REHABILITATION DE SITES POLLUES DANS UN CONTEXTE D'INCERTITUDE : STRATEGIES ADAPTATIVES POUR L'OPTIMISATION DE PROJETS

*(***Contaminated Sites Remediation under Uncertainty: Adaptive Strategies for Project Optimization***)*

Résumé en français

Ignacio Guridi

SARPI Remediation | Bordeaux INP

Table des matières

Chapitre 1 - Introduction *2*

Contexte Général 2

Classification de l'Incertitude 2

Gestion de l'Incertitude dans la Gestion des Sites Pollués 2

Portée de thèse 5

Chapitre 2 - Site d'étude *7*

Pollution des sols 8

Pollution des eaux 9

Discussion et conclusions 10

Chapter 3 - Estimation du volume de sol pollue *11*

Théorique de l'Estimation du Volume de Pollution et de l'Incertitude 11

Prédictions en présence de données biaisées 12

Estimation du Volume avec les Méthodes Classiques 13

Le Rôle des Variables de Jeu de Données et des Réseaux Neuronaux Artificiels dans l'Amélioration de l'Estimation de Volume 14

Méthodologie Evol : Estimation de l'Incertitude avec des Caractéristiques Descriptives et l'Utilisation de Réseaux Neuronaux Profonds 15

Résultats 17

Discussion 19

Conclusions 19

Chapitre 4 - Estimation du volume d'huile *20*

Vue d'ensemble théorique sur l'estimation de la saturation et du volume d'huile 20

Méthodologie EMob : Estimation de l'huile sur un grand site hétérogène 21

Résultats 23

Discussion 24

Conclusion 25

Chapter 5 - Analyse et prévision des flux du projet de réhabilitation *26*

Vue d'Ensemble Théorique sur l'Incertitude des Résultats Analytiques, en Laboratoire et Échantillonnage 26

Description du Projet de Remédiation 27

Méthodologie de prévision des volumes catégorisés 27

Resultats 30

Chapitre 1- Introduction

Contexte Général

En France, en 2020, il existe 8 254 sites pollués, principalement dus à la pollution industrielle par des hydrocarbures et des métaux lourds. L'huile minérale est un polluant majeur dans les environnements souterrains. L'incertitude quant à l'emplacement et au type de contamination affecte les plans de remédiation et les coûts. Les méga-sites avec de multiples utilisateurs compliquent davantage l'évaluation de la pollution et la remédiation. La revitalisation des sites industriels nécessite un équilibre entre les intérêts des parties prenantes, la valeur foncière, les coûts et le développement urbain. Les réglementations européennes, comme la Directive IED, mettent l'accent sur la prévention mais rencontrent toujours des problèmes de contamination. La France suit les réglementations européennes et dispose d'une méthodologie basée sur les risques pour la gestion des sites pollués. Il y a une préoccupation croissante concernant l'incertitude dans l'évaluation des sites contaminés, et des efforts sont en cours pour intégrer l'estimation de l'incertitude dans le processus d'évaluation.

Classification de l'Incertitude

L'incertitude dans les sites contaminés peut être classée en trois types principaux : l'inexactitude, la non-fiabilité et la frontière de l'ignorance. Elle découle de facteurs tels qu'une compréhension incomplète des médias ou de la pollution, des divergences dans les mesures, et des procédures d'estimation. Les évaluations environnementales doivent prendre en compte l'incertitude car elle impacte la précision et la dépendance des résultats. Les incertitudes dans la gestion des sites pollués peuvent résulter de données insuffisantes, d'une compréhension incomplète du contexte du site, d'erreurs d'échantillonnage, d'erreurs dans les modèles numériques, ou du manque d'informations réglementaires.

L'incertitude peut être qualitative ou quantitative. L'incertitude qualitative découle d'un manque de connaissance ou d'une aléa naturel, tandis que l'incertitude quantitative résulte de l'utilisation d'instruments, de méthodes ou de modèles.

Les erreurs dans les mesures scientifiques peuvent s'accumuler considérablement, avec des erreurs dans la préparation des échantillons variant de 100% à 300%, des erreurs de mesure des instruments de 2% à 20%, et des erreurs d'évaluation des données et de modélisation allant jusqu'à 50%. L'erreur totale dans les représentations peut atteindre jusqu'à 1000%, posant des défis pour l'obtention de résultats scientifiques précis.

Les sources d'incertitude peuvent être catégorisées comme aléatoires ou systématiques. L'incertitude aléatoire découle de fluctuations imprévisibles dans les données ou de la variabilité inhérente au système mesuré, tandis que l'incertitude systématique résulte d'erreurs constantes et prévisibles, généralement dues à des imperfections dans l'instrument de mesure, des biais ou des modèles.

Gestion de l'Incertitude dans la Gestion des Sites Pollués

La gestion environnementale des sites englobe une multitude d'activités, allant de l'évaluation initiale à la remédiation ultérieure et à la fermeture du site. À chaque étape de ce processus,

il existe une multitude de sources d'incertitude, provenant à la fois de facteurs aléatoires et systématiques

La phase d'évaluation prend une place centrale, où une analyse méticuleuse des éléments procéduraux de la gestion des sites environnementaux est effectuée. Chaque composant, de la caractérisation du site à la clôture post-remédiation, subit une dissection pour mettre en lumière les origines spécifiques de l'incertitude dans le processus. Ces complexités inhérentes sont détaillées de manière systématique, éclairant les sources de variabilité et de biais potentiels.

Au sein de la phase d'évaluation, le voyage commence par les évaluations de la pollution, une série d'étapes comprenant l'échantillonnage, l'analyse en laboratoire, la prédiction de la pollution, le transport, la conservation, et l'utilisation de modèles de précision. Chacune de ces étapes introduit un degré d'incertitude, à la fois aléatoire et systématique. L'échantillonnage, une étape critique, implique la collecte d'échantillons physiques sur le site pour analyser l'étendue et le type de contamination.

L'incertitude aléatoire dans le processus d'échantillonnage découle de la variabilité inhérente aux systèmes naturels, englobant la distribution de la pollution dans l'espace et dans le temps. Les caractéristiques uniques de chaque site, influencées par les paramètres naturels, peuvent avoir un impact significatif sur la distribution et la concentration de la pollution.

D'autre part, les erreurs systématiques ou les biais résultent principalement des décisions humaines prises lors du processus d'échantillonnage. Des facteurs tels que la stratégie d'échantillonnage choisie, la méthode d'échantillonnage et la méthodologie appliquée sont tous influencés par diverses considérations, y compris les ressources disponibles, la couverture temporelle de la zone d'étude, et les particularités spécifiques du site.

Le texte ajoute davantage de détails alors qu'il explore les sous-sections de la phase d'évaluation. Cela inclut la sélection de la zone et les paramètres, où la sélection d'une zone d'échantillonnage implique une recherche méticuleuse des informations historiques sur le site et des techniques complémentaires telles que les méthodes de dépistage analytique et géophysique. Les paramètres à analyser dans les échantillons sont généralement identifiés en fonction du problème spécifique en question ou d'une revue des informations de base.

Le nombre d'échantillons et leurs emplacements émergent comme des considérations critiques. Le texte cite le nombre optimal d'emplacements d'échantillonnage comme un facteur clé pour obtenir des prédictions fiables. Des facteurs tels que l'espacement, l'étendue, et le support jouent un rôle essentiel dans la conception d'un plan d'échantillonnage efficace. Le texte clarifie habilement ces termes, décrivant l'espacement comme la distance entre les paires d'échantillons, l'étendue comme l'étendue totale de la zone échantillonnée, et le support comme la superficie réelle de l'échantillon.

La variabilité spatiale de la matrice échantillonnée et la variabilité de la variable à mesurer influencent les ajustements à ces mesures. Un espacement trop grand peut passer à côté des zones à forte variabilité, tandis qu'une petite étendue pourrait négliger des régions importantes. Si le support est trop grand, la variabilité pourrait être dissimulée.

L'incertitude liée à la conception de l'échantillonnage est un autre aspect exploré dans le texte. Elle souligne l'importance de s'assurer que les échantillons collectés sont vraiment représentatifs de la population plus vaste ou du système environnemental sous étude. Ce concept s'aligne sur l'objectif plus large d'adresser l'incertitude au sein de la phase d'évaluation.

L'analyse en laboratoire vient ensuite, dans l'intention de fournir des données précises sur la présence et la concentration des contaminants. Cependant, des incertitudes systématiques et des biais découlent d'inexactitudes potentielles dans l'étalonnage des instruments, les limites de détection des méthodes, et même des facteurs humains tels que les erreurs de manipulation ou la contamination croisée.

L'incertitude aléatoire découle souvent de diverses actions dans le laboratoire, y compris la sélection de la méthode dans la préparation des échantillons, la manipulation des échantillons, les conditions de stockage, et la variabilité inhérente de l'instrument. Le texte cite des facteurs tels que l'homogénéisation des échantillons de sol, les fluctuations de température, l'exposition à la lumière, ou les conditions de stockage inappropriées comme des sources potentielles d'incertitude aléatoire.

Le sous-échantillonnage et l'extraction, des étapes critiques dans les procédures de laboratoire, contribuent de manière significative à l'incertitude globale des résultats. De légères déviations dans ces processus peuvent entraîner des disparités substantielles dans les résultats. L'extraction, qui isole un composant spécifique d'un sous-échantillon, aggrave encore cette incertitude. Son efficacité dépend de facteurs tels que le solvant utilisé, la durée de l'extraction, et la technique employée.

L'incertitude de mesure, une autre dimension de l'incertitude, découle du processus de mesure réel dans le laboratoire. Cela inclut des facteurs tels que l'étalonnage des instruments, la précision et l'exactitude de la méthode analytique. L'incertitude de mesure peut se manifester sous forme d'incertitude aléatoire due à la reproductibilité de l'instrument, et de biais lié au choix de la méthode et de l'instrument, à la performance de l'instrument (précision), à l'interférence avec d'autres composés, ou aux limitations des techniques analytiques.

Les modèles, essentiels pour prédire la distribution et le comportement des contaminants, sont également examinés. Ils utilisent des données d'entrée et des techniques mathématiques pour simuler le mouvement des polluants dans divers environnements. Les incertitudes aléatoires résultent souvent de la qualité et de l'exhaustivité des données d'entrée. Les incertitudes systématiques et les biais peuvent découler du choix du modèle, des hypothèses et des simplifications qui ne représentent peut-être pas pleinement les processus du monde réel.

L'incertitude inhérente au modèle, un aspect critique, découle des simplifications et des hypothèses nécessaires pour rendre les systèmes environnementaux complexes gérables sur le plan computationnel. Les analyses de sensibilité et d'incertitude jouent un rôle essentiel

dans l'identification et la quantification de ces sources d'incertitude inhérente, garantissant que les résultats sont fiables et peuvent éclairer les processus de prise de décision.

Le texte explore ensuite l'incertitude de la réponse structurelle dans le contexte des charges de base et des paramètres de résistance de base, soulignant que prédire avec précision les réponses structurelles typiques reste un défi, même avec des données d'entrée détaillées. Ce type d'incertitude intrinsèque, appelé incertitude du modèle, découle d'hypothèses générales, de conditions aux limites non spécifiées et des impacts imprévisibles de variables non prises en compte.

Dans les modèles d'interpolation, l'incertitude découle de l'utilisation de ces modèles pour prédire des valeurs dans des emplacements non échantillonnés dans le champ spatial. Cette incertitude émane de diverses sources, notamment le choix du modèle de variogramme dans les techniques géostatistiques ou la sélection de la méthode d'interpolation.

En passant à la phase de remédiation, le texte souligne comment l'incertitude lors de la phase d'évaluation peut entraîner des résultats différents, notamment la surestimation ou la sous-estimation de l'étendue de la contamination, le choix incorrect de stratégies de remédiation, l'augmentation du temps de projet et l'escalade des coûts. Une représentation graphique illustre le risque et la valeur en jeu au fil du temps, mettant en évidence que le risque le plus élevé et l'incertitude la plus grande surviennent pendant l'exécution.

La conception dans la phase de remédiation implique des décisions cruciales pour résoudre les problèmes de pollution et restaurer les sites à un état acceptable sur le plan environnemental. Cependant, l'incertitude subsiste, et la gestion appropriée de cette incertitude est essentielle pour prendre des décisions éclairées et efficaces tout au long du processus de gestion des sites pollués.

Portée de thèse

Cette recherche doctorale constitue une contribution majeure au domaine de la remédiation environnementale en se penchant sur la question complexe de la gestion de l'incertitude au sein des phases d'investigation et de remédiation des sites contaminés. L'objectif principal de cette étude est de prendre en compte le rôle essentiel de l'incertitude dans de tels projets et d'élargir son analyse au-delà des limites traditionnelles. Plutôt que d'examiner simplement l'incertitude dans des segments isolés des projets de remédiation, cette recherche s'efforce de comprendre de manière exhaustive et de quantifier l'influence de l'incertitude tout au long du cycle de vie de ces projets, de l'évaluation initiale à la remédiation.

Objectifs : La méthodologie de la recherche suit étroitement le développement d'un projet de remédiation de site réel, garantissant que les résultats sont solidement ancrés dans des applications pratiques du monde réel. La recherche est structurée en deux phases distinctes :

Évaluation Initiale et Prédiction : Dans la première phase du projet, la recherche se concentre sur une étude complète de la contamination du sol et des données sur le LNAPL obtenues avant le début de la remédiation. Des modèles prédictifs sont développés et appliqués pour estimer le volume du sol contaminé et l'étendue de la propagation du LNAPL. Cette phase pose les bases pour les analyses ultérieures.

Processus de Remédiation : À mesure que le projet progresse, la recherche déplace son attention vers l'étude complète de l'ensemble du processus de remédiation du site. Cette phase synthétise toutes les données collectées, les analyses et les résultats pour évaluer l'efficacité globale des stratégies de remédiation et l'exactitude des modèles prédictifs.

Utilisation des Résultats : Les résultats attendus de cette recherche ont le potentiel de bénéficier considérablement au domaine de la remédiation environnementale de plusieurs manières. Ces résultats incluent des modèles et des méthodologies validés qui peuvent être appliqués à des projets de remédiation similaires, un cadre complet pour comprendre et gérer les complexités et les incertitudes associées à la remédiation de sites, des informations et des outils précieux pour les praticiens et les décideurs dans le domaine de la remédiation environnementale, et une contribution au développement de méthodes plus efficaces et fiables pour le nettoyage des sites contaminés.

Structure du Manuscrit : Le manuscrit est méticuleusement organisé en six chapitres, chacun étant conçu pour servir un double objectif. Chaque chapitre fonctionne à la fois comme une entité de recherche indépendante et comme une partie intégrante de l'étude globale sur l'optimisation des projets de remédiation. Les chapitres suivent une structure cohérente, commençant par une introduction pour mettre en place le contexte, suivie d'un exposé détaillé du cadre théorique pour fournir un contexte académique. La section méthodologie de chaque chapitre décrit l'approche adoptée pour la collecte et l'analyse des données, spécifique à la focalisation du chapitre. Cela est suivi par une présentation des résultats, offrant des éclairages sur les conclusions de cet aspect particulier de l'étude. Enfin, chaque chapitre se conclut par une discussion qui résume les conclusions du chapitre et les lie aux implications plus larges de l'ensemble du projet. Cette structure réfléchie permet au manuscrit d'intégrer en douceur les différents aspects de la recherche en une histoire unifiée, garantissant une exploration approfondie et une connexion cohérente avec le thème global de la recherche.

Arrangement des Chapitres :

Chapitre 2 : Offre une description détaillée du site d'étude, y compris ses aspects géologiques et hydrogéologiques, la composition du sol, les détails sur la contamination, les enquêtes sur site et la contamination des eaux souterraines. Les statistiques descriptives dérivées des données recueillies sont également présentées.
Chapitre 3 : Se concentre sur l'étude de l'estimation du volume du sol contaminé par les hydrocarbures, en discutant des concepts fondamentaux, des modèles classiques et de l'impact de leur application dans des données biaisées. Il introduit une nouvelle méthodologie qui combine des techniques d'interpolation non paramétrique traditionnelles avec l'apprentissage profond pour aborder les incertitudes.
Chapitre 4 : Met l'accent sur l'estimation du volume du Liquide Non Aqueux en Phase Légère (LNAPL). Il commence par élucider les concepts généraux impliqués dans l'estimation de la saturation en huile. Ensuite, il examine les modèles les plus classiques utilisés à cette fin, explore les dernières avancées en matière d'estimation du LNAPL et propose une nouvelle méthodologie pour affiner les estimations.

Chapitre 5 : Explore l'exécution du projet de remédiation, examinant les incertitudes liées à l'échantillonnage, à l'analyse et à la gestion des polluants. Il introduit une méthode innovante pour simuler et optimiser le processus d'excavation et de rééchantillonnage, prévoyant la distribution des données sur les contaminants.
Chapitre 6 : Sert de discussion complète qui étend les conclusions tirées des chapitres précédents, fournissant un résumé cohérent des principaux résultats et conclusions de l'ensemble du travail, offrant d'autres éclairages et implications.

Chapitre 2- Site d'étude

Dans le sud-ouest de la France, la région du site étudié se trouve dans l'immense bassin d'Aquitaine, caractérisé géologiquement comme un bassin sédimentaire. Ce bassin, bordé par les collines armoricaines et vendéennes au nord, les Pyrénées au sud, et le "Massif Central" et la "Montagne noire" à l'est, présente une forme triangulaire, faisant face à l'océan Atlantique à l'ouest.

L'histoire géologique de la région est marquée par sa transformation d'un environnement marin à des conditions terrestres pendant le Crétacé inférieur. Cependant, l'époque du Crétacé supérieur a vu un retour aux conditions marines, coïncidant avec l'activité tectonique liée à l'ouverture du golfe de Gascogne. Cette incursion marine a contribué à une nouvelle déposition de sédiments, façonnant le paysage géologique de la région. À l'ère du Cénozoïque, une hétérogénéité notable des couches stratigraphiques est apparue, avec des formations distinctes des époques de l'Éocène et de l'Oligocène, caractérisées par des dépôts de grès, de calcaire et d'argile. Les périodes ultérieures, notamment le Miocène, le Pliocène et le Quaternaire, ont vu une érosion significative due à des facteurs tels que les fluctuations du niveau de la mer et les cycles glaciaires-interglaciaires, modifiant davantage les caractéristiques géologiques de la région.

La géologie de surface de la région d'étude désignée est principalement caractérisée par des dépôts fluviaux, comprenant des sables, des graviers, des limons et des argiles. De plus, des formations éoliennes, principalement composées de dépôts sableux, contribuent à l'hétérogénéité de la couche superficielle de la région.

La géologie immédiate en subsurface de la zone du site local est caractérisée par des sédiments du Pléistocène peu profonds, s'étendant à des profondeurs de 12 à 18 mètres. Ces sédiments se composent principalement de gravier et de sable, formant une couche supérieure. Sous les sédiments du Pléistocène se trouve une couche substantielle de marne du Tertiaire, caractérisée par son importante épaisseur. Les conditions de la nappe phréatique dans le site restent principalement non confinées, fluctuant saisonnièrement entre environ -2 et -3,5 mètres par rapport à la surface du sol.

Cette couche supérieure présente une hétérogénéité significative, tant verticalement qu'horizontalement, en raison des variations de la distribution de la taille des grains et de la présence sporadique de lentilles limoneuses. La composition locale du site s'aligne sur la zone environnante, mais présente également des variations, notamment des zones avec des

matériaux de remblai agrégés et des couches limoneuses entre les couches de sable fin et de sable moyen.

Pour valider les types de sols trouvés dans les puits et mesurer la porosité, 15 échantillons de sol ont été prélevés dans la zone nord du bâtiment de l'usine, à proximité des coupes géologiques C1 et C2. Les types de sols identifiés comprennent le remblai noir, les sables moyens, les sables grossiers, les graviers, les sables fins, les limons sableux, les sables silteux et les limons. Ces types de sols présentent des porosités variables.

Pollution des sols

La pollution des sols sur le site d'étude, qui englobe une ancienne usine automobile d'une superficie considérable de 496 000 m², est principalement due aux hydrocarbures lourds, en particulier l'huile minérale, qui affecte plusieurs zones.

Au fil du temps, l'utilisation de divers produits pétroliers a conduit à la contamination du sol par l'huile, qui s'infiltre et subit différentes étapes de dégradation. Cela a entraîné des différences dans les propriétés physiques de l'huile selon les emplacements. Des mesures de la viscosité de l'huile, de la densité relative et de la tension interfaciale ont été effectuées sur 25 échantillons, tandis qu'une tension interfaciale huile/eau a été mesurée sur 4 échantillons. Une campagne de référence complète axée sur les sols et les eaux souterraines a révélé l'existence de quatre zones principales où la contamination par l'huile a été identifiée, couvrant ensemble une superficie d'environ 15 000 m².

Entre octobre 2019 et septembre 2020, une vaste campagne de forage a été menée dans la zone d'étude, impliquant un total de 1 067 forages avec une taille de grille moyenne de 300 m², et des grilles de 100 m² dans les zones les plus polluées. Pour les grands sites, la règle théorique pour l'échantillonnage suggère de collecter un échantillon pour chaque 250 à 400 m² sur une grille espacée d'environ 15 à 20 mètres. À partir de chaque emplacement de forage, des échantillons individuels ont été prélevés à diverses profondeurs, notamment à des intervalles de 1, 2, 3, 4 et 5 mètres. De plus, un échantillon composite a été inclus pour l'analyse des éléments en trace. Dans l'ensemble, le nombre d'échantillons obtenus pour l'analyse s'élevait à 5802. En Europe, les sites à grande échelle dépassent rarement 100 à 200 forages, et ne dépassent jamais 400 à 500 échantillons. D'après l'expérience des experts consultés, les plus grands sites n'avaient pas plus de 1000 échantillons.

Le processus d'échantillonnage a été réalisé en deux campagnes distinctes, chacune correspondant à une entreprise différente impliquée dans l'étude, mais utilisant le même laboratoire et sur une période relativement courte en 2019. La stratégie employée a suivi une approche en deux phases. Initialement, les zones suspectes ont été examinées et échantillonnées avec soin. Ensuite, dans les zones où les concentrations les plus élevées ont été détectées, une approche d'échantillonnage en grille uniforme a été mise en œuvre, avec des grilles de 10, 15 et 25 en fonction de la zone. L'analyse visait à déterminer la présence et les niveaux de concentration des hydrocarbures totaux (TPH), des composés organiques volatils halogénés (COVH), des composés BTEX (benzène, toluène, éthylbenzène et xylènes), des hydrocarbures aromatiques polycycliques (HAP) et des métaux lourds.

Ces analyses montrent que les hydrocarbures totaux (TPH) étaient le groupe de contaminants présentant les concentrations les plus élevées. Pour cette raison, les TPH dans la plage de C10 à C40 ont été utilisés pour déterminer la zone à remédier. Certaines zones du site présentent également des concentrations élevées en composés chlorés (COVH).

La zone affectée par la pollution s'étend sur environ 50 000 m², avec une région fortement polluée couvrant 15 000 m² où les concentrations dépassent 15 000 mg/kg. Cependant, il est important de noter que d'autres zones polluées existent au-delà de cette région fortement touchée.

Une analyse détaillée des échantillons présentant des concentrations dépassant 5 000 mg/kg révèle des variations de distribution en fonction du type de sol et de la profondeur. Dans les deux premiers mètres de sol, principalement composés de matériaux de remblayage, les valeurs maximales n'excèdent pas 40 000 mg/kg, avec une moyenne de 20 000 mg/kg. À mesure que l'on descend à une profondeur de trois mètres, différents types de sols deviennent plus courants, et les niveaux d'hydrocarbures totaux (TPH) dépassent 80 000 mg/kg. Enfin, sur les deux derniers mètres, des sables grossiers et des graviers augmentent jusqu'à une profondeur de cinq mètres, où les niveaux de contaminants chutent à des valeurs proches de 60 000 mg/kg.

L'influence de la géologie sur la concentration de la pollution est analysée en détail dans les échantillons du site d'étude présentant des concentrations supérieures à 5 000 mg/kg TPH. Il existe une relation directe entre la concentration des hydrocarbures, le type de sol et la profondeur. Les deux premiers mètres de sol se composent principalement de matériaux de remblayage, avec une présence significative de graviers, de sables et de limons (SS, SF/SM et SG/G). Dans cette section, les valeurs maximales n'excèdent pas 40 000 mg/kg. À partir d'une profondeur de 2 mètres, différents types de sols deviennent plus courants, avec une augmentation des couches de sable fin et de limon sableux (SF/SM et Ls). Notamment, les niveaux de TPH augmentent et dépassent 80 000 mg/kg. La plage de concentration de 50 000 mg/kg.

Pollution des eaux

Le site a fait l'objet d’un suivi continue pendant plus d'une décennie, évaluant 250 puits installés entre 2006 et 2019. Les évaluations ont couvert les hydrocarbures, les composés chlorés, les niveaux d'eau souterraine et l'épaisseur du LNAPL. Notre principal axe de recherche porte sur l'épaisseur du LNAPL et les niveaux d'eau souterraine, certains puits ayant des mesures supplémentaires.

Les activités d'échantillonnage ont eu lieu de 2006 à 2019, coïncidant avec le début des efforts de remédiation des eaux souterraines en 2011. Au fil du temps, le panache de contamination a diminué progressivement. Cependant, des "points chauds" persistants ont indiqué la mobilité limitée du LNAPL sur le site.

Les investigations sur la relation entre le LNAPL et les eaux souterraines ont révélé des disparités dans le comportement des puits. La catégorisation en Catégorie 1 (formations homogènes) et Catégorie 2 (formations hétérogènes) a dévoilé des schémas distinctifs. Les

puits de Catégorie 1 ont montré une forte corrélation entre le LNAPL et les eaux souterraines, tandis que les puits de Catégorie 2 ont affiché des réponses complexes et imprévisibles.

De plus, des facteurs externes, tels que la présence de bâtiments, ont eu peu d'influence sur le comportement des puits. Dans des conditions extrêmes, notamment avec des niveaux d'eau souterraine élevés, le LNAPL pouvait disparaître entièrement des puits, ce qui revêt une importance particulière pour comprendre les défis liés à la récupération et à la surveillance de l'huile dans de tels scénarios.

L'analyse des variations de l'épaisseur du LNAPL dans 243 puits entre 2011 et 2019 a montré un lien avec les fluctuations des eaux souterraines. L'analyse de régression a catégorisé les puits en "augmentation", "diminution" ou "pas de changement" en fonction des tendances du LNAPL. La majorité des puits ont montré une diminution de l'épaisseur, tandis que certains présentaient des comportements associés à l'"effet lentille".

L'exploration de la corrélation entre les catégories géologiques et les tendances du LNAPL a indiqué que la géologie seule ne dictait pas les tendances à long terme de l'épaisseur pendant la remédiation. Une enquête plus approfondie dans les puits de la Catégorie 2 a révélé que la présence et la profondeur des couches d'argile/silt jouaient un rôle significatif dans l'influence sur les pentes des tendances du LNAPL.

Les données de 2019, choisies pour leur proximité avec l'échantillonnage des sols, ont offert des informations sur l'épaisseur du LNAPL. La distribution des données a révélé que plus de la moitié des échantillons avaient une épaisseur de LNAPL inférieure à 1 cm, tandis que d'autres dépassaient 4 cm. L'épaisseur du LNAPL présentait un schéma de distribution uniforme dans cette plage, avec une épaisseur moyenne de 35,90 cm. En excluant les valeurs de 0 cm, les données restantes ont affiché une distribution uniforme avec une augmentation linéaire après la marque des 50%.

Discussion et conclusions

Le site d'étude, couvrant 500 000 mètres carrés, détient un vaste ensemble de données provenant de plus d'une décennie de recherche. Il comprend 1 000 foreuses de sol, 5 000 échantillons et l'analyse de 130 puits. La géologie du site, influencée par les chaînes de montagnes proches, se compose principalement de sols sablonneux avec des couches intermittentes de limon, ce qui affecte la circulation des contaminants.

Les couches limoneuses ont un impact significatif sur la distribution des contaminants, en particulier des hydrocarbures, avec des concentrations plus élevées dans les sols plus fins et les couches de limon, atteignant plus de 80 000 mg/kg pour les hydrocarbures totaux (TPH). Les conditions de la nappe phréatique sont principalement libres, avec des fluctuations saisonnières d'environ -2 à -3,5 mètres.

Le comportement de LNAPL dépend des facteurs hydrogéologiques et géologiques, variant selon les catégories de puits (Catégorie 1 et 2). Les puits de la Catégorie 1, avec des formations adjacentes homogènes, suivent généralement les variations du niveau de la nappe

phréatique. Les puits de la Catégorie 2, avec des tailles de grains variables et des couches de limon, présentent des réponses complexes de LNAPL.

Bien que le LNAPL ait une mobilité limitée à l'échelle du site, certains puits montrent une augmentation de l'épaisseur tandis que d'autres puits voisins montrent une diminution. Les facteurs influençant le mouvement latéral de l'huile comprennent les caractéristiques géologiques, les couches de limon, les flux d'eau souterraine et les facteurs environnementaux. Le transfert d'huile augmente lorsque les couches de limon sont peu profondes (<3 mètres) mais diminue à des profondeurs >4 mètres, avec une imprévisibilité autour de 3-4 mètres.

L'analyse statistique de l'épaisseur du LNAPL révèle une large gamme de valeurs, avec un nombre significatif de mesures faibles.

Chapter 3- Estimation du volume de sol pollue

Théorique de l'Estimation du Volume de Pollution et de l'Incertitude

Méthodes d'Interpolation

Les méthodes d'interpolation, notamment les techniques géostatistiques (Kriging Ordinaire, Kriging Simple, Kriging Indicateur) et les méthodes d'apprentissage automatique (Forêt Aléatoire, Régression des Moindres Carrés Partiels, Réseau de Neurones Artificiels), ainsi que les méthodes traditionnelles comme la Régression Linéaire Multiple et le Poids Inverse de la Distance, sont essentielles pour prédire la variation spatiale de la pollution. La géostatistique, avec son accent sur les données spatiales, l'estimation de précision et la quantification des erreurs à travers les variogrammes, est particulièrement appréciée pour sa justification mathématique de la quantification de l'incertitude, ce qui la rend indispensable dans les études minières et environnementales. Les méthodes d'apprentissage automatique, adaptables aux relations non linéaires, influencent de plus en plus la modélisation géostatistique, montrant un potentiel prometteur dans les applications d'analyse spatiale.

Entrées et Observations

L'échantillonnage est crucial pour les modèles d'interpolation, car il fournit les données observationnelles nécessaires pour la délimitation des motifs spatiaux. Les stratégies d'échantillonnage, incluant l'échantillonnage aléatoire, systématique et stratifié, ainsi que les conceptions d'échantillonnage (régulier, adaptatif ou orienté), ont un impact significatif sur l'efficacité des modèles d'estimation de la pollution. Le choix de la stratégie affecte la capture de la variabilité spatiale, avec une préférence pour l'échantillonnage stratifié et systématique pour leur capacité à utiliser les connaissances préalables pour une analyse plus homogène des sous-groupes.

Kriging

Le Kriging est une technique géostatistique fondamentale pour l'interpolation spatiale, utilisant un modèle stochastique pour estimer les valeurs d'attributs aux emplacements non échantillonnés sur la base des observations à proximité. Il s'appuie sur des variogrammes pour analyser les dépendances spatiales, effectuant des prédictions précises en considérant à la fois la proximité spatiale et la variabilité observée. La capacité du Kriging à quantifier

l'incertitude d'interpolation le rend un outil précieux dans divers domaines, soutenu par sa robustesse méthodologique et les avancées computationnelles.

Le Variogramme
Le variogramme est essentiel pour comprendre les relations spatiales dans le Kriging, traçant les semi-variances contre les distances de décalage pour analyser la variabilité des points de données. Les composants clés incluent la distance de décalage, la semi-variance, le seuil, la portée et le pépite, chacun contribuant à l'interprétation de la dépendance spatiale. L'hypothèse de stationnarité est cruciale pour la précision du variogramme, avec la non-stationnarité abordée par des modifications ou des méthodes alternatives pour assurer des estimations de corrélation spatiale valides.

Kriging Ordinaire
Le Kriging Ordinaire, privilégié pour sa précision de prédiction et la robustesse de son modèle, suppose une stationnarité d'ordre deux—des propriétés statistiques constantes dans l'espace, dépendant uniquement de la distance entre les points d'échantillonnage. Il distingue entre les prédictions de point et de bloc, répondant à différents besoins d'analyse spatiale. La méthode dépend d'un variogramme bien construit et de transformations de données appropriées pour aborder les distributions non normales, soulignant la nécessité d'une sélection de modèle soigneuse et d'une validation croisée pour confirmer la fiabilité de la prédiction.

Prédictions en présence de données biaisées

Même si les techniques mentionnées sont les plus couramment utilisées et donnent d'excellents résultats, elles ne peuvent pas toujours être appliquées et nécessitent que la distribution des données d'entrée soit normale. Les prédictions peuvent être inexactes lorsque les données sont biaisées et contiennent des valeurs aberrantes en raison de la mauvaise représentation de la structure spatiale dans le variogramme ou de l'hypothèse de stationnarité. Pour cette raison, des techniques géostatistiques alternatives telles que le Kriging Ordinaire Logarithmique (LOK), le Kriging Indicateur (IK) et le Kriging Indicateur Multiple (MIK) ont été utilisées dans ces cas. Le Kriging Indicateur et le Kriging Indicateur Multiple sont largement appliqués pour l'estimation de la pollution lorsque les données sont biaisées. En effet, les techniques non paramétriques peuvent gérer un mélange modéré de données diverses sans supposition sur la distribution des données. L'Estimation de Densité par Noyau (KDE) avec le Réseau de Neurones est une méthode non géostatistique alternative largement utilisée pour les données biaisées qui peut être utilisée pour créer des surfaces de distribution avec des points chauds de pollution en utilisant une fonction de surface symétrique sur un point.

Kriging Ordinaire Lognormal (LOK)
Le Kriging Ordinaire Lognormal adapte les données biaisées en les transformant en une distribution log-normale, facilitant ainsi l'application des techniques de kriging. Cette transformation log-normale corrige le biais en normalisant la distribution des données, permettant une estimation plus précise des valeurs dans des emplacements non échantillonnés. Le processus implique l'application de transformations logarithmiques avant le kriging et l'utilisation d'une étape de rétro-transformation pour convertir les résultats estimés en leur échelle originale. Cette technique est particulièrement utile dans les cas où

une proportion significative des données est proche de zéro ou présente une distribution fortement asymétrique.

Kriging Indicateur (IK)
Le Kriging Indicateur est une méthode robuste pour estimer des distributions spatiales de données non gaussiennes. En convertissant les valeurs continues en indicateurs binaires basés sur un seuil spécifique, l'IK facilite l'analyse de la probabilité qu'une certaine valeur soit dépassée à différents emplacements. Cette approche est avantageuse pour modéliser des phénomènes à seuils, comme la contamination, où l'intérêt réside dans la probabilité de dépassement d'un certain niveau de pollution. Le Kriging des valeurs indicatrices génère des cartes de probabilité, offrant une visualisation intuitive des risques sur un site.

Estimation de Densité par Noyau (KDE)
L'Estimation de Densité par Noyau est une technique non paramétrique avancée pour modéliser la fonction de densité de probabilité d'une variable aléatoire. Contrairement aux histogrammes traditionnels, le KDE applique une fonction de noyau lisse autour de chaque point de données, produisant une estimation continue de la densité. Cette méthode est particulièrement efficace pour identifier les structures spatiales des données, comme les points chauds de pollution, en ajustant la lissitude de l'estimation à travers le paramètre de bande passante. L'utilisation de noyaux gaussiens est commune, offrant un compromis entre précision et souplesse dans l'analyse de données spatiales complexes.

Estimation du Volume avec les Méthodes Classiques

L'estimation du volume de sol contaminé par des concentrations d'hydrocarbures totaux en pétrole (TPH) supérieures à un seuil de 5 000 mg/kg est cruciale pour évaluer l'impact environnemental et planifier les mesures de remédiation. Ce seuil, déterminé par les autorités environnementales en fonction des études de risques pour la santé et l'environnement, sert de référence pour l'utilisation future du site. Les techniques d'interpolation décrites antérieurement permettent d'analyser les variations du volume contaminé à travers différentes profondeurs, employant des blocs d'interpolation de 5 à 10 mètres de côté.

Outils de Configuration pour l'Interpolation
Afin d'assurer une méthodologie cohérente et reproductible pour les calculs de volume via différentes techniques d'interpolation, un ensemble de conditions fondamentales a été mis en place. L'utilisation des bibliothèques Python SKGstat et GSTools permet d'appliquer diverses formes de kriging, tandis que la bibliothèque Scikit-learn facilite l'estimation de densité par noyau (KDE) grâce à des structures de données efficaces comme les Ball Trees et KD Trees. Ces outils offrent une flexibilité pour l'optimisation des paramètres et l'exportation des résultats, garantissant des analyses précises et comparables des volumes contaminés.

Estimation et Ajustement du Variogramme
Le Variogram Cloud (VCloud) joue un rôle essentiel dans l'évaluation de la dissimilarité entre les paires d'échantillons, révélant l'influence des distances sur les variations de concentration. Cette visualisation permet d'identifier les comportements spécifiques du variogramme et de choisir le modèle le plus approprié pour les données. Malgré les défis posés par les distributions non normales, l'analyse comparative des modèles de variogrammes sur les données brutes et transformées (log+1) aide à déterminer le modèle offrant la meilleure

adéquation, soulignant l'importance de la transformation des données pour améliorer la précision de l'interpolation.

Surfaces et Estimation du Volume Pollué
L'estimation du volume des sols avec des valeurs TPH dépassant le seuil de 5 000 mg/kg utilise les techniques d'interpolation susmentionnées. Les calculs sont effectués pour des blocs de grille à différentes profondeurs, avec une attention particulière à la dimension et à la densité des blocs d'interpolation. Cette approche permet d'estimer avec précision le volume moyen de sol contaminé, reflétant la variabilité spatiale du TPH et fournissant une base solide pour les décisions de remédiation.

Les résultats montrent des variations significatives des estimations de volume entre les techniques d'interpolation, mettant en lumière l'impact du choix méthodologique sur l'évaluation de la contamination. La sélection du modèle sphérique, après une analyse comparative approfondie, élimine les problèmes potentiels de valeurs négatives et assure une estimation fiable du volume contaminé.

Ce processus d'estimation du volume souligne l'importance des techniques d'interpolation avancées dans la caractérisation précise des sites contaminés. En fournissant une évaluation détaillée de la distribution spatiale de la contamination, ces méthodes facilitent la planification des interventions de remédiation, permettant une gestion environnementale efficace et informée du site étudié.

Le Rôle des Variables de Jeu de Données et des Réseaux Neuronaux Artificiels dans l'Amélioration de l'Estimation de Volume

Les Variables Descriptives comme Indicateurs de Performance dans les Interpolations Spatiales
Diverses études ont examiné comment les variables descriptives au sein d'un jeu de données influencent la performance des interpolations spatiales. Ces études ont régulièrement conclu que des facteurs tels que le schéma de distribution des échantillons, le nombre d'échantillons, et la distribution des données influencent significativement l'exactitude des résultats de prédiction. Cependant, il est important de noter que ces variables ne sont pas uniformément faciles à mesurer et peuvent ne pas affecter de manière cohérente les résultats de l'interpolation dans tous les scénarios.

Une découverte clé est l'influence du coefficient de variation (CV) de la variable sur la performance de l'interpolation. Le CV, qui quantifie la variabilité relative au sein d'un jeu de données, est calculé comme le rapport de l'écart-type à la moyenne, souvent exprimé en pourcentage. La relation entre le CV et la performance de l'interpolation montre que lorsque le CV augmente (indiquant une plus grande variabilité des données), la performance des méthodes d'interpolation a tendance à se détériorer.

Exploiter les Réseaux Neuronaux Artificiels pour des Prédictions Améliorées
Les Réseaux Neuronaux Artificiels (RNA), en particulier les Réseaux Neuronaux Profonds (RNP), ont gagné une traction significative dans divers domaines en raison de leur polyvalence et de leur précision prédictive. Les RNP ont la capacité de gérer divers types de données, y

compris des variables quantitatives et catégorielles. Ils sont particulièrement précieux dans les scénarios où de grands ensembles de données sont disponibles, car ils peuvent exploiter ces données étendues pour produire des prédictions fiables.

Les RNP sont des modèles informatiques inspirés de la structure et du fonctionnement du cerveau humain. Ils se composent de neurones artificiels interconnectés, organisés en couches. Les composants clés d'un RNP comprennent les couches d'entrée, les couches cachées, et les couches de sortie. L'information circule à travers le réseau lors d'une passe avant, où chaque neurone reçoit des entrées, effectue des calculs, et génère des sorties. Les connexions entre les neurones, représentées par des poids, déterminent l'influence de chaque entrée sur la sortie. L'application de fonctions non linéaires à ces valeurs de sortie améliore la capacité du réseau à capturer des motifs de données complexes.

Une catégorie notable de RNA est le modèle de Réseaux Neuronaux Profonds (RNP), qui comprend plusieurs couches cachées entre les couches d'entrée et de sortie. Cette architecture permet aux RNP d'apprendre des représentations hiérarchiques des données d'entrée, capturant à la fois les caractéristiques de bas niveau et de haut niveau. La passe avant implique la transformation des données d'entrée à travers les couches à l'aide de connexions pondérées et de fonctions d'activation. Les RNP apprennent en ajustant ces poids lors de l'entraînement, ce qui leur permet de généraliser et de faire des prédictions sur de nouvelles données non vues.

L'efficacité des RNP est attribuée à leur capacité à modéliser efficacement les relations complexes des données. Leur capacité à apprendre à partir d'ensembles de données étendus et diversifiés les rend adaptés à la capture de motifs de données complexes et à la production de prédictions précises. Cette nature gourmande en données, tout en posant des défis, permet aux RNP de se démarquer dans divers domaines, y compris l'interpolation spatiale.

Les applications récentes des RNP en combinaison avec les méthodes d'interpolation spatiale traditionnelles ont montré des résultats prometteurs, en particulier dans les études environnementales telles que l'évaluation de la qualité de l'air. Ces applications exploitent la puissance des RNP pour analyser de multiples variables avec des corrélations non linéaires, offrant ainsi des résultats de prédiction améliorés.

Méthodologie Evol : Estimation de l'Incertitude avec des Caractéristiques Descriptives et l'Utilisation de Réseaux Neuronaux Profonds

De nombreuses études de recherche ont examiné l'efficacité des techniques d'interpolation par rapport aux caractéristiques descriptives des ensembles de données et à l'utilisation de réseaux neuronaux profonds pour améliorer les méthodes d'interpolation. Cependant, aucune étude antérieure n'a spécifiquement ciblé l'estimation du volume contaminé concernant le réseau d'échantillonnage.

Proposition de la Méthodologie Evol

Nous proposons une méthodologie appelée "Evol" pour estimer les volumes de sol contaminé avec des plages d'incertitude. Cette approche combine des techniques classiques d'interpolation spatiale non paramétrique avec l'apprentissage profond. La méthodologie consiste à mesurer les variables descriptives (caractéristiques) dans un ensemble

d'échantillons pour estimer l'erreur dans le calcul du volume de sol contaminé à l'aide d'un modèle de réseau neuronal profond.

Modèle Synthétique
Un modèle synthétique basé sur une simulation stochastique a été créé en utilisant la bibliothèque Python GSTools. Ce modèle reproduit la distribution de la variable principale de pollution du site, qui est la concentration d'hydrocarbures pétroliers totaux C10-C40 en mg/kg. La simulation était basée sur le modèle de covariance de Webster, en utilisant le variogramme expérimental de la variable du site transformée en logarithme.

Échantillonnage
Des ensembles d'échantillons ont été créés en subdivisant la zone du modèle synthétique en cellules de 10 m x 10 m, à partir desquelles un échantillon de 100 cellules est prélevé selon quatre stratégies d'échantillonnage : Systématique Régulier, Aléatoire, Cluster (base régulière) et Cluster (base aléatoire). Au total, 400 ensembles de données d'échantillons différents, chacun contenant 100 échantillons, ont été prélevés à partir du modèle synthétique.

Estimation du Volume et Erreur Associée
Des techniques d'interpolation, l'Interpolation Inverse (IK) et l'Estimation de la Densité du Noyau (KDE), ont été utilisées pour effectuer des prédictions à partir de chacun des 400 ensembles de données d'échantillons. Le calcul du volume a été effectué en additionnant le nombre de blocs de 1 m^3 interpolés dépassant le seuil de 5000 mg/kg. Un volume a été calculé pour chaque technique, et un Volume le Plus Probable (MPV) a été calculé comme la valeur moyenne dans chaque bloc.

Description des Caractéristiques
De nombreuses caractéristiques ont été extraites et divisées en deux groupes : les caractéristiques de l'ensemble d'échantillons et les caractéristiques d'analyse des prédictions. Les caractéristiques de l'ensemble d'échantillons comprenaient des statistiques descriptives de base ainsi qu'une analyse spatiale, tandis que les caractéristiques d'analyse des prédictions étaient liées aux prédictions faites par les techniques d'interpolation.

Sélection et Extraction des Caractéristiques
Les caractéristiques ayant une corrélation supérieure à 0,25 avec la valeur cible, ont été sélectionnées pour le modèle de réseau neuronal profond (DNN). Un modèle de régression à entrées multiples utilisant un DNN a été construit pour établir la relation entre les caractéristiques sélectionnées parmi les 400 ensembles de données d'échantillons et la prédite.

Structure du Modèle DNN
Le modèle DNN se compose d'une couche d'entrée avec 21 caractéristiques normalisées, de trois couches cachées denses avec 32 et 64 neurones chacune, et d'une couche de sortie avec une forme de 2 correspondant à l’erreur prédite. Le modèle a été entraîné en utilisant l'algorithme d'optimisation Adam et l'erreur absolue moyenne (MAE) comme fonction de perte.

Résultats

Variabilité du Volume dans les Ensembles d'Échantillons

L'analyse de cette section se penche sur les subtilités de la variabilité du volume au sein de différents ensembles d'échantillons générés à partir du modèle synthétique. Il est important de noter que le degré de variabilité dans les volumes prédits est influencé par plusieurs facteurs critiques. Ces facteurs comprennent la qualité globale de l'ensemble d'échantillons, les valeurs spécifiques des points de données individuels, leur distribution statistique et leur arrangement spatial au sein de l'ensemble de données. Un point clé de cette analyse est que plus l'erreur dans la prédiction du volume augmente, plus la gamme des volumes calculés obtenus par différentes techniques de prédiction s'élargit significativement. Comprendre cette variabilité est crucial pour obtenir des estimations fiables du volume dans l'évaluation et la gestion des sites contaminés.

Techniques de Prédiction et Leurs Implications

Diverses techniques de prédiction ont été appliquées aux ensembles d'échantillons synthétiques pour estimer les volumes. Il est à noter que les techniques de Kriging Inverse (IK) et d'Estimation de la Densité Noyau (KDE) ont montré leur efficacité dans l'estimation des volumes, en particulier pour les ensembles de données avec un faible. Ces techniques ont démontré des prédictions de volume relativement stables et cohérentes. Cependant, à mesure que l'erreur dans la prédiction du volume augmentait, ces méthodes présentaient également une plage plus large d'estimations de volume, mettant en évidence la sensibilité de la précision de la prédiction à la qualité des données en entrée.

En revanche, différentes stratégies d'échantillonnage ont été explorées dans l'analyse, chacune ayant ses propres caractéristiques. Les stratégies d'échantillonnage régulières, caractérisées par leur répartition uniforme des points de données, ont montré de la stabilité et de la cohérence dans les volumes prédits, même lorsque l'erreur était plus élevée. Ces stratégies ont montré une gamme plus large de résultats mais avec moins de valeurs aberrantes. Les stratégies d'échantillonnage aléatoires ont donné une plage plus étroite d'estimations de volume mais avec une fréquence plus élevée de valeurs aberrantes. Cette observation met en évidence les compromis entre les différentes stratégies d'échantillonnage dans le contexte de l'estimation du volume, en tenant compte à la fois de la précision et de la robustesse.

Sélection des Caractéristiques pour une Précision Améliorée

Pour améliorer la précision des prédictions de volume, l'analyse a impliqué la sélection des caractéristiques qui présentaient une forte corrélation linéaire avec l'erreur de volume. Les coefficients de corrélation de Pearson (CCP) ont été calculés pour chaque caractéristique, et celles avec des valeurs de CCP dépassant 0,25 ont été sélectionnées pour être incluses dans le modèle prédictif. Ce seuil a été soigneusement choisi pour donner la priorité aux caractéristiques qui présentaient une corrélation linéaire modérée avec la variable cible.

Le Rôle des Réseaux de Neurones Profonds (DNN)

Le modèle prédictif utilisé dans cette analyse a été construit en utilisant une architecture de Réseaux de Neurones Profonds (DNN). Le modèle DNN était composé de couches de normalisation, de couches denses et d'une couche de sortie. Au cours du processus d'entraînement, le modèle a montré des performances remarquables, atteignant une

précision de 91 % sur l'ensemble d'entraînement et de 89 % sur l'ensemble de validation. Pour éviter le surajustement, deux stratégies essentielles ont été mises en œuvre. Premièrement, le nombre de paramètres du modèle a été réduit en fonction des valeurs de CCP des caractéristiques sélectionnées, en donnant la priorité aux caractéristiques qui présentaient une corrélation linéaire plus forte avec l'erreur. Deuxièmement, un mécanisme d'arrêt anticipé a été introduit pendant l'entraînement pour garantir que le modèle généralise bien sur les données non vues. Cette combinaison de stratégies a contribué à la robustesse et à la fiabilité du modèle dans la prédiction des volumes.

Test sur de Nouvelles Données

Le véritable test de l'efficacité du modèle prédictif est venu lorsque celui-ci a été appliqué à 79 nouveaux échantillons du modèle synthétique. Les résultats étaient prometteurs, le modèle montrant une forte corrélation entre les volumes prédits et les valeurs mesurées, comme l'indique un coefficient de détermination (r^2) de 0,9882. Cette valeur élevée souligne la capacité du modèle à prédire avec précision les volumes pour de nouvelles données, renforçant ainsi la confiance dans son applicabilité pratique.

Performance et Estimation de l'Incertitude

Un aspect crucial de l'analyse a consisté à comparer la méthodologie Evol avec la Simulation Conditionnelle en termes de performance et d'estimation de l'erreur de volume pour l'estimation du volume des sites contaminés. Les deux méthodes ont été utilisées pour calculer le Volume le Plus Probable (MPV) et un Intervalle de Confiance à 90 % (CI) pour chaque échantillon.

Dans l'analyse des données synthétiques, Evol a constamment surpassé la Simulation Conditionnelle en présentant des tailles de CI plus petites. Cela indique que la méthodologie Evol a atteint une plus grande précision dans l'estimation du volume par rapport à la Simulation Conditionnelle, en particulier dans les cas avec un faible. Cependant, il est important de noter que la performance d'Evol variait en fonction des caractéristiques des ensembles d'échantillons. Pour les stratégies d'échantillonnage régulières et aléatoires, Evol a démontré une précision et une cohérence supérieures. En revanche, pour l'échantillonnage regroupé, Evol a montré une précision légèrement inférieure, mais a quand même fourni des résultats précieux.

Comparaison de la Taille de l'intervalle d'Incertitude

Un examen détaillé des tailles de CI a révélé un schéma convaincant. Evol a constamment réduit l'incertitude dans les estimations de volume par rapport à la Simulation Conditionnelle. Les réductions de la taille du CI allaient de 43 % à 80 % pour les échantillons synthétiques, mettant en évidence l'amélioration substantielle de la précision obtenue par Evol. Il est à noter que la réduction de l'incertitude était une tendance constante, à l'exception d'un échantillon (S4) qui a connu une augmentation de la taille du CI. Cette observation souligne le potentiel de la méthodologie Evol pour améliorer considérablement la fiabilité et la précision des estimations de volume, en particulier dans les scénarios où la réduction de l'incertitude est cruciale.

Estimation Complète du Volume du Site

L'application du modèle DNN et de la méthodologie Evol a permis d'estimer le volume de la couche de sol pollué couche par couche sur le site d'étude. Bien qu'Evol ait rencontré des défis pour générer des calculs d'incertitude pour les couches initiales, il a efficacement réduit l'incertitude dans les couches plus profondes par rapport à la Simulation Conditionnelle. Cette amélioration de la réduction de l'incertitude est un résultat prometteur pour une estimation précise et détaillée du volume des sites contaminés, ce qui est essentiel pour la remédiation efficace du site et la gestion de l'environnement.

Discussion

Deux facteurs principaux font l'objet d'un examen attentif : la précision des points d'échantillonnage et la conception de l'ensemble d'échantillons. La précision de l'échantillonnage fait référence à la précision et à la justesse des échantillons dans la représentation des caractéristiques ciblées du site. Il est essentiel que les échantillons soient non seulement représentatifs mais également collectés aux emplacements corrects. Pendant ce temps, la conception de l'ensemble d'échantillons concerne l'arrangement stratégique et la distribution de ces échantillons sur le terrain, englobant le modèle global et la méthodologie d'échantillonnage pour garantir une représentation exhaustive et impartiale de la zone ou du phénomène étudié.

Conception et Précision de l'Échantillonnage

La précision des échantillons concerne la précision avec laquelle ils reflètent la pollution réelle à chaque point d'échantillonnage et leur caractère représentatif. Nous découvrons que les ensembles d'échantillons regroupés, bien que bénéfiques pour les méthodes géostatistiques classiques selon fournissent moins d'erreurs et d'intervalles de confiance plus étroits pour la méthodologie Evol.

Caractéristiques Sélectionnées

Traditionnellement, le Coefficient de Variation (CV) et la Variance sont utilisés pour estimer l'incertitude. Cependant, notre approche repose sur une sélection de caractéristiques qui influencent l'erreur sur le volume, sans qu'il y ait une relation linéaire entre elles. Par conséquent, une combinaison de plusieurs caractéristiques analysées par un Réseau de Neurones Artificiels (RNA) est le choix le plus précis, garantissant une grande précision dans la prédiction de l'erreur de volume.

Applicabilité et Limitations

L'applicabilité de la méthodologie Evol s'étend à la plupart des sites contaminés, grâce à sa généricité et à sa capacité à traiter des distributions statistiques non standard. Cependant, il est essentiel que le modèle synthétique représente au mieux l'hétérogénéité du site. La taille de l'ensemble de données en entrée est cruciale, car les petits ensembles de données n'ont pas suffisamment d'informations pour construire le modèle RNA avec précision. Les ensembles de données plus volumineux augmentent la précision de l'estimation, mais peuvent nécessiter des ressources de calcul spécifiques.

Conclusions

En conclusion, la méthodologie Evol combine avec succès des techniques d'interpolation spatiale classiques avec l'apprentissage en profondeur pour estimer les volumes de sol contaminé avec une grande précision. Elle réduit considérablement les erreurs d'estimation

de volume et les intervalles de confiance. Cette méthodologie peut être intégrée à des méthodes géostatistiques traditionnelles pour guider la sélection de nouvelles zones d'échantillonnage. Cependant, il est essentiel de valider cette approche sur différents sites. En fin de compte, Evol montre que l'apprentissage en profondeur peut compléter les méthodes classiques sans les remplacer, ouvrant ainsi de nouvelles perspectives pour l'estimation des volumes de sol contaminé.

Chapitre 4- Estimation du volume d'huile

Vue d'ensemble théorique sur l'estimation de la saturation et du volume d'huile

Cette section offre un examen approfondi du comportement de l'huile au sein des milieux poreux, soulignant les facteurs critiques qui influencent la saturation et l'estimation du volume d'huile dans les environnements souterrains. Elle plonge dans la progression des modèles simplistes initiaux vers des représentations plus complexes et précises qui tiennent compte de la nature hétérogène des formations souterraines.

Distribution de la Saturation en Huile

Initialement, la migration de l'huile dans les milieux poreux était conceptualisée en utilisant le modèle de pancake, qui supposait une distribution uniforme de l'huile flottant au-dessus des eaux souterraines. Cependant, ce modèle ne tenait pas compte de la variation de taille des grains et de la porosité trouvées dans les formations souterraines naturelles, conduisant à des représentations inexactes de la distribution de l'huile. Les recherches, notamment par Lenhard (1990), ont mis en lumière l'importance de considérer la tension de surface, la mouillabilité et la pression capillaire pour comprendre la distribution de l'huile. Ces facteurs, dictés par les interactions entre l'huile, l'eau et les particules du sol, déterminent le comportement de l'huile dans la matrice poreuse, influençant son mouvement et son accumulation.

La tension de surface et la pression capillaire sont particulièrement significatives dans la définition des modèles de distribution de l'huile. La pression capillaire, résultant de la différence de pressions entre deux phases fluides dans les pores, dicte le mouvement et la saturation de l'huile. Ce concept est crucial pour comprendre comment l'huile est retenue ou déplacée dans le sol, avec le modèle de Van Genuchten étant largement utilisé pour l'analyse quantitative dans les investigations environnementales.

Saturation Résiduelle en Huile

La saturation résiduelle en huile se réfère au volume d'huile restant dans le milieu poreux après un processus de déplacement, tel que l'inondation par l'eau. Ce concept est crucial pour la réhabilitation environnementale car il indique l'étendue de l'huile qui peut être naturellement retenue dans le sol. Cependant, il est noté que les conditions réelles sur le terrain révèlent souvent des niveaux de saturation en huile inférieurs à ceux déterminés dans des conditions de laboratoire, soulignant l'écart entre les modèles théoriques et les scénarios réels. Cet écart souligne le besoin de mesures de terrain précises et de modèles capables de s'adapter aux complexités des systèmes naturels.

Distribution de l'Huile dans les Puits de Surveillance

Les puits de surveillance jouent un rôle vital dans l'identification et la quantification de la présence d'huile dans les environnements souterrains. En mesurant les niveaux d'huile et d'eau dans ces puits, les scientifiques peuvent déduire les états de saturation et les modèles de distribution de l'huile dans le sol environnant. Les conditions d'équilibre dans le puits indiquent l'équilibre entre les niveaux de fluide dans le puits et la formation adjacente, fournissant des aperçus de la saturation en huile souterraine et de l'efficacité des stratégies de remédiation potentielles.

Distributions de la Saturation en Huile
Comprendre la distribution de la saturation en huile au sein des milieux poreux nécessite une connaissance détaillée des pressions capillaires en jeu entre les phases d'huile, d'eau et d'air. Le modèle de Van Genuchten offre un cadre pour estimer la saturation effective de ces phases, en tenant compte de la porosité du sol et des tensions interfaciales des fluides. Ces calculs sont essentiels pour prédire comment l'huile se comporte sous différentes conditions souterraines, facilitant le développement de techniques de récupération d'huile et de remédiation environnementale plus efficaces.

Saturation Résiduelle en Huile à Partir des Puits de Surveillance
Les mesures de terrain, telles que l'épaisseur des couches d'huile dans les puits de surveillance, fournissent des données essentielles pour estimer la saturation résiduelle en huile. Cette information, couplée avec des modèles théoriques et des courbes de pression capillaire, permet une compréhension nuancée des modèles de distribution de l'huile, qui est inestimable pour les évaluations environnementales et la détermination de l'efficacité des efforts de remédiation.

Méthodologie EMob : Estimation de l'huile sur un grand site hétérogène

Les approches précédemment développées considèrent généralement un sol homogène, mais à l'échelle du terrain, la présence d'hétérogénéité introduit des erreurs significatives, posant un défi considérable dans l'estimation précise du volume d'huile en raison des saturations variables et des flux préférentiels. Dans les applications pratiques, les modèles présentés nécessitent souvent des données étendues, qui ne sont pas toujours disponibles, conduisant à l'utilisation de paramètres issus de la littérature. Par conséquent, estimer le volume d'huile devient très difficile en raison du grand nombre de variables inconnues. Selon l'expérience des experts, la plupart des modèles populaires ont tendance à sous-estimer le volume d'huile résiduelle, entraînant une surestimation de l'huile récupérable.

Aucune étude n'a été menée pour estimer le facteur f au niveau du champ pour la détermination de l'huile totale et résiduelle. Pour combler cette lacune, nous proposons une méthodologie nommée "EMob", qui utilise les mesures de terrain pour évaluer la relation entre l'huile totale et résiduelle et estimer le facteur f. Notre approche est mise en œuvre dans un scénario réel caractérisé par une hétérogénéité du sol significative et des dimensions considérables, dépassant les volumes typiquement rencontrés dans les processus de remédiation standard. La méthode présentée est ensuite comparée aux modèles classiques pour estimer le contenu en huile libre et résiduelle à l'échelle du site. Grâce à un suivi très détaillé du processus de remédiation, les volumes estimés peuvent ensuite être comparés aux valeurs obtenues après excavation.

Transfert de saturation en huile

Le site de recherche, présenté au Chapitre 2, subit des procédures de remédiation depuis 2010. Le processus de remédiation a été caractérisé par trois phases principales. Initialement, la première étape impliquait la mise en œuvre de techniques de pompage et de traitement, activement employées pendant environ 10 ans tandis que le site était opérationnel. La deuxième étape impliquait l'excavation du sol contaminé, suivie par le retrait de l'huile résiduelle par des méthodes d'écrémage. Enfin, le volume entier de sol extrait a subi un traitement sur site utilisant deux techniques principales : le traitement biologique et le lavage du sol, suivi par un traitement thermique.

Acquisition de nouvelles données

Bien que des données antérieures sur la contamination du sol et de l'huile dans la nappe phréatique étaient disponibles, les points de mesure dans les coordonnées géographiques différaient. Pour assurer des mesures précises aux mêmes emplacements, une décision a été prise de collecter de nouvelles données du terrain.

L'acquisition de nouvelles données a eu lieu pendant le processus de remédiation en cours et a été divisée en deux parties principales. D'abord, des échantillons de sol ont été collectés, et les caractéristiques physiques du sol, telles que la porosité, la densité apparente, et le pourcentage de particules fines et grossières, ont été déterminées. Ensuite, des mesures de l'épaisseur de l'huile et de la concentration totale en hydrocarbures dans la zone saturée ont été effectuées.

Au total, 35 puits ont été installés dans la zone vadose à la limite entre les zones non saturées et saturées. Ces puits ont été manuellement installés à l'aide d'une tarière hélicoïdale, et le sol extrait a été utilisé pour créer des échantillons composites pour l'analyse chimique des hydrocarbures totaux pétroliers (C10-C40) et leurs fractions.

Estimation du facteur (f') de terrain (fr')

Suivant la ligne de pensée d'Adamski (2005) et Charbeneau (2007), l'idée était d'obtenir un facteur (f) de terrain, appelé (fr'), comme estimation pour chaque type de sol en supposant une relation linéaire entre (So) et (Sor). La méthodologie repose sur deux hypothèses : la position du haut et du bas de l'huile arrivée au puits en 24h représente la position des interfaces eau-huile et air dans la formation adjacente en quasi-équilibre et a été utilisée pour calculer la saturation en huile (So). Un modèle de régression linéaire a été utilisé pour estimer le facteur (f) avec les saturations de terrain calculées pour chaque type de sol. Par conséquent, la corrélation entre (So) et (Sor) pour chaque type de sol sur le terrain aboutira à (fr'). (fr') sera déterminé à partir de la pente du modèle de régression linéaire à partir de l'équation suivante :

$$Sor' = fr' \, So + E$$

Où So est la saturation en huile calculée,Sor' est la saturation résiduelle relative en huile à partir de la mesure de TPH sur site, fr' est le facteur f de terrain, et E est l'erreur.

Résultats

Paramètres physiques du sol

La porosité, la densité apparente et le pourcentage de particules fines et grossières ont été soigneusement déterminés pour chaque sol à partir des 35 profils de terrain. Les données complètes sur les propriétés du sol ont été résumées, fournissant des informations précieuses sur quatre types de sol distincts rencontrés sur le site, allant du Gravier/Sable grossier au plus fin, le Limon.

Ces données présentent les valeurs moyennes et les écarts-types de la porosité, de la densité apparente et du pourcentage de particules fines pour chaque type de sol rencontré sur le site. L'analyse des données montre une gamme de porosité de 0.21 à 0.44, tandis que la densité apparente varie de 1.45 à 1.89 kg/dm³. La fraction de particules fines varie également significativement, de 0.06 à 0.39. Certains types de sol affichent des valeurs de porosité similaires, le Limon et le Loam sableux présentant la porosité la plus élevée.

Pour explorer les relations potentielles entre ces propriétés du sol, une série d'analyses de corrélation a été menée. Les résultats montrent une relation entre la composition du sol des mini-piezomètres et les caractéristiques mesurées, avec une reclassification des types de sol due à l'hétérogénéité verticale.

Épaisseur de l'huile

Suite à la classification des types de puits, les tests de rabattement ont été regroupés de la même manière. Les résultats de tous les tests de rabattement montrent que chaque type se comporte d'une manière particulière, atteignant un plateau d'épaisseur à environ 24 heures. Comme prévu, les sols plus grossiers ont montré une épaisseur d'huile libre plus importante que les sols plus fins.

Estimation du facteur (f')

Sur la base des résultats des tests de rabattement, une comparaison a été faite entre l'épaisseur de l'huile à la marque des 24 heures et les concentrations d'Hydrocarbures Pétroliers Totaux (HPT) dans le sol. Un modèle linéaire a été créé pour chaque type de sol, établissant une relation entre les valeurs calculées de (Sor') et (So) à partir des données de terrain. Les différents modèles de régression pour les types de sol montrent que le facteur (f') calculé varie selon la porosité du sol et le nombre de particules fines présentes.

Estimation du volume d'huile

Après avoir réalisé le calcul du facteur (f), nous avons comparé les résultats des deux modèles classiques d'estimation du volume d'huile et les résultats estimés par type de sol avec le modèle modifié. L'application des modèles classiques a révélé des différences substantielles

dans les volumes estimés de liquide non aqueux de phase légère libre. Le modèle développé avec le facteur (fr) modifié présente l'estimation de volume la plus proche de la valeur réelle récupérée sur le terrain.

Comparaison entre les modèles et les données du site

Durant le projet de remédiation, l'huile a été récupérée en deux étapes. La première étape impliquait l'extraction d'huile des puits de récupération avant la démolition du bâtiment de fabrication. La seconde étape a utilisé un système de récupération d'huile par écrémage après la phase d'excavation. En total, 953 m^3 d'huile ont été récupérés.

L'analyse des résultats du terrain et des modèles employés montre que tous les modèles surestiment la quantité d'huile présente sur le site. Le modèle développé avec le paramètre (f'r) modifié présente l'estimation de volume la plus proche de la valeur réelle récupérée sur le terrain, indiquant une approche plus précise pour l'estimation du volume d'huile dans des conditions de terrain hétérogènes.

Discussion

Disparités du Volume d'Huile

Malgré l'ajustement des valeurs pour l'huile libre et résiduelle en utilisant le facteur (fr) de terrain, une comparaison avec les valeurs in situ révèle que le volume estimé par le modèle ajusté s'écarte toujours significativement des valeurs récupérées sur le site. Cela suggère que certaines zones de sol avec des valeurs basses de (Sor) pourraient connaître une augmentation de (Sor) à travers les interactions avec des zones de sol adjacentes exhibant une haute mobilité d'huile.

Les calculs approfondis démontrent une présence significative d'huile dans les biopiles et les zones de traitement thermique, où aucune huile libre n'a été observée malgré des concentrations de TPH atteignant 100,000 mg/kg. Ceci renforce l'hypothèse que les sols avec un contenu plus élevé de limon possèdent une plus grande capacité à retenir l'huile, menant à des fractions résiduelles considérablement plus élevées comparées aux sols sableux.

Utilisation de l'Huile Mesurée à Partir des Puits

L'évaluation directe de l'huile à partir des puits, sans considérer l'influence des propriétés de formation adjacentes, est reconnue comme susceptible d'erreurs. Les mesures d'épaisseur obtenues à partir des puits qui ont été fermés pendant une période prolongée ne devraient pas être utilisées pour estimer les volumes, même si ces modèles incorporent les caractéristiques du sol. Les tests de rabattement effectués pour évaluer l'épaisseur de l'huile sur une plage de temps spécifiée, en particulier 24 heures, permettent d'observer la stabilisation de l'épaisseur.

Niveau de la Nappe Phréatique et Volume Calculé

Le volume d'huile est influencé par le niveau de la nappe phréatique, et bien que la saturation totale reste constante, les quantités d'huile libre et résiduelle changent en fonction de la table d'eau. Cela implique que le volume d'extraction requis variera selon le niveau de la nappe phréatique.

Formes de So, Sor et Som

Sur la base des études de terrain du site et de l'implémentation de modèles mathématiques, nous avons validé ce qu'Adamski (2005) a démontré en laboratoire : il existe une relation linéaire entre la saturation résiduelle en huile (Sor) et la saturation totale en huile (So), située entre (Zoa) et (Zow). Cependant, au-delà de cette plage, il existe une portion de saturation résiduelle qui n'est pas liée à la présence de phase libre d'huile mesurée dans des conditions de quasi-équilibre.

Dans les deux cas, le volume libre d'huile dans le sol est le même, mais la concentration moyenne en TPH est différente, en relation avec la valeur différente de (Sor). Cette distinction joue un rôle crucial dans l'estimation précise du volume total d'huile récupérable, car elle peut soit surestimer soit sous-estimer ce volume.

Cette discussion met en lumière l'importance de considérer la dynamique complexe de l'huile dans les sols hétérogènes pour une estimation précise du volume d'huile récupérable. Elle souligne également les défis associés à l'utilisation des données de puits et les implications du niveau de la nappe phréatique sur les volumes calculés d'huile.

Conclusion

La méthodologie "EMob", basée sur des mesures de terrain, améliore la précision de l'estimation du volume d'huile en évaluant la relation entre l'huile totale et résiduelle pour estimer le facteur (f). L'étude des paramètres physiques du sol sur 35 profils a mis en évidence d'importantes variations entre les différents types de sols.

Les analyses de corrélation entre (So) et (Sor) ont permis de développer des modèles linéaires prédisant le facteur (f) pour chaque type de sol, révélant que la porosité et les particules fines influencent la rétention d'huile. Les sols plus grossiers présentent une mobilité d'huile plus élevée et une saturation résiduelle plus faible, tandis que les sols limoneux montrent une saturation résiduelle plus élevée.

Les comparaisons entre les volumes d'huile estimés et les données de terrain indiquent que tous les modèles surestiment la quantité d'huile présente, avec le modèle "EMob" fournissant l'estimation la plus proche de la réalité.

Le niveau de la nappe phréatique s'est avéré être un facteur significatif, influençant le volume d'huile, avec des variations des quantités d'huile libre et résiduelle en fonction de la table d'eau.

L'étude a également identifié l'augmentation de la saturation résiduelle due au mélange pendant l'excavation et a souligné que les sols limoneux ont une plus grande capacité à retenir l'huile, entraînant des fractions résiduelles plus élevées que dans les sols sableux.

Chapter 5- Analyse et prévision des flux du projet de réhabilitation

Vue d'Ensemble Théorique sur l'Incertitude des Résultats Analytiques, en Laboratoire et Échantillonnage

Ellison et al. (2012) et l'ISO (1995) détaillent la démarche pour estimer l'incertitude des mesures analytiques, soulignant l'importance de spécifier l'analyte, la méthode d'obtention des résultats, et les sources d'incertitude. L'incertitude combinée et l'incertitude élargie sont calculées pour refléter la fiabilité des résultats analytiques. Dans les analyses environnementales, l'échantillonnage, le prétraitement en laboratoire, et le processus analytique lui-même sont identifiés comme principales sources d'incertitude.

Incertitude et Biais de Laboratoire

L'incertitude en laboratoire englobe toutes les incertitudes liées aux étapes analytiques, depuis l'arrivée de l'échantillon jusqu'à la génération du rapport final. Cette incertitude se manifeste à plusieurs niveaux :

1. Erreur Méthodologique : Écart systématique introduit par la méthode analytique choisie.
2. Erreur de Laboratoire : Erreur systématique propre à chaque laboratoire, qui peut être due aux spécificités des instruments ou des procédures du laboratoire.
3. Variation Jour après Jour : Erreur aléatoire se produisant entre des déterminations répétées effectuées différents jours sur une période prolongée.
4. Répétabilité : Erreur aléatoire entre des déterminations répétées dans un court intervalle de temps, où l'inhomogénéité des échantillons peut jouer un rôle.

Sous-échantillonnage et Variabilité

Le sous-échantillonnage soulève la question de la représentativité lorsqu'un petit échantillon est prélevé d'un lot plus grand. La taille de l'échantillon et la méthode de réduction de l'échantillon sont essentielles pour garantir que l'échantillon de laboratoire reflète de manière fiable les caractéristiques du lot plus large.

Distribution des Données sous le Limite de Détection

Les données en dessous de la limite de détection posent des défis statistiques. Les valeurs sous cette limite sont généralement indiquées comme non détectées et doivent être traitées avec soin pour éviter de fausser l'analyse. Les méthodes modernes MLE, les techniques d'imputation et la méthode Kaplan-Meier sont utilisées pour traiter ces données.

Incertitude d'Échantillonnage

L'incertitude d'échantillonnage dans les études environnementales provient de la variabilité naturelle, de la sélection des sites d'échantillonnage et de la profondeur des échantillons. Le processus de passage d'un échantillon d'un mètre cube à un pot de 300 grammes pour l'analyse de laboratoire est cruciale et source potentielle d'incertitude.

Description du Projet de Remédiation

Le but du projet de remédiation est de traiter les sols contaminés par des hydrocarbures au-delà d'un seuil spécifié et d'éliminer tout LNAPL excédant 1 cm. La stratégie de remédiation repose sur le traitement in-situ, visant à minimiser le besoin de traitement hors site. Le projet se divise en deux phases principales, avec une attention particulière sur l'excavation des sols contaminés et leur traitement sur place.

Excavation et Rééchantillonnage

L'excavation a impliqué le retrait et la classification des sols selon leur niveau de contamination. Un processus détaillé de rééchantillonnage et d'analyse a été mis en œuvre pour confirmer les catégories de contamination et orienter les sols vers des traitements appropriés, en fonction de leur concentration en hydrocarbures.

Validation des Quantités de Sol Pollué : Planifié vs Excavé

Une comparaison des concentrations d'Hydrocarbures Pétroliers Totaux (TPH) avant et après l'excavation révèle des changements significatifs dans les niveaux de contamination. La proportion de sols 'propres' a diminué, tandis que le volume de sol nécessitant un traitement biologique a augmenté de manière significative, soulignant les répercussions de l'excavation sur les niveaux de contamination et l'ajustement nécessaire des stratégies de remédiation.

Méthodologie de prévision des volumes catégorisés

Méthodologie de prévision des volumes catégorisés

Cette section se penche sur l'estimation de l'incertitude des stocks et des laboratoires, ainsi que sur l'impact de l'hétérogénéité des piles de sol sur l'échantillonnage composite aléatoire. De plus, la recherche aborde l'estimation de l'incertitude dans les deux scénarios, en s'appuyant sur des statistiques descriptives pour étayer les conclusions.

Pour réaliser cette analyse, des piles de sol représentatives ont été sélectionnées en fonction de critères garantissant leur représentativité de la population cible. Cette approche a permis une évaluation complète de la variabilité et de l'incertitude associées à l'échantillonnage et à l'analyse des sols, mettant en évidence les effets de la technique de l'opérateur et du traitement en laboratoire sur la cohérence des données sur les sols.

Analyse de l'hétérogénéité de l'échantillonnage composite aléatoire

La première analyse visait à évaluer l'hétérogénéité des échantillons de sol collectés par deux opérateurs distincts, dénommés Opérateur A et Opérateur B, à partir du même stock de 200

m3. L'objectif était d'investiguer la variabilité de la composition des échantillons en utilisant une approche classique basée sur le calcul de la moyenne et de l'écart-type pour quantifier la variabilité.

Pour chaque stock sélectionné, spécifiquement de 200 m3, tant l'Opérateur A que l'Opérateur B ont collecté indépendamment un échantillon composite. Un échantillon composite, dans ce contexte, consiste à agréger le sol à partir de cinq points prédéterminés dans le stock pour former un seul échantillon représentant les caractéristiques globales de la pile entière. Le processus a été soigneusement conçu pour garantir que la méthode d'échantillonnage de chaque opérateur était cohérente, dans le but de fournir une comparaison claire de l'hétérogénéité attribuable à la technique d'échantillonnage de chaque opérateur. Les deux échantillons ont été envoyés le même jour, au même moment, au même laboratoire.

Pour le calcul de la variabilité, nous avons d'abord calculé la moyenne de chaque paire d'échantillons OpA et OpB. De même, nous avons calculé les écarts-types de chaque paire d'échantillons à l'aide des formules décrites ci-dessous.

Avec la variabilité calculée, nous calculons l'incertitude de mesure U en utilisant un facteur de couverture k et l'écart des échantillons, ainsi que l'incertitude relative en pourcentage U%.

Introduction d'un biais dû au changement de laboratoire

En plus de comparer l'hétérogénéité des piles avec l'échantillonnage composite, cette étude explore également le biais introduit par l'analyse en laboratoire. Concrètement, des échantillons collectés par un seul opérateur ont été divisés en deux et envoyés à deux laboratoires distincts, Lab D et Lab E, pour analyse. Cette approche visait à identifier d'éventuelles différences significatives dans les résultats de composition du sol qui pourraient être attribuées à l'environnement du laboratoire plutôt qu'à l'hétérogénéité et au processus d'échantillonnage.

Un échantillon composite unique, collecté par l'un des opérateurs, a été homogénéisé et divisé en deux parties égales. Ces parties ont ensuite été envoyées simultanément à Lab D et Lab E pour analyse le même jour, au même moment, avec les mêmes conditions d'emballage et de conservation.

Pour l'analyse de la variabilité, la valeur moyenne pour un paramètre donné a été calculée comme la moyenne des résultats de Lab D et Lab E.

Avec la variabilité calculée, l'incertitude de mesure pour les résultats en laboratoire est déterminée à l'aide d'un facteur de couverture et de l'écart type des échantillons, ainsi que l'incertitude relative en pourcentage.

Analyse statistique supplémentaire: Pour évaluer davantage la variabilité et l'accord entre les échantillons collectés par l'Opérateur A et B et les Laboratoires D et E, les analyses statistiques suivantes ont été effectuées :

Analyse de corrélation: Le coefficient de corrélation de Pearson (\(r \)) a été calculé pour déterminer la force et la direction de la relation linéaire entre les mesures obtenues par les deux opérateurs.

Histogramme des différences: Les différences entre les mesures ont été calculées, et un histogramme a été construit pour visualiser la distribution de ces différences.

Graphique de Bland-Altman: Les mêmes différences entre les mesures ont été utilisées pour créer un graphique de Bland-Altman. Ce graphique permet de visualiser l'accord entre les deux méthodes de mesure et d'identifier d'éventuels biais systématiques ou valeurs aberrantes.

Tests statistiques: Un test t apparié a été effectué pour évaluer s'il existait une différence statistiquement significative entre les moyennes des mesures appariées. De plus, un test de rang signé de Wilcoxon, un test non paramétrique, a été réalisé pour évaluer la signification statistique des différences entre les mesures appariées.

Création de l'échantillon composite virtuel pour des distributions connues

Pour analyser l'influence des échantillons composites sur la distribution des données, des distributions virtuelles ont été créées. Le processus méthodologique s'est appuyé sur l'utilisation de distributions de probabilité bien établies créées par des simulations de données.

Création de distribution

Trois distributions de 10 000 échantillons aléatoires ont été générées à partir de trois distributions de probabilité distinctes.

Test 1: Échantillonnage aléatoire

Dans ce test, nous prenons des échantillons aléatoires à partir de différentes distributions de données. Pour chaque échantillon, nous sélectionnons un point de départ aléatoire, puis nous choisissons quatre autres points à proximité de celui-ci. Ensuite, nous calculons la moyenne de ces cinq points. Nous répétons ce processus de nombreuses fois pour voir comment les moyennes varient.

Test 2: Échantillonnage basé sur un intervalle

Dans ce test, au lieu de choisir des points complètement aléatoires, nous commençons par un point de départ aléatoire. Ensuite, nous créons un intervalle autour de ce point, de manière à capturer une plage spécifique. Ensuite, nous choisissons quatre autres points aléatoires à l'intérieur de cet intervalle et calculons la moyenne de ces cinq points. Nous répétons cela plusieurs fois pour voir comment les moyennes varient en fonction de l'intervalle.

Test 3: Intégration de la contrainte spatiale

Dans ce test, nous ajoutons une contrainte spatiale. Cela signifie que les points que nous choisissons doivent être proches les uns des autres dans l'espace. Nous commençons toujours par un point de départ aléatoire, mais les autres points doivent être à une certaine distance de ce point. Ensuite, nous calculons la moyenne de ces points et répétons le processus plusieurs fois pour comprendre comment la contrainte spatiale affecte les moyennes.

Ces tests nous aident à voir comment différents types d'échantillonnage influencent les moyennes de nos échantillons composites et nous aident à mieux comprendre comment ces échantillons se comportent dans des situations réelles.

Exploration statistique et analyse comparative

Une exploration statistique des caractéristiques locales au sein des ensembles de données créés a été menée. Elle comprenait des statistiques descriptives, des visualisations, des tests statistiques et la création de fonctions de densité de probabilité pour comparer les ensembles de données originaux et les ensembles de données composites créés.

Résultats

Analyse d'Incertitude des Stocks et du Laboratoire

Analyse de l'Hétérogénéité de l'Échantillonnage Composite

Dans cette sous-section, l'incertitude liée à l'hétérogénéité de l'échantillonnage a été explorée en comparant deux ensembles de 155 échantillons, désignés comme Échantillons A et Échantillons B, prélevés par deux opérateurs différents suivant le même protocole. L'analyse a révélé des perspectives intéressantes :

1. Comparaison des Statistiques de Base : Les Échantillons B ont montré une concentration moyenne légèrement plus élevée (environ 4351 mg/kg) par rapport aux Échantillons A (environ 4041 mg/kg). Cependant, les Échantillons B présentaient un écart type plus important (3913 mg/kg) par rapport aux Échantillons A (3118 mg/kg), indiquant une plus grande variabilité dans les mesures pour les Échantillons B.

2. Analyse Visuelle : La comparaison visuelle des concentrations de TPH entre le groupe d'échantillons A et le groupe d'échantillons B n'a révélé aucune différence significative, indiquant que les variations provenaient probablement de l'hétérogénéité inhérente du tas plutôt que de facteurs spécifiques à l'opérateur.

3. Corrélation: Il y avait une corrélation robuste de 0,63 entre l'Échantillon A et l'Échantillon B, suggérant une cohérence dans les mesures, avec seulement quelques points de données montrant des valeurs extrêmes.

4. Analyse des Différences : Les différences entre les résultats de l'Échantillon A et de l'Échantillon B suivaient une distribution normale allant de -20 000 à 10 000 mg/kg, sans déséquilibres de données discernables.

5. Sta Le test t apparié (p-valeur ≈ 0,196) et le test de Wilcoxon sur les rangs signés (p-valeur ≈ 0,110) ont tous deux indiqué aucune différence significative entre les moyennes des deux groupes appariés, soutenant l'idée que les différences n'étaient pas statistiquement significatives.

6. Évaluation de l'Hétérogénéité : L'étude a évalué l'hétérogénéité des échantillons de sol collectés par deux groupes d'échantillons différents, Échantillon A et Échantillon B, à partir du même stock. L'échantillon A avait un écart type de 3118 mg/kg, l'échantillon B avait un écart type de 3913 mg/kg, et lorsqu'ils étaient combinés, l'écart type global augmentait à 5003 mg/kg. L'incertitude de mesure avec un facteur de couverture (k) de 2 a été calculée à 4352, avec une incertitude en pourcentage de 81,56 %.

Analyse de l'Incertitude en Laboratoire

Cette partie de la sous-section se concentre sur l'analyse de l'incertitude en laboratoire, où 34 échantillons ont été divisés en deux groupes et analysés dans deux laboratoires différents, Lab A et Lab B :

1. Comparaison des Statistiques de Base : Le Lab A avait une concentration moyenne plus élevée (environ 1768 mg/kg) par rapport au Lab B (environ 1620 mg/kg). Le Lab B avait un écart type plus élevé (2322 mg/kg) par rapport au Lab A (629 mg/kg), indiquant plus de variabilité dans les mesures pour le Lab B.

2. Analyse Visuelle : La comparaison visuelle des concentrations de TPH entre le Lab A et le Lab B a indiqué qu'il y avait de légères différences, mais elles n'étaient pas visuellement significatives.

3. Corrélation : Une corrélation relativement forte de 0,82 a été observée entre le Lab A et le Lab B, suggérant une cohérence dans les mesures.

4. Analyse des Différences : Les différences entre les résultats du Lab A et du Lab B suivaient une distribution normale allant de -4000 mg/kg à 3500 mg/kg.

5. Tests Statistiques : Le test t apparié (p-valeur ≈ 0,21) et le test de Wilcoxon sur les rangs signés (p-valeur ≈ 0,70) ont tous deux indiqué aucune différence significative entre les médianes des deux groupes appariés.

6. Calcul de l'Incertitude : La valeur d'incertitude estimée, calculée avec un intervalle de confiance de 95 %, a été déterminée à 105,32 %, indiquant un niveau élevé d'incertitude provenant principalement de la variabilité du site et du laboratoire.

Composite: Validation de la méthode

1. Test 1 : Échantillonnage Aléatoire

Le Test 1 a lancé la séquence expérimentale en prélevant des échantillons de trois distributions statistiques connues : Normale, Log-Normale et Gamma. Chaque échantillon

composite était composé de cinq observations individuelles, dont la moyenne était la principale statistique d'analyse.

Pour la Distribution Normale, le Test 1 a légèrement augmenté la moyenne par rapport à la distribution originale, indiquant un léger décalage de la tendance centrale. Dans la Distribution Log-Normale, le Test 1 a résulté en une moyenne plus élevée, suggérant que l'échantillon composite avait tendance à avoir des valeurs plus élevées par rapport à la distribution originale. Pour la Distribution Gamma, le Test 1 a réduit l'écart-type et augmenté la moyenne, signifiant un changement à la fois dans la tendance centrale et la variabilité.

2. Test 2 : Échantillonnage Basé sur un Intervalle

En développant le Test 1, le Test 2 a introduit un échantillonnage basé sur des intervalles, définissant une plage de 20% autour de la valeur de l'échantillon primaire. Les échantillons suivants ont été prélevés dans cette plage, permettant une exploration contrôlée des caractéristiques de la distribution.

Pour la Distribution Normale, le Test 2 a maintenu une moyenne et une médiane similaires à celles du Test 1, mais l'écart-type a légèrement augmenté, indiquant une légère augmentation de la variabilité. Dans la Distribution Log-Normale, le Test 2 a résulté en un écart-type plus bas, suggérant une réduction de la variabilité par rapport au Test 1. La moyenne et la médiane sont restées similaires. Pour la Distribution Gamma, le Test 2 a réduit l'écart-type mais a causé une diminution significative de la médiane, impliquant un changement de tendance centrale vers des valeurs plus basses.

3. Test 3 : Intégration de Contraintes Spatiales

Le Test 3 a intégré des contraintes spatiales dans la stratégie d'échantillonnage, tenant compte à la fois des paramètres statistiques et de la proximité spatiale.

Pour la Distribution Normale, le Test 3 a maintenu une moyenne, une médiane et un écart-type similaires à ceux du Test 1, indiquant un impact minimal sur les caractéristiques de la distribution. Dans la Distribution Log-Normale, le Test 3 a montré une légère augmentation de la moyenne et de l'écart-type par rapport au Test 2, mais la médiane est restée similaire, suggérant un léger déplacement vers des valeurs plus élevées. Pour la Distribution Gamma, le Test 3 a légèrement augmenté la moyenne tout en maintenant un écart-type similaire au Test 2.

Résumé de la sous-section "Application sur des Données Réelles" avec des descriptions détaillées de 20% de la taille du texte original :

Application sur des Données Réelles

Remplacement des Non-Détectés et Ajustement des Données

Dans cette sous-section, les données d'entrée du modèle ont été ajustées en fonction des biais identifiés dans les analyses en laboratoire. Les échantillons initiaux du projet de

remédiation ont été réalisés par le Laboratoire A, tandis que l'excavation a été effectuée par le Laboratoire B. La distribution des données est restée inchangée, mais la moyenne a été ajustée en utilisant uniquement l'incertitude en laboratoire.

Jusqu'à présent, toutes les valeurs situées en dessous de la limite de détection avaient été remplacées par la valeur de la limite de détection. Cependant, comme le pourcentage de non-détectés est de 37,7%, ce qui dépasse le seuil établi et les recommandations de l'EPA dans son document d'orientation "U.S. EPA's 2004 Local Limits Development Guidance Appendices", il a été décidé de procéder au remplacement en utilisant la méthode d'imputation pour gérer les non-détectés de l'ensemble de données d'évaluation. Pour la sélection des valeurs d'imputation, une distribution des données a été créée en suivant la distribution originale pour les données allant de 0 à la limite de détection. La modification a été effectuée sans altérer la forme de la distribution pour préserver l'écart-type des données.

Adapter la Distribution

Adapter la Distribution Gamma

Pour continuer, nous avons utilisé des outils statistiques Python spécialisés pour adapter la distribution gamma aux données réelles du site. Ce processus d'ajustement utilise une estimation du maximum de vraisemblance basée sur la bibliothèque Scipy Stats. La technique identifie les valeurs de k et de theta et utilise un optimiseur pour ajuster les données empiriques au modèle.

Évaluation de l'ajustement et estimation des paramètres

Après avoir ajusté la distribution gamma, nous évaluons la qualité de l'ajustement. Cette évaluation comprend une inspection visuelle, en comparant la distribution gamma ajustée aux données observées, ainsi que des évaluations statistiques qui mesurent le niveau de concordance entre le modèle et la réalité. Les techniques courantes incluent les graphiques de probabilité et les tests d'hypothèses tels que les tests de Kolmogorov-Smirnov ou du Chi-carré.

Nous documentons les valeurs estimées de k et de θ dérivées du processus d'ajustement. Ces paramètres servent de base pour les étapes ultérieures, car ils définissent la distribution gamma qui servira de fondement à nos analyses.

Nous avons utilisé la bibliothèque scipy.stats en Python pour ajuster les données sélectionnées à une distribution gamma. Le processus d'ajustement consistait à trouver les meilleurs paramètres de la distribution gamma (forme, loc et échelle) qui minimisaient la différence entre les données observées et la distribution gamma. Le processus d'ajustement a estimé une forme k de 0,22 et une échelle theta de 18 577.

Échantillonnage et Création de Composites

Une fois que nous avons ajusté avec succès les données réelles du site à une distribution gamma et estimé les paramètres cruciaux k et theta, nous procédons à l'application de notre

méthodologie décrite et menons les trois tests distincts (Test 1, Test 2 et Test 3) sur les données réelles du site.

De la même manière que dans les distributions connues, nous avons généré des moyennes composites à partir d'ensembles de cinq échantillons dérivés de la distribution originale. Ce processus a été répété 1 000 fois. L'indépendance à la fois des échantillons individuels au sein d'un échantillon composite et des échantillons composites eux-mêmes est soigneusement maintenue.

Comparaison entre les Données Composites Modélisées et les Données Réelles d'Excavation

Lorsque nous comparons les résultats du Test 3 avec les données d'excavation, nous observons que le Test 3 présente une distribution plus proche de celle de l'excavation. Bien qu'il subsiste quelques différences, telles qu'une légère surestimation dans la plage de moins de 5000 mg/kg et une sous-estimation similaire dans la plage de 5000 à 15 000 mg/kg, la similitude est évidente. La plage de plus de 15 000 mg/kg reste également assez similaire entre le Test 3 et l'excavation. Cela suggère que le Test 3 parvient efficacement à reproduire les concentrations de TPH observées lors de l'excavation, ce qui en fait une méthode précieuse pour modéliser et analyser de manière précise les données de pollution du sol.

Volumes par Date

Tout au long du projet de remédiation, un système complet a été mis en place pour suivre quotidiennement les volumes de sol excavés, transportés et traités. Ce processus a débuté dès la source, avec chaque chargement de camion enregistré, ainsi que l'origine et la destination. Parallèlement, les volumes effectivement excavés ont été validés à l'aide de la technologie des drones, offrant une vue d'ensemble pour assurer la précision et la précision des mesures. De plus, les tas de sol accumulés en attente de traitement ont été évalués par des experts du domaine, corroborant davantage les données volumétriques collectées.

Ces données brutes, collectées consciencieusement quotidiennement, ont ensuite été regroupées dans des rapports hebdomadaires. Cependant, la transition des données de terrain brutes en informations exploitables n'était pas simple. Les données nécessitaient un nettoyage et une homogénéisation approfondis pour garantir la cohérence et la fiabilité. La traçabilité du sol excavé était primordiale. Tous les excavateurs étaient équipés de GPS et des cartes d'excavation. Par conséquent, bien que la collecte initiale de données comprenait une géolocalisation pour cartographier précisément la source de la contamination et les emplacements spécifiques ciblés pour l'excavation, le volume massif de sol impliqué présentait d'importants défis pour maintenir un haut niveau de traçabilité pour chaque chargement de camion. Il est devenu peu pratique de suivre l'origine de chaque chargement de sol individuellement en raison de l'ampleur de l'opération. Cependant, pour garantir une approche structurée du suivi et de la responsabilité, le projet a mis en place un système de suivi catégoriel. Ce système classifiait le sol en différentes catégories en fonction de la méthode de traitement prévue : 'Propre', 'Bio' et 'Lavage du Sol'. En organisant l'origine et la destination du sol dans ces grandes catégories, le projet maintenait une vue d'ensemble du parcours du sol de l'excavation au traitement. Cette méthode facilitait un niveau de surveillance gérable, garantissant que, bien que les chargements de camions individuels ne

puissent pas être suivis isolément, le flux général et le traitement des volumes de sol étaient systématiquement surveillés et documentés.

Discussion et Conclusion

Principaux points de notre étude et conclusions importantes :

Changement de Laboratoire et Remplacement des Non-Détectés :
- Nous avons utilisé une méthode d'imputation pour remplacer les valeurs non-détectées dans nos données de laboratoire.
- Cette méthode a modifié la moyenne et la médiane des données, mais n'a pas fondamentalement changé la classification de la pollution du sol en raison des seuils élevés que nous avons utilisés.

Création de Distributions Connues et Validation de la Méthode :
- Nous avons utilisé des distributions connues pour valider notre technique, mais ces distributions ne doivent pas être considérées comme fixes.
- Il est important d'adapter la méthode aux caractéristiques spécifiques de chaque étude.

Impact sur Variabilité et Moyenne des Composites :
- L'utilisation d'échantillons composites a réduit la variabilité des données, mais peut affecter la moyenne, en particulier pour les distributions log-normales et gamma.
- Nous avons utilisé trois tests, dont le test 3 semble être le plus approprié.

Volumes Finaux, Stratégie d'Excavation et Coûts :
- Nous avons comparé les volumes estimés à l'aide de notre méthode avec les volumes réels excavés et avons constaté une bonne similitude.
- La stratégie d'excavation doit être adaptée à la réalité opérationnelle d'un projet de remédiation, et les coûts doivent être considérés comme indicatifs.

Applicabilité et Limitations de la Méthode :
- Notre méthode peut être appliquée à la plupart des sites pollués, mais nécessite un ensemble de données volumineux.
- Il est essentiel de bien traiter les données d'entrée, notamment en ce qui concerne les non-détectés et les changements de laboratoire.

En conclusion, notre méthode est efficace pour prédire ce qui se passera dans un projet de remédiation impliquant une excavation et une réanalyse. Cependant, une étude minutieuse des données d'entrée est essentielle, et il faut tenir compte des différentes limitations et considérations spécifiques à chaque projet.