Traitement géostatistique des résultats de mesure pour la caractérisation radiologique dans le cadre de l'assainissement/démantèlement de sites nucléaires

Martin Wieskotten

THÈSE DE DOCTORAT D'AVIGNON UNIVERSITÉ

École Doctorale n°536
Agrosciences et Sciences

Mention de doctorat :
Doctorat en Mathématiques

Laboratoire de Mathématiques d'Avignon, UPR 2151

Présentée par
Martin WIESKOTTEN

Traitement géostatistique des résultats de mesure pour la caractérisation radiologique dans le cadre de l'assainissement/démantèlement de sites nucléaires

---

Soutenue publiquement le 31 janvier 2022 à Avignon, devant le jury :

Robert Faivre	Directeur de Recherche	INRAE	Rapporteur
Alberto Pasanisi	Directeur R&D	Edison SpA	Rapporteur
Céline Helbert	Maîtresse de Conférence	Centrale Lyon	Examinatrice
Delphine Blanke	Professeure	Avignon Université	Examinatrice
Céline Lacaux	Professeure	Avignon Université	Co-directrice
Bertrand Iooss	Chercheur Senior	EDF R&D	Co-directeur
Marielle Crozet	Directrice de Recherche	CEA	Encadrante
Yvon Desnoyers	Ingénieur Expert	Géovariances	Examinateur
Amandine Marrel	Experte Senior	CEA	Invitée
Magali Saluden	Cheffe de service	CEA	Invitée

Résumé

Dans les projets d'assainissement/démantèlement de sites nucléaires, l'étape de caractérisation radiologique initiale a pour objectif d'estimer la quantité et la répartition spatiale de la contamination en différents radionucléides. Pour réaliser cette estimation, des mesures sont réalisées sur site et en laboratoire. Cependant compte-tenu des environnements et de la nature des mesures, le nombre de mesures peut être réduit et/ou ces mêmes mesures peuvent êtres censurées. Avec ce type de jeu de données, les approches classiques de géostatistique n'offrent pas de solutions pratiques pour les traiter. Dans cette thèse, plusieurs méthodes permettant de traiter les problématiques posées par ces jeux de données sont étudiées. Parmi ces méthodes, le krigeage bayésien sera particulièrement approfondi, puisqu'il permet de construire des modèles efficaces lorsque peu d'observations sont disponibles. Ce krigeage étant singulier, de nouveaux outils de validation seront introduits, et le krigeage bayésien sera comparé à d'autres méthodes usuelles comme le krigeage ordinaire. Nous proposons également une variante d'un algorithme MCMC (Méthode de Monte-Carlo par Chaînes de Markov) pour la résolution des équations du krigeage bayésien ainsi que pour l'augmentation de données, méthode permettant le traitement de données censurées.

Abstract

In decommissioning projects of nuclear facilities, the radiological characterisation step aims to estimate the quantity and spatial distribution of radionucleides. Statistical tools such as ordinary kriging (stemming from geostatistics) is part of usual industrial methods for soils sanitizing. To carry out the estimation, measurements are performed *in situ* or in laborator. However, due to the constrained environment and the nature of measurements, the data set's size can be reduced and/or measurement results can be censored. With these types of data sets, usual geostatistical practices do not offer practical solutions. In this thesis we study several methods allowing the treatment of these specific problematics. We will especially dwell on one of these methods, called Bayesian kriging, since it allows to treat efficiently these kind of data sets. This kriging type being singular, new validation criterion will be introduced and Bayesian kriging will be compared to other usual methods like ordinary kriging. We also suggest a variant of a MCMC algorithm to solve Bayesian kriging equations and data augmentation (which allows the processing of censored data).

Remerciements

Je souhaite tout d'abord remercier Robert Faivre et Alberto Pasanisi pour avoir accepter d'être les rapporteurs de ma thèse. Leurs commentaires et avis ont été précieux, et leurs questions durant la soutenance sont à l'origine de discussions passionnantes.

Parvenir à cette soutenance et arriver au bout de cette thèse aura été une grande victoire. Je le dois en grande partie à mes directeurs de thèse Céline Lacaux et Bertrand Iooss et à mon encadrante CEA Marielle Crozet qui m'ont encouragé, conseillé et accompagné jusqu'au bout de ces travaux. Je ne les remercierai jamais assez pour leur soutien et leur patience à mon égard, et ce malgré certains moments difficiles. Je souhaite également remercier Nadia Pérot, qui n'aura pas suivi la thèse jusqu'à la fin, mais qui m'a aidé à formuler les axes de recherche de cette thèse. Je remercie ensuite Amandine Marrel d'avoir accepté de travailler sur la rédaction d'un article et de nous avoir fourni ses commentaires et remarques précieuses afin d'enrichir le contenu de celui-ci. Merci aussi à Danièle Roudil qui m'a guidé durant ma thèse en tant que présidente de la CETAMA, à Manuel Saez pour son soutien et ses conseils lors de ma présence à Cadarache. Je souhaite tout particulièrement remercier Cédric Rivier qui m'aura à la fois accompagné en tant que président de la CETAMA, mais également en tant que partenaire d'escalade (et grâce à qui j'ai appris à grimper en tête !). Enfin je remercie Magali Saluden et Yvon Desnoyers d'avoir participé au jury de ma soutenance, ainsi que pour leurs remarques pertinentes.

Je remercie également Delphine Blanke et Céline Helbert, à la fois en tant que membres de mon CST que de mon jury de thèse, qui m'ont rassuré, conseillé et encouragé dans cette longue tâche qu'est la thèse. Je remercie également Delphine d'avoir accepté de présider le jury de ma soutenance de thèse.

Ensuite merci évidemment à l'équipe de la CETAMA : Giacomo, mon conseiller culture, partenaire d'escalade et de bus pour toutes les discussions passionnantes, Fabienne pour sa gentillesse et sa patience avec moi, Sébastien pour ses réponses à mes questions sur le nucléaire et les discussions politiques, Véronique pour son soutien et ses conseils dans les moments difficiles de ma thèse, Alexandre mon « co-bureau » de Marcoule avec qui j'ai pu longuement discuter de mangas et jeux videos, Caroline et Igor pour leur sympathie et la visite du laboratoire d'Atalante !

Je n'oublie pas non plus ma première année à Cadarache ni l'équipe du SESI, en particulier les autres doctorants Thibauld, Aude et Marlène et camarades du Hameau, que je remercie pour toutes les soirées et les moments passés ensemble. Je remercie bien sûr Nicolas, mon « co-bureau » de Cadarache pour tous les bons moments passés ensemble et son aide pour réparer mon vélo, sous oublier Reda pour ses réponses à mes questions géostatistiques.

Bien sûr je pense aussi à mes partenaires d'Internet et amis de toujours, Philippe et Thomas avec qui j'ai pu me défouler dès que j'en avais besoin lors de parties en ligne endiablées.

Merci aussi à mes amis d'Avignon, qu'ils soient doctorants, alternants, ou sous contrat. En particulier Lucas grâce à qui j'ai pu découvrir une passion pour le skate et l'escalade

(et même changer de coupe de cheveux !). Je remercie également toute l'équipe des jeudis gourmets, Simon, Gabriel, Vincent, Florian et Laurianne (mais pas pour les Terraforming Mars, ce jeu est nul !). Je remercie enfin l'Aisther dont les soirées m'ont permis de rencontrer tous les autres doctorants du CEA.

Il faut également que je cite mes amis d'école (voire de classe préparatoire), Hirvin, Camille et Anatole chez qui je pouvais toujours faire escale lors de mes retours à Paris, et Théodore et Valentin pour les balades à vélo et les randonnées de l'extrême !

J'ai une pensée pour mes amis doctorants Anas et Mohammed, pour les rares mais bons moments passés au LMA, et à qui je souhaite le meilleur pour leur thèse !

Enfin je remercie ma famille pour son soutien et amour durant ces trois années si spéciales. Entre les déménagements, les retours rapides en Île de France et les confinements, ils ont été d'un soutien sans faille.

Table des matières

Glossaire

Sigles	Définitions
ACP	Analyse en Composantes Principales
A&D	Assainissement et Démantèlement
AIC	Akaike Information Criterion / Critère d'information d'Akaike
AIEA	Agence Internationale de l'Energie Atomique
ASN	Autorité de Sûreté Nucléaire
BIC	Bayesian Information Criterion / Critère d'information bayésien
CEA	Commissariat à l'Energie Atomique et aux énergies alternatives
EDF	Electricité De France
EM	Espérance-Maximisation
GUM	Guide pour l'expression de l'incertitude de mesure
ICP-MS	Inductively Coupled Plasma Mass Spectrometry / Spectrométrie de masse à plasma à couplage inductif
INB	Installation Nucléaire de Base
i.i.d.	indépendant(e)s et identiquement distribué(e)s
INSIDER	Improved Nuclear SIte characterization for waste minimization in Dismantling and decommissioning operations under constrainded EnviRonment / Caractérisation améliorée de sites nucléaires pour la minimisation de déchets dans les opérations d'assainissement et de démantèlement dans un environnement contraint.
KM	Kaplan-Meier
LD	Limite de Détection
LOO	Leave-One-Out
MCMC	Monte-Carlo par Chaînes de Markov
MSE	Mean Square Error / Moyenne des erreurs quadratiques
PVA	Predictive Variance Adequacy / Adéquation de la Variance Prédictive
REP	Réacteurs à eau préssurisée
ROS	Regression on Order Statistics / Régression sur des statistiques d'ordre
SD	Seuil de Décision
TIMS	Thermal Ionisation Mass Spectrometry / Spectrométrie de masse à ionisation thermique

Symboles	**Définitions**
$\mathbb{R}$	Ensemble des réels
$\mathbb{P}$	Probabilité
P	Valeur de probabilité
Z	Variable aléatoire
z	Réalisation de la variable aléatoire Z
$\boldsymbol{Z}$	Vecteur
$F(.)$	Fonction de répartition
$S(.)$	Fonction de survie
$f_Z(.)$	Densité de Z
q_s	Quantile d'ordre s
$\mathbb{E}[Z]$	Espérance de Z
$Var[Z]$	Variance de Z
$Cov(Z, Y)$	Covariance de Z et Y
$\boldsymbol{R}$	Matrice de covariance
$C(.)$	Fonction de covariance
$\gamma(.)$	Variogramme
$L(\boldsymbol{z}; \theta)$	Vraisemblance des observations selon le paramètre θ
$\hat{\theta}$	Estimateur ou prédicteur de θ
$\hat{\theta}^*$	Réalisation de l'estimateur ou du prédicteur
$\boldsymbol{D}$	Matrice de design
μ	Paramètre de tendance ou moyenne
σ^2	Paramètre de dispersion ou variance
ϕ	Portée
τ^2	Effet de pépite
η	Rapport effet de pépite/variance
π	Loi *a priori*
$\propto$	Proportionnel à
λ_{BC}	Paramètre de transformation de Box-Cox
$\boldsymbol{x}$	Position spatiale
$\boldsymbol{\lambda}$	Pondérateur de krigeage
ν	Limite de détection

Introduction

I.1 Les installations nucléaires en cours de démantèlement ou à l'arrêt en France

En France, l'industrie du nucléaire civil a connu un fort développement initié dans les années 1960, dotant le pays d'un grand nombre d'installations nucléaires de base (INB) à des fins industrielles ou de recherche. De nombreuses installations arrivent maintenant en fin de vie et doivent passer par un ensemble d'opérations de démontage et d'assainissement. On parle ici de démantèlement. Une fois ce démantèlement terminé, le site est déclassé et réhabilité.

Démantèlement [ASN, 2022] :
Le démantèlement couvre l'ensemble des activités, techniques et administratives, réalisées après l'arrêt définitif d'une installation nucléaire, afin d'atteindre un état prédéfini où la totalité des substances dangereuses et radioactives a été évacuée de l'installation.

Déclassement [ASN, 2022] :
Le déclassement est l'ensemble des opérations administratives et réglementaires destinées soit à classer une installation nucléaire dans une catégorie inférieure, soit à en supprimer le classement initial. Ici le déclassement consiste à supprimer une installation de la liste des INB, ou du moins à la classer dans une catégorie différente.

L'Autorité de Sûreté Nucléaire (ASN) recense l'ensemble des sites à l'arrêt ou en cours de démantèlement avec la carte donnée en Figure 1. Ces sites sont principalement répartis entre les trois grands acteurs du nucléaire : le CEA (Commissariat à l'Energie atomique et aux énergies alternatives), EDF (Electricité de France) et Orano. Trente-cinq sites sont donc concernés en 2021 par des projets d'assainissement et démantèlement (A&D). Davantage de détails sur ces différentes installations (comme leur fonction, le début de leur démantèlement, etc.) peuvent être trouvés dans [IRSN, 2020].

Les démantèlements de ces installations peuvent être extrêmement longs. Ils s'étalent souvent sur plusieurs décades, selon la taille de l'installation en question. Par exemple, le démantèlement du réacteur prototype à eau lourde de Brennilis a été démarré en 1985, et doit se terminer avant 2040. Néanmoins il est important de noter que selon la technologie de l'installation en question, son historique et les retours d'expérience sur d'autres installations, ce temps peut fortement varier.

En plus d'être planifiés sur le long terme, ces projets sont également coûteux. [Chevet et al., 2020] estime que les charges de démantèlement du CEA au 31 décembre 2018 représentent 16, 5 milliards d'euros. EDF estime que le démantèlement d'un REP (réacteur à eau pressurisée dit de deuxième génération formant la majorité du parc de réacteurs français, voir [Chevet et al., 2020]) est de 400 millions d'euros. A nouveau ce montant va dépendre du type d'installation, de sa technologie, etc.

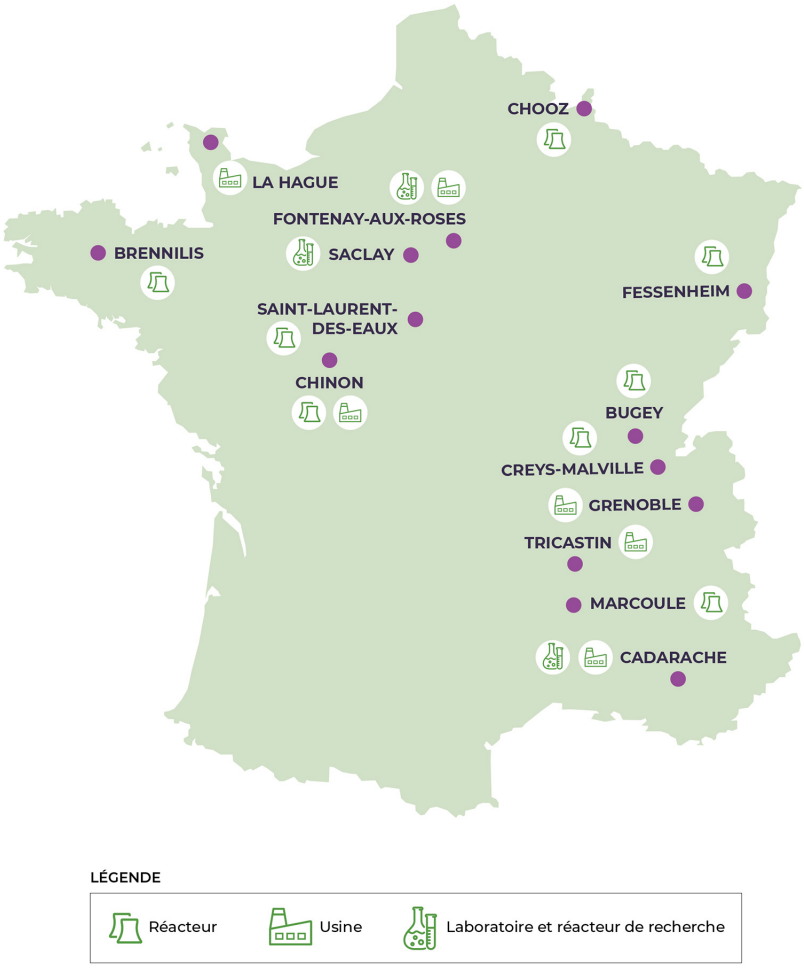


FIGURE 1 – Carte des installations à l'arrêt définitif ou en cours de démantèlement en 2021 [ASN, 2022].

Enfin le nombre de projets de démantèlement va rapidement augmenter avec l'arrêt progressif des REP. Le gouvernement avait initialement prévu la fermeture de 14 réacteurs d'ici 2035. Néanmoins la crise de l'énergie récente a amené le gouvernement actuel à revenir sur cette décision, et à prévoir la construction de nouveaux réacteurs. Ces sites nécessiteront à leur tour d'être démantelés, accroissant ainsi le nombre de chantiers de démantèlement à l'avenir.

Les projets de démantèlement sont donc des projets très longs et très coûteux, mais essentiels pour l'acceptabilité de l'industrie nucléaire. La stratégie du CEA est l'assainissement et le démantèlement immédiat d'un site nucléaire à la fin de son fonctionnement pour garantir une gestion pérenne des INB ([Chevet et al., 2020]). Bien entendu, ces projets

étant importants d'un point de vue de la santé et de l'opinion publique, ils sont fortement encadrés par la réglementation en vigueur.

I.2 Cadre réglementaire

Le cadre réglementaire pour l'A&D de sites nucléaires est en grande partie défini par l'ASN qui réglemente et contrôle l'industrie nucléaire en France. Ce cadre est réalisé en accord avec les recommandations internationales de l'AIEA ([AIEA, 2017]). Ainsi la vie d'une INB passe par les différentes étapes présentées dans la Figure 2 qui sont encadrées par la législation.

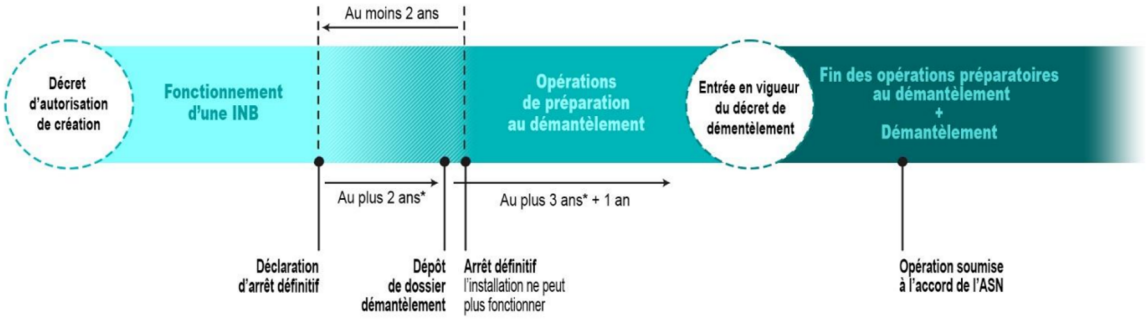


FIGURE 2 – Phases de vie d'une INB [ASN, 2016c].

Au moins deux ans avant la date d'arrêt envisagée, l'exploitant déclare son intention d'arrêter son installation. Après cette déclaration d'arrêt, l'exploitant dispose de deux ans maximum pour fournir un plan de démantèlement présentant les différentes opérations de démantèlement envisagées, ainsi que les dispositions prises concernant la sûreté nucléaire et la radioprotection.

Sûreté nucléaire [DEN, 2017] :
La sûreté nucléaire est l'ensemble des dispositions techniques et des mesures d'organisation appliquées aux installations ou aux activités nucléaires, en vue de prévenir les accidents ou d'en limiter les effets.

Radioprotection [DEN, 2017] :
La radioprotection est l'ensemble des règles, des procédures et des moyens de prévention et de surveillance visant à empêcher ou à réduire les effets nocifs des rayonnements ionisants sur les personnes directement ou indirectement, y compris lors des atteintes portées à l'environnement.

A partir de la date d'arrêt définitif, l'exploitant n'est plus autorisé à faire fonctionner l'installation. Il commence alors les opérations de préparation au démantèlement, notamment l'évacuation des substances radioactives et chimiques encore présentes sur le site, et l'aménagement des locaux (zone de stockages, barrières de protection, etc.) en vue des travaux prévus. Le démantèlement de l'INB est prescrit par un décret spécifiant plusieurs éléments clés du projet de démantèlement, comme les principales étapes du chantier, sa date de fin et les objectifs de l'assainissement. Ce décret est réalisé à partir du plan de

démantèlement déposé par l'exploitant et après avis de l'ASN. Les opérations spécifiées peuvent alors commencer. Une fois ces opérations terminées et les objectifs fixés par le décret atteints, le site peut être déclassé.

L'ASN fournit plusieurs guides importants pour la conception du plan de démantèlement. Le guide n°6 [ASN, 2016c] donne les recommandations générales pour le démantèlement, en détaillant notamment les différentes étapes obligatoires pour le déclassement ainsi que les phases de vie de l'INB. Ici deux guides sont d'un intérêt particulier pour les thématiques de cette thèse. Le guide n°14 ([ASN, 2016a]) présente des recommandations pour l'assainissement des structures (par exemple les bétons de l'installation, la plomberie, etc.), tandis que le second, le guide n°24 ([ASN, 2016b]), traite des sols pollués. Ils viennent donner plusieurs éléments et solutions concernant certaines contraintes de démantèlement rencontrées lors de ces chantiers. Ces deux guides discutent de la caractérisation radiologique, une étape essentielle des opérations de préparation au démantèlement ainsi qu'au déclassement du site.

I.3 Focus sur la caractérisation radiologique d'une installation

I.3.1 Définition et principe général

La caractérisation radiologique est une étape déterminante des opérations de démantèlement.

> **Caractérisation** [ISO, 2017] :
> La caractérisation consiste à déterminer la nature, la concentration et l'étendue spatiale du contenu radiologique et chimique présent dans un lieu donné.

De manière générale, cette caractérisation est réalisée lors des opérations de préparation au démantèlement, mais également à la fin du chantier afin de s'assurer que les objectifs fixés soient atteints. Ce type de caractérisation peut être réalisé sur des sols, mais également sur certaines structures de l'installation (comme par exemple l'échangeur de chaleur du réacteur G3, présenté dans la section 2.12.1). On peut schématiser cette caractérisation en trois étapes données dans la Figure 3.

FIGURE 3 – Schéma des étapes principales de caractérisation radiologique initiale, [Granier et al., 2017].

La première phase consiste à préciser les objectifs de cette caractérisation, en accord avec les directives de l'ASN et les objectifs donnés dans le plan de démantèlement. Ensuite vient une étape d'échantillonnage dont le but est de recueillir suffisamment d'informations pour répondre aux objectifs définis à l'étape précédente. Enfin les résultats de mesure sont interprétés et analysés afin d'identifier les étapes de démantèlement nécessaires, mesures de protection à mettre en place, etc.

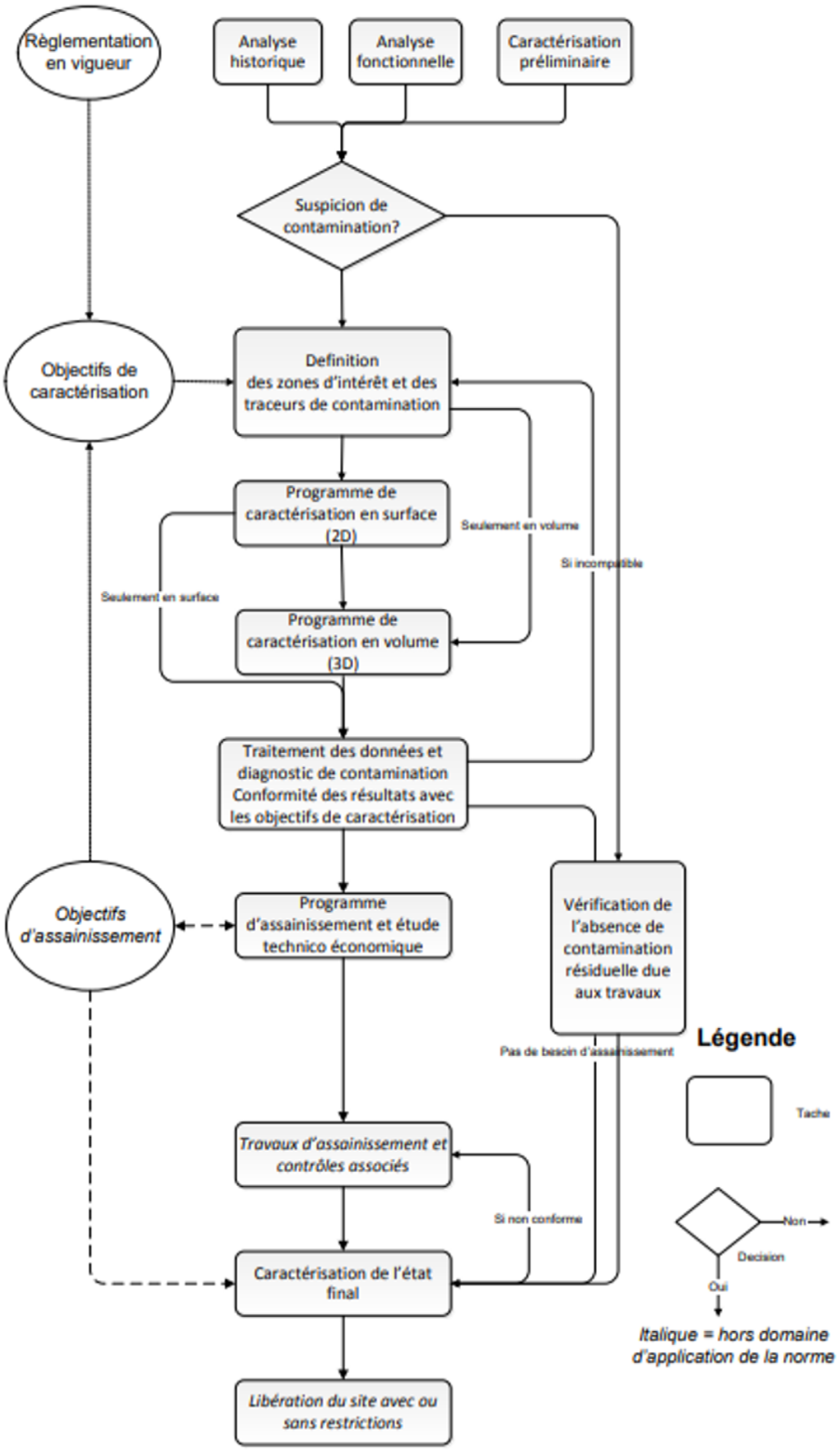


FIGURE 4 – Logigramme de la stratégie de caractérisation pour l'assainissement de sites contaminés, [Granier et al., 2017].

La norme ISO 18557 ([ISO, 2017]) fournit une méthodologie à l'exploitant pour définir et

dimensionner des plans de démantèlement efficaces. Elle fait appel à différentes techniques de mesure et d'échantillonnage, ainsi qu'à une analyse statistique des données collectées. Le sujet de cette thèse concerne cette analyse statistique. La Figure 4 présente un logigramme pour la mise en place d'une stratégie de caractérisation efficace.

L'analyse fonctionnelle vient identifier les zones contaminées par le fonctionnement normal de l'installation, tandis que l'analyse historique identifie les zones contaminées par d'éventuels accidents. Ces deux analyses permettent alors de définir les zones d'intérêt dont la caractérisation doit être faite. Des zones de référence sont également définies pour réaliser des comparaisons avec les zones d'intérêt. Davantage de détails sur la définition de ces zones peuvent être trouvés dans [ISO, 2017] et [Granier et al., 2017]. Une fois ces zones et les objectifs de caractérisation définis, une stratégie d'échantillonnage peut être mise en place.

I.3.2 Stratégie d'échantillonnage

La stratégie d'échantillonnage vient définir la méthode des prélèvements et des mesures ainsi que leur position. Ces positions sont définies par la géométrie d'échantillonnage. [Belbeze et al., 2013] donne plusieurs exemples différents de plans d'échantillonnage. Un premier exemple classique est le plan d'échantillonnage préférentiel.

Les plans d'échantillonnage préférentiels font appel à une connaissance d'expert sur le site (nature du site, historique, etc.) pour choisir la position des mesures et des prélèvements (connaissances *a priori*). Ce type de plan permet de limiter le nombre de mesures selon l'objectif. Par exemple, si l'objectif de l'étude est de caractériser les zones polluées (en identifiant les concentrations en contamination), ce type de plan permet de diminuer le nombre de mesures nécessaires en se concentrant sur les zones *a priori* contaminées. La Figure 5 illustre un exemple d'un plan d'échantillonnage préférentiel.

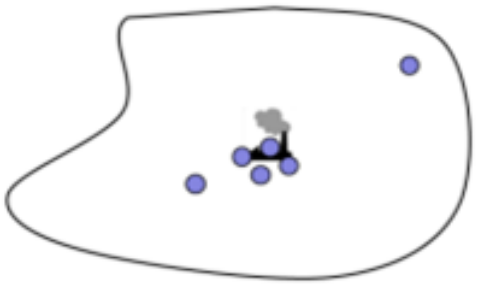


FIGURE 5 – Exemple de plan préférentiel [Belbeze et al., 2013].

Néanmoins, si l'objectif est de caractériser l'ensemble du site étudié (installation, chantier, etc.), ces plans d'échantillonnage induisent des biais d'échantillonnage (car l'échantillon alors choisi n'est pas représentatif du domaine). Par exemple, si l'on cherche à mesurer une radioactivité surfacique et que les mesures et prélèvements sont principalement réalisés dans les zones polluées *a priori*, alors la moyenne des résultats de mesures de radioactivité surfacique sera surestimée par rapport à celle du domaine d'étude.

Domaine d'étude :
Le domaine d'étude est une zone d'intérêt (installation, chantier...) définie par le plan de démantèlement sur laquelle la grandeur physique (radioactivité massique ou surfacique, concentration en un radionucléide) sera étudiée. Cette région est notée D.

Dans le cas d'un plan préférentiel, des traitements particuliers des données sont mis en œuvre en amont de l'analyse statistique pour limiter les biais introduits par le choix des positions. Parmi ces traitements, existent les techniques de dégroupement, qui permettent de limiter l'influence des points présents dans les régions de D dont la densité d'échantillonnage (nombre de mesures par unité de surface ou de volume) est forte. Cette méthode est développée dans [Deutsch, 1989].

Dans le cas de plans d'échantillonnage aléatoires, les positions des mesures sont tirées aléatoirement selon une distribution *a priori* (souvent uniforme). La Figure 6 donne un exemple de plan aléatoire.

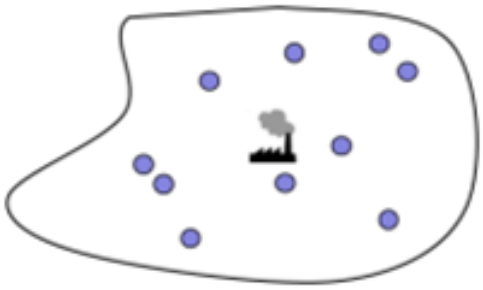


FIGURE 6 – Exemple de plan aléatoire [Belbeze et al., 2013].

Enfin un dernier exemple classique de plan d'échantillonnage est le plan systématique qui correspond à une grille régulière. Ce plan est parfois associé en géostatistique aux plans aléatoires, la justification étant que l'origine de cette grille est aléatoire. Il est l'un des plus répandus puisqu'il permet une couverture homogène de la région D et qu'il est souvent plus simple à mettre en place. La Figure 7 illustre le plan d'échantillonnage systématique selon une maille régulière carrée.

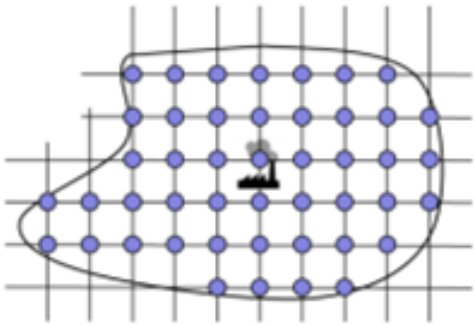


FIGURE 7 – Exemple de plan systématique [Belbeze et al., 2013].

Dans la suite du document, on considérera un échantillonnage de ce type, sauf mention contraire. Des inventaires et des discussions sur les schémas d'échantillonnage peuvent être trouvés dans [Wang et al., 2012] ou dans les travaux du projet « INSIDER » [Rogiers et al., 2018].

I.3.3 Mesures et méthodes de caractérisation

Une fois la stratégie d'échantillonnage planifiée, il est nécessaire de réaliser les mesures spécifiées par la stratégie. La brève présentation faite ici des méthodes de caractérisation est non-exhaustive et sert de contexte pour certains aspects du sujet de thèse. Le(a) lecteur(rice) intéressé(e) peut se tourner vers les références suivantes pour davantage de détails : [Granier et al., 2017, ISO, 2017, Amgarou et al., 2018].

Il n'est pas rare d'observer sur les chantiers d'A&D la présence de plusieurs radionucléides différents (souvent issus du processus de fission nucléaire ou d'activations neutroniques). De plus, il est rare qu'une quantité en radionucléide soit directement disponible, et ce sont souvent d'autres grandeurs physiques qui sont mesurées. Ainsi pour réaliser l'inventaire des radionucléides présents sur le site, plusieurs mesures identifiant les différentes contaminations doivent être réalisées.

Grandeur [BIPM et al., 2012] :
Une grandeur est une propriété d'un phénomène, d'un corps ou d'une substance, que l'on peut exprimer quantitativement sous forme d'un nombre et d'une référence.

Par exemple, les grandeurs physiques rencontrées dans les projets de démantèlement peuvent être des activités (un nombre de désintégrations de noyaux radioactifs par seconde, en Bq), une activité massique, surfacique ou encore volumique (respectivement en Bq/kg, Bq/m^2 ou Bq/m^3). Ces activités peuvent correspondre à différents rayonnements : α, β, γ et neutronique, selon la nature des radionucléides présents. Des explications sur la radioactivité et les différents rayonnements peuvent être trouvées dans l'ouvrage [Foos, 1994]. Des analyses chimiques peuvent également être réalisées, fournissant ainsi des concentrations en radionucléides. Tous ces éléments rendent parfois difficile l'identification de la nature et de la concentration des radionucléides présents. C'est pour cette raison que de nombreuses méthodes de mesure sont employées, afin de pouvoir identifier séparément les différents contaminants présents.

Chaque méthode de mesure aboutit à un résultat venant estimer la valeur de la grandeur étudiée. Quelques définitions utiles pour les problématiques de cette thèse sont données ici, et le(a) lecteur(rice) intéressé(e) par la mesure et la métrologie peut se référer au GUM, le guide pour l'expression de l'incertitude de mesure ([BIPM et al., 2008]) ou au VIM, le Vocabulaire International de Métrologie ([BIPM et al., 2012]).

Résultat de mesure [Crozet et al., 2015] :
Un résultat de mesure est une estimation de la valeur vraie de la grandeur physique que l'on veut mesurer.

Ce résultat de mesure étant une estimation, on lui associe une incertitude dite de mesure.

Incertitude de mesure [BIPM et al., 2012] :
L'incertitude de mesure est un paramètre non négatif qui caractérise la dispersion des valeurs attribuées à une grandeur physique, à partir des informations utilisées. Si ce paramètre est exprimé sous la forme d'un écart-type, on parle d'incertitude-type.

La notion de limite de détection est extrêmement importante entre autres pour les projets de démantèlement. Elle est liée à l'incertitude de mesure, et nous reviendrons sur sa définition ainsi que sur son importance dans le chapitre 3.1.

On peut distinguer les mesures réalisées sur site, dites *in-situ*, et celles réalisées en laboratoire. Les mesures *in-situ* sont généralement réparties en 3 catégories ([Granier et al., 2017]) :

— mesures de débit de dose ;
— mesures surfaciques de contamination sur frottis ;
— mesures d'activité massique de radionucléides par spectrométrie de rayonnement ionisant (souvent de rayonnement γ).

Ces mesures sont relativement simples à mettre en place. Un inventaire non-exhaustif des méthodes *in-situ* peut être trouvé dans [DEN, 2017], document écrit dans le cadre du projet européen « INSIDER ».

On parle de mesures en laboratoire lorsque des prélèvements sont réalisés sur site puis transmis à des laboratoires pour analyse. Le panel de méthodes de mesure est alors souvent plus varié, avec des méthodes fournissant des biais et des incertitudes de mesure souvent plus faibles que les mesures *in-situ*. Pour l'identification et/ou la quantification de radionucléides, on trouve ([Granier et al., 2017]) :

— le comptage de rayonnement α et β total ;
— la scintillation liquide pour les radionucléides β ;
— la spectrométrie de masse par plasma à couplage inductif (ICP-MS pour Inductively Coupled Plasma Mass Spectrometry) ou avec thermoionisation (TIMS) ;
— la spectrométrie α et γ.

Un inconvénient de ce type de mesure est le transport des prélèvements, qui dans le cas de prélèvements radioactifs est soumis à certaines contraintes réglementaires. De plus le temps de transport et de traitement par le laboratoire impose généralement des délais supplémentaires avant l'obtention des résultats de mesure. Elle ne permettent pas non plus de quantifier l'incertitude d'échantillonnage, qui peut être significative. Enfin ces mesures sont généralement plus performantes mais plus coûteuses que les mesures *in-situ*.

I.3.4 Le traitement statistique des informations obtenues par les résultat de mesure

L'ensemble des résultats de mesure ou encore observations constitue un jeu de données. Une fois ces données obtenues avec l'échantillonnage et les résultats de mesure, une première analyse statistique peut être réalisée. Le choix de l'outil statistique va dépendre de l'objectif de caractérisation, de la stratégie d'échantillonnage ainsi que des caractéristiques

des données obtenues. Par exemple un jeu de données présentant des corrélations spatiales pourra être analysé à l'aide de la géostatistique. Cette approche permettra par exemple de réaliser une cartographie radiologique à partir des observations.

L'utilisation de la géostatistique est relativement récente et a fait l'objet de plusieurs études afin de vérifier son utilité pour les projets de démantèlement. Le rapport [EPRI, 2016] donne des recommandations générales pour son utilisation dans l'A&D de sites nucléaires, avec de nombreux cas d'application. Il présente également de nombreux logiciels actuellement disponibles pour une application industrielle. Les travaux de [Desnoyers, 2010] donnent également plusieurs applications et retours d'expérience sur la géostatistique dans ce contexte. Enfin hors du cadre des sites nucléaires, le rapport [Belbeze et al., 2013] fait un retour d'expériences sur l'utilisation de la géostatistique pour l'assainissement de sites et sols pollués (contamination en hydrocarbures et métaux lourds).

L'un des intérêts de ces approches est de pouvoir quantifier les incertitudes associées à certaines estimations. Prenons l'exemple d'un lot de déchets dont on souhaite déterminer la catégorie (par exemple déchets haute activité ou moyenne activité). Si l'incertitude dans la catégorisation de ce lot est trop grande, il est alors possible (si le contexte du démantèlement le permet) de réaliser des mesures supplémentaires jusqu'à obtenir une incertitude satisfaisante sur le choix de la catégorie. Cela permet de mettre au point des approches itératives pour optimiser le nombre de mesures et garantir une certaine qualité des estimations.

Différents outils statistiques utiles dans l'A&D seront présentés dans le chapitre 1 de ce manuscrit.

I.3.5 Evaluation techno-économique et contraintes de démantèlement

Comme évoqué dans la section I.3.1, la caractérisation radiologique passe par une étape de collecte d'informations. L'analyse fonctionnelle et historique permet d'obtenir des informations avec des coûts très faibles. Cependant la réalisation des mesures représente un coût non négligeable. Ce coût est d'autant plus grand que les résultats de mesure obtenus sont exacts. La Figure 8 vient illustrer cette relation entre qualité, quantité et coût d'obtention des informations.

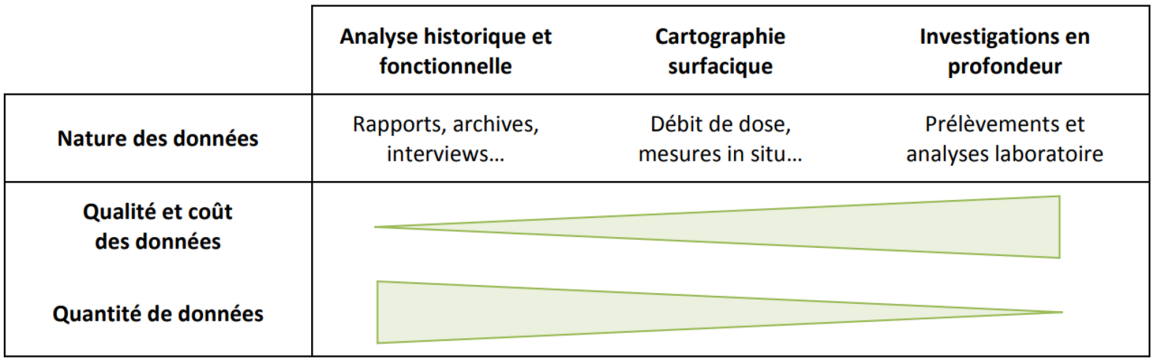


FIGURE 8 – Coût de l'information et qualité des données [Granier et al., 2017].

Il est donc nécessaire de trouver un équilibre entre les objectifs de caractérisation et les coûts imposés par la réalisation de cette collecte d'informations. Il est souvent judicieux d'investir suffisamment de temps et de moyens dans une collecte d'informations pour garantir le bon déroulement des étapes de démantèlement, plutôt que de réaliser une caractérisation rapide entraînant une mauvaise catégorisation des zones d'intérêt et donc d'éventuels retards et coûts supplémentaires.

Dans le cas d'une installation nucléaire de base, il existe plusieurs contraintes spécifiques supplémentaires. Aux risques classiques d'un chantier de démantèlement s'ajoutent des risques liés à la radioprotection. Par exemple, les niveaux radiologiques sur certains sites sont faibles et permettent aux opérateurs de se déplacer librement, tandis que sur d'autres, ils sont particulièrement élevés et empêchent l'intervention prolongée d'opérateurs. Il est parfois nécessaire de faire intervenir des robots si les opérateurs ne peuvent tout simplement pas accéder au site. Des méthodes permettant la réalisation de mesures sous contraintes sont données dans [Amgarou, 2017]. Ces difficultés supplémentaires ont pour effet d'augmenter le coût et le temps nécessaire à la réalisation de ces mesures. Une conséquence directe de l'augmentation de ces coûts est la diminution du nombre de données disponibles.

I.4 Synthèse et problématiques du contexte du sujet de thèse

Le contexte de l'A&D de sites nucléaires est donc singulier, en particulier en ce qui concerne l'application d'outils statistiques pour caractériser les sites en question. De manière générale, il est possible de n'avoir accès qu'à un nombre réduit d'observations. Pour les analyses statistiques, ce fait est particulièrement problématique puisque peu de méthodes statistiques classiques sont adaptées au traitement d'un faible nombre d'observations.

Pour la plupart des projets d'A&D, des mesures *in-situ* sont d'abord réalisées pour une première caractérisation radiologique. De part la nature de la grandeur physique mesurée et du protocole de mesure employé, des incertitudes de mesure élevées peuvent être obtenues. De plus, dans la plupart des sites nucléaires, si certaines zones peuvent être hautement contaminées, d'autres peuvent être à des niveaux de contamination très faibles. Ces deux faits conduisent à des censures des observations, qui posent à nouveau problème lors de l'analyse statistique, la plupart des méthodes usuelles (en particulier en géostatistique) ne les prenant pas en compte.

Dans cette thèse nous chercherons donc à tester et proposer des méthodes pour le traitement de ces deux problématiques, en particulier en géostatistique.

I.5 Plan de la thèse et contributions aux problématiques

Le premier chapitre donne plusieurs rappels et définitions de statistique et de géostatistique importantes pour la compréhension des différentes méthodes présentées. Ce chapitre fait donc figure d'un état de l'art des méthodes classiques utilisées en A&D de sites nucléaires. Les méthodes de géostatistique et ses outils associés comme le variogramme (pour caractériser la corrélation spatiale des données) et le krigeage (pour les prédictions) seront détaillés dans ce chapitre.

Le chapitre suivant s'intéresse à une technique utile dans le contexte de l'A&D : le krigeage bayésien. Ce krigeage particulier est une méthode robuste face à la problématique d'un faible nombre d'observations. Une nouvelle méthode adaptée à la prise en compte d'un effet de pépite inconnu (paramètre des modèles statistiques permettant de quantifier certaines incertitudes) dans la résolution des équations du krigeage bayésien est présentée. Cette méthode y est décrite en détails, avec ses avantages et inconvénients. Ce krigeage est ensuite comparé à une méthode usuelle, le krigeage ordinaire. Cette comparaison permet de montrer l'intérêt du premier pour l'A&D et ses applications industrielles. Le krigeage bayésien ayant une formulation particulière, certains critères de validation usuels des modèles statistiques ne sont plus adaptés à l'évaluation de ses performances. De nouveaux critères de validation sont alors introduits dans ce chapitre pour quantifier ces performances. Cette comparaison et ces nouveaux critères ont fait l'objet d'un article soumis [Wieskotten et al., 2022] donné en annexe B.1.

Enfin le troisième chapitre s'attache à la présentation d'outils pour le traitement de données censurées, notamment l'algorithme EM (Espérance Maximisation) et sa version bayésienne, l'augmentation de données. Ce type de données étant très présent dans les chantiers d'A&D, l'application de ces méthodes permettrait une meilleure caractérisation de la contamination. Nous proposons ici également un nouvel algorithme de traitement des données censurées avec la prise en compte d'un effet de pépite, qui combine l'approche classique de l'augmentation de données avec la méthode de résolution des équations du krigeage bayésien proposée dans le chapitre 2. En plus de cette méthode, nous proposons également une méthode basée sur une approche Monte-Carlo pour le traitement des données censurées. Des comparaisons entre méthodes existantes et méthodes proposées (appliquées à des simulations et un jeu de données réelles) sont également présentées dans ce chapitre.

Un dernier chapitre conclut ce mémoire de thèse et fournit quelques perspectives et axes de recherche mis en évidence par les travaux présentés dans ce document. Les annexes fournissent quelques résultats complémentaires, la liste des communications effectuées durant la thèse et ainsi que l'article soumis émanant du chapitre deux.

CHAPITRE 1

Inférence et outils statistiques classiques

L'inférence statistique consiste à déduire de quelques observations les caractéristiques de l'ensemble dont sont extraites les observations. Dans notre cas, cela consiste à déduire à partir des résultats de mesures le comportement de la grandeur physique qui nous intéresse. Il est alors également possible *a posteriori* d'estimer plusieurs informations comme la probabilité de dépassement de seuil, le volume contaminé ou irradié, etc.

L'objectif de ce chapitre est de fournir une liste non-exhaustive des différents outils statistiques utiles pour l'A&D de sites nucléaires. Pour cela nous commencerons par faire des rappels sur des définitions importantes pour la compréhension de ces outils. Nous donnerons quelques outils généraux de statistique classique, puis une partie importante de ce chapitre sera dédiée à présenter la méthodologie de géostatistique, avec en particulier le krigeage qui sera le principal outil étudié de cette thèse.

1.1 Variables aléatoires et loi de probabilités

Dans cette partie, l'objectif est de modéliser la grandeur physique étudiée par une variable aléatoire réelle. Cette grandeur peut correspondre à une mesure de radioactivité (comme expliqué dans l'introduction), mais également à toute autre grandeur physique d'intérêt pour l'A&D (comme des concentrations en métaux lourds ou en radionucléides) et sera notée Z. Sa réalisation sera quant à elle notée z.

Chaque variable aléatoire est caractérisée par sa fonction de répartition.

Fonction de répartition [Saporta, 1990] :
La fonction de répartition d'une variable aléatoire Z est l'application F de $\mathbb{R}$ dans $[0, 1]$ définie par :

$$F(z) = \mathbb{P}(Z \leq z).$$

On considère en particulier le cas d'une variable aléatoire à densité.

Variable aléatoire à densité réelle unidimensionnelle [Saporta, 1990] :
Une variable aléatoire ou sa loi est dite à densité si il existe f_Z telle que pour tout intervalle $I \in \mathbb{R}$,

$$\mathbb{P}(Z \in I) = \int_I f_Z(z)dz.$$

Pour les variables aléatoires réelles à densité, la fonction de répartition et la densité de probabilité sont liées par la relation suivante :

$$\mathbb{P}(a < Z < b) = \int_a^b f_Z(z)dz = F(b) - F(a).$$

Dans la suite et sauf mention contraire, on considérera que les observations sont des réalisations des variables aléatoires $Z_1, \ldots, Z_n$ indépendantes et identiquement distribuées (i.i.d.) de même loi que Z et que z_i est la réalisation de Z_i.

1.2 Modélisation de la densité de probabilité

Pour poursuivre dans l'analyse statistique, il est alors nécessaire d'identifier et de modéliser la densité de probabilité de la variable aléatoire étudiée. Cette identification passe souvent par une modélisation de la densité. Deux catégories de méthodes pour la modélisation de densité sont alors considérées :

- les méthodes paramétriques qui considèrent que la densité de probabilité appartient à une famille de lois paramétrées ;
- les méthodes non-paramétriques qui ne font pas appel à des familles de densités paramétrées, et qui sont plus flexibles.

L'approche paramétrique utilise des familles de lois paramétrées afin de modéliser la loi de probabilité de la variable aléatoire. Cette approche se déroule en plusieurs étapes : choix de la famille paramétrique (loi gaussienne, gamma, etc.), puis estimation des paramètres de cette loi et validation ou test d'adéquation de la loi obtenue aux données. Les méthodes non-paramétriques pour l'expression de la densité de probabilité permettent de gagner en flexibilité dans la modélisation et ne demandent que peu d'hypothèses sur la densité de probabilité. Elles nécessitent cependant souvent davantage de données et sont très dépendantes de l'échantillon étudié. Dans la suite du document, nous ne considérerons que les méthodes paramétriques car elles offrent de nombreux outils théoriques facilement manipulables. Plus de détails sur les méthodes non-paramétriques peuvent être trouvés dans [Parzen, 1962].

1.2.1 Choix de la famille paramétrée

La sélection de la densité paramétrique peut faire appel à l'histogramme, dont la forme donne une indication sur la loi de probabilité. Néanmoins dans le cas où peu de données sont disponibles, l'histogramme ne donne pas une bonne « image » de la densité recherchée. La Figure 1.1 donne un exemple de deux histogrammes représentant deux échantillons de taille différente, tous deux issus d'une loi gaussienne centrée réduite. Le graphe en haut

contient trente données et ne donne pas une image « nette » de la loi, contrairement au graphe en bas dont l'histogramme est bien plus proche de la densité gaussienne.

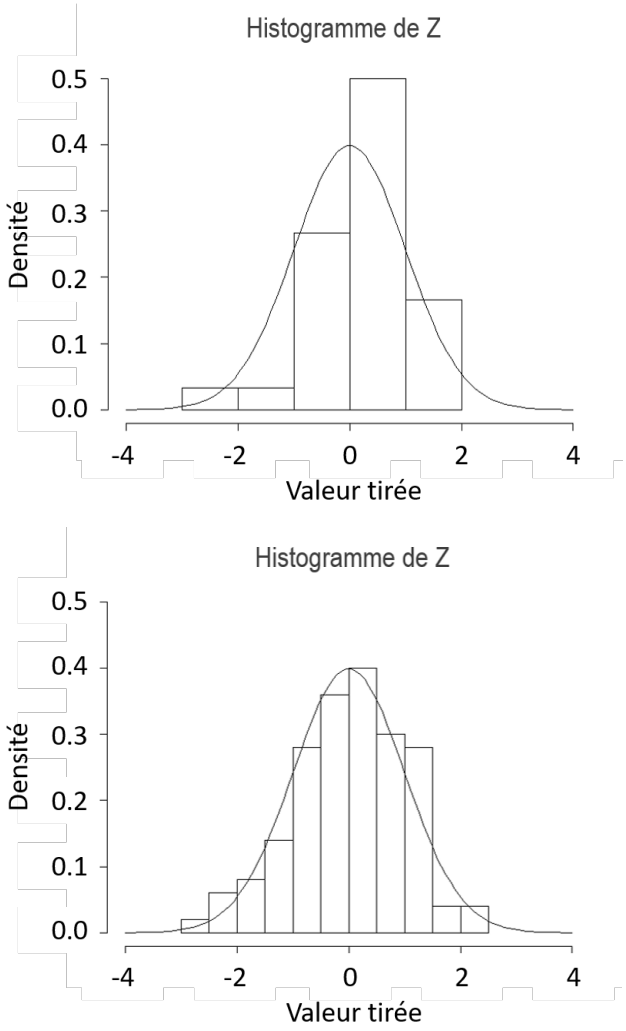


FIGURE 1.1 – Deux échantillons issus de la loi $\mathcal{N}(0,1)$ de taille respective 30 et 100.

Il est également possible de faire cette sélection à l'aide de connaissances *a priori* sur le phénomène, comme par exemple la durée de vie d'un objet qui est réputée être bien représentée par une loi exponentielle ou de Weibull.

Plusieurs familles de lois peuvent être proches de la loi de la variable aléatoire. Les tests statistiques ainsi que les critères AIC (Akaike Information Criterion pour critère d'information d'Akaike) et BIC (Bayesian Information Criterion pour critère d'information bayésien), détaillés dans la suite, permettent de choisir et comparer les différentes familles possibles. Une fois la famille de loi sélectionnée, l'étape suivante consiste à en estimer les paramètres.

1.2.2 Maximum de vraisemblance

Le maximum de vraisemblance est une méthode très répandue en statistique pour l'estimation des paramètres d'un modèle. Il permet d'estimer les paramètres en maximisant la vraisemblance d'observer les données conditionnellement à la loi et aux paramètres choisis.

Estimation du maximum de vraisemblance [Saporta, 1990] :
Soit un échantillon $\boldsymbol{z} = (z_1, \dots, z_n)'$ de taille n issu d'une variable aléatoire de densité f_Z paramétrée par θ. La vraisemblance est définie par :

$$L(\boldsymbol{z}; \theta) = \prod_{i=1}^{n} f_Z(z_i; \theta).$$

L'estimation $\hat{\theta}^*$ par maximum de vraisemblance est alors obtenue en maximisant la vraisemblance :

$$\hat{\theta}^* = \arg\max_{\theta} L(\boldsymbol{z}; \theta),$$

avec $\arg\max_\theta$ l'argument du maximum correspondant à l'ensemble des θ pour lesquels $L(\boldsymbol{z}; \theta)$ atteint sa valeur maximale. Il est important de noter que cette estimation n'existe pas toujours et n'est pas forcément unique.

Dans le cas où la vraisemblance est dérivable et explicite, il est possible de calculer analytiquement l'estimation du maximum de vraisemblance à partir de l'équation dite équation de la vraisemblance :

$$\frac{\partial L(\boldsymbol{z}; \theta)}{\partial \theta} = 0.$$

Pour résoudre cette équation, il est possible de faire appel à des méthodes numériques d'optimisation.

En substituant les réalisations $z_1, \dots, z_n$ par les variables aléatoires $Z_1, \dots, Z_n$, la vraisemblance $L(Z_1, \dots, Z_n; \theta)$ devient aléatoire. On obtient alors l'estimateur du maximum de vraisemblance $\hat{\theta}$:

$$\hat{\theta} = \arg\max_{\theta} L(\boldsymbol{z}; \theta).$$

Il est assez simple de généraliser le maximum de vraisemblance au cas de plusieurs paramètres. En notant $\boldsymbol{\theta}$ le vecteur des paramètres $\theta_1, \dots, \theta_k$, il suffit alors de résoudre le système :

$$\frac{\partial L(\boldsymbol{z}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}$$

avec $\mathbf{0}$ le vecteur colonne nul de taille k.

Une des limites du maximum de vraisemblance est qu'il peut être biaisé pour un faible nombre de données et que la plupart de ses propriétés intéressantes sont vérifiées pour un grand nombre de données, comme expliqué dans [Saporta, 1990]. Néanmoins, cet outil est utilisé dans la plupart des méthodes qui seront présentées plus loin. Une version applicable aux champs aléatoires gaussiens existe et est présentée dans [Kitanidis and Lane, 1985]. Elle sera décrite dans la section 1.9.4.1 de ce document.

Il est souvent utile de travailler avec la log-vraisemblance qui transforme le produit en une somme pour faciliter le problème d'optimisation ou l'expression de la vraisemblance :

$$\hat{\boldsymbol{\theta}}^* = \arg\max(\ln L(\boldsymbol{z};\boldsymbol{\theta})) = \arg\max_{\boldsymbol{\theta}} \left(\sum_{i=1}^{n} \ln f_Z(z_i;\boldsymbol{\theta}) \right).$$

1.2.3 Tests statistiques

De nombreux tests statistiques sont disponibles pour étudier différentes hypothèses, comme la valeur d'un paramètre ou le choix de la famille de densité. Le principe classique d'un test est de confronter deux hypothèses H_0 et H_1. Seule l'une de ces deux hypothèses est vraie. Selon la décision prise, quatre scénarios sont possibles. Le Tableau 1.1 résume ces différents scénarios selon leur probabilité respective d'occurrence.

Décision \ Vérité	H_0	H_1
H_0	$1-P_1$	P_2
H_1	P_1	$1-P_2$

TABLE 1.1 – Probabilités des risques d'un test statistique.

Les valeurs P_1 et P_2 correspondent respectivement à la probabilité d'erreur de première espèce (rejeter H_0 alors que H_0 est vraie) et celle d'erreur de seconde espèce (ne pas rejeter H_0 alors que H_0 est fausse). Pour orienter la prise de décision, les tests utilisent une variable de décision (ou statistique de test) dont la loi est connue. Le protocole d'un test statistique classique est le suivant :

- choix du risque P_1 (généralement égal à 5%) sous l'hypothèse H_0 et définition de la région critique ;
- calcul de la statistique de test ;
- vérification de la présence ou non de la statistique de test dans la région critique.

Si lors de la vérification, la statistique de test appartient à la région critique, on rejette alors l'hypothèse H_0 avec un risque P_1 de rejeter à tort cette hypothèse. Dans le cas contraire, on ne peut rejeter l'hypothèse H_0. De nombreux tests existent pour vérifier une variété d'hypothèses différentes. Voici quelques exemples de tests classiques :

- le test de Kolmogorov-Smirnov qui vérifie si un échantillon suit une certaine loi en comparant la fonction de répartition empirique avec celle de la loi testée ;
- le test du χ^2 qui permet de vérifier si un échantillon suit une certaine loi (test d'adéquation) en comparant les effectifs de l'échantillon (regroupés par classes) avec ceux issus de la loi testée ;
- le test de Shapiro-Wilk qui teste si un échantillon est bien issu d'une loi gaussienne en comparant les quantiles empiriques à ceux d'une loi gaussienne. Il est particulièrement efficace dans le cas d'un petit échantillon.

Plus d'informations sur les tests statistiques peuvent être trouvées dans [Saporta, 1990].

1.2.4 Critères d'information

Comme les tests ne permettent pas de valider un modèle mais seulement de le rejeter, d'autres critères sont nécessaires pour sélectionner le modèle parmi ceux non rejetés. Deux critères basés sur la vraisemblance sont disponibles : les critères AIC et BIC.

Soit k le nombre de paramètres de la famille de lois paramétrées et $L(\boldsymbol{z};\boldsymbol{\theta})$ la vraisemblance.
Aikake Information Criterion (AIC) [Akaike, 1998] :
L'AIC est égal à :

$$AIC = -2\ln\left(L(\boldsymbol{z};\boldsymbol{\theta})\right) + 2k.$$

Bayesian Information Criterion (BIC) [Schwarz, 1978] :
Le BIC est égal à :

$$BIC = -2\ln\left(L(\boldsymbol{z};\boldsymbol{\theta})\right) + k\ln(n).$$

où l'on rappelle que n est la taille de l'échantillon.

Ces deux critères quantifient l'équilibre entre ajustement du modèle et le nombre de paramètres du modèle. Ils appliquent ainsi le principe de parcimonie visant à limiter la complexité des modèles. Le BIC a la particularité de prendre en compte le nombre de données dans son calcul et de fonctionner dans le domaine asymptotique [Schwarz, 1978]. Il est donc préférable de privilégier l'AIC lorsque peu de données sont disponibles. Le meilleur modèle au sens du critère considéré est celui qui le minimise.

1.3 Inférence bayésienne

L'inférence bayésienne fait appel au théorème de Bayes ([Bayes, 1763]) pour intégrer des connaissances *a priori* dans la modélisation. Un des points forts de cette approche est qu'elle permet d'estimer des incertitudes sur l'estimation du ou des paramètres grâce à la loi *a posteriori*.

1.3.1 Estimation bayésienne

Pour estimer le paramètre θ, l'inférence bayésienne le considère comme une variable aléatoire. Si θ est une variable aléatoire à valeurs dans Θ de densité $\pi(\theta)$, alors la densité de probabilité de θ sachant z est :

$$f_\theta(\theta|\boldsymbol{z}) = \frac{L(\boldsymbol{z};\theta)\pi(\theta)}{\int_\Theta L(\boldsymbol{z};\theta) f_\theta(\theta) d\theta}.$$

$\pi(\theta)$ est appelée la densité de la loi *a priori* et $f_\theta(\theta|\boldsymbol{z})$ est appelée la densité de la loi *a posteriori*.

La distribution $f_\theta(\theta|\boldsymbol{z})$ est souvent réécrite de la manière suivante :

$$f_\theta(\theta|\boldsymbol{z}) \propto L(\boldsymbol{z};\theta)\pi(\theta)$$

car la valeur de $\int_\Theta L(\boldsymbol{z};\theta)\pi(\theta)d\theta$ ne dépend pas de θ et est donc une constante de normalisation.

Le calcul de loi *a posteriori* est parfois impossible, puisque la constante de normalisation n'est pas forcément connue et est parfois très difficile à estimer. Des solutions à ce problème sont données dans la section 1.3.3 avec les lois *a priori* conjuguées. Une fois la loi *a posteriori* obtenue, le paramètre θ peut être estimé de plusieurs manières, par exemple en prenant le mode ou l'espérance de la loi *a posteriori* (ou encore l'estimation par maximum de vraisemblance). Des intervalles de confiance sur l'estimation de θ peuvent également être obtenus à partir de $f_\theta(\theta|\boldsymbol{z})$.

1.3.2 Choix de la loi *a priori*

Le choix de la loi *a priori* est un élément crucial de l'approche bayésienne, puisqu'il modifie directement le calcul de la loi *a posteriori*. Le choix de la loi *a priori* étant une décision prise indépendamment des données, il s'agit d'une décision (potentiellement arbitraire) du statisticien. Il existe plusieurs approches pour faire ce choix :

- le statisticien possède une intuition ou un savoir d'expert (émanant souvent d'experts du domaine « métier ») sur le paramètre à estimer et choisit une loi *a priori* représentant ses croyances sur le paramètre (il s'agit d'une loi *a priori* subjective) ;
- le statisticien souhaite des calculs qui soient possibles et simples à exécuter (le choix de la loi *a priori* est alors une décision pragmatique) ;
- le statisticien souhaite apporter peu ou pas d'informations avec le choix de la loi *a priori* (il s'agit d'une loi *a priori* objective ou non-informative).

1.3.3 Les lois conjuguées

Pour faciliter les calculs et obtenir une forme analytique pour $f_\theta(\theta|\boldsymbol{z})$, il est possible de faire appel aux lois *a priori* conjuguées.

Loi *a priori* conjuguée :
Une loi *a priori* est dite conjuguée si la loi *a posteriori* appartient à la même famille de lois paramétrées que la loi *a priori*.

Elles permettent donc de rendre certains calculs possibles analytiquement et facilitent l'intégration de nombreux paramètres dans le modèle. Un exemple classique est celui d'une variable aléatoire Z suivant une loi gaussienne et dont la moyenne à estimer suit une loi *a priori* également gaussienne. La loi *a posteriori* suit alors également une loi gaussienne :

$$\text{Si } Z \sim \mathcal{N}(\theta, \sigma^2) \text{ et si } \theta \sim \mathcal{N}(\mu, \tau^2) \text{ alors } \theta|\theta^* \sim \mathcal{N}\left(\frac{\theta^*}{\sigma^2} + \frac{\mu}{\tau^2}, \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right).$$

où θ^* est une estimation de la moyenne θ. Des exemples et détails supplémentaires sur les lois conjuguées peuvent être trouvés dans [Berger, 1985].

1.4 Méthodes de ré-échantillonnage

Les méthodes de ré-échantillonnage ont été introduites afin d'estimer la qualité d'un estimateur en calculant son biais, sa variance et des intervalles de confiance. Deux méthodes de ré-échantillonnage seront décrites ici : le jackknife et le bootstrap.

1.4.1 Jackknife

Méthode jackknife [Saporta, 1990] :
Soit $\hat{\theta}$ un estimateur calculé avec l'échantillon $Z_1, \ldots, Z_n$ et $\hat{\theta}_{-i}$ l'estimateur calculé à partir de l'échantillon $Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n$ de taille $n-1$. Ainsi $\hat{\theta}$ et $\hat{\theta}_{-i}$ correspondent au même estimateur, le premier étant calculé sur n observations tandis que le second est estimé sur $n-1$ observations. L'estimateur jackknife est alors :

$$\hat{\theta}_J = \frac{1}{n}\sum_{i=1}^{n}\left(n\hat{\theta} - (n-1)\hat{\theta}_{-i}\right).$$

L'estimateur de sa variance $\hat{S}_J^2$ est égal à :

$$\hat{S}_J^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{(n\hat{\theta} - (n-1)\hat{\theta}_{-i} - \hat{\theta}_J)^2}{n-1}.$$

Le jackknife consiste donc à ré-échantillonner sans remise $n-1$ valeurs. Le principal intérêt du jackknife est de réduire le biais de certaines statistiques. Cela est particulièrement important lorsque ce biais est difficile à estimer (comme pour le coefficient de corrélation par exemple), d'après [Saporta, 1990].

1.4.2 Bootstrap

Le bootstrap consiste à échantillonner avec remise l'échantillon initial et permet d'estimer des distributions d'estimateurs d'intérêt. L'estimateur d'intérêt est noté $\hat{\theta}$. Un échantillon bootstrap de taille n, noté $\boldsymbol{Z}^{bs} = (Z_1^{bs}, \ldots, Z_n^{bs})$, est tiré aléatoirement avec remise de l'échantillon. $\hat{\theta}$ est calculé à partir de ce nouvel échantillon et donne l'estimation bootstrap $\hat{\theta}^{bs,*}$. Ce tirage est répété B fois (de manière générale très grand). A partir des B valeurs de $\hat{\theta}_1^{bs,*}, \ldots, \hat{\theta}_B^{bs,*}$ est obtenue la distribution de $\hat{\theta}^{bs}$. Le biais, la variance et l'intervalle de confiance à 95% de l'estimateur sont obtenus à partir de cette distribution :

- le biais de l'estimateur : $b = \mathbb{E}[\hat{\theta}^{bs} - \hat{\theta}]$,
- la variance de l'erreur d'estimation : $\sigma_{err}^2 = \mathbb{E}[(\hat{\theta}^{bs} - \hat{\theta})^2]$,
- l'intervalle de confiance à 95% sur θ : $I_c = \left[\hat{\theta}^{bs} - q_{2.5\%}(\hat{\theta}^{bs} - \hat{\theta}), \hat{\theta}^{bs} + q_{97.5\%}(\hat{\theta}^{bs} - \hat{\theta})\right]$ avec $q_{2.5\%}$ et $q_{97.5\%}$ les quantiles d'ordre 2.5% et 97.5% de la distribution de $\hat{\theta}^{bs} - \hat{\theta}$.

Davantage de détails sur le bootstrap peuvent être trouvés dans [Efron, 1979]. Le bootstrap présenté ici est une méthode simple à implémenter et non-paramétrique, ce qui en fait un outil très général. Une application intéressante du bootstrap est proposée par [Pérot and Iooss, 2008] pour vérifier la représentativité d'un échantillon. Soit un échantillon $\boldsymbol{Z}$ de taille n et une statistique θ d'intérêt. L'algorithme proposé est le suivant :

— choix d'une taille initiale $n_0 < n$ et d'un incrément k ;
— tirage aléatoire sans remise dans $\boldsymbol{Z}$ d'un échantillon $\boldsymbol{Z}^{bs}_{(0)}$ de taille n_0,
— puis pour $j \in [\![1, \frac{(n-n_0)}{k}]\!]$;

1. construction de l'échantillon $\boldsymbol{z}^{bs}_{(j)}$ de taille $n_0 + kj$ en tirant sans remise k valeurs parmi les données restantes, soit les données n'appartenant pas à $\boldsymbol{Z}^{bs}_{(j-1)}$;
2. estimation de θ et de son intervalle bootstrap à 95% pour chaque $\boldsymbol{Z}^{bs}_{(j)}$.

Il est ainsi possible de vérifier graphiquement l'évolution des estimations de θ selon le nombre de données utilisées. Cela permet de vérifier l'intérêt d'acquérir des données supplémentaires ou de juger de la représentativité de l'échantillon, dont le graphe convergera.

1.5 Estimation robuste de probabilités d'évènements extrêmes

Dans les travaux d'analyse de risque, il est souvent nécessaire de calculer des quantiles pour identifier la probabilité d'un évènement extrême (par exemple un niveau de radioactivité mesuré dépassant un seuil réglementaire dans le cas l'A&D). Ce calcul se met sous la forme suivante :

$$\mathbb{P}(Z > s) = P_s. \tag{1.1}$$

Le problème peut ici être vu de deux manières. On peut chercher la probabilité P_s pour un seuil s connu, par exemple la probabilité que la radioactivité d'un fût de déchets dépasse un seuil fixé par la législation (ce qui revient à estimer $1 - P_s$). On peut également chercher le seuil s pour lequel la probabilité est égale à P_s (ce qui revient à estimer le quantile d'ordre $1 - P_s$). Prenons le cas d'une estimation de probabilité de dépassement d'un seuil réglementaire. Cette estimation est généralement complexe lorsque la valeur s correspond à une valeur extrême, puisque qu'obtenir des valeurs supérieures ou égales à s sont peu probables, rendant l'estimation peu robuste. Ce problème est rendu particulièrement difficile lorsque le nombre de données disponibles est faible. Deux exemples présentés ici permettent de répondre à cette problématique : les inégalités robustes et la méthode de Wilks.

1.5.1 Inégalités robustes

Dans le cas d'un échantillon de petite taille, l'inégalité (1.1) doit être robuste pour permettre de réaliser des estimations réalistes sur les quantiles ou les ordres des quantiles d'intérêt. Il existe plusieurs inégalités de ce type dans la littérature, comme l'inégalité de Markov ou celle de Bienaymé-Tchebychev ([Savage, 1961]).

Inégalité de Bienaymé-Tchebychev :
Soit Z une variable aléatoire d'espérance $\mathbb{E}[Z]$ et de variance $Var[Z]$. Alors pour tout $a \in \mathbb{R}^{+*}$:

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| \geq a\right) \leq \frac{Var[Z]}{a^2}$$

Dans le cas de ces deux inégalités, aucune hypothèse n'est nécessaire sur la loi (mis à part l'existence des moments mis en jeu dans les inégalités). Néanmoins elles donnent des résultats peu fiables car elles surestiment le quantile réel. D'autres inégalités avec des hypothèses assez classiques comme l'unimodalité ou la convexité de la queue de distribution de Z permettent de produire des résultats plus fiables. Des exemples et applications de ces inégalités peuvent être trouvés dans [Savage, 1961] et [Pérot et al., 2017].

Le calcul des inégalités robustes nécessite la connaissance des moments de la variable aléatoire. Il est possible de remplacer ces moments par leur estimation empirique ; néanmoins, dans le cas de l'estimation d'une probabilité de dépassement de seuil, [Blatman et al., 2017] montre expérimentalement que ces estimations donnent des résultats sous-estimant les probabilités de dépassement de seuil. La solution employée est la méthode bootstrap pour l'estimation des moments. Pour chaque estimation bootstrap, les moments sont estimés et la probabilité de dépassement est calculée. L'estimation est ensuite « pénalisée » en ne conservant que les valeurs supérieures à un quantile d'ordre élevé (ordre arbitraire). Les estimations de la probabilité ainsi obtenues sont très conservatives et ne sous-estiment donc plus le risque réel.

1.5.2 Méthode de Wilks

La méthode de Wilks utilise les statistiques d'ordre pour estimer le quantile q_P d'ordre P d'une variable aléatoire avec un niveau de confiance ζ. Pour un échantillon donné de taille n, l'objectif est de calculer la probabilité qu'au moins m éléments soient plus grands que q_P. Par définition et indépendance :

$$\begin{cases} \mathbb{P}(Z \leq q_P) = P \\ \mathbb{P}(Z_1 \leq q_P, \dots, Z_n \leq q_P) = P^n. \end{cases}$$

La probabilité qu'au moins une observation soit plus grande que q_P est alors :

$$\zeta = 1 - P^n.$$

La probabilité d'observer au moins deux valeurs supérieurs à q_P est alors :

$$\zeta = 1 - P^n - nP^{n-1}(1 - P).$$

De manière plus générale, la formule de Wilks correspond à une loi binomiale de paramètres P et n, comptant le nombre d'observations supérieures à un certain quantile :

Formule de Wilks [Wilks, 1941] :
La probabilité d'observer m éléments supérieurs à q_P est :

$$\zeta = 1 - \sum_{j=n-m+1}^{n} \binom{n}{j} P^j (1 - P)^{n-j}.$$

Cette formule peut être vue de plusieurs manières. Si l'on considère que P et ζ sont connus, elle permet alors d'estimer le nombre d'observations nécessaires pour obtenir m majorants de l'ensemble des valeurs inférieures au quantile d'ordre P avec une probabilité

ζ. En fixant n et P, il est possible de calculer la probabilité d'obtenir un majorant du quantile d'ordre P. Un seul majorant est souvent considéré, mais l'utilisation d'un nombre plus élevé peut permettre de se rapprocher du quantile réel en prenant le minorant des majorants (néanmoins cela implique de réaliser davantage d'observations). La formule de Wilks étant indépendante de la loi de Z, il est possible d'en déduire un tableau général pour des P, ζ et n variés. Le Tableau 1.2 donne des exemples classiques du triplet (P, ζ, n) pour un majorant $(m = 1)$.

P	0.99	0.95	0.95	0.90	0.90	0.70
ζ	0.99	0.95	0.90	0.95	0.90	0.70
n	459	59	45	29	22	4

TABLE 1.2 – Exemples de valeurs pour la formule de Wilks au premier ordre $(m = 1)$.

Plus d'explications sur la méthode peuvent être trouvées dans [Nutt and Wallis, 2004]. Cette méthode est robuste et nécessite peu d'hypothèses, mais elle donne une surestimation du vrai quantile. Cette surestimation est souvent recherchée en analyse de risque, mais elle peut impliquer un surdimensionnement des solutions au problème étudié si le quantile est pris comme seul indicateur. Des exemples d'applications numériques peuvent être trouvés dans [Pérot et al., 2017] et [Blatman et al., 2017].

1.6 Régression linéaire

Lorsque l'on se place dans un cas multivariable, par exemple lorsque des résultats de mesure correspondent à une activité surfacique tandis que d'autres correspondent à une activité massique (radioactivité mesurée par unité de masse), il peut être intéressant d'exprimer certaines variables (principales) en fonction d'autres (secondaires ou explicatives). Il est dans ce cas possible de faire appel à la régression linéaire pour quantifier les corrélations entre les différentes variables. S'il est validé, le modèle de régression linéaire pourra ensuite être utilisé à des fins prédictives, la relation entre variable(s) étudiée(s) et la ou les variable(s) explicative(s) étant connue. C'est notamment l'outil employé pour la méthode ROS (Regression on Order Statistics pour régression sur statistiques d'ordre) présentée dans la section 3.3.4.

1.6.1 Modèle de régression linéaire multivariable

Régression linéaire [Kleijnen, 2015] :

Soient $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$ les variables aléatoires i.i.d, $\boldsymbol{D} = \begin{pmatrix} 1 & d_{1,1} & \ldots & d_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & d_{n,1} & \ldots & d_{n,p} \end{pmatrix}$ la matrice des p variables explicatives déterministes (aussi appelée matrice de design). Le modèle de régression linéaire classique s'écrit alors :

$$\boldsymbol{Z} = \boldsymbol{D}\boldsymbol{\mu} + \boldsymbol{\epsilon}$$

avec $\boldsymbol{\mu}$ le vecteur des paramètres et $\boldsymbol{\epsilon}$ le vecteur des résidus vérifiant :

$$\forall i, j \in [\![1, n]\!] : \begin{cases} \mathbb{E}[\epsilon_i] = 0, \\ Var[\epsilon_i] = \sigma^2, \\ \forall i \neq j, Cov(\epsilon_i, \epsilon_j) = 0. \end{cases}$$

L'estimateur des moindres carrés est alors le meilleur estimateur linéaire non-biaisé de $\boldsymbol{\mu}$ (au sens de la variance d'estimation la plus faible parmi l'ensemble des estimateurs linéaires non-biaisés). Cet estimateur $\hat{\boldsymbol{\mu}}$ est :

$$\hat{\boldsymbol{\mu}} = (\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{Z}.$$

L'écriture de l'estimateur des moindres carrés suppose que la matrice $\boldsymbol{D}'\boldsymbol{D}$ est inversible, propriété qui peut être vérifiée à l'aide d'un plan d'expérience adapté. Des informations sur les plans d'expérience peuvent être trouvés dans [Kleijnen, 2015]. Il existe des extensions de la régression linéaire dans le cas où les résidus ne vérifient pas les hypothèses données précédemment. Par exemple, il est possible que les résidus soient corrélés. L'estimateur des moindres carrés généralisés est alors employé :

Moindres carrés généralisés [Kleijnen, 2015] :
Le modèle précédent est modifié de telle sorte à autoriser des corrélations entre variables aléatoires. On note la matrice de covariance $\boldsymbol{R} = (Cov(\epsilon_i, \epsilon_j))_{1 \leq i,j \leq n}$. Le modèle vérifie l'hypothèse suivante :

$$\forall i \in [\![1, n]\!], \mathbb{E}[\epsilon_i] = 0.$$

L'estimateur des moindres carrés généralisés est alors :

$$\hat{\boldsymbol{\mu}} = (\boldsymbol{D}'\boldsymbol{R}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{R}^{-1}\boldsymbol{Z}.$$

La réalisation de cette estimation nécessite néanmoins la connaissance de la matrice de covariance $\boldsymbol{R}$.

1.6.2 Validation de la régression

La qualité du modèle de régression peut être quantifiée à l'aide du coefficient R^2 appelé coefficient de détermination.

Coefficient de détermination [Kleijnen, 2015] :
Soient $z_1, \ldots, z_n$ les observations, $\hat{z}_1, \ldots, \hat{z}_n$ les prédictions du modèle sur les observations (autrement dit pour tout i dans $[\![1, n]\!]$, $\hat{z}_i$ est la prédiction de z_i) et $\bar{z}$ la moyenne empirique des observations. Le coefficient de détermination R^2 est alors défini par :

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(z_i - \hat{z}_i)^2}{\sum_{i=1}^{n}(z_i - \bar{z})^2}.$$

Si l'on compare deux modèles dont le nombre de variables explicatives est identique, celui dont le R^2 est le plus proche de 1 sera meilleur selon ce critère d'adéquation. Pour

prendre en compte le nombre de variables explicatives du modèle, il est possible d'utiliser le coefficient de détermination ajusté :

$$R^2_{\text{ajusté}} = R^2 - \frac{p(1-R^2)}{n-p-1}$$

avec p le nombre de variables explicatives. Ce critère vient donc pénaliser les modèles complexes (avec p élevé). A nouveau, plus le critère est proche de 1, meilleur il est. Cependant, il est fortement déconseillé d'utiliser uniquement ce critère R^2 pour sélectionner un modèle du fait des problèmes de surapprentissage.

Une seconde approche consiste à séparer le jeu de données initial en deux parties. La première partie sert à construire le modèle (appelée ensemble d'apprentissage) et la seconde à le vérifier (appelée ensemble de validation ou de test). Cela permet de calculer le coefficient de prédictivité, défini de manière analogue au coefficient de détermination.

Coefficient de prédictivité [Marrel et al., 2008] :
Soient $y_1, \dots, y_n$ les observations de l'ensemble de validation, $\widehat{y}_1, \dots, \widehat{y}_n$ les prédictions du modèle sur les observations de l'ensemble de validation et $\bar{y}$ la moyenne des $y_1, \dots, y_n$. Le coefficient de prédictivité Q^2 est alors défini par :

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Lors d'une comparaison entre modèles, celui qui possède le Q^2 le plus proche de 1 sera considéré comme le meilleur. Par exemple, on peut considérer dans certaines applications qu'un modèle dont le Q^2 est inférieur à 0.5 est mauvais, tandis qu'un modèle ayant un Q^2 supérieur à 0.9 sera considéré comme satisfaisant.

La validation croisée à k-blocs est une autre technique de validation du modèle consistant à séparer le jeu de données en k échantillons. Elle prend un des échantillons comme bloc de validation et les autres comme blocs d'apprentissage. Cette étape de séparation du jeu de données est répétée pour les k blocs et donne le coefficient de prédictivité pour chacun des blocs, dont on calcule la moyenne pour obtenir un Q^2 global. Ce processus permet également de vérifier la robustesse du modèle et d'étudier l'évolution de l'estimation des paramètres. Dans le cas où l'ensemble de validation est composé d'un seul élément, on parle de version « leave-one-out ». La validation croisée pour le modèle de krigeage sera présentée dans la section 1.9.6.

Enfin il est possible de comparer différents modèles (avec différentes variables explicatives) à l'aide des critères AIC et BIC présentés auparavant (cf. section 1.2.4).

1.7 Analyse en composantes principales (ACP)

L'analyse en composantes principales se fait classiquement sur des données mises sous la forme d'un tableau :

$$\boldsymbol{X} = (x_{i,j})_{1 \leq i \leq n, 1 \leq j \leq p}$$

Les colonnes correspondent aux p variables explicatives, tandis que les lignes correspondent aux n individus. Aucune hypothèse d'indépendance n'est faite sur les variables explicatives. L'étude de ce tableau de données nécessite de travailler dans des espaces de dimensions p ou n, ce qui rend difficile la visualisation de graphiques synthétiques.

La méthode de l'ACP réalise des projections dans des espaces de dimension réduite (classiquement de dimension deux) en perdant le moins d'informations possibles. L'analyse en composantes principales permet d'étudier un grand nombre de variables aléatoires différentes simultanément en les exprimant sous la forme de combinaisons linéaires de variables non corrélées appelées facteurs. Il s'agit d'une analyse factorielle linéaire qui permet d'identifier les relations entre les variables explicatives (corrélations, anti-corrélations, non-corrélation) et de mettre en évidence des groupes dans les individus étudiés.

La Figure 1.2 donne en exemple deux graphes les plus répandus pour la synthèse de résultats des ACP. Le graphe de gauche représente les corrélations des variables explicatives (projetées sur le plan formé par les facteurs F1 et F2), tandis que le graphe de droite représente les individus (représentés sur le plan formé par les facteurs F1 et F2).

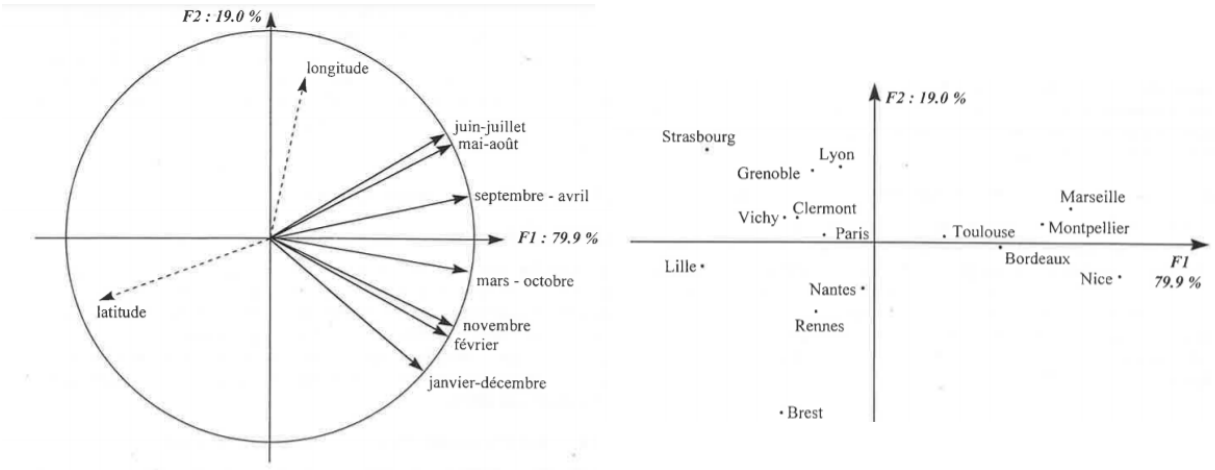


FIGURE 1.2 – Exemples d'ACP sur des données de températures selon la saison et la ville [Escofier and Pagès, 2008].

Cette analyse est particulièrement utile pour l'A&D puisque plusieurs grandeurs physiques différentes peuvent être mesurées. Un exemple d'application sur des résultats de mesure de radioactivité de plusieurs radionucléides peut être trouvé dans [Desnoyers, 2010]. Enfin les méthodes géostatistiques multivariables nécessitent une première analyse des corrélations pour être appliquées, ce que l'ACP permet de faire de manière efficace. Plus de détails sur l'ACP peuvent être trouvés dans [Escofier and Pagès, 2008]. Des versions prenant en compte des données qualitatives comme l'analyse en correspondances multiples existent également et sont décrites dans [Saporta, 1990].

1.8 Transformations des données

De manière générale, on cherche à approcher les données expérimentales par une loi gaussienne, puisque de nombreux outils statistiques et développements théoriques sont disponibles dans le cas où la loi de la variable aléatoire est une loi gaussienne.

1.8.1 Transformation gaussienne graphique

La première transformation classique est la transformation gaussienne graphique. Cette transformation associe la fonction de répartition de la variable aléatoire étudiée à celle de la loi gaussienne centrée réduite. La Figure 1.3 donne un exemple de transformation gaussienne graphique. Le graphe de droite correspond à la courbe d'une fonction de répartition d'une variable aléatoire de loi gaussienne centrée réduite, et celui de gauche à la courbe de la fonction de répartition de la variable aléatoire initiale avant transformation. Il s'agit d'une anamorphose gaussienne.

Anamorphose gaussienne [Chilès and Delfiner, 2012] :
Soient une variable aléatoire Z et une variable aléatoire de loi gaussienne centrée réduite Y. Une anamorphose gaussienne, notée Φ est alors l'application vérifiant :

$$Z = \Phi(Y)$$

avec $\Phi = F^{-1} \circ G$ avec F la fonction de répartition de Z et G la fonction de répartition de la loi gaussienne centrée réduite.

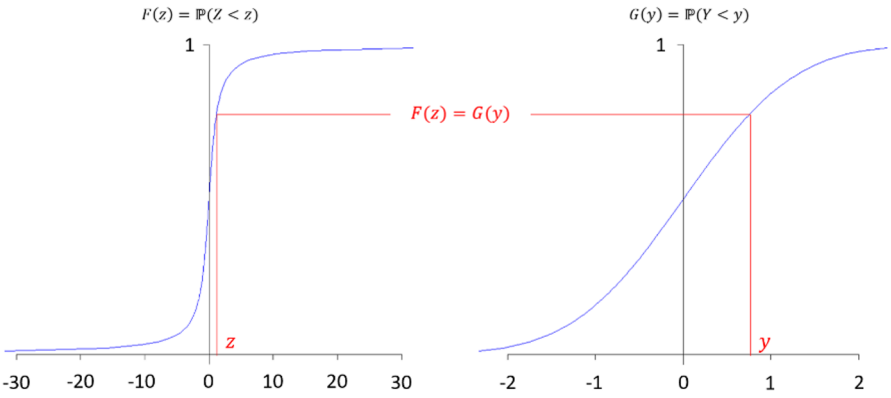


FIGURE 1.3 – Exemple de transformation gaussienne graphique.

Cette transformation nécessite cependant que la fonction F soit inversible. Enfin la fonction de répartition de la variable aléatoire étudiée est rarement connue, et il est souvent nécessaire d'utiliser la fonction de répartition empirique. Cette fonction est alors lissée afin d'obtenir une fonction de répartition approchée permettant de satisfaire la bijectivité de l'anamorphose gausienne.

1.8.2 Transformation de Box-Cox

Il existe d'autres transformations permettant de rapprocher les distributions asymétriques vers des distributions symétriques et de limiter l'influence des valeurs extrêmes sur

les estimateurs peu robustes comme la moyenne ou la variance. Une de ces transformations est celle de Box-Cox :

Transformation de Box-Cox [Box and Cox, 1964] :
Soient $\lambda_{BC} \in \mathbb{R}$ et Z une variable aléatoire à valeurs dans $\mathbb{R}^{+*}$. La transformation de Box-Cox est définie par :

$$BC(Z, \lambda_{BC}) = \begin{cases} \dfrac{Z^{\lambda_{BC}} - 1}{\lambda_{BC}} \text{ si } \lambda_{BC} \neq 0, \\ \log(Z) \text{ si } \lambda_{BC} = 0. \end{cases}$$

Le paramètre λ_{BC} peut être estimé par une méthode de maximum de vraisemblance en utilisant comme densité le produit de la loi gaussienne par le jacobien de la transformation de Box-Cox ou par une approche bayésienne (voir [Box and Cox, 1964]). D'autres alternatives pour le choix de λ_{BC} sont évoquées dans [Sakia, 1992]. Un des inconvénients de la transformation de Box-Cox est qu'elle ne permet pas nécessairement d'obtenir une loi symétrique après transformation dans le cas de distribution à queue lourde et qu'elle ne permet pas de transformer des variables aléatoires prenant des valeurs négatives ou nulles.

1.9 Outils géostatistiques

Commençons par donner une définition générale de la géostatistique :

Géostatistique :
La géostatistique est une branche des statistiques spatiales s'intéressant à l'étude de champs aléatoires, principalement en dimension 2 et 3. Elle est souvent utilisée pour réaliser des prédictions du champ étudié en des points non-observés afin d'en obtenir une cartographie complète.

La géostatistique a été développée par [Matheron, 1970] à la suite des travaux [Krige, 1978] sur des gisements miniers en Afrique du Sud. Depuis, elle est appliquée dans de nombreux domaines, comme par exemple la météorologie, la climatologie, l'hydrogéologie, la géologie et l'agronomie. Notre cadre d'application est celui de l'A&D de sites nucléaires, les mesures étant principalement des mesures de contaminations (radioactives, chimiques) sur des sols ou des matériaux. Les premiers travaux de géostatistiques pour l'A&D d'INB sont présentés dans [Desnoyers, 2010]. Des descriptions plus précises de la méthodologie peuvent être trouvées dans [Chauvet, 2006, Chilès and Delfiner, 2012].

1.9.1 Modélisation du phénomène

Champ aléatoire réel :
Soit un domaine d'étude $D \subset \mathbb{R}^d$ avec $d \in \mathbb{N}^*$ et un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$. Un champ aléatoire réel est une application $Z(.)$ qui associe une variable aléatoire $Z(\boldsymbol{x})$ à valeurs dans $\mathbb{R}$ à chaque $\boldsymbol{x} \in D$, avec $\boldsymbol{x}' = (x^{(1)}, \ldots, x^{(d)})$. Toutes ces variables aléatoires sont définies sur le même espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$.

Par ailleurs, un champ aléatoire $Z(.)$ défini sur D est dit gaussien si pour tout n dans $\mathbb{N}^*$, tout $\alpha_1, ..., \alpha_n$ dans $\mathbb{R}$ et tout $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ dans D, la loi de la combinaison linéaire $\sum_{i=1}^{n} \alpha_i Z(\boldsymbol{x}_i)$ est une loi gaussienne.

Dans la suite du document, la grandeur physique est modélisée par un champ aléatoire réel $\{Z(\boldsymbol{x}), \boldsymbol{x} \in D \subset \mathbb{R}^d\}$ avec $d \in \{1, 2, 3\}$. La position d'une observation est notée $\boldsymbol{x} = (x^{(1)}, \ldots, x^{(d)})'$. Le champ aléatoire est souvent décomposé en un terme déterministe et un terme aléatoire, de la manière suivante :

$$Z(\boldsymbol{x}) = \mu(\boldsymbol{x}) + \epsilon(\boldsymbol{x})$$

avec $\mu(.)$ le terme correspondant à moyenne (ou tendance) de Z et $\epsilon(.)$ le terme aléatoire, lui-même un champ aléatoire d'espérance nulle.

Le domaine d'étude D est discrétisé selon un plan d'échantillonnage (cf. section I.3.2). Les résultats de mesure sont alors les observations $z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_n)$ de la grandeur physique aux positions $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in D$. Les observations correspondent à la réalisation d'une trajectoire discrétisée sur le domaine d'étude du champ aléatoire $Z(.)$. L'hypothèse i.i.d. faite dans la section 1.1 ne tient plus ici : les variables aléatoires $Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)$ sont supposées corrélées. Enfin nous ne considérerons dans la suite que des modèles paramétriques, à la fois pour la tendance, et pour la matrice de covariance du vecteur $\boldsymbol{Z} = (Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n))'$.

1.9.2 Stationnarité

Pour réaliser l'analyse géostatistique, il est nécessaire d'identifier la structure spatiale du champ $Z(.)$, par exemple en estimant sa moyenne, sa variance etc... Néanmoins, sans hypothèse supplémentaire, ces estimations ne sont pas réalisables car nous ne disposons que d'une seule réalisation du champ aléatoire $Z(.)$. De plus cette réalisation est fragmentaire car elle n'est disponible que sur les points échantillonnés. Les hypothèses de stationnarité permettent de réaliser l'inférence statistique nécessaire pour l'estimation des paramètres du modèle ainsi que les prédictions. Ces hypothèses consistent à considérer que certains moments de la loi du champ aléatoire $Z(.)$ sont invariants par translation. Dans la suite, nous considérons les deux types de stationnarité suivants : la stationnarité d'ordre 2 et la stationnarité intrinsèque.

Stationnarité d'ordre deux [Chilès and Delfiner, 2012] :
Supposons le champ aléatoire $Z(.)$ de carré intégrable. Le champ aléatoire $Z(.)$ est dit stationnaire d'ordre deux lorsque :

$$\forall \boldsymbol{x}, \boldsymbol{x} + \boldsymbol{h} \in D, \begin{cases} \mathbb{E}[Z(\boldsymbol{x})] = \mu \\ Cov(Z(\boldsymbol{x}), Z(\boldsymbol{x} + \boldsymbol{h})) = C(\boldsymbol{h}) \end{cases}$$

où μ est la moyenne constante du champ et où la fonction $C(.)$ est la fonction de covariance et ne dépend que de $\boldsymbol{h}$.

Stationnarité intrinsèque [Chilès and Delfiner, 2012] :
Supposons les accroissements du champ aléatoire $Z(.)$ de carré intégrable. Le champ aléatoire $Z(.)$ est dit intrinsèquement stationnaire lorsque :

$$\forall \boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h} \in D, \begin{cases} \mathbb{E}[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})] = 0 \\ Var\left[Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})\right] = 2\gamma(\boldsymbol{h}) \end{cases}$$

où l'espérance des accroissements est nulle et où la fonction $\gamma(.)$, qui ne dépend que de $\boldsymbol{h}$, est appelée le (semi)variogramme.

De plus dans le cas de la stationnarité d'ordre deux, variogramme et covariance sont liés par la formule suivante :

Relation entre variogramme et covariance [Chauvet, 2006] :
Sous l'hypothèse de stationnarité d'ordre 2, pour tout $\boldsymbol{x} \in D$ et pour tout $\boldsymbol{h}$ tel que $\boldsymbol{x}+\boldsymbol{h} \in D$, on a :

$$C(\boldsymbol{h}) = C(0) - \gamma(\boldsymbol{h}).$$

Enfin la stationnarité d'ordre deux implique la stationnarité intrinsèque. La propriété réciproque n'est cependant pas vraie, le contre-exemple classique étant le mouvement brownien dont la covariance dépend de la position et non pas uniquement de la distance entre deux points. Cela rend donc l'hypothèse de stationnarité intrinsèque plus générale. En ce qui concerne le choix de l'utilisation du variogramme ou de la fonction de covariance, la plupart des auteurs préfèrent le variogramme puisqu'il est l'outil utilisé dans le cas de l'hypothèse de stationnarité la plus générale et qu'il permet également de se libérer de l'estimation de l'espérance du champ aléatoire lors de l'estimation des paramètres du modèle de covariance ou du variogramme. Ce point sera davantage discuté dans la section 1.9.4.

Les hypothèses de stationnarité peuvent être vérifiées expérimentalement. Dans le cas de stationnarité d'ordre deux, la moyenne et la variance empirique doivent être constantes sur n'importe quel sous-ensemble de l'ensemble formé par les observations. Par exemple, elles peuvent être estimées expérimentalement à l'aide du graphe d'évolution directionnelle du moment étudié. Ces graphes consistent à représenter la moyenne ou la variance selon une direction. Dans le cas de données échantillonnées selon une grille non-régulière, des tolérances peuvent être ajoutées pour obtenir des ensembles contenant un nombre satisfaisant de valeurs.

La Figure 1.4 en donne un exemple, avec le graphe de gauche représentant un jeu de données issu d'une simulation, et le graphe de droite représentant l'évolution de la moyenne selon chacune des directions (ici $x^{(1)}$ et $x^{(2)}$). Le graphe de droite illustre un cas où l'hypothèse d'une moyenne constante est raisonnable.

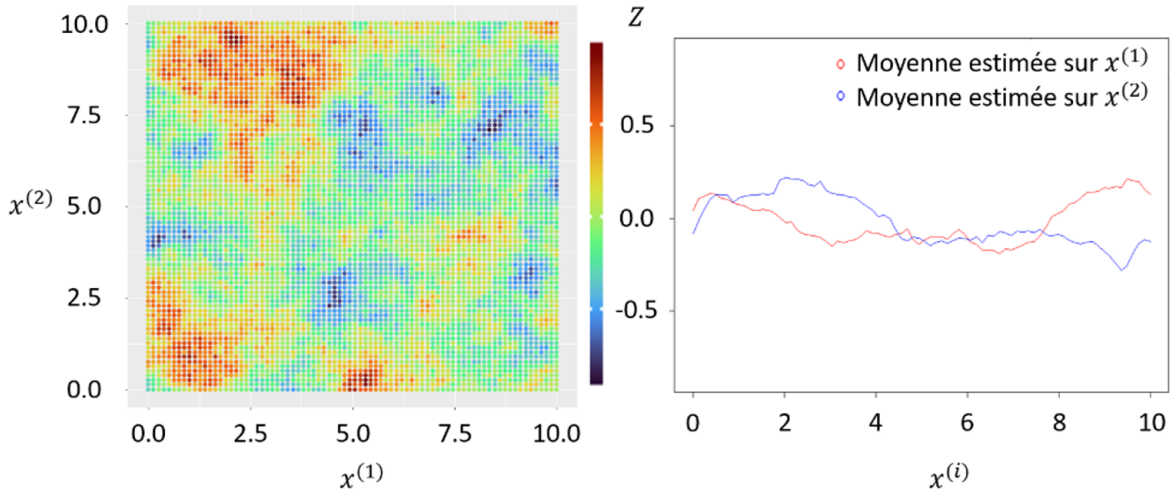


FIGURE 1.4 – Exemple de graphe d'évolution directionnelle des moments d'un champ aléatoire [Emery, 2001].

Cette analyse exploratoire permet également de mettre en évidence des tendances si les moments évoluent régulièrement selon certaines directions, ou des phénomènes quasi-stationnaires (phénomènes stationnaires sur certaines régions du domaine d'étude).

1.9.3 Propriétés du variogramme

1.9.3.1 Variogrammes usuels

Un variogramme γ est une fonction de type négatif conditionnel :

$$\forall n \in \mathbb{N}^*, \forall \boldsymbol{y}_1, \dots, \boldsymbol{y}_n \in D \text{ et } \forall a_1, \dots, a_n \in \mathbb{R}, \sum_{i=1}^{n} a_i = 0 \implies \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(\boldsymbol{y}_i - \boldsymbol{y}_j) \leq 0.$$

Cette propriété est une condition nécessaire et suffisante pour que $\gamma(.)$ soit un variogramme. Voici quelques exemples classiques de familles de fonctions utilisées en tant que variogramme :

Par simplicité on se place dans le cas où $d = 1$. Soient $\sigma^2, \tau^2, h \in \mathbb{R}^+$ et $\psi, \phi \in \mathbb{R}^{+*}$.
Famille de Matérn :
On note Γ la fonction gamma et K_ψ la fonction de Bessel modifiée de seconde espèce (voir [Watson, 1995]). La famille est définie par :

$$\gamma_{\phi,\sigma^2}(h) = \sigma^2 \left(1 - \frac{1}{2^{\psi-1}\Gamma(\psi)} \left(\frac{h}{\phi}\right)^{\psi} K_\psi \left(\frac{h}{\phi}\right)\right).$$

Famille sphérique :

La famille est définie par :

$$\gamma_{\phi,\sigma^2}(h) = \begin{cases} \sigma^2 \left(\frac{3h}{2\phi} - \frac{1}{2} \left(\frac{h}{\phi} \right)^3 \right) & \text{si } h \leq \phi, \\ \sigma^2 \text{ si } h > \phi. \end{cases}$$

Famille gaussienne :
La famille est définie par :

$$\gamma_{\phi,\sigma^2}(h) = \sigma^2 \exp\left(-\frac{h^2}{\phi^2} \right).$$

Effet de pépite :
L'effet de pépite est modélisé par :

$$\gamma_{\tau^2}(h) = \begin{cases} 0 \text{ si } h = 0, \\ \tau^2 \text{ si } h \neq 0. \end{cases}$$

Les fonctions présentées ici sont parfois paramétrées par ϕ (éventuellement un vecteur de paramètres), souvent appelé portée ou longueur de corrélation, et σ^2 souvent assimilé à une variance. La constante τ^2 est appelée pépite ou effet de pépite. Ce modèle particulier sera discuté plus en détails dans la section 1.9.5.6. L'estimation de ces paramètres sera détaillée dans la section 1.9.4.

La famille de Matérn est l'une des familles classiques les plus utilisées. Le choix du paramètre ψ vient modifier la régularité du champ ainsi paramétré. Plus ψ est faible, plus les réalisations du champ apparaîtront irrégulières. Au contraire, lorsque ψ tend vers $+\infty$, les réalisations du champ apparaîtront régulières. En particulier pour $\psi \to +\infty$, la famille de Matérn ainsi paramétrée est équivalente à la famille gaussienne. On précise ici quelques cas particuliers usuels de la fonction de Matérn dont l'expression analytique est disponible.

Famille exponentielle (Matérn $\psi = \frac{1}{2}$) :
La famille est définie par :

$$\gamma_{\phi,\sigma^2}(h) = \sigma^2 \left(1 - \exp\left(-\frac{h}{\phi} \right) \right).$$

Famille Matérn-$\frac{3}{2}$:
La famille est définie par :

$$\gamma_{\phi,\sigma^2}(h) = \sigma^2 \left(1 - \left(1 + \frac{\sqrt{3}h}{\phi} \right) \exp\left(-\frac{\sqrt{3}h}{\phi} \right) \right).$$

Famille Matérn-$\frac{5}{2}$:

La famille est définie par :

$$\gamma_{\phi,\sigma^2}(h) = \sigma^2 \left(1 - \left(1 + \frac{\sqrt{5}h}{\phi} + \frac{\sqrt{5}h^2}{3\phi^2}\right) \exp\left(-\frac{\sqrt{5}h}{\phi}\right)\right).$$

Des familles équivalentes (sous hypothèse de stationnarité d'ordre 2) pour les fonctions de covariance sont données ici.

Famille sphérique :
La famille est définie par :

$$C_{\phi,\sigma^2}(h) = \begin{cases} \sigma^2 \left(1 - \frac{3h}{2\phi} + \frac{1}{2}\left(\frac{h}{\phi}\right)^3\right) & \text{si } h \leq \phi, \\ 0 \text{ si } h > \phi. \end{cases}$$

Famille exponentielle (Matérn $\psi = \frac{1}{2}$) :
La famille est définie par :

$$C_{\phi,\sigma^2}(h) = \sigma^2 \exp\left(-\frac{h}{\phi}\right).$$

Famille Matérn-$\frac{3}{2}$:
La famille est définie par :

$$C_{\phi,\sigma^2}(h) = \sigma^2 \left(1 + \frac{\sqrt{3}h}{\phi}\right) \exp\left(-\frac{\sqrt{3}h}{\phi}\right).$$

Famille Matérn-$\frac{5}{2}$:
La famille est définie par :

$$C_{\phi,\sigma^2}(h) = \sigma^2 \left(1 + \frac{\sqrt{5}h}{\phi} + \frac{\sqrt{5}h^2}{3\phi^2}\right) \exp\left(-\frac{\sqrt{5}h}{\phi}\right).$$

Famille gaussienne :
La famille est définie par :

$$C_{\phi,\sigma^2}(h) = \sigma^2 \left(1 - \exp\left(-\frac{h^2}{\phi^2}\right)\right).$$

Effet de pépite :
L'effet de pépite est modélisé par :

$$C_{\tau^2}(h) = \begin{cases} 0 \text{ si } h \neq 0, \\ \tau^2 \text{ si } h = 0. \end{cases}$$

Dans la suite tous les variogrammes considérés seront paramétrés par une variance $\boldsymbol{\sigma^2}$, une portée $\boldsymbol{\phi}$ et un effet de pépite $\boldsymbol{\tau^2}$.

1.9.3.2 Isotropie

Le variogramme ne dépend que du vecteur $\boldsymbol{h}$ qui est caractérisé par sa norme et sa direction. Si le variogramme ne dépend que de la norme, il est dit isotrope. S'il dépend également de la direction, il est dit anisotrope. Une définition similaire existe pour la fonction de covariance.

Covariance et variogramme isotrope :
Le champ aléatoire $\epsilon(.)$ intrinsèquement stationnaire possède un variogramme isotrope si et seulement si :

$$\gamma(h) = \gamma(\|h\|).$$

Le champ aléatoire $\epsilon(.)$ stationnaire d'ordre 2 possède une fonction de covariance isotrope si et seulement si :

$$C(h) = C(\|h\|).$$

Dans la suite du document, tous les variogrammes et covariances considérés seront isotropes. Le cas anisotrope ne sera pas détaillé ici, mais est développé dans [Chilès and Delfiner, 2012] et [Allard et al., 2015].

1.9.3.3 Décomposition de la variance

Reprenons le modèle proposé par [Cressie, 1993] pour le champ aléatoire :

$$Z(.) = \mu(.) + \omega(.) + \xi(.). \tag{1.2}$$

Les différents termes peuvent être décrits de la manière suivante :

- $\mu(.)$ le terme de tendance déjà présenté correspondant aux variations déterministes du champ à grande échelle ;
- $\omega(.)$ le terme de variations aléatoires lisses à courte échelle de moyenne nulle, intrinsèquement stationnaire, de variogramme continu avec une portée ϕ_ω, si elle existe, telle que $\phi_\omega \geq \min_{1\leq i<j\leq n}(\|\boldsymbol{x}_i - \boldsymbol{x}_i\|)$;
- $\xi(.)$ le terme de variations aléatoires à micro-échelle de moyenne nulle, intrinsèquement stationnaire de variogramme de portée ϕ_ξ, si elle existe, est telle que $\phi_\xi \leq \min_{1\leq i<j\leq n}(\|\boldsymbol{x}_i - \boldsymbol{x}_i\|)$.

Ces termes sont considérés indépendants. En reprenant les notations précédentes, on a alors :

$$\gamma(.) = \gamma_\omega(.) + \gamma_\xi(.)$$

avec $\gamma(.)$ le variogramme de Z, γ_ω le variogramme de ω et γ_ξ le variogramme de ξ. En pratique le terme ξ est historiquement nommé « effet de pépite » et est modélisé avec le modèle d'effet de pépite présenté dans la section 1.9.3.1. En supposant que les portées ϕ_ω et ϕ_ξ existent, la portée ϕ de $Z(.)$ est alors le maximum des deux portées.

Enfin si ces variogrammes admettent un palier (autrement dit leur variogramme $\gamma(.)$ vérifie $\lim_{h\to+\infty}\gamma(h) = l$, avec l une constante positive), on peut alors écrire asymptotiquement :

$$\lim_{h\to+\infty}\gamma(h) = \sigma^2 + \tau^2,$$

où $\sigma^2 = \lim_{h\to+\infty}\gamma_\omega(h)$ et $\tau^2 = \lim_{h\to+\infty}\gamma_\xi(h)$.

La question de l'existence des paliers peut se poser puisque certains phénomènes ne possèdent pas de variogrammes bornés (comme le mouvement brownien). Ces cas ne seront pas traités ici mais plus d'informations sur ces variogrammes et leur modélisation peuvent être trouvés dans [Chilès and Delfiner, 2012].

La Figure 1.5 présente graphiquement les paramètres σ^2, ϕ et τ^2.

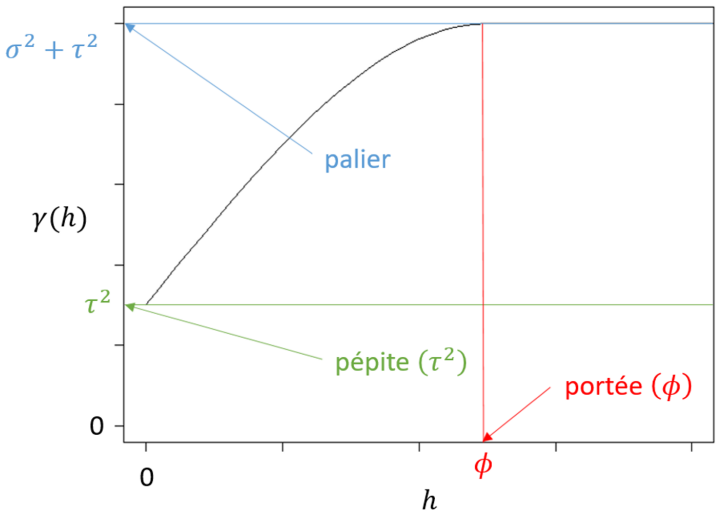


FIGURE 1.5 – Illustration graphique de l'ensemble des paramètres du variogramme.

1.9.4 Modélisation et estimation des paramètres du variogramme

En pratique le variogramme du champ $Z(.)$ n'est pas connu. Il est donc nécessaire de le modéliser et de l'estimer pour pouvoir décrire la structure spatiale du champ. Dans la suite on se limitera à des variogrammes issus des familles classiques présentées dans la section 1.9.3.1 pour modéliser le variogramme du champ étudié. Ainsi les paramètres σ^2, τ^2 et ϕ forment généralement les paramètres du modèle de variogramme, et correspondent respectivement à la variance, à la pépite et la portée.

Deux approches sont présentées ici pour estimer les paramètres de notre modèle : l'estimateur du maximum de vraisemblance et l'analyse variographique. Il est important de noter que l'analyse variographique n'estime pas directement la moyenne, qui est alors estimée par moindres carrés généralisés.

1.9.4.1 Maximum de vraisemblance

Pour l'application du maximum de vraisemblance, on considère le cas spécifique où la tendance vérifie :

$$\mu(\boldsymbol{x}) = \sum_{k=1}^{p} d_k(\boldsymbol{x})\mu_k = \boldsymbol{D}\boldsymbol{\mu},$$

avec $d_k(.)$ les covariables connues, p le nombre de covariables considérées et μ_k les coefficients de régression à estimer. Le champ $Z(.)$ est supposé gaussien et stationnaire d'ordre 2. Le vecteur $\boldsymbol{Z} = (Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n))'$ suit alors la loi multivariée gaussienne $\mathcal{N}_n(\boldsymbol{D}\boldsymbol{\mu}, \boldsymbol{R})$ avec

$\boldsymbol{D\mu}$ la moyenne et $\boldsymbol{R}$ la matrice de covariance. Cette matrice est supposée inversible et se met sous la forme :

$$\boldsymbol{R}(\sigma^2, \phi, \tau^2) = \sigma^2 \boldsymbol{V}(\phi) + \tau^2 \boldsymbol{I}_n,$$

avec la matrice $\boldsymbol{V}(\phi) = (C_\phi(\boldsymbol{x}_i - \boldsymbol{x}_j))_{1 \leq i,j \leq n}$, C_ϕ une fonction de covariance et $\boldsymbol{I}_n$ la matrice identité de taille n. Dans la suite (et sauf mention contraire) la fonction C_ϕ vérifie $C_\phi(0) = 1$, le paramètre σ^2 étant exclu de la fonction pour simplifier certaines expressions. Le modèle est donc paramétré par $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma^2, \phi, \tau^2)$.

Log-vraisemblance pour l'estimation des paramètres de la covariance (cas gaussien) :
L'expression de la log-vraisemblance est donnée par :

$$\begin{aligned}\ln\left(L(\boldsymbol{z}, \boldsymbol{\mu}, \sigma^2, \phi, \tau^2)\right) = &-\frac{1}{2}(n \ln(2\pi) + \ln\left(\det(\boldsymbol{R}(\sigma^2, \phi, \tau^2))\right) \\ &+ (\boldsymbol{z} - \boldsymbol{D\mu})' \boldsymbol{R}(\sigma^2, \phi, \tau^2)^{-1} (\boldsymbol{z} - \boldsymbol{D\mu})).\end{aligned}$$

L'estimateur du maximum de vraisemblance est alors :

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left(\ln\left(L(\boldsymbol{Z}, \boldsymbol{\mu}, \sigma^2, \phi, \tau^2)\right)\right).$$

A nouveau, le maximum de la vraisemblance ou de la log-vraisemblance n'est pas forcément unique ou n'existe pas. Cette méthode permet l'estimation simultanée de l'ensemble des paramètres, mais il est néanmoins important d'être vigilant puisque cette estimation peut introduire des biais non-négligeables. [Chilès and Delfiner, 2012] indique que ces biais sont introduits à cause de la difficulté de séparer la tendance des écarts à la moyenne. Pour cela il est possible de faire appel au maximum de vraisemblance restreint détaillé par exemple dans [Webster and Oliver, 2007].

Trouver les paramètres maximisant la fonction de vraisemblance est parfois difficile. [Diggle and Ribeiro, 2007] donne une approche simplifiant l'estimation des paramètres $\boldsymbol{\theta}$. Tout d'abord la matrice $\boldsymbol{R}$ est re-paramétrée en considérant $\eta^2 = \frac{\tau^2}{\sigma^2}$ et en écrivant $\boldsymbol{R} = \sigma^2 \boldsymbol{\Sigma}(\phi, \tau^2) = \sigma^2(\boldsymbol{V}(\phi) + \eta^2 \boldsymbol{I}_n)$. Dans la suite les dépendances selon les paramètres σ^2, ϕ et τ^2 seront omises. $\boldsymbol{\Sigma}$ est supposée connue et la vraisemblance est maximisée par les estimateurs des moindres carrés généralisés donnés dans la partie 1.6.1 :

$$\begin{cases} \hat{\boldsymbol{\mu}}(\boldsymbol{\Sigma}) &= (\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z}, \\ \hat{\sigma}^2(\boldsymbol{\Sigma}) &= \frac{1}{n}(\boldsymbol{Z} - \boldsymbol{D}\hat{\boldsymbol{\mu}}(\boldsymbol{\Sigma}))'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - \boldsymbol{D}\hat{\boldsymbol{\mu}}(\boldsymbol{\Sigma})). \end{cases}$$

Ensuite ces expressions sont directement injectées dans celle de la log-vraisemblance qui ne dépend alors plus que des paramètres η^2, ϕ :

$$\ln\left(L(\boldsymbol{z}, \eta^2, \phi)\right) = -\frac{1}{2}\left(n \ln(2\pi) + n \ln(\hat{\sigma}^2(\boldsymbol{\Sigma})) + \ln(\det(\boldsymbol{\Sigma})) + n\right).$$

Ensuite les η^2 et ϕ maximisant la log-vraisemblance sont obtenus par optimisation, puis les paramètres $\boldsymbol{\mu}$ et σ^2 sont estimés en injectant les estimations de η^2 et ϕ dans les expressions des moindres carrés généralisés.

1.9.4.2 Analyse variographique

Le variogramme γ peut être estimé par le variogramme expérimental $\hat{\gamma}$.

Variogramme expérimental [Emery, 2001] :
Soit $h \in \mathbb{R}^{+*}$, $N(h) = \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \|\boldsymbol{x}_i - \boldsymbol{x}_j\| = h\}$ et $|N(h)| = Card(N(h))$. Le variogramme expérimental est alors défini par :

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in N(h)} [Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_j)]^2 .$$

Ainsi en calculant le variogramme pour différentes distances, des estimations ponctuelles du variogramme sont obtenues. En pratique les couples de points $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ correspondant aux positions des observations sont rarement distants exactement de h. Cela implique un faible nombre de couples disponibles pour l'estimation du variogramme. Pour limiter ce problème, on introduit des tolérances sur h en terme de distance. Les couples pris pour estimer $\hat{\gamma}$(h) vérifient alors :

$$\{(\boldsymbol{x}_i, \boldsymbol{x}_j) | \, \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \in [h - dh, h + dh]\},$$

avec dh la tolérance choisie. Ce choix de tolérance n'est pas trivial et peut modifier l'estimation du variogramme. Plus de détails sur le choix des tolérances peuvent être trouvés dans [Emery, 2001, Webster and Oliver, 2007]. De manière plus générale, il est souvent nécessaire d'avoir à disposition un grand nombre de points pour pouvoir estimer le variogramme sur de nombreuses distances et de limiter l'influence du choix de la tolérance sur ces estimations.

L'estimation du variogramme donne seulement une description discontinue du variogramme. Un modèle paramétrique est donc ajusté sur le variogramme expérimental. Pour cela une famille de variogrammes est choisie puis les paramètres peuvent être estimés à l'aide d'une méthode des moindres carrés ou selon une pénalisation particulière (voir [Diggle and Ribeiro, 2007] pour d'autres exemples d'ajustement du modèle). La moyenne est ensuite estimée en utilisant l'expression des moindres carrés généralisés.

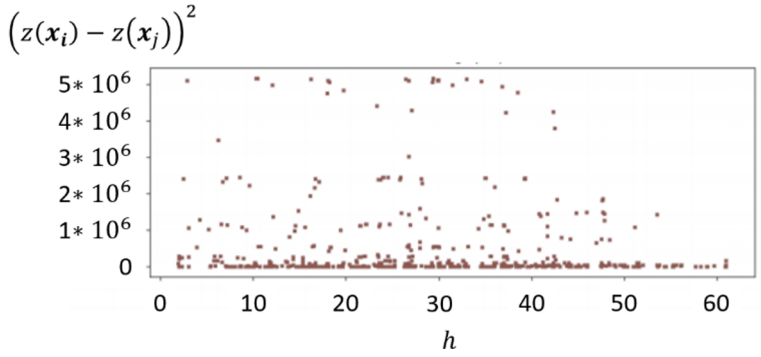


FIGURE 1.6 – Exemple de nuée variographique.

La réalisation de la nuée variographique indique la répartition des données lors du calcul du variogramme expérimental. Cette nuée correspond à l'ensemble des valeurs $[z(\boldsymbol{x}_i) - z(\boldsymbol{x}_j)]^2$ obtenues pour chaque couple $(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Elle est utilisée lors de l'analyse exploratoire afin d'identifier d'éventuelles valeurs aberrantes (qui apparaîtront comme des points très éloignés du reste du nuage). Le variogramme expérimental étant très sensible aux valeurs extrêmes, la réalisation de cette nuée variographique évite d'introduire des biais dans l'estimation du variogramme. La Figure 1.6 illustre un exemple de nuée variographique où plusieurs points apparaissent éloignés du reste du nuage, mettant en évidence des valeurs extrêmes éventuellement aberrantes.

1.9.5 Prédictions par krigeage

1.9.5.1 Définition et contraintes de krigeage

Pour la suite de cette section, on suppose que le champ aléatoire Z est intrinsèquement stationnaire et de variogramme γ. Le krigeage est une méthode d'interpolation linéaire qui prédit la valeur d'une variable étudiée en un point en utilisant une moyenne pondérée des observations. La valeur de la variable aléatoire $Z(x_0)$ en un point non-observé $\boldsymbol{x}_0$ est alors prédite par :

$$\hat{Z}(\boldsymbol{x}_0) = \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{x}_i) = \boldsymbol{\lambda}' \boldsymbol{Z}$$

où le vecteur $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$ est à estimer. Les prédictions d'un krigeage regroupent deux termes, un premier correspondant à la valeur prédite et un second correspondant à la variance de l'erreur de prédiction, aussi appelée variance de prédiction qui sont obtenues à l'aide des contraintes de krigeage. Ces contraintes sont énoncées sans hypothèse sur le champ Z et permettent de définir les équations vérifiées par les coefficients λ_i :

Contraintes de krigeage [Chauvet, 2006] :
- **La contrainte de linéarité** : la prédiction de krigeage doit s'écrire sous la forme d'une combinaison linéaire des données.
- **La contrainte d'autorisation** : l'espérance et la variance de l'erreur de prédiction $\hat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)$ doivent exister.
- **La contrainte de non-biais** : la prédiction par krigeage doit être sans biais i.e. $\mathbb{E}[\hat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)] = 0$.
- **La contrainte d'optimalité** : les λ_i sont calculés pour minimiser $\sigma^2_{pred} = Var[\hat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)]$.

Dans la suite du document, seules les contraintes d'optimalité et de non-biais seront considérées pour expliciter les formules de krigeage.

Selon les hypothèses initiales considérées pour le modèle (type de stationnarité, présence ou non d'une tendance, etc.), les contraintes donneront des équations différentes qui définiront des krigeages différents. Le krigeage simple consiste à considérer la moyenne connue dans le cas stationnaire d'ordre 2. Le krigeage ordinaire permet de travailler dans le cas d'une moyenne inconnue toujours sous stationnarité d'ordre 2. C'est l'un des krigeages les plus répandus grâce à sa facilité d'utilisation. Le krigeage universel permet de travailler

avec la présence d'une tendance spatiale $\mu(\boldsymbol{x})$ dans les données en ajustant un modèle de régression linéaire, puis en appliquant une analyse des corrélations spatiales sur estimations de $\epsilon(\boldsymbol{x}_i) = Z(\boldsymbol{x}_i) - \mu(\boldsymbol{x}_i)$ (pour tout $1 \leq i \leq n$). Le krigeage par bloc permet de prédire une valeur sur un certain support (surface ou volume) plutôt que de réaliser une estimation ponctuelle. Le krigeage bayésien considère les paramètres comme des variables aléatoires et prend en compte les incertitudes dans leur estimation. Il sera détaillé dans le chapitre 2. Le cokrigeage permet de répondre au cas multivariable en utilisant les corrélations entre les différentes variables. Plus de détails peuvent être trouvés sur le krigeage simple, ordinaire, et universel dans [Cressie, 1993], plus de détails sur le krigeage par bloc et le co-krigeage peuvent être trouvés dans [Webster and Oliver, 2007] et davantage de détails sur le krigeage bayésien peuvent être trouvés dans [Diggle and Ribeiro, 2002].

La formulation donnée ici sous-entend une prédiction avec l'ensemble des observations. Néanmoins il peut être intéressant de ne considérer que certaines valeurs proches du point à prédire, notamment dans le cas de phénomènes quasi-stationnaires. Les points considérés pour la prédiction appartiennent alors à une région appelée voisinage de krigeage. Plus d'informations sur les problèmatiques de voisinage de krigeage peuvent être trouvées dans [Emery, 2001]. Dans la suite nous ne restreindrons pas le voisinage de krigeage : l'ensemble des données sera utilisé pour les prédictions.

1.9.5.2 Krigeage simple

Le krigeage simple est la première forme de krigeage. Son utilisation nécessite la connaissance de la moyenne μ supposée constante. De plus on suppose que les paramètres ϕ, σ^2 et τ^2 sont connus. Les équations du krigeage simple sont obtenues pour la stationnarité d'ordre deux. Le modèle est alors :

$$\forall \boldsymbol{x} \in D, Z(\boldsymbol{x}) = \mu + \epsilon(\boldsymbol{x}).$$

A nouveau la matrice de covariance du vecteur $\boldsymbol{Z}$ des observations est notée $\boldsymbol{R}$ en omettant la dépendance en ϕ, σ^2 et τ^2. La matrice $\boldsymbol{R}$ et le vecteur $\boldsymbol{r} = (Cov(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_0)))_{1 \leq i \leq n}$ sont alors connus. $\boldsymbol{R}$ est supposée inversible. Enfin les conditions de linéarité et de non-biais amènent :

$$\hat{Z}(\boldsymbol{x}_0) = \mu + \boldsymbol{\lambda}'(\boldsymbol{Z} - \mu \mathbf{1}_n),$$

avec $\mathbf{1}_n$ le vecteur colonne de taille n composé uniquement de 1.

Les poids $\boldsymbol{\lambda}$ sont ensuite calculés à l'aide de la condition d'optimalité :

$$\begin{aligned} Var\left[\hat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)\right] &= Var\left[\mu + \boldsymbol{\lambda}'(\boldsymbol{Z} - \mu \mathbf{1}_n) - Z(\boldsymbol{x}_0)\right] \\ &= Var[\boldsymbol{\lambda}'\boldsymbol{Z}] + Var[Z(\boldsymbol{x}_0)] - 2Cov(\boldsymbol{\lambda}'\boldsymbol{Z}, Z(\boldsymbol{x}_0)) \\ &= \boldsymbol{\lambda}'\boldsymbol{R}\boldsymbol{\lambda} + \sigma^2 - 2\boldsymbol{\lambda}'\boldsymbol{r}. \end{aligned}$$

En dérivant par rapport à $\boldsymbol{\lambda}$, on obtient :

$$\frac{\partial}{\partial \boldsymbol{\lambda}}\left(Var\left[\hat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)\right]\right) = \frac{\partial}{\partial \boldsymbol{\lambda}}(\boldsymbol{\lambda}'\boldsymbol{R}\boldsymbol{\lambda} + \sigma^2 - 2\boldsymbol{\lambda}'\boldsymbol{r}) = 2(\boldsymbol{R}\boldsymbol{\lambda} - \boldsymbol{r}).$$

Ce terme s'annule pour $\boldsymbol{\lambda} = \boldsymbol{R}^{-1}\boldsymbol{r}$. Le discriminant hessien de la fonction $g(.)$ définie par

$$g(\lambda) = \boldsymbol{\lambda}'\boldsymbol{R}\boldsymbol{\lambda} + \sigma^2 - 2\boldsymbol{\lambda}'\boldsymbol{r}$$

étant $2\boldsymbol{R}$, et cette matrice étant semi-définie positive, $g(.)$ est convexe et le point ainsi obtenu est un minimum global. En injectant $\boldsymbol{\lambda}$ dans l'expression de l'estimateur, les équations du krigeage simple sont obtenues.

Equations du krigeage simple :
Les équations du krigeage simple sont données par :

$$\begin{cases} \hat{Z}(\boldsymbol{x}_0) = \mu + \boldsymbol{r}'\boldsymbol{R}^{-1}(\boldsymbol{Z} - \mu\mathbf{1}_n), \\ Var\left[\hat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)\right] = \sigma^2 - \boldsymbol{r}'\boldsymbol{R}^{-1}\boldsymbol{r}. \end{cases}$$

1.9.5.3 Krigeage ordinaire

Le modèle du krigeage ordinaire est similaire à celui du krigeage simple, à la différence que la moyenne μ est inconnue et quasi-constante, ce qui autorise des valeurs variables pour la moyenne, à la condition que cette moyenne reste constante dans le voisinage de krigeage. Dans notre cas le voisinage de krigeage correspond à l'ensemble des points, et μ est donc considérée comme constante sur D. Le champ $Z(.)$ est supposé comme intrinsèquement stationnaire. On pose $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}(\boldsymbol{x}_i - \boldsymbol{x}_j))_{1\leq i,j\leq n}$ et $\boldsymbol{\Gamma}_0 = (\boldsymbol{\gamma}(\boldsymbol{x}_i - \boldsymbol{x}_0))'_{1\leq i\leq n}$. Les contraintes de non-biais et d'autorisation permettent d'ajouter une contrainte sur les paramètres $\boldsymbol{\lambda}$ (voir [Baillargeon, 2005]) :

$$\sum_{i=1}^{n} \lambda_i = 1.$$

La variance de l'erreur de prédiction est alors :

$$Var\left[\hat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)\right] = 2\boldsymbol{\lambda}'\boldsymbol{\Gamma}_0 - \boldsymbol{\lambda}'\boldsymbol{\Gamma}\boldsymbol{\lambda}.$$

Pour satisfaire la contrainte d'optimalité, il est nécessaire de minimiser la variance de prédiction sous la contrainte $\sum_{i=1}^{n} \lambda_i = 1$. Pour cela il est possible de faire appel au multiplicateur de Lagrange ψ. Cette minimisation sous contrainte permet d'aboutir au résultat suivant (voir [Matheron, 1970] pour la preuve détaillée) :

$$\boldsymbol{\lambda} = \boldsymbol{\Gamma}^{-1}(\psi\mathbf{1}_n + \boldsymbol{\Gamma}_0).$$

L'expression du multiplicateur est alors :

$$\psi = \frac{1 - \mathbf{1}_n^{'}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_0}{\mathbf{1}_n^{'}\boldsymbol{\Gamma}^{-1}\mathbf{1}_n}.$$

On en déduit :

$$\boldsymbol{\lambda} = \boldsymbol{\Gamma}^{-1}\left(\frac{1 - \mathbf{1}_n^{'}\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_0}{\mathbf{1}_n^{'}\boldsymbol{\Gamma}^{-1}\mathbf{1}_n}\mathbf{1}_n + \boldsymbol{\Gamma}_0\right).$$

Equations du krigeage ordinaire :
Les équations du krigeage ordinaire sont données par :

$$\begin{cases} \hat{Z}(\boldsymbol{x}_0) = \left(\dfrac{1 - \mathbf{1}_n^{'} \boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}_0}{\mathbf{1}_n^{'} \boldsymbol{\Gamma}^{-1} \mathbf{1}_n} \mathbf{1}_n + \boldsymbol{\Gamma}_0 \right)^{'} \boldsymbol{\Gamma}^{-1} \boldsymbol{Z}, \\ Var\left[\hat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0) \right] = \boldsymbol{\Gamma}_0^{'} \boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}_0 - \dfrac{\left(1 - \mathbf{1}_n^{'} \boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}_0\right)^2}{\mathbf{1}_n^{'} \boldsymbol{\Gamma}^{-1} \mathbf{1}_n}. \end{cases}$$

1.9.5.4 Krigeage universel

Le krigeage universel permet de traiter le cas d'un phénomène présentant une tendance. Le modèle est le suivant :

$$\forall \boldsymbol{x} \in D, Z(\boldsymbol{x}) = \sum_{k=1}^{p} d_k(\boldsymbol{x}) \mu_k + \epsilon(\boldsymbol{x}),$$

avec d_k les p covariables et μ_k les coefficients associés. Les d_k sont considérées connues, tandis que les μ_k sont inconnus. On considère que le champ $Z(.)$ est intrinsèquement stationnaire et que le prédicteur est :

$$\hat{\boldsymbol{Z}}(\boldsymbol{x}_0) = \boldsymbol{D}_0 \boldsymbol{\mu} + \boldsymbol{\lambda}'(\boldsymbol{Z} - \boldsymbol{D}\boldsymbol{\mu}),$$

avec $\boldsymbol{D}_0 = (d_k(\boldsymbol{x}_0))_{1 \leq k \leq p}$. L'expression de la variance de prédiction est alors identique à celle du krigeage ordinaire :

$$Var\left[\hat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0) \right] = 2\boldsymbol{\lambda}' \boldsymbol{\Gamma}_0 - \boldsymbol{\lambda}' \boldsymbol{\Gamma} \boldsymbol{\lambda}.$$

Lors de la minimisation sous contraintes de cette variance de prédiction, les contraintes de non-biais et d'autorisation donnent les relations suivantes sur le vecteur $\boldsymbol{\lambda}$:

$$\begin{cases} \sum_{i=1}^{n} \lambda_i = 1, \\ \boldsymbol{\lambda}' \boldsymbol{D} = \boldsymbol{D}_0^{'}. \end{cases}$$

La méthode de résolution du système est alors identique à celle du krigeage ordinaire, mais faisant cette fois-ci appel à $p+1$ multiplicateurs de Lagrange. Les équations du krigeage universel sont alors :

Equations du krigeage universel :
Les équations du krigeage universel sont données par :

$$\begin{cases} \hat{Z}(\boldsymbol{x}_0) = \left(\boldsymbol{\Gamma}_0 + \boldsymbol{D}(\boldsymbol{D}' \boldsymbol{\Gamma}^{-1} \boldsymbol{D})^{-1} (\boldsymbol{D}_0 - \boldsymbol{D}' \boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}_0) \right)' \boldsymbol{\Gamma}^{-1} \boldsymbol{Z}, \\ Var\left[\hat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0) \right] = \boldsymbol{\Gamma}_0^{'} \boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}_0 - (\boldsymbol{D}_0 - \boldsymbol{D}' \boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}_0)' (\boldsymbol{D}' \boldsymbol{\Gamma}^{-1} \boldsymbol{D})^{-1} (\boldsymbol{D}_0 - \boldsymbol{D}' \boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}_0). \end{cases}$$

Une autre version du krigeage universel appelée krigeage avec dérive externe utilise des covariables ne dépendant pas de la position, mais d'autres variables externes. Les équations correspondantes sont identiques à celle du krigeage universel, à la différence du terme $\boldsymbol{D_0}$.

1.9.5.5 Krigeage par bloc

Plutôt que de prédire une expression ponctuelle, on cherche la valeur moyenne $Z(V)$, aussi appelée régularisée, du champ aléatoire sur un volume $V \subset D$:

$$Z(V) = \frac{1}{|V|} \int_V Z(\boldsymbol{x}) d\boldsymbol{x}.$$

Le variogramme γ_V de bloc obtenu entre un point $\boldsymbol{y} \in D$ et le bloc V est alors défini par :

$$\gamma_V(V, \boldsymbol{y}) = \frac{1}{|V|} \int_V \gamma(\boldsymbol{x} - \boldsymbol{y}) d\boldsymbol{x},$$

avec $|V|$ le volume du bloc.

Le prédicteur est toujours de la forme :

$$\hat{Z}(V) = \sum_{i=1}^{n} \lambda_i Z(\boldsymbol{x}_i).$$

Ainsi lorsqu'un krigeage ordinaire est appliqué, les équations sont identiques à celles du krigeage ordinaire classique, en remplaçant $\boldsymbol{\Gamma}_0$ par $\boldsymbol{\Gamma}_{V,0} = (\boldsymbol{\gamma}_V(V, \boldsymbol{x}_i))'_{1 \leq i \leq n}$.

Equations du krigeage par bloc ordinaire :
Les équations du krigeage par bloc ordinaire sont données par :

$$\begin{cases} \hat{Z}(V) = \left(\frac{1 - \mathbf{1}_n' \boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}_{V,0}}{\mathbf{1}_n' \boldsymbol{\Gamma}^{-1} \mathbf{1}_n} \mathbf{1}_n + \boldsymbol{\Gamma}_{V,0} \right)' \boldsymbol{\Gamma}^{-1} \boldsymbol{Z}, \\ Var\left[\hat{Z}(V) - Z(V) \right] = \boldsymbol{\Gamma}_{V,0}' \boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}_{V,0} - \frac{\left(1 - \mathbf{1}_n' \boldsymbol{\Gamma}^{-1} \boldsymbol{\Gamma}_{V,0}\right)^2}{\mathbf{1}_n' \boldsymbol{\Gamma}^{-1} \mathbf{1}_n}, \end{cases}$$

Le krigeage par bloc est très utile notamment pour prédire la variable d'intérêt sur certains volumes ou surfaces (par exemple pour l'A&D lors de séparation du chantier en différentes zones). Enfin le terme de krigeage par bloc est général et peut exister sous différentes formes, comme le krigeage ordinaire par bloc ou le krigeage simple par bloc.

1.9.5.6 Krigeage avec des incertitudes de mesure

En géostatistique, une des hypothèses de construction du krigeage est qu'il s'agit d'un interpolateur linéaire sans biais, et donc qu'une prédiction sur la position d'une observation aboutit la valeur observée. Cela suppose implicitement que cette observation est faite sans erreur. Ces erreurs sont souvent ignorées lorsqu'elles restent petites devant les variations du champ $Z(.)$. Cependant dans le cas de mesures de radioactivité, ces erreurs entraînent des incertitudes qui peuvent être élevées. Il est alors intéressant de considérer les observations comme la somme du champ que l'on cherche à prédire et d'une erreur de

mesure, indépendante et issue du protocole de mesure. Les n observations peuvent alors s'écrire pour tout $i \in [\![1, n]\!]$ sous la forme suivante :

$$T(\boldsymbol{x}_i) = Z(\boldsymbol{x}_i) + \kappa(\boldsymbol{x}_i)$$

avec $\kappa(\boldsymbol{x}_i)$, $1 \leq i \leq n$, i.i.d. et de loi $\mathcal{N}(0, \sigma_\kappa^2)$ et indépendantes du champ $Z(.)$. On peut donc réécrire l'effet de pépite du champ $T(.)$ comme :

$$\tau_T^2 = \tau^2 + \sigma_\kappa^2$$

avec τ^2 l'effet de pépite de $Z(.)$ et σ_κ^2 l'incertitude de mesure.

On cherche alors à prédire le champ $Z(.)$ à partir des observations de $(T(\boldsymbol{x}_i))_{1 \leq i \leq n}$. Si l'on considère un effet de pépite dont le terme σ_κ^2 a été identifié (par exemple lorsque qu'un mesureur fournit les incertitudes estimées sur un résultat de mesure), les équations de krigeage sont modifiées. Prenons l'exemple du krigeage ordinaire. Son système d'équations de krigeage est :

$$\boldsymbol{\Gamma}\boldsymbol{\lambda}^* = \boldsymbol{\Gamma}_0$$

avec :

$$\begin{cases} \boldsymbol{\Gamma} = \begin{vmatrix} (\gamma(|\boldsymbol{x}_i - \boldsymbol{x}_j|))_{1 \leq i,j \leq n} & \mathbf{1}_n \\ \mathbf{1}_n' & 0 \end{vmatrix}, \\ \boldsymbol{\Gamma}_0 = (\gamma(|\boldsymbol{x}_1 - \boldsymbol{x}_0|), \ldots, \gamma(|\boldsymbol{x}_n - \boldsymbol{x}_0|), 1), \\ \boldsymbol{\lambda}^* = (\lambda_1, \ldots, \lambda_n, \psi)' \end{cases}$$

où ψ est le multiplicateur de Lagrange introduit à la section 1.9.5.3.

Avec des incertitudes de mesure, ce système devient :

$$\boldsymbol{\Gamma}\boldsymbol{\lambda}^* = \boldsymbol{\Gamma}_0^*$$

où $\boldsymbol{\Gamma}_0^* = (\gamma^*(|\boldsymbol{x}_1 - \boldsymbol{x}_0|), \ldots, \gamma^*(|\boldsymbol{x}_n - \boldsymbol{x}_0|), 1)$, avec $\gamma^*(.)$ la fonction telle que :

$$\forall h \in [0, +\infty[, \gamma^*(h) = \begin{cases} \gamma(h) \text{ si } h \neq 0 \\ \sigma_\kappa^2 \text{ si } h = 0. \end{cases}$$

Enfin la variance de prédiction est également affectée et est diminuée des incertitudes de mesure :

$$Var\left[\hat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)\right] = \boldsymbol{\lambda}^{*\prime}\boldsymbol{\Gamma}_0 - \sigma_\kappa^2.$$

Cette modification des équations de krigeage modifie également les propriétés du krigeage, puisqu'avec ce choix de modélisation, le krigeage n'est plus un interpolateur exact. Les prédictions en des points observés peuvent aboutir à des prédictions différentes de l'observation.

En pratique et sans informations supplémentaires sur les incertitudes de mesure, il est impossible de distinguer un champ à micro-échelle présenté dans la section 1.2 des incertitudes de mesure. Ainsi le choix de la composition de τ^2 est alors un choix de modélisation arbitraire, aboutissant à un modèle de krigeage légèrement différent. C'est également pour cette raison que le terme d'effet de pépite est employé pour définir les variations micro-échelles ainsi que les incertitudes de mesure.

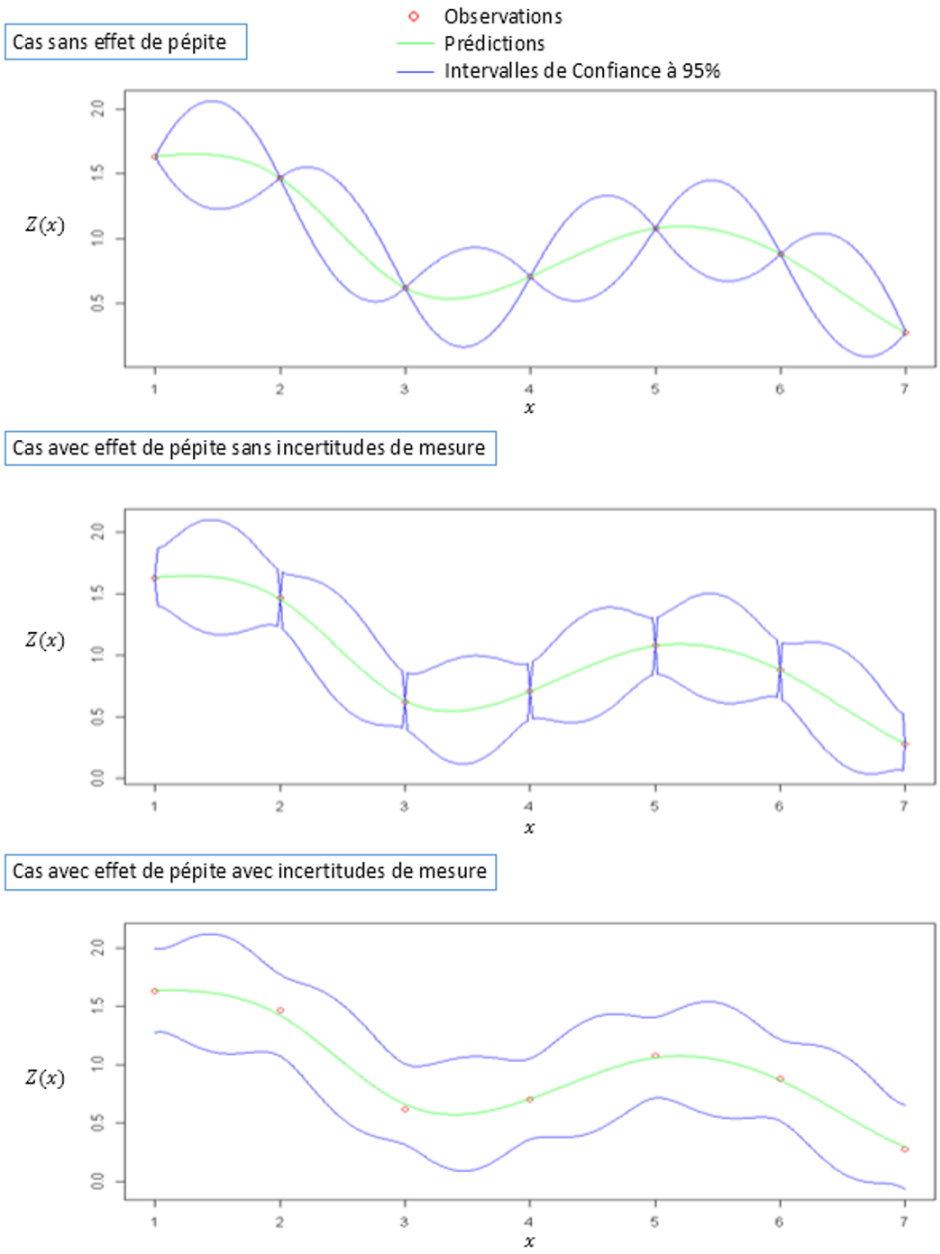


FIGURE 1.7 – Comparaison des différents modèles selon la modélisation de l'effet de pépite.

La Figure 1.7 illustre ces différences dans les prédictions faites par krigeage ordinaire selon le choix de modélisation de l'effet de pépite. Le premier graphe représente les prédictions et les observations dans le cas d'un modèle sans effet de pépite ($\tau_T^2 = 0$). Le second graphe présente un modèle avec effet de pépite comprenant uniquement le variogramme de ξ (voir section 1.2) ($\tau_T^2 = \tau^2$), sans erreurs de mesure. Les prédictions aux observations sont alors identiques aux observations. Enfin le dernier graphe correspond à un modèle avec effet de pépite comprenant uniquement les incertitudes de mesure ($\tau_T^2 = \sigma_\kappa^2$). On remarque que sur ce dernier graphe les prédictions ne sont plus centrées sur les observations.

Une discussion plus approfondie sur ce sujet peut être trouvée dans [Cressie, 1993] et dans [Cressie, 1986]. Le choix de modélisation de l'effet de pépite et des incertitudes de

mesure dépend donc de la situation et des informations à disposition sur les observations, le cas idéal étant d'avoir une estimation des incertitudes de mesure pour chaque observation disponible. En pratique ce n'est pas forcément le cas et les incertitudes de mesure doivent parfois être ré-estimées.

Enfin il est possible de considérer des incertitudes de mesure dépendantes de l'observation (cas hétéroscédastique). Par exemple, le krigeage stochastique ([Ankenman et al., 2010]) ré-estime les incertitudes de fidélité de mesure en chaque point à l'aide de répétitions. Nous ne considérerons pas le cas hétéroscédastique dans la suite, mais renvoyons vers [Binois et al., 2018] pour davantage de détails sur le traitement de ce cas particulier.

1.9.6 Validation du choix de modèle

Le krigeage fait appel au variogramme (ou à la fonction de covariance) pour réaliser ses prédictions. Il est donc nécessaire de vérifier la validité du modèle statistique afin de s'assurer des prédictions réalistes en accord avec le phénomène observé. La validation croisée (voir section 1.6.2) est l'un des outils les plus répandus pour vérifier la validité du modèle ([Chilès and Delfiner, 2012] et [Webster and Oliver, 2007]). La version « Leave-One-Out » est la plus employée en géotatistique.

Une étude des résidus peut ensuite être réalisée pour vérifier quelles sont les limites du modèle. La prédiction de la i^{e} valeur en supprimant la i^{e} donnée du jeu initial est notée $\hat{z}_{-i}$. De même la variance de prédiction sans la i^{e} donnée est notée $\hat{s}^2_{-i}$.

Le coefficient de prédictivité Q^2 (cf. section 1.6.2) est souvent utilisé comme critère de validation d'un prédicteur de krigeage. Sa version en validation croisée est identique à sa version en régression, à la différence que la prédiction usuelle est remplacée par $\hat{z}_{-i}$:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n} \left(z(\boldsymbol{x}_i) - \hat{z}_{-i}\right)^2}{\sum_{i=1}^{n} \left(z(\boldsymbol{x}_i) - \bar{z}\right)^2},$$

avec $\bar{z}$ la moyenne empirique de l'ensemble des observations.

Une des particularités du krigeage est qu'en plus de donner une prédiction, il fournit également une variance sur cette prédiction. Il est donc également possible de quantifier les performances du modèle en utilisant la variance de prédiction fournie par le krigeage. L'adéquation de la variance prédictive (PVA : Predictive Variance Adequacy) est définie de la manière suivante :

Adéquation de la variance prédictive (Predictive Variance Adequacy) [Bachoc, 2013] :
Le PVA est défini par :

$$PVA = \left| \log \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\left(z(\boldsymbol{x}_i) - \hat{z}_{-i}\right)^2}{\hat{s}^2_{-i}} \right) \right|.$$

Plus le PVA est faible, meilleure est la prédiction de la variance.

Ce critère vérifie donc que la variance de prédiction calculée par krigeage est du même ordre que l'écart au carré observé entre les prédictions et les observations. De la même manière que la validation croisée peut se faire sur des k-blocs, l'adéquation de la variance peut être calculée sur différents k-blocs. Ce calcul de variance est important pour valider les intervalles de confiance calculés à partir de la variance de prédiction puisque ces variances déterminent la largeur des intervalles de confiance.

Il est important de noter que le PVA ne prend pas en compte une éventuelle asymétrie dans la loi prédictive. Dans le cas gaussien, la moyenne et la variance décrivent complètement la distribution. Cependant dans le cas du krigeage bayésien, cette distribution n'est pas nécessairement gaussienne. Les critères Q^2 et PVA deviennent donc insuffisants pour quantifier la qualité du modèle et de ses prédictions. En conséquence nous introduisons un nouveau critère (voir [Wieskotten et al., 2022]) appelé "Predictive Interval Adequacy" (PIA) pour "adéquation de l'intervalle prédit" et défini de la manière suivante :

Adéquation de l'intervalle prédit (Predictive Interval Adequacy) :
Le PIA est défini par :

$$PIA = \left| \log \left(\frac{1}{n} \sum_{i=1}^{n} \frac{(z(\boldsymbol{x}_i) - \widehat{z}_{-i})^2}{(\widehat{q}_{0.31,-i} - \widehat{q}_{0.69,-i})^2} \right) \right|,$$

où $\widehat{q}_{0.31,-i}$ (respectivement $\widehat{q}_{0.69,-i}$) est l'estimation du quantile d'ordre 0.31 (respectivement 0.69) de la loi prédictive (à la position $\boldsymbol{x}_i$) sans la i^{e} observation.

Les quantiles $\widehat{q}_{0.31,-i}$ et $\widehat{q}_{0.69,-i}$ ont été choisis puisque dans le cas gaussien, on a :

$$(\widehat{q}_{0.31,-i} - \widehat{q}_{0.69,-i})^2 = \hat{s}^2_{-i}.$$

Le PIA a été donc été défini de telle sorte qu'il soit identique au PVA pour une distribution gaussienne. Contrairement au PVA, le PIA compare la largeur des intervalles de prédiction avec les erreurs quadratiques. De plus le dénominateur n'est plus centré sur la moyenne comme le PVA mais sur la médiane de la distribution. Enfin une estimation de la distribution est nécessaire pour le calcul du PIA (du fait de la présence des quantiles dans son expression), tandis que le PVA ne nécessite que la moyenne et variance de prédiction, rendant ce critère plus complexe à calculer.

Pour juger de la qualité des intervalles de confiance, il est possible de faire appel au graphe α-CI ([Demay et al., 2022]). Le principe de ce critère graphique est d'évaluer empiriquement le nombre d'observations appartenant à l'intervalle de confiance prédit, puis de comparer cette estimation à celle attendue empiriquement :

Graphe α-CI :
Pour $\alpha \in [0,1]$, on introduit :

$$\Delta_\alpha = \frac{1}{n}\sum_{i=1}^{n} \delta_i \text{ où } \delta_i = \begin{cases} 1 & \text{si } z(\boldsymbol{x}_i) \in CI_\alpha(z(\boldsymbol{x}_i)), \\ 0 & \text{sinon,} \end{cases}$$

avec $CI_\alpha(z(\boldsymbol{x}_i))$ l'intervalle de confiance de niveau α prédit pour $\hat{z}_{-i}$. Si la loi prédictive est gaussienne, CI_α est donné par :

$$CI_\alpha(z(\boldsymbol{x}_i)) = [\widehat{z}_{-i} - \widehat{s}_{-i}\, q^{\mathcal{N}}_{\frac{1+\alpha}{2}} \,;\, \widehat{z}_{-i} + \widehat{s}_{-i}\, q^{\mathcal{N}}_{\frac{1+\alpha}{2}}],$$

où $q^{\mathcal{N}}_{\frac{1+\alpha}{2}}$ est le quantile d'ordre $\frac{1+\alpha}{2}$ de la loi gaussienne.

De manière plus générale, si la loi prédictive n'est pas gaussienne, CI_α est alors donné par :

$$CI_\alpha(z(\boldsymbol{x}_i)) = [\widehat{q}_{\frac{1-\alpha}{2}}; \widehat{q}_{\frac{1+\alpha}{2}}],$$

où $\widehat{q}_{\frac{1-\alpha}{2}}$ (respectivement $\widehat{q}_{\frac{1+\alpha}{2}}$) est l'estimation du quantile d'ordre $\frac{1-\alpha}{2}$ (respectivement $\frac{1+\alpha}{2}$) de la distribution prédictive (au point $\boldsymbol{x}_i$) du modèle construit sans la i^{e} observation.

Δ_α est ensuite estimé pour le niveau de confiance α, et est représenté en fonction de α, donnant ce que ([Demay et al., 2022]) appelle le graphe α-CI, dont un exemple est donné dans la Figure 1.8.

Pour obtenir un critère quantitatif sur le graphe α-CI, nous introduisons un nouveau critère calculé à partir du graphe α-CI appelé « Moyenne des erreurs quadratiques α » (Mean Squared Errors α) noté $MSE\alpha$ et défini par :

Moyenne des erreurs quadratiques α (MSEα) :
On considère une discrétisation régulière $\alpha_1 < \ldots < \alpha_{n_\alpha}$ de l'intervalle $]0,1[$. Le $MSE\alpha$ est alors défini par :

$$MSE\alpha = \frac{1}{n_\alpha}\sum_{j=1}^{n_\alpha} (\Delta_{\alpha_j} - \alpha_j)^2.$$

En pratique on considère une discrétisation régulière pour calculer le $MSE\alpha$. Plus ce critère est proche de 0, meilleurs sont les intervalles de confiance en moyenne. Pour illustrer les valeurs prises par ce critère, la Figure 1.8 donne un exemple de de graphe α-CI avec un « bon » et un « mauvais » ajustement du modèle avec les $MSE\alpha$ correspondant. Le « mauvais » modèle (au sens de ce critère) donne un $MSE\alpha$ de 0.035 contre 0.004 pour le « bon » modèle. De manière plus général, un $MSE\alpha$ de 0.01 correspondra à un modèle dont les intervalles de confiance ne sont pas satisfaisants, tandis qu'un $MSE\alpha$ de 0.001 correspondra à un modèle avec de bons intervalles de confiance.

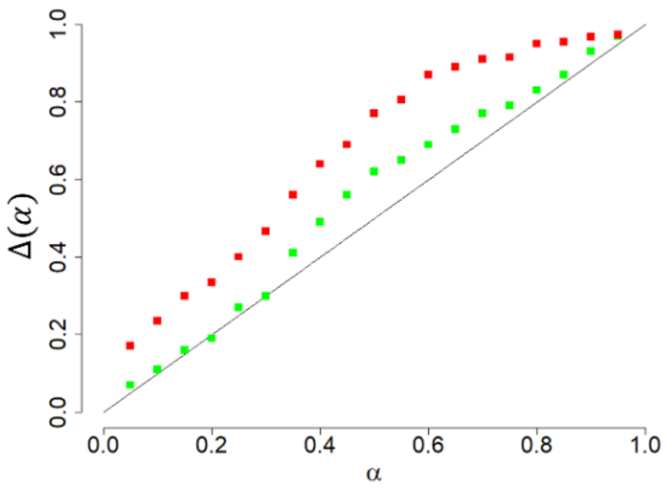


FIGURE 1.8 – Exemple de graphe α-CI avec un bon modèle (en vert) et un mauvais modèle (en rouge) au sens du $MSE\alpha$.

Ce critère permet de différencier différents modèles lorsque les α-CI plots générés sont trop proches graphiquement pour être différenciés, tout en apportant un critère quantitatif sur la qualité des intervalles de confiance.

1.9.7 Simulations conditionnelles

Le krigeage a pour effet de lisser les prédictions et ne représente pas correctement la variabilité réelle du phénomène (quantifiée par la relation de lissage donnée dans [Chilès and Delfiner, 2012]). Pour réintroduire cette variabilité dans les prédictions, on fait appel aux simulations conditionnelles. Ces simulations utilisent le krigeage pour recréer des réalisations du champ aléatoire tout en les conditionnant aux observations. Néanmoins il est important de noter que ces simulations conditionnelles n'offrent que certaines réalisations du champ aléatoire et ne donnent pas directement l'espérance et la variance des variables aléatoires prédites. Ainsi lors d'une analyse géostatistique, krigeage et simulations conditionnelles sont souvent employés simultanément pour obtenir le plus d'informations possibles sur le champ.

Ces simulations doivent vérifier plusieurs propriétés :

— les données simulées doivent avoir la même espérance que le processus original ;
— le variogramme des données simulées doit rester identique à celui de l'original ;
— l'erreur de krigeage doit être nulle en chaque point observé.

Il est important de noter que certaines des simulations conditionnelles ne fonctionnent que pour un processus gaussien. Il est donc souvent nécessaire de transformer les données vers une distribution gaussienne avant d'appliquer les simulations. Les simulations conditionnelles les plus répandues sont la simulation séquentielle gaussienne, le recuit simulé et la simulation par bandes tournantes. Des descriptions de ces algorithmes peuvent être trouvées dans [Webster and Oliver, 2007].

1.9.8 Approche multivariable

La géostatistique multivariable se présente comme une extension de la géostatistique classique et utilise donc des méthodes et objets mathématiques similaires, comme le variogramme et le krigeage. Son intérêt réside dans l'utilisation d'un champ aléatoire plus simple à échantillonner pour prédire un autre champ dont le nombre d'observations est réduit, et ce en utilisant les corrélations entre les deux champs. Cette approche est décrite avec plus de détails dans [Rivoirard, 2003, Wackernagel, 1993].

1.9.8.1 Analyse variographique

On considère toujours notre champ aléatoire $Z(.)$, mais également des champs aléatoires correspondant à n_Y autres variables étudiées $Y_j(.)$. $Z(.)$ est la variable primaire, tandis que les champs $Y_j(.)$ correspondent aux variables secondaires. Pour simplifier les formules on note également $Z(.) = Y_0(.)$.

De plus, les plans d'échantillonnage des différentes variables ne sont pas nécessairement identiques. On note donc pour tout $j \in [\![1, n_Y]\!]$, $\{\boldsymbol{x}_{Y_j,1}, \dots, \boldsymbol{x}_{Y_j,n_{Y_j}}\}$ les points d'observation du champ $Y_j(.)$.

De nouvelles hypothèses de stationnarité sont introduites sur les champs $Y_j(.)$: les hypothèses de stationnarité conjointes.

Stationnarité d'ordre deux conjointe :
Supposons que pour tout $j \in [\![0, n_Y]\!]$, le champ aléatoire $Y_j(.)$ est de carré intégrable. Les champs aléatoires $Y_j(.), j \in [\![0, n_Y]\!]$ sont conjointement stationnaires d'ordre 2 si pour tout $j, k \in [\![0, n_Y]\!]$,

$$\forall \boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h} \in D, Cov(Y_j(\boldsymbol{x}), Y_k(\boldsymbol{x}+\boldsymbol{h})) = C_{j,k}(\boldsymbol{h})$$

où la covariance $C_{j,k}$ ne dépend que de $\boldsymbol{h}$.

Stationnarité intrinsèque multivariable :
Les champs aléatoires $Y_j(.), j \in [\![0, n_Y]\!]$ sont conjointement intrinsèquement stationnaires si pour tout $j, k \in [\![0, n_Y]\!]$:

$$\forall \boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h} \in D, Cov(Y_j(\boldsymbol{x}) - Y_j(\boldsymbol{x}+\boldsymbol{h}), Y_k(\boldsymbol{x}) - Y_k(\boldsymbol{x}+\boldsymbol{h})) = 2\gamma_{j,k}(\boldsymbol{h})$$

où le variogramme croisé $\gamma_{j,k}$ ne dépend que de $\boldsymbol{h}$.

Si n_Y variables secondaires sont disponibles, il est nécessaire d'analyser les $\binom{n_Y}{2} + n_Y + 1$ variogrammes. L'analyse individuelle de ces variogrammes est identique à celle présentée à la section 1.9.4.

Enfin ces covariances et variogrammes vérifient l'inégalité de Cauchy-Schwarz :

$$\begin{cases} C_{j,k}(\boldsymbol{h})^2 \leq C_j(\boldsymbol{h}) C_k(\boldsymbol{h}), \\ \gamma_{j,k}(\boldsymbol{h})^2 \leq \gamma_j(\boldsymbol{h}) \gamma_k(\boldsymbol{h}). \end{cases}$$

La Figure 1.9 illustre cet ajustement simultané des variogrammes, avec ici deux champs aléatoires représentant respectivement des teneurs en uranium et des mesures d'activité surfaciques $\beta\gamma$.

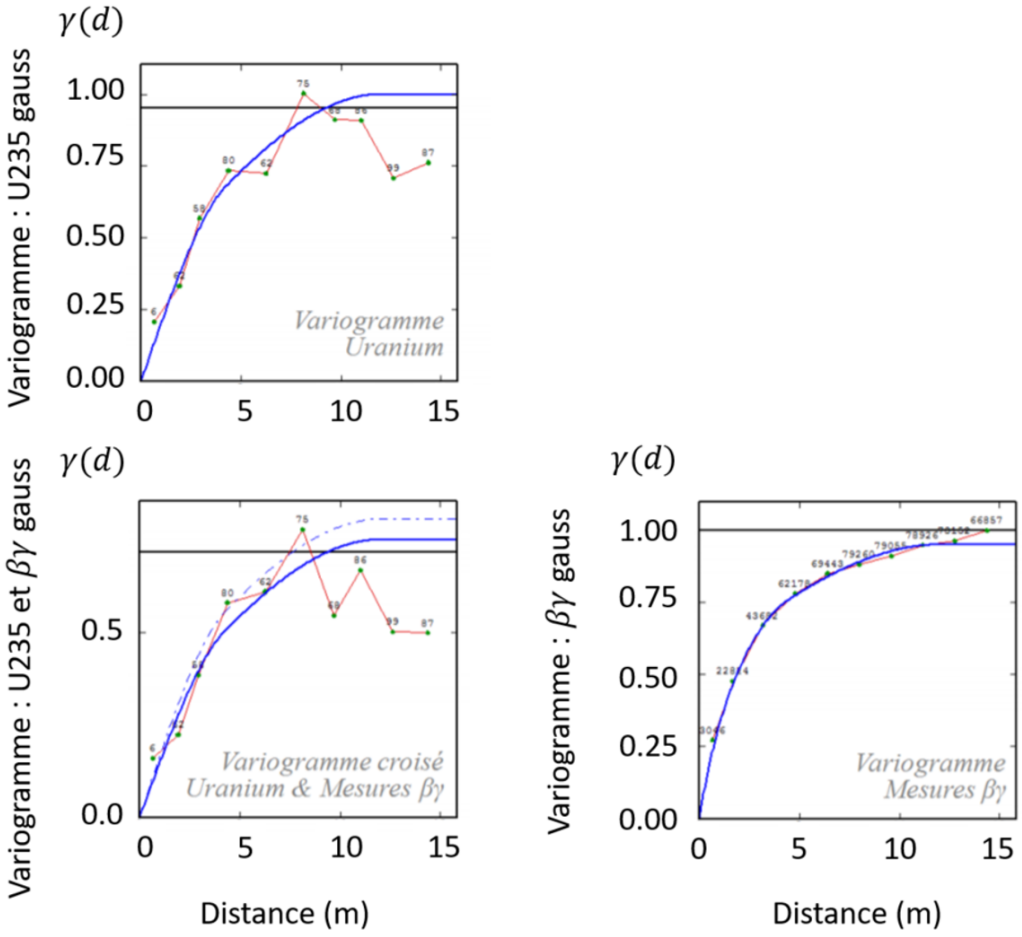


FIGURE 1.9 – Exemple d'analyse variographique multivariable avec en rouge les variogrammes expérimentaux et en bleu les modèles ajustés [Desnoyers, 2010]

1.9.8.2 Cokrigeage

Le cokrigeage est l'extension du krigeage au cas multivariable. Il permet de prédire une variable aléatoire en un point non échantillonné à l'aide des observations ainsi que de celles des autres variables qui lui sont corrélées. Le terme de cokrigeage est général, et plusieurs variantes existent. De la même manière que le krigeage peut être séparé en krigeage simple, ordinaire, universel etc., le cokrigeage existe en version simple, ordinaire, universelle etc. Le prédicteur est de la forme suivante :

$$\hat{Z}(\boldsymbol{x}_0) = \sum_{i=1}^{n} \lambda_{0,i} Z(\boldsymbol{x}_i) + \sum_{j=1}^{n_Y} \sum_{i=1}^{n_{Y_j}} \lambda_{j,i} Y(\boldsymbol{x}_{Y_j,i}).$$

Des informations plus complètes ainsi que des exemples d'application dans le domaine de la modélisation des expériences numériques peuvent être trouvés dans [Le Gratiet, 2013].

1.10 Conclusion

De nombreux outils sont disponibles pour le traitement des données, résultats des mesures réalisées pour l'A&D de sites nucléaires. En particulier, les méthodes multivariables se prêtent bien à ce contexte, puisque de nombreuses grandeurs physiques différentes peuvent être mesurées. Pour une même grandeur physique, il peut exister plusieurs méthodes de mesure (par exemple mesure d'activité surfacique *in-situ* ou en laboratoire). Ces différentes méthodes de mesure utilisées peuvent justifier l'introduction de nouvelles variables aléatoires (ou nouveaux champs aléatoires) et de travailler avec leurs corrélations.

Cependant certains des outils proposés ici sont mis en défaut vis à vis des problématiques évoquées dans la section I.4. Par exemple les estimations par maximum de vraisemblance sont peu robustes et biaisées pour un faible nombre d'observations. En géostatistique, certains problèmes inhérents à la méthodologie classique sont exacerbés dans ce contexte. En particulier dans le krigeage usuel où les paramètres estimés sont directement injectés dans les équations, on suppose implicitement que ces paramètres sont connus. Deux problèmes se combinent ici : l'estimation des paramètres est peu robuste du fait du faible nombre d'observations, et les incertitudes dans leur estimation ne sont pas répercutées dans les prédictions du krigeage. Cela peut donc conduire à des prédictions erronées, à la fois en terme d'espérance que de variance de prédiction.

Pour répondre à cette problématique le krigeage bayésien est introduit, et est l'objet d'étude du chapitre 2.

CHAPITRE 2

Mise en œuvre et étude du krigeage bayésien

Le chapitre précédent a consisté en un bref rappel des outils statistiques et géostatistiques classiques utilisés notamment pour l'A&D de sites nucléaires. Une problématique récurrente en géostatistique concerne les hypothèses implicites des krigeages. En effet les krigeages classiques présentés dans la section 1.9.5 considèrent les paramètres comme connus. Les estimations des paramètres sont donc directement injectées dans les équations de krigeage, ignorant ainsi les incertitudes sur ces estimations. Le krigeage est le meilleur interpolateur linéaire non biaisé si ces paramètres sont connus, mais perd cette propriété importante si les paramètres injectés ne sont pas appropriés. De manière générale, les paramètres sont inconnus et estimés à partir des données disponibles. Ainsi les variances de prédiction obtenues avec un krigeage classique peuvent sous-estimer les variances de prédiction réelles en ne prenant pas en compte les incertitudes sur les estimations des paramètres.

Une solution à ce problème fait appel à l'approche bayésienne, et en particulier au krigeage bayésien. L'une des premières formalisations du krigeage bayésien se trouve dans [Kitanidis, 1986] dont l'objectif était la prise en compte d'incertitudes dans l'estimation des paramètres de moyenne et de variance. Les travaux de [Omre, 1987] présentent le lien entre krigeage et modèle linéaire bayésien classique en considérant des incertitudes dans l'estimation de la moyenne. Dans un autre article, [Omre and Halvorsen, 1989] fait le lien entre krigeage simple et krigeage universel en considérant le krigeage bayésien comme un pont marquant la connaissance plus ou moins exacte de la moyenne.

L'approche bayésienne est ensuite complétée dans [Handcock and Stein, 1993] qui ajoute une incertitude sur le paramètre de portée. Les travaux de [De Oliveira et al., 1997] étendent l'application du krigeage bayésien en considérant le krigeage transgaussien. Dans ce cas les données sont transformées en utilisant la famille de transformations de Box-Cox, et un paramètre de transformation est ajouté à l'ensemble des paramètres à estimer. Plus récemment, l'approche empirique bayésienne pour le krigeage bayésien est utilisée pour donner une approche multi-fidélité dans [De Risi et al., 2021] pour modéliser des données sismiques. Ici les paramètres des lois *a priori* sont estimés à partir d'un jeu de données initial peu informatif, puis sont mis à jour à l'aide du théorème de Bayes. Cette approche rentre dans le cas du krigeage empirique bayésien présenté dans [Krivoruchko and Gribov,

2019], qui construit des lois *a priori* en utilisant des simulations non-conditionnelles (issues d'une estimation initiale des paramètres). Ce krigeage empirique bayésien permet d'éviter le choix parfois difficile de lois *a priori*. Enfin, le krigeage bayésien est présenté dans [Diggle and Ribeiro, 2007] avec une extension au modèle linéaire généralisé. De nombreuses applications récentes existent, comme par exemple dans le domaine de l'extraction pétrolière [Al-Mudhafar, 2019] ou celui de la météorologie [Gupta et al., 2017].

Pour décrire le krigeage bayésien, nous revenons d'abord rapidement sur le théorème de Bayes. Ensuite nous explicitons les équations du krigeage bayésien selon les hypothèses faites sur les paramètres. Lorsque tous les paramètres sont considérés inconnus, ces équations ne possèdent plus de forme explicite. Nous détaillons alors plusieurs méthodes pour la résolution numérique de ces équations, tout en comparant deux de ces approches. Nous revenons ensuite sur le choix des spécifications de ces méthodes (propre à la méthode ou au choix de lois *a priori*). Enfin nous comparons le krigeage bayésien avec le krigeage ordinaire (la méthode usuelle industrielle) pour des jeux de données simulés et un jeu de données réelles.

2.1 Le théorème de Bayes

Le théorème de Bayes est le principe fondamental du cadre bayésien. Son expression permet d'inverser les conditions sur la probabilité d'un évènement, et permet de formuler une expression dite *a posteriori*.

Théorème de Bayes :
Soient deux variables aléatoires X et Y à valeurs dans D_X et D_Y respectivement, dont la loi jointe de (X, Y) a pour densité $f_{X,Y}$. Nous notons f_X et f_Y les densités respectives de X et Y. Le théorème donne :

$$f_{X|Y}(x|Y = y) = \frac{f_{Y|X}(y|X = x)f_X(x)}{f_Y(y)}.$$

avec en convention $\forall x, y \in D_X \times D_Y, f_{X|Y}(x|Y = y) = 0$ si $f_X(x) = 0$ ou $f_Y(y) = 0$.

$f_{X|Y}(.)$ est appelée la densité *a posteriori* de X et $f_X(.)$ est la densité *a priori* de X.

Pour l'inférence bayésienne, l'estimation des paramètres est différente de l'inférence statistique classique puisque cette estimation correspond à une « mise à jour » de la loi *a priori* sur le ou les paramètres recherchés. Cette mise à jour est réalisée à partir d'observations pour formuler une loi *a posteriori*. Si l'on considère qu'un paramètre à estimer est une variable aléatoire, il est alors possible de lui appliquer le théorème de Bayes, selon une loi *a priori*. Si l'on note θ le paramètre à estimer, $\boldsymbol{z}$ le vecteur contenant les n observations des variables $Z_1, \ldots, Z_n$ i.i.d., on a :

$$f_\theta(\theta|\boldsymbol{z}) = \frac{L(\boldsymbol{z};\theta)\pi(\theta)}{f_Z(\boldsymbol{z})}.$$

où $\pi(\theta)$ est la densité de la loi *a priori* de ϕ et $L(\boldsymbol{z};\theta)$ est la vraisemblance définie dans la section 1.2.2.

Le réel $f_Z(z)$ étant indépendant de θ, ce terme est considéré comme une constante de normalisation. Le théorème de Bayes est donc souvent réécrit sous la forme suivante :

$$f_\theta(\theta|\boldsymbol{z}) \propto f_Z(\boldsymbol{z}|\theta)\pi(\theta).$$

Le théorème de Bayes permet d'obtenir la distribution de θ (conditionnée aux observations) plutôt qu'une valeur unique, contrairement aux estimateurs classiques comme celui du maximum de vraisemblance. Il est alors possible de donner une valeur à θ en prenant le mode ou l'espérance de $f(\theta|\boldsymbol{z})$ et de donner un intervalle de crédibilité (équivalent à un intervalle de confiance) sur cette estimation (voir [Carlin and Louis, 2009]).

2.2 Prédictions par krigeage bayésien

Dans la suite, le champ aléatoire considéré est noté :

$$\{Z(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^d\}.$$

Pour un ensemble de points $\boldsymbol{x_1}, ..., \boldsymbol{x_n}$ et $\boldsymbol{Z} = (Z(\boldsymbol{x_1}), ..., Z(\boldsymbol{x_n}))'$, on a :

$$\boldsymbol{Z}|\boldsymbol{\mu}, \sigma^2, \phi, \tau^2 \sim \mathcal{N}_n(\boldsymbol{D\mu}, \boldsymbol{R}(\sigma^2, \phi, \tau^2)),$$

où l'on rappelle que

$$\begin{cases} \boldsymbol{R}(\sigma^2, \phi, \tau^2) = ((Cov(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j))))_{1\leq i,j\leq n} = \sigma^2 \left(\boldsymbol{V}(\phi) + \frac{\tau^2}{\sigma^2}\boldsymbol{I}\right) = \sigma^2\boldsymbol{\Sigma}(\phi, \tau^2) \\ \boldsymbol{D} = (d_i(\boldsymbol{x}_j))_{1\leq i\leq p, 1\leq j\leq n} \end{cases}$$

avec d_i des covariables connues.

Pour simplifier l'écriture, la dépendance des matrices de covariance aux paramètres sera omise. La vraisemblance se met alors sous la forme :

$$L(\boldsymbol{z}; \boldsymbol{\mu}, \sigma^2, \phi, \tau^2) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{R}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{D\mu})'\boldsymbol{R}^{-1}(\boldsymbol{z} - \boldsymbol{D\mu})\right).$$

Dans le cas du krigeage bayésien, les paramètres du modèle ne sont plus considérés comme des scalaires, mais comme des variables aléatoires. En utilisant le théorème de l'espérance totale et de la variance totale, on obtient les équations du krigeage bayésien.

Equations du krigeage bayésien [Kitanidis, 1986] :
Soit $Z(\boldsymbol{x}_0)$ la variable aléatoire que l'on souhaite prédire. Son espérance par krigeage bayésien est donnée par :

$$\mathbb{E}[Z(\boldsymbol{x}_0)|\boldsymbol{z}] = \mathbb{E}_{\boldsymbol{\theta}}[\mathbb{E}[Z(\boldsymbol{x}_0)|\boldsymbol{z}, \boldsymbol{\theta}]] = \int_\Theta \mathbb{E}[Z(\boldsymbol{x}_0)|\boldsymbol{z}, \boldsymbol{\theta}] f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{z}) d\boldsymbol{\theta},$$

avec $\boldsymbol{\theta'} = (\boldsymbol{\mu}, \sigma^2, \phi, \tau^2)$ et $\Theta = \mathbb{R}^p \times]0, +\infty[^2$.

De même en utilisant la formule de la variance complète (ou totale) :

$$\begin{aligned} Var[Z(\boldsymbol{x}_0)|\boldsymbol{z}] &= \mathbb{E}_{\boldsymbol{\theta}}[Var[Z(\boldsymbol{x}_0)|\boldsymbol{z},\boldsymbol{\theta}]] + Var_{\boldsymbol{\theta}}[\mathbb{E}[Z(\boldsymbol{x}_0)|\boldsymbol{z},\boldsymbol{\theta}]] \\ &= \int_\Theta Var[Z(\boldsymbol{x}_0)|\boldsymbol{z},\boldsymbol{\theta}] f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{z}) d\boldsymbol{\theta} \\ &+ \int_\Theta (\mathbb{E}[Z(\boldsymbol{x}_0)|\boldsymbol{z},\boldsymbol{\theta}] - \mathbb{E}[Z(\boldsymbol{x}_0)|\boldsymbol{z}])^2 f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{z}) d\boldsymbol{\theta}. \end{aligned}$$

Ces expressions peuvent ensuite être explicitées en utilisant les équations du krigeage simple :

$$\mathbb{E}[Z(\boldsymbol{x}_0)|\boldsymbol{z}] = \int_\Theta (\boldsymbol{D}_0\boldsymbol{\mu} + \boldsymbol{r}'\boldsymbol{R}^{-1}(\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\mu})) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{z}) d\boldsymbol{\theta},$$

et

$$\begin{aligned} Var[Z(\boldsymbol{x}_0)|\boldsymbol{z}] &= \int_\Theta (\sigma^2 - \boldsymbol{r}'\boldsymbol{R}^{-1}\boldsymbol{r}) f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{z}) d\boldsymbol{\theta} \\ &+ \int_\Theta (\boldsymbol{D}_0\boldsymbol{\mu} + \boldsymbol{r}'\boldsymbol{R}^{-1}(\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\mu}) - \mathbb{E}[Z(\boldsymbol{x}_0)|\boldsymbol{z}])^2 f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{z}) d\boldsymbol{\theta}, \end{aligned}$$

où l'on rappelle que $\boldsymbol{D}_0 = (d_i(\boldsymbol{x}_0))_{1\leq i\leq p}$ et $\boldsymbol{r} = (Cov(Z(\boldsymbol{x}_0), Z(\boldsymbol{x}_i)))_{1\leq i\leq n}$.

Les équations sont données pour un seul point à prédire, mais peuvent être facilement étendues au cas d'un vecteur à prédire.

Les équations du krigeage bayésien peuvent être interprétées comme une moyenne réalisée sur toute les prédictions obtenues pour chaque jeu de paramètres possibles. Cette moyenne est pondérée par la densité *a posteriori* des paramètres. Dans le cas de la variance de prédiction, le second terme vient quantifier l'incertitude sur l'estimation des paramètres. Ce terme étant positif, les variances d'estimation du krigeage bayésien sont en moyenne plus grandes que celles obtenues avec d'autres krigeages classiques. Il est important de noter que selon l'expression de la densité *a posteriori* et des hypothèses faites sur les paramètres, l'évaluation de ces intégrales peut être difficile. Pour décrire ces équations plus en détails, nous commencerons par un cas simple où seule la moyenne est considérée inconnue. Nous considérerons ensuite des cas de plus en plus complexes, jusqu'à obtenir le cas où tous les paramètres du modèle sont inconnus. Les prochaines sections sont inspirées de [Diggle and Ribeiro, 2002].

La différence classique présente en géostatistique entre estimation des paramètres et prédictions n'est plus aussi claire pour le krigeage bayésien. Le terme de lois *a posteriori* correspond à la fois aux paramètres et aux prédictions. Dans la suite, lorsque nous évoquerons la loi *a posteriori*, nous parlerons de la loi *a posteriori* des paramètres. Quant à la loi *a posteriori* des prédictions, nous la nommons loi prédictive.

2.3 Paramètre de moyenne inconnu

2.3.1 Loi *a posteriori*

La loi multivariée gaussienne conjuguée est le choix usuel pour la loi *a priori* sur $\boldsymbol{\mu}$:

$$\boldsymbol{\mu}|\sigma^2 \sim \mathcal{N}_p(\boldsymbol{m}_0, \sigma^2 \boldsymbol{R}_0).$$

La loi *a posteriori* est également une loi gaussienne [Diggle and Ribeiro, 2002] de paramètres :

$$\boldsymbol{\mu}|\boldsymbol{z}, \sigma^2, \phi \sim \mathcal{N}_p((\boldsymbol{R}_0^{-1} + \boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})^{-1}(\boldsymbol{R}_0^{-1}\boldsymbol{m}_0 + \boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z}), \sigma^2(\boldsymbol{R}_0^{-1} + \boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})^{-1}).$$

On se place maintenant dans le cas d'une loi *a priori* impropre :

$$\pi(\boldsymbol{\mu}) \propto 1.$$

Dans ce cas la loi *a posteriori* est :

$$\boldsymbol{\mu}|\boldsymbol{z}, \sigma^2, \phi \sim \mathcal{N}_p((\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z}, \sigma^2(\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})^{-1}).$$

On récupère ici l'estimateur de la moyenne par moindres carrés généralisés :

$$\hat{\mu}_{\boldsymbol{\Sigma}} = (\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z}.$$

2.3.2 Loi prédictive

La densité de la loi prédictive est ensuite donnée par la relation suivante :

$$f_{Z(\boldsymbol{x}_0)}(z(\boldsymbol{x}_0)|\boldsymbol{z}, \sigma^2, \phi) = \int_{\mathbb{R}^p} f_{Z(\boldsymbol{x}_0)}(z(\boldsymbol{x}_0)|\boldsymbol{z}, \boldsymbol{\mu}, \sigma^2, \phi) f_{\boldsymbol{\mu}}(\boldsymbol{\mu}|\boldsymbol{z}, \sigma^2, \phi) d\boldsymbol{\mu}.$$

Dans le cas d'une loi *a priori* gaussienne, on obtient à nouveau une loi gaussienne :

$$Z(x_0)|\boldsymbol{z}, \sigma^2, \phi \sim \mathcal{N}(m_1, R_1),$$

avec :

$$\begin{cases} m_1 = (\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})\boldsymbol{H}\boldsymbol{R}_0^{-1}\boldsymbol{m}_0 + (\boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1} + (\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})\boldsymbol{H}\boldsymbol{D}'\boldsymbol{\Sigma}^{-1})\boldsymbol{z} \\ R_1 = \sigma^2 \left(1 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0 + (\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})\boldsymbol{H}(\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})'\right). \end{cases}$$

où $\boldsymbol{H} = (\boldsymbol{R}_0^{-1} + \boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})^{-1}$ et $\sigma^2\boldsymbol{\Sigma}_0 = (Cov(Z(\boldsymbol{x}_0), Z(\boldsymbol{x}_i)))_{1 \leq i \leq n}$.

Si l'on considère une loi *a priori* impropre, on récupère alors à nouveau une loi gaussienne paramétrée par :

$$\begin{cases} m_1 = (\boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1} + (\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})(\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{\Sigma}^{-1})\boldsymbol{z}, \\ R_1 = \sigma^2 \left(1 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0 + (\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})(\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})^{-1}(\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})'\right). \end{cases}$$

2.3.3 Lien avec la géostatistique classique

Les équations données ici sont une généralisation des équations de krigeage lorsqu'une incertitude est autorisée pour le paramètre de moyenne. Si l'on considère que le paramètre de moyenne est connu, les paramètres m_1, R_1 se simplifient :

$$\begin{cases} m_1 = (\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0' \boldsymbol{\Sigma}^{-1} \boldsymbol{D})\boldsymbol{\mu} + \boldsymbol{\Sigma}_0' \boldsymbol{\Sigma}^{-1} \boldsymbol{z} \\ R_1 = \sigma^2 \left(1 - \boldsymbol{\Sigma}_0' \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_0\right). \end{cases}$$

On retrouve ici les équations du krigeage simple, qui suppose que tous les paramètres sont connus.

Si l'on considère la loi *a priori* impropre $\pi(\boldsymbol{\mu}) \propto 1$ (qui revient à considérer la moyenne comme inconnue), on obtient alors les équations du krigeage universel (ou celle du krigeage ordinaire si $\boldsymbol{D}\boldsymbol{\mu} = \mathbf{1}\mu$), qui estime le ou les paramètres de moyenne à l'aide de l'estimateur des moindres carrés généralisés :

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\Sigma}} = (\boldsymbol{D}' \boldsymbol{\Sigma}^{-1} \boldsymbol{D})^{-1} \boldsymbol{D}' \boldsymbol{\Sigma}^{-1} \boldsymbol{z}.$$

La loi *a posteriori* obtenue pour $\boldsymbol{\mu}$ correspond alors à celle de l'estimateur des moindres carrés généralisés :

$$\boldsymbol{\mu}_{\boldsymbol{\Sigma}} | \sigma^2, \phi \sim \mathcal{N}_p(\hat{\boldsymbol{\mu}}, \sigma^2 (\boldsymbol{D}' \boldsymbol{\Sigma}^{-1} \boldsymbol{D})^{-1}).$$

Cette relation entre krigeage bayésien, krigeage simple et krigeage universel est décrit en détails chez [Omre and Halvorsen, 1989] et une comparaison entre krigeage universel et krigeage bayésien est réalisée par [Helbert et al., 2009].

2.4 Paramètres de moyenne et de variance inconnus

2.4.1 Loi *a posteriori*

La loi gaussienne est conservée pour $\boldsymbol{\mu}$. Pour le paramètre σ^2, la loi usuelle est la loi inverse-gamma, souvent paramétrée sous la forme d'une loi inverse du χ^2 pour faciliter l'interprétation des choix des paramètres *a priori*. La loi jointe *a priori* est alors explicitée par l'expression suivante :

$$\pi(\boldsymbol{\mu}, \sigma^2 | \phi) = \pi(\boldsymbol{\mu} | \sigma^2, \phi) \pi(\sigma^2 | \phi)$$

et

$$\boldsymbol{\mu} | \sigma^2, \phi \sim \mathcal{N}_p(\boldsymbol{m}_0, \boldsymbol{R}_0) \ , \ \sigma^2 | \phi \sim \mathcal{IG}\left(a_\sigma, S_\sigma^2\right),$$

où $\mathcal{IG}$ désigne la loi inverse-gamma. On rappelle que $X \sim \mathcal{IG}(a_\sigma, S_\sigma^2)$ a pour densité la fonction f_X définie par

$$f_X(x) = \frac{(S_\sigma^2)^{a_\sigma}}{\Gamma(a_\sigma)} \left(\frac{1}{x}\right)^{a_\sigma + 1} \exp\left(-\frac{S_\sigma^2}{x}\right), \ x > 0.$$

La loi *a posteriori* est une loi issue du produit d'une loi multivariée gaussienne et d'une loi inverse-gamma notée $\mathcal{N}_p\mathcal{IG}$:

$$\boldsymbol{\mu}, \sigma^2 | \boldsymbol{z}, \phi \sim \mathcal{N}_p\mathcal{IG}(\boldsymbol{m}_1, \boldsymbol{R}_1, c_1, S_1^2) \tag{2.1}$$

avec

$$\begin{cases} c_1 = 2a_\sigma + n, \\ \boldsymbol{R}_1 = (\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D} + \boldsymbol{R}_0^{-1})^{-1}, \\ \boldsymbol{m}_1 = \boldsymbol{R}_1(\boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D} + \boldsymbol{m}_0\boldsymbol{R}_0^{-1}), \\ S_1^2 = c_1^{-1}\left(2S_\sigma^2 + \boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{z} + \boldsymbol{m}_0'\boldsymbol{R}_0^{-1}\boldsymbol{m}_0 - \boldsymbol{m}_1'\boldsymbol{R}_1^{-1}\boldsymbol{m}_1\right). \end{cases}$$

Dans le cas de loi *a priori* impropre, on a classiquement :

$$\pi(\boldsymbol{\mu}, \sigma^2) \propto \frac{1}{\sigma^2}.$$

De plus le nombre de degrés de liberté de la loi *a posteriori* est alors égal à $n - p$. On obtient à nouveau une loi gaussienne inverse-gamma, paramétrée par :

$$\begin{cases} c_1 = n - p, \\ \boldsymbol{R}_1 = (\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})^{-1}, \\ \boldsymbol{m}_1 = \boldsymbol{R}_1(\boldsymbol{z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D}), \\ S_1^2 = c_1^{-1}\left(\boldsymbol{z} - \boldsymbol{D}'\boldsymbol{m}_1\right)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{z} - \boldsymbol{D}'\boldsymbol{m}_1). \end{cases}$$

2.4.2 Loi prédictive

De manière similaire au cas précédent, cette prédiction est obtenue à l'aide de la loi *a posteriori* :

$$f_{Z(\boldsymbol{x}_0)}(z(\boldsymbol{x}_0)|\boldsymbol{z}, \phi) = \int_{]0,+\infty[}\int_{\mathbb{R}^p} f_{Z(\boldsymbol{x}_0)}(z(\boldsymbol{x}_0)|\boldsymbol{z}, \boldsymbol{\mu}, \sigma^2, \phi) f_{\boldsymbol{\mu},\sigma^2}(\boldsymbol{\mu}, \sigma^2|\boldsymbol{z}, \phi) d\boldsymbol{\mu} d\sigma^2.$$

La densité obtenue correspond à celle d'une loi de Student à c_1 degrés de liberté, notée $\mathcal{T}_{c_1}$:

$$Z(\boldsymbol{x}_0)|\boldsymbol{z}, \phi \sim \mathcal{T}_{c_1}(m_2, S_1^2 R_2) \tag{2.2}$$

dont les paramètres sont

$$\begin{cases} m_2 = (\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})\boldsymbol{H}\boldsymbol{R}_0^{-1}\boldsymbol{m}_0 + (\boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1} + (\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})\boldsymbol{H}\boldsymbol{D}'\boldsymbol{\Sigma}^{-1})\boldsymbol{z} \\ R_2 = 1 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0 + (\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})'(\boldsymbol{R}_0^{-1} + \boldsymbol{R}_1^{-1})^{-1}(\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{D}) \end{cases}$$

où $\boldsymbol{H} = (\boldsymbol{R}_0^{-1} + \boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})^{-1}$.

Dans le cas d'une loi *a priori* impropre $f_{\boldsymbol{\mu},\sigma^2}(\boldsymbol{\mu}, \sigma^2|\phi) \propto \frac{1}{\sigma^2}$, le nombre de degrés de liberté est égal à $n - p$ avec p la dimension de $\boldsymbol{\mu}$ [Diggle and Ribeiro, 2007]. On trouve donc une loi de Student à $n - p$ degrés de liberté, paramétrée de la manière suivante :

$$Z(\boldsymbol{x}_0)|\boldsymbol{z}, \phi \sim \mathcal{T}_{c_1}(m_2, S_1^2 R_2)$$

avec :

$$\begin{cases} c_1 = n - p \\ m_2 = (\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})\boldsymbol{m}_1 + \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{z} \\ R_2 = 1 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_0 + (\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})'\boldsymbol{H}(\boldsymbol{D}_0 - \boldsymbol{\Sigma}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{D}) \end{cases}$$

où $\boldsymbol{H} = (\boldsymbol{R}_0^{-1} + \boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D})^{-1}$.

2.5 Paramètres de moyenne, de variance et de portée inconnus

2.5.1 Loi *a posteriori*

De manière plus générale, ϕ est désormais également considéré comme inconnu. Plutôt que de formuler le problème de manière similaire aux deux sections précédentes, on considère la loi *a posteriori* de la portée uniquement. En effet ce paramètre est indépendant des autres, et l'expression de la densité *a priori* jointe de la moyenne et la variance étant connue, on peut écrire :

$$\pi(\boldsymbol{\mu}, \sigma^2, \phi) = \pi(\boldsymbol{\mu}, \sigma^2|\phi)\pi(\phi),$$

avec $\pi(\phi)$ la densité de la loi *a priori* sur la portée. Cette formule permet ensuite d'exprimer la densité *a posteriori* :

$$f_\phi(\phi|\boldsymbol{z}) \propto \frac{\pi(\boldsymbol{\mu}, \sigma^2|\phi)\pi(\phi)L(\boldsymbol{z}; \boldsymbol{\mu}, \sigma^2, \phi)}{f_{\boldsymbol{\mu},\sigma^2}(\boldsymbol{\mu}, \sigma^2|\boldsymbol{z}, \phi)} \tag{2.3}$$

où $f_{\boldsymbol{\mu},\sigma^2}(\boldsymbol{\mu}, \sigma^2|\boldsymbol{z}, \phi)$ correspond à la densité *a posteriori* de la moyenne et de la variance donnée dans l'équation (2.1), tandis que $\pi(\boldsymbol{\mu}, \sigma^2|\phi)$ correspond à la loi *a priori* donnée dans la section 2.4.

2.5.2 Loi prédictive

La densité de la loi prédictive peut donc être écrite sous la forme d'une intégrale du produit de la densité de la loi prédictive sachant la portée donnée dans l'équation 2.2 avec la densité *a posteriori* de la portée :

$$f_{Z(\boldsymbol{x}_0)}(z(\boldsymbol{x}_0)|\boldsymbol{z}) = \int_{]0,+\infty[} f_{Z(\boldsymbol{x}_0)}(z(\boldsymbol{x}_0)|\boldsymbol{z}, \phi) f_\phi(\phi|\boldsymbol{z}) d\phi \tag{2.4}$$

Cette intégrale ne dispose généralement pas d'une expression explicite, la densité de la loi *a posteriori* des paramètres n'étant pas usuelle. Des approximations numériques sont donc nécessaires pour obtenir la loi prédictive. Pour cela plusieurs approches sont disponibles. La première consiste à réaliser une approximation par méthodes de Monte-Carlo de cette intégrale en échantillonnant la densité *a posteriori* des paramètres (voir [Diggle and Ribeiro, 2007]). Une autre approche échantillonne la densité *a posteriori* à l'aide d'un algorithme MCMC de type Metropolis-Hasting. Enfin d'autres approches font appel à l'analyse de la vraisemblance pour certains choix de lois *a priori* pour éviter de faire appel à des algorithmes MCMC (par exemple [Gu et al., 2018]). Des solutions plus récentes utilisent des équations différentielles stochastiques (voir [Lindgren and Rue, 2015] et [Blangiardo et al., 2013] pour un cas spatio-temporel) pour construire les prédictions par krigeage bayésien. Dans la suite nous étudierons les deux premières approches.

2.6 Krigeage bayésien avec effet de pépite inconnu

Pour compléter le modèle on ajoute l'effet de pépite τ^2. Ce paramètre étant considéré comme indépendant des autres paramètres, il n'apporte pas de difficultés particulières lors

de sa prise en compte. On considère la densité jointe *a priori* de la portée ϕ et de l'effet de pépite τ^2 :

$$\pi(\boldsymbol{\mu}, \sigma^2, \phi, \tau^2) = \pi(\boldsymbol{\mu}, \sigma^2|\phi, \tau^2)\pi(\phi)\pi(\tau^2),$$

avec $\pi(\tau^2)$ la densité *a priori* sur l'effet de pépite. On obtient alors la densité *a posteriori* :

$$f_{\phi,\tau^2}(\phi, \tau^2|\boldsymbol{z}) \propto \frac{\pi(\boldsymbol{\mu}, \sigma^2|\phi, \tau^2)\pi(\phi)\pi(\tau^2)L(\boldsymbol{z}; \boldsymbol{\mu}, \sigma^2, \phi, \tau^2)}{f_{\boldsymbol{\mu},\sigma^2}(\boldsymbol{\mu}, \sigma^2|\boldsymbol{z}, \phi, \tau^2)}.$$

La densité de la loi prédictive est donnée par la double intégrale suivante :

$$f_{Z(\boldsymbol{x}_0)}(z(\boldsymbol{x}_0)|\boldsymbol{z}) = \int_{]0,+\infty[}\int_{]0,+\infty[} f_{Z(\boldsymbol{x}_0)}(z(\boldsymbol{x}_0)|\boldsymbol{z}, \phi, \tau^2) f_{\phi,\tau^2}(\phi, \tau^2|\boldsymbol{z}) d\phi d\tau^2.$$

Cette intégrale pose des difficultés similaires à l'intégrale de l'équation 2.4. Les méthodes présentées dans la section 2.5 peuvent également être appliquées en échantillonnant un paramètre supplémentaire. Cet ajout de paramètre vient cependant au prix de temps de calculs plus élevés et d'une spécification supplémentaire de la loi *a priori* de l'effet de pépite.

2.7 Approche Monte-Carlo pour le krigeage bayésien

L'intégrale présente dans l'expression de la densité de la loi prédictive peut être approchée par une méthode de Monte-Carlo. Cette approche (présentée dans [Diggle and Ribeiro, 2002]) est disponible dans le package `geoR` de `R` et consiste à venir échantillonner les différentes valeurs possibles des paramètres, puis à les injecter dans les équations présentées précédemment afin de réaliser un échantillonnage de la loi prédictive.

Krigeage bayésien avec approche Monte-Carlo (`geoR`) :

- **Discrétisation de la loi *a priori*** : la densité $\pi(\phi)$ est discrétisée sur un ensemble de valeurs raisonnables pour ce paramètre.
- **Calcul de la loi *a posteriori de* ϕ** : à partir de l'équation (2.3), la loi approchée discrète $\widehat{f}_\phi(\phi|\boldsymbol{z})$ de $f_\phi(\phi|\boldsymbol{z})$ est estimée.
- **Tirages aléatoires** : Ces étapes sont répétées M fois :
  - Échantillonnage de ϕ depuis $\widehat{f}_\phi(\phi|\boldsymbol{z})$,
  - Échantillonnage de $Z(\boldsymbol{x}_0)$ depuis la loi donnée à l'équation (2.4) en injectant la valeur de ϕ précédemment tirée.
- **Inférences sur l'échantillon obtenu** : les caractéristiques de la loi prédictive sont alors obtenues par inférence statistique sur l'échantillon obtenu.

Enfin l'effet de pépite peut être pris en compte en considérant la loi *a priori* jointe $\pi(\phi, \tau^2) = \pi(\phi)\pi(\tau^2)$ puis en la discrétisant de manière similaire. A nouveau cela implique une approximation supplémentaire dans la modélisation de l'effet de pépite, et des travaux

supplémentaires pour identifier un ensemble raisonnable de valeurs pour cette approximation discrète. [Berger et al., 2001] indique que les espérances et variances de prédiction sont sensibles à ces discrétisations supplémentaires. Cela complexifie donc fortement l'analyse de sensibilité aux choix de loi *a priori*.

Le choix d'une approximation discrète de la densité $\pi(\phi)$ (ou de la densité $\pi(\phi, \tau^2)$) permet de calculer analytiquement les moments de la loi prédictive, qui est alors une somme de lois de Student (voir section 2.5). En pratique, il est possible d'éviter cet échantillonnage en approximant les lois de Student. Ces lois étant, avec suffisamment de données, proches de lois gaussiennes, leur somme forme une distribution proche d'une distribution gaussienne. Les moments étant connus à l'aide de la discrétisation de la loi *a priori* de la portée, il suffit alors d'estimer espérance et variance de prédiction pour obtenir assez d'informations pour caractériser la loi prédictive. Avec cette approche le temps de calcul est considérablement réduit, au prix d'une approximation supplémentaire.

La méthode proposée fait donc plusieurs approximations pour faciliter les calculs de la loi prédictive. La première est réalisée lors de la discrétisation de la loi *a priori* $\pi(\phi)$, puis lorsque la somme de loi de Student est approchée par une loi gaussienne. Les moments obtenus analytiquement sont eux-mêmes des valeurs approchées des vrais moments du fait de la discrétisation de la loi *a priori*. L'importance de ces différentes approximations dans les estimations peut être difficile à quantifier. Dans la section 2.10, nous comparerons cette approche avec un algorithme MCMC afin de de tester l'efficacité de ces approximations.

2.8 Algorithme MCMC pour le krigeage bayésien

2.8.1 Algorithmes MCMC classiques

Les algorithmes MCMC sont des algorithmes permettant de simuler des distributions. Ils font appel à des échantillons corrélés qui sont par construction issus de la distribution cible. Ils sont particulièrement utilisés en statistique bayésienne puisqu'ils permettent d'échantillonner selon la loi *a posteriori*, dont la constante de normalisation (voir section 1.3.1) est souvent inconnue.

L'algorithme Metropolis-Hastings initialement présenté dans [Metropolis et al., 1953], puis généralisé dans [Hastings, 1970], est formulé de la manière suivante :

Algorithme de Metropolis-Hastings :
Soit θ une variable aléatoire dont on souhaite un échantillon de taille $M - m$ et possédant comme distribution $f_\theta(\theta)$. On note $\theta^{(j)}$ le j^{e} échantillon.

- **Initialisation** : Choix arbitraire de $\theta^{(0)}$ dans un un espace de valeurs raisonnables et choix d'une distribution de transition $g_\theta(.)$.
- **A l'itération** $j \in [\![1, M]\!]$:
  - Échantillonnage de θ_{nouv} depuis $g_\theta(\theta_{nouv}|\theta^{(j)})$

— Calcul de la probabilité d'acceptation $\alpha = \min\left(1, \frac{f_\theta(\theta_{nouv}) g_\theta(\theta|\theta_{nouv})}{f_\theta(\theta^{(j)}) g_\theta(\theta_{nouv}|\theta)}\right)$

— Attribution de $\theta^{(j+1)}$:

$$\theta^{(j+1)} = \begin{cases} \theta_{nouv} & \text{avec une probabilité } \alpha \\ \theta^{(j)} & \text{avec une probabilité } 1-\alpha \end{cases}$$

— **Approximation de la loi** : les m premières valeurs sont supprimées, puis une estimation de la densité $f_\theta(\theta)$ est obtenue à partir des $M - m$ échantillons.

M correspond au nombre d'itérations à réaliser (la taille de la chaîne), tandis que m correspond au « burn-in ». Ce second paramètre vient supprimer les premières itérations, dont les échantillons sont éventuellement éloignés de la distribution cible (le point de départ de la chaîne étant arbitraire). Le choix des valeurs M et m dans le cas du krigeage bayésien sera discuté dans la section 2.8.2.2.

Une des difficultés d'application de l'algorithme est le choix de la distribution de transition $g_\theta(.)$ dont va dépendre l'évolution de la chaîne. Le choix usuel pour cette distribution est la distribution gaussienne, mais d'autres choix comme la loi uniforme sont possibles (voir [Gelman et al., 1996]). La distribution gaussienne est centrée sur la valeur $\theta^{(j)}$ et de variance à spécifier. Ce choix de variance va également venir influencer le taux d'acceptation et modifier l'exploration de l'espace des paramètres. Le choix de spécification de cette variance est discuté dans [Gelman et al., 1996]. Ces particularités demandent donc de réaliser plusieurs tests sur la chaîne afin d'identifier correctement la variance de la distribution de transition. Des graphes d'évolution des $\theta^{(j)}$ peuvent être réalisés afin d'identifier la convergence de la chaîne et d'adapter les valeurs de M et m. Des critères de convergence peuvent être trouvés dans [Robert and Casella, 2004].

L'échantillonneur de Gibbs est un algorithme MCMC utilisé lorsqu'il est plus simple d'échantillonner selon les marginales plutôt que directement selon la loi jointe.

Echantillonneur de Gibbs :
Soit $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ un vecteur dont on souhaite un échantillon de taille $M - m$ et de distribution $f_{(\theta_1,\ldots,\theta_n)}(\theta_1, \ldots, \theta_n)$. On note $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, \ldots, \theta_n^{(j)})$ le i^{e} échantillon.

— **Initialisation** : Choix arbitraire de $\boldsymbol{\theta}^{(0)}$ selon des valeurs raisonnables.

— **A l'itération** $j \in [\![1, M]\!]$:
Pour tout $k \in [\![1, n]\!]$, on tire aléatoirement une valeur $\theta_k^{(j+1)}$ depuis :

$$f_\theta(\theta_k|\theta_1^{(j+1)}, \theta_2^{(j+1)}, \ldots, \theta_{k-1}^{(j+1)}, \theta_{k+1}^{(j)}, \ldots, \theta_n^{(j)}).$$

— **Approximation de la loi :** les m premières valeurs sont supprimées, puis une estimation de la loi $f_{(\theta_1,\ldots,\theta_n)}(\theta_1, \ldots, \theta_n)$ est obtenue à partir des $M - m$ échantillons.

L'échantillonneur de Gibbs peut être vu comme un cas spécifique de l'algorithme de Metropolis-Hastings dont la probabilité d'acceptation est fixée à 1. Son application est également bien plus simple puisqu'aucun paramètre (mis à part la taille de la chaîne M et le « burn-in ») n'a besoin d'être spécifié. Plus de détails sur ces deux algorithmes ainsi que

des preuves théoriques de leur convergence peuvent être trouvés dans [Robert and Casella, 2004].

2.8.2 Application au krigeage bayésien

Les algorithmes présentés dans la section précédente 2.8.1 peuvent être appliqués au krigeage bayésien. En effet l'ensemble des marginales sont connues et sont pour la plupart faciles à échantillonner. Un échantillonneur de Gibbs semble donc être adapté pour la résolution des équations du krigeage bayésien. La difficulté provient de l'échantillonnage de la loi *a posteriori* de la portée. Cette loi étant connue à une constante près, l'algorithme de Metropolis-Hastings permet de résoudre cette difficulté. Les algorithmes MCMC pour le krigeage bayésien font donc appel à une combinaison de l'échantillonneur de Gibbs pour les paramètres de moyenne et de variance (dont les marginales sont connues), puis à une étape Metropolis-Hastings pour la portée et l'effet de pépite (dont la distribution jointe est plus complexe à échantillonner).

2.8.2.1 Principe général

Le champ aléatoire gaussien est considéré stationnaire d'ordre 2 et peut se mettre sous la forme suivante :

$$Z(\boldsymbol{x}) = \mu + \omega(\boldsymbol{x}) + \kappa(\boldsymbol{x})$$

avec μ une moyenne déterministe, $\omega(.)$ un champ aléatoire centré de variance σ^2 et de portée ϕ, où les variables aléatoires $\kappa(\boldsymbol{x})$, $\boldsymbol{x} \in D$, sont i.i.d. de loi $\mathcal{N}(0, \tau^2)$. De plus, $\omega(.)$ et $\kappa(.)$ sont indépendants. On note également $\boldsymbol{z} = (z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_n))'$ les n observations. On rappelle également que la matrice de covariance du champ vérifie $\boldsymbol{R} = \sigma^2\boldsymbol{\Sigma}(\phi, \tau^2) = \sigma^2(\boldsymbol{V}(\phi) + \eta^2\boldsymbol{I}_n)$, avec $\eta^2 = \frac{\tau^2}{\sigma^2}$.

Les lois *a priori* sur les différents paramètres sont choisies de la manière suivante :

$$\mu \sim \mathcal{N}(m_0, R_0), \tau^2 \sim \mathcal{IG}(a_\tau, S_\tau^2), \sigma^2 \sim \mathcal{IG}(a_\sigma, S_\sigma^2) \text{ et } \phi \sim \mathcal{IG}(a_\phi, S_\phi^2).$$

Une première version adaptée de [De Oliveira, 2005] est formulée dans l'algorithme suivant.

Krigeage bayésien avec algorithme MCMC :

- **Initialisation** : Choix aléatoire de $\mu^{(0)}, \tau^{2(0)}, \sigma^{2(0)}, \phi^{(0)}$.
- **A l'itération** j : Ces étapes sont répétées M fois :
  - Échantillonnage de $\mu^{(j+1)}$ depuis $\mathcal{N}_p(m_1, R_1)$ avec :

$$\begin{cases} m_1 = \left(R_0^{-1} + \frac{1}{\sigma^{2(j)}}\mathbf{1}'\boldsymbol{\Sigma}^{-1}(\phi^{(j)})\mathbf{1}\right)^{-1}\left(R_0^{-1}m_0 + \frac{1}{\sigma^{2(j)}}\mathbf{1}'\boldsymbol{\Sigma}^{-1}(\phi^{(j)})\boldsymbol{z}\right), \\ R_1 = \left(R_0^{-1} + \frac{1}{\sigma^{2(j)}}\mathbf{1}'\boldsymbol{\Sigma}^{-1}(\phi^{(j)})\mathbf{1}\right)^{-1}, \end{cases}$$

  - Échantillonnage de $\sigma^{2(j+1)}$ depuis :

$$\mathcal{IG}\left(\frac{n}{2} + a_\sigma, \frac{1}{2}\left((\boldsymbol{z} - \mathbf{1}\mu^{(j+1)})'\boldsymbol{\Sigma}^{-1}(\phi^{(j)})(\boldsymbol{z} - \mathbf{1}\mu^{(j+1)}) + S_\sigma^2\right)\right),$$

— Échantillonnage de $\phi^{(j+1)}$ et $\tau^{(j+1)}$ à l'aide d'un algorithme Metropolis-Hastings avec la distribution :

$$f_{\phi,\tau^2}(\phi,\tau^2|\boldsymbol{z},\mu^{(j+1)},\sigma^{2(j+1)}) \propto \pi(\phi)\pi(\tau^2)\exp\left(-K\right),$$

où $K = \frac{1}{\sigma^{2(j+1)}}(\boldsymbol{z}-\mathbf{1}\mu^{(j+1)})'\boldsymbol{\Sigma}^{-1}(\phi)(\boldsymbol{z}-\mathbf{1}\mu^{(j+1)})$,

— Prédiction au point d'intérêt par krigeage simple avec les paramètres $\mu^{(j+1)}$, $\tau^{2(j+1)}$, $\sigma^{2(j+1)}$, et $\phi^{(j+1)}$. Tirage aléatoire depuis la loi obtenue par prédiction de $z(\boldsymbol{x}_0)^{(j+1)}$.

— **Approximation de la loi** : m premières valeurs sont supprimées, puis une estimation de la loi prédictive est obtenue à partir des $M-m$ échantillons.

L'étape Metropolis-Hastings de l'algorithme vient comparer de nouvelles valeurs ϕ_{nouv} et τ^2_{nouv} tirées aléatoirement avec celles obtenues à l'étape précédente. Nous explicitons ici cette étape :

A l'étape j : Échantillonnage de :

$$\begin{cases} \phi_{nouv} \sim \mathcal{N}\left(\phi^{2(j)}, s^2_{\phi}\right) = q_{\phi}(\phi_{nouv}|\phi^{2(j)}), \\ \tau^2_{nouv} \sim \mathcal{N}\left(\tau^{2(j)}, s^2_{\tau^2}\right) = q_{\tau^2}(\tau^2_{nouv}|\tau^{2(j)}). \end{cases}$$

On pose ensuite la probabilité d'acceptation du nouvel échantillon α défini par :

$$\alpha(\phi_{nouv},\tau^2_{nouv},\phi^{(j)},\tau^{2(j)}) = \min\left(1, \frac{f_{\phi,\tau^2}(\phi_{nouv},\tau^2_{nouv}|\boldsymbol{z},\mu^{(j+1)},\sigma^{2(j+1)})q_{\phi}(\phi^{(j)}|\phi_{nouv})q_{\tau^2}(\tau^{2(j)}|\tau^2_{nouv})}{f_{\phi,\tau^2}(\phi^{(j)},\tau^{2(j)}|\boldsymbol{z},\mu^{(j+1)},\sigma^{2(j+1)})q_{\phi}(\phi_{nouv}|\phi^{(j)})q_{\tau^2}(\tau^2_{nouv}|\tau^{2(j)})}\right).$$

Le nouvel échantillon est ensuite choisi aléatoirement selon la probabilité α :

$$(\phi^{(j+1)},\tau^{2(j+1)}) = \begin{cases} (\phi_{nouv},\tau^2_{nouv}) & \text{avec une probabilité } \alpha, \\ (\phi^{(j)},\tau^{2(j)}) & \text{avec une probabilité } 1-\alpha. \end{cases}$$

L'algorithme présenté ici combine donc un échantillonneur de Gibbs lorsque les marginales sont connues et simples à échantillonner et un algorithme de Metropolis-Hastings lorsqu'elles deviennent plus complexes à échantillonner (ici dans le cas de la portée et de l'effet de pépite). Si le paramètre de portée est multidimensionnel (dans le cas anisotrope), il est alors possible de venir échantillonner chaque composante du vecteur $\boldsymbol{\phi}$ (voir [De Oliveira, 2005] pour un cas en deux dimensions). Une approche similaire est réalisée par [De Oliveira and Ecker, 2002] en prenant en compte des données transformées, avec une approximation MC plutôt qu'un algorithme MCMC pour échantillonner la loi *a posteriori*. Des articles plus récents comme [Tadayon, 2017] considèrent les incertitudes de mesure comme connues. Ce choix de modélisation est réaliste puisque les résultats de mesure sont normalement communiqués avec leur incertitude associée. Cependant nous nous plaçons dans un cas où ces incertitudes ne sont pas nécessairement fournies.

Une seconde version de cet algorithme est proposée par [Fridley, 2003] avec des données censurées. La présence des données censurées ajoute une étape supplémentaire de prédiction de ces donnée. L'algorithme donné ici est modifié pour un modèle classique sans données censurées. Le modèle du champ reste identique. Sous les hypothèses de stationnarité d'ordre deux, on a alors :

$$\begin{cases} \boldsymbol{Z} \sim \mathcal{N}_n(\mu\mathbf{1}, \boldsymbol{R}(\sigma^2,\phi,\tau^2)), \\ \boldsymbol{Z}|\boldsymbol{\omega} \sim \mathcal{N}_n(\mu\mathbf{1}+\boldsymbol{\omega}, \tau^2\boldsymbol{I}). \end{cases}$$

où $\boldsymbol{\omega} = (\omega(\boldsymbol{x}_1), \ldots, \omega(\boldsymbol{x}_n))'$. Ce modèle considère un champ latent $\boldsymbol{\omega}(.)$ pour réaliser son algorithme MCMC. Le terme de champ latent est issu du vocabulaire des algorithmes d'augmentation de données et vient ajouter une variable inconnue (ici le champ $\boldsymbol{\omega}(.)$). Cette modification permet d'échantillonner l'effet de pépite directement à l'aide de l'échantillonneur de Gibbs sans passer par une étape Metropolis-Hastings.

Krigeage bayésien avec champ latent et algorithme MCMC :

- **Initialisation** : Choix aléatoire de $\mu^{(0)}, \tau^{2(0)}, \sigma^{2(0)}, \phi^{(0)}$ et $\boldsymbol{\omega}^{(0)}$,
- **A l'itération** $j \in [\![1, M]\!]$:
  - Échantillonnage de $\mu^{(j+1)}$ depuis $\mathcal{N}_p(m_1, R_1)$ avec :

$$m_1 = \frac{R_0 \tau^{2(j)}}{R_0 + \tau^{2(j)}} \left(\frac{m_0}{R_0} + \frac{1}{\tau^{2(j)}} (\overline{z} - \overline{\omega}^{(j)}) \right) \text{ et } R_1 = \frac{1}{n} \frac{R_0 \tau^{2(j)}}{R_0 + \tau^{2(j)}},$$

    où $\overline{z} = \frac{1}{n} \sum_{i=1}^{n} z(\boldsymbol{x}_i)$ et $\overline{\omega}^{(j)} = \frac{1}{n} \sum_{i=1}^{n} \omega^{(j)}(\boldsymbol{x}_i)$,
  - Échantillonnage de $\tau^{2(j+1)}$ depuis :

$$\mathcal{IG}\left(\frac{n}{2} + a_\tau, \frac{1}{2} \left(\left(\boldsymbol{z} - \left(\mu^{(j+1)} \mathbf{1} + \boldsymbol{\omega}^{(j)} \right) \right)' \left(\boldsymbol{z} - \left(\mu^{(j+1)} \mathbf{1} + \boldsymbol{\omega}^{(j)} \right) \right) + S_\tau^2 \right) \right),$$

  - Échantillonnage de $\sigma^{2(j+1)}$ depuis :

$$\mathcal{IG}\left(\frac{n}{2} + a_\sigma, \frac{1}{2} \left(\boldsymbol{\omega}^{(j)'} \boldsymbol{V}^{-1}(\phi^{(j)}) \boldsymbol{\omega}^{(j)} + S_\sigma^2 \right) \right),$$

  - Échantillonnage de $\boldsymbol{\omega}^{(j+1)}$ depuis $\mathcal{N}_n(\boldsymbol{m}_W, \boldsymbol{R}_W)$ avec :

$$\begin{cases} \boldsymbol{m}_W = \left(\frac{1}{\sigma^{2(j+1)}} \boldsymbol{V}^{-1}(\phi^{(j)}) + \frac{1}{\tau^{2(j+1)}} \boldsymbol{I} \right)^{-1} \left(\frac{1}{\tau^{2(j+1)}} (\boldsymbol{Z} - \mu^{(j+1)}) \right), \\ \boldsymbol{R}_W = \left(\frac{1}{\sigma^{2(j+1)}} \boldsymbol{V}^{-1}(\phi^{(j)}) + \frac{1}{\tau^{2(j+1)}} \boldsymbol{I} \right)^{-1}, \end{cases}$$

  - Échantillonnage de ϕ^{j+1} à l'aide d'un algorithme Metropolis-Hastings avec la distribution :

$$f_\phi(\phi | \mu^{(j+1)}, \tau^{2(j+1)}, \sigma^{2(j+1)}, \boldsymbol{\omega}^{(j+1)}, \boldsymbol{Z}) \propto \frac{\phi^{a_\phi - 1}}{\det(\boldsymbol{V}(\phi))^{\frac{1}{2}}} \exp(-K),$$

    où $K = \frac{1}{\sigma^{2(j+1)}} \boldsymbol{\omega}^{(j+1)'} \boldsymbol{V}^{-1}(\phi) \boldsymbol{\omega}^{(j+1)} - S_\phi^2 \phi$,
  - Prédiction au point d'intérêt par krigeage simple avec les paramètres $\mu^{(j+1)}$, $\tau^{2(j+1)}$, $\sigma^{2(j+1)}$, et $\phi^{(j+1)}$. Tirage aléatoire depuis la loi obtenue par prédiction de $z(\boldsymbol{x}_0)^{(j+1)}$.
- **Approximation de la loi** : m premières valeurs sont supprimées, puis une estimation de la loi prédictive est obtenue à partir des $M - m$ échantillons.

Cette version de l'augmentation de données permet de simplifier l'étape de Metropolis-Hastings en évitant d'échantillonner selon deux paramètres simultanément. Cependant elle ajoute un étape de prédiction du vecteur $\boldsymbol{\omega}$ et ajoute donc une étape dans l'algorithme MCMC. C'est pourquoi dans la suite nous considérerons la première version sans champ

latent.

2.8.2.2 Choix des paramètres pour la convergence de l'algorithme

Hormis le choix de lois *a priori* sur les différents paramètres, plusieurs paramètres propres à l'algorithme doivent être spécifiés. En premier lieu le nombre d'échantillons M de la chaîne ainsi que le « burn-in » m doivent être précisés. Dans la littérature, [Fridley and Dixon, 2007] et [De Oliveira, 2005] prennent $M = 10000$ et $m = 500$, tandis que [Tadayon, 2017] utilise $M = 150000$ et $m = 100000$ en ne conservant qu'un échantillon sur 100. Plusieurs choix sont donc possibles, sans que des valeurs recommandées existent. Afin de choisir ces paramètres, il est possible de tracer des graphes représentant l'évolution des échantillons en fonction de l'itération. L'identification de la convergence dans l'évolution de la chaîne permet alors de déterminer M et m. Des outils pour l'identification de cette convergence sont fournis dans [Robert and Casella, 2004]. La Figure 2.1 illustre l'évolution des différents paramètres pour un jeu de données simulé selon le modèle présenté dans la section 2.11.1.1. Dans cet exemple $M = 10000$ itérations est suffisant, tandis que $m = 3000$ semble être ici approprié.

Le second choix d'importance est la famille de lois utilisée lors de l'étape Metropolis-Hastings et l'échantillonnage des paramètres ϕ et τ^2. [Gelman et al., 1996] considèrent plusieurs choix possibles, comme la famille gaussienne ou uniforme. [Fridley and Dixon, 2007] utilise une loi gamma, tandis que [De Oliveira, 2005] utilise une loi gaussienne. Dans la suite nous utiliserons la distribution de transition gaussienne.

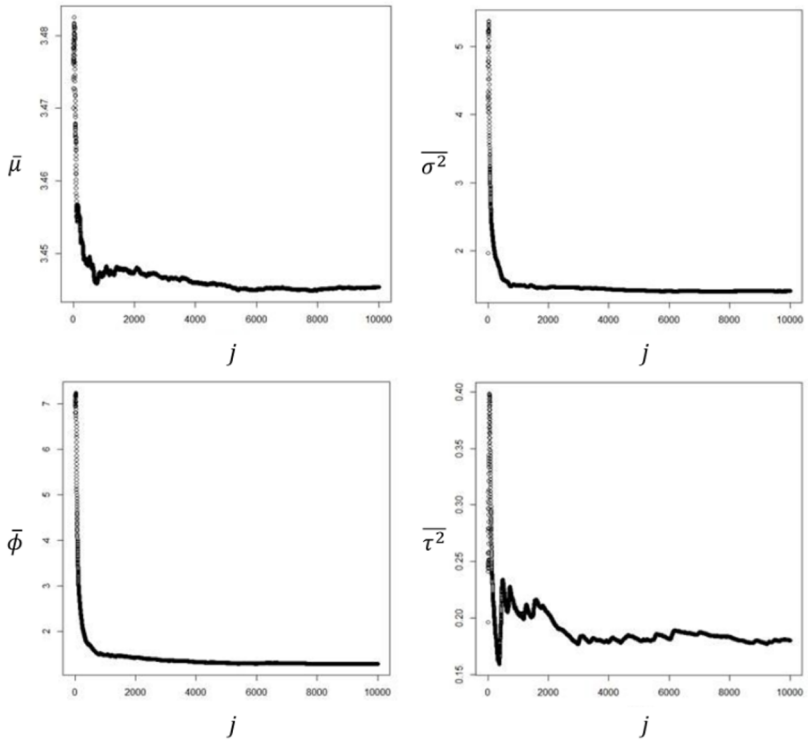


FIGURE 2.1 – Évolution de la valeur moyenne de la moyenne μ, de la variance σ^2, de la portée et de l'effet de pépite τ^2 en fonction de l'itération i.

2.9 Choix des lois *a priori* et leur paramétrisation

2.9.1 Choix classiques

Le choix des lois *a priori* et leur paramétrisation est une étape difficile de l'application des méthodes bayésiennes. Le choix de ces lois influence les inférences et prédictions réalisées et ne dépendent pas des données. Ce choix est alors arbitraire, pouvant occasionner des erreurs dans la spécification, entraînant de possibles erreurs dans la modélisation statistique.

Une option classique est le choix de lois *a priori* impropres (souvent appelées non-informatives). En effet ces lois ne nécessitent pas de spécifications de paramètres supplémentaires et n'apportent que peu d'informations *a priori* sur les paramètres, limitant ainsi les risques d'erreur. Cette approche est notamment défendue par [Helbert et al., 2009], où le gain d'information apporté par des lois informatives est faible devant le risque d'erreur de spécification. Pour notre exemple d'application, la densité de la loi jointe impropre classique est la suivante :

$$\pi(\mu, \sigma^2, \phi, \tau^2) \propto \frac{\pi(\phi)\pi(\tau^2)}{\sigma^2}.$$

Dans le cas de l'effet de pépite τ^2, la loi impropre $\pi(\tau^2) \propto \frac{1}{\tau^2}$ peut être considérée, ce paramètre correspondant comme σ^2 à une variance.

La situation du paramètre de portée ϕ est cependant plus complexe. En effet il n'existe pas de loi *a priori* usuelle pour la portée. Par exemple, [Diggle and Ribeiro, 2002] propose plusieurs choix différents comme une loi uniforme ou une loi *a priori* de la forme $\pi(\phi) \propto \frac{1}{\phi}$. [Berger et al., 2001] discute des options classiques de lois *a priori* et propose une nouvelle loi *a priori* objective de référence pour la portée et dont le domaine d'application est étendu par les travaux de [Muré, 2018], notamment aux cas anisotropes. Cette loi *a priori* non-informative de référence est la suivante :

$$\pi(\phi) \propto \left(tr(\boldsymbol{W}_\phi) - \frac{1}{n-p}(tr(\boldsymbol{W}_\phi)^2) \right)^{\frac{1}{2}}, \tag{2.5}$$

où $\boldsymbol{W}_\phi = \frac{\partial}{\partial\phi}(\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\boldsymbol{J}(\phi)$, $\boldsymbol{J}(\phi) = \boldsymbol{I} - \boldsymbol{D}\left(\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}\boldsymbol{D}\right)^{-1}\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}$ et $\frac{\partial}{\partial\phi}(\boldsymbol{\Sigma})$ la matrice obtenue en dérivant la matrice $\boldsymbol{\Sigma}$ terme à terme selon ϕ et où $\boldsymbol{\Sigma} = \frac{1}{\sigma^2}\boldsymbol{R}$.

Cependant si des informations sont disponibles sur certains paramètres, il est possible d'utiliser des lois *a priori* propres (souvent appelées subjectives ou encore informatives). Dans le cas de la moyenne, de la variance et de l'effet de pépite, les lois informatives classiques sont respectivement la loi normale (ou normale multivariée) et la loi inverse-gamma pour la variance et l'effet de pépite :

$$\mu \sim \mathcal{N}(m_0, R_0), \sigma^2 \sim \mathcal{IG}\left(a_\sigma, S_\sigma^2\right), \phi \sim \mathcal{IG}\left(a_\phi, S_\phi^2\right), \tau^2 \sim \mathcal{IG}\left(a_\tau, S_\tau^2\right).$$

En pratique les différents paramètres de ces lois peuvent être estimés à partir des données. m_0 correspond à une première estimation de la moyenne, tandis que R_0 correspond à une variance faible (éventuellement une matrice diagonale dont les coefficients sont petits dans le cas d'un champ avec tendance). Dans la littérature, [De Oliveira, 2005] et [Toscas, 2010] utilisent avec des données censurées $R_0 = 0.00005$. S_σ^2 correspond souvent à une estimation

de la variance. Pour la portée et l'effet de pépite, leur loi inverse-gamma peut être centrée sur une première estimation des paramètres ϕ et τ^2 en utilisant par exemple une analyse variographique ([De Oliveira, 2005]). Quant au choix de a_σ, a_ϕ et a_τ, ils sont souvent fixés de manière à ce que la variance de la loi inverse-gamma ainsi paramétrée soit centrée sur une première estimation du paramètre en question. Cela peut aussi être modifié pour obtenir une loi de variance infinie en spécifiant par exemple $a_\sigma = 2$ ([Ecker and Gelfand, 1997, De Oliveira, 2005]).

Dans la suite nous considérerons comme loi *a priori* pour la portée la loi inverse-gamma ainsi que la loi uniforme. [Berger et al., 2001] met en garde contre l'utilisation de la loi uniforme comme loi objective, pouvant engendrer des lois *a posteriori* impropres. Cependant la discrétisation de la loi *a priori* de la portée utilisée dans l'approche MC permet d'éviter cet écueil. De plus dans nos algorithmes MCMC nous ne ferons appel qu'à des lois *a priori* propres afin d'éviter de générer des lois *a posteriori* impropres.

2.9.2 Brève analyse de sensibilité

L'analyse de la sensibilité des prédictions et des estimations du modèle est une étape primordiale des statistiques bayésiennes. Elle permet de s'assurer que le choix de lois *a priori* est raisonnable et ne vient pas biaiser les estimations et prédictions. Ici une brève analyse de sensibilité est menée afin de vérifier le choix de lois *a priori* impropres donné dans la section 2.9.1.

Un jeu initial dense de $129 * 129 = 16641$ données est simulé à partir d'une fonction de covariance exponentielle selon un modèle de champ aléatoire gaussien, stationnaire d'ordre deux sur une grille régulière dans l'espace $[0, 10]^2$. Le calcul des prédictions est réalisé à l'aide du package `geoR`.

Un premier tirage aléatoire de 2000 observations est réalisé. A partir de ces 2000 observations sont estimés par maximum de vraisemblance les paramètres de moyenne, de variance et de portée :

$$\mu^*_{init} = 0.52, \sigma^{2*}_{init} = 0.07, \phi^*_{init} = 3.33.$$

Ces valeurs servent de référence comme spécifications « correctes » pour les lois *a priori*.

L'influence des lois *a priori* sur les prédictions et estimations dépend du nombre d'observations (peu d'observations implique une plus grande influence des lois *a priori*). Deux tailles de jeux de données seront donc considérées dans la suite : $n = 20$ et $n = 50$. Pour chaque taille d'échantillon, 100 tirages aléatoires sans remise sont réalisés sur le jeu de données initial.

On souhaite considérer 5 situations. Dans la première, la loi *a priori* est impropre, tandis que la seconde est spécifiée par des lois centrées sur les valeurs estimées précédemment avec variances faibles. Ensuite on considère les mêmes lois mais de variances plus grandes. Finalement, les 3 situations précédentes sont considérées cette fois-ci avec des lois mal centrées. On obtient donc les 5 lois *a priori* suivantes :

— impropre :

$$\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2} \text{ et } \phi \sim \mathcal{U}([0, d_{max}]),$$

— fidèle et juste :

$$\sigma^2 \sim \mathcal{IG}\left(\frac{n}{2}, \frac{n\sigma^{2*}_{init}}{2}\right), \ \mu|\sigma^2 \sim \mathcal{N}\left(\mu^*_{init}, \frac{\sigma^{2*}_{init}}{n}\right) \text{ et } \phi \sim \mathcal{U}([2.33, 4.33]),$$

— fidèle et peu juste :

$$\sigma^2 \sim \mathcal{IG}\left(\frac{n}{2}, \frac{3n\sigma^{2*}_{init}}{2}\right), \mu|\sigma^2 \sim \mathcal{N}\left(3\mu^*_{init}, \frac{\sigma^{2*}_{init}}{n}\right) \text{ et } \phi \sim \mathcal{U}([3, 5]),$$

— peu fidèle et juste :

$$\sigma^2 \sim \mathcal{IG}\left(\frac{n}{6}, \frac{n\sigma^{2*}_{init}}{6}\right), \mu|\sigma^2 \sim \mathcal{N}\left(\mu_{init}, \frac{\sigma^{2*}_{init}}{n}\right) \text{ et } \phi \sim \mathcal{U}([1.33, 5.33]),$$

— peu fidèle et peu juste :

$$\sigma^2 \sim \mathcal{IG}\left(\frac{n}{6}, \frac{n\sigma^{2*}_{init}}{2}\right), \mu|\sigma^2 \sim \mathcal{N}\left(3\mu^*_{init}, \frac{\sigma^{2*}_{init}}{n}\right) \text{ et } \phi \sim \mathcal{U}([2, 6]).$$

Pour le choix des bornes pour le paramètre de portée, elles sont choisies de telle sorte à inclure ϕ^*_{init} et d'être centrées (ou non) sur ϕ^*_{init}. Une validation croisée est ensuite appliquée sur chacun des jeux de données avec l'ensemble des lois *a priori* présentées et les critères de validation présentés dans la partie 1.9.6 sont estimés. Les résultats sont représentés sous la forme d'un box-plot donné par la Figure 2.2.

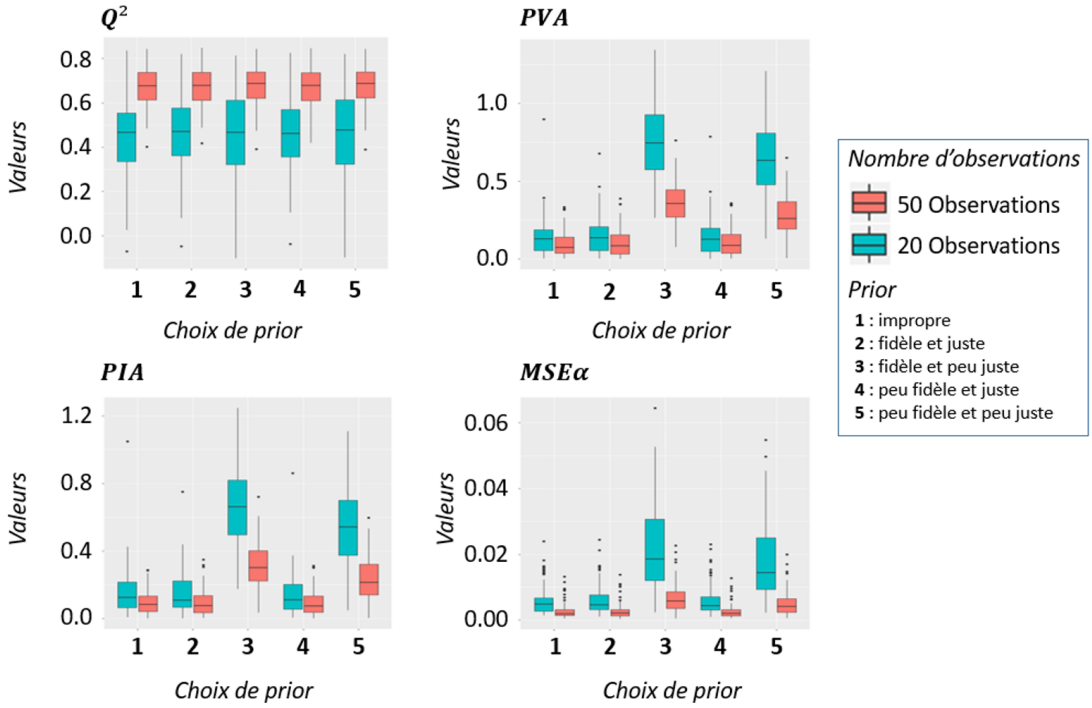


FIGURE 2.2 – Distribution des critères de validation (Q^2, PVA, PIA, et $MSE\alpha$) selon le choix de loi *a priori*.

On remarque que le critère Q^2 est insensible au choix de loi *a priori*. C'est un résultat attendu puisque les prédictions sont avant tout sensibles au nombre de données. Par contre les critères PVA et PIA (ainsi que le critère $MSE\alpha$ dans une moindre mesure) sont très sensibles à ce choix, puisque la variance de prédiction dépend fortement de l'estimation des paramètres de variance et de portée. Une loi *a priori* impropre permet de mitiger le biais introduit lors d'un choix de loi incorrectement centrée, puisque le cas 3 donne en moyenne de pires résultats que le cas 5. Enfin on remarque que même si la loi *a priori* est correctement spécifiée (juste et fidèle, cas 2), le gain obtenu pour nos différents critères reste relativement faible, notamment comparé au cas impropre (cas 1).

En conclusion, le choix d'une loi *a priori* impropre est généralement raisonnable. Il est possible d'améliorer les performances du krigeage bayésien en spécifiant une loi *a priori* juste et fidèle. Cependant, cela fait courir un risque de mauvaise spécification de la loi *a priori*, entraînant de mauvaises variances de prédiction. Ces résultats et conclusions sont en accord avec [Helbert et al., 2009].

2.10 Comparaison algorithmes MCMC et Monte-Carlo

Afin de valider notre algorithme MCMC et de comparer la méthode proposée par [Diggle and Ribeiro, 2002] fournie dans le package `geoR` du langage `R`, nous appliquons les méthodes sur un jeu de données simulé et réalisons une validation croisée. Un krigeage ordinaire est réalisé également pour référence, ses paramètres étant estimés par maximum de vraisemblance.

Le jeu de données est simulé selon un processus gaussien de fonction de covariance exponentielle (voir la section 1.9.3.1), et dont les paramètres sont :

$$\mu = 3.5, \sigma^2 = 1, \phi = 3.5, \tau^2 = 0.$$

Cette simulation est faite pour 100 observations sur une grille régulière dans le domaine $D = [0, 10]^2$.

En ce qui concerne le choix de lois *a priori*, les deux krigeages sont paramétrés de manière similaire pour la comparaison :

$$\mu \sim \mathcal{N}_p\left(\widehat{\mu}, 0.00005\right), \sigma^2 \sim \mathcal{IG}\left(1.5, \frac{\widehat{\sigma^2}}{2}\right) \text{ et } \phi \sim \mathcal{IG}(3, 1),$$

avec $\widehat{\mu}$ une estimation de la moyenne et $\widehat{\sigma^2}$ une estimation de la variance.

L'approche Monte-Carlo utilise une version discrétisée en 200 points dans $]0, 100]$ de la loi $\mathcal{IG}(3, 1)$. En ce qui concerne le choix des paramètres de l'algorithme MCMC, des premiers essais ont permis de déterminer les valeurs de telle sorte que le taux d'acceptation de l'algorithme soit d'environ 0.35. De même, les paramètres de « burn-in » et d'itérations ont été fixés à $m = 3000$ et $M = 10000$ (après la réalisation de plusieurs essais afin de vérifier ces choix de paramètres).

Krigeage	Q^2	PVA	$MSE\alpha$
Bayésien MC	0.53	0.011	0.0008
Bayésien MCMC	0.53	0.019	0.0008
Ordinaire	0.54	0.038	0.0014

TABLE 2.1 – Critères de validation pour le krigeage bayésien par Monte-Carlo, algorithme MCMC et krigeage oridnaire.

Les résultats sont donnés dans le Tableau 2.1 et la Figure 2.3.

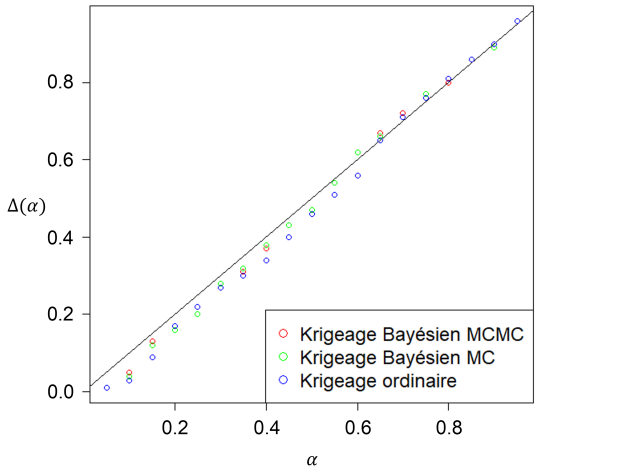


FIGURE 2.3 – Graphes α-CI du krigeage bayésien par MCMC, par Monte-Carlo (MC) et par krigeage ordinaire.

On remarque que les trois approches donnent des résultats extrêmement similaires. Le krigeage ordinaire fournit ici de moins bonnes performances, et les deux krigeages bayésiens fournissent des résultats très proches. Les deux méthodes apparaissent donc comme équivalentes.

Dans la suite nous privilégions la méthode Monte-Carlo dont les temps de calcul sont plus faibles (avec l'approximation présentée dans la section 2.7 évitant un échantillonnage direct). Néanmoins la méthode MCMC sera ensuite choisie dans le chapitre suivant puisqu'elle permet de traiter les données censurées.

2.11 Application à des simulations et comparaison avec le krigeage ordinaire

Dans le cadre de l'A&D de sites nucléaires, le krigeage bayésien pourrait être un outil utile pour la réalisation de cartographie radioactive notamment lorsque le nombre de données disponibles est faible (du fait d'un environnement contraignant). L'objectif de cette partie est donc de vérifier l'utilité du krigeage bayésien lorsque peu de données sont disponibles par rapport aux méthodes de krigeage classique utilisées dans le même contexte (comme le krigeage ordinaire utilisé dans [Desnoyers, 2010]). Pour réaliser cette comparaison, des jeux de données de tailles variables seront simulés, puis des données

issues d'une fonction déterministe seront utilisées. Enfin, un jeu de données réelles sera employé. Dans ces trois cas un nombre variable d'observations sera considéré et la qualité des krigeages sera jugée selon les critères de validation présentés dans la section 1.9.6.

Dans la suite, le krigeage bayésien utilisé fait appel au package `geoR`, impliquant l'utilisation de la méthode Monte-Carlo pour la résolution des équations du krigeage bayésien. Le problème de sensibilité au choix de loi *a priori* évoqué dans la section 2.7 n'est pas contraignant puisque la spécification de loi *a priori* ne change pas durant le protocole et que l'objectif ici n'est pas la construction d'un modèle satisfaisant mais la comparaison des deux krigeages.

2.11.1 Jeux de données simulés

Le jeu de données simulé est un cas idéal d'application des deux krigeages. En effet, toutes les hypothèses initiales (gaussienne et stationnarité) sont respectées par construction et permettent de comparer dans une situation optimale les performances de chaque krigeage.

2.11.1.1 Génération des jeux de données

Les données utilisées sont issues de simulations non-conditionnelles générées à l'aide de la fonction `grf` du package `geoR` du langage `R`. Ces jeux sont gaussiens, de fonction de covariance exponentielle et stationnaires d'ordre deux. Ainsi pour n observations, on obtient le vecteur gaussien :

$$\boldsymbol{Z} \sim \mathcal{N}_n(\mu\mathbf{1}, \boldsymbol{R}(\sigma^2, \phi)).$$

Les valeurs utilisées pour les différents paramètres sont :

$$\mu = 0.5, \sigma^2 = 0.1, \phi = 4.5.$$

2.11.1.2 Protocole

Les jeux de données sont simulés pour différentes tailles d'échantillons, de 16 à 81 observations dans l'espace $[0, 10]^2$. Ces valeurs sont choisies de manière à générer une maille carrée sur le domaine étudié. 100 simulations indépendantes sont réalisées pour chaque taille d'échantillon. Une validation croisée utilisant chacun des krigeages est ensuite appliquée à chaque jeu de données, et les dispersions des critères de validation sont ensuite représentées selon la taille de l'échantillon par un box-plot.

Les paramètres sont estimés par maximum de vraisemblance pour le krigeage ordinaire. Pour le krigeage bayésien, la densité de la loi *a priori* jointe pour la moyenne et la variance est la suivante :

$$\pi(\mu, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Pour la portée, une loi *a priori* uniforme impropre dans $\mathbb{R}^{+*}$ est choisie :

$$\pi(\phi) \propto 1.$$

La densité $\pi(\phi)$ de la portée est ensuite discrétisée en 50 valeurs différentes.

2.11.1.3 Résultats et interprétations

Les box-plots des résultats sont donnés dans la Figure 2.4 où les résultats obtenus par krigeage ordinaire sont en rouge et ceux obtenus par krigeage bayésien en bleu.

Le krigeage bayésien est plus performant sur l'ensemble des critères de validation pour les jeux de données de taille inférieure ou égale à 40 observations. Ce résultat est particulièrement visible pour le PVA et le PIA et explique que la différence principale entre les deux modèles dans les prédictions provient de la variance prédite, les Q^2 calculés étant relativement proches (à l'exception du jeu de données avec 16 observations). En prenant en compte les incertitudes dans les estimations des paramètres, le krigeage bayésien fournit des variances de prédiction plus élevées et des intervalles de prédiction plus justes et fidèles puisqu'il fournit de meilleurs PVA, PIA et $MSE\alpha$.

Si l'on regarde en détails ces résultats, la valeur médiane du critère Q^2 augmente de 0 à 0.63 selon la taille de l'échantillon pour le krigeage ordinaire, tandis que celle du krigeage bayésien passe de 0.16 à 0.63, fournissant de meilleures prédictions. La dispersion du Q^2 pour une taille de jeu de données fixe est proche pour chaque méthode de krigeage. Cela s'explique par le nombre d'observations qui reste fixe entre les deux méthodes. Les prédictions en moyenne par krigeage étant principalement sensibles à la taille de l'échantillon, il est normal d'obtenir des valeurs de Q^2 proches.

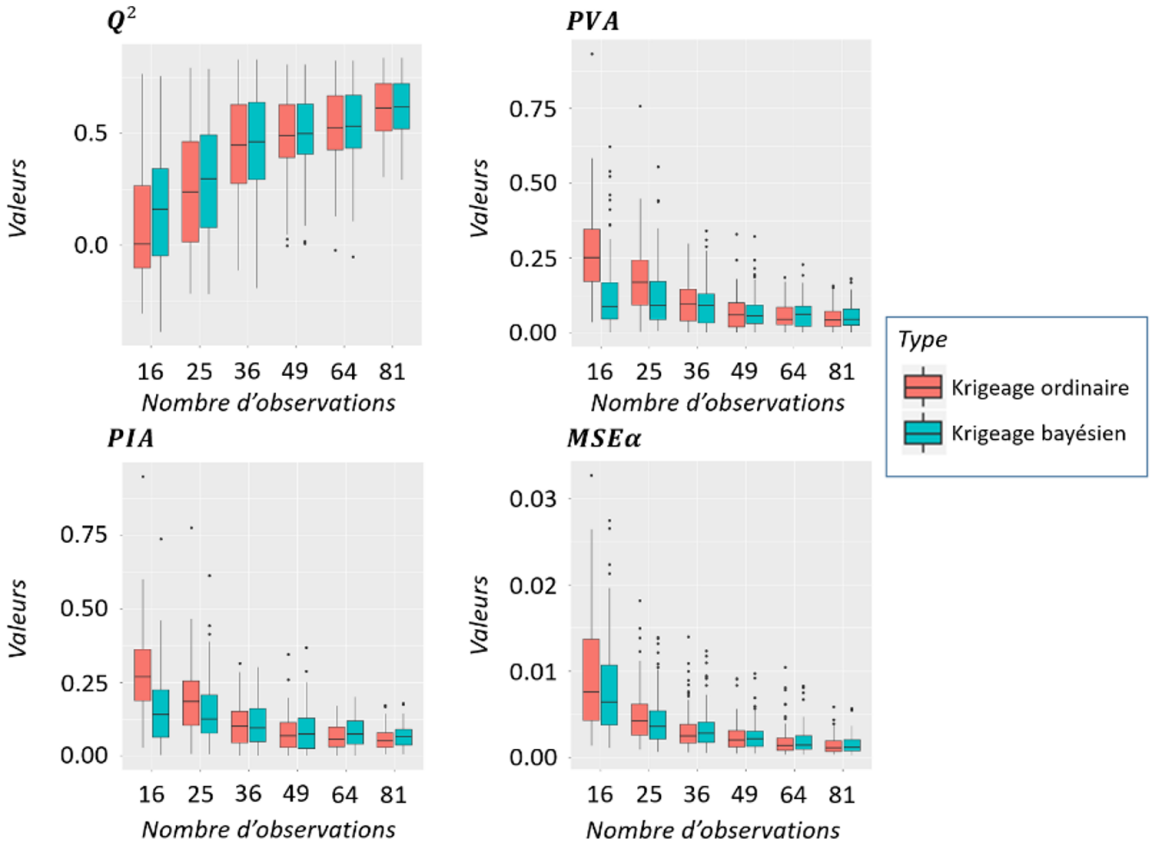


FIGURE 2.4 – Distribution des critères de validation (Q^2, PVA, PIA et $MSE\alpha$) contre la taille de l'échantillon pour des données issues d'un processus gaussien simulé.

Concernant la valeur médiane du PVA, elle évolue de 0.25 à 0.04 pour le krigeage ordinaire, et de 0.15 à 0.05 pour le krigeage bayésien. De plus, contrairement au Q^2, les dispersions des PVA et PIA sont différentes pour de petits jeux de données entre les deux krigeages. Le PVA et le PIA étant très sensibles à l'estimation de la variance et de la portée, lorsque l'estimation des paramètres est peu robuste (comme c'est le cas pour le maximum de vraisemblance lorsque peu d'observations sont disponibles), il est normal d'obtenir des critères très différents d'un jeu de données à un autre. Pour le krigeage bayésien, l'estimation des paramètres étant plus robuste pour de petits jeux de données (du fait de la loi *a priori*), cette dispersion est plus faible. Enfin on remarque que ces différences s'estompent lorsque la taille de l'échantillon est supérieure à 36 observations, et que le krigeage ordinaire semble même plus performant. Cependant la différence est trop faible pour être considérée comme significative.

Pour le $MSE\alpha$, on observe des résultats similaires aux graphes précédents. Cela s'explique par le fait que le $MSE\alpha$ est un critère jugeant les intervalles de confiance, et juge donc simultanément l'espérance et la variance de prédiction. Pour le krigeage ordinaire, la médiane du $MSE\alpha$ évolue de 0.0075 à 0.0025, tandis que pour le krigeage bayésien, cette médiane évolue de 0.006 à 0.0025. A nouveau le krigeage bayésien semble offrir de meilleures performances.

Les résultats obtenus avec des données simulées selon une fonction de Matérn$-\frac{3}{2}$ et Matérn$-\frac{5}{2}$ sont donnés dans l'annexe A.1. Les évolutions des différents critères sont très similaires à celles obtenues pour la fonction de Matérn$-\frac{1}{2}$. Cependant les différences entre krigeage bayésien et ordinaire semblent tout de même s'amoindrir lorsque le paramètre ψ augmente (autrement dit lorsque la régularité du champ simulé augmente).

2.11.2 Jeux de données issus d'une fonction déterministe

Les données seront maintenant créées à partir d'une fonction déterministe. Ce cas s'éloigne du cas idéal des simulations où la fonction de covariance et les paramètres du modèle ne sont plus connus. De plus, les hypothèses gaussienne et de stationnarité ne sont pas nécessairement respectées. Ce contexte est donc plus délicat à traiter pour les deux krigeages et peut engendrer de mauvaises prédictions.

2.11.2.1 Présentation de la fonction déterministe

La fonction déterministe employée est une fonction continue sur $[-1,1]^2$ de la forme suivante [Iooss et al., 2010] :

$$t(x,y) = \frac{e^x}{5} - \frac{y}{5} + \frac{y^6}{3} + 4y^4 - 4y^2 + \frac{7x^2}{10} + x^4 + \frac{3}{4x^2+4y^2+1}$$

La Figure 2.5 illustre cette fonction sur le domaine $[-1,1]^2$.

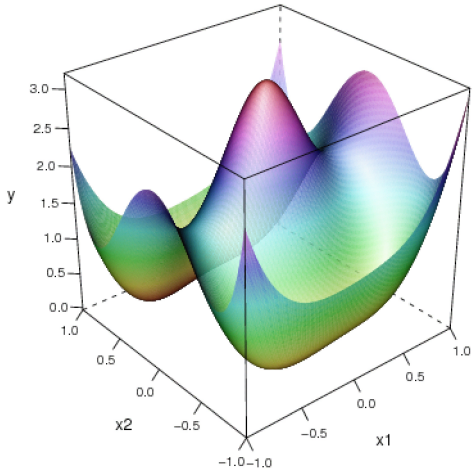


FIGURE 2.5 – Illustration de la fonction déterministe t.

Cette fonction étant déterministe, il n'est pas possible de générer différents jeux de données à partir d'une géométrie fixe. La géométrie carrée utilisée dans la section 2.11.1.2 n'est donc plus utilisée pour la comparaison mais uniquement pour le choix de fonction de covariance détaillé dans la section suivante 2.11.2.2. A la place de la géométrie carrée, on fait appel à un plan d'échantillonnage aléatoire, les positions des observations étant tirées uniformément sans remise dans l'espace $[-1, 1]^2$.

2.11.2.2 Protocole

Le protocole est ici séparé en deux étapes. La première étape consiste à identifier la fonction de covariance la plus adaptée à ce problème. Pour cela nous comparons les modèles exponentiel, Matérn-1.5, Matérn-2.5 et gaussien (pour le modèle gaussien un effet de pépite de 10^{-6} est ajouté pour des raisons de stabilité numérique), voir la section 1.9.3.1 pour les différentes expressions de ces fonctions. Un jeu de données comportant 144 observations selon une grille carrée est employé ici pour comparer les fonctions de covariance (ce nombre d'observations est choisi grand pour correctement identifier la fonction de covariance).

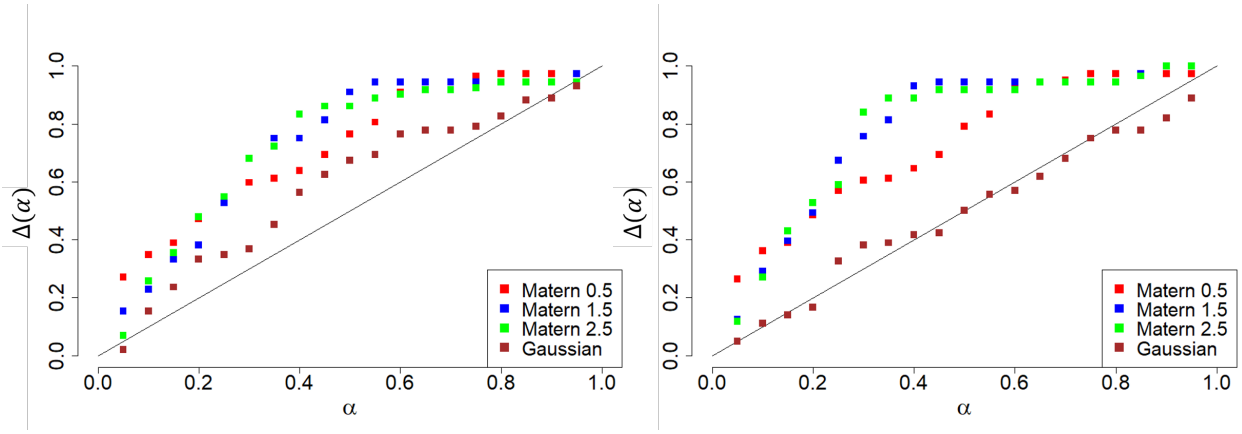
Une fois la fonction de covariance déterminée, la fonction déterministe est ensuite évaluée pour chaque position tirée aléatoirement, formant ainsi le jeu de données. Ce processus est répété 100 fois pour chaque taille d'échantillon jusqu'à 150 observations (nombre d'observations choisi de manière arbitraire, sans lien avec le choix du paragraphe précédent). A nouveau pour chacun de ces jeux de données une validation croisée est appliquée, les paramètres étant estimés par maximum de vraisemblance pour le krigeage ordinaire et les lois *a priori* restant identiques à celles données dans la section 2.11.1.2, c'est à dire des lois *a priori* impropres.

2.11.2.3 Résultats et interprétations

Pour comparer les différentes fonctions de covariance considérées, les critères de validation sont estimés et les graphes α-CI pour chacune des fonctions sont tracés. Ceci est répété pour les deux krigeages. Les résultats sont résumés dans la Figure 2.6, le Tableau 2.2 pour le krigeage ordinaire et le krigeage bayésien.

Covariance	Q^2_{ord}	PVA_{ord}	PIA_{ord}	$MSE\alpha_{ord}$	Q^2_{bay}	PVA_{bay}	PIA_{bay}	$MSE\alpha_{bay}$
Matérn 1/2	0.95	0.99	0.98	0.056	0.95	1.09	1.06	0.061
Matérn 3/2	0.99	0.91	0.90	0.073	0.99	1.62	1.60	0.106
Matérn 5/2	1.00	0.65	0.63	0.073	1.00	1.58	1.55	0.106
Gaussian	1.00	0.05	0.07	0.011	1.00	0.13	0.16	0.002

TABLE 2.2 – Critères de validation pour le krigeage ordinaire et bayésien selon différentes fonctions de covariance, sur l'échantillon de 144 observations de la fonction t.



(a) Krigeage ordinaire.

(b) Krigeage bayésien.

FIGURE 2.6 – Graphes α-CI pour les krigeages ordinaire et bayésien selon différentes fonctions de covariances, sur l'échantillon de 144 observations de la fonction t.

Les résultats montrent que la fonction de covariance gaussienne est la plus performante. Ce résultat est cohérent, puisque la fonction de covariance permet de modéliser des champs présentant de fortes corrélations entre observations. La fonction $t(.)$ étant très « lisse », la fonction gaussienne est efficace pour reproduire cette forte régularité. Si les valeurs prises par le critère Q^2 sont relativement proches entre fonctions de covariance, des différences importantes apparaissent pour le PVA et le PIA. Ces différences sont par ailleurs réduites pour le $MSE\alpha$. Ces comportements différents des critères montrent l'importance de les utiliser simultanément lors d'un choix de modèle.

Dans la suite la fonction de covariance gaussienne est utilisée pour appliquer les deux krigeages. Les résultats sont représentés sous la forme de box-plots dans la Figure 2.7.

Les valeurs obtenues pour le critère Q^2 nous amènent à des conclusions identiques à celles de la section 2.11.1.3. Le krigeage bayésien semble offrir de meilleures performances, en particulier pour un faible nombre d'observations. On remarque tout de même que les valeurs de Q^2 sont en moyenne plus élevées que dans l'application de la section 2.11.1.3 à cause de la régularité de la fonction $t(.)$.

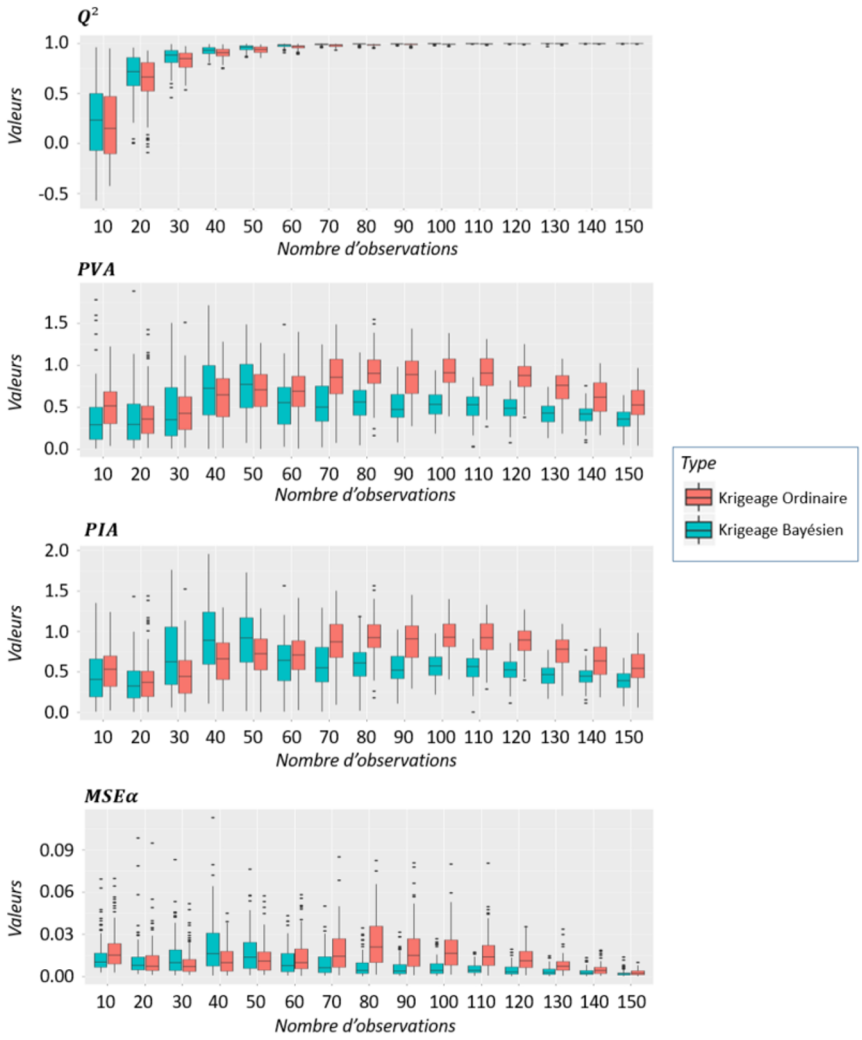


FIGURE 2.7 – Distribution des critères de validation (Q^2, PVA, PIA et $MSE\alpha$) contre la taille de l'échantillon pour la fonction déterministe t.

Le cas des critères PVA, PIA et $MSE\alpha$ est cependant bien plus complexe. En effet, ces critères ne décroissent pas strictement avec le nombre d'observations, et l'évolution des critères en fonction du nombre d'observations n'est pas le même pour les deux krigeages. Pour le krigeage bayésien, les critères croissent entre 20 et 50 observations avant de décroître, tandis que ces critères continuent de croître jusqu'à 110 observations environ pour le krigeage ordinaire avant de décroître. Dans certains cas particuliers (les jeux de données avec 30 et 50 observations), le krigeage bayésien se montre moins performant que le krigeage ordinaire, mais a de meilleures performances pour plus de 50 observations. Enfin pour plus de 80 observations, on observe une évolution similaire à celle décrite dans la section 2.11.1.3.

Pour expliquer ces résultats, il est important de rappeler ici l'hypothèse gaussienne qui n'est pas vérifiée ici, les données étant obtenues à partir d'une fonction déterministe (inférieur à 50 observations). Dans cette situation, il est possible d'obtenir des critères qui empirent lorsque le nombre d'observations augmente. Les moyennes des prédictions restent bonnes, comme l'indique les excellentes valeurs du Q^2 dont la médiane reste supérieure

à 0.7 pour la plupart des tailles d'échantillon considérées. Cependant les variances de prédiction semblent être mal estimées, donnant de mauvais intervalles de confiance et de crédibilité. Pour plus de 80 observations, le comportement obtenu est en accord avec l'hypothèse gaussienne.

En conclusion, dans cette application où l'hypothèse gaussienne n'est valide, le krigeage bayésien semble en moyenne plus robuste que le krigeage ordinaire. Cependant la prudence est recommandée, puisque dans certains exemples le krigeage ordinaire semble être plus performant que le krigeage bayésien, comme illustré avec les cas où le nombre d'observations est égal à 40 ou 50 observations.

2.12 Application à des données réelles : réacteur G3 et comparaison avec le krigeage ordinaire

Pour cette dernière application on se place dans un cas réel issu d'un démantèlement industriel du réacteur G3 à Marcoule. Ce réacteur plutonigène a été arrêté en 1984 et a fait l'objet d'un démantèlement partiel.

2.12.1 Présentation du jeu de données

Ce jeu de données contient 70 observations de radioactivité surfacique issues de G3 (plus précisément d'un échangeur de chaleur du réacteur). Les unités sont en Bq/cm^2 et les observations sont réparties selon une maille régulière dans $[0, 6] \times [0, 4]$. La Figure 2.8 donne la carte des observations.

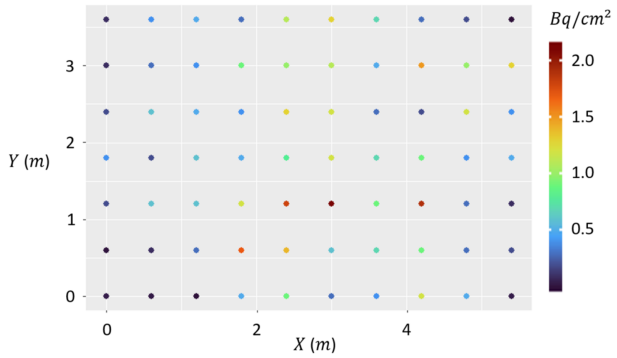


FIGURE 2.8 – Carte des observations issues de G3.

Aucune analyse exploratoire n'est réalisée ici. Le but de cette application étant la comparaison des deux krigeages, le jeu de données n'est pas modifié avant application des deux méthodes. Cette situation se place donc dans un cas désavantageux, où les hypothèses initiales (stationnarité et gaussienne) ne peuvent pas être vérifiées. L'analyse exploratoire sera réalisée dans la section 3.5.4.1.

2.12.2 Protocole

De manière similaire au protocole présenté dans la section 2.11.2.2, deux étapes sont ici nécessaires. La première vient d'abord comparer différentes fonctions de covariance tandis que la seconde compare les deux krigeages pour un nombre variable d'observations.

Pour cette première étape l'ensemble des 70 observations est utilisé. Une validation croisée est appliquée pour différentes fonctions de covariance (exponentielle, Matérn-$\frac{3}{2}$, Matérn-$\frac{5}{2}$ et gaussienne), voir la section 1.9.3.1.

Pour obtenir divers jeux de données de tailles différentes, on tire aléatoirement sans remise dans le jeu de données initial (on perd ainsi la maille régulière). Les différentes tailles d'échantillon sont comprises entre 10 et 70 et ce tirage est répété 100 fois pour chaque taille différente. Pour $n = 70$, on retrouve le jeu initial complet, qui est donc unique. Sur chacun de ces jeux de données est appliquée une validation croisée.

A nouveau les paramètres sont estimés par maximum de vraisemblance pour le krigeage ordinaire, tandis que les lois *a priori* du krigeage bayésien sont identiques à celles données dans la section 2.11.1.2 :

$$\pi(\mu, \sigma^2, \phi) \propto \frac{1}{\sigma^2} \quad \text{avec } \phi \in \mathbb{R}^{+*}.$$

2.12.3 Résultats et Interprétations

Les résultats concernant le choix de fonction de covariance sont synthétisés dans le Tableau 2.3 pour le krigeage ordinaire et le krigeage bayésien et enfin la Figure 2.9 pour les graphes α-CI des deux krigeages.

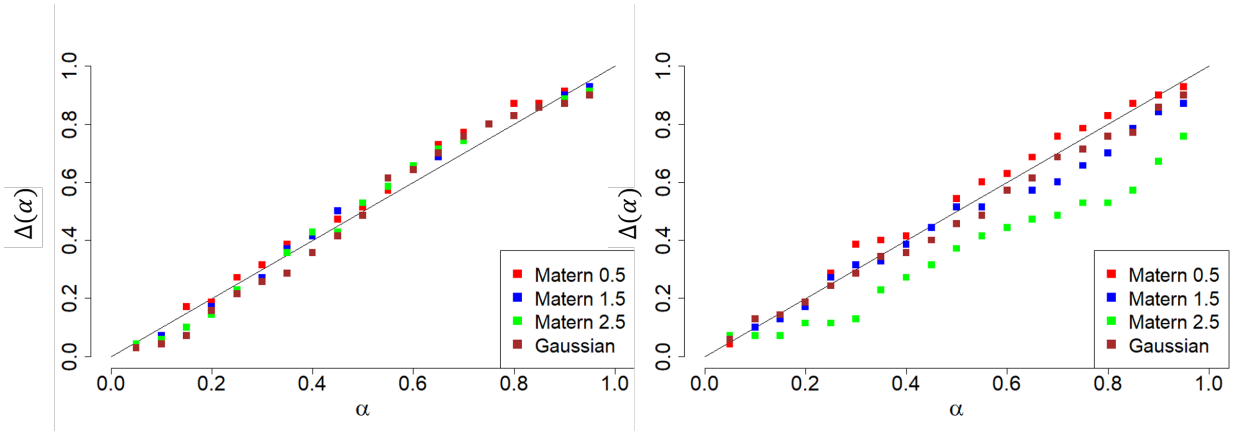
Covariance	Q^2_{ord}	PVA_{ord}	PIA_{ord}	$MSE\alpha_{ord}$	Q^2_{bay}	PVA_{bay}	PIA_{bay}	$MSE\alpha_{bay}$
Matérn 1/2	0.37	0.06	0.07	0.0015	0.38	0.12	0.07	0.0013
Matérn 3/2	0.33	0.12	0.14	0.0010	0.20	0.51	0.55	0.0028
Matérn 5/2	0.31	0.14	0.15	0.0014	0.16	1.19	1.25	0.0284
Gaussien	0.24	0.16	0.18	0.0021	0.15	0.36	0.40	0.0015

TABLE 2.3 – Critères de validation pour le krigeage ordinaire et bayésien avec différentes fonctions de covariance, appliqués sur les données de G3 pour $n = 70$.

La fonction de covariance exponentielle (voir la section 1.9.3.1, équivalente à la covariance de Matérn-$\frac{1}{2}$) fournit les meilleures performances selon nos critères, puisqu'elle maximise le critère Q^2 tout en minimisant les critères PVA et PIA. Seule sa valeur de $MSE\alpha$ n'est pas optimale mais reste correcte comparée aux $MSE\alpha$s des autres fonctions de covariance. Elle sera donc utilisée dans la suite pour le krigeage ordinaire et le krigeage bayésien.

Les résultats de la comparaison des deux krigeages pour différentes tailles de données réelles issues de G3 sont donnés dans les box-plots de la Figure 2.10.

La médiane du critère Q^2 croît d'environ 0 pour 10 observations à 0.38 pour 70 observations pour les deux krigeages. Les valeurs sont légèrement meilleures pour le krigeage bayésien, en particulier pour de petits jeux de données. La dispersion des valeurs est quant à elle identique d'une méthode de krigeage à l'autre. Enfin on remarque que les valeurs obtenues sont très faibles, inférieures à 0.5, ce qui implique que plus de la moitié de la variance prédite n'est pas expliquée par le modèle. [Demay et al., 2022] indique qu'un modèle dont le Q^2 est inférieur à 0.5 n'est pas valide. En pratique il serait donc nécessaire d'obtenir de nouvelles observations ou d'adapter le modèle. Néanmoins notre objectif étant



(a) Ordinary kriging. (b) Bayesian kriging.

FIGURE 2.9 – Graphes α-CI pour le krigeage bayésien avec différentes fonctions de covariance, appliqués sur les données de G3 pour $n = 70$.

la comparaison des deux méthodes et non la construction d'un excellent modèle prédictif, ce point n'est pas davantage discuté.

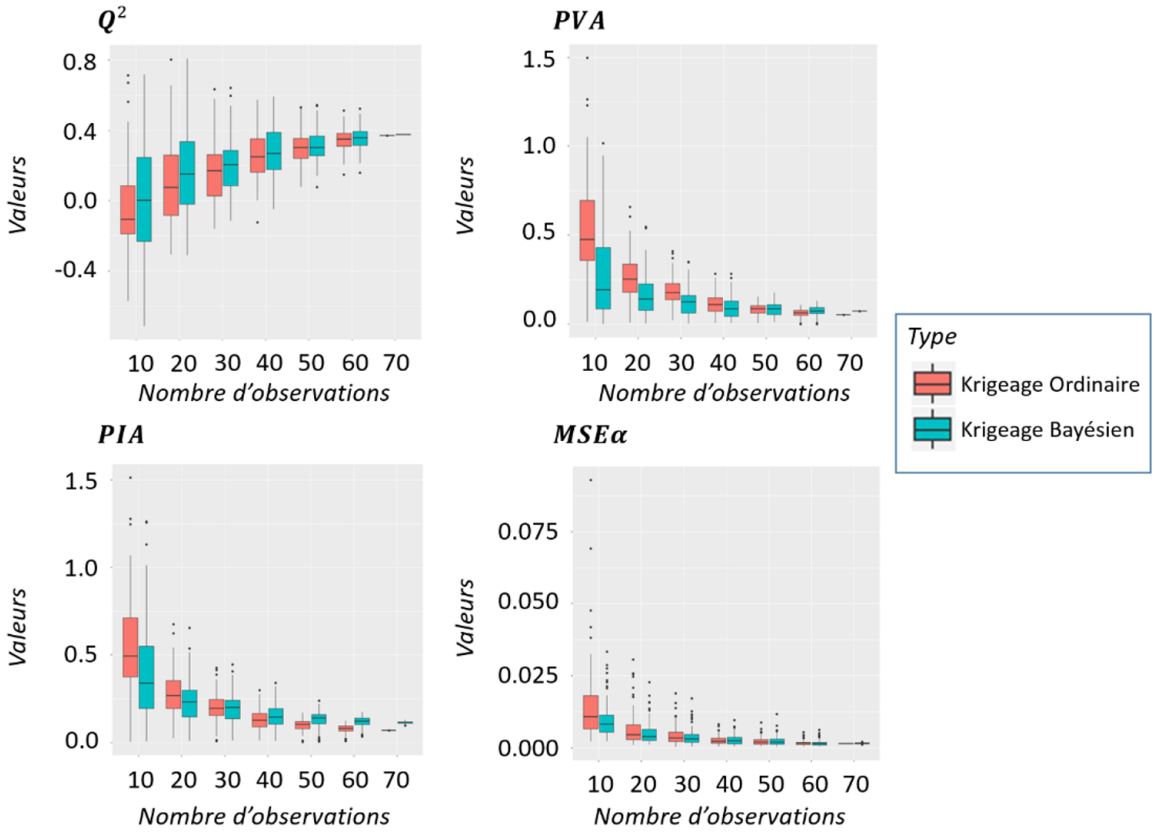


FIGURE 2.10 – Distribution des critères de validation (Q^2, PVA, PIA et $MSE\alpha$) contre la taille de l'échantillon pour les données issues de G3.

Pour le PVA, les valeurs médianes décroissent de 0.47 à 0.16 pour le krigeage ordinaire, des valeurs bien plus grandes que celles observées pour le krigeage bayésien, dont les

médianes décroissent de 0.19 à 0.06. Les valeurs du PIA sont très proches de celles du PVA. Pour le $MSE\alpha$, la médiane décroît de 0.011 à 0.0017 pour le krigeage ordinaire, et de 0.008 à 0.0017 pour le krigeage bayésien. Le krigeage bayésien donne à nouveau de meilleures performances, en particulier pour un faible nombre d'observations.

On peut également remarquer que la dispersion des critères diminue avec la taille de l'échantillon. Cela peut être expliqué par la taille de l'échantillon, mais également par notre protocole, puisque les observations sont tirées sans remise depuis le jeu original de 70 observations. De ce fait, les jeux de données pour un grand nombre d'observations sont de plus en plus similaires, aboutissant à des critères de validation plus proches.

2.13 Conclusion

L'utilisation du krigeage bayésien pour l'interpolation spatiale dans le cadre de l'A&D offre des résultats prometteurs. Il est particulièrement efficace pour un faible nombre d'observations, situation dans laquelle il fournit de meilleures performances que le krigeage ordinaire en terme de moyennes, variances et intervalles de confiance prédits. Néanmoins cet avantage diminue avec l'augmentation du nombre d'observations : le krigeage ordinaire, moins coûteux en terme de calculs, est alors préférable. Le krigeage bayésien a aussi comme défaut de nécessiter la spécification d'une loi *a priori*, qui peut être difficile à choisir et peut fortement influencer les prédictions. Il est pour cette raison recommandé de n'utiliser le krigeage bayésien que pour de petits jeux de données ou dans les situations où l'information *a priori* sur les paramètres du modèle est bien connue.

Dans nos travaux nous n'avons pas utiliser l'effet de pépite comme un outil de modélisation, mais uniquement à des fins de régularisation de la fonction de covariance gaussienne. Des travaux futurs chercheront à ajouter ce paramètre dans le modèle. Cela pourrait être approfondi en considérant un modèle hétéroscédastique ([Ng and Yin, 2012]), puisque l'effet de pépite classique considère un modèle homoscédastique. Ce nouveau modèle pourrait être extrêmement utile dans le cadre de l'A&D de sites nucléaires puisque les mesures sont souvent sujettes à des incertitudes de mesure variables, selon le protocole de mesure employé.

De nombreux outils présentés dans le chapitre précédent, comme le krigeage de bloc, les approches multivariables ou encore les simulations conditionnelles peuvent être adaptées dans l'approche bayésienne, le krigeage bayésien se présentant comme une généralisation de ces méthodes existantes.

CHAPITRE 3

Traitement des données censurées

Dans les sciences environnementales, certaines mesures réalisées correspondent à des concentrations très faibles en composées chimiques. Il est souvent nécessaire d'identifier la présence ou non d'espèces chimiques à des fins d'assainissement ou de contrôle de qualité. Un exemple classique est le contrôle de la qualité de l'eau (voir [Helsel and Gilliom, 1986] ou [Toscas, 2010] pour un exemple en statistiques spatiales). Les normes environnementales évoluant, des contraintes de plus en plus fortes sont appliquées afin de garantir le contrôle des contaminations des milieux naturels par les activités humaines. De plus, certaines espèces chimiques sont dangereuses même sous formes de traces et imposent des évaluations de plus en plus exactes des contaminations dans l'environnement. Néanmoins toute méthode d'analyse physique ou chimique présente des limites à partir desquelles les résultats de mesure sont trop incertains pour être considérés comme des valeurs fiables. C'est dans ce cas de figure que certains laboratoires d'analyse censurent leurs résultats de mesure, ne pouvant pas garantir l'exactitude du résultat. En effet, pour des valeurs de mesure proches de zéro (dites mesures à bas niveaux), la nature aléatoire d'une mesure peut conduire à déclarer à tort l'absence d'un constituant, ou à déclarer à tort sa présence.

3.1 Définition d'une donnée censurée

Les données censurées sont alors définies de la manière suivante :

Censures :
Soient $\nu_1, \nu_2 \in \mathbb{R}$. Une donnée est censurée à gauche si elle est reportée sous la forme :

$$]-\infty, \nu_1] \text{ ou }]-\infty, \nu_1[.$$

Une donnée est censurée à droite si elle est reportée sous la forme :

$$[\nu_2, +\infty[\text{ ou }]\nu_2, +\infty[.$$

Une donnée est censurée par intervalle si elle est reportée sous la forme :

$$]\nu_1, \nu_2[, [\nu_1, \nu_2[,]\nu_1, \nu_2] \text{ ou } [\nu_1, \nu_2].$$

Puisque les données une fois censurées ne sont plus des valeurs quantitatives, des méthodes particulières sont développées pour leur traitement statistique. Initialement, ces méthodes étaient utilisées pour l'étude de la mortalité de patients ou la durée de vie d'une machine. En pratique, ces données étaient censurées à droite lorsqu'un patient n'était plus suivi ou s'il survivait à l'étude. Ces données peuvent être traitées par des méthodes issues d'une branche de la statistique nommée « analyse de survie ». La plupart des méthodes de l'analyse de survie sont néanmoins applicables aux données censurées à gauche par symétrie (voir section 3.3.3).

Ce type de données est souvent rencontré dans l'A&D puisque de nombreuses données peuvent être censurées. Les négliger peut être néfaste, comme expliqué dans [Helsel, 2012]. En effet cela peut conduire à une mauvaise estimation de la radioactivité réelle dans les sols étudiés et éventuellement au relâchement de matière dangereuse dans l'environnement. Le traitement de cette problématique est donc primordial afin de garantir la bonne réalisation des travaux d'A&D et un respect rigoureux des objectifs formulés dans le décret de démantèlement.

Pour traiter cette problématique, la censure de données est d'abord décrite à l'aide des définitions du seuil de décision et de la limite de détection. Ensuite quelques méthodes classiques de traitement des données censurées lorsqu'un échantillon i.i.d. est disponible sont présentées. A la suite de cette brève présentation, la géostatistique et les méthodes permettant le traitement des données censurées lorsque les données présentent une structure spatiale sont décrites. Plusieurs approches seront ici proposées, certaines se basant sur des inférences classiques tandis que d'autres reposent sur des inférences bayésiennes. Enfin ces différentes méthodes sont comparées sur des jeux de données simulés ainsi que le jeu de données de G3.

3.2 Seuil de décision et limite de détection

Une question importante soulevée dans l'introduction précédente est la définition de ν_1 et/ou ν_2. Comment définir ces valeurs, et sur quel critère se baser pour considérer un résultat de mesure comme trop « incertain » ? La réponse à cette question n'est pas unique. [Currie, 1968] fait une liste non-exhaustive des différentes méthodes employées pour censurer un résultat de mesure à l'époque de la rédaction de cet article. Certaines limites étaient calculées selon un pourcentage de l'écart-type du bruit de fond, tandis que d'autres prenaient en compte l'écart-type des répétitions d'une mesure. Face à ces définitions différentes aboutissant à des protocoles de censure différents, [Currie, 1968] propose deux nouvelles définitions, le seuil de décision et la limite de détection, en utilisant la démarche des tests statistiques.

La méthodologie proposée par [Currie, 1968] considère qu'un résultat de mesure est la réalisation d'une variable aléatoire. Plus précisément, un résultat de mesure est considéré comme la réalisation d'une variable aléatoire Z suivant une loi normale d'espérance μ et

de variance σ_0^2 inconnues. On suppose que n observations i.i.d. ont été obtenues pour les résultat de mesure.

Seuil de décision :
On considère que $\mu = 0$. Le seuil de décision, noté z_C, est défini comme la valeur critique pour laquelle la probabilité d'obtenir une valeur de mesure supérieure à cette valeur est égale à P_1. Il est estimé de la manière suivante :

$$z_C \approx t_{1-P_1,c}\sqrt{s_0^2},$$

avec s_0 l'estimation de l'incertitude de mesure σ_0 et $t_{1-P_1,c}$ le quantile d'ordre $1 - P_1$ de la loi de Student à $c = n - 1$ degrés de liberté.

La Figure 3.1 reprend les graphiques de [Crozet et al., 2015] et donne une illustration du concept de seuil de décision.

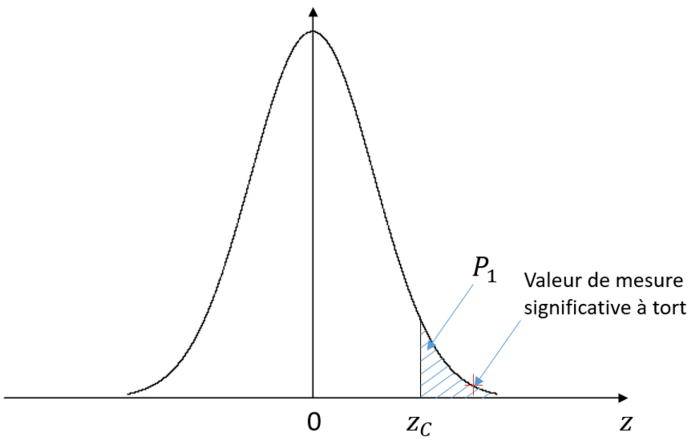


FIGURE 3.1 – Illustration du concept de seuil de décision.

Il permet d'éviter de déclarer à tort l'absence d'un constituant, de radioactivité, etc. avec une probabilité P_1.

Limite de détection :
La limite de détection, notée z_D, est la valeur de μ pour laquelle la probabilité qu'une valeur de mesure soit inférieure au seuil de décision, est égale à P_2. Elle est estimée avec la formule suivante :

$$z_D \approx t_{1-P_2,c}\sqrt{s_0^2} + z_C.$$

La Figure 3.2 reprend également un graphique de [Crozet et al., 2015] et représente graphiquement la limite de détection par rapport au seuil de décision.

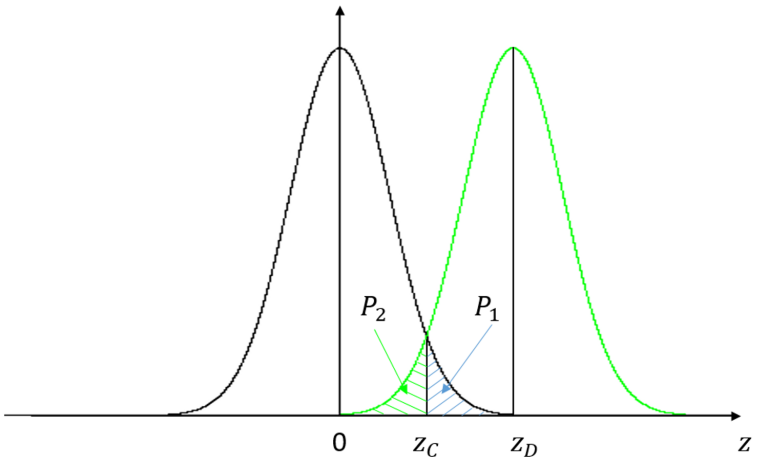


FIGURE 3.2 – Illustration du concept de limite de détection.

La limite de détection permet d'éviter de déclarer à tort la présence d'un constituant avec une probabilité P_2.

Ainsi pour exprimer un résultat de mesure, le protocole suivant est appliqué. Si le résultat de mesure est supérieur au seuil de décision (SD), on peut affirmer, avec un risque P_1 de se tromper, la présence, par exemple, de contamination. La valeur déclarée est alors le résultat de mesure accompagné de son incertitude de mesure. Si le résultat de mesure est inférieur au SD, on peut seulement affirmer qu'elle est inférieure à la limite de détection (LD), avec un risque P_2 de se tromper. Ce protocole est résumé dans la Figure 3.3, en notant $u(z)$ l'estimation de l'incertitude sur z.

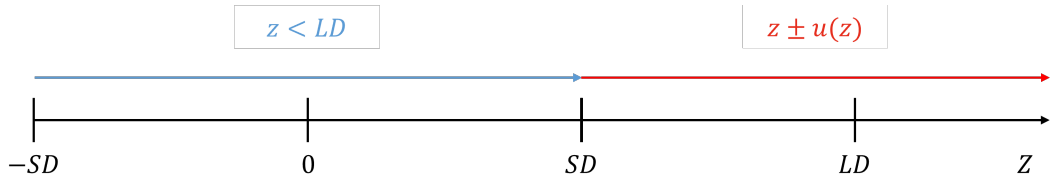


FIGURE 3.3 – Expression du résultat de mesure en fonction de SD et LD.

Dans certains cas, il arrive qu'un résultat de mesure inférieur à la limite de détection soit déclaré inférieur à la limite de détection. Dans tous les cas, au lieu d'une valeur quantitative, un intervalle contenant la valeur mesurée est retenu. Cela implique une perte d'informations non-négligeable, et nécessite ultérieurement des outils statistiques adaptés au traitement de données non-quantitatives lors de l'analyse des résultats de mesure.

Il s'agit d'une situation assez courante pour les mesures environnementales, particulièrement présente dans l'A&D de sites nucléaires (puisque les contaminations des sols sont souvent très faibles).

Ces limites, seuil de décision et limite de détection, sont définies par le mesureur pour chaque méthode de mesure pour la grandeur mesurée. Ainsi il est possible de rencontrer plusieurs limites de détection dans un seul jeu de données selon la méthode de mesure

utilisée, le mesureur ou la grandeur mesurée. Par exemple, pour une même grandeur physique mesurée, une mesure *in-situ* est de manière générale plus incertaine (incertitudes de mesure plus élevées) qu'une mesure en laboratoire. Ces deux types de mesure fournissent donc des limites de détection différentes. Le reste du document ne discutera que de données censurées à gauche et supposera que la limite de détection n'est pas unique pour un jeu de données sauf mention contraire.

3.3 Traitement des données censurées dans le cas d'un échantillon i.i.d.

3.3.1 Notations et définitions

Les données forment donc un jeu contenant des observations quantifiées et des données censurées. Quand un résultat de mesure a été censuré, seul un intervalle est disponible. On note alors :

— le vecteur des données censurées inconnues $\boldsymbol{z}_c$ de taille n_ν ;
— le vecteur des observations $\boldsymbol{z}_o$ de taille $n - n_\nu$;
— le vecteur des limites de détection $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{n_\nu})'$ de taille n_ν.

Dans le cas d'une censure à gauche, la seule information disponible sur les données $\boldsymbol{z}_c$ est que pour tout $i \in [\![1, n_\nu]\!]$, $z_{i,c} \leq \nu_i$. Le vecteur $\boldsymbol{z} = (\boldsymbol{z}_c', \boldsymbol{z}_o')'$ est supposé issu de variables aléatoires i.i.d., mais dont une partie des observations est inconnue. On introduit enfin le vecteur $\boldsymbol{z}^* = (\boldsymbol{\nu}', \boldsymbol{z}_o')'$, identique au vecteur $\boldsymbol{z}$, à la différence que les données censurées inconnues $\boldsymbol{z}_c$ ont été remplacées par leur limite de détection respective.

[Helsel, 2012] décrit l'intérêt et l'importance de prendre en compte correctement les données censurées lors de l'application d'outils statistiques. Dans le cas particulier de la mesure de radioactivité, [Kim et al., 2020] fait un inventaire des méthodes employées pour traiter ces données censurées et donne trois méthodes couramment utilisées qui sont détaillées ici :

— la méthode de maximum de vraisemblance ;
— la méthode Kaplan-Meier (KM) ;
— la méthode Regression on Order Statistics (ROS).

Certaines de ces méthodes sont non-paramétriques, comme la méthode Kaplan-Meier qui repose sur les statistiques d'ordre, tandis que d'autres supposent une distribution pour la variable aléatoire étudiée comme la méthode ROS ou celle du maximum de vraisemblance pour estimer des paramètres.

A ces trois méthodes s'ajoutent les algorithmes EM (Expectation Maximisation) et l'augmentation de données qui reposent sur des imputations réalisées dans un algorithme itératif.

Imputation :
L'imputation est une procédure statistique consistant à remplacer des données manquantes à l'aide d'estimations ou de prédictions réalisées à partir d'informations

supplémentaires issues du jeu de données initial.

3.3.2 Maximum de vraisemblance

La méthode par maximum de vraisemblance est proche de celle présentée dans le chapitre 1. L'approche est paramétrique et les données sont associées à une famille de densités paramétrées.

Vraisemblance pour des données censurées [Helsel, 2012] :
On suppose que les données avant censure sont issues d'une variable aléatoire Z de densité de probabilité $f_Z(.|\theta)$ et de fonction de répartition F. Dans le cas d'une censure à gauche, avec un vecteur contenant les limites de détection $\boldsymbol{\nu}$, l'expression de la vraisemblance des données $\boldsymbol{z} = (\boldsymbol{z}_c', \boldsymbol{z}_o')'$ est alors :

$$L(\boldsymbol{z}, \theta) = \prod_{i=1}^{n_\nu} F(\nu_i) \prod_{i=n_\nu+1}^{n} f_Z(z_i|\theta).$$

De la même manière que pour le maximum de vraisemblance usuel (voir section 1.2.2), ce maximum de vraisemblance nécessite un grand nombre de données pour donner une estimation robuste du paramètre θ (environ 50 pour des distributions symétriques, de l'ordre de 70 pour des distributions asymétriques selon [Helsel, 2012]).

3.3.3 Méthode Kaplan-Meier

La méthode de Kaplan-Meier (KM) est une méthode non-paramétrique utilisant les approches de l'analyse de survie. Dans sa conception initiale, cette méthode s'applique à des données censurées à droite et utilise la fonction de survie pour son estimateur (voir [Kaplan and Meier, 1958]). On suppose alors ici que nos observations sont positives.

Fonction de survie [Saporta, 1990] :
La fonction de survie d'une variable aléatoire Z est l'application S de $\mathbb{R}$ dans $[0, 1]$ définie par :

$$S(z) = \mathbb{P}(Z \geq z).$$

Dans le cas d'une censure à gauche, il est donc d'abord nécessaire de « retourner » le vecteur $\boldsymbol{z}^*$ à l'aide d'une constante m pour pouvoir appliquer la méthode à des données censurées à droite. Cette constante m est choisie telle que $m = max\{z_1^*, \ldots, z_n^*\}$. On introduit alors le vecteur $\boldsymbol{y}^* = (y_1^*, \ldots, y_n^*)'$ des données retournées tel que pour tout $i \in [\![1, n]\!]$:

$$y_i^* = m - z_i^*.$$

La méthode nécessite ensuite d'ordonner ces données. On note alors $\boldsymbol{y}_{ord}^*$ le vecteur contenant l'ensemble des données ordonnées de manière croissante. Le vecteur $\tilde{\boldsymbol{y}}_{ord}^*$ de taille $\tilde{n}$ contenant toutes les valeurs distinctes des $y_{i,ord}^*$ est ensuite introduit.

Estimation de Kaplan-Meier :
Soit un jeu de données ordonné $\boldsymbol{y}^*_{ord}$ et $\tilde{\boldsymbol{y}}^*_{ord}$ le vecteur contenant toutes les valeurs distinctes des $y^*_{i,ord}$. On définit ensuite pour tout $i \in [\![1, \tilde{n}]\!]$:

— n_i le nombre de données y^*_i supérieures ou égales à $\tilde{y}^*_{i,ord}$,
— o_i le nombre d'observations (sans les limites de détection) y^*_i égales à $\tilde{y}^*_{i,ord}$.

L'estimation de Kaplan-Meier est alors :

$$\forall y \in \mathbb{R},\ \hat{S}^*_{KM}(y) = \begin{cases} 1 \text{ si } y \leq \tilde{y}^*_{1,ord}, \\ \prod_{j=1}^{i} \frac{n_j - o_j}{n_j} \text{ si } y \in \left]\tilde{y}^*_{i,ord}, \tilde{y}^*_{i+1,ord}\right] \text{ avec } i \in [\![1, \tilde{n}-1]\!], \\ 0 \text{ si } y > \tilde{y}^*_{\tilde{n},ord}. \end{cases}$$

La fonction de survie est ensuite utilisée pour réaliser l'estimation de statistiques descriptives comme la moyenne, la variance ou les quantiles. Par exemple, l'estimation KM de la moyenne de y^* est donnée par :

$$\hat{\mu}^*_{KM} = \int_{[0, \tilde{y}^*_{\tilde{n},ord}]} \hat{S}^*_{KM}(y)dy.$$

La fonction $\hat{S}^*_{KM}(.)$ étant constante par morceaux, cette intégrale peut se réécrire sous la forme d'une somme :

$$\hat{\mu}^*_{KM} = \sum_{i=1}^{\tilde{n}} (\tilde{y}^*_{i,ord} - \tilde{y}^*_{i-1,ord}) \hat{S}^*_{KM}(\tilde{y}^*_{i,ord}),$$

avec la convention $y_0 = 0$.

La fonction de répartition de Z estimée par la méthode KM est alors obtenue avec la relation suivante, pour tout $z \in \mathbb{R}$:

$$\hat{F}^*_{KM}(z) = \hat{S}^*_{KM}(m - z).$$

La Figure 3.4 (réalisée avec le package `R NADA`) donne un aperçu de l'estimation Kaplan-Meier, le jeu de données utilisé étant le suivant :

$$(0.3, 0.3, 0.4, 0.5, < 1, 1.2, 1.3, < 1.5, < 1.5, < 1.5, 1.6, 1.7, 1.9, 1.9, < 2, < 2, < 2, < 2, 2.5, 3)$$

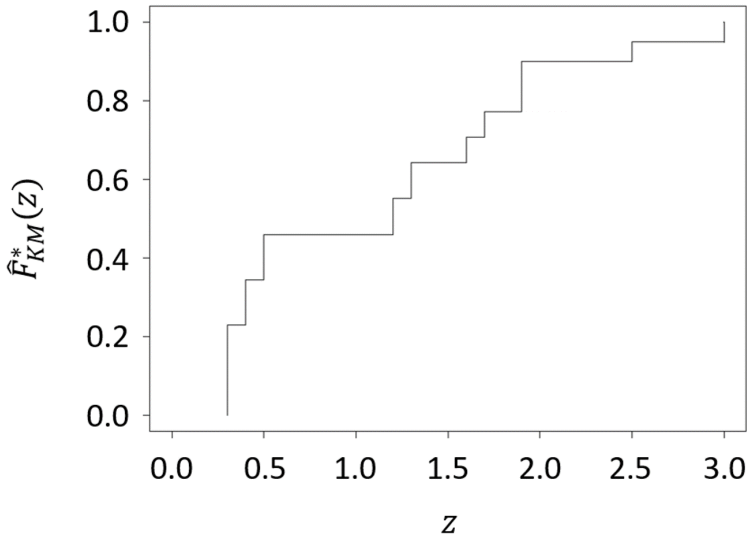


FIGURE 3.4 – Exemple d'application de l'estimateur de Kaplan-Meier pour l'estimation d'une fonction de répartition empirique.

Il existe une version de la méthode KM permettant le traitement de censures par intervalle appelée méthode des intervalles de Turnbull. Néanmoins, les censures par intervalle rencontrées en A&D sont le plus souvent de la forme $[0, \text{limite}[$, et sont équivalentes dans le traitement de la méthode KM à une censure à gauche. La méthode des intervalles de Turnbull ne sera donc pas développée ici, mais davantage d'informations peuvent être trouvées dans [Helsel, 2012].

3.3.4 Méthode de régression sur des statistiques d'ordre (Regression on Order Statistics)

La méthode ROS (« Regression on Order Statistics ») fait appel, comme son nom l'indique, aux statistiques d'ordre décrites dans [Wilks, 1948]. Il s'agit d'une méthode paramétrique utilisant des prédictions sur les données censurées pour estimer les statistiques d'intérêt.

Dans un premier temps, les données censurées ne sont pas prises en compte (on ne travaille alors qu'avec l'échantillon $\boldsymbol{z}_o$). Les observations sont alors ordonnées pour former le vecteur $\boldsymbol{z}_{o,ord}$. On suppose que les observations sont i.i.d. de loi normale, de moyenne μ et de variance σ^2. Cette hypothèse n'est pas utilisée pour l'estimation finale de statistiques par la méthode, mais uniquement pour réaliser des prédictions qui viendront remplacer les données censurées lors de l'estimation des statistiques d'intérêt. Les paramètres μ et σ^2 sont estimés à l'aide d'un ajustement par moindres carrés sur un graphe quantile-quantile (graphe représentant les quantiles empiriques contre les quantiles théoriques d'une loi).

Il s'agit alors d'un modèle de régression :

$$z_{i,o,ord} = \mu + \sigma F_{\mathcal{N}}^{-1}\left(q_{\frac{i}{n-n_\nu}}\right),$$

où $z_{i,o,ord}$ est la i^{e} composante du vecteur ordonné $\boldsymbol{z}_{o,ord}$, $F_{\mathcal{N}}^{-1}$ est la fonction de répartition inverse de la loi normale centrée réduite et $q_{\frac{i}{n-n_\nu}}$ un quantile d'ordre $\frac{i}{n-n_\nu}$ des données non-censurées.

Les données censurées sont ensuite prédites à l'aide des paramètres estimés. C'est à partir de ces prédictions et des observations que les statistiques d'intérêt sont estimées. La méthode de prédiction ne sera pas développée ici, mais davantage de détails peuvent être trouvés dans [Helsel, 2012]. La Figure 3.5 (également réalisée avec le package `R NADA`) donne un exemple de graphe obtenu par la méthode ROS avec les données présentées dans la section 3.3.3.

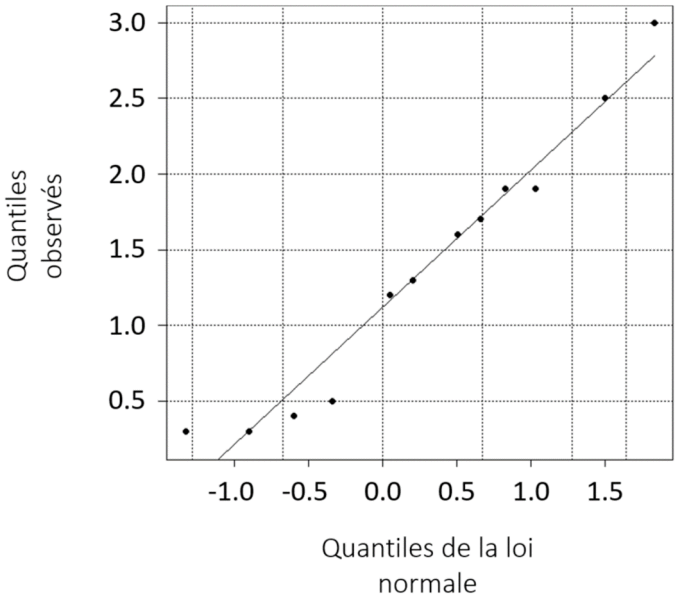


FIGURE 3.5 – Exemple de graphe Quantile-Quantile obtenu dans le cadre de la méthode ROS.

3.3.5 Algorithme EM pour les données censurées

Cet algorithme initialement conçu par [Dempster et al., 1977] a pour objectif l'estimation de statistiques ou de paramètres par maximum de vraisemblance en présence de variables latentes inconnues ; un exemple classique d'application est la classification automatique, où la variable latente correspond à un indice identifiant l'appartenance d'une observation à un certain groupe. L'algorithme calcule la vraisemblance des données complètes, qui prend en compte des données observées et latentes, puis ré-estime par maximum de vraisemblance les paramètres de la distribution de la variable aléatoire étudiée. En considérant les données censurées comme des variables latentes, il est alors possible d'appliquer l'algorithme EM au traitement de données censurées.

On considère que les observations sont issues d'une variable aléatoire à valeurs dans $\mathbb{R}$ et de densité de probabilité paramétrée par θ. On sépare le vecteur des données en un vecteur observé $\boldsymbol{z}_o$ et un vecteur de données censurées inconnu $\boldsymbol{z}_c$ de taille n_ν.

Algorithme EM

— **Initialisation** : Choix de ϵ et choix aléatoire de $\theta^{(0)}$.

— **A l'étape** j : Tant que $|L(\boldsymbol{z}_o; \boldsymbol{z}_c, \theta^{(j)}) - L(\boldsymbol{z}_o; \boldsymbol{z}_c, \theta^{(j+1)})| < \epsilon$, avec $L(\boldsymbol{z}_o; \boldsymbol{z}_c, \theta)$ la vraisemblance de l'ensemble des données :

— Etape E, évaluation de $Q(\theta, \theta^{(j)})$:

$$\begin{aligned} Q(\theta, \theta^{(j)}) &= \mathbb{E}_{\boldsymbol{Z}_c|\boldsymbol{z},\theta^{(j)}}[\log(L(\boldsymbol{z}_o; \boldsymbol{z}_c, \theta))] \\ &= \int_{\mathbb{R}^{n_\nu}} \log(L(\boldsymbol{z}_o; \boldsymbol{z}_c, \theta)) L(\boldsymbol{z}_c; \theta^{(j)}) d\boldsymbol{z}_c, \end{aligned}$$

où $L(\boldsymbol{z}_c; \theta^{(j)})$ est la vraisemblance des données censurées.

— Etape M : Maximisation de $Q(\theta, \theta^{(j)})$ selon θ. La valeur θ pour laquelle $Q(\theta, \theta^{(j)})$ atteint son maximum est noté $\theta^{(j+1)}$.

Une étape implicite de l'algorithme est l'évaluation de $\boldsymbol{Z}_c$ conditionnellement à $\boldsymbol{z}_o$ et θ pour le calcul de $L(\boldsymbol{z}_o; \boldsymbol{z}_c, \theta)$. Cette étape est propre au contexte et peut rendre l'application de l'algorithme difficile. De manière générale, l'étape E n'est pas réalisable analytiquement. Des approximations par simulation sont alors souvent utilisées pour résoudre ce problème.

Une variante de l'algorithme EM appelé MCEM (Monte-Carlo Expectation Maximisation) utilise des tirages Monte-Carlo pour obtenir cette approximation. Une autre variante, l'algorithme SAEM (Stochastic Approximation Expectation Maximisation) utilise une approche similaire mais ajoute une approximation stochastique à l'étape E limitant le nombre de tirages nécessaires pour réaliser l'évaluation de $Q(\theta, \theta^{(j)})$. Un autre intérêt de l'algorithme SAEM est qu'il permet d'éviter de converger vers un maximum local de la vraisemblance (qui est un problème récurrent des algorithmes EM) en utilisant un paramètre de mémoire lors de l'approximation statistique. Ce paramètre permet alors de prendre en compte les précédents tirages pour la nouvelle évaluation de $Q(\theta, \theta^{(j)})$. Plus de détails sur ces variantes peuvent être trouvées dans [McLachlan and Krishnan, 1997] et [Kuhn and Lavielle, 2004].

3.3.6 Augmentation de données

[Tanner and Wong, 1987] propose une méthode d'estimation de la distribution *a posteriori* des paramètres de la loi de la variable aléatoire étudiée lorsque certaines informations d'intérêt sont absentes. Cela correspond donc à une version bayésienne de l'algorithme EM. Nous reprenons les notations de la section précédente. On cherche ici à estimer le paramètre $\theta \in \Theta$ avec des imputations pour les données censurées.

L'expression de la loi *a posteriori* du paramètre θ peut s'écrire :

$$f_\theta(\theta|\boldsymbol{z}_o) = \int_{\mathbb{R}^{n_\nu}} f_\theta(\theta|\boldsymbol{z}_o, \boldsymbol{z}_c) f_{\boldsymbol{Z}_c}(\boldsymbol{z}_c|\boldsymbol{z}_o) d\boldsymbol{z}_c.$$

De même pour la loi *a posteriori* des données censurées, on a :

$$f_{\boldsymbol{Z}_c}(\boldsymbol{z}_c|\boldsymbol{z}_o) = \int_\Theta f_{\boldsymbol{Z}_c}(\boldsymbol{z}_c|\boldsymbol{z}_o, \boldsymbol{\phi}) f_\theta(\theta|\boldsymbol{z}_o) d\theta.$$

En combinant ces deux expressions et en introduisant la variable d'intégration θ_{int} pour la différencier de la variable θ, on peut ré-exprimer la loi *a posteriori* de θ :

$$\begin{aligned} f_\theta(\theta|\boldsymbol{z}_o) &= \int_\Theta \left(\int_{\mathbb{R}^{n_\nu}} f_\theta(\theta|\boldsymbol{z}_o, \boldsymbol{z}_c) f_{\boldsymbol{Z}_c}(\boldsymbol{z}_c|\boldsymbol{z}_o, \theta_{int}) d\boldsymbol{z}_c \right) f_\theta(\theta_{int}|\boldsymbol{z}_o) d\theta_{int} \\ &= \int_\Theta K(\theta, \theta_{int}) f_\theta(\theta_{int}|\boldsymbol{z}_o) d\theta_{int}. \end{aligned}$$

On définit la transformation T telle que :

$$T(f_\theta)(\theta|\boldsymbol{z}_o) = \int_\Theta K(\theta, \theta_{int}) f_\theta(\theta_{int}|\boldsymbol{z}_o) d\theta_{int}.$$

A l'aide de cette transformation on définit le processus itératif :

$$f_{(j+1)}(\theta|\boldsymbol{z}_o) = T(f_{(j)})(\theta|\boldsymbol{z}_o),$$

qui converge simplement vers $f_\theta(\theta|\boldsymbol{z}_o)$ ([Tanner and Wong, 1987]).

L'algorithme nécessaire au calcul de ces intégrales est présenté ici.

Algorithme d'augmentation de données :

- **Initialisation** : Choix aléatoire de $\theta^{(0)}$ et $\boldsymbol{z}_c^{(0)}$. Calcul de $f_{(1)}(.)$,
- **A l'itération** $j \in [\![1, M]\!]$:
  - Échantillonnage de θ depuis $f_{\theta^{(j)}}(\theta)$,
  - Échantillonnage de m valeurs $\boldsymbol{z}_c^{(j,k)}$, $k \in [\![1, m]\!]$, issues de la loi de densité $f_{\boldsymbol{Z}_c}(.|\boldsymbol{z}_o, \theta^{(j)})$,
- **Approximation de la loi** : Calcul de l'approximation $f_{(j+1)}(.)$ avec

$$f_{(j+1)}(\theta) = \frac{1}{m} \sum_{k=1}^{m} f_\theta(\theta|\boldsymbol{z}_o, \boldsymbol{z}_c^{(k)}).$$

Présenté ainsi l'algorithme suppose qu'il est possible d'échantillonner selon la loi *a posteriori* complète des paramètres $f_\theta(\theta|\boldsymbol{z}_o, \boldsymbol{z}_c)$. S'il est possible d'obtenir cette distribution à l'aide du théorème de Bayes, elle peut être difficile à appliquer. Ainsi à l'instar de l'algorithme EM, l'augmentation de données peut être difficile à mettre en place selon le contexte et les distributions mises en jeu.

Il est possible d'étendre le parallèle entre algorithme EM et augmentation de données. L'étape d'imputation (échantillonnage de θ et $\boldsymbol{z}_c$) est similaire à l'étape E, tandis que l'étape M est similaire à l'étape d'approximation (appelée étape « postérieure » dans [Tanner and Wong, 1987]). Enfin l'algorithme EM fournit une valeur pour le paramètre θ, tandis que l'augmentation de données fournit la distribution *a posteriori* du paramètre.

Cet algorithme appartient à la classe plus générale des échantillonneurs de Gibbs ([van Dyk and Meng, 2001]) et son utilisation est très répandue pour la construction d'algorithmes MCMC. En effet l'introduction d'une variable latente peut faciliter l'expression de certaines marginales en rajoutant un conditionnement à une variable supplémentaire (ici la variable correspondant aux données censurées). Dans certains cas cette étape est cruciale pour la réalisation de l'algorithme MCMC (voir les exemples donnés dans [Tanner and Wong, 1987]).

3.4 Traitement des données censurées dans le cas de présence d'une structure spatiale.

Si une structure spatiale existe, certains outils présentés ne peuvent être appliqués, puisque l'hypothèse d'indépendance n'est plus respectée. D'autres méthodes doivent donc être employées. Plusieurs approches sont données ici, en commençant par les solutions les plus simples.

Les approches présentées dans la suite se basent principalement sur l'idée d'imputer des valeurs aux données censurées grâce à des prédictions par krigeage. Ces imputations seront ensuite utilisées pour le reste des inférences statistiques comme l'estimation des paramètres du modèle ou encore la réalisation de prédictions. Ces imputations et estimations sont ensuite répétées jusqu'à validation d'un critère de convergence. Il est cependant important de noter que ces méthodes perdent en efficacité si le taux de données censurées est grand ($< 50\%$).

3.4.1 Quelques approches simples

Une première approche consiste à ne pas traiter ces données censurées, et donc à ne travailler qu'avec les données observées. Bien sûr cette approche n'a de sens que si le taux de données censurées est faible et que le nombre de données est grand, ce qui est rare dans le cas de l'A&D.

Une approche plus nuancée consiste à remplacer les données censurées par des valeurs arbitraires. Ainsi il est courant de remplacer les données censurées par LD, $LD/2$ ou encore par 0. Les outils classiques présentés dans le chapitre 1 sont alors appliqués à ces données modifiées, malgré des hypothèses de modélisation non respectées. Ce choix surestime sciemment le résultat de l'estimation, par exemple pour une contamination radioactive (pour l'A&D) en remplaçant les données censurées par LD ou sous-estime une estimation de la teneur en métaux (dans le cas de l'industrie minière) en les remplaçant par 0. Néanmoins comme démontré dans [Helsel, 2012], ces pratiques, si elles sont mal employées, peuvent avoir des conséquences importantes. Elles peuvent conduire à une surestimation massive des coûts de démantèlement ou plus problématique encore, à l'oubli de matières contaminées sur un site en cours d'assainissement. Ces méthodes sont donc à éviter, malgré leur facilité d'utilisation (voir [Crozet et al., 2015] pour des démonstrations des biais induits par l'application de ces méthodes lors de cumuls de mesure).

3.4.2 Notations et modèle spatial

Réintroduisons les notations du chapitre 1. On considère que les observations sont issues d'un champ aléatoire réel $\{Z(\boldsymbol{x}), \boldsymbol{x} \in D \subset \mathbb{R}^d\}$, avec $d \in \{1,2,3\}$, modélisant la grandeur physique étudiée. On suppose de plus que le champ peut se réécrire sous la forme :

$$\forall \boldsymbol{x} \in D, Z(\boldsymbol{x}) = \sum_{k=1}^{p} d_k(\boldsymbol{x})\mu_k + \omega(\boldsymbol{x}) + \kappa(\boldsymbol{x}),$$

avec d_k les p covariables, μ_k les coefficients associés, $\omega(.)$ un champ aléatoire centré, sans pépite, stationnaire d'ordre 2 et isotrope, et κ les erreurs de mesure (i.i.d. de loi normale centrée, de variance τ^2), de telle sorte que

$$\forall \boldsymbol{x}, \boldsymbol{x'} \in D, \mathrm{Cov}(Z(\boldsymbol{x}), Z(\boldsymbol{x'})) = \sigma^2 C_\phi(\|\boldsymbol{x} - \boldsymbol{x'}\|) + \tau^2,$$

où C_ϕ est la fonction de covariance de $\omega(.)$ et $\boldsymbol{\mu}, \sigma^2, \phi, \tau^2$ correspondent respectivement à la moyenne, la variance, la portée et l'effet de pépite. Les données sont regroupées dans le vecteur $\boldsymbol{z} = (z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_n))'$ avec $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in D$ les positions des n données. Enfin on note $\boldsymbol{R}$ la matrice de covariance du vecteur gaussien $\boldsymbol{Z} = (Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n)'$, de telle sorte que :

$$\boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{D\mu}, \boldsymbol{R}),$$

avec $\boldsymbol{D} = (d_k(\boldsymbol{x}_i))_{1\leq i\leq n, 1\leq k\leq p}$, $\boldsymbol{\mu} = (\mu_k)'_{1\leq k\leq p}$ et $\boldsymbol{R} = (\sigma^2 (C(|\boldsymbol{x}_i - \boldsymbol{x}_j|) + \eta^2))_{1\leq i,j\leq n}$ où $\eta^2 = \frac{\tau^2}{\sigma^2}$. On note enfin $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma^2, \phi, \eta^2)'$.

De plus on suppose que certaines observations ont été censurées à gauche. Sans perte de généralité on réorganise les données de telle sorte que :

$$\begin{cases} \boldsymbol{z} = (\boldsymbol{z}'_c, \boldsymbol{z}'_o)', \\ \boldsymbol{D} = (\boldsymbol{D}_c, \boldsymbol{D}_o)', \\ \boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_{cc} & \boldsymbol{R}_{co} \\ \boldsymbol{R}_{oc} & \boldsymbol{R}_{oo} \end{pmatrix}, \end{cases}$$

avec $\boldsymbol{z}_o$ le vecteur colonne des observations, $\boldsymbol{z}_c$ le vecteur colonne des données censurées, $\boldsymbol{D}_o$ la matrice des covariables des observations, $\boldsymbol{D}_c$ la matrice des covariables des données censurées, $\boldsymbol{R}_{oo}$ la matrice de covariance des observations, $\boldsymbol{R}_{cc}$ la matrice de covariance des données censurées et $\boldsymbol{R}_{oc} = \boldsymbol{R}'_{co}$ la matrice de covariance entre observations et données censurées. Les termes $\boldsymbol{z}$ et $\boldsymbol{R}$ sont de manière générale partiellement connus, les données censurées n'étant pas connues.

De manière similaire aux notations de la section 3.3.1, on introduit le vecteur $\boldsymbol{\nu}$ de taille n_ν, contenant les limites de détection associées aux données censurées $z_{i,c}$. Enfin on note A_i les intervalles de censure tels que pour tout $i \in [\![1, n_\nu]\!]$, $z_{i,c} \in A_i$. Dans le cas de censures à gauche, on a pour tout $i \in [\![1, n_\nu]\!]$, $A_i =]-\infty, \nu_i]$.

Enfin dans le cas d'approches bayésiennes, les lois *a priori* propres considérées sont paramétrées de la manière suivante :

$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{m}_0, \boldsymbol{R}_0), \sigma^2 \sim \mathcal{IG}(a_\sigma, S^2_\sigma), \phi \sim \mathcal{IG}(a_\phi, S^2_\phi), \tau^2 \sim \mathcal{IG}(a_\tau, S^2_\tau).$$

où l'on rappelle que $\mathcal{IG}$ correspond à loi inverse-gamma décrite dans la section 2.4.

Nous reviendrons sur le choix des paramètres et des lois dans la section 3.4.6.

3.4.3 Approche par simulation de Monte-Carlo

Une première idée de traitement des données censurées consiste à simuler les résultats de mesure censurés par tirage aléatoire à partir de certaines hypothèses. Par exemple sous l'hypothèse gaussienne, il est possible de connaître la loi des données qui ont été censurées à partir des observations et à l'aide du krigeage simple :

$$\boldsymbol{Z}_c|\boldsymbol{z}_o, \boldsymbol{\theta} \sim \mathcal{N}_{n_\nu}(\boldsymbol{D}_c\boldsymbol{\mu} - \boldsymbol{R}_{co}\boldsymbol{R}_{oo}^{-1}(\boldsymbol{z}_o - \boldsymbol{D}_o\boldsymbol{\mu}), \boldsymbol{R}_{cc} - \boldsymbol{R}_{co}\boldsymbol{R}_{oo}^{-1}\boldsymbol{R}_{co}').$$

Ces données ayant été censurées, cette loi multivariée normale est restreinte à $A_1 \times \cdots \times A_{n_\nu}$ pour obtenir :

$$\boldsymbol{Z}_c|\boldsymbol{z}_o, \boldsymbol{\theta}, (A_i)_{1\leq i\leq n_\nu} \sim \mathcal{NT}_{n_\nu}(\boldsymbol{D}_c\boldsymbol{\mu} - \boldsymbol{R}_{co}\boldsymbol{R}_{oo}^{-1}(\boldsymbol{z}_o - \boldsymbol{D}_o\boldsymbol{\mu}), \boldsymbol{R}_{cc} - \boldsymbol{R}_{co}\boldsymbol{R}_{oo}^{-1}\boldsymbol{R}_{co}'; (A_i)_{1\leq i\leq n_\nu}),$$

où $\mathcal{NT}_{n_\nu}$ est la loi multivariée normale tronquée.

La méthode est décrite dans l'algorithme suivant.

Approche Monte-Carlo pour le traitement des données censurées :

— **Etape 0** : Calcul des paramètres $\boldsymbol{\theta}^{(0)}$ à partir des données observées par maximum de vraisemblance :

$$\boldsymbol{\theta}^{(0)} = \underset{\boldsymbol{\theta}\in D_{\boldsymbol{\theta}}}{\operatorname{argmax}} \left(-\frac{1}{2}(\log(\det(\boldsymbol{R}_{oo}))) - \frac{1}{2}\left(\boldsymbol{z}_o - \boldsymbol{D}_o\boldsymbol{\mu}\right)' \boldsymbol{R}_{oo}^{-1}\left(\boldsymbol{z}_o - \boldsymbol{D}_o\boldsymbol{\mu}\right) \right),$$

avec $D_{\boldsymbol{\theta}} = \mathbb{R} \times \mathbb{R}^{+*} \times \mathbb{R}^{+*} \times \mathbb{R}^{+}$.

— **Etape** $j \in [\![1, M]\!]$:
  — Échantillonnage de $\boldsymbol{Z}_c|\boldsymbol{\theta}^{(0)}, (A_i)_{1\leq i\leq n_\nu}$ depuis :

$$\mathcal{NT}_{n_\nu}(\boldsymbol{D}_c\boldsymbol{\mu}^{(0)} - \boldsymbol{R}_{co}\boldsymbol{R}_{oo}^{-1}(\boldsymbol{z}_o - \boldsymbol{D}_o\boldsymbol{\mu}^{(0)})\boldsymbol{R}_{cc} - \boldsymbol{R}_{co}\boldsymbol{R}_{oo}^{-1}\boldsymbol{R}_{co}'; (A_i)_{1\leq i\leq n_\nu}).$$

  — Estimation du paramètre $\boldsymbol{\theta}$ par l'estimateur $\boldsymbol{\theta}^{(j)}$ du maximum de vraisemblance.

— **Prédictions** : Utilisation du krigeage simple en injectant les valeurs moyennes des paramètres :

$$\begin{cases} \overline{\boldsymbol{\mu}} = \frac{1}{M}\sum_{k=1}^{M} \boldsymbol{\mu}^{(k)} \\ \overline{\sigma}^2 = \frac{1}{M}\sum_{k=1}^{M} \sigma^{2(k)} \\ \overline{\phi} = \frac{1}{M}\sum_{k=1}^{M} \phi^{(k)} \\ \overline{\tau}^2 = \frac{1}{M}\sum_{k=1}^{M} \tau^{2(k)} \end{cases}$$

Cette approche fonctionne donc de manière similaire au krigeage bayésien empirique (voir [Krivoruchko and Gribov, 2019]) en réalisant des simulations des données censurées (avec les estimations des paramètres faites avec les données observées). Cependant les différentes simulations sont indépendantes, de poids identiques dans le calcul des paramètres. Ainsi la vraisemblance de ces différentes simulations n'est pas prise en compte, et la variabilité de ces simulations dépend entièrement des paramètres estimés initialement.

Cet algorithme possède plusieurs avantages. Tout d'abord il ne nécessite qu'une seule spécification (celle du nombre d'itérations), et est extrêmement simple à mettre en place et à utiliser. Un de ses défauts est ne pas utiliser les imputations des données censurées dans les prédictions, ce qui évite tout de même d'introduire des variances ou des biais injustifiés par les simulations des données censurées.

3.4.4 Algorithme EM en géostatistique

L'application de l'algorithme EM en statistique spatiale a été initialement proposée par [Militino and Ugarte, 1999] dans le cas d'un champ aléatoire gaussien en considérant la structure de corrélation connue. Les données sont transformées de manière à être décorrélées afin d'appliquer l'algorithme EM. Avec cette approche, seules la moyenne et la variance sont considérées comme inconnues et à estimer. Une autre approche proposée par [Ordoñez et al., 2018] utilise une variante de l'algorithme EM appelé SAEM (pour « Stochastic Approximation Expectation Maximisation »). Cette méthode est disponible dans le package `R CensSpatial`. Les notations utilisées dans la suite reprennent celles de la section précédente.

Pour réaliser l'estimation des paramètres, on écrit la log-vraisemblance des données (observations et données censurées) $\boldsymbol{Z} = (\boldsymbol{Z}_c', \boldsymbol{Z}_o')'$:

$$\log(L(\boldsymbol{z}, \boldsymbol{\theta})) = -\frac{1}{2}\left(n\log(2\pi) + \log(\det(\boldsymbol{R})) + (\boldsymbol{z} - \boldsymbol{D\mu})'\boldsymbol{R}^{-1}(\boldsymbol{z} - \boldsymbol{D\mu})\right).$$

Enfin on note les deux premiers moments conditionnels estimés à l'étape j :

$$\begin{cases} \widehat{\boldsymbol{Z}}^{(j)} = \mathbb{E}[\boldsymbol{Z}|\boldsymbol{\theta}^{(j)}, (A_i)_{1\leq i\leq n_\nu}] \\ \widehat{\boldsymbol{ZZ}}^{(j)} = \mathbb{E}[\boldsymbol{ZZ'}|\boldsymbol{\theta}^{(j)}, (A_i)_{1\leq i\leq n_\nu}]. \end{cases}$$

L'étape E de l'algorithme EM permet de calculer, pour une estimation $\boldsymbol{\theta}^{(j)}$ à l'itération j des paramètres, l'espérance de cette log-vraisemblance :

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(j)}) &= \mathbb{E}\left[\log(L(\boldsymbol{Z}, \boldsymbol{\theta}))|\boldsymbol{z}_o, \boldsymbol{\delta}, \boldsymbol{\theta}^j\right] \\ &= -\frac{1}{2}\left(n\log(2\pi) + \log(\det(\boldsymbol{R})) + \mathbb{E}\left[(\boldsymbol{Z} - \boldsymbol{D\mu})'\boldsymbol{R}^{-1}(\boldsymbol{Z} - \boldsymbol{D\mu})|\boldsymbol{z}_o, \boldsymbol{\delta}, \boldsymbol{\theta}^j\right]\right) \\ &= -\frac{1}{2}\left(n\log(2\pi) + \log(\det(\boldsymbol{R})) + K^{(j)}\right) \end{aligned}$$

où $K^{(j)} = tr\left(\widehat{\boldsymbol{ZZ}}^{(j)}\boldsymbol{R}^{-1}\right) - 2\left(\widehat{\boldsymbol{Z}}^{(j)}\right)'\boldsymbol{R}^{-1}\boldsymbol{D\mu} + \boldsymbol{\mu'D'R}^{-1}\boldsymbol{D\mu}$.

Les moments $\widehat{\boldsymbol{Z}}^{(j)}$ et $\widehat{\boldsymbol{ZZ}}^{(j)}$ ne sont pas connus complètement. Néanmoins leurs composantes inconnues peuvent être estimées par la loi multivariable normale de $\boldsymbol{Z}$, comme présenté dans la section 3.4.3.

L'algorithme complet est décrit ici :

Algorithme CensSpatial :

— **Initialisation** : Choix de ϵ et choix aléatoire de $\boldsymbol{\theta}^{(0)}$.

— **A l'itération** j : Tant que $|\log(L(\boldsymbol{Z}, \boldsymbol{\theta}^{(j)})) - \log(L(\boldsymbol{Z}, \boldsymbol{\theta}^{(j-1)}))| < \epsilon$,

— Échantillonnage de N réalisations de $\boldsymbol{Z}_c$ depuis la loi normale tronquée :

$$\mathcal{NT}_{n_\nu}(\boldsymbol{D}_c\boldsymbol{\mu} - \boldsymbol{R}_{co}\boldsymbol{R}_{oo}^{-1}(\boldsymbol{z}_o - \boldsymbol{D}_o\boldsymbol{\mu}), \boldsymbol{R}_{cc} - \boldsymbol{R}_{co}\boldsymbol{R}_{oo}^{-1}\boldsymbol{R}_{co}'; (A_i)_{1\leq i\leq n_\nu}).$$

Ces imputations des données censurées sont indexées par $k \in [\![1, N]\!]$ et sont regroupées avec les observations pour former les vecteurs $\boldsymbol{Z}^{(j,k)}$.

— Approximation stochastique des moments : $\widehat{\boldsymbol{Z}}^{(j)}$ et $\widehat{\boldsymbol{ZZ}}^{(j)}$ sont ensuite approchés par :

$$\begin{cases} \widehat{\boldsymbol{Z}}^{(j)} = \widehat{\boldsymbol{Z}}^{(j-1)} + \alpha_j \left[\frac{1}{N}\sum_{k=1}^{N} \boldsymbol{Z}^{(j,k)} - \widehat{\boldsymbol{Z}}^{(j-1)}\right], \\ \widehat{\boldsymbol{ZZ}}^{(j)} = \widehat{\boldsymbol{ZZ}}^{(j-1)} + \alpha_j \left[\frac{1}{N}\sum_{k=1}^{N} \boldsymbol{Z}^{(j,k)}\boldsymbol{Z}^{(j,k)'} - \widehat{\boldsymbol{ZZ}}^{(j-1)}\right]. \end{cases}$$

où α_j est un paramètre de mémoire qui sera détaillé plus loin.

— Estimation des paramètres par maximum de vraisemblance à l'aide des formules suivantes :

$$\begin{cases} \boldsymbol{\mu}^{(j+1)} = \left(\boldsymbol{D}'\left(\boldsymbol{R}^{(j)}\right)^{-1}\boldsymbol{D}\right)^{-1}\boldsymbol{D}'\left(\boldsymbol{R}^{(j)}\right)^{-1}\widehat{\boldsymbol{Z}}^{(j)} \\ \sigma^{2(j+1)} = \frac{1}{n}\left(K^{(j)}\right) \\ (\phi, \eta^2)^{(j+1)} = \underset{(\phi,\eta^2)\in\mathbb{R}^{+*}\times\mathbb{R}^{+}}{\operatorname{argmax}} \left(-\left(\log\left(\det\left(\boldsymbol{R}^{(j)}\right)\right) + K^{(j)}\right)\right) \end{cases}$$

$$\text{où } K^{(j)} = tr\left(\left(\boldsymbol{R}^{(j)}\right)^{-1}\widehat{\boldsymbol{ZZ}}^{(j)}\right) - 2\left(\widehat{\boldsymbol{Z}}^{(j)}\right)'\left(\boldsymbol{R}^{(j)}\right)^{-1}\boldsymbol{D}\boldsymbol{\mu}^{(j+1)} + \boldsymbol{\mu}^{(j+1)'}\boldsymbol{D}'\left(\boldsymbol{R}^{(j)}\right)^{-1}\boldsymbol{D}\boldsymbol{\mu}^{(j+1)}$$

— **Prédictions** : Utilisation du krigeage simple avec les données comprenant les observations et les valeurs imputées aux données censurées ainsi que les paramètres estimés à la dernière itération.

Le choix du paramètre de mémoire α_j modifie le comportement de l'algorithme. Si pour tout j, α_j est proche de 1, l'algorithme aura peu de mémoire et convergera rapidement vers un maximum local, tandis que si α_j est proche de 0, l'algorithme convergera plus lentement mais vers un maximum global. [Ordoñez et al., 2018] détaille le choix de valeurs pour ce paramètre et les propriétés théoriques qu'il doit vérifier.

Selon le choix des α_j, cet algorithme peut avoir l'avantage de converger rapidement, mais pose des difficultés notamment lors de la spécification de certains paramètres, comme les paramètres de mémoire α_j ou encore le choix du critère d'arrêt ϵ.

3.4.5 Augmentation de données en géostatistique

3.4.5.1 Sans effet de pépite

A l'aide d'une approche bayésienne et en utilisant un algorithme MCMC, [De Oliveira, 2005] intègre le paramètre de portée. Cette version est similaire à l'algorithme MCMC présenté dans la section 2.8.2.1, mais ajoute une étape de prédiction des données censurées dans l'échantillonneur de Gibbs (et ne prend pas en compte l'effet de pépite).

Augmentation de données spatiales (sans effet de pépite) :

— **Initialisation** : Choix aléatoire de $\boldsymbol{\mu}^{(0)}, \sigma^{2(0)}, \phi^{(0)}$. Les données censurées sont fixées à la limite de détection : $\boldsymbol{z}_c^{(0)} = \boldsymbol{\nu}$ avec $\boldsymbol{\nu}$ le vecteur des limites de détection.

— **A l'itération** $j \in [\![1, M]\!]$:

— Échantillonnage de $\boldsymbol{Z}_c$ depuis :

$$\mathcal{NT}_{n_\nu}(\boldsymbol{D}_c\boldsymbol{\mu} - \boldsymbol{R}_{co}\boldsymbol{R}_{oo}^{-1}(\boldsymbol{z}_o - \boldsymbol{D}_o\boldsymbol{\mu}), \boldsymbol{R}_{cc} - \boldsymbol{R}_{co}\boldsymbol{R}_{oo}^{-1}\boldsymbol{R}_{co}'; (A_i)_{1\leq i\leq n_\nu}).$$

— Échantillonnage de $\boldsymbol{\mu}^{(j+1)}$ depuis $\mathcal{N}_p(\boldsymbol{m}_1, \boldsymbol{R}_1)$ avec :

$$\begin{cases} \boldsymbol{m}_1 = \left(\boldsymbol{R}_0^{-1} + \frac{1}{\sigma^{2(j)}}\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}(\phi^{(j)})\boldsymbol{D}'\right)^{-1}\left(\boldsymbol{R}_0^{-1}\boldsymbol{m_0} + \frac{1}{\sigma^{2(j)}}\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}(\phi^{(j)})\boldsymbol{z}\right) \\ \boldsymbol{R}_1 = \left(\boldsymbol{R}_0^{-1} + \frac{1}{\sigma^{2(j)}}\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}(\phi^{(j)})\boldsymbol{D}\right)^{-1}. \end{cases}$$

— Échantillonnage de $\sigma^{2(j+1)}$ depuis :

$$\mathcal{IG}\left(\frac{n}{2} + a_\sigma, \frac{1}{2}\left((\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\mu}^{(j+1)})'\boldsymbol{\Sigma}^{-1}(\phi)(\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\mu}^{(j+1)}) + S_\sigma^2\right)\right),$$

— Échantillonnage de $\phi^{(j+1)}$ à l'aide d'un algorithme Metropolis-Hastings avec la distribution :

$$f_\phi(\phi, \tau^2 | \boldsymbol{\mu}^{(j+1)}, \sigma^{2(j+1)}, \boldsymbol{z}) \propto \pi(\phi)\exp(-K)$$

où $K = \frac{1}{\sigma^{2(j+1)}}(\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\mu}^{(j+1)})'\boldsymbol{\Sigma}^{-1}(\phi)(\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\mu}^{(j+1)})$

— Prédiction au point d'intérêt par krigeage simple avec les paramètres $\boldsymbol{\mu}^{(j+1)}$, $\sigma^{2(j+1)}$, et $\phi^{(j+1)}$. Tirage aléatoire depuis la loi obtenue par prédiction de $z(\boldsymbol{x}_0)^{(j+1)}$.

— **Approximation de la loi** : les m premières valeurs sont supprimées, puis une estimation de la loi prédictive est obtenue à partir des $M - m$ échantillons.

Dans la suite nous nous intéressons à prendre en compte un effet de pépite. Pour cela nous proposons un nouvel algorithme inspiré de celui de [De Oliveira, 2005] ajoutant une étape d'échantillonnage de l'effet de pépite dans l'étape de Metropolis-Hastings.

3.4.5.2 Augmentation de données avec effet de pépite

L'algorithme d'augmentation de données est donc similaire à l'algorithme précédent, et combine une étape d'imputation des données censurées avec l'algorithme que nous avons proposé dans la section 2.8.2.1.

Augmentation de données spatiales :

— **Initialisation** : Choix aléatoire de $\mu^{(0)}, \sigma^{2(0)}, \phi^{(0)}$ et $\tau^{2(0)}$. Les données censurées sont fixées à la limite de détection : $\boldsymbol{z}_c^{(0)} = \boldsymbol{\nu}$ avec $\boldsymbol{\nu}$ le vecteur des limites de détection.

— **A l'itération** $j \in [\![1, M]\!]$:

— Échantillonnage de $\boldsymbol{Z}_c$ depuis :

$$\mathcal{NT}_{n_\nu}(\boldsymbol{D}_c\boldsymbol{\mu} - \boldsymbol{R}_{co}\boldsymbol{R}_{oo}^{-1}(\boldsymbol{z}_o - \boldsymbol{D}_o\boldsymbol{\mu}), \boldsymbol{R}_{cc} - \boldsymbol{R}_{co}\boldsymbol{R}_{oo}^{-1}\boldsymbol{R}'_{co}; (A_i)_{1\leq i\leq n_\nu}).$$

— Échantillonnage de $\boldsymbol{\mu}^{(j+1)}$ depuis $\mathcal{N}_p(\boldsymbol{m}_1, \boldsymbol{R}_1)$ avec :

$$\begin{cases} \boldsymbol{m}_1 = \boldsymbol{R}_1\left(\boldsymbol{R}_0^{-1}\boldsymbol{m}_0 + \frac{1}{\sigma^{2(j)}}\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}(\phi^{(j)}, \tau^{2(j)})\boldsymbol{z}\right) \\ \boldsymbol{R}_1 = \left(\boldsymbol{R}_0^{-1} + \frac{1}{\sigma^{2(j)}}\boldsymbol{D}'\boldsymbol{\Sigma}^{-1}(\phi^{(j)}, \tau^{2(j)})\boldsymbol{D}\right)^{-1}. \end{cases}$$

— Échantillonnage de $\sigma^{2(j+1)}$ depuis :

$$\mathcal{IG}\left(\frac{n}{2} + a_\sigma, \frac{1}{2}\left((\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\mu}^{(j+1)})'\boldsymbol{\Sigma}^{-1}(\phi^{(j)}, \tau^{2(j)})(\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\mu}^{(j+1)}) + S_\sigma^2\right)\right),$$

— Échantillonnage de $\phi^{(j+1)}$ et $\tau^{(j+1)}$ à l'aide d'un algorithme Metropolis-Hastings avec la distribution :

$$f_{\phi,\tau^2}(\phi, \tau^2|\boldsymbol{\mu}^{(j+1)}, \sigma^{2(j+1)}, \boldsymbol{z}) \propto \pi(\phi)\pi(\tau^2)\exp(-K)$$

où $K = \frac{1}{\sigma^{2(j+1)}}(\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\mu}^{(j+1)})'\boldsymbol{\Sigma}^{-1}(\phi^{(j)}, \tau^{2(j)})(\boldsymbol{z} - \boldsymbol{D}\boldsymbol{\mu}^{(j+1)})$

— Prédiction au point d'intérêt par krigeage simple avec les paramètres $\boldsymbol{\mu}^{(j+1)}$, $\tau^{2(j+1)}$, $\sigma^{2(j+1)}$, et $\phi^{(j+1)}$. Tirage aléatoire depuis la loi obtenue par prédiction de $z(\boldsymbol{x}_0)^{(j+1)}$.

— **Approximation de la loi** : les m premières valeurs sont supprimées, puis une estimation de la loi prédictive est obtenue à partir des $M - m$ échantillons.

3.4.6 Choix des lois *a priori*

Ici se pose à nouveau le choix de la loi *a priori*. La spécification de cette loi ou des lois possibles est identique aux recommandations données dans la section 2.9.1, à une exception près. En effet, [De Oliveira, 2005] indique que l'utilisation de lois *a priori* impropres avec des données censurées peut conduire à des lois *a posteriori* impropres. Il recommande donc

de n'utiliser que des lois propres mais éventuellement vagues pour les différents paramètres (comme une loi inverse-gamma de variance infinie).

Néanmoins après plusieurs tests réalisés (sur des jeux de données simulés) pour la spécification des différents paramètres pour l'augmentation de données, il apparaît qu'un choix de variance infinie conduit de manière quasi-systématique à de très mauvaises estimations des paramètres σ^2, ϕ et τ^2. En effet la chaîne de l'algorithme MCMC tend alors à évoluer vers des valeurs extrêmes pour un paramètre, tandis qu'un autre paramètre évolue vers l'extrême opposé. Cela conduit à des variances de prédiction très mal estimées, et lorsque le paramètre de portée est touché, à d'éventuelles mauvaises espérances de prédiction. En conséquence, les recommandations de la section 2.9.1 sont légèrement modifiées de telle sorte que les variances des lois *a priori* soient grandes mais définies. Les paramètres a_σ, a_ϕ et a_τ sont ici fixés à 2.1, contrairement à 2 comme expliqué dans la section 2.9.1. Cet ajout de 0.1 permet alors de rendre la variance finie sans pour autant considérer une variance faible. Ce choix de modélisation est également réalisé dans plusieurs articles de la littérature, comme dans [Toscas, 2010] et [Fridley and Dixon, 2007].

3.5 Comparaison des différentes méthodes

On souhaite maintenant comparer les différentes méthodes obtenues afin de choisir celle qui semble le mieux répondre aux problématiques de l'A&D, tout en mettant en évidence leurs éventuels avantages et inconvénients.

3.5.1 Application à un jeu de données simulé

3.5.1.1 Construction du jeu de données et protocole

De manière similaire aux simulations réalisées dans le chapitre précédent pour la comparaison du krigeage bayésien avec le krigeage ordinaire, on simule 100 jeux de données dans l'espace $[0, 10]^2$ sur une grille régulière avec 225 points.

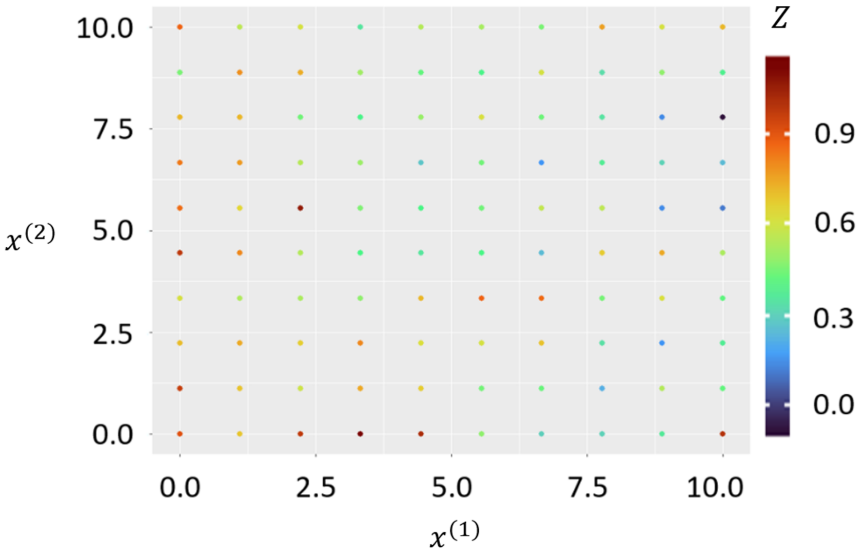


FIGURE 3.6 – Carte des observations simulées.

Cependant la validation croisée utilisée dans le chapitre précédent n'est plus directement applicable pour juger de la qualité des prédictions. En effet si les données sont censurées, il est difficile mais envisageable en pratique de les comparer à une prédiction étant donné que l'observation est inconnue. Pour construire notre protocole nous nous inspirons des applications réalisées dans [Fridley and Dixon, 2007, Toscas, 2010]. Une donnée sur deux est utilisée pour servir de cible de prédictions (base de test), tandis que le reste est utilisé pour conditionner le modèle et estimer les paramètres (base d'apprentissage). La forme du plan d'échantillonnage est représentée sur la Figure 3.6. Les paramètres utilisés pour la simulation sont les suivants :

$$\mu = 3, \sigma^2 = 2, \phi = 2.5, \tau^2 = 0.2.$$

Une fois ces jeux de données obtenus, un taux fixe de $0.1, 0.2$ et 0.4 de données censurées est généré. Pour cela les quantiles empiriques q d'ordre $0.1, 0.2$ et 0.4 sont estimés, puis pour chacun des quantiles q obtenus, toutes les données inférieures ou égales à q sont censurées. Le jeu de données ainsi construit $\boldsymbol{z}_{simu}$ est défini par :

$$\forall i \in [\![1, n]\!], z_{simu}(\boldsymbol{x}_i) = \max(z(\boldsymbol{x}_i), q).$$

où les $z(\boldsymbol{x}_i)$ sont les données obtenues après la simulation initiale.

L'idée derrière cette méthodologie est de reproduire un jeu de données qui pourrait être obtenu dans un projet d'A&D. Si l'on suppose que tous les résultats de mesure ont été obtenus selon le même protocole de mesure, alors le SD est la même pour toutes les observations. Ce quantile vient remplacer le SD et censure toutes les données qui y sont inférieures. L'ordre du quantile est ensuite modifié pour obtenir différentes taux de censure. Dans la suite on a donc $SD = q$.

Après ces censures réalisées, la base d'apprentissage est utilisée pour estimer les paramètres, puis la base de test est prédite à l'aide des paramètres estimés et de la base d'apprentissage.

Comme expliqué dans les paragraphes précédents, la validation croisée n'est pas applicable ici. Les critères de validation présentés dans la section 1.9.6 sont donc adaptés pour des données prédites différentes du jeu de données initial. En pratique les formules restent identiques, les prédictions étant comparées à la base de test plutôt qu'aux observations.

3.5.1.2 Comparaison des résultats et interprétations

Les résultats sont présentés ici sous la forme de boxplots et répartis selon le taux de données censurées. Pour 225 données (soit 113 observations pour la base d'apprentissage et 112 observations pour la base de test), on obtient la Figure 3.7.

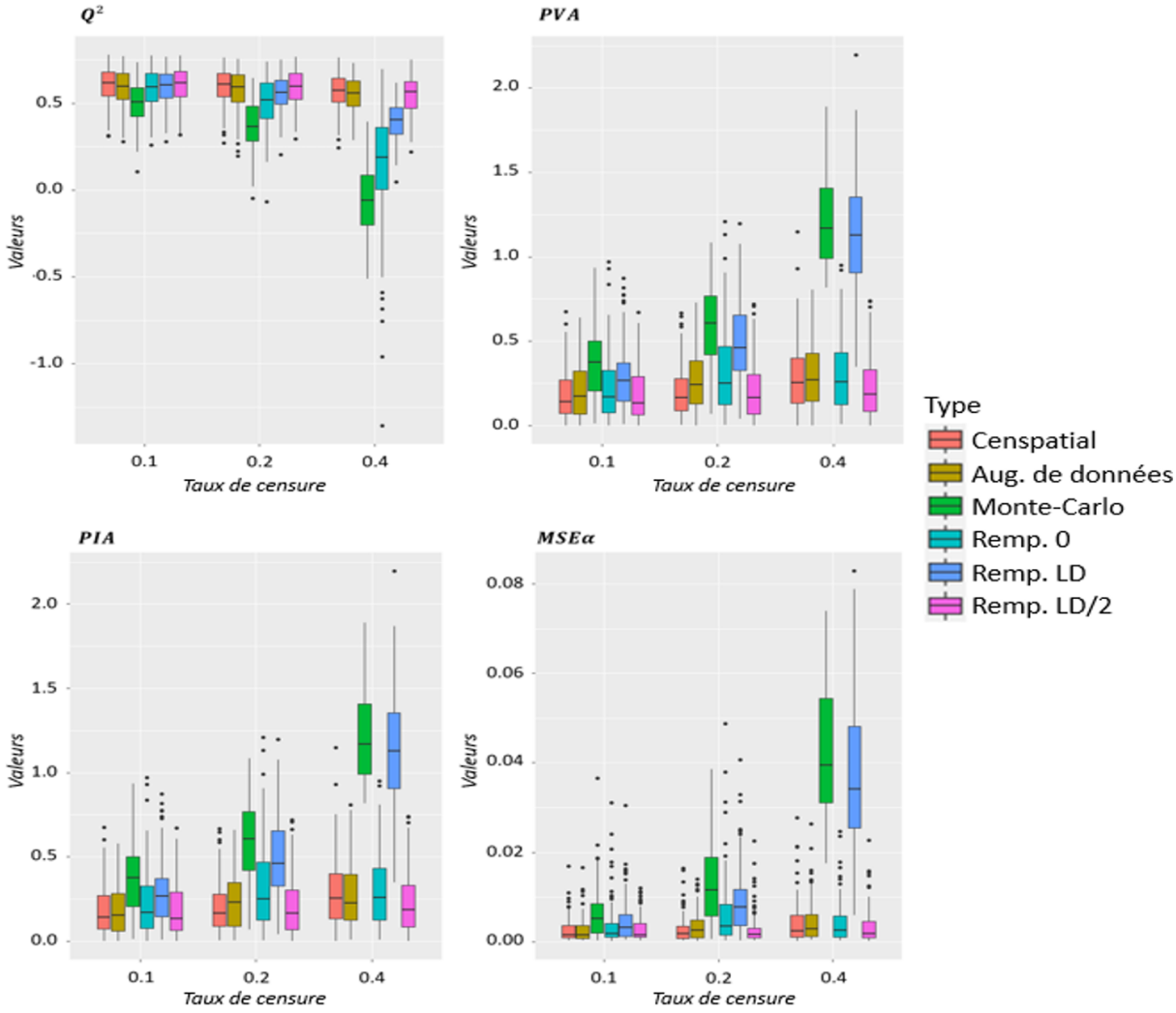


FIGURE 3.7 – Distribution des critères de validation (Q^2, PVA, PIA et $MSE\alpha$) contre le taux de données censurées pour différentes approches de traitement des données censurées et 225 données.

Commençons par la méthode Monte Carlo qui apparaît immédiatement comme une très mauvaise méthode. En effet ses performances sont extrêmement mauvaises, et empirent rapidement avec le taux de censure. Plusieurs éléments permettent d'expliquer ces résultats. Tout d'abord la méthode est assez simple et ne vient jamais utiliser les données censurées dans les prédictions, ce que toutes les autres méthodes font. De plus l'estimation initiale des paramètres servant à construire les imputations des données censurées se fait uniquement à partir des observations. Il semblerait donc que ce ré-échantillonnage soit ici insuffisant pour rendre compte de la variabilité réelle introduite par les données censurées. Plusieurs améliorations de l'algorithme sont possibles, comme l'utilisation des imputations des données censurées dans les prédictions, mais nous choisissons de mettre de côté cette méthode pour nous concentrer sur l'augmentation de données.

Pour le reste des méthodes, commençons par comparer les Q^2. On remarque que pour un taux de censure faible (ici 0.1), les méthodes sont toutes équivalentes, avec des valeurs médianes autour de 0.60. Cependant les méthodes de remplacement perdent rapidement

en efficacité avec le taux de censure, à l'exception de la méthode de remplacement par $LD/2$ qui reste stable. La méthode de remplacement par 0 évolue d'une valeur médiane de 0.59 pour un taux de censure à 0.1 à une valeur de 0.19 pour un taux de censure à 0.4. Celle de remplacement par LD évolue de 0.60 à 0.40. Les méthodes d'augmentation de données, CensSpatial et de remplacement par $LD/2$ restent quant à elle stables en terme de médiane. L'augmentation de données évolue de 0.60 à 0.56, CensSpatial évolue de 0.62 à 0.57 et la méthode de remplacement par $LD/2$ de 0.62 à 0.56. Les différences entre ces méthodes ne paraissent pas significatives, et elles fournissent des performances similaires en termes de prédictivité.

En ce qui concerne le PVA et le PIA, on retrouve des résultats similaires au Q^2, avec l'augmentation de données, CensSpatial et le remplacement par $LD/2$ qui restent stables pour différents taux de censure, donnant des médianes de 0.17 à 0.27, de 0.14 à 0.25, et de 0.13 à 0.18 (et de 0.15 à 0.22 pour le PIA de l'augmentation de données). La méthode de remplacement par 0 se révèle également performante avec une évolution des médianes de 0.17 à 0.26. Enfin la méthode de remplacement par LD donne de très mauvaises performances au sens du PVA et du PIA, passant de valeurs médianes de 0.27 à 1.13.

Enfin le $MSE\alpha$ fournit des résultats similaires au PVA et PIA, avec l'augmentation de données, CensSpatial, les méthodes de remplacement par $LD/2$ et par 0 fournissant d'excellents résultats. Leurs évolutions respectives sont de 0.0014 à 0.0028, de 0.0015 à 0.0025, de 0.0015 à 0.0018 et de 0.0018 à 0.0026. A nouveau, la méthode de remplacement par LD se montre bien inférieure en qualité en passant de 0.0031 à 0.0341.

Comme attendu, les différents critères ont tendance à empirer avec l'augmentation du taux de censure des données. Lorsque le taux de censure est faible, les méthodes fournissent des résultats relativement proches, puisque les données censurées sont trop peu nombreuses pour créer une différence importante. Ensuite, lorsque ce taux augmente, la plupart des méthodes fournissent de moins bonnes performances. Un point intéressant observé ici est que la méthode de remplacement par $LD/2$ semble, contrairement à ce qui est expliqué dans la section 3.4.1, donner de bonnes performances. Ces performances vont dépendre du jeux de données, et peuvent fournir de très mauvais résultats selon la méthode de construction des censures. Cette affirmation peut être vérifiée lorsque les données sont censurées selon le protocole présenté dans la section 3.2 et est illustrée dans la section 3.5.4.3.

Pour 49 données (soit 25 observations pour la base d'apprentissage et 24 observations pour la base de test), on obtient la Figure 3.8.

A nouveau les performances de la méthode Monte Carlo et de la méthode de remplacement par LD se révèlent être très mauvaises. Nous nous intéressons donc dans la suite davantage à la méthode d'augmentation de données, CensSpatial et aux méthodes de remplacement par $LD/2$ et 0.

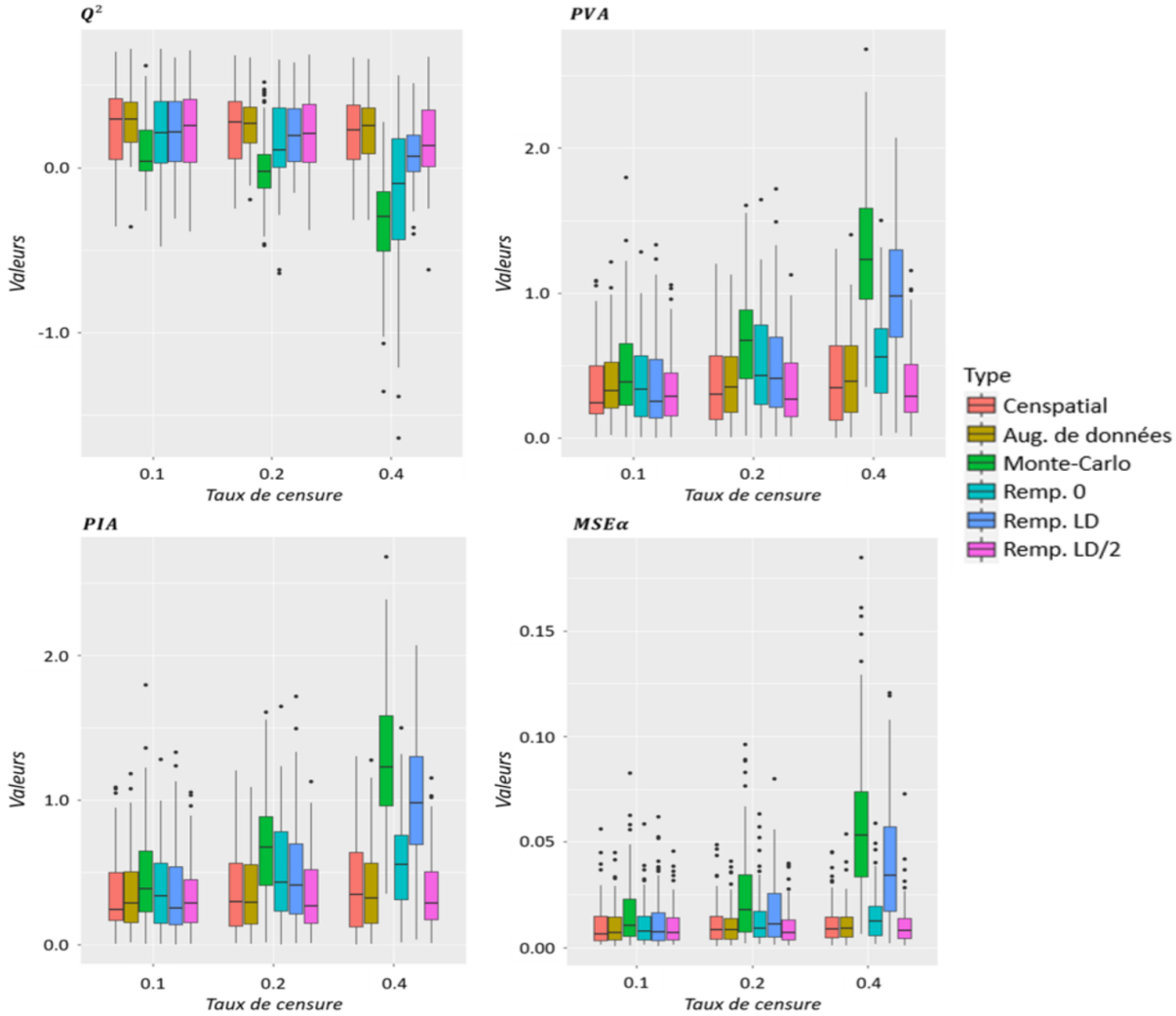


FIGURE 3.8 – Distribution des critères de validation (Q^2, PVA, PIA et $MSE\alpha$) contre le taux de données censurées pour différentes approches de traitement des données censurées et 49 données.

Pour le Q^2, les méthodes d'augmentation de données, CensSpatial et les méthodes de remplacement par $LD/2$ et 0 donnent des valeurs médianes qui évoluent respectivement de 0.29 vers 0.25, de 0.29 vers 0.23, de 0.25 vers 0.13 et enfin de 0.21 vers -0.01. Les largeurs des boxplots pour différents taux de censure sont également proches. On peut également remarquer que les valeurs obtenues pour le Q^2 sont en moyenne moins bonnes comparées aux valeurs obtenues avec 225 données, conséquence logique de la diminution du nombre de données.

Pour le PVA, les méthodes d'augmentation de données, CensSpatial et les méthodes de remplacement par $LD/2$ et 0 donnent respectivement les valeurs médianes suivantes : de 0.33 vers 0.39, de 0.24 vers 0.34, de 0.28 vers 0.29 et enfin de 0.34 vers 0.56. Pour le PIA de l'augmentation de données, on obtient une médiane de 0.29 augmentant vers 0.32. Ainsi on observe bien une baisse dans les performances des différentes méthodes avec l'augmentation du taux de censure. Une autre observation importante est ici que le PVA de CensSpatial semble être meilleur que celui de l'augmentation de données et ce

pour tous les taux de censure. Cependant ce résultat est à relativiser avec le PIA où les différences ne sont plus aussi significatives, notamment avec de meilleures performances pour l'augmentation de données avec un taux de censure de 0.4.

Enfin concernant le $MSE\alpha$, pour les méthodes d'augmentation de données, CensSpatial et les méthodes de remplacement par $LD/2$ et 0, on obtient respectivement des évolutions de valeurs médianes de 0.0070 vers 0.0090, de 0.0065 vers 0.0089, de 0.0070 vers 0.0082 et de 0.0076 vers 0.0124. On obtient donc à nouveau des évolutions similaires à celles observées pour le PVA et PIA.

Le même protocole a été réalisé pour 81 données (soit 41 observations pour la base d'apprentissage et 40 observations pour la base de test), et les résultats sont présentés en annexe, voir A.2.

En conclusion, les méthodes d'augmentation de données et CensSpatial donnent des résultats quasiment identiques, même si les Q^2 obtenus avec l'augmentation de données semblent généralement être inférieurs à ceux obtenus avec CensSpatial. Dans le chapitre précédent, des applications avaient mis en évidence l'avantage de faire appel au krigeage bayésien lorsque peu de données sont disponibles. Il était donc possible d'imaginer un résultat similaire avec l'augmentation de données (approche bayésienne) avec CensSpatial (approche classique). Néanmoins les résultats obtenus pour différentes tailles d'échantillon sont extrêmement similaires entre les deux méthodes. Plusieurs éléments pourraient expliquer cette observation. Tout d'abord nous évoquons quelques problèmes dans la convergence de la chaîne de Markov dans la section 3.5.3, pouvant engendrer de moins bonnes performances pour l'augmentation de données. Ce problème est notamment issu du fait que chaque chaîne des différentes simulations n'a pas été ajustée manuellement (dans la spécification des paramètres de l'algorithme). Ce problème est moins présent dans un cas appliqué puisqu'un seul jeu de données est étudié et que l'ajustement de la chaîne peut alors être réalisé sans problème. Néanmoins l'approche bayésienne semble offrir plus de flexibilité à l'aide de la spécification de loi *a priori* pour intégrer des informations supplémentaires, ce qui laisse penser que le choix de méthode employée dépendra principalement du contexte de l'étude.

3.5.2 Analyse de sensibilité

Une analyse de sensibilité est ici conduite pour tester la sensibilité de notre algorithme à la spécification des lois *a priori*. Nous utilisons ici une approche similaire à celle présentée dans la section 2.9.1. Des données sont simulées selon un processus gaussien avec les paramètres suivants :

$$\mu = 3, \sigma^2 = 2, \phi = 2.5, \tau^2 = 0.2.$$

On simule 100 jeux de données composés de 100 observations, dont 20% des données sont censurées. Comme expliqué dans la section 3.4.6, le choix de lois *a priori* impropres n'est pas considéré ici. De plus une validation croisée étant plus difficile à appliquer avec les données censurées, une approche similaire à la section 3.5 est employée, en séparant le jeu de données en une base d'apprentissage et une base de test. Pour cela une donnée sur deux est prise pour la base d'apprentissage, et le reste constitue la base de test, soient 50 données pour la base de test et 50 données pour la base d'apprentissage.

Pour les différentes lois *a priori*, on considère les 3 choix suivants :
— peu fidèle et juste :

$$\mu \sim \mathcal{N}(3, 0.1), \sigma^2 \sim \mathcal{IG}(2.1, 2), \phi \sim \mathcal{IG}(2.1, 2.5) \text{ et } \tau^2 \sim \mathcal{IG}(2.1, 0.2),$$

— fidèle et peu juste :

$$\mu \sim \mathcal{N}(4, 0.0001), \sigma^2 \sim \mathcal{IG}(20, 3), \phi \sim \mathcal{IG}(20, 4.5) \text{ et } \tau^2 \sim \mathcal{IG}(20, 0.5),$$

— peu fidèle et peu juste :

$$\mu \sim \mathcal{N}(4, 0.1), \sigma^2 \sim \mathcal{IG}(2.1, 3), \phi \sim \mathcal{IG}(2.1, 4.5) \text{ et } \tau^2 \sim \mathcal{IG}(2.1, 0.5).$$

On répète ensuite la même chose pour un jeu initial de 49 données, soit 24 données pour la base d'apprentissage et 25 pour la base de test. Les résultats sont présentés dans la Figure 3.9.

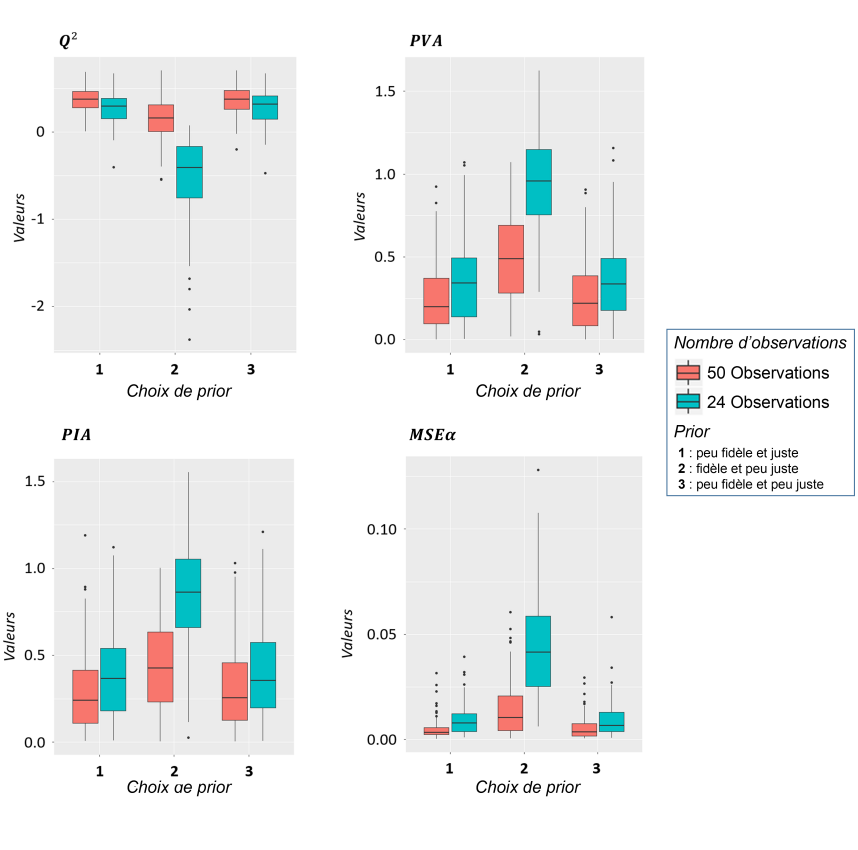


FIGURE 3.9 – Distribution des critères Q^2, PVA, PIA et $MSE\alpha$ pour différents choix de lois *a priori* et différentes tailles d'échantillons.

Les résultats observés ici sont identiques à ceux de la section 2.2. Il semble plus intéressant de préciser des lois *a priori* vagues. Même si ces lois sont mal centrées, la variance élevée de la loi *a priori* vient compenser cette erreur. Le problème vient principalement d'une loi fidèle (ici au sens de variance faible) et peu juste, où les erreurs deviennent importantes et les performances de l'augmentation de données baissent de

manière drastique. On remarque également que plus la taille de l'échantillon est grande, moins les résultats sont sensibles aux choix de lois *a priori*. Enfin le critère Q^2 semble plus sensible à la spécification *a priori*. Comme nous l'avions indiqué dans la section 2.2, le critère Q^2 est principalement sensible au nombre de données. Cependant l'augmentation de données présentée vient imputer des valeurs aux données censurées, qui sont ensuite utilisées pour les prédictions. Ces valeurs imputées étant tirées aléatoirement depuis la loi prédite, elles sont sensibles à l'estimation des paramètres σ^2, ϕ et τ^2. Or ces estimations sont très sensibles aux choix *a priori*, rendant par conséquence les prédictions des méthodes bien plus sensibles aux choix de lois *a priori*.

3.5.3 Problèmes rencontrés avec l'algorithme proposé

Quelques difficultés apparaissent lors de l'utilisation de notre algorithme pour l'augmentation de données. En effet la chaîne MCMC ne semble pas nécessairement converger pour le paramètre de pépite. La Figure 3.10 en donne ici un exemple. Cependant la valeur de l'effet de pépite étant relativement faible comparée à la variance estimée, ce problème n'a pas nécessairement une grande influence dans les prédictions. Au contraire, si l'effet de pépite estimé est grand, cette absence de convergence pourrait poser des problèmes dans l'estimation des variances de prédiction.

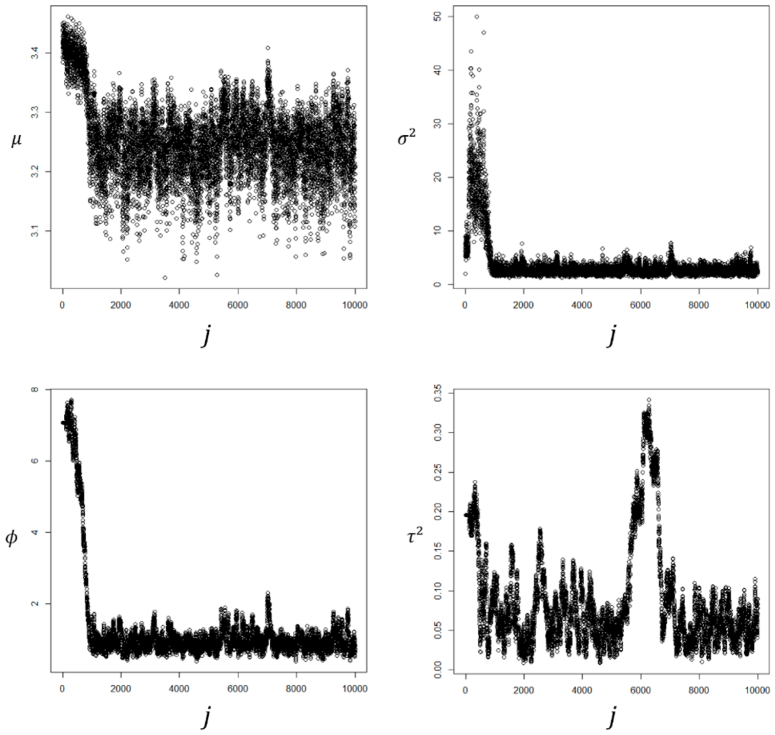


FIGURE 3.10 – Évolution des tirages des différents paramètres lors de la chaîne de Markov.

Ce problème peut être en partie corrigé en précisant une loi *a priori* plus fidèle. Cependant cela implique de posséder des informations supplémentaires sur l'effet de pépite. Sans informations supplémentaires, une spécification plus fidèle peut entraîner un risque d'introduction de biais si les lois *a priori* ne sont pas justes (voir la section précédente 3.5.2).

3.5.4 Application aux données de G3

Les données présentées rapidement dans la section 2.12.1 sont utilisées pour cette application.

3.5.4.1 Analyse exploratoire de G3

Une analyse exploratoire est initialement réalisée pour étudier la présence d'anisotropie, de stationnarité etc. Un histogramme donné Figure 3.11 est d'abord tracé pour évaluer la forme de la distribution.

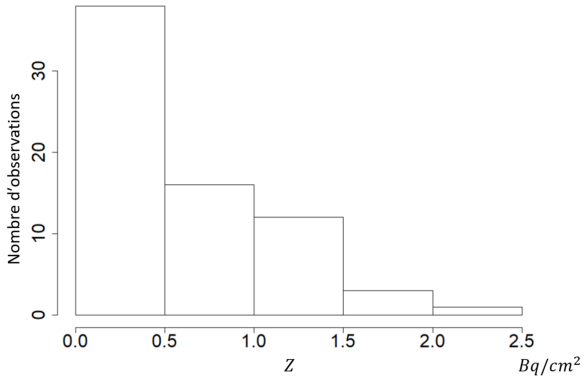


FIGURE 3.11 – Histogramme des données de G3.

La forme de cette distribution montre la présence d'une queue lourde. Pour s'approcher de l'hypothèse gaussienne nécessaire à l'application des méthodes présentées plus haut, une transformation de Box-Cox est appliquée (voir section 1.8.2). Avec $\lambda_{BC} = 0.418$, on obtient l'histogramme de la Figure 3.12.

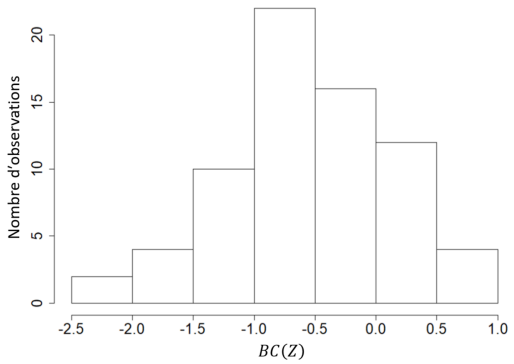


FIGURE 3.12 – Histogramme des données de G3 après transformation.

L'hypothèse gaussienne semble donc être vérifiée en partie dans la forme de l'histogramme. Enfin aucune valeur extrême n'est observée. Il n'est donc pas nécessaire de faire appel à des méthodes permettant de traiter des extrêmes.

La suite de l'analyse exploratoire s'intéresse à l'étude d'une éventuelle anisotropie dans le champ aléatoire. Pour cela des variogrammes empiriques directionnels sont réalisés afin d'étudier l'évolution de la portée et de la variance en fonction de la direction prise par les couples d'observations. Le variogramme omnidirectionnel ainsi que les variogrammes horizontal et vertical (respectivement 0° et 90°) sont représentés sur la Figure 3.13.

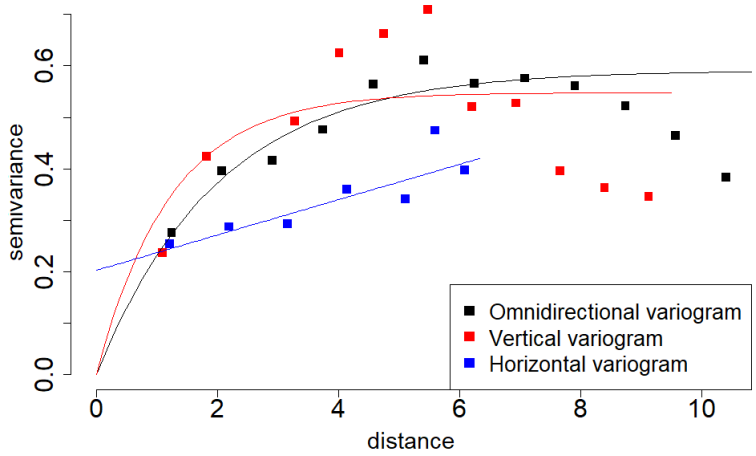


FIGURE 3.13 – Variogrammes empiriques omnidirectionnel et directionnels (horizontal et vertical).

On remarque ici que les variogrammes vertical et omnidirectionnel présentent une variance (respectivement environ 0.55 et 0.59) et une portée très similaires (respectivement environ 0.7 et 1.2). On remarque cependant que l'estimation du variogramme horizontal est très différente des deux autres variogrammes. Cela peut s'expliquer par le nombre de données plus restreint dans la direction horizontale, et donc une estimation moins robuste du variogramme. Il semble tout de même raisonnable de supposer que le champ considéré est isotrope pour notre application.

3.5.4.2 Protocole

Nous appliquons maintenant notre approche aux données de G3 présentées dans la partie 2.12.1. Ici nous comparons les différentes méthodes de traitement des données censurées avec une méthodologie similaire à celle de la section 3.5.1.1.

Pour construire nos jeux de données, nous tirons aléatoirement sans remise la moitié des données (parmi les 70 observations) pour former une base d'apprentissage, tandis que le reste des données sert de base de test. Ce processus est répété 100 fois.

Dans cette partie on cherche à simuler un jeu de données qui pourrait être issu de l'A&D. Ainsi pour construire les censures, le protocole présenté dans la section 3.2 pour la censure résultat de mesure est appliqué. On considère alors que le seuil de décision (SD) est le même pour toutes les données. Ce seuil est fixé ici de telle sorte à obtenir un certain pourcentage de données censurés (en l'estimant à l'aide du quantile empirique approprié). Toute donnée de la base d'apprentissage inférieure à SD est censurée et est est remplacée par la limite de détection (LD), en supposant que $LD = 2SD$ (comme c'est souvent le cas, [Demongeot et al., 2011]). A nouveau des jeux de données comportant $10\%, 20\%$ et 40% de données censurées sont simulés.

Les données sont ensuite transformées à l'aide d'une transformation Box-Cox (et ce pour chaque tirage) pour respecter l'hypothèse gaussienne (voir section 2.8). Pour chaque

jeu de données et pour chaque taux de censure, la base de test correspondante est prédite, et les critères de validation Q^2, PVA, PIA et $MSE\alpha$ (voir section 2.2) sont calculés. Au total 300 jeux de données différents sont construits. Pour les prédictions, les méthodes employées sont : l'augmentation de données, la méthode CensSpatial et la méthode de remplacement par $LD/2$.

3.5.4.3 Résultats et interprétation

Les résultats sont synthétisés dans les différents box-plots de la Figure 3.14.

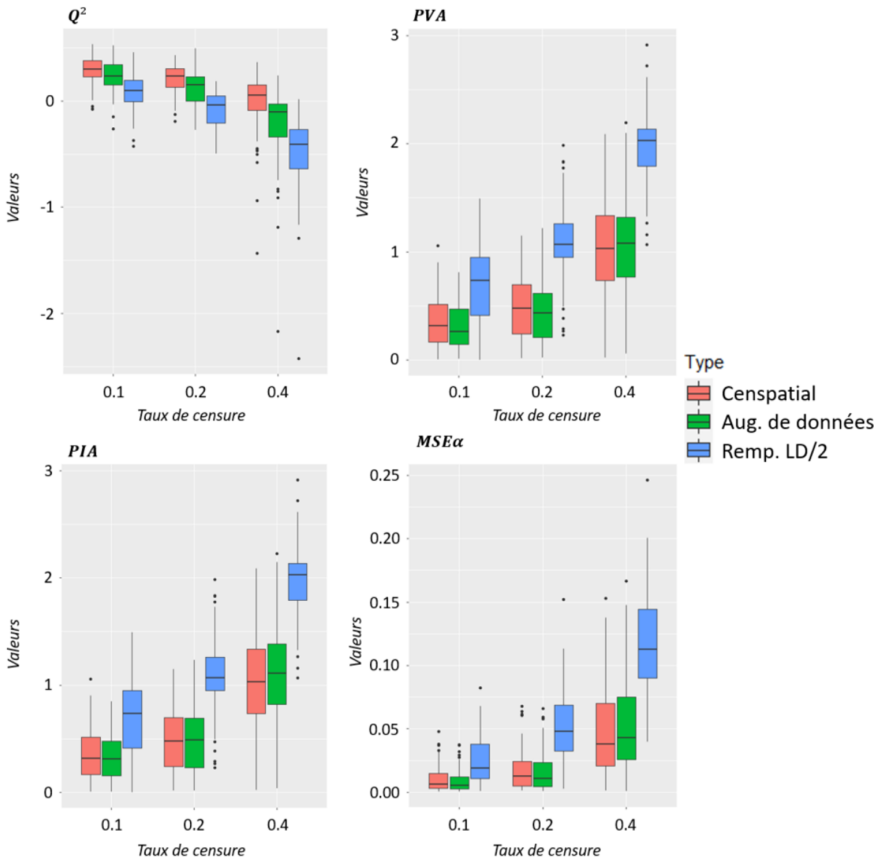


FIGURE 3.14 – Distribution des critères de validation (Q^2, PVA, PIA et $MSE\alpha$) contre le taux de données censurées pour différentes approches de traitement des données censurées et 49 données.

On remarque tout de suite que la méthode de remplacement par $LD/2$ donne de très mauvaises performances dans ce cas, contrairement à ce qui était observé dans la section 3.5.1.2. Remplacer les données censurées par $LD/2$ ne permet donc pas de se rapprocher du résultat de mesure avant censure. Cela s'explique par la construction des données censurées, qui ici ne donne pas l'avantage à cette méthode de remplacement. Cela rejoint donc l'argument évoqué dans la section 3.5.1.2. Dans la suite on se concentrera donc sur les méthodes d'augmentation de données et CensSpatial.

Commençons par comparer les Q^2. Pour l'augmentation de données, la valeur médiane évolue de 0.23 à -0.11, tandis que pour CensSpatial elle évolue de 0.29 à 0.05. On retrouve

des évolutions logiques, notamment avec le critère empirant avec l'augmentation du taux de censure. Cependant, on remarque ici une différence importante avec les résultats de la section 3.5.1.2 : l'augmentation de données fournit de moins bons critères que CensSpatial. Néanmoins les résultats restent très mauvais avec des Q^2 souvent négatifs. Cela paraît cohérent avec la comparaison réalisée en section 3.5.4.4, où les meilleurs Q^2 étaient de l'ordre de 0.35. Ici les jeux construits contiennent moitié moins de données, qui sont de plus partiellement censurées. On obtient donc des modèles prédictifs peu satisfaisants. Les dispersions restent proches entre les différentes méthodes.

Pour les critères PVA et PIA, l'augmentation de données fournit une valeur médiane de 0.26 à 1.08, tandis que CensSpatial donne une valeur médiane de 0.32 à 1.03. Pour le PIA de l'augmentation de données, les valeurs vont de 0.31 à 1.11. Les valeurs obtenues restent stables quel que soit le taux de censure.

Enfin pour le critère $MSE\alpha$, l'augmentation de données fournit une valeur médiane de 0.0055 à 0.0429, tandis que CensSpatial donne une valeur médiane de 0.0062 à 0.0381.

Les résultats obtenus ici sont extrêmement similaires à ceux obtenus précédemment (notamment dans la sections 3.5.1.2). Les performances des différentes méthodes sont inférieures aux jeux de données simulées avec 225 observations, mais supérieures aux performances obtenues avec les jeux simulées de 49 observations. On retrouve donc bien des résultats cohérents. On remarque tout de même que l'augmentation de données semble être légèrement inférieure en terme de performances que la méthode CensSpatial, notamment concernant le critère Q^2.

3.5.4.4 Correction de la loi *a priori*

Les performances obtenues ici sont donc en moyenne moins bonnes que celles obtenues lors des simulations. Il apparaît cependant que le paramètre μ de moyenne était mal estimé après transformation. En effet la moyenne était très largement sur-estimée, et les prédictions étaient donc biaisées. Une analyse de la chaîne de Markov a mis en évidence une loi *a priori* mal spécifiée. Ces résultats peuvent s'expliquer par le faible nombre d'observations, qui rend l'estimation des paramètres par l'augmentation de données très dépendante des lois *a priori*. Ainsi si ces lois sont mal spécifiées, l'estimation des paramètres sera très mauvaise. Or dans le cas des données censurées, l'estimation de la moyenne avec les données observées et les limites de détection est biaisée. La loi *a priori* de μ est donc mal spécifiée, entraînant une mauvaise estimation des paramètres et donc des prédictions.

Une solution considérée ici est de choisir pour le paramètre m_0 (la moyenne de la loi *a priori* de μ) la médiane du jeu de données complet (observations et limites de détection). Cette statistique d'ordre est ici plus robuste aux modifications apportées par les censures, et permet de mieux centrer la loi *a priori* sur μ. Le protocole de la section 3.5.4.2 est appliqué, avec comme seule modification la nouvelle spécification de la loi *a priori* de μ. Les résultats sont présentés dans la Figure 3.15.

L'ensemble des critères obtenus pour l'augmentation de données sont donc améliorés vis à vis des résultats de la section 3.5.4.3, avec des performances obtenues bien plus proches de celles de CensSpatial. Cependant on retrouve pour le critère Q^2 des performances

inférieures à celles de CensSpatial. Il semblerait donc que l'augmentation de données ne prédise pas de manière efficace les espérances de prédiction.

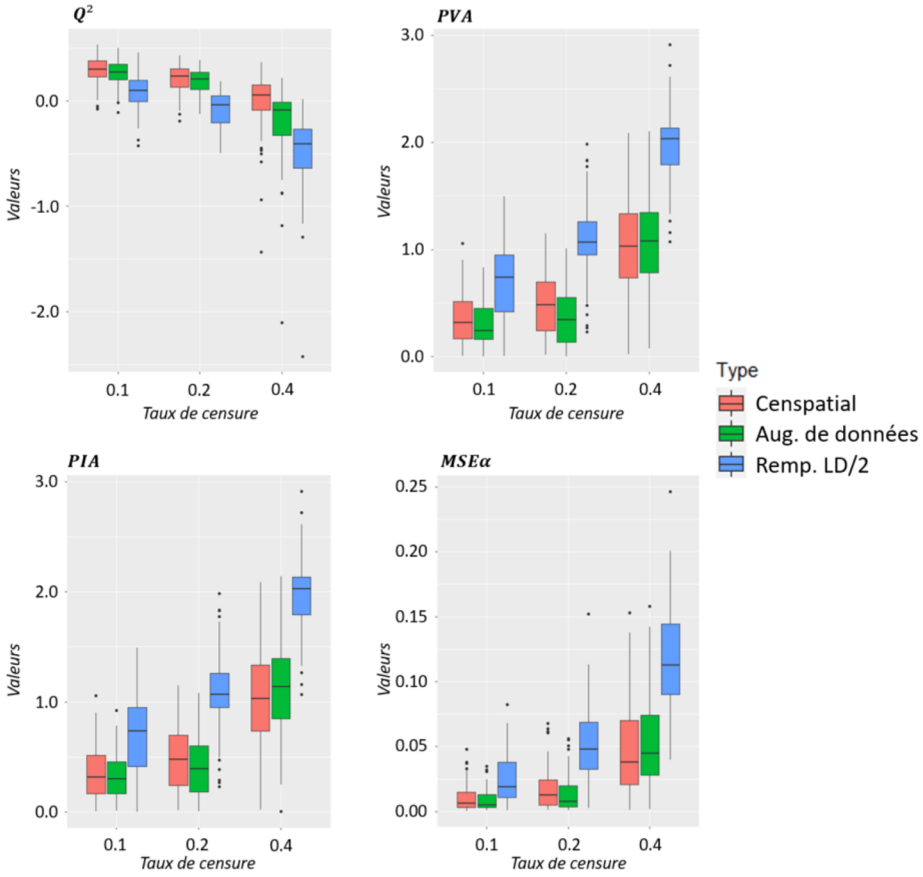


FIGURE 3.15 – Distribution des critères de validation (Q^2, PVA, PIA et $MSE\alpha$) contre le taux de données censurées pour différentes approches de traitement des données censurées, 49 données et avec modification de la loi *a priori* usuelle.

3.6 Conclusion

Les outils que nous proposons dans cette partie apportent des améliorations mais posent également certaines difficultés dans leur application. La méthode par approche Monte Carlo fournit de mauvaises performances, mais pourrait être facilement améliorée. L'augmentation de données nécessite quant à elle de nombreuses spécifications parfois difficiles à fournir.

Elle est de plus très sensible aux lois *a priori*. C'est particulièrement remarquable si l'on compare la sensibilité du krigeage bayésien avec celle de l'augmentation de données. Le critère Q^2 est ici un excellent exemple de ce phénomène, étant insensible au choix de loi *a priori* dans le cas du krigeage bayésien mais sensible dans le cas de l'augmentation de données. En plus de cette sensibilité accrue, les lois *a priori* impropres sont à éviter dans le cas de données censurées. Cela force donc une spécification des lois *a priori* qui est plus difficile à réaliser puisque les prédictions y sont sensibles. Les sections 3.5.4.3 et 3.5.4.4 illustrent également la sensibilité accrue aux spécifications *a priori* de l'augmentation de

données si peu d'observations sont disponibles. Néanmoins les recommandations fournies dans la littérature ainsi que dans ce document permettent d'avoir quelques choix de référence garantissant les performances de l'augmentation de données.

Concernant les méthodes de remplacement, il semble clair qu'elles sont à proscrire. Elle introduisent de manière générale des biais importants dans les estimations et prédictions (voir [Helsel, 2012],[Crozet et al., 2015] et les sections 3.5.1.2 et 3.5.4.3).

Notre version de l'augmentation de données semble prometteuse et fournit de bons résultats comparée aux autres méthodes disponibles dans la littérature. Cependant certains problèmes de convergence semblent apparaître avec l'effet de pépite, et explique l'approche suivi par [Fridley, 2003], qui exclut l'estimation de la pépite de l'algorithme Metropolis-Hastings. Des travaux supplémentaires sont donc encore nécessaires pour justifier son utilisation dans un cadre industriel, malgré des premiers résultats encourageants.

Conclusions et Perspectives

Le développement d'une industrie de démantèlement de sites nucléaires est en cours en France. La construction de nouveaux réacteurs et la fin de service des installations actuelles garantissent un besoin important dans l'expertise du démantèlement de ce type de sites industriels. Cette industrie étant relativement récente et les retours d'expérience étant peu nombreux, le développement de méthodologies efficaces permettrait une optimisation des durées des démantèlements et de leur coût.

Cette thèse se place dans le cadre des opérations de caractérisation du site en amont des étapes de démantèlement. Si les méthodes statistiques usuelles ont déjà fait leurs preuves dans ce domaine, certaines problématiques n'étaient pas traitées et les solutions appropriées n'avaient pas encore été mises en place ou testées. Les résultats de cette thèse justifient l'utilisation d'approches bayésiennes comme le krigeage bayésien ou l'augmentation de données dans le traitement des résultats de mesure, pour garantir une meilleure robustesse des estimations et prédictions réalisées, en particulier pour la répartition spatiale de la contamination.

Les résultats obtenus ont permis de mettre en évidence un nombre d'observations à partir duquel le krigeage bayésien semble être plus performant que l'approche classique, le krigeage ordinaire. Ces résultats ont également été vérifiés pour différentes fonctions de covariance et différents jeux de données. Cependant rien ne garantit la répétabilité de ces résultats à des cas différents de ceux proposés. Ainsi de nouvelles applications à des cas anisotropes, non-gaussiens ou non-stationnaires seraient utiles pour justifier une utilisation systématique de ces méthodes. Par exemple les cas anisotropes, souvent présents dans l'analyse de contaminations pour l'assainissement de sols, justifieraient des études approfondies. Le krigeage bayésien peut en effet prendre en compte des incertitudes dans l'estimation de paramètres d'anisotropie, comme les paramètres d'une transformation géométrique. De plus certains éléments théoriques existent déjà dans la littérature, comme par exemple les travaux de [Muré, 2018] qui vérifient la validité de la loi *a priori* de référence de [Berger et al., 2001] pour les cas anisotropes. Nous avons également rapidement évoqué dans le chapitre 2 le cas hétéroscédastique. Sa modélisation consiste à considérer des incertitudes de mesure différentes en chaque point mesuré. Cela peut se produire si le protocole de mesure évolue, ou si l'incertitude de la méthode de mesure dépend de la valeur de la grandeur mesurée. Pour le krigeage bayésien, ce cas a été étudié dans [Ng and Yin, 2012] et permettrait donc son application.

L'augmentation de données, lorsque des données censurées sont présentes, a également fait ses preuves au travers de nos différentes applications. Sur des jeux de données simulées,

l'augmentation de données conduit à des résultats similaires à ceux des méthodes existantes récentes comme la méthode proposée par [Ordoñez et al., 2018]. Néanmoins les résultats observés sur le jeu de données de G3 sont décevants pour les espérances de prédiction. Des recherches supplémentaires permettraient de mieux comprendre les résultats observés et de valider l'utilisation de cette méthode pour des données de l'A&D. L'application de l'augmentation de données aux cas hétéroscédastiques ou anisotropes semble possible, mais nécessite également des travaux supplémentaires.

Les applications ont mis en évidence quelques difficultés pour la version de l'augmentation de données avec estimation de l'effet de pépite (présentée dans cette thèse). Le premier est une mauvaise estimation de l'effet de pépite, notamment lorsque la chaîne de Markov ne converge pas. Une solution à ce problème est bien sûr de considérer les incertitudes de mesure comme connues (voir [Tadayon, 2017]). Un autre problème rencontré est la sensibilité accrue aux choix de lois *a priori* de l'augmentation de données, elle-même plus difficile puisqu'elle nécessite le choix de lois *a priori* propres. Des recherches plus poussées sur ces problèmes permettraient de mettre au point des recommandations usuelles pour un usage systématique de la méthode.

L'utilisation de l'augmentation de données et du krigeage bayésien peut également être justifiée par la prise en compte des informations supplémentaires à l'aide des lois *a priori*. Si des informations supplémentaires sont disponibles (autres que les résultats de mesure), par exemple l'historique des contaminations accidentelles, elles peuvent être intégrées de manière quantitative dans les approches bayésiennes à l'aide des lois *a priori*. Cependant l'approche usuelle recommandée est d'utiliser des lois impropres ou vagues. Il semble donc souhaitable de réaliser, si possible, une analyse avec des lois contenant ces informations supplémentaires et de comparer les résultats obtenus avec ceux obtenus avec des lois *a priori* impropres ou peu fidèles. Cela permet de confirmer l'utilité de ces informations et de tester les modifications qu'elles apportent aux prédictions. Néanmoins cette question reste une étape difficile des approches bayésiennes, et une analyse de sensibilité aux choix *a priori* reste un standard à appliquer. Ce point est donc un sujet de discussion important et mériterait son propre axe de recherche.

Plusieurs jeux de données ont été testés dans cette thèse. La plupart des résultats a été obtenue à partir de jeux de données simulées et du jeu de données du réacteur G3. Il serait donc intéressant de réaliser davantage d'applications, notamment à d'autres jeux de données issus de l'A&D de sites nucléaires. Ces autres jeux de données permettraient notamment d'obtenir des exemples dont les incertitudes de mesure sont connues, et éviteraient un des problèmes rencontrés dans notre version de l'augmentation de données. Il serait alors intéressant de comparer l'efficacité de l'approche considérant les incertitudes de mesure connues à celle présentée dans cette thèse ré-estimant ces incertitudes.

Ces deux derniers points font ressortir une remarque importante : il est impératif d'avoir accès aux méthodes utilisées par le mesureur pour censurer les données et aux incertitudes de mesures. Dans le cas idéal, une valeur de résultat de mesure d'une méthode validée, même inférieure au seuil de décision, ne devrait jamais être censurée par le mesureur, et devrait être fournie telle qu'obtenue, avec son incertitude associée pouvant être très élevée (100% par exemple). Trop souvent les jeux de données sont recueillis sans les incertitudes

de mesure associées, ou, si les données sont censurées, sans description des choix de censure. Ce manque d'information est particulièrement présent pour la décomposition de l'effet de pépite (voir 1.9.5.6). Selon le choix fait par le statisticien, les variances de prédiction peuvent alors fortement varier. Les critères de validation comme le PVA étant extrêmement sensibles à ces variances, une erreur de modélisation peut également conduire à une mauvaise évaluation de la qualité du modèle. Un travail d'échange avec les experts en métrologie apparaît donc comme primordial pour une amélioration significative de la caractérisation radiologique initiale et du traitement statistique des données collectées.

Enfin une étape classique de la géostatistique évoquée dans la section 1.9.7 n'a pas été étudiée : les simulations conditionnelles. Ces simulations ont pour avantage de permettre de réaliser des analyses de risque à l'aide de réalisations possibles (conditionnellement aux observations) du champ aléatoire. Elles sont souvent réalisées après avoir estimé les paramètres pour le krigeage. Ici elles n'ont pas été traitées, mais mériteraient d'être étudiées plus en détails. Les méthodes comme CensSpatial et l'augmentation de données réalisent déjà partiellement des simulations, en particulier pour les données censurées. De même le krigeage bayésien repose également sur des simulations puisque l'on vient échantillonner la loi prédictive pour différentes réalisations. Il pourrait donc être intéressant d'étendre ces simulations à des points non-observés et de construire des simulations bayésiennes conditionnelles. Cette approche possède quelques similitudes avec le krigeage empirique bayésien [Krivoruchko and Gribov, 2019] qui réalise des simulations non-conditionnelles. Cependant le coût calculatoire d'une telle méthode serait très élevé, et poserait donc des défis intéressants dans sa réalisation.

Table des figures

Liste des tableaux

ANNEXE A

Quelques résultats complémentaires

A.1 Résultats de la comparaison krigeage ordinaire et krigeage bayésien pour des jeux de données simulés

Les résultats de la comparaison entre krigeage bayésien et krigeage ordinaire pour des données simulées avec une fonction de covariance de Matérn$-\frac{3}{2}$ sont présentés dans la figure A.1.

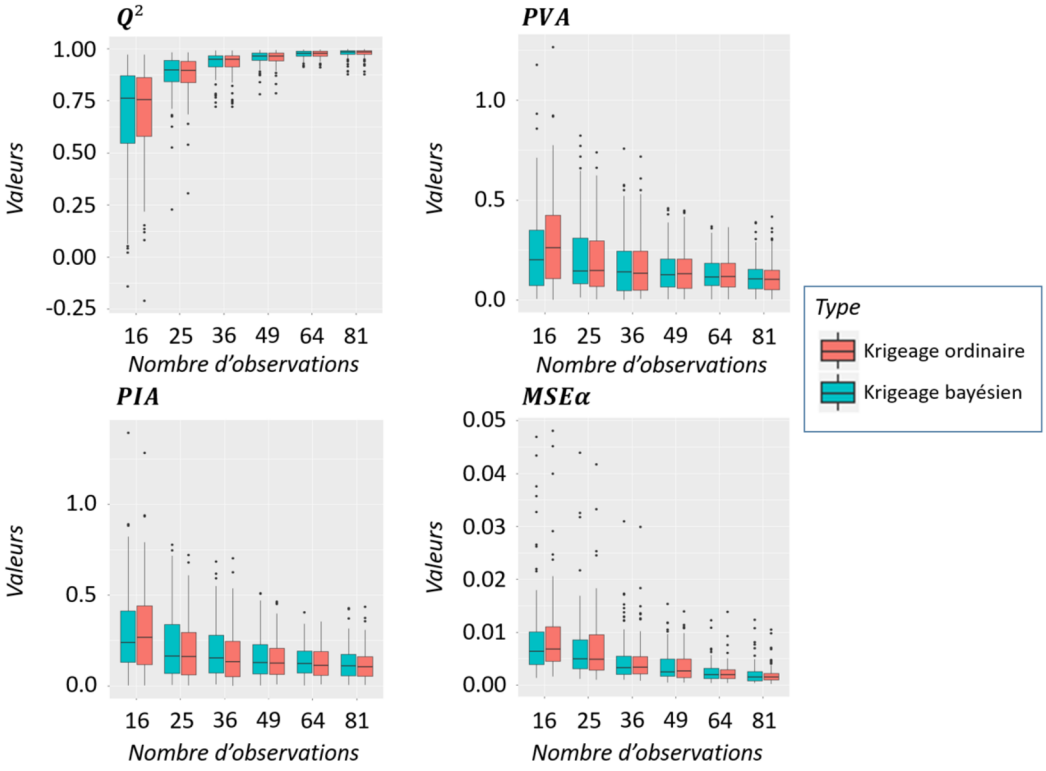


FIGURE A.1 – Distribution des critères de validation (Q^2, PVA, PIA et $MSE\alpha$) contre le taux de données censurées pour différentes approches de traitement des données censurées et 81 données.

Pour des données simulées avec une fonction de covariance de Matérn$-\frac{5}{2}$, les résultats de la comparaison entre krigeage bayésien et krigeage oridnaire sont présentés dans la figure A.2.

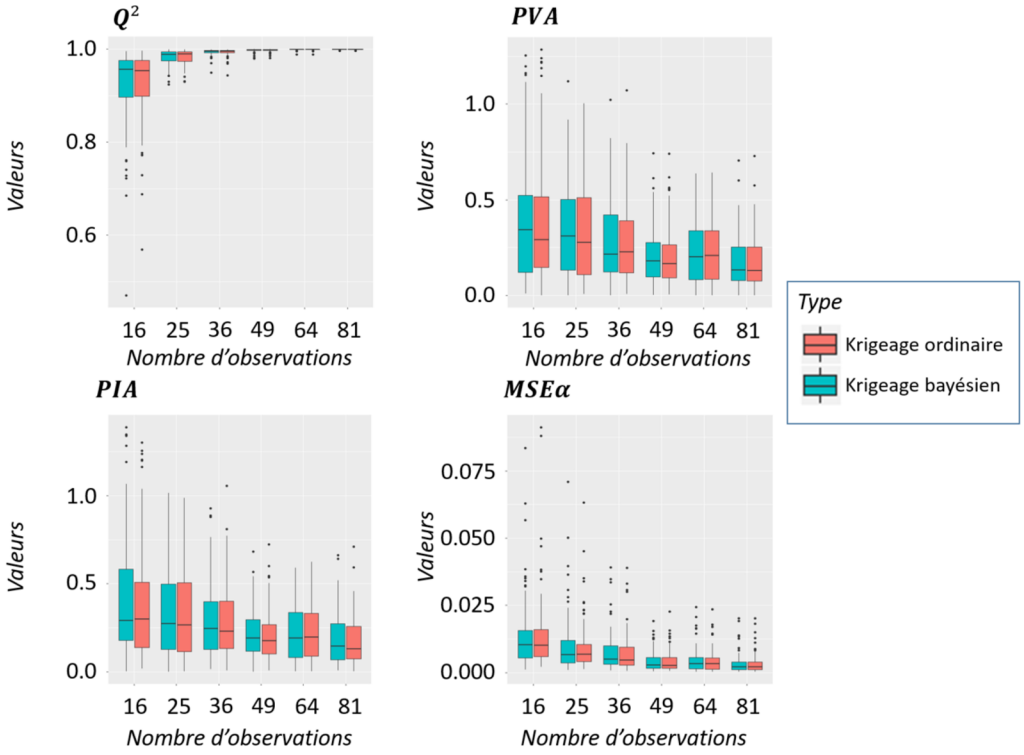


FIGURE A.2 – Distribution des critères de validation (Q^2, PVA, PIA et $MSE\alpha$) contre le taux de données censurées pour différentes approches de traitement des données censurées et 81 données.

A.2 Résultats de la comparaison des méthodes de comparaison des données censurées pour 81 observations

Les résultats de la comparaison de méthodes de traitement des données censurées pour 81 observations, sont donnés dans la Figure A.3.

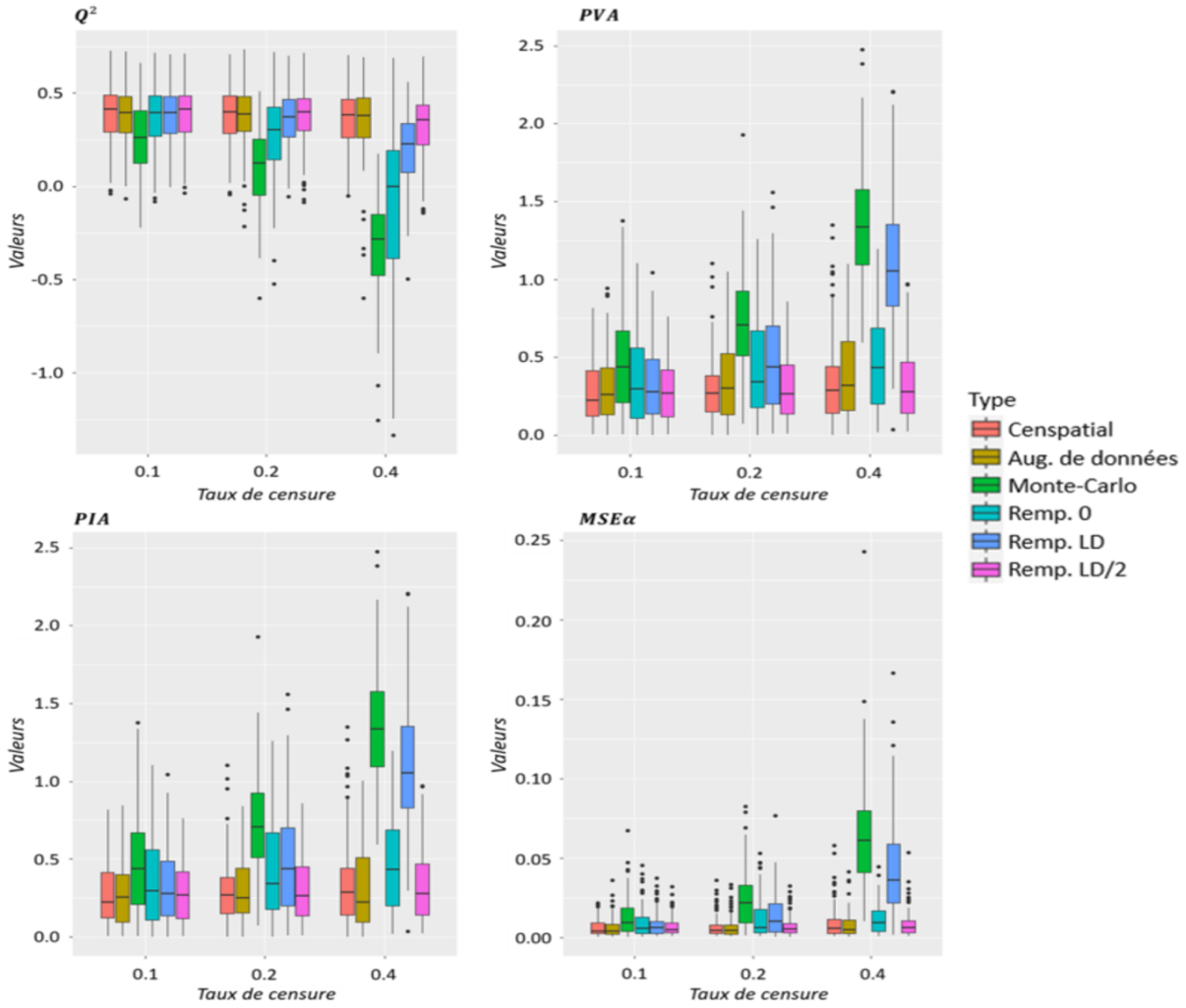


FIGURE A.3 – Distribution des critères de validation (Q^2, PVA, PIA et $MSE\alpha$) contre le taux de données censurées pour différentes approches de traitement des données censurées et 81 données.

ANNEXE B

Publications et communications

B.1 Publication soumise : "A comparison between Bayesian and ordinary kriging based on validation criteria : application to radiological characterisation"

A comparison between Bayesian and ordinary kriging based on validation criteria: application to radiological characterisation

Martin Wieskotten
CEA, DES, ISEC, DMRC, Univ. Montpellier, Marcoule, France
LMA Université d'Avignon, EA 2151, 84029, Avignon, France

Marielle Crozet
CEA, DES, ISEC, DMRC, Univ. Montpellier, Marcoule, France

Bertrand Iooss
EDF R&D, 6 quai Watier, 78400, Chatou, France
Institut de Mathématiques de Toulouse, France

Céline Lacaux
LMA Université d'Avignon, EA 2151, 84029, Avignon, France

Amandine Marrel
CEA, DES, IRESNE, DER, Cadarache, Saint-Paul-Lez-Durance, France
Institut de Mathématiques de Toulouse, France

November 24, 2022

Abstract

In decommissioning projects of nuclear facilities, the radiological characterisation step aims to estimate the quantity and spatial distribution of different radionuclides. To carry out the estimation, measurements are performed on site to obtain preliminary information. The usual industrial practice consists in applying spatial interpolation tools (as the ordinary kriging method) on these data to predict the value of interest for the contamination (radionuclide concentration, radioactivity, etc.) at unobserved positions. This paper questions the ordinary kriging tool on the well-known problem of the overoptimistic prediction variances due to not taking into account uncertainties on the estimation of the kriging parameters (variance and range). To overcome this issue, the practical use of the Bayesian kriging method, where the model parameters are considered as random variables, is deepened. The usefulness of Bayesian

kriging, whilst comparing its performance to that of ordinary kriging, is demonstrated in the small data context (which is often the case in decommissioning projects). This result is obtained via several numerical tests on different toy models, and using complementary validation criteria: the predictivity coefficient (Q^2), the Predictive Variance Adequacy (PVA), the α-Confidence Interval plot (and its associated Mean Squared Error α ($MSE\alpha$)), and the Predictive Interval Adequacy (PIA). The latter is a new criterion adapted to the Bayesian kriging results. Finally, the same comparison is performed on a real dataset coming from the decommissioning project of the CEA Marcoule G3 reactor. It illustrates the practical interest of Bayesian kriging in industrial radiological characterisation.

Keywords: Geostatistics, Bayesian kriging, Ordinary kriging, Validation criteria, Radiological characterisation

1 Introduction

Radiological characterisation is one of the main challenges encountered in the nuclear industry for the decommissioning and dismantling (D&D) of old infrastructures such as buildings (see, e.g., Attiogbe et al. (2014), EPRI (2016) and CEA/DEN (2017)). Its main goal is to evaluate the quantity and spatial distribution of radionuclides. As such, measurements are made to constitute a dataset and obtain preliminary information. While measurements are made, many problems can arise. The radioactivity present on site can be dangerous for operators and does not allow for many measurements. In some extreme cases, drones and robots have to be used, making measurements more expensive and reducing the size of the datasets (see, e.g., Goudeau et al. (2015) and CEA/DEN (2017)). It is therefore quite common in nuclear D&D characterisation to have only a small number of available data: a balance has to be found between data acquisition costs and provided information from data. Statistical tools make it possible to optimise the information extracted from the data, within a rigorous mathematical framework and with associate confidence intervals (in the D&D field, see, e.g., Zaffora et al. (2016), Blatman et al. (2017) and Pérot et al. (2020)).

More precisely, as in many other environmental and industrial fields (see, e.g., Webster and Oliver (2007) and Daya Sagar et al. (2018)), spatial statistics and geostatistical methods are used to predict the variables of interest at an unobserved location (prediction of the expected value), with an indication of the expected error in prediction (prediction variance). The methodology is often based on two steps: first the construction of a statistical model with the estimation of its parameters, followed by the prediction with the statistical model for any unobserved point. The ordinary kriging model (see, e.g., Chilès and Delfiner (2012) and Cressie (1993)) is one of the most widely used models in industrial practice of D&D (see, e.g., Attiogbe et al. (2014), Goudeau et al. (2015) and EPRI (2016)). However, a common criticism is that its predictions do not take into account the uncertainty in the estimation of the model parameters. As a result, the variances of the predictions are often too optimistic and these

neglected uncertainties in the model parameters can have a significant impact. This problem is made worse for smaller datasets, which can be common in D&D projects. For the radiological characterisation in D&D projects, the first examples of kriging shown in Jeannée et al. (2008), Desnoyers (2010) and Desnoyers et al. (2011) have studied practical cases based on many measurements and did not considered this issue. The more realistic studies by Boden et al. (2013), Lajaunie et al. (2020) and Desnoyers et al. (2020), carried out on smaller datasets, have instead highlighted the errors generated by the estimation errors of the kriging parameters.

To overcome this kriging issue, a Bayesian approach was first proposed by Kitanidis (1986). Its main goal was to take into account the uncertainties in the scale and mean parameters of the kriging model. The work of Handcock and Stein (1993) then completed the full Bayesian approach which considers all the parameters of the model as unknown. More recently, a slightly different approach was presented by Krivoruchko and Gribov (2019) and is called empirical Bayesian kriging. While the equations are similar to the ones of regular Bayesian kriging, the choice on the prior distributions of kriging parameters are obtained through unconstrained simulations of the random field. This approach was adapted to allow for multi-fidelity applications, where Bayesian theory is used to update the initial data with new, more accurate data (classically used with cokriging if correlations between old and new data exist). Some examples can be found in meteorology with Gupta et al. (2017) or for oil extraction in Al-Mudhafar (2019). Note that a more complete description of Bayesian kriging with an extension to generalised linear models are presented in Diggle and Ribeiro (2007).

In this framework, our work aims to understand the usefulness of the Bayesian kriging approach, compared to the ordinary kriging one, for the radiological characterisation of contaminated buildings. In particular, the specification of a priori laws for the parameters in Bayesian kriging, which allows a more robust estimation of these parameters when only a few observations are available, is studied. The performance of ordinary and Bayesian kriging is compared on several numerical examples. For this, we not only focus on the kriging predictor accuracy but also on the kriging predictive variance accuracy. Indeed, the kriging variance is often used by practitioners to estimate confidence intervals on predicted quantities, to justify their choice of sampling, or to find locations of new (potentially expensive) measurements (Bechler et al. 2013). To ensure a certain level of confidence in the use of the predictive variance, the works of Marrel et al. (2012), Bachoc (2013a) and Demay et al. (2022), about kriging model validation, have emphasised the usefulness of several validation criteria, as the Predictive Variance Adequacy (PVA) and the α-Confidence Interval (α-CI) plot. In addition to allow a more accurate comparison in the case of the Bayesian kriging model, new validation criteria are required and are proposed in the present work.

The following section describes the different studied kriging models, while Section 3 develops the associate classical validation criteria before introducing the newly proposed ones. Section 4 presents the results of the model comparison

obtained on several numerical tests. Section 5 then illustrates the application on a real case study coming from the decommissioning project of the CEA Marcoule G3 reactor. Section 6 gives some conclusions.

2 The ordinary and Bayesian kriging models

This section provides reminders on kriging principles, within the framework of Gaussian random field model.

2.1 The Gaussian random field model

The variable of interest is assumed to be a random field $\{Z(\boldsymbol{x}), \boldsymbol{x} \in D\}$, with $D \subset \mathbb{R}^2$. $Z(.)$ is supposed to be isotropic and stationary, meaning that

$$\forall \boldsymbol{x} \in D, E[Z(\boldsymbol{x})] = \beta,$$

$$\forall \boldsymbol{x}, \boldsymbol{x'} \in D, \mathrm{Cov}(Z(\boldsymbol{x}), Z(\boldsymbol{x'})) = \sigma^2 C_\phi(|\boldsymbol{x} - \boldsymbol{x'}|),$$

where C_ϕ is the correlation function where $C_\phi(0) = 1$, and β, σ^2, ϕ denote the mean, variance and range (or correlation length) parameters, respectively. For ease of notation, densities will be denoted as $p(.)$ and the conditioning to parameters will be simplified from $Z|\beta = \widehat{\beta}$ to $Z|\beta$. The term C_ϕ corresponds to a semi-definite positive function. Moreover, by definition of a Gaussian process, every finite set of Z is a multivariate normal distribution (denoted $\mathcal{N}(.,.)$). Thus for n observations at positions $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, we obtain the Gaussian random vector $Z = (Z(\boldsymbol{x}_1), \ldots, Z(\boldsymbol{x}_n))'$:

$$Z|\beta, \sigma^2, \phi \sim \mathcal{N}(\beta \mathbf{1}_n, \sigma^2 \boldsymbol{R}_\phi),$$

where $\mathbf{1}_n$ is the vector $(1, \ldots, 1)'$ of length n, and the covariance matrix is $\sigma^2 \boldsymbol{R}_\phi = (\mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)))_{1 \leq i,j \leq n}$. The observation sample of Z is written $\boldsymbol{z} = (z(\boldsymbol{x}_1), \ldots, z(\boldsymbol{x}_n))'$.

The semi-definite positive function C_ϕ is often modeled using common covariance function. In this work, two covariance models will be used (see, e.g., Chilès and Delfiner (2012) for an extensive list of covariance functions). The first one is the Gaussian covariance function written

$$\forall h \in \mathbb{R}, C_\phi(h) = e^{-h^2/\phi^2},$$

while the second one is the Matérn covariance function written

$$\forall h \in \mathbb{R}, C_{\phi,\nu}(h) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{h}{\phi}\right)^\nu K_\nu \left(\sqrt{2\nu}\frac{h}{\phi}\right), \quad (1)$$

with ν a strictly positive parameter, $\Gamma(.)$ the gamma function and $K_\nu(.)$ the modified Bessel function of second type and order ν. The parameter ν, that drives the regularity of the process trajectories, is not estimated. It is chosen

from a set of possible values, the most commonly used being $\nu \in \{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\}$. In addition, we have the nugget effect, written

$$\forall h \in \mathbb{R}, C_{\tau^2}(h) = \tau^2 \delta(h),$$

with τ^2 a variance and δ the Dirac function where $\delta(h) = 1$ if $h = 0$ and $\delta(h) = 0$ otherwise. The nugget effect is often used to model micro-scale variations and measurements uncertainties. In our case studies, it will only be used to improve the conditioning of the matrix $\boldsymbol{R}_\phi$, in order to improve the stability of its numerical inversion (especially in the case of Gaussian covariance function).

The model is therefore specified by three different parameters: the trend parameter $\beta \in D_\beta$, the variance parameter $\sigma^2 \in D_{\sigma^2}$ and the range parameter $\phi \in D_\phi$. In the case of ordinary kriging and for the covariance functions considered here, the parameter spaces are

$$D_\beta = \mathbb{R}, D_{\sigma^2} =]0, +\infty[, D_\phi =]0, +\infty[.$$

The first step of the kriging methodology in practice is to estimate these parameters. Two main procedures are commonly used: variographic analysis and maximum likelihood estimation. An extensive literature is available about parameter estimation with variographic analysis, such as Chilès and Delfiner (2012) and Webster and Oliver (2007). In this work, we will use maximum likelihood estimation to take advantage of the probabilistic framework and to avoid manual or automatic fitting of variograms, especially since our numerical tests will require parameter estimation for many simulated data sets. Moreover, the automatic fitting of variograms is strongly discouraged in most of the literature (see, e.g., Chilès and Delfiner (2012) and Webster and Oliver (2007)). Note that, for the maximum likelihood estimation, an optimization algorithm with a multi-start procedure could be used to avoid the known pitfall of local extrema and better explore the parameter space. However, this procedure will not be used here because preliminary studies have shown that in our case it is not really necessary, probably due to the small dimension of the problem (2D, i.e., two-dimensional, random field) and the regularity of the likelihood function. This decision allowed to reduce computation times without compromising parameter estimation.

2.2 Kriging model principles

The kriging predictor is a linear interpolator whose expressions are derived from supplementary conditions, such as minimizing the prediction variance. For a detailed description of kriging and its construction, the reader can refer to the reference books of Chilès and Delfiner (2012), Cressie (1993) for geostatistics, but also Rasmussen and Williams (2006) for the Gaussian process regression point of view. Let $\boldsymbol{x}_0$ be an unobserved position at which we wish to predict the expected value and the variance of $Z(\boldsymbol{x}_0)|\sigma^2, \phi, \boldsymbol{Z} = \boldsymbol{z}$ (the mean is considered

unknown). The ordinary kriging equations are then

$$\mathbb{E}[Z(\boldsymbol{x}_0)|\sigma^2,\phi,\boldsymbol{Z}=\boldsymbol{z}] = \left(\boldsymbol{r}+\mathbf{1}_n\frac{1-\mathbf{1}_n'\boldsymbol{R}_\phi^{-1}\boldsymbol{r}}{\mathbf{1}_n'\boldsymbol{R}_\phi^{-1}\mathbf{1}_n}\right)'\boldsymbol{R}_\phi^{-1}\boldsymbol{Z},$$

$$\mathrm{Var}[Z(\boldsymbol{x}_0)|\sigma^2,\phi,\boldsymbol{Z}=\boldsymbol{z}] = \sigma^2\left(1-\boldsymbol{r}'\boldsymbol{R}_\phi^{-1}\boldsymbol{r}+\frac{(1-\mathbf{1}_n'\boldsymbol{R}_\phi^{-1}\boldsymbol{r})^2}{\mathbf{1}_n'\boldsymbol{R}_\phi^{-1}\mathbf{1}_n}\right),$$

with $\sigma^2\boldsymbol{r} = (\mathrm{Cov}(Z(\boldsymbol{x}_0),Z(\boldsymbol{x}_j))_{1\leq j\leq n}$.

A major concern for applications of these equations is that they are conditional on the knowledge of the variance and range parameters, which is mostly unrealistic since they are estimated. This assumption yields overoptimistic prediction variances and narrower confidence intervals. This problem is made worse in the case of a small dataset where parameter estimation is sensitive to each observation. To address this issue, more robust methods exist such as cross-validation estimation (Bachoc 2013b). Another solution is to consider the parameters as random variables. Bayesian approach seems natural in this case and leads to Bayesian kriging.

2.3 Bayesian kriging principles

Bayesian kriging deals simultaneously with estimation and predictions by considering the parameters as random variables that must be predicted conditionally to the observed data (Diggle and Ribeiro 2002). Bayesian kriging predictions are derived from the predictive distribution as follows:

$$\begin{aligned} p(Z(\boldsymbol{x}_0)|\boldsymbol{Z}=\boldsymbol{z}) &= \int_{D_\beta\times D_{\sigma^2}\times D_\phi} p(Z(\boldsymbol{x}_0),\beta,\sigma^2,\phi|\boldsymbol{Z}=\boldsymbol{z})d\beta d\sigma^2 d\phi \\ &= \int_{D_\beta\times D_{\sigma^2}\times D_\phi} p(Z(\boldsymbol{x}_0)\beta,\sigma^2|\phi,\boldsymbol{Z}=\boldsymbol{z})p(\phi|\boldsymbol{Z}=\boldsymbol{z})d\beta d\sigma^2 d\phi \\ &= \int_{D_\phi} p(Z(\boldsymbol{x}_0)|\phi,\boldsymbol{Z}=\boldsymbol{z})p(\phi|\boldsymbol{Z}=\boldsymbol{z})d\phi. \end{aligned}$$

The density $p(Z(\boldsymbol{x}_0)|\phi,\boldsymbol{Z}=\boldsymbol{z})$ is known to be a Student's t-density under usual assumptions for the prior distribution (as demonstrated in Le and Zidek (1992)), but the density $p(\phi|\boldsymbol{Z}=\boldsymbol{z})$ is not an usual one, and the integral is usually intractable. In practice, it must therefore be estimated numerically. The most commonly used method for this is a Metropolis-Hasting type algorithm. More details about this algorithm can be found in Tanner (1993) and Carlin and Louis (2013). Another approach given in Tanner (1993), and used in the `geoR` package (Ribeiro and Diggle 2001) of the `R` software, is a Monte-Carlo approximation of the predictive distribution. This approximation is done by sampling the prior distribution of ϕ and injecting these sampled values into the density $p(Z(\boldsymbol{x}_0)|\phi,\boldsymbol{Z}=\boldsymbol{z})$. It should be noted that this approach is also presented in Gaudard et al. (1999), where it is applied to compare different prior choices of

ϕ distribution for modeling and predicting precipitation measurements. So, the algorithm described by Algorithm 1 will be used in the following to estimate the Bayesian prediction.

Algorithm 1 Monte-Carlo approximation for Bayesian kriging

Choose a prior specification and a prediction target $\boldsymbol{x}_0$
Sample a value ϕ from $p(\phi|\boldsymbol{Z} = \boldsymbol{z})$
Compute $p(Z(\boldsymbol{x}_0)|\phi, \boldsymbol{Z} = \boldsymbol{z})$ by injecting the sampled value ϕ
Sample a value $z(\boldsymbol{x}_0)$ from $p(Z(\boldsymbol{x}_0)|\phi, \boldsymbol{Z} = \boldsymbol{z})$
Reiterate the three previous steps as many times as needed to approximate the predictive distribution $p(Z(\boldsymbol{x}_0)|\boldsymbol{Z} = \boldsymbol{z})$
Take $\mathbb{E}[Z(\boldsymbol{x}_0)|\boldsymbol{Z} = \boldsymbol{z}]$ and $Var(Z(\boldsymbol{x}_0)|\boldsymbol{Z} = \boldsymbol{z})$ as prediction mean and prediction variance.

As per usual in Bayesian framework, a joint prior distribution is chosen for β, σ^2:

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}.$$

For ϕ, the prior is reduced to a uniform law between the minimum $d_{\phi,min}$ and the maximum distance $d_{\phi,max}$ allowed by the dataset:

$$\phi \sim \mathcal{U}([d_{\phi,min}, d_{\phi,max}]).$$

According to the priors thus specified, the resulting parameter space is:

$$D_\beta = \mathbb{R}, D_{\sigma^2} =]0, +\infty[, D_\phi = [d_{\phi,min}, d_{\phi,max}].$$

Note that a sensitivity analysis is presented in the Appendix (Section 7) to explain our choice of priors.

3 Validation criteria

Choosing an "optimal" covariance model for geostatistical predictions is a classical issue in geostatistics (Chilès and Delfiner 2012). It has been recently deepened in Demay et al. (2022), where interpretable validation criteria allow to assess the quality of both the predictions of the model and the associated prediction variances. The expressions of these criteria, with some new adaptations, are given in this section in their leave-one-out cross-validation form. Extension to K-fold cross-validation or to test set cases are immediate.

3.1 Predictivity coefficient (Q^2)

The main goal of this coefficient, often called "Nash-Sutcliffe criterion" (Nash and Sutcliffe 1970), is to evaluate the predictive accuracy of the model by nor-

malising the errors, allowing a direct interpretation in terms of explained variance. Its practical definition (Marrel et al. 2008) is:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(z(\boldsymbol{x}_i) - \widehat{z}_{-i})^2}{\sum_{i=1}^{n}(z(\boldsymbol{x}_i) - \widehat{\mu})^2},$$

where $\widehat{z}_{-i}$ is the value predicted at location $\boldsymbol{x}_i$ by the model built without the i-th observation (the one located at $\boldsymbol{x}_i$) and $\widehat{\mu}$ is the empirical mean of the dataset. Its theoretical definition can be found in Fekhari et al. (2022).

The Q^2 coefficient measures the quality of the predictions and how near they are to the observed values. Its formula is similar to the coefficient of determination used for regression (with independent observations), but estimated here in prediction (by using cross-validation residuals). The closer its value is to 1, the better the predictions are (relatively to the observations). As a rule of thumb, if the Q^2 is smaller than 0.5 (i.e., less than 50% of output variance explained), the model is not considered to be good, and should be used with precaution if no other model offers better predictions.

3.2 Predictive variance adequacy (PVA)

This second criterion aims to quantify the quality of the prediction variances given by the kriging model. Finely studied in Bachoc (2013a;b) and Demay et al. (2022), it is defined by

$$PVA = \left| \log \left(\frac{1}{n} \sum_{i=1}^{n} \frac{(z(\boldsymbol{x}_i) - \widehat{z}_{-i})^2}{\widehat{s}^2_{-i}} \right) \right|,$$

where $\widehat{s}^2_{-i}$ is the prediction variance (at location $\boldsymbol{x}_i$) of the model built without the i-th observation (the one located at $\boldsymbol{x}_i$).

This coefficient estimates the average ratio between the squared observed prediction error and the prediction variance. It therefore gives an indication of how much a prediction variance is larger or smaller than the one expected. The closer the PVA is to 0, the better the prediction variances are. For example, a PVA close to 0.7 indicates prediction variances that are on average two times larger or smaller than the squared errors.

3.3 Predictive interval adequacy (PIA)

The PVA is a criterion of variance adequacy but does not take into account a possible skewness in the predictive distribution. In the Gaussian case (like ordinary kriging), mean and variance completely characterise the distribution. But in the case of Bayesian kriging where the predictive distribution is no longer Gaussian, the Q^2 and PVA are not sufficient to evaluate the quality of the model and its prediction. As such, we propose a new complementary geometrical criterion called the predictive interval adequacy (PIA) and defined as

$$PIA = \left| \log \left(\frac{1}{n} \sum_{i=1}^{n} \frac{(z(\boldsymbol{x}_i) - \widehat{z}_{-i})^2}{\left(\widehat{q}_{0.31,-i} - \widehat{q}_{0.69,-i}\right)^2} \right) \right|,$$

where $\widehat{q}_{0.31,-i}$ (respectively $\widehat{q}_{0.69,-i}$) is the estimation of the quantile of order 0.31 (respectively 0.69) of the predictive distribution (at location $\boldsymbol{x}_i$) without the i-th observation.

The PIA has been defined to be identical to the PVA for a Gaussian distribution. However, rather than comparing squared errors to the predictive variance, it compares the width of prediction intervals with the squared errors. Another main difference is that the intervals considered by the PIA are centered on the median while those of the PVA are centered around the mean. Finally, an estimation of the predictive distribution is necessary to compute in practice this criterion, whereas the PVA only requires the computation of predictive mean and variance.

3.4 α-CI plot

The Gaussian process model allows to build prediction intervals of any level $\alpha \in]0,1[$:

$$CI_\alpha(z(\boldsymbol{x}_i)) = [\widehat{z}_{-i} - \widehat{s}_{-i} q^{\mathcal{N}}_{(1+\alpha)/2} \,;\, \widehat{z}_{-i} + \widehat{s}_{-i} q^{\mathcal{N}}_{(1+\alpha)/2}],$$

where $q^{\mathcal{N}}_{(1+\alpha)/2}$ is the quantile of order $(1+\alpha)/2$ of the standard normal distribution. This expression is only valid if all parameters are known. For example, if the variance parameter is poorly estimated, the width of the predicted confidence intervals will not reflect what we might observe. But how can we validate a confidence interval without prior knowledge of the model parameters? The idea behind this criterion (see Marrel et al. (2012) and Demay et al. (2022)) is to evaluate empirically the number of observations falling into the predicted confidence intervals and to compare this empirical estimation to the theoretical ones expected:

$$\Delta_\alpha = \frac{1}{n}\sum_{i=1}^{n} \phi_i \text{ where } \delta_i = \begin{cases} 1 & \text{if } z(\boldsymbol{x}_i) \in CI_\alpha(z(\boldsymbol{x}_i)) \\ 0 & \text{else.} \end{cases}$$

This value can be computed for varying α, and can then be visualised against the theoretical values, yielding what Demay et al. (2022) calls the α-CI plot, with an example given in Figure 1.

Similarly to the PIA, the α-CI plot must be adapted to the Bayesian kriging since the posterior distribution is not Gaussian. We therefore introduce a slightly different criterion based on the quantiles of the predictive distribution. More precisely, this modified α-CI plot relies now on credible intervals defined as

$$\widetilde{CI}_\alpha(z(\boldsymbol{x}_i)) = [\widehat{q}_{\frac{1-\alpha}{2}} ; \widehat{q}_{\frac{1+\alpha}{2}}],$$

where $\widehat{q}_{\frac{1-\alpha}{2}}$ (respectively $\widehat{q}_{\frac{1+\alpha}{2}}$) is the estimation of the quantile of order $\frac{1-\alpha}{2}$ (respectively $\frac{1+\alpha}{2}$) of the predictive distribution (at location $\boldsymbol{x}_i$) of the model built without the i-th observation. Once again, we obtain a criterion that is identical for both methods when the predictive distribution is Gaussian. Illustrations of α-CI plot can be found in Demay et al. (2022).

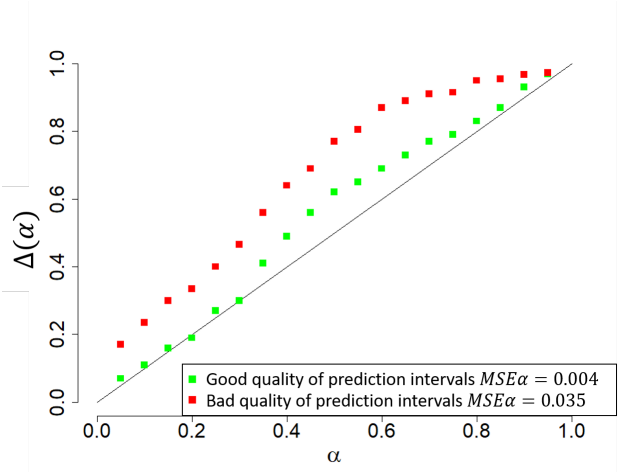


Figure 1: Example of two α-CI plots and corresponding values of MSEα for a good (in green) and bad (in red) predictive error model.

3.5 Mean Squared Error α ($MSE\alpha$)

Finally, to summarise the α-CI plot, we also introduce a quantitative criterion called "Mean Squared Error α" and defined as

$$MSE\alpha = \frac{1}{n_\alpha}\sum_{j=1}^{n_\alpha}(\Delta_{\alpha_j} - \alpha_j)^2,$$

where the considered levels α are discretised over $]0,1[$ in n_α possible values. In practice a regular discretization will be considered to compute $MSE\alpha$. The closer this criterion is to 0, the better the confidence/credible intervals are on average. To illustrate the values taken by the criterion, Figure 1 gives the α-CI plot corresponding to a "good" and "bad" model fitting.

In this graph, the bad model yields a $MSE\alpha$ of 0.035 against 0.004 for a model with more accurate prediction intervals. As a rule of thumb, a $MSE\alpha$ of 0.01 will be considered to correspond to a model with wrong predictive intervals, while a model with a $MSE\alpha$ of 0.001 will be deemed to have correct predictive intervals. Similarly to the PVA, the $MSE\alpha$ does not explain if the poorly fitted predictive intervals are due to badly centered predictive intervals or if the predictive variance was badly estimated (and whether or not this variance was underestimated or overestimated). This criterion must therefore be used in conjunction with the previous criteria to better assert the model qualities and weaknesses. Finally, this criterion also offers a quantitative tool for comparing different models if the α-CI plots do not allow to clearly distinguish the performances of competing models. This will be illustrated in particular in the numerical tests in Section 4.2 (Fig. 8).

The different aforementioned criteria provide complementary information to evaluate the prediction quality of the kriging model, either in terms of mean, variance or confidence/credible intervals. They will be used in the following to compare the performance of ordinary and Bayesian kriging.

4 Numerical tests and results

Our goal is to compare Bayesian and ordinary kriging (the latter being the more commonly used kriging method). To do so, the different criteria mentioned in Section 3 will be computed on datasets (i.e., samples of observations), coming from different models, of different sizes.

4.1 Datasets from 2D Gaussian process simulations

First, we consider samples simulated from an analytical Gaussian process model with known parameters. More precisely, the samples are simulated in the input space $[0, 10]^2$ from a Gaussian process with an exponential covariance (i.e., the Matérn covariance of Eq. (1) with $\nu = 0.5$) and the following parameters:

$$\beta = 0.5, \sigma^2 = 0.1, \phi = 4.5.$$

We simulate datasets of different sizes, varying from 16 to 81 observations, sampled on a square grid in the input space. Here, the sampling designs will be regular squared grids. This choice is made to comply with the application purpose which deals with D&D constraints of buildings. Indeed, most of the times, the radiological measurements inside buildings are made regularly (equidistant location) along lines of investigations (see, e.g. Attiogbe et al. (2014) and EPRI (2016)). For each size, the process is repeated 100 times with independent random Gaussian process simulations.

For each dataset, Bayesian and ordinary kriging models are estimated and the different validation criteria are computed by cross-validation. Every kriging predictions (Bayesian and ordinary) are made with the `R` package `geoR` (Ribeiro and Diggle 2001). Results are given in Figure 2 with boxplots (corresponding to the 100 random replicates) w.r.t. the dataset sizes. The results indicate that Bayesian kriging performs better in terms of both mean and prediction variance for small sample sizes. More precisely, Bayesian kriging outperforms ordinary kriging on all the four criteria for datasets with less than 40 observations. This result is especially visible for the PVA and PIA and shows that the main difference between both kriging methods still lies in the predictive variance accuracy. By taking into account the additional uncertainty in the estimation of the Gaussian process parameters, Bayesian kriging therefore yields larger and more accurate prediction intervals, and as a result better PVA, PIA, and $MSE\alpha$ criteria.

More precisely, if we first look at the median values of Q^2 estimation, these increase from 0 to 0.63, according to the data size, for ordinary kriging. Bayesian kriging gives better Q^2 for smaller datasets, starting from a median value of 0.16

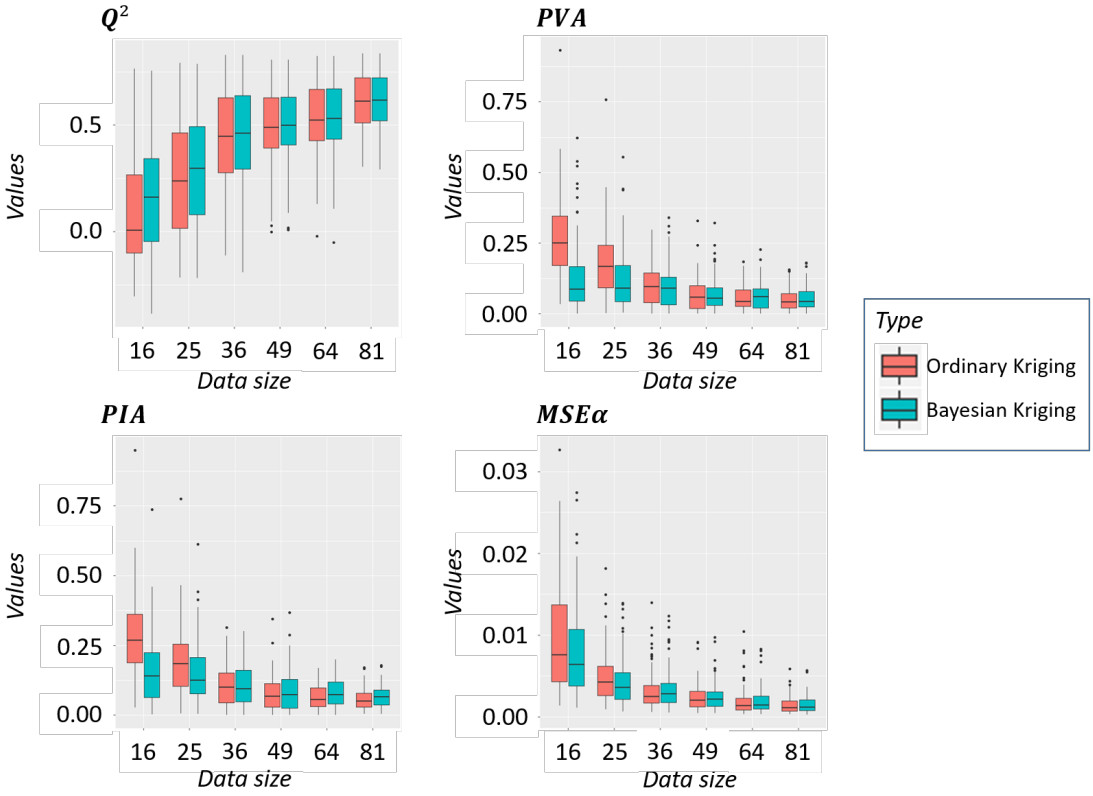


Figure 2: Distribution of validation criteria (Q^2, PVA, PIA, and $MSE\alpha$) w.r.t. the size of datasets, for Gaussian process simulation datasets.

up to 0.63. For a fixed sample size, the dispersion of Q^2 is quite similar between both kriging methods (the geometry of sampling being fixed).

Regarding the median of PVA, the value range from 0.25 to 0.04 for ordinary kriging, compared to 0.08 to 0.04 for Bayesian kriging. For the PIA, the results are identical for ordinary kriging, but Bayesian kriging performs slightly worse, starting at 0.15 up to 0.05. We can also see that the dispersion of PIA and PVA estimates is different for small datasets between both kriging methods. This is explained by the fact that PVA and PIA are sensitive to the parameter estimation process. Since the number of observation is low, maximum likelihood estimations are not robust, yielding large variations in parameter estimations, and therefore in PVA and PIA estimations. Finally, we observe that for datasets larger or equal to 49, Bayesian kriging seems to perform slightly worse, but the differences are too small to be considered significant.

The $MSE\alpha$ graph shares similarities with the other graphs, since confi-

dence and credible intervals both depends on prediction mean and variance. For the ordinary kriging, the median $MSE\alpha$ goes from 0.0075 to 0.0025, while for Bayesian kriging the values are lower, from 0.006 to 0.0025. The evolution observed is similar between the PVA and PIA, with Bayesian kriging yielding better results for smaller datasets.

It can also be noted that for larger datasets, Bayesian and ordinary kriging yield similar results. This observation was to be expected, since Bayesian and inferential methodology coincide for larger datasets. It can therefore be argued that Bayesian kriging becomes less advantageous and relevant for datasets with more than 40 observations. Note that Q^2 values are also extremely low for 49 observations or fewer, but again this is to be expected for very small datasets.

4.2 Datasets from a 2D deterministic function

In order to test the kriging models on cases that do not fall within the theoretical framework of the Gaussian process hypothesis, we consider a sample coming from the following two-dimensional deterministic function (Iooss et al. 2010):

$$f(x,y) = \frac{e^x}{5} - \frac{y}{5} + \frac{y^6}{3} + 4y^4 - 4y^2 + \frac{7x^2}{10} + x^4 + \frac{3}{4x^2+4y^2+1}, \qquad (2)$$

where (x, y) are the function inputs. Figure 3 shows this function over the $D = [-1, 1]^2$ input space.

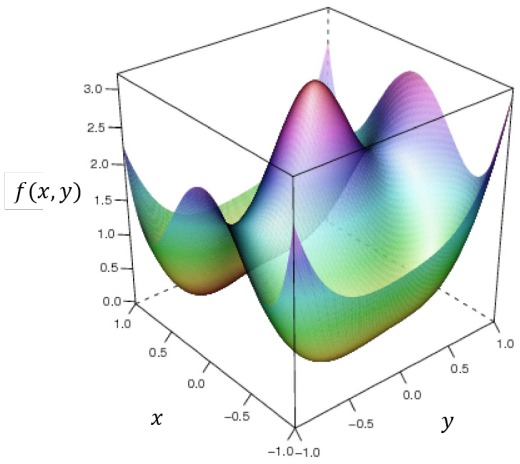


Figure 3: Illustration of the deterministic function f.

We consider two steps for studying this test function. First, the validation criteria are used to compare the results obtained by using different covariance functions in order to identify the most appropriate one for the dataset (as done in Demay et al. (2022)).

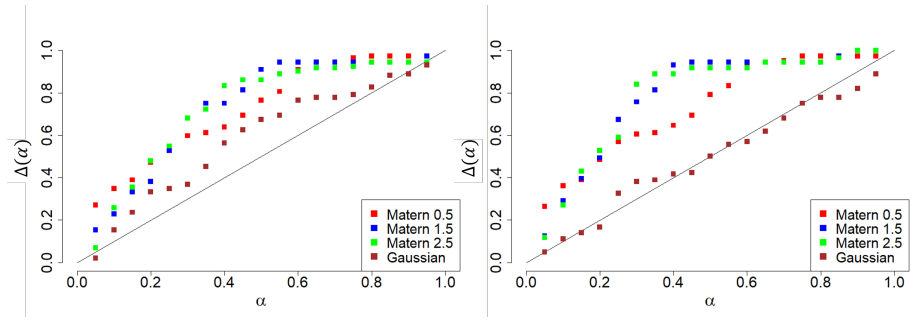
Then, a regular squared grid is considered to sample the input space, composed here of 144 observations. On this dataset, the ordinary kriging model is fitted with different covariance functions, namely three Matérn covariances and the Gaussian covariance with a nugget effect for the latter of 10^{-6} (to improve the numerical stability of the covariance matrix inversion). For each of these covariances, the validation criteria are estimated by a cross validation process. The results are presented in Table 1 for ordinary kriging, in Table 2 for Bayesian kriging and in Figure 4.

Covariance	Q^2	PVA	PIA	$MSE\alpha$
Matérn 1/2	0.95	0.99	0.98	0.056
Matérn 3/2	0.99	0.91	0.90	0.073
Matérn 5/2	1.00	0.65	0.63	0.073
Gaussian	1.00	0.05	0.07	0.011

Table 1: Validation criteria for the ordinary kriging with different covariance functions, on the sample of $n = 144$ observations of function f.

Covariance	Q^2	PVA	PIA	$MSE\alpha$
Matérn 1/2	0.95	1.09	1.06	0.061
Matérn 3/2	0.99	1.62	1.60	0.106
Matérn 5/2	1.00	1.58	1.55	0.106
Gaussian	1.00	0.13	0.16	0.002

Table 2: Validation criteria for the Bayesian kriging with different covariance functions, on the sample of $n = 144$ observations of function f.



(a) Ordinary kriging.

(b) Bayesian kriging.

Figure 4: α-CI plots for the ordinary and Bayesian kriging with different covariances functions, on the sample of $n = 144$ observations of function f.

The results show that, in this case, a Gaussian covariance function is the most appropriate covariance function w.r.t. to the different criteria. This result is not

surprising since the test function is smooth and shows large correlations between observations. Although the differences between Q^2 are very small between the Gaussian and Matérn models (except for the Matérn 1/2 model), significant differences appear for the PVA and PIA. These differences become smaller for the $MSE\alpha$. This shows the importance of using simultaneously various criteria for a better assessment of the model performance and accuracy.

Once our covariance model is chosen (the Gaussian one in this case), we can apply a similar test protocol than in Section 4.1. In order to generate datasets, we have to slightly modify the protocol. Since the function is deterministic, choosing a specific geometry for a fixed dataset size will not allow to generate different datasets. Therefore we discard here the regular grid and choose to sample random positions in the input space. The observed dispersion in the results of this section is affected by that choice. This sampling is repeated 100 times for each data set sizes, up to 150 observations.

The results are presented in Figure 5. The values of the Q^2 criterion lead to the same conclusions as for the data from Gaussian process trajectories, in the previous section. We again find better performance with Bayesian kriging, especially for small sample sizes. Note that we have higher Q^2's than for the previous test case due to the high regularity of the function f.

Significant differences arise with the PVA, PIA and $MSE\alpha$ criteria. Indeed, these criteria do not decrease steadily and monotonically with the number of observations. Moreover, they behave differently depending on the type of kriging. More precisely, for Bayesian kriging, the PVA, PIA and $MSE\alpha$ increase between 20 observations and 50 observations, before decreasing, whereas they keep increasing for ordinary kriging. For datasets made of 50 observations or less, Bayesian kriging seems to under-perform when compared to ordinary kriging but outperform ordinary kriging for more than 50 observations. Still, once the size of the datasets exceed 80 observations, we observe similar results to those obtained with the simulated datasets.

To explain these results, we recall that the initial assumption whereby the function f is a trajectory of a Gaussian process is not verified here, at least for datasets of 50 or less observations. It is therefore possible to obtain poorer criteria as the dataset size increases. We still get good prediction accuracy, since the median of the Q^2 criterion stays between 0.7 and 1 for all dataset sizes and kriging methods, but the predicted variances do not seem to be very accurate, yielding poorly estimated confidence and credible intervals. We can observe that once the dataset size exceeds 80 observations, the evolution of the validation criteria shows that the initial assumption is now valid.

In conclusion, in this deterministic function case for which the initial assumption of a random field is not ensured, Bayesian kriging seems to be more robust than ordinary kriging. Caution is still advised, since in some cases ordinary kriging seems to perform better than Bayesian kriging, as illustrated with the $n = 40$ or $n = 50$ observations' dataset.

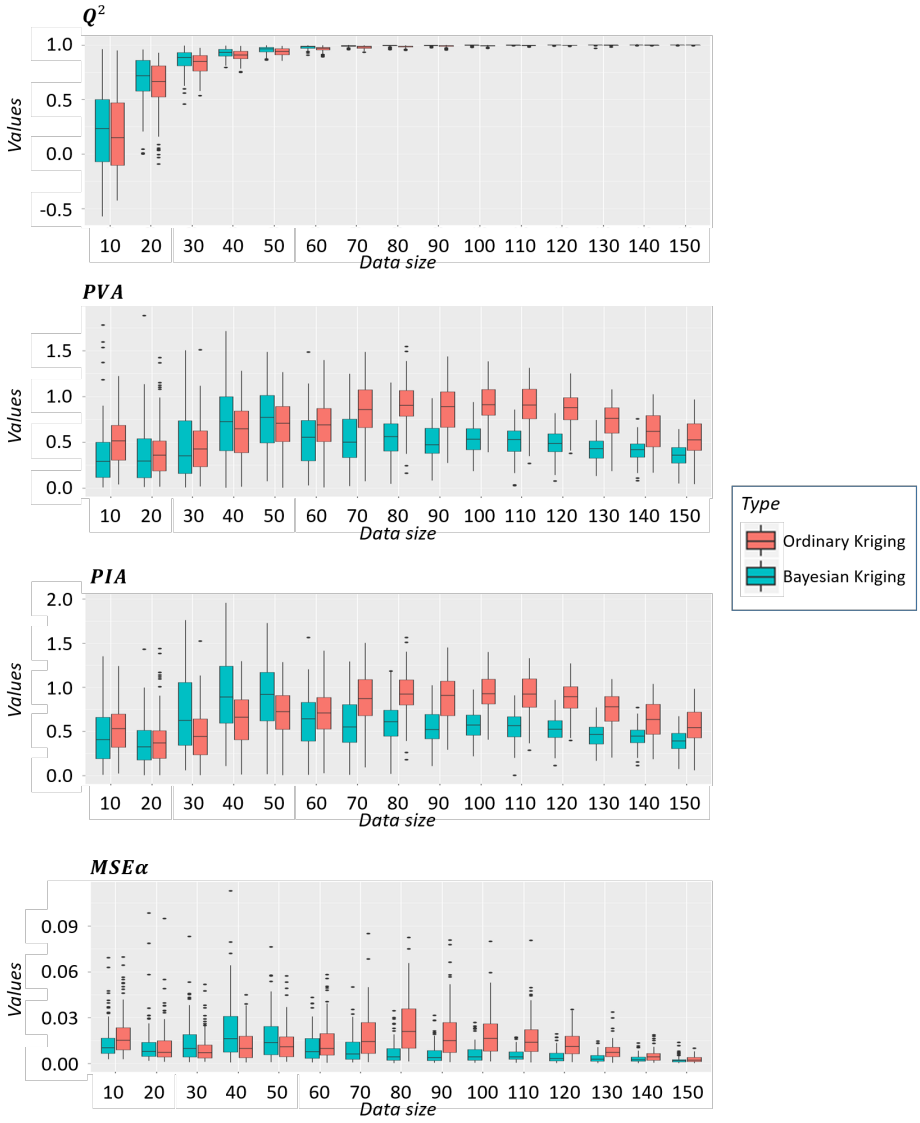


Figure 5: Distribution of validation criteria (Q^2, PVA, PIA, and $MSE\alpha$) w.r.t. the size of datasets, for the deterministic function f.

5 Real application case: G3's dataset

This dataset is made of 70 observations of radioactivity measurements coming decommissioning project of the CEA Marcoule G3 reactor (CEA 2009). They

are sampled in the input domain $[0, 10] \times [0, 7]$. The dataset is mapped in Figure 6. A first kriging modeling using the full dataset is given by Figure 7. More precisely, the prediction maps obtained with ordinary and Bayesian kriging with an exponential covariance for both are given. These prediction maps show small apparent differences between both methods, since the number of observations is high. Let us now see if the same is true if we reduce the sample size and vary the covariance.

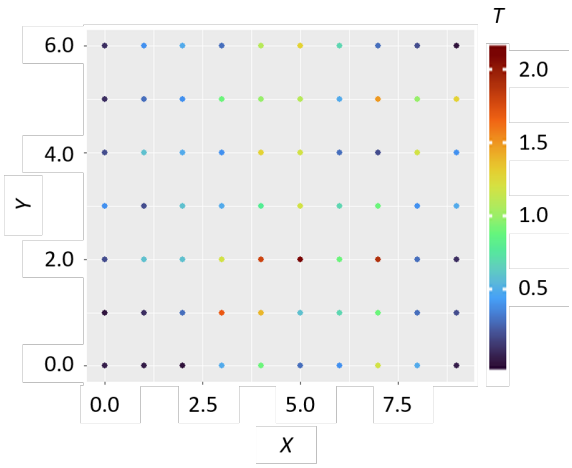


Figure 6: Mapping of G3 observations

Then, a similar test protocol as in Section 4 is applied to assess the behaviour of kriging models according to n. First, let us consider ordinary kriging for different covariance functions, applied to the initial set of 70 observations. The validation criteria estimated by cross-validation are given in Table 3 and Figure 8. For Bayesian kriging, they are given in Table 4 and Figure 8. The results indicate that the Matérn 1/2 model is the best choice in regards of our different criteria since it maximizes the Q^2 criterion while minimizing both PVA and PIA criteria (it also performs well for the $MSE\alpha$ criterion, while not being the function minimizing it overall). Therefore only the Matérn 1/2 covariance function is considered.

To generate multiple datasets, we resample without replacement datasets of various sizes $n = 20, 30, 40, 50, 60, 70$, with the last one being the original dataset. Once again, the process is repeated 100 times for each sample size (except for 70 observations) and for each sample a cross-validation is applied to estimate the validation criteria.

The obtained results are summarised in Figure 9. For the Q^2 criterion, the median values increase from about 0 ($n = 10$) to 0.38 ($n = 70$) for both kriging methods. Slightly higher results are obtained for Bayesian kriging, especially for small sample sizes. The dispersion of Q^2 is similar between the two kriging

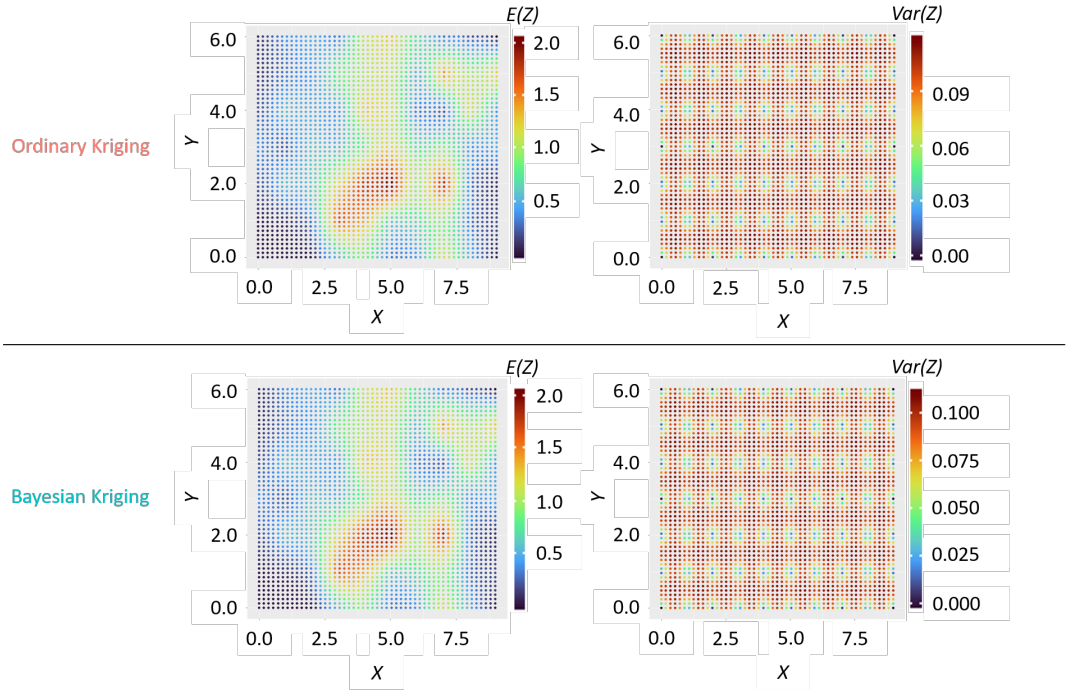


Figure 7: Mapping of predictions obtained with ordinary (top) and Bayesian (bottom) kriging. Kriging predictor (left) and predictive variance (right) are plotted.

Covariance	Q^2	PVA	PIA	$MSE\alpha$
Matérn 1/2 (exponential)	0.37	0.06	0.07	0.0015
Matérn 3/2	0.33	0.12	0.14	0.0010
Matérn 5/2	0.31	0.14	0.15	0.0014
Gaussian	0.24	0.16	0.18	0.0021

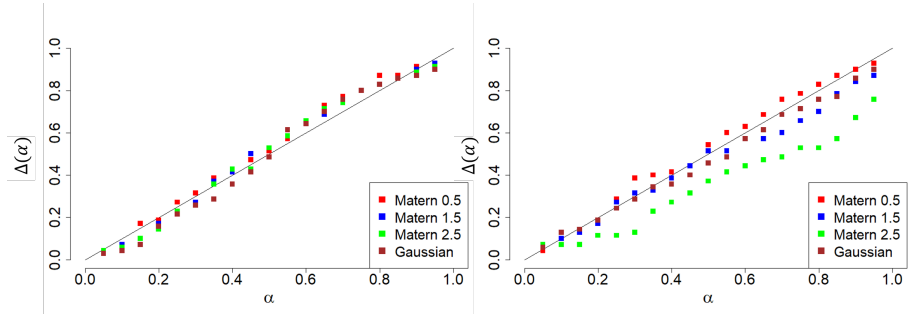
Table 3: Validation criteria for the ordinary kriging with different covariance functions, on the G3 sample of $n = 70$ observations.

methods. The obtained Q^2 estimates here are very low, which normally means that the model is not predictive enough. As our objective is only to compare the kriging methods, this problem is not further investigated here.

Regarding now the PVA, the median values decrease from 0.47 to 0.16 for ordinary kriging, compared to much lower values for Bayesian kriging, namely from 0.19 to 0.06 For the PIA, the values are very close to the ones of PVA. For the $MSE\alpha$, the median values go from 0.011 to 0.0017 for the ordinary kriging,

Covariance	Q^2	PVA	PIA	$MSE\alpha$
Matérn 1/2 (exponential)	0.38	0.12	0.07	0.0013
Matérn 3/2	0.20	0.51	0.55	0.0028
Matérn 5/2	0.16	1.19	1.25	0.0284
Gaussian	0.15	0.36	0.40	0.0015

Table 4: Validation criteria for the Bayesian kriging with different covariance functions, on the G3 sample of $n = 70$ observations.



(a) Ordinary kriging.

(b) Bayesian kriging.

Figure 8: α-CI plots for the ordinary and Bayesian kriging with different covariance functions, on the G3 sample of $n = 70$ observations.

against 0.008 down to 0.0017 for Bayesian kriging. Once again, Bayesian kriging yields better results, especially for smaller datasets. The results of both methods then become almost identical for data sets of 40 or more observations. This is especially visible for the $MSE\alpha$.

We can also remark that the variance of each validation criterion is reduced as the datasets size grows. This is both explained by the larger datasets, but also by our protocol, since observations are randomly drawn without replacement among the original 70 observations. As a result, the samples differ less and less as the dataset sizes increases.

6 Discussion and Conclusions

In conclusion, the use of Bayesian kriging for spatial interpolation of datasets in support of decommissioning and dismantling projects shows promising results. It allows to take into account the uncertainty on the parameters of the kriging model. It is particularly relevant for small datasets for which it outperforms the ordinary kriging in terms of accuracy of predictive mean, variance and predictive intervals. This advantage becomes less important as the sample size increases: ordinary kriging, less computationally expensive, is then preferable for large datasets. Bayesian kriging has also the drawback of requiring a prior specifica-

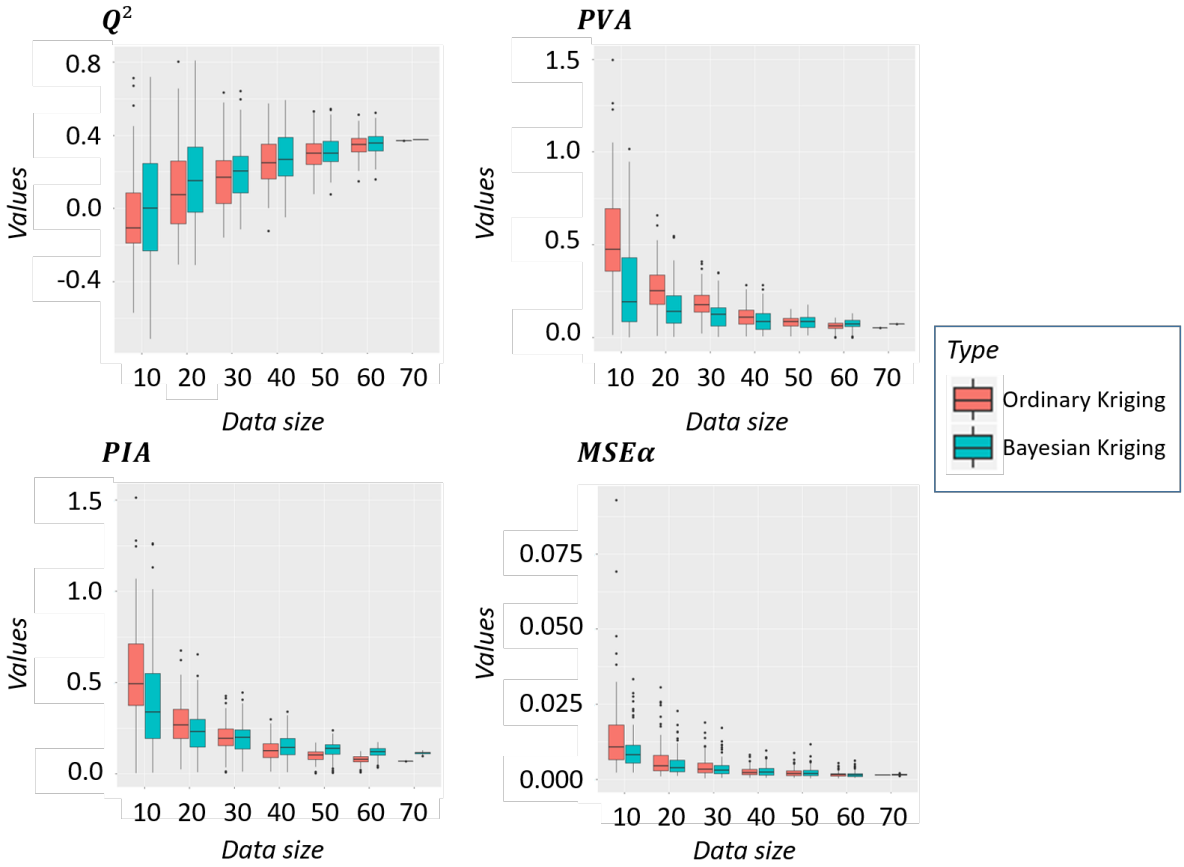


Figure 9: Distribution of validation criteria (Q^2, PVA, PIA, and $MSE\alpha$) w.r.t. the size of datasets, for the G3 dataset.

tion, which is often difficult to choose and can strongly influence the predictions. Therefore, the use of Bayesian kriging should be restricted to smaller datasets or cases in which prior information on parameters is well known.

In our work we did not use the nugget effect as a modelling tool but only as a regularisation of the Gaussian covariance function. Future works will aim at adding this parameter in the model. This could be taken further by considering a heteroscedastic model (Ng and Yin 2012), since the usual nugget effect is formulated as a homoscedastic model. This could be extremely useful and show promising results in the framework of D&D of nuclear sites since radioactive measurements are prone to varying measurement uncertainties, depending on the measuring technique.

7 Appendix: sensitivity analysis to the prior distribution of parameters

The choice of prior specifications is a complicated step in Bayesian analysis. We therefore conduct a sensitivity analysis to justify our use of an improper prior on the mean and variance parameters. Note that the range will not be described here, since no usual specification is available.

First, it could be argued that the prior on the parameter β is chosen improper since this choice is implicitly made in ordinary kriging:

$$\pi(\beta) \propto 1.$$

Second, for the variance parameter σ^2, several choices for priors can be considered. To give a quick overview of our test protocol, we used a simulated dataset, defined as random trajectories of the same Gaussian process model as in Section 4.1. An initial dataset of 16641 observations is simulated, on which the parameters β_{init} and σ^2_{init} are estimated. These estimations will be considered as reference values for our prior specifications. From these 16641 observations, samples of $n = 20$ and $n = 50$ observations are randomly drawn. This sampling is then repeated 100 times, generating a total of 200 datasets. Then, for each dataset, the Bayesian kriging is applied considering five different prior specifications:

1. vague with
$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2},$$

2. correctly centred and informative with
$$\sigma^2 \sim \text{ Scaled-Inv-}\chi^2(\sigma^2_{\text{init}}, n) \text{ and } \beta|\sigma^2 \sim \mathcal{N}(\beta_{\text{init}}, \frac{\sigma^2}{n}),$$

3. incorrectly centred and informative with
$$\sigma^2 \sim \text{ Scaled-Inv-}\chi^2(3\sigma^2_{\text{init}}, n) \text{ and } \beta|\sigma^2 \sim \mathcal{N}(3\beta_{\text{init}}, \frac{\sigma^2}{n}),$$

4. correctly centred and non-informative with
$$\sigma^2 \sim \text{ Scaled-Inv-}\chi^2(\sigma^2_{\text{init}}, \frac{n}{3}) \text{ and } \beta|\sigma^2 \sim \mathcal{N}(\beta_{\text{init}}, \frac{\sigma^2}{n}),$$

5. incorrectly centred and non-informative with
$$\sigma^2 \sim \text{ Scaled-Inv-}\chi^2(3\sigma^2_{\text{init}}, \frac{n}{3}) \text{ and } \beta|\sigma^2 \sim \mathcal{N}(3\beta_{\text{init}}, \frac{\sigma^2}{n}).$$

For each prior specification, validation criteria are estimated by cross-validation. The obtained results are given in Figure 10. First, we observe that the Q^2 criterion is not sensitive to the prior specification. This is expected since the prediction performances depend mostly on the number of observations and on the geometry of the dataset. On contrary, the PVA and PIA criterion are very sensitive to the prior specification since prediction variance highly depends on parameter estimation. A vague prior allows to mitigate the bias introduced with an incorrectly centred prior, as case 3 shows a worse result than case 5. We can also see that even with a correctly centred and informative prior (case 2), the gains in parameter estimation are small if we compare it to a vague specification (case 1).

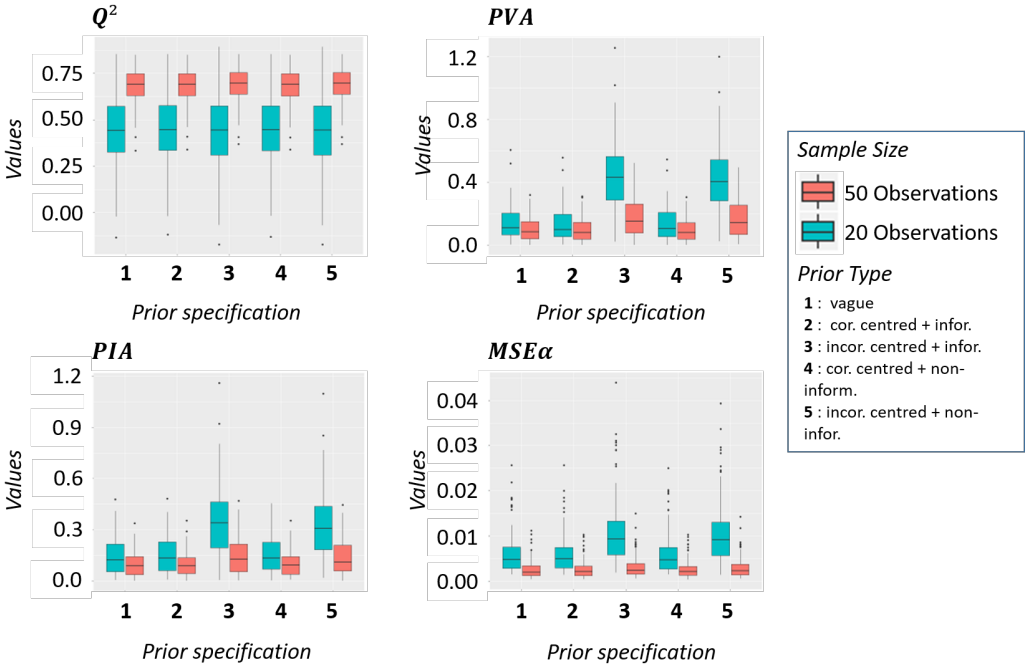


Figure 10: Distribution of validation criteria (Q^2, PVA, PIA, and $MSE\alpha$) w.r.t. to the prior specification.

In conclusion, the choice of a vague or improper prior seems to be reasonable, as the improvements provided by a correctly specified prior do not seem good enough in comparison to the pitfall of a bad prior specification. These results are also similar to the one obtained by Helbert et al. (2009).

References

Al-Mudhafar W J (2019) Bayesian kriging for reproducing reservoir heterogeneity in a tidal depositional environment of a sandstone formation. Journal of Applied Geophysics 160:84–102

Attiogbe J, Aubonnet E, Maquille L D, Moura P D, Desnoyers Y, Dubot D, Feret B, Fichet P, Granier G, Iooss B, Nokhamzon J G, Dehaye C O, Pillette-Cousin L, Savary A (2014) Soil radiological characterisation methodology. CEA-R–6386, Commissariat à l'énergie atomique et aux énergies alternatives (CEA). CEA Marcoule Center, Analytical Methods Committee (CETAMA), France

Bachoc F (2013a) Cross validation and maximum likelihood estimations of hyperparameters of Gaussian processes with model misspecification. Computational Statistics and Data Analysis 66:55–69

Bachoc F (2013b) Parametric estimation of covariance function in Gaussian-process based kriging models. Application to uncertainty quantification for computer experiments. PhD Thesis, Université Paris Diderot - Paris VII

Bechler A, Romary T, Jeannée N, Desnoyers Y (2013) Geostatistical sampling optimization of contaminated facilities. Stochastic Environmental Research and Risk Assessment 27:1967–1974

Blatman G, Delage T, Iooss B, Pérot N (2017) Probabilistic risk bounds for the characterization of radiological contamination. The European Journal of Physics - Nuclear Sciences & Technology (EPJ-N) 3:23

Boden S, Rogiers B, Jacques D (2013) Determination of ^{137}cs contamination depth distribution in building structures using geostatistical modeling of ISOCS measurements. Applied Radiation and Isotopes 79:25–36

Carlin B, Louis T (2013) Bayesian methods for data analysis, Third Edition. CRC Press

CEA (2009) Marcoule : dismantling the G1, G2 and G3 reactors. http://www.francetnp.gouv.fr/IMG/pdf/D-Dem_G1_G2_G3.pdf

CEA/DEN (2017) L'assainissement-démantèlement des installations nucléaires. Monographie CEA, CEA et Editions Le Moniteur

Chilès J P, Delfiner P (2012) Geostatistics : Modeling spatial uncertainty, second edition. Wiley

Cressie N (1993) Statistics for spatial data. John Wiley & Sons

Daya Sagar B, Cheng Q, Agterberg F (2018) Handbook of Mathematical Geosciences: Fifty Years of IAMG. Springer

Demay C, Iooss B, Le Gratiet L, Marrel A (2022) Model selection based on validation criteria for Gaussian process regression: An application with highlights on the predictive variance. Quality and Reliability Engineering International 38(3):1482–1500

Desnoyers Y (2010) Approche méthodologique pour la caractérisation géostatistique des contaminations radiologiques dans les installations nucléaires. Phd thesis, Ecole Nationale Supérieure des Mines de Paris

Desnoyers Y, Chilès J P, Dubot D, Jeannée N, Idasiak J M (2011) Geostatistics for radiological evaluation: study of structuring of extreme values. Stochastic Environmental Research and Risk Assessment 25:1031–1037

Desnoyers Y, Faucheux C, Pérot N (2020) Use case 3: post accidental site remediation. The European Journal of Physics - Nuclear Sciences & Technology (EPJ-N) 6:13

Diggle P J, Ribeiro P J (2002) Bayesian inference in Gaussian model-based geostatistics. Geographical and Environmental Modelling 6(2):129–146

Diggle P J, Ribeiro P J (2007) Model-based geostatistics. Springer

EPRI (2016) Guidance for using geostatistics in developing a site final status survey program for plant decommissionning. 3002007554, Electic Power Research Institute (EPRI), USA

Fekhari E, Iooss B, Muré J, Pronzato L, Rendas J (2022) Model predictivity assessment: incremental test-set selection and accuracy evaluation. In Salvati N, Perna C, Marchetti S, Chambers R, editors, Studies in Theoretical and Applied Statistics, SIS 2021, Pisa, Italy, June 21-25, Springer

Gaudard M, Karson M, Linder E, Sinha D (1999) Bayesian spatial prediction. Environmental and Ecological Statistics 6(2):147–171

Goudeau V, Galet N, Dubot D, Attiogbe J, Aubonnet E, Lalanne J Y (2015) Mobile platform for radiological characterization of sites under or after decommissioning. In WM2015 Conference Proceedings - Waste Management Symposia, Phoenix, Arizona, USA

Gupta A, Kamble T, Machiwal D (2017) Comparison of ordinary and Bayesian kriging techniques in depicting rainfall variability in arid and semi-arid regions of North-West India. Environmental Earth Sciences 76(15):512

Handcock M S, Stein M L (1993) A Bayesian analysis of kriging. Technometrics 35(4):403–410

Helbert C, Dupuy D, Carraro L (2009) Assessment of uncertainty in computer experiments from Universal to Bayesian kriging. Applied Stochastic Models in Business and Industry 25(2):99–113

Iooss B, Boussouf L, Feuillard V, Marrel A (2010) Numerical studies of the metamodel fitting and validation processes. International Journal On Advances in Systems and Measurements 3:11–21

Jeannée N, Desnoyers Y, Lamadie F, Iooss B (2008) Geostatistical sampling optimization of contaminated premises. In DEM - Decommissionning challenges: an industrial reality?, Avignon, France

Kitanidis P (1986) Parameter uncertainty in estimation of spatial functions: Bayesian analysis. Water Resources Research 22(4):499–507

Krivoruchko K, Gribov A (2019) Evaluation of empirical Bayesian kriging. Spatial Statistics 32:100368

Lajaunie C, Renard D, Quentin A, Le Guen V, Caffari Y (2020) A non-homogeneous model for kriging dosimetric data. Mathematical Geosciences 52:847–863

Le N D, Zidek J V (1992) Interpolation with uncertain spatial covariances: A Bayesian alternative to kriging. Journal of Multivariate Analysis 43(2):351–374

Marrel A, Iooss B, Da Veiga S, Ribatet M (2012) Global sensitivity analysis of stochastic computer models with joint metamodels. Statistics and Computing 22:833–847

Marrel A, Iooss B, Van Dorpe F, Volkova E (2008) An efficient methodology for modeling complex computer codes with Gaussian processes. Computational Statistics and Data Analysis 52:4731–4744

Nash J, Sutcliffe J (1970) River flow forecasting through conceptual models part I–A discussion of principles. Journal of Hydrology 10(3):282–290

Ng S H, Yin J (2012) Bayesian kriging analysis and design for stochastic simulations. ACM Transactions on Modeling and Computer Simulation 22(3):17:1–17:26

Pérot N, Le Cocguen A, Carré D, Lamotte H, Duhard-Baronne A, Pointeau I (2020) Sampling strategy and statistical analysis for radioactive waste characterization. Nuclear Engineering and Design 364:110647

Rasmussen C, Williams C (2006) Gaussian processes for machine learning. MIT Press

Ribeiro P, Diggle P (2001) geoR: a package for geostatistical analysis. R-NEWS 1(2):14–18

Tanner M A (1993) Tools for statistical inference. New York, NY: Springer US

Webster R, Oliver M A (2007) Geostatistics for environmental scientists. John Wiley & Sons

Zaffora B, Magistris M, Saporta G, Torre F L (2016) Statistical sampling applied to the radiological characterization of historical waste. The European Journal of Physics - Nuclear Sciences & Technology (EPJ-N) 2:11

B.2 Liste des communications réalisées

Voici toutes les communications (écrites, orales et posters) réalisées lors de cette thèse :

1. Communication orale "Prédictions géostatistiques avec des données censurées : application à la caractérisation radiologique pour le démantèlement des installations nucléaires" aux 52èmes Journées de Statistique de la SFdS (online conference), Nice, mai 2020, https://jds2020.sciencesconf.org/, [Wieskotten et al., 2020] ;
2. Présentation d'un poster aux PhD Days Mascot-Num 2020, Grenoble, septembre 2020, https://www.gdr-mascotnum.fr/mascotphd20.html ;
3. Présentation orale aux Journées des Doctorants du DMRC 2021, CEA Marcoule ;
4. Présentation orale au séminaire de la compétence "Proba/Stats" d'EDF R&D, EDF Lab Chatou, juin 2021 ;
5. Communication orale "Study of the effectiveness of Bayesian kriging for the decommissioning and dismantling of nuclear sites" à la conférence ENBIS-21 (online conference), septembre 2021, https://conferences.enbis.org/event/11/ ;
6. Présentation d'un poster aux Journées de Géostatistique 2021, Fontainebleau, septembre 2021, https://geostat21.sciencesconf.org/ ;
7. Communication orale "A comparison between Bayesian and ordinary kriging based on validation criteria : application to radiological characterisation" à la 14th International Conference on Geostatistics for Environmental Applications (geoENV2022), Parma, Italie, juin 2022, https://2022.geoenvia.org ;
8. Soumission d'un article en revue internationale à comité de lecture, 2022, [Wieskotten et al., 2022], https://hal.archives-ouvertes.fr/hal-03806713, cf. Annexe B.1 ;
9. Présentation orale aux Journées de la Recherche de l'ISEC 2022, CEA Marcoule ;
10. Présentation d'un poster au workshop "Statistical methods for safety and decommisionning" du GdR Mascot-Num, Avignon, novembre 2022, https://www.gdr-mascotnum.fr/nov22.html

Bibliographie

[AIEA, 2017] AIEA (2017). Release of Sites from Regulatory Control on Termination of Practices. Text WS-G-5.1, AIEA.

[Akaike, 1998] Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In Parzen, E., Tanabe, K., and Kitagawa, G., editors, *Selected Papers of Hirotugu Akaike*, Springer Series in Statistics, pages 199–213. Springer, New York, NY.

[Al-Mudhafar, 2019] Al-Mudhafar, W. J. (2019). Bayesian kriging for reproducing reservoir heterogeneity in a tidal depositional environment of a sandstone formation. *Journal of Applied Geophysics*, 160 :84–102.

[Allard et al., 2015] Allard, D., Senoussi, R., and Porcu, E. (2015). Anisotropy models for spatial data. *Mathematical Geosciences*, 48(3) :24 p.

[Amgarou, 2017] Amgarou, K. (2017). Criteria for characterization, RN & materials-cartography. Technical Report D2.2, INSIDER European Project.

[Amgarou et al., 2018] Amgarou, K., Herranz, M., Csöme, C., and Aspe, F. (2018). Inventory of existing methodologies for constrained environments. Technical Report D5.1, INSIDER European Project.

[Ankenman et al., 2010] Ankenman, B., Nelson, B. L., and Staum, J. (2010). Stochastic Kriging for Simulation Metamodeling. *Operations Research*, 58(2) :371–382.

[ASN, 2016a] ASN (2016a). Guide n°14-Assainissement des structures dans les installations nucléaires de base. Technical Report 14, ASN.

[ASN, 2016b] ASN (2016b). Guide n°24-Gestion des sols pollués par les activités d'une installation nucléaire de base. Technical Report 24, ASN.

[ASN, 2016c] ASN (2016c). Guide n°6-Arrêt définitif, démantèlement et déclassement des installations nucléaires de base. Technical Report 6, ASN.

[ASN, 2022] ASN (2022). Le démantèlement des installations nucléaires. URL https ://www.asn.fr/l-asn-informe/dossiers-pedagogiques/le-demantelement-des-installations-nucleaires.

[Bachoc, 2013] Bachoc, F. (2013). Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66 :55–69.

[Baillargeon, 2005] Baillargeon, S. (2005). *Le krigeage : revue de la théorie et application à l'interpolation spatiale de données de précipitations*. Mémoire, Université Laval.

[Bayes, 1763] Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53 :370–418.

[Belbeze et al., 2013] Belbeze, S., Demougeot-Renard, H., Faucheux, C., and Jeannee, N. (2013). Retour d'expérience critique sur l'utilisation de méthodes géostatistiques pour la caractérisation des sites et sols pollués. Technical Report 11-0514/1A, RECORD.

[Berger, 1985] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer New York, New York, NY.

[Berger et al., 2001] Berger, J. O., Pericchi, L. R., Ghosh, J. K., Samanta, T., De Santis, F., Berger, J. O., and Pericchi, L. R. (2001). Objective Bayesian methods for model selection : Introduction and comparison. *Lecture Notes-Monograph Series*, pages 135–207.

[Binois et al., 2018] Binois, M., Gramacy, R. B., and Ludkovski, M. (2018). Practical heteroscedastic gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4) :808–821.

[BIPM et al., 2008] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML (2008). *Evaluation des données de mesure - Guide pour l'expression de l'incertitude de mesure*. JCGM.

[BIPM et al., 2012] BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML (2012). *Vocabulaire international de métrologie - Concepts fondamentaux et généraux et termes associés (VIM)*. JCGM, 3rd edition.

[Blangiardo et al., 2013] Blangiardo, M., Cameletti, M., Baio, G., and Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology*, 4 :33–49.

[Blatman et al., 2017] Blatman, G., Delage, T., Iooss, B., and Pérot, N. (2017). Probabilistic risk bounds for the characterization of radiological contamination. *EPJ Nuclear Sci. Technol.*, 3(23).

[Box and Cox, 1964] Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society : Series B (Methodological)*, 26(2) :211–243.

[Carlin and Louis, 2009] Carlin, B. and Louis, T. (2009). Bayesian Methods for Data Analysis, Third Edition.

[Chauvet, 2006] Chauvet, P. (2006). *Aide-mémoire de géostatistique linéaire*. Les Presses de L'Ecole des mines de Paris.

[Chevet et al., 2020] Chevet, P.-F., Duranthon, J.-P., and Follenfant, P. (2020). Le démantelement des installations nucléaires. Technical Report 012756-01, CGEDD.

[Chilès and Delfiner, 2012] Chilès, J.-P. and Delfiner, P. (2012). *Geostatistics : Modeling Spatial Uncertainty*. Wiley Series In Probability and Statistics. Wiley, second edition edition.

[Cressie, 1986] Cressie, N. (1986). Kriging Nonstationary Data. *Journal of the American Statistical Association*, 81(395) :625–634. Publisher : [American Statistical Association, Taylor & Francis, Ltd.].

[Cressie, 1993] Cressie, N. (1993). *Statistics for spatial data*. John Wiley & Sons.

[Crozet et al., 2015] Crozet, M., Rivier, C., and Puydarrieux, S. (2015). Cumul de mesures. *Techniques de l'ingénieur*.

[Currie, 1968] Currie, L. A. (1968). Limits for qualitative detection and quantitative determination. Application to radiochemistry. *Analytical chemistry*, 40(3) :586–593.

[De Oliveira, 2005] De Oliveira, V. (2005). Bayesian Inference and Prediction of Gaussian Random Fields Based on Censored Data. *Journal of Computational and Graphical Statistics*, 14(1) :95–115.

[De Oliveira and Ecker, 2002] De Oliveira, V. and Ecker, M. D. (2002). Bayesian hot spot detection in the presence of a spatial trend : application to total nitrogen concentration in Chesapeake Bay. *Environmetrics*, 13(1) :85–101.

[De Oliveira et al., 1997] De Oliveira, V., Kedem, B., and Short, D. A. (1997). Bayesian Prediction of Transformed Gaussian Random Fields. *Journal of the American Statistical Association*, 92(440) :1422–1433.

[De Risi et al., 2021] De Risi, R., De Luca, F., Gilder, C. E., Pokhrel, R. M., and Vardanega, P. J. (2021). The SAFER geodatabase for the Kathmandu valley : Bayesian kriging for data-scarce regions. *Earthquake Spectra*, 37(2) :1108–1126.

[Demay et al., 2022] Demay, C., Iooss, B., Le Gratiet, L., and Marrel, A. (2022). Model selection based on validation criteria for Gaussian process regression : An application with highlights on the predictive variance. *Quality and Reliability Engineering International*, 38(3) :1482–1500.

[Demongeot et al., 2011] Demongeot, S., Girard, V., Onody, C., Picolo, J.-L., Quinio, C., Scapolan, S., and Tillie, J.-L. (2011). Guide de bonnes pratiques des laboratoires de mesure de radioactivité en situation post-accidentelle. Technical Report 2011-02, IRSN.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1) :1–22.

[DEN, 2017] DEN, C. (2017). *L'assainissement-démantèlement des installations nucléaires*. Monographie CEA. Le Moniteur.

[Desnoyers, 2010] Desnoyers, Y. (2010). *Approche méthodologique pour la caractérisation géostatistique des contaminations radiologiques dans les installations nucléaires.* Thèse, École Nationale Supérieure des Mines de Paris.

[Deutsch, 1989] Deutsch, C. (1989). DECLUS : a fortran 77 program for determining optimum spatial declustering weights. *Computers & Geosciences*, 15(3) :325–332.

[Diggle and Ribeiro, 2002] Diggle, P. J. and Ribeiro, P. J. (2002). Bayesian Inference in Gaussian Model-based Geostatistics. *Geographical and Environmental Modelling*, 6(2) :129–146.

[Diggle and Ribeiro, 2007] Diggle, P. J. and Ribeiro, P. J. (2007). *Model-based Geostatistics.* Springer Series in Statistics. Springer.

[Ecker and Gelfand, 1997] Ecker, M. D. and Gelfand, A. E. (1997). Bayesian Variogram Modeling for an Isotropic Spatial Process. *Journal of Agricultural, Biological, and Environmental Statistics*, 2(4) :347–369.

[Efron, 1979] Efron, B. (1979). Bootstrap Methods : Another look at the Jackknife. *The Annals of Statistics*, 7(1) :1–26.

[Emery, 2001] Emery, X. (2001). *Géostatistique Linéaire.* Centre de Géostatistique, Ecole des Mines de Paris.

[EPRI, 2016] EPRI (2016). Guidance for Using Geostatistics in Developing a Site Final Status Survey Program for Plant Decommissioning. Technical Report 3002007554, EPRI.

[Escofier and Pagès, 2008] Escofier, B. and Pagès, J. (2008). *Analyses factorielles simples et multiples : objectifs, méthodes et interprétation.* Dunod, Paris.

[Foos, 1994] Foos, J. (1994). *Manuel de la RADIOACTIVITE à l'usage des utilisateurs : Tome 2, les désintégrations radioactives.* Formascience.

[Fridley, 2003] Fridley, B. L. (2003). *Data augmentation for the handling of censored spatial data.* Thèse, Iowa State University.

[Fridley and Dixon, 2007] Fridley, B. L. and Dixon, P. (2007). Data augmentation for a Bayesian spatial model involving censored observations. *Environmetrics*, 18(2) :107–123.

[Gelman et al., 1996] Gelman, A., Roberts, G., and Gilks, W. (1996). Efficient Metropolis jumping rules. *Bayesian Statistics*, pages 599–608.

[Granier et al., 2017] Granier, G., Aubonnet, E., Courbet, C., Desnoyers, Y., Dubot, D., Fichet, P., Nokhamzon, J.-G., Olivier Dehaye, C., Pillette-Cousin, L., and Mahe, C. (2017). Evaluation de l'état radiologique initial et final d'une installation nucléaire en situation d'assainissement. Technical Report CEA-R-6455, CEA.

[Gu et al., 2018] Gu, M., Wang, X., and Berger, J. O. (2018). Robust Gaussian stochastic process emulation. *The Annals of Statistics*, 46(6A).

[Gupta et al., 2017] Gupta, A., Kamble, T., and Machiwal, D. (2017). Comparison of ordinary and Bayesian kriging techniques in depicting rainfall variability in arid and semi-arid regions of north-west India. *Environmental Earth Sciences*, 76(15) :512.

[Handcock and Stein, 1993] Handcock, M. S. and Stein, M. L. (1993). A Bayesian Analysis of Kriging. *Technometrics*, 35(4) :403–410.

[Hastings, 1970] Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 57(1) :97–109.

[Helbert et al., 2009] Helbert, C., Dupuy, D., and Carraro, L. (2009). Assessment of uncertainty in computer experiments from Universal to Bayesian Kriging. *Applied Stochastic Models in Business and Industry*, 25(2) :99–113.

[Helsel, 2012] Helsel, D. R. (2012). *Statistics for Censored Environmental Data Using Minitab and R.* Statistics in Practice. Wiley, second edition.

[Helsel and Gilliom, 1986] Helsel, D. R. and Gilliom, R. J. (1986). Estimation of Distributional Parameters for Censored Trace Level Water Quality Data : 2. Verification and Applications. *Water Resources Research*, 22(2) :147–155.

[Iooss et al., 2010] Iooss, B., Boussouf, L., Feuillard, V., and Marrel, A. (2010). Numerical studies of the metamodel fitting and validation processes. *International Journal On Advances in Systems and Measurements*, 3 :11–21.

[IRSN, 2020] IRSN (2020). Les démantèlements en cours chez EDF, Areva et au CEA. URL https ://www.irsn.fr/FR/connaissances/Installations_nucleaires/demantelement/demantelement-France-centrales-installations-nucleaires-EDF-recherche-militaire/Pages/2-centrales-installations-nucleaires-en-cours-France.aspx ?dId=6ae3535e-f095-4e8c-9a02-a71f0e893ad7&dwId=c4827b15-a982-4837-a56b-d66c0804e759#.Y3OGiHbMJaQ.

[ISO, 2017] ISO (2017). 18557 Principes de caractérisation des sols, bâtiments et infrastructures contaminés par des radionucléides à des fins de réhabilitation.

[Kaplan and Meier, 1958] Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53(282) :457–481.

[Kim et al., 2020] Kim, J. H., Hornibrook, C., and Yim, M.-S. (2020). The impact of below detection limit samples in residual risk assessments for decommissioning nuclear power plant sites. *Journal of Environmental Radioactivity*, 222 :106340.

[Kitanidis, 1986] Kitanidis, P. (1986). Parameter Uncertainty in Estimation of Spatial Functions : Bayesian Analysis. *Water Resources Research*, 22(4) :499–507.

[Kitanidis and Lane, 1985] Kitanidis, P. K. and Lane, R. W. (1985). Maximum likelihood parameter estimation of hydrologic spatial processes by the Gauss-Newton method. *Journal of Hydrology*, 79(1) :53–71.

[Kleijnen, 2015] Kleijnen, J. P. C. (2015). *Design and Analysis of Simulation Experiments*, volume 230 of *International Series in operations Research and Management Science*. Springer.

[Krige, 1978] Krige, D. G. (1978). *Lognormal-de Wijsian geostatistics for ore evaluation*. Number 1 in Geostatistics. South African Institute of Mining and Metallurgy, Johannesburg.

[Krivoruchko and Gribov, 2019] Krivoruchko, K. and Gribov, A. (2019). Evaluation of empirical Bayesian kriging. *Spatial Statistics*, 32 :100368.

[Kuhn and Lavielle, 2004] Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM : Probability and Statistics*, 8 :115–131.

[Le Gratiet, 2013] Le Gratiet, L. (2013). *Multi-fidelity Gaussian process regression for computer experiments*. Thèse, Université Paris-Diderot - Paris VII.

[Lindgren and Rue, 2015] Lindgren, F. and Rue, H. (2015). Bayesian Spatial Modelling with R-INLA. *Journal of Statistical Software*, 63 :1–25.

[Marrel et al., 2008] Marrel, A., Iooss, B., Van Dorpe, F., and Volkova, E. (2008). An efficient methodology for modeling complex computer codes with Gaussian processes. *Computational Statistics and Data Analysis*, 52 :4731–4744.

[Matheron, 1970] Matheron, G. (1970). *La théorie des variables régionalisées*. Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau. ENSMP.

[McLachlan and Krishnan, 1997] McLachlan, G. J. and Krishnan, T. (1997). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

[Metropolis et al., 1953] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6) :1087–1092.

[Militino and Ugarte, 1999] Militino, A. F. and Ugarte, M. D. (1999). Analyzing Censored Spatial Data. *Mathematical Geology*, 31(5) :551–561.

[Muré, 2018] Muré, J. (2018). *Objective Bayesian analysis of Kriging models with anisotropic correlation kernel*. Thèse, Université Sorbonne Paris Cité.

[Ng and Yin, 2012] Ng, S. H. and Yin, J. (2012). Bayesian Kriging Analysis and Design for Stochastic Simulations. *ACM Transactions on Modeling and Computer Simulation*, 22(3) :17 :1–17 :26.

[Nutt and Wallis, 2004] Nutt, W. T. and Wallis, G. B. (2004). Evaluation of nuclear safety from the outputs of computer codes in the presence of uncertainties. *Reliability Engineering & System Safety*, 83(1) :57–77.

[Omre, 1987] Omre, H. (1987). Bayesian kriging—Merging observations and qualified guesses in kriging. *Mathematical Geology*, 19(1) :25–39.

[Omre and Halvorsen, 1989] Omre, H. and Halvorsen, K. B. (1989). The Bayesian bridge between simple and universal kriging. *Mathematical Geology*, 21(7) :767–786.

[Ordoñez et al., 2018] Ordoñez, J. A., Bandyopadhyay, D., Lachos, V. H., and Cabral, C. R. B. (2018). Geostatistical estimation and prediction for censored responses. *Spatial Statistics*, 23 :109–123.

[Parzen, 1962] Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics*, 33(3) :1065–1076.

[Pérot et al., 2017] Pérot, N., Desnoyers, Y., Augé, G., Aspe, O., Boden, S., Bart, S., Sevbo, O., Nitzsche, O., and de Groot, J. (2017). Sampling strategy State-of-the-art report. Technical Report D3.1, INSIDER European Project.

[Pérot and Iooss, 2008] Pérot, N. and Iooss, B. (2008). Quelques problématiques d'échantillonnage statistique pour le démantèlement d'installations nucléaires. In *Conférence $\lambda\mu$16, Avignon, France.*

[Rivoirard, 2003] Rivoirard, J. (2003). *Cours de Géostatistique Multivariable.* Centre de Géostatistique, Ecole des Mines de Paris.

[Robert and Casella, 2004] Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods.* Springer Texts in Statistics. Springer New York, New York, NY.

[Rogiers et al., 2018] Rogiers, B., Boden, S., Pérot, N., Desnoyers, Y., Sevbo, O., and Nitzsche, O. (2018). Report on the sampling strategy development. Technical Report D3.2, INSIDER European Project.

[Sakia, 1992] Sakia, R. (1992). The Box-Cox Transformation Technique : A Review. *The Statistician*, 41.

[Saporta, 1990] Saporta, G. (1990). *Probabilités, analyse des données et statistique.* Editions Technip.

[Savage, 1961] Savage, I. (1961). Probability Inequalities of the Tchebycheff Type. *Journal of Research of the National Bureau of Standards-B. Mathematics and Mathematical Physics*, 65B(3) :211–226.

[Schwarz, 1978] Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6(2) :461–464.

[Tadayon, 2017] Tadayon, V. (2017). Bayesian Analysis of Censored Spatial Data Based on a Non-Gaussian Model. *Journal of Statistical Research of Iran*, 13 :155–180.

[Tanner and Wong, 1987] Tanner, M. A. and Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82(398) :528–540.

[Toscas, 2010] Toscas, P. J. (2010). Spatial modelling of left censored water quality data. *Environmetrics*, 21(6) :632–644.

[van Dyk and Meng, 2001] van Dyk, D. A. and Meng, X.-L. (2001). The Art of Data Augmentation. *Journal of Computational and Graphical Statistics*, 10(1) :1–50.

[Wackernagel, 1993] Wackernagel, H. (1993). *Cours de Géostatistique Multivariable.* 4ième édition. Centre de Géostatistique, Ecole des Mines de Paris.

[Wang et al., 2012] Wang, J.-F., Stein, A., Gao, B.-B., and Ge, Y. (2012). A review of spatial sampling. *Spatial Statistics*, 2 :1–14.

[Watson, 1995] Watson, G. N. (1995). *A Treatise on the Theory of Bessel Functions.* Cambridge University Press.

[Webster and Oliver, 2007] Webster, R. and Oliver, M. A. (2007). *Geostatistics for environmental scientists.* John Wiley & Sons.

[Wieskotten et al., 2022] Wieskotten, M., Crozet, M., Iooss, B., Lacaux, C., and Marrel, A. (2022). A comparison between Bayesian and ordinary kriging based on validation criteria : application to radiological characterisation. *Preprint URL https ://hal.archives-ouvertes.fr/hal-03806713/document.*

[Wieskotten et al., 2020] Wieskotten, M., Crozet, M., Iooss, B., Lacaux, C., and Pérot, N. (2020). Prédictions géostatistiques avec des données censurées : Application à la caractérisation radiologique pour le démantèlement des installations nucléaires. In *jds2020 : 52èmes Journées de Statistiques de la Société Française de Statistique : Recueil des soumissions*, pages 782–787, Nice.

[Wilks, 1948] Wilks, S. (1948). Order Statistics. *Bulletin of the American Mathematical Society*, 54(1) :6–50.

[Wilks, 1941] Wilks, S. S. (1941). Determination of Sample Sizes for Setting Tolerance Limits. *Annals of Mathematical Statistics*, 12(1) :91–96.