Données Censurées dans les Statistiques Spatiales : Comparaison de Différentes Approches

M. Wieskotten[1,3]*

Equipe encadrante: M. Crozet[1] , B. Iooss[2] , C. Lacaux[3]

[1] CEA/DES/ISEC/DMRC/CETAMA [2] EDF R&D [3] LMA Univ. Avignon *martin.wieskotten@cea.fr

Contexte and Problématique:

La caractérisation spatiale des radionucléides est l'un des défis de l'assainissement/démantèlement de sites nucléaires. Pour établir cette carte de radioactivité, la géostatistique est une méthode classique, mais nécessite des observations avec des valeurs et positions connues. Si la valeur mesurée est trop faible (inférieure au seuil de détection voire à la limite de détection LD), la donnée est généralement censurée. La valeur mesurée est alors indisponible et appartient à un intervalle compris entre $-\infty$ et la limite de détection. Pour gérer ces données censurées, des méthodes de remplacement sont souvent utilisées : la donnée censurée est alors remplacée par une valeur arbitraire. Ces méthodes introduisent un biais dans les estimations et prédictions des modèles statistiques (voir [1],[2]). Notre objectif ici est de comparer différentes méthodes de traitement des données censurées, dont une approche bayésienne : l'augmentation de données adaptée de [3].

Notations :

$Z, \boldsymbol{Z}$	Champ aléatoire, Vecteur de observations	$E[\boldsymbol{Z}] = \boldsymbol{\mu}$	Espérance du vecteur d'observations.
x_i	Coordonnées du $i^{ème}$ point.	$Var[\boldsymbol{Z}] = \sigma^2\boldsymbol{\Sigma} + \tau^2\boldsymbol{I}$	Matrice de covariance et variance.
x_0	Point non-observé à prédire.	$\boldsymbol{\theta}' = (\mu, \sigma^2, \phi, \tau^2)$	Ensemble des paramètres du modèle.
$\hat{z}_i$	Prédiction au point $\boldsymbol{x}_i$ de la base de test.	$\boldsymbol{k} = (Cov(Z(\boldsymbol{x}_0), Z(\boldsymbol{x}_i)))_i$	Vecteur de covariance entre les données et le point $\boldsymbol{x}_0$.
${s_i}^2$	Variance de prédiction au point $\boldsymbol{x}_i$ de la base de test.	α	Niveau de confiance de l'intervalle de confiance (IC).
$\hat{q}_{\epsilon,i}$	Estimation du quantile d'ordre ϵ au point $\boldsymbol{x}_i$ de la base de test.		

Méthodes de traitement des données censurées:

- ***Méthode de remplacement :***

  Les données censurées sont simplement remplacées par une valeur arbitraire : 0, LD ou LD/2. Le reste de l'analyse est réalisé de manière classique.

- ***Méthode Monte-Carlo :***

  Les paramètres sont estimés sans les données observées. Ensuite les données censurées sont simulées N fois (avec tirage aléatoire). Les paramètres sont ré-estimés pour chacune des simulations et la moyenne est ensuite prise pour faire un krigeage ordinaire avec les données non-censurées.

- ***Méthode CensSpatial [3]:***

  Mise en place d'un algorithme EM (Espérance Maximisation): Les données censurées sont ré-estimées à chaque étape par krigeage puis tirées aléatoirement. Ces étapes sont répétées jusqu'à convergence.

- ***Méthode d'Augmentation de données (adaptée de [4]):***

  Mise en place d'un algorithme MCMC (Monte Carlo Markov Chain) : Une chaîne de Markov est construite en tirant de manière itérative les données censurées, puis les différents paramètres du modèle. Un échantillonneur de Gibbs vient échantillonner les données censurées, la moyenne et la variance, tandis que la portée et l'effet de pépite sont échantillonnés par Metropolis-Hastings.

Modèle:

$$\boldsymbol{Z} \sim \mathcal{N}(\mu\mathbf{1}, \sigma^2\boldsymbol{\Sigma} + \tau^2\boldsymbol{I})$$

-Modèle exponentiel pour la covariance

-Champ aléatoire stationnarité d'ordre 2

Critères de Validation (dérivés de [5]):

- $Q^2 = 1 - \frac{\sum_{i=1}^{n}(z(\boldsymbol{x}_i) - \hat{z}_i)^2}{\sum_{i=1}^{n}(z(\boldsymbol{x}_i) - \hat{\mu})^2}$

  où $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} z(\boldsymbol{x}_i)$
- $PVA = \left|\log\left(\frac{1}{n}\sum_{i=1}^{n}\frac{(z(\boldsymbol{x}_i) - \hat{z}_i)^2}{{s_i}^2}\right)\right|$
- $PIA = \left|\log\left(\frac{1}{n}\sum_{i=1}^{n}\frac{(z(\boldsymbol{x}_i) - \hat{z}_i)^2}{(\hat{q}_{0,31,i} - \hat{q}_{0,69,i})^2}\right)\right|$
- $\Delta_\alpha = \frac{1}{n}\sum_{i=1}^{n}\phi_i\,,\ \phi_i = \begin{cases} 1 \text{ si } z(\boldsymbol{x}_i) \in CI_{\alpha,i} \\ 0 \text{ sinon} \end{cases}$

  où $CI_{\alpha,i} = \left[\hat{q}_{\frac{1-\alpha}{2},i}; \hat{q}_{\frac{1+\alpha}{2},i}\right]$
- $MSE\alpha = \frac{1}{m}\sum_{i}^{m}(\Delta_{\alpha_i} - \alpha_i)^2$

Protocole:

1. 225 données sont simulées selon un processus gaussien avec :

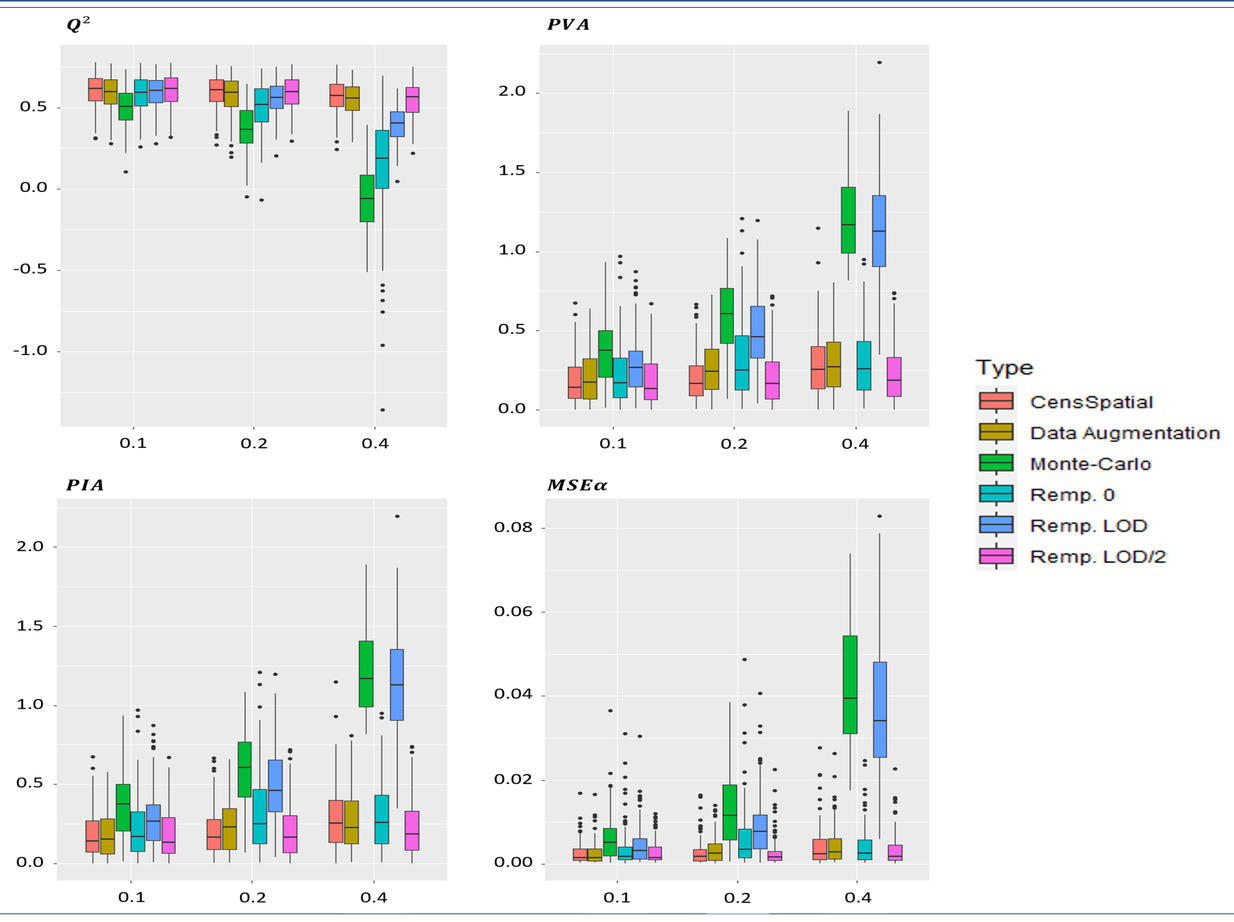
$$\mu = 3, \sigma^2 = 2, \phi = 2.5, \tau^2 = 0.2$$

2. Séparation en 2 : une base de données et une base d'apprentissage.

3. Pour un taux de censure donné $s = (0.1, 0.2, 0.4)$, toutes les valeurs inférieures au quantile d'ordre s sont censurées.

4. 100 jeux de données sont ainsi générés, et les critères de validation sont calculés.

Résultats



Conclusion et Interprétation :

-Les méthodes de remplacement sont inefficaces et dépendent de la construction des données censurées. Elles sont donc à éviter le plus possible.

-La méthode de Monte-Carlo est très mauvaise mais peut être améliorée, avec l'utilisation des tirages des données censurées.

-Les méthodes CensSpatial et d'augmentation de données donnent des résultats extrêmement similaires avec de bonnes performances.

Recommandation : L'utilisation des méthodes CensSpatial et d'augmentation de données pour l'assainissement/démantèlement de sites nucléaires sont plus performantes que les méthodes classiques de remplacement.

L'augmentation offre une approche bayésienne tandis que CensSpatial utilise une approche classique.

References:

- [1] Helsel, D.R., 2012. Statistics for Censored Environmental Data Using Minitab and R, Second edition. ed, Statistics in Practice. Wiley.
- [2] Crozet, M., Rivier, C., Puydarrieux, S., 2015. Cumul de mesures. Techniques de l'ingénieur.
- [3] Ordoñez, J.A., Bandyopadhyay, D., Lachos, V.H., Cabral, C.R.B., 2017. Geostatistical estimation and prediction for censored responses. Spatial Statistics 23 109–123.
- [4] Fridley, B.L., Dixon, P., 2006. Data Augmentation for a Bayesian spatial model involving censored observations. Statistics and Probability Commons.
- [5] 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, 1482–1500. https://doi.org/10.1002/qre.2973