prédictions géostatistiques avec des données censurées : application à la caractérisation radiologique pour le démantèlement des installations nucléaires

Martin Wieskotten, Marielle Crozet, Bertrand Iooss, Céline Lacaux, Nadia Perot

Prédictions géostatistiques avec des données censurées : application à la caractérisation radiologique pour le démantèlement des installations nucléaire

Martin Wieskotten[1,4], Marielle Crozet[2], Bertrand Iooss[3], Céline Lacaux[4], Nadia Pérot[1]

[1] *CEA, DER, SESI, Cadarache, France et nadia.perot@cea.fr/martin.wieskotten@cea.fr*
[2] *CEA, DEN, DMRC, Univ. Montpellier, Marcoule, France et marielle.crozet@cea.fr*
[3] *EDF R&D, 6 quai Watier, 78400, Chatou, France et bertrand.iooss@edf.fr*
[4] *LMA Université d'Avignon, 84029, Avignon, France et celine.lacaux@univ-avignon.fr*

Résumé. La caractérisation spatiale de la radioactivité ou de radionucléides d'intérêt est un sujet primordial dans les problématiques d'assainissement et de démantèlement d'infrastructures nucléaires. Les statistiques spatiales offrent des solutions pour prédire la présence de contamination, mais ne permettent pas toujours de gérer des données censurées. Ces données correspondent à des résultats de mesure inférieure à la limite de détection et sont souvent supprimées ou remplacées par la valeur de la limite de détection. Ces pratiques introduisent un biais dans les prédictions en modifiant la variance et la moyenne du jeu de données étudié. Une alternative aux méthodes de remplacement est la méthode du package R CensSpatial qui à l'aide de calculs de maximum de vraisemblance permet la prise en compte de données censurées. Notre objectif est de comparer les méthodes courantes de remplacement de données à la méthode CensSpatial développée par Ordoñez et al. (2018) sur des mesures provenant d'un projet du CEA Marcoule.

Mots-clés. Statistique Spatiale, Géostatistique, Données Censurées, Algorithme SAEM, Package CensSpatial

Abstract. The spatial radioactive contamination's characterization is one of the main topic in sanitation and dismantling projects of nuclear infrastructures. Spatial statistics offer solutions for predicting the location of contamination, but can not always take into account censored responses. These responses correspond to measurement inferior to detection threshold and are often discarded or replaced with the value of the detection threshold. These practices insert bias in predictions for they change the variance and the mean of studied data set. An alternative to these practices is the R package CensSpatial's method that allows to take censored responses into account with the use of maximum likelihood calculation. Our goal is to compare the usual practices of replacement to the CensSpatial methods implemented by Ordoñez et al. (2018) on a data set from a project of the CEA Marcoule.

Keywords. Spatial Statistics, Geostatistics, Censored Data, SAEM Algorithm, CensSpatial Package

1 Introduction

La caractérisation spatiale d'une contamination est une étape primordiale dans les projets d'assainissement et de démantèlement des installations nucléaires, en particulier lorsque les déchets sont radiotoxiques et coûteux à traiter. Cette carte de contamination est fondée sur l'analyse statistique de quelques points et est donc incomplète, là où une connaissance complète serait préférable. Les statistiques spatiales offrent des solutions pour reconstituer ce type de phénomène, notamment avec la géostatistique qui utilise les corrélations entre observations pour réaliser des prédictions sur des sites non observés. Néanmoins la géostatistique n'est pas capable de prendre en compte les données censurées, données qui sont très répandues en assainissement/démantèlement. Elles correspondent à des valeurs de mesure inférieures au seuil de décision et qui sont rendues comme inférieures à la limite de détection, sans valeur numérique associée, comme expliqué par Rivier et Crozet (2014). Ces données apportent des informations limitées mais non négligeables. Les pratiques actuelles en géostatistique remplacent simplement ces données par une valeur arbitraire, ce qui selon la valeur choisie introduit un biais modifiant les prédictions finales. Le package R CensSpatial développé par Ordoñez et al. (2018) permet à l'aide d'un algorithme SAEM (Stochastic Aproximation Expectation Maximisation) d'estimer les valeurs censurées et de limiter le biais lors de la réalisation de prédictions. Pour vérifier les avantages de cette méthode par rapport aux méthodes de remplacement, nous avons choisi de comparer la méthode implémentée par CensSpatial avec une méthode de remplacement par 0, et une autre de remplacement par la limite de détection.

Pour réaliser cette comparaison, nous avons choisi un jeu de données provenant d'un projet du CEA Marcoule et correspondant à 70 mesures surfaciques d'activité (en Bq/cm^2). Afin de se ramener à l'hypothèse de normalité, une transformation des données est réalisée par un logarithme translaté : $Y = \ln(1 + X)$. Pour faciliter les comparaisons, des points supplémentaires ont été obtenus à l'aide d'un krigeage simple. Un maillage de 100×100 points est ainsi créé et permet d'augmenter le nombre de points sur lesquels nous pouvons travailler. Dans la suite ces points supplémentaires seront considérés comme des observations, et seront donc considérés comme vrais. De plus on fait l'hypothèse que la limite de détection est la même pour chaque observation, ce qui implique un protocole exactement identique lors de la mesure d'activité surfacique en chaque point. Cette hypothèse est raisonnable puisque ce jeu de données correspond à un échantillonnage in-situ effectué avec le même appareil de mesure.

Considérons un processus gaussien réel $Z(x), x \in D$, avec D la région étudiée ($D \subset \mathbb{R}^2$), stationnaire à l'ordre 2, dont la moyenne μ et la covariance C vérifient:

$$\mu = \mathbf{E}[Z(x)] \quad , \quad C(Z(x), Z(x\prime)) = C(x - x\prime)$$

. On observe les réalisations de ce processus $\mathbf{Z} = (Z(x_1), Z(x_2), ..., Z(x_n))$ en des points connus $x_i, i = 1, ..., n$, correspondant ainsi à n variables aléatoires. L'expression de la matrice de covariance de ces variables aléatoires $Z(x_i)$ peut s'écrire $\mathbf{\Sigma} = \tau^2\mathbf{I}_n + \sigma^2\mathbf{R}(\phi)$

d'après Diggle et Ribeiro (2007), τ^2 correspondant au terme de pépite, σ^2 à la variance, ϕ à la portée et le terme $\mathbf{R}$ à une fonction vérifiant plusieurs conditions détaillées par Chilès et Delfiner (1999). Ici nous avons choisi une fonction exponentielle isotrope en accord avec le modèle utilisé pour construire les points supplémentaires par krigeage simple:

$$\mathbf{R}(\phi) = [R(\phi, \|x_i - x_j\|)] = \left[\exp\left(-\frac{\|x_i - x_j\|}{\phi}\right)\right],\ i = 1, ..., n \text{ et } j = 1, ..., n$$

Par soucis de simplicité, nous considérons dans la suite que $\tau^2 = 0$, ce qui revient à considérer qu'il n'y a aucune incertitude de mesure. Cette hypothèse a pour but de simplifier la comparaison des méthodes. De plus elle est en accord avec la construction des points supplémentaires ce qui la rend donc vraie, mais en pratique elle n'est pas raisonnable. Nous reviendrons sur ce point dans la conclusion.
L'intégration des données censurées se fait en réorganisant les expressions du vecteur des données et de la covariance en ordonnant les données observées (d'indice o) et les données censurées (d'indice c), avec *vec* la fonction qui concatène les matrices de colonnes identiques:

$$\mathbf{Z} = vec(\mathbf{Z}^o, \mathbf{Z}^c), \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}^{oo} & \mathbf{\Sigma}^{oc} \\ \mathbf{\Sigma}^{co} & \mathbf{\Sigma}^{cc} \end{bmatrix}$$

$\mathbf{Z}$ correspond aux données observées complétées par les données censurées. Cette mise en forme sera utile lors de l'application de l'algorithme SAEM, notamment lors de son étape E. L'algorithme est décrit dans le prochain paragraphe. On note $\theta = (\mu, \sigma^2, \phi)$ l'ensemble des paramètres à estimer. En notant K une constante indépendante de θ et $\boldsymbol{\mu} = [\mu]$, l'expression de la fonction de vraisemblance l_c utilisée pour l'estimation des paramètres de la covariance et de la moyenne est la suivante:

$$l_c(\theta) \propto -\frac{1}{2}\left[\log(|\mathbf{\Sigma}|) + (\mathbf{Z} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{Z} - \boldsymbol{\mu})\right] + K$$

L'optimisation de cette fonction de vraisemblance se fait à l'aide de l'algorithme SAEM détaillé par Delyon et al. (1999). Cet algorithme est une variante de l'algorithme EM : "Expectation Maximisation" proposé par Dempster et al. (1977), qui est un algorithme itératif dont l'objectif est l'estimation des paramètres d'une loi de probabilité par maximum de vraisemblance. Sa particularité est sa capacité à prendre en compte une variable latente, dans notre cas les données censurées, pour réaliser cette estimation. Ces algorithmes se déroulent en 2 temps : la première étape estime l'espérance de la fonction de vraisemblance conditionnellement aux données (étape E) et la seconde estime les paramètres permettant de maximiser cette vraisemblance (étape M). Ordoñez et al. (2018) détaillent l'algorithme SAEM de la façon suivante:
Etape E-1 : On effectue un tirage aléatoire d'un vecteur $\mathbf{Z}^c$ représentant les données censurées d'une loi normale tronquée sur l'intervalle $[0; V_{lim}]$. On forme ainsi le vecteur $\mathbf{Z}^k = vec(\mathbf{Z}^c, \mathbf{Z}^o)$ qui contient le tirage des nouvelles données (censurées) et les données

observées. On répète ce tirage un nombre M de fois pour obtenir une séquence de vecteurs aléatoires: $(\mathbf{Z}^{(k,l)})_{l=1,\ldots,M}$.

Etape E-2 : On estime la valeur de l'espérance de la fonction de vraisemblance conditionnellement aux paramètres estimés à l'étape précédente et aux données complètes à l'aide d'une approximation stochastique comme décrit par Ordoñez et al. (2018).

Etape M : On maximise la log-vraisemblance et on obtient une nouvelle estimation des paramètres $\theta^{(k+1)}$ contenant la moyenne et la variance du processus ainsi que les paramètres de la covariance. Ces étapes sont répétées jusqu'à ce que la différence en valeur absolue des fonctions de vraisemblance aux étapes (k) et $(k+1)$ soit inférieure à un seuil donné. La convergence de l'algorithme est prouvée sous des hypothèses générales par Wu (1983).

La prédiction de la variable régionalisée en des points non observés est ensuite faite par krigeage simple.

2 Protocole des tests numériques

La Figure 1 correspond à la carte de la contamination (après transformation par le logarithme translaté, sans unité) obtenue avec les données construites initialement. Cette carte correspond à la "réalité" par rapport à laquelle nous comparons les 3 méthodes. La Figure 2 donne un exemple de carte de prédiction de la contamination à l'aide de la méthode CensSpatial pour une limite de détection fixée à 0.3.

Les calculs réalisés se font en 2 temps, estimation des paramètres de la covariance (et

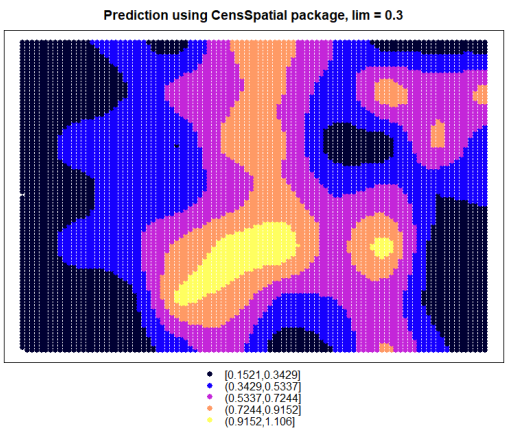

Figure 1: Carte initiale des données construites

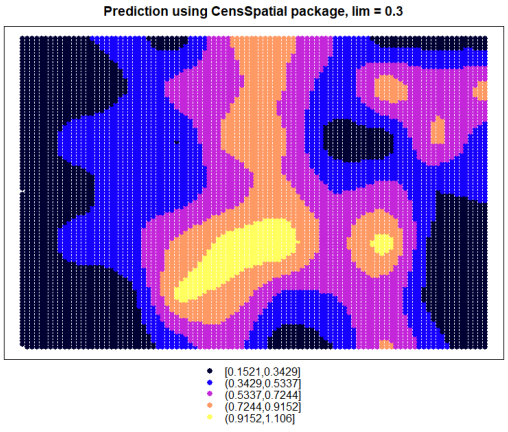

Figure 2: Carte des prédictions par CensSpatial

des paramètres de la distribution pour la méthode CensSpatial) puis prédictions selon un maillage prédéfini par krigeage simple.

La moyenne des erreurs $m_{erreurs}$ est calculée pour différentes limites de détection correspondant à des pourcentages de données censurées différents, et ce pour chaque méthode. Cette moyenne est calculée en faisant la différence entre la prédiction de la méthode z_{predit} et la valeur vraie z_{vrai} et ce en chaque point de notre maillage de n_{predit} points : $m_{erreurs} = \sum \frac{z_{predit} - z_{vrai}}{n_{predit}}$.

3 Résultats

La Figure 4 représente les prédictions en fonction des valeurs vrais (observations) pour une limite de détection égale à 0.4, avec la droite identité correspondant à une prédiction parfaite. Cette figure permet donc d'évaluer les tendances qui peuvent apparaître dans les prédictions par les 3 méthodes étudiées. Les différentes erreurs moyennes obtenues ont été représentées sur la Figure 3. Cette figure permet d'étudier l'évolution de la moyenne des erreurs avec la variation du pourcentage de données censurées.

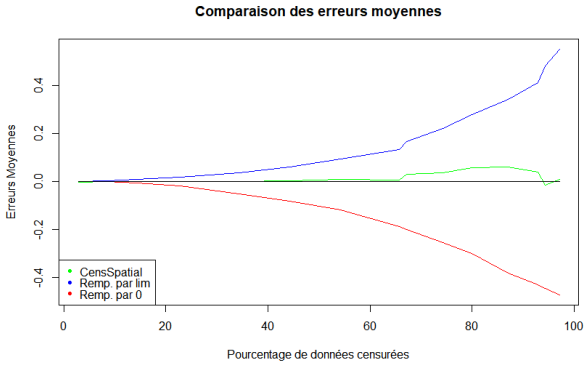

Figure 3: Moyenne des erreurs selon la limite de détection

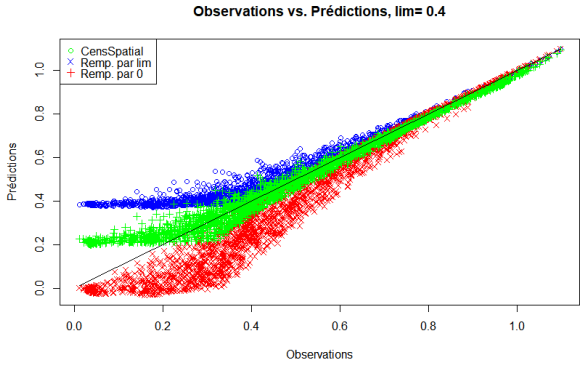

Figure 4: Graphe des Prédictions en fonction des observations

Pour comparer les différents résultats, nous avons isoler plusieurs éléments importants: les paramètres estimés de la covariance (variance et portée) et la qualité des prédictions faites, et ce selon le pourcentage de données censurées.
On observe pour le remplacement par la limite une sur-estimation de la variance et pour le remplacement par 0 une sous-évaluation de la variance de la variable régionalisée quand le pourcentage de données censurées augmente. La méthode CensSpatial estime une variance proche de celle réellement observée. De plus cette variance est plus stable lorsque le pourcentage de données censurées varie. La portée suit des évolutions similaires mais inversées : elle est sous-estimée par la méthode de remplacement par 0 et et sur-estimée par la méthode de remplacement par la limite.

Lorsque le nombre de données en limite de détection est faible, les 3 méthodes donnent des résultats quasi identiques. En effet la technique de prédiction (krigeage simple) étant

commune aux 3 méthodes et les jeux de données étant proches pour les 3 méthodes, les prédictions sont toutes quasi parfaites. Par contre lorsque la limite de détection augmente, les 3 méthodes offrent des résultats assez différents, la méthode de remplacement par 0 sous-estimant les valeurs réelles et la méthode de remplacement par la limite surestimant ces valeurs. La méthode CensSpatial offre un compromis aux 2 autres méthodes en générant des estimations comprises entre 0 et la limite de détection (conditionnellement aux données). La Figure 4 indique également que la méthode CensSpatial a tendance à surestimer les valeurs faibles, ce qui implique tout de même un gain en terme de sûreté dans le cadre de l'assainissement démantèlement (mais une perte en terme d'optimisation des coûts).

En conclusion, cette application de la méthode développée par Ordoñez et al. montre l'intérêt de la prise en compte des données censurées lors de réalisation de prédictions ou l'évaluation d'une structure de covariance. La méthode CensSpatial fournit des résultats similaires aux méthodes de remplacement pour un faible nombre de données censurées mais son intérêt se fait ressentir pour des pourcentages de données censurées supérieurs à 15%. Cette étude n'est cependant qu'une première étape. Comme nous l'avons déjà évoqué, l'hypothèse d'incertitude de mesure nulle n'est pas raisonnable, en particulier avec un échantillonnage in-situ où le cadre expérimental est moins contrôlé qu'en laboratoire. La suite envisagée est d'étudier l'effet de cette incertitude sur les modélisations et les méthodes comme CensSpatial.

Bibliographie

Ordoñez et al. (2018), Geostatistical estimation and prediction for censored responses , *Spatial Statistics*, 23, 109-123.

Rivier C., Crozet M. (2014) Limite de détection de méthodes d'analyse et termes apparentés, *Techniques de l'ingénieur*, p262

Diggle P.J. et Ribeiro P.J. (2007). Model-based Geostatistics, *Springer Series in Statistics*, Springer.

Chilès J.P. et Delfiner P. (1999). Geostatistics : Modeling Spatial Uncertainty, *Wiley Series in Probability and Statistics*, Wiley.

Delyon B., Lavielle M., Moulines E., (1999). Convergence of a stochastic approximation version of the EM algorithm, Ann. Statist. 27(1), 94-128.

Dempster A., Laird N., Rubin D. (1977), Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society*, 39, 1-38.

Wu C.J. (1983) On the convergence properties of the EM algorithm, Ann. Statist., 11(1)

Ordoñez A., Galarza C.E., Lachos V.H. (2017), CensSpatial : Censored Spatial Models. R package version 2.1 URL https://CRAN.R-project.org/package=CensSpatial