics and soil mapping : A blossoming partnership under pressure. Spatial Statistics , 100639.

Irwin, A., 1995. Citizen science : A study of people, expertise and sustainable development. Routledge.

Kempen, B., Dalsgaard, S., Kaaya, A.K., Chamuya, N., Ruipérez-González, M., Pekkarinen, A., Walsh, M.G., 2019. Mapping topsoil organic carbon concentrations and stocks for Tanzania. Geoderma 337, 164–180.

Krige, D.G., 1951. A statistical approach to some basic mine valuation problems on the Witwatersrand. Journal of the Southern African Institute of Mining and Metallurgy 52, 119–139.

Künsch, H., Papritz, A.J., Schwierz, C., Stahel, W.A., 2013. Robust estimation of the external drift and the variogram of spatial data, in : ISI 58th World Statistics Congress of the International Statistical Institute, Eidgenössische Technische Hochschule Zürich.

Lagacherie, P., Arrouays, D., Bourennane, H., Gomez, C., Martin, M., Saby, N.P.A., 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 337, 1320–1328.

Lal, R., 2000. Soil management in the developing countries. Soil Science 165, 57–72.

Lark, R.M., 2000. Estimating variograms of soil properties by the method-of-moments and maximum likelihood. European Journal of Soil Science 51, 717–728.

Lark, R.M., Marchant, B.P., 2018. How should a spatial-coverage sample design for a geostatistical soil survey be supplemented to support estimation of spatial covariance parameters ? Geoderma 319, 89–99.

van Leeuwen, C.C.E., Mulder, V.L., Batjes, N.H., Heuvelink, G.B.M., 2022. Statistical modelling of measurement error in wet chemistry soil data. European Journal of Soil Science 73, e13137.

Lugato, E., Panagos, P., Bampa, F., Jones, A., Montanarella, L., 2014. A new baseline of organic carbon stock in european agricultural soils using a modelling approach. Global Change Biology 20, 313–326.

Marchant, B.P., Lark, R.M., 2004. Estimating variogram uncertainty. Mathematical Geology 36, 867–898.

Martin, M.P., Dimassi, B., Román Dobarco, M., Guenet, B., Arrouays, D., Angers, D.A., Blache, F., Huard, F., Soussana, J.f., Pellerin, S., 2021. Feasibility of the 4 per 1000 aspirational target for soil carbon : A case study for France. Global Change Biology 27, 2458–2477.

Matheron, G., 1963. Principles of geostatistics. Economic Geology 58, 1246–1266.

McBratney, A.B., Field, D.J., Koch, A., 2014. The dimensions of soil security. Geoderma 213, 203–213.

Meinshausen, N., 2006. Quantile regression forests. Journal of Machine Learning Research 7.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences 32, 1378–1388.

Nenkam, A.M., Wadoux, A.M.J.C., Minasny, B., McBratney, A.B., Traore, P.C.S., Falconier, G.N., Whitbread, A.M., 2022. Using homosoils for quantitative extrapolation of soil mapping models. European Journal of Soil Science , e13285.

Pietsch, S.A., Hasenauer, H., 2002. Using mechanistic modeling within forest ecosystem restoration. Forest Ecology and Management 159, 111–131.

Potschin, M.B., Haines-Young, R.H., 2011. Ecosystem services : Exploring a geographical perspective. Progress in Physical Geography 35, 575–594.

Rabot, E., Guiresse, M., Pittatore, Y., Angelini, M., Keller, C., Lagacherie, P., 2022. Development and spatialization of a soil potential multifunctionality index for agriculture (Agri-SPMI) at the regional scale. Case study in the Occitanie region (France). Soil Security 6, 100034.

Ramirez-Lopez, L., Stevens, A., Orellano, C., Rossel, R.V., Shen, Z., Lobsey, C., Wadoux, A.M.J.C., 2022. resemble : Regression and similarity evaluation for memory-based learning in spectral chemometrics. URL : https://CRAN.R-project.org/package=resemble. r package version 2.1.2.

Ramirez-Lopez, L., Wadoux, A.M.J.C., Franceschini, M.H.D., Terra, F.S., Marques, K.P.P., Sayão, V.M., Demattê, J.A.M., 2019. Robust soil mapping at the farm scale with vis–NIR spectroscopy. European Journal of Soil Science 70, 378–393.

Randall, A., 1988. What mainstream economists have to say about the value of biodiversity. Biodiversity 217, 217–23.

Refsgaard, J.C., van der Sluijs, J.P., Højberg, A.L., Vanrolleghem, P.A., 2007. Uncertainty in the environmental modelling process–a framework and guidance. Environmental Modelling & Software 22, 1543–1556.

Renard, B., Kavetski, D., Leblois, E., Thyer, M., Kuczera, G., Franks, S.W., 2011. Toward a reliable decomposition of predictive uncertainty in hydrological modeling : Characterizing rainfall errors using conditional simulation. Water Resources Research 47.

Rittel, H.W.J., Webber, M.M., 1973. Dilemmas in a general theory of planning. Policy Sciences 4, 155–169.

Robinson, D.A., Hockley, N., Dominati, E., Lebron, I., Scow, K.M., Reynolds, B., Emmett, B.A., Keith, A.M., de Jonge, L.W., Schjønning, P., et al., 2012. Natural capital, ecosystem services, and soil change : Why soil science must embrace an ecosystems approach. Vadose Zone Journal 11.

Rossiter, D.G., Liu, J., Carlisle, S., Zhu, A.X., 2015. Can citizen science assist digital soil mapping? Geoderma 259, 71–80.

Royle, J.A., Nychka, D., 1998. An algorithm for the construction of spatial coverage designs with implementation in SPLUS. Computers & Geosciences 24, 479–488.

Rumpel, C., Lehmann, J., Chabbi, A., 2018. '4 per 1,000' initiative will boost soil carbon for climate and food security. Nature 553.

Schindewolf, M., Schmidt, J., 2012. Parameterization of the EROSION 2D/3D soil erosion model using a small-scale rainfall simulator and upstream runoff simulation. CATENA 91, 47–55.

Schmidt, K., Behrens, T., Daumann, J., Ramirez-Lopez, L., Werban, U., Dietrich, P., Scholten, T., 2014. A comparison of calibration sampling schemes at the field scale. Geoderma 232, 243–256.

Schmidt-Traub, G., 2021. National climate and biodiversity strategies are hamstrung by a lack of maps. Nature Ecology & Evolution 5, 1325–1327.

Shirk, J.L., Ballard, H.L., Wilderman, C.C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, B.V., Krasny, M.E., et al., 2012. Public participation in scientific research : a framework for deliberate design. Ecology and Society 17.

Strasser, B., Baudry, J., Mahr, D., Sanchez, G., Tancoigne, E., 2019. "citizen science"? rethinking science and public participation. Science & Technology Studies 32, 52–76.

Stumpf, F., Goebes, P., Schmidt, K., Schindewolf, M., Schönbrodt-Stitt, S., Wadoux, A., Xiang, W., Scholten, T., 2017a. Sediment reallocations due to erosive rainfall events in the Three Gorges Reservoir Area, Central China. Land Degradation & Development 28, 1212–1227.

Stumpf, F., Schmidt, K., Behrens, T., Schönbrodt-Stitt, S., Buzzo, G., Dumperth, C., Wadoux, A., Xiang, W., Scholten, T., 2016. Incorporating limited field operability and legacy soil samples in a hypercube sampling design for digital soil mapping. Journal of Plant Nutrition and Soil Science 179, 499–509.

Stumpf, F., Schmidt, K., Goebes, P., Behrens, T., Schönbrodt-Stitt, S., Wadoux, A., Xiang, W., Scholten, T., 2017b. Uncertainty-guided sampling to improve digital soil maps. CATENA 153, 30–38.

Tibi, A., Therond, O., 2017. Évaluation des services écosystémiques rendus par les écosystèmes agricoles. une contribution au programme EFESE. INRA.

Van Groenigen, J.W., Stein, A., 1998. Constrained optimization of spatial sampling using continuous simulated annealing. Journal of Environmental Quality 27, 1078–1086.

Vaysse, K., Lagacherie, P., 2017. Using quantile regression forest to estimate uncertainty of digital soil mapping products. Geoderma 291, 55–64.

Voltz, M., Webster, R., 1990. A comparison of kriging, cubic splines and classification for predicting soil properties from sample information. European Journal of Soil Science 41, 473–490.

Von Korff, Y., Daniell, K.A., Moellenkamp, S., Bots, P., Bijlsma, R.M., 2012. Implementing participatory water management : recent advances in theory, practice, and evaluation. Ecology and Society 17.

Vrebos, D., Jones, A., Lugato, E., O'Sullivan, L., Schulte, R., Staes, J., Meire, P., 2021. Spatial evaluation and trade-off analysis of soil functions through Bayesian networks. European Journal of Soil Science 72, 1575–1589.

Wadoux, A.M.J.C., 2019. Using deep learning for multivariate mapping of soil with quantified uncertainty. Geoderma 351, 59–70.

Wadoux, A.M.J.C., Brus, D.J., 2021. How to compare sampling designs for mapping? European Journal of Soil Science 72, 35–46.

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.

Wadoux, A.M.J.C., Brus, D.J., Heuvelink, G.B.M., 2019a. Sampling design optimization for soil mapping with random forest. Geoderma 355, 113913.

Wadoux, A.M.J.C., Brus, D.J., Rico-Ramirez, M.A., Heuvelink, G.B.M., 2017. Sampling design optimisation for rainfall prediction using a non-stationary geostatistical model. Advances in Water Resources 107, 126–138.

Wadoux, A.M.J.C., Heuvelink, G.B.M., 2023. Uncertainty of spatial averages and totals of natural resource maps. Under Review .

Wadoux, A.M.J.C., Heuvelink, G.B.M., De Bruin, S., Brus, D.J., 2021a. Spatial cross-validation is not the right way to evaluate map accuracy. Ecological Modelling 457, 109692.

Wadoux, A.M.J.C., Heuvelink, G.B.M., Uijlenhoet, R., De Bruin, S., 2020a. Optimization of rain gauge sampling density for river discharge prediction using Bayesian calibration. PeerJ 8, e9558.

Wadoux, A.M.J.C., Malone, B., Minasny, B., Fajardo, M., McBratney, A.B., 2021b. Soil Spectral Inference with R : Analysing Digital Soil Spectra Using the R Programming Environment. Springer Nature, Cham.

Wadoux, A.M.J.C., Marchant, B.P., Lark, R.M., 2019b. Efficient sampling for geostatistical surveys. European Journal of Soil Science 70, 975–989.

Wadoux, A.M.J.C., McBratney, A.B., 2021a. Digital soil science and beyond. Soil Science Society of America Journal 85, 1313–1331.

Wadoux, A.M.J.C., McBratney, A.B., 2021b. Hypotheses, machine learning and soil mapping. Geoderma 383, 114725.

Wadoux, A.M.J.C., McBratney, A.B., 2023. Participatory approaches for soil research and management : A literature-based synthesis. Under Review .

Wadoux, A.M.J.C., Minasny, B., McBratney, A.B., 2020b. Machine learning for digital soil mapping : Applications, challenges and suggested solutions. Earth-Science Reviews 210, 103359.

Wadoux, A.M.J.C., Molnar, C., 2022. Beyond prediction : methods for interpreting complex models of soil variation. Geoderma 422, 115953.

Wadoux, A.M.J.C., Padarian, J., Minasny, B., 2019c. Multi-source data integration for soil mapping using deep learning. SOIL 5, 107–119.

Wadoux, A.M.J.C., Román Dobarco, M., Malone, B., Minasny, B., McBratney, A.B., Searle, R., 2023. Baseline maps of organic carbon of Australian soils. Under Review .

Wadoux, A.M.J.C., Román-Dobarco, M., McBratney, A.B., 2021c. Perspectives on data-driven soil research. European Journal of Soil Science 72, 1675–1689.

Wadoux, A.M.J.C., Saby, N.P., Martin, M.P., 2022a. Shapley values reveal the drivers of soil organic carbon stocks prediction. EGUsphere 2022, 1–25.

Wadoux, A.M.J.C., Samuel-Rosa, A., Poggio, L., Mulder, V.L., 2020c. A note on knowledge discovery and machine learning in digital soil mapping. European Journal of Soil Science 71, 133–136.

Wadoux, A.M.J.C., Walvoort, D.J.J., Brus, D.J., 2022b. An integrated approach for the evaluation of quantitative soil maps through Taylor and solar diagrams. Geoderma 405, 115332.

Walter, C., Bispo, A., Chenu, C., Langlais, A., Schwartz, C., 2015. Les services écosystémiques des sols : du concept à sa valorisation. Cahiers Demeter , 53–68.

Westman, W.E., 1977. How Much Are Nature's Services Worth ? Measuring the social benefits of ecosystem functioning is both controversial and illuminating. Science 197, 960–964.

Zwetsloot, M.J., van Leeuwen, J., Hemerik, L., Martens, H., Simo Josa, I., Van de Broek, M., Debeljak, M., Rutgers, M., Sandén, T., Wall, D.P., et al., 2021. Soil multifunctionality : Synergies and trade-offs across European climatic zones and land uses. European Journal of Soil Science 72, 1640–1654.

ANNEXE A

Tirés à part

Tiré à part relatif au Chapitre 2.2

Article original (A) : Wadoux, A.M.J-C., Marchant, B.P. and Lark, R.M. (2019). Efficient sampling for geostatistical surveys. European Journal of Soil Science, 70, 975-989.

Tiré à part relatif au Chapitre 2.3

Article original (A) : Wadoux, A.M.J-C., Minasny, B. and McBratney, A.B. (2020). Machine learning for digital soil mapping : Applications, challenges and suggested solutions. Earth-Science Reviews, 210, 103359.

Tiré à part relatif au Chapitre 2.4

Article original (A) : Wadoux, A.M.J-C., Heuvelink, G.B.M., Uijlenhoet, R. and De Bruin, S. (2020). Optimization of rain gauge sampling density for river discharge prediction using Bayesian calibration. PeerJ, 8, e9558.

Tiré à part relatif au Chapitre 2.5

Article original (A) : Wadoux, A.M.J-C., Walvoort, D.J.J. and Brus, D.J. (2022). An integrated approach for the evaluation of quantitative soil maps through Taylor and solar diagrams. Geoderma, 405, 115332.

Tiré à part relatif au Chapitre 2.6

Article original (A) : Wadoux, A.M.J-C., McBratney, A.B. (2021). Digital soil science and beyond. Soil Science Society of America Journal, 85, 1313-1331.

A.1 Tiré à part du Chapitre 2.2

D'après :
Wadoux, A.M.J-C., Marchant, B.P. and Lark, R.M. (2019). Efficient sampling for geostatistical surveys. European Journal of Soil Science, 70, 975-989. https://doi.org/10.1111/ejss.12797

A.2 Tiré à part du Chapitre 2.3

D'après :
Wadoux, A.M.J-C., Minasny, B. and McBratney, A.B. (2020). Machine learning for digital soil mapping : Applications, challenges and suggested solutions. Earth-Science Reviews, 210, 103359.
https://doi.org/10.1016/j.earscirev.2020.103359

A.3 Tiré à part du Chapitre 2.4

D'après :
Wadoux, A.M.J-C., Heuvelink, G.B.M., Uijlenhoet, R. and De Bruin, S. (2020). Optimization of rain gauge sampling density for river discharge prediction using Bayesian calibration. PeerJ, 8, e9558. https://doi.org/10.7717/peerj.9558

A.4 Tiré à part du Chapitre 2.5

D'après :
Wadoux, A.M.J-C., Walvoort, D.J.J. and Brus, D.J. (2022). An integrated approach for the evaluation of quantitative soil maps through Taylor and solar diagrams. Geoderma, 405, 115332.
https://doi.org/10.1016/j.geoderma.2021.115332

A.5 Tiré à part du Chapitre 2.6

D'après :
Wadoux, A.M.J-C., McBratney, A.B. (2021). Digital soil science and beyond. Soil Science Society of America Journal, 85, 1313-1331. https://doi.org/10.1002/saj2.20296