Echantillonnage spatial basé sur le krigeage pour la reconstruction de carte d'occurrence

Mathieu Bonneau[1], Nathalie Peyrard[1], Régis Sabbadin[1]

INRA-Toulouse ; UBIA UR875 ; BP 52627 – 31326 Castanet-Tolosan, France
{mathieu.bonneau,nathalie.peyrard,regis.sabbadin}@toulouse.inra.fr

Mots-Clés : *Echantillonnage spatial optimal, krigeage conditionnel.*

1 Motivation, définition du problème

En épidémiologie ou en écologie, le contrôle d'une maladie ou la gestion d'une espèce dans une zone d'étude donnée repose souvent sur l'établissement d'une carte d'occurrence (présence/absence) du phénomène. Une exploration exhaustive de la zone est en général impossible et la carte doit être reconstruite à partir d'un échantillon spatial. Par ailleurs, les observations recueillies peuvent être bruitées. Le problème est alors celui du choix d'un échantillon puis de la reconstruction d'une carte d'occurrence à partir d'observations bruitées et incomplètes. Nous proposons une méthode, basée sur le krigeage, pour la sélection d'un échantillon spatial, adaptée au cas d'un champ à valeurs binaires et à des observations bruitées également binaires (détecté / non détecté). Notre démarche est la suivante : à partir du krigeage, nous définissons la valeur d'un échantillon. Ensuite, nous définissons la question du choix du "meilleur" échantillon comme un problème d'optimisation. Enfin, une carte d'occurrence est reconstruite (toujours par krigeage) à partir des observations résultant de l'échantillon sélectionné et d'un jeu d'observations initiales. L'échantillonnage peut être "statique" ou "adaptatif". Dans ce dernier cas, l'échantillon n'est pas choisi une fois pour toute au début de la campagne mais de manière séquentielle, en prenant en compte les observations intermédiaires pour le choix de l'échantillon à venir. Nous ne présentons ici les aspects méthodologiques que dans le cas statique, les méthodes proposées dans les cas statique et adaptatif reposant sur les mêmes formalisme et heuristique.

2 Krigeage conditionnel

La zone d'étude est représentée par un compact W de $\mathbb{R}^2$. Une carte d'occurence est la réalisation d'un champ aléatoire binaire Z défini sur W, isotrope et stationnaire du second ordre, de moyenne et de fonction de covariance connues. Les observations ne sont pas fiables : Y est une observation bruitée de Z, au sens où l'on peut "manquer" le phénomène. Y est également un champ aléatoire binaire sur W. Enfin, Δ est un sous-ensemble fini de points échantillonnables de W (par exemple une grille). Nous supposons l'indépendance conditionnelle des observations $\{Y_s\}_{s\in\Delta}$ sachant le champ Z. Pour des variables binaires, le krigeage simple[1] fournit une estimation de $\mathbb{P}(Z_s = 1 \mid Z^{obs})$. Cependant,

[1]J.P Chilès and P.Delfiner. *Modeling Spatial Uncertainty.* Wiley series in Probability and Statistics, 1999.

nous devons introduire ici deux adaptations de cette méthode car (i) les observations sont bruitées et (ii) nous considérons deux types d'observations : $Y^{obs} = \{Y_{\alpha_1}, \ldots, Y_{\alpha_K}\}$, obtenues après observation sur les sites de l'échantillon $\{\alpha_1, \ldots, \alpha_K\}$ dont on veut optimiser le choix, et $y^{init} = \{y_{\beta_1}, \ldots, y_{\beta_L}\}$ obtenues suite à un échantillonnage initial arbitraire et/ou au cours des étapes précédentes d'un échantillonnage adaptatif. On souhaite que le choix de $\{\alpha_1, \ldots, \alpha_K\}$ dépende de $\{y_{\beta_1}, \ldots, y_{\beta_L}\}$ et pas seulement des positions $\{\beta_1, \ldots, \beta_L\}$. Nous proposons donc une méthode de *krigeage conditionnel* qui consiste à calculer $p^{\lambda^*,\gamma^*}(Y^{obs}, y^{init}, s) \simeq \mathbb{P}(Z_s = 1 \mid Y^{obs}, y^{init})$, tel que :

$$\{\lambda^*, \gamma^*\} = \underset{\lambda,\gamma}{\operatorname{argmin}} \; \mathbb{E}\left[(p^{\lambda,\gamma}(Y^{obs}, y^{init}, s) - Z_s)^2 \mid y^{init}\right] \tag{1}$$

avec $p^{\lambda,\gamma}(Y^{obs}, y^{init}, s) = \lambda_s + \sum_{k=1}^{K} \lambda_k Y_{\alpha_k} + \sum_{l=1}^{L} \gamma_l \, y_{\beta_l}$. La quantité $\mathbb{E}[(p^{\lambda^*,\gamma^*}(Z^{obs}, y^{init}, s) - Z_s)^2 \mid y^{init}]$ est la *variance du krigeage conditionnel*, qui permet d'associer une erreur locale au point s à un échantillon $\{\alpha_1, \ldots, \alpha_K\}$ étant donné y^{init}.

3 Echantillonnage spatial approché basé sur le krigeage

Etant donné un échantillon initial $\{\beta_1, \ldots, \beta_L\}$ et les observations associées y^{init}, nous définissons la valeur d'un échantillon $\{\alpha_1, \ldots, \alpha_K\}$ comme la somme, sur l'ensemble des points de Δ, de l'opposé des variances du krigeage conditionnel en tous points de Δ :

$$U^{kri}(\alpha_1, \ldots, \alpha_K) = -\sum_{s \in \Delta} \mathbb{E}\left[(p^{\lambda^*,\gamma^*}(Y^{obs}, y^{init}, s) - Z_s)^2 \mid y^{init}\right] \tag{2}$$

Un “bon” échantillon (au sens de U^{kri}) minimise l'erreur de prédiction, mesurée par la variance du krigeage conditionnel, en tout point de la grille. Optimiser l'échantillon consiste alors à résoudre le problème de maximisation suivant :

$$\alpha^* = \underset{\{\alpha_1,\ldots,\alpha_K\} \subseteq \Delta}{\operatorname{argmax}} U^{kri}(\alpha_1, \ldots, \alpha_K) \tag{3}$$

La résolution exacte de (3) est trop coûteuse si $|\Delta|$ et K sont grands. Nous proposons donc une méthode de résolution approchée de (3). Une méthode d'échantillonnage approchée “naturelle” consiste à échantillonner les sites de Δ où l'incertitude reste la plus grande après prise en compte de l'information y^{init}. Elle consiste à échantillonner les points s pour lesquels la valeur $\min\Big\{\mathbb{P}(Z_s = 1 \mid y^{init}), \mathbb{P}(Z_s = 0 \mid y^{init})\Big\}$ est la plus élevée. Cela revient à modifier le critère U^{kri} de la façon suivante :

$$\tilde{U}^{kri}(\alpha_1, \ldots, \alpha_K) = \sum_{k=1}^{K} \min\Big\{\mathbb{P}(Z_{\alpha_k} = 1 \mid y^{init}), \mathbb{P}(Z_{\alpha_k} = 0 \mid y^{init})\Big\} \tag{4}$$

La mise en œuvre de cette méthode approchée requiert uniquement le calcul des probabilités conditionnelles que nous calculons de manière approchée en utilisant la méthode du krigeage simple : $\mathbb{P}(Z_s = 1 \mid y^{init}) \simeq p^{\gamma^*}(y^{init}, s)$, où :

$$\gamma^* = \underset{\gamma}{\operatorname{argmin}} \, \mathbb{E}\left[(p^{\gamma}(Y^{init}, s) - Z_s)^2\right] \tag{5}$$

Cette méthode approchée résulte de la méthode exacte, en faisant les deux hypothèses suivantes :

1. Les observations $\{Y_{\alpha_1}, \ldots, Y_{\alpha_K}\}$ sont exactes : $Y_{\alpha_i} = Z_{\alpha_i} \; \forall i \in \{1, \ldots, K\}$.
2. Pour tout $s \in \Delta$, pour tout $k \in \{1, \ldots, K\}$: $\mathbb{E}\left[Z_s Z_{\alpha_k} \mid y^{init}\right] = \mathbb{E}\left[Z_s \mid y^{init}\right] \mathbb{E}\left[Z_{\alpha_k} \mid y^{init}\right]$.

Les conclusions de nos expérimentations, menées sur des données simulées, sont les suivantes : (i) même avec cette heuristique très simple nous obtenons de très bons résultats en terme d'erreur de reconstruction et (ii) la méthode adaptative donne de meilleurs résultats que la méthode statique, car elle permet de retourner échantillonner les endroits où il reste de l'incertitude.