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

M. Bonneau[1] N. Peyrard[1] R. Sabbadin[1]

[1] INRA-Toulouse - Unité de Biométrie et Intelligence Artificielle UR875
BP 52627 – 31326 Castanet-Tolosan – FRANCE
{mathieu.bonneau,nathalie.peyrard,regis.sabbadin}@toulouse.inra.fr

Résumé

Le contrôle ou la gestion d'un phénomène spatial repose souvent sur l'établissement d'une carte d'occurrence du phénomène. Celle-ci doit être construite à partir d'un échantillonnage spatial de la zone d'étude, une exploration exhaustive étant trop coûteuse. De plus, le phénomène pouvant être difficile à observer, les données d'observation sont souvent bruitées. Dans le cas de variables à valeurs continues, la géostatistique fournit des outils pour le choix d'un échantillon. Ici, en restant dans ce cadre méthodologique, nous étudions le même problème de choix d'un échantillon mais afin de reconstruire une carte d'occurrence (variables binaires). A partir de la méthode du krigeage, nous définissons la valeur d'un échantillon, puis nous exprimons le problème de choix d'un échantillon comme un problème d'optimisation. Nous considérons deux types d'échantillonnage : statique ou adaptatif. Dans les deux cas, la résolution de ce problème étant trop complexe, nous proposons des méthodes de résolution approchée. Les méthodes exactes et approchées sont définies pour toute distribution de cartes binaires, stationnaire d'ordre 2. Une analyse sur données simulées dans le cas d'un modèle booléen montre que les taux d'erreur de reconstruction sont très raisonnables. De plus, elle montre que le gain obtenu avec la stratégie adaptative, qui intègre les observations recueillies aux étapes d'échantillonnage intermédiaires pour sélectionner l'échantillon suivant, est significatif.

Mots Clef

Echantillonnage spatial optimal, échantillonnage adaptatif, krigeage conditionnel, modèle booléen.

Abstract

The control of spatial processes often requires to build an occurrence map of this process. This map is in general built from a spatial sampling of the study zone, since an exhaustive exploration of the area is too costly. Moreover, observations of the process to map can also be noisy. When the variables to map have a continuous domain, geostatistical tools can be used for spatial sampling design. Here, within this methodological framework, we study the same problem of sample choice but for building occurrence maps (finite variable domains). More precisely, we develop a method based on kriging for the definition of the value of a sample and then we define the sample choice as an optimisation problem. We consider two types of sampling problems : static and adaptive. In both cases, since exact resolution is out of reach, we propose approximate solution methods. Exact and approximate methods are defined for any binary map distribution, provided that it is stationary of order 2. An anaysis on data simulated from a boolean model shows low error rates in the map reconstruction. Furthermore, the analysis reveals that the adaptive sampling method, which takes previous sampling steps results into account in order to choose the next sample, outperforms the static sampling method.

Keywords

Optimal spatial sampling, adaptive sampling, conditional kriging, boolean model.

1 Introduction

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 du phénomène. Cependant une exploration exhaustive de la zone est en général impossible du fait du coût en temps et en argent de cette exploration. La carte doit donc être reconstruite à partir d'un échantillon spatial. Par ailleurs, les observations recueillies peuvent être bruitées, car certaines espèces sont difficiles à observer ou, pour certaines maladies, les cas sont mal diagnostiqués ou répertoriés. Le problème posé est alors celui du choix d'un échantillon puis de la reconstruction d'une carte d'occurrence à partir de données bruitées et incomplètes.

Dans ce contexte spatial, les outils de la géostatistique permettent d'apporter des réponses méthodologiques [4], [5] ou [6]. Ainsi, dans [7] les auteurs proposent une méthode basée sur une modélisation par champ gaussien et un critère de type entropie pour la sélection d'un sous-réseau optimal dans un réseau initial de surveillance, dans le cas d'observations non bruitées. Dans [1], [2] ou [9], les

mêmes outils sont utilisés pour développer une méthode d'extension ou réduction d'un réseau de surveillance. Dans les deux travaux précédents, le champ spatial à reconstruire est à valeurs continues, cadre classique de la géostatistique. Or, lorsque l'on souhaite reconstruire une carte d'occurrence, le champ spatial est à valeurs binaires (présence / absence). Nous proposons ici une méthode 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). Pour cela nous restons dans le cadre de la géostatique et utilisons le krigeage [3]. Nous nous intéressons aux cas d'échantillonnage statique et 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.

Notre démarche est la suivante : à partir du krigeage, nous définissons la valeur d'un échantillon (ici un échantillon est un ensemble de point de $\mathbb{R}^2$, et non les observations associées). Ensuite, à partir de cette notion de valeur nous définissons la question du choix du "meilleur" échantillon comme un problème d'optimisation. Enfin, toujours en utilisant le krigeage, une carte d'occurrence est reconstruite à partir des observations reccueillies pour l'échantillon sélectionné. D'autres méthodes d'échantillonnages se servant du krigeage sont envisageable, *e.g* [11].

Le krigeage classique fournit une prédiction en un site (un point de $\mathbb{R}^2$) à partir d'observations obtenues en d'autres sites. Il associe à cette prédiction une notion d'erreur qui ne dépend que de la position des sites échantillonnés et pas de leur valeur. Ainsi l'application directe du krigeage pour la sélection d'un échantillon ne permet pas de prendre en compte les valeurs des observations acquises, soit lors des étapes précédentes dans le cas adaptatif, soit par un échantillonnage initial ad-hoc de la zone d'étude. Or ces informations peuvent guider le choix du nouvel échantillon. Aussi, nous proposons d'utiliser une adaptation du krigeage que nous appelons krigeage conditionnel pour définir la valeur d'un échantillon lorsque des informations (observations résultant des échantillons précédents) sont déjà disponibles. Ensuite, nous présentons la définition du problème de choix de l'échantillon optimal en utilisant le krigeage conditionnel, dans les cas statique et adaptatif. La résolution du problème d'optimisation étant trop complexe, même pour des cartes de petite taille, nous présentons dans les deux cas une méthode de résolution approchée.

Nous présentons tout d'abord les hypothèses faites sur le modèle de carte d'occurrence et des observations (section 2), puis nous définissons le krigeage conditionnel (section 3). La méthode exacte pour le choix d'un échantillon optimal, ainsi que la méthode de résolution approchée sont décrites dans la section 4 pour le cas statique et dans la section 5 pour le cas adaptatif. Enfin, les performances relatives de ces deux méthodes sont étudiées dans le cas de données simulées selon un modèle booléen (section 6). Nous concluons cet article par une discussion sur des perspectives ouvertes par ces travaux.

2 Le modèle

Nous supposons que la zone d'étude peut être représentée comme un compact W de $\mathbb{R}^2$ et qu'une carte d'occurence est la réalisation d'un champ aléatoire binaire Z défini sur W comme suit : $Z_s = 1$ si le phénomène est présent au point $s \in W$ et 0 sinon. La seule hypothèse faite est que Z est un champ stationnaire du second ordre, de moyenne m et de fonction de covariance $\sigma_{ij} = \mathrm{Cov}[Z_i, Z_j], \forall i, j \in W$. Nous supposons de plus que les observations effectuées sur le terrain ne sont pas exactes. Y est un champ aléatoire binaire sur W tel que $Y_s = 1$ si le phénomène a été observé au point $s \in W$ et 0 sinon. Y est une observation bruitée de la "réalité" Z, au sens où l'on peut "manquer" le phénomène. Nous faisons également une hypothèse d'indépendance conditionnelle des observations Y sachant le champ Z. La loi conditionnelle de Y_s sachant Z_s est définie comme suit :

$$\begin{aligned} \mathbb{P}(Y_s = 0 \mid Z_s = 0) &= 1, \\ \mathbb{P}(Y_s = 1 \mid Z_s = 1) &= \theta, 0 < \theta < 1. \end{aligned} \quad (1)$$

Définissons Δ, un sous-ensemble fini de W, qui représente l'ensemble des N sites échantillonnables de la zone d'étude. En effet, en pratique il n'est pas possible de chercher l'échantillon optimal parmi l'ensemble des sous-ensembles de W. Δ sera, par exemple, l'ensemble des nœuds d'une grille régulière sur W. L'hypothèse d'indépendance conditionnelle des observations implique alors :

$$\mathbb{P}(Y_\Delta = y_\Delta \mid Z_\Delta = z_\Delta) = \prod_{s \in \Delta} \mathbb{P}(Y_s = y_s \mid Z_s = z_s).$$

où l'indice Δ indique la restriction de la variable ou de sa réalisation aux points de Δ.

3 Krigeage conditionnel

Soient $Z^{obs} = \{Z_{\alpha_1}, \ldots, Z_{\alpha_K}\}$ les valeurs du champ Z en K points $\{\alpha_1, \ldots, \alpha_K\}$ de Δ. Le problème de l'estimation des paramètres du modèle, lorsque les observations sont bruitées et rares, est un problème difficile. Nous supposerons, afin de nous concentrer sur l'exposé de notre méthode d'échantillonnage, que ces paramètres sont connus, en particulier la moyenne du champ aléatoire Z. Dans ce cas, le krigeage simple [3] fournit une prédiction, $p^*(Z^{obs}, s)$, de la valeur de Z au point s, comme le meilleur prédicteur linéaire sans biais au sens des moindres carrés : $p^*(Z^{obs}, s) = p^{\lambda^*}(Z^{obs}, s)$, où

$$\lambda^* = \arg \min_{\lambda \in \mathbb{R}^{K+1}} \mathbb{E}\left[(p^\lambda(Z^{obs}, s) - Z_s)^2\right] \quad (2)$$

$p^{\lambda}(Z^{obs}, s) = \lambda_0 + \sum_{k=1}^{K} \lambda_k Z_{\alpha_k}$ et $\lambda = \{\lambda_0, \ldots, \lambda_K\}$. Un calcul simple montre que : $\mathbb{E}[(p^*(Z^{obs}, s) - Z_s)^2] = \mathrm{Var}[(p^*(Z^{obs}, s) - Z_s)] + \mathbb{E}[p^*(Z^{obs}, s) - Z_s]^2$. Le terme $\mathbb{E}[p^*(Z^{obs}, s) - Z_s]$ est appelé *biais* de la prédiction, le poid λ_0^* est choisi afin de l'éliminer : $\lambda_0^* = \mathbb{E}[Z_s](1 - \sum_{k=1}^{K} \lambda_k)$. Le terme $\mathrm{Var}[(p^*(Z^{obs}, s) - Z_s)]$ est la *variance du krigeage*, qui permet d'associer une erreur locale au point s à un échantillon $\{\alpha_1, \ldots, \alpha_K\}$. Ainsi, la résolution de (2) consiste à trouver les poids optimaux $\{\lambda_1^*, \ldots, \lambda_K^*\}$ minimisant la variance du krigeage.

Nous devons introduire deux adaptations du krigeage simple. D'une part les observations sont bruitées. D'autre part, 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. Les premières ne sont pas connues au moment de la sélection de l'échantillon (variables aléatoires) alors que les secondes le sont (réalisations). On souhaite que le choix de $\{\alpha_1, \ldots, \alpha_K\}$ dépende de $\{y_{\beta_1}, \ldots, y_{\beta_L}\}$ et pas uniquement des positions $\{\beta_1, \ldots, \beta_L\}$. Nous proposons donc de considérer ce que nous appelerons *le krigeage conditionnel* qui consiste à calculer $p^{\lambda^*, \gamma^*}(Y^{obs}, y^{init}, s)$, tel que

$$\{\lambda^*, \gamma^*\} = \arg\min_{\lambda, \gamma} \mathbb{E}\left[(p^{\lambda,\gamma}(Y^{obs}, y^{init}, s) - Z_s)^2 \mid y^{init}\right]. \tag{3}$$

avec $p^{\lambda,\gamma}(Y^{obs}, y^{init}, s) = \lambda_0 + \sum_{k=1}^{K} \lambda_k Y_{\alpha_k} + \sum_{l=1}^{L} \gamma_l y_{\beta_l}$.

Les termes γ_l^* disparaissent de l'expression de $p^{\lambda^*, \gamma^*}(Y^{obs}, y^{init}, s)$ (par l'équation 4) et le calcul de λ^* revient à résoudre le système suivant :

$$\begin{cases} \sum_{k=1}^{K} \lambda_k \sigma'_{\alpha_k \alpha_1} = & \sigma'_{s\alpha_1} \\ \quad \vdots & \quad \vdots \\ \sum_{k=1}^{K} \lambda_k \sigma'_{\alpha_k \alpha_K} = & \sigma'_{s\alpha_K} \end{cases}$$

avec

$$\sigma'_{\alpha_i \alpha_j} = \mathrm{Cov}[Y_{\alpha_i}, Y_{\alpha_j} \mid y^{init}]$$

$$\sigma'_{s\alpha_i} = \mathrm{Cov}[Y_{\alpha_i} Z_s \mid y^{init}]$$

et λ_0^* vaut

$$\lambda_0^* = m'_s - \sum_{l=1}^{L} \gamma_l^* y_{\beta_l} - \sum_{k=1}^{K} \lambda_k^* m'_{\alpha_k} \tag{4}$$

avec $m'_{\alpha_k} = \mathbb{E}\left[Y_{\alpha_k} \mid y^{init}\right]$ et $m'_s = \mathbb{E}\left[Z_s \mid y^{init}\right]$. Les observations y^{init} interviennent via les espérances conditionnelles m'_{α_k} et m'_s. Le système obtenu est similaire au système du krigeage simple, les covariances et espérances étant remplacées par des covariances et espérances conditionnelles et les observations portant sur le champ Y et non Z. Même si l'expression mathématique de m'_{α_k}, m'_s, $\sigma'_{\alpha_i \alpha_j}$ et $\sigma'_{s\alpha_i}$ peut être obtenue pour une distribution donnée de Z, il n'est généralement pas possible en pratique de calculer ces valeurs de manière exacte. Dans la suite nous définissons le problème de choix d'un échantillon en nous appuyant sur le krigeage conditionnel et nous proposons une méthode de résolution approchée qui ne fait intervenir que m'_s. Cette espérance sera calculée de manière approchée, là encore par krigeage.

4 Echantillonnage statique

Nous présentons d'abord l'utilisation du krigeage conditionnel dans le cas de l'échantillonnage statique, où l'ensemble des sites à échantillonner est déterminé une fois pour toute au début de la campagne d'échantillonnage.

4.1 Valeur d'un échantillon

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 :

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

Cela correspond à dire qu'un bon échantillon minimise l'erreur de prédiction, mesurée par la variance du krigeage conditionnel, en tout point de la grille. Chaque $p^*(Y^{obs}, y^{init}, s)$ est obtenu par résolution du problème (3). Remarquons que la valeur U^{kri} d'un échantillon peut aussi être définie en deux étapes, si l'on définit d'abord la valeur d'un ensemble d'observations $y^{obs} = \{y_{\alpha_1}, \ldots, y_{\alpha_K}\}$ par

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

La valeur d'un échantillon est alors l'espérance de V^{kri} sur toutes les observations possibles y^{obs} :

$$U^{kri}(\alpha_1, \ldots, \alpha_K) = \mathbb{E}\left[V^{kri}(Y_{\alpha_1}, \ldots, Y_{\alpha_K}) \mid y^{init}\right].$$

4.2 Echantillonnage optimal

Nous considérons la prise en compte d'un coût d'échantillonnage uniquement à travers la contrainte d'une taille fixe (K) de l'échantillon. Le choix de l'échantillon $\{\alpha_1, \ldots, \alpha_K\}$, optimal au sens du critère de qualité défini à partir du krigeage conditionnel, consiste alors à résoudre le problème de maximisation suivant :

$$\alpha^* = \arg \max_{\{\alpha_1, \ldots, \alpha_K\} \subseteq \Delta} U^{kri}(\alpha_1, \ldots, \alpha_K). \tag{5}$$

La résolution exacte de (5) demande de résoudre un système de la forme (3) pour chaque sous-ensemble de taille K de Δ, ce qui est trop coûteux en pratique si N et K sont grands. Nous proposons donc une méthode de résolution approchée de (5).

4.3 Méthode de résolution approchée

Rappelons que la seule information disponible pour choisir l'échantillon "optimal" est $y^{init} = \{y_{\beta_1}, \ldots, y_{\beta_L}\}$. Une méthode d'échantillonnage approchée "naturelle" consiste à aller échantillonner les sites de Δ où l'incertitude reste la plus grande après prise en compte de l'information y^{init}. Il s'agit donc d'aller 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 :

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

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 $k \in \{1, \ldots, K\}$:

$$\mathbb{E}\left[Z_s Z_{\alpha_k} \mid y^{init}\right] = \mathbb{E}\left[Z_s \mid y^{init}\right] \mathbb{E}\left[Z_{\alpha_k} \mid y^{init}\right]$$

(voir Annexe pour la démonstration).

La mise en œuvre de cette méthode approchée requiert uniquement le calcul des probabilités conditionnelles $\mathbb{P}(Z_s = 1 \mid y^{init})$ pour tout point s de Δ. Nous les calculons de manière approchée en utilisant la méthode du krigeage simple. En effet, le prédicteur linéaire obtenu en remplaçant Z^{obs} par Y^{init} dans (2) est une approximation de l'espérance conditionnelle $\mathbb{E}\left[Z_s \mid y^{init}\right]$ qui, dans le cas de variables à valeurs dans $\{0, 1\}$, est égale à $\mathbb{P}(Z_s = 1 \mid y^{init})$. La résolution de ce problème de krigeage simple au point $s \in \Delta$ revient à résoudre le système suivant :

$$\begin{cases} \sum_{l=1}^{L} \lambda_l \theta^2 \sigma_{\beta_l \beta_1} = & \theta \sigma_{s\beta_1} \\ \qquad\vdots & \qquad\vdots \\ \sum_{l=1}^{L} \lambda_l \theta^2 \sigma_{\beta_l \beta_L} = & \theta \sigma_{s\beta_L} \end{cases}$$

Le coefficient λ_0^* vaut $m - \sum_{l=1}^{L} \lambda_l^* \theta m$. On retrouve le même système que dans le cas d'observations exactes, mais les covariances et espérances sont remplacées par celles des observations bruitées : $\mathbb{E}[Y_s] = \theta m$, $Cov(Y_i, Y_j) = \theta^2 \sigma_{ij}$ et $Cov(Y_i, Z_s) = \theta \sigma_{is}$.

5 Echantillonnage adaptatif

Nous considérons maintenant le cas de l'échantillonnage adaptatif : le choix des sites à échantillonner est fait de manière séquentielle. Les observations recueillies aux étapes précédentes sont prises en compte pour le choix de l'échantillon dans l'étape courante.

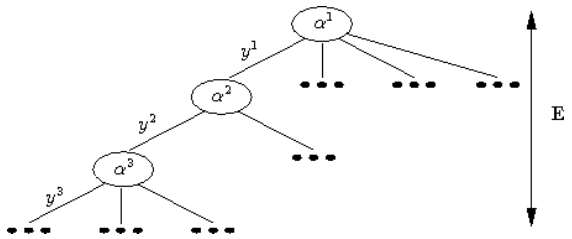


FIG. 1 – Représentation d'une politique d'échantillonnage par un arbre.

5.1 Politique d'échantillonnage

Supposons que l'on répartisse le choix des K sites à échantillonner en E étapes et qu'à chaque étape Q sites sont échantillonnés. Dans le cas adaptatif on ne parle plus d'un échantillon (liste de sites de Δ) mais d'une *politique d'échantillonnage*. Si α^1 est l'ensemble des sites échantillonnés à l'étape 1 et y^1 les observations correspondantes, alors pour une politique d'échantillonnage δ donnée, l'ensemble des sites échantillonnés à l'étape 2, α^2, est une fonction de α^1 et y^1 : $\alpha^2 = \delta^2\big((\alpha^1, y^1)\big)$. Plus généralement si α^q est l'ensemble des sites échantillonnés à l'étape q, alors $\alpha^q = \delta^q\big((\alpha^1, y^1), \ldots, (\alpha^{q-1}, y^{q-1})\big)$. Une telle politique peut être représentée sous forme d'un arbre (voir Figure 1) de profondeur E : les nœuds de l'arbre sont des ensembles de sites échantillonnés et les arêtes issues d'un même nœud représentent les différentes valeurs possibles pour les observations correspondantes. Une trajectoire τ_δ de δ est un chemin dans l'arbre, c'est à dire un ensemble de sites et d'observations reliant la racine de l'arbre à une feuille :

$$\tau_\delta = \{(\alpha^1, y^1), \ldots, (\alpha^E, y^E)\}.$$

5.2 Echantillonnage optimal

Dans le cas adaptatif, la définition du problème de choix de la politique d'échantillonnage optimale demande de redéfinir ce que l'on appelle observation. En effet, une des hypothèses du krigeage est que les sites échantillonnés sont deux à deux distincts. Or, rien n'interdit dans une politique adaptative de retourner plusieurs fois explorer un même site. Pour une trajectoire donnée $\tau_\delta = \{(\alpha^1, y^1), \ldots, (\alpha^E, y^E)\}$ qui visite au cours des E étapes les sites $\{s_1, \ldots, s_T\}$, respectivement $\{n_1, \ldots, n_T\}$ fois (avec $n_t > 0,\ \forall 1 \leq t \leq T$), nous définissons $u^{obs} = \{u_{s_1}, \ldots, u_{s_T}\}$ avec $u_{s_t} = 1$ si on a observé le phénomène à la dernière visite du site s_t et 0 sinon (on exclut le cas où une politique retournerait explorer un site où le phénomène a déjà été observé). La valeur d'une trajectoire τ_δ s'écrit

$$V^{kri}(\tau_\delta) = -\sum_{s \in \Delta} \mathbb{E}\left[(p^*(u^{obs}, y^{init}, s) - Z_s)^2 \mid y^{init}, u^{obs}\right]$$

$$\text{avec } p^*(U^{obs}, y^{init}, s) = \lambda_0^* + \sum_{t=1}^{T} \lambda_t^* U_{s_t}$$

où les λ_t^* sont solutions du krigeage conditionnel où l'on a remplacé Y^{obs} par U^{obs} (voir section 3). Remarquons que le cas où un site est exploré dans y^{init} et dans u^{obs} ne pose pas de problème car en pratique on ne krige pas sur y^{init} (y^{init} intervient uniquement via λ_0^*). La valeur $U^{kri}(\delta)$ de la politique δ s'obtient alors comme l'espérance de $V^{kri}(\tau_\delta)$ sur l'ensemble des trajectoires possibles :

$$U^{kri}(\delta) = \mathbb{E}\left[V^{kri}(\tau_\delta) \mid \delta, y^{init}\right].$$

5.3 Méthode de résolution approchée

La méthode approchée dans le cas adaptatif est une répétition de E étapes de la méthode approchée dans le cas statique. Au début de l'étape d'échantillonnage e ($1 < e \leq E$), l'échantillon initial et les observations correspondantes y^{init} sont incrémentés des sites visités et des observations recueillies lors de l'étape $e-1$. L'observation y^{init} ainsi augmentée est transformée en u^{init} selon le même principe de transformation de Y^{obs} en U^{obs} dans le cas de la méthode adaptative exacte. On calcule alors de manière approchée les probabilités conditionnelles $\mathbb{P}(Z_s = 1 \mid u^{init})$ par la méthode du krigeage simple. Le prédicteur en s est $p^*(U^{init}, s) = \lambda_0^* + \sum_{t=1}^T \lambda_t^* U_{s_t}$ où les λ_t^* sont solutions du système suivant :

$$\begin{cases} \theta \sum_{t=1}^T \lambda_t \big((1-\theta)^{n_t - 1}\sigma_{s_t s_1}\big) & = \sigma_{s\alpha_1} \\ \qquad\qquad \vdots & \qquad\quad \vdots \\ \theta \sum_{t=1}^T \lambda_t \big((1-\theta)^{n_t - 1}\sigma_{s_t s_T}\big) & = \sigma_{s\alpha_T} \end{cases}$$

et

$$\lambda_0^* = m\big(1 - \theta \sum_{t=1}^T \lambda_t^* (1-\theta)^{n_t - 1}\big).$$

Ce système découle du fait que $\mathbb{E}[U_{s_t}] = \theta m (1-\theta)^{n_t - 1}$, $\mathbb{E}\left[U_{s_i}, U_{s_j}\right] = \theta^2 \sigma_{s_i s_j} (1-\theta)^{n_i + n_j - 2}$ et $\mathbb{E}[U_{s_i} Z_s] = \theta(1-\theta)^{n_i - 1}\sigma_{s_i s}$.

6 Illustration dans le cas du modèle booléen

Les méthodes exactes et approchées présentées jusqu'ici sont décrites de manière générale sans spécification d'une distribution de probabilité sur le champ Z. Seules sont requises la stationnarité d'ordre 2 et la connaissance de la moyenne m et la fonction de covariance σ associées à Z. Nous illustrons maintenant les méthodes approchées sur des données simulées, dans le cas particulier où Z est modélisé comme la fonction indicatrice d'un ensemble aléatoire, réalisation d'un modèle booléen.

6.1 Le modèle booléen

Le modèle booléen [10] est construit à partir d'un processus stationnaire de Poisson, d'intensité Λ qui forme les germes du processus. A chaque germe s est attaché un grain primaire Ξ_s. Ces grains sont des variables aléatoires i.i.d. et indépendantes du processus de Poisson. Le modèle booléen Ξ est défini à partir des germes $\{s_n\}$ et des grains primaires $\{\Xi_{s_n}\}$ de la façon suivante :

$$\Xi = \cup_{n=1}^{\infty} (\Xi_{s_n} \; + \; s_n)$$

Le champ Z est alors défini comme $Z_s = \mathbb{1}_{\{s \in \Xi\}} \quad \forall s \in W$. En épidémiologie ou écologie, ce modèle peut s'interpréter comme un ensemble de sites d'infection/apparition initiaux (les germes) à partir desquels la maladie/l'espèce se propage pour former des patchs (les grains). Par souci de simplicité, nous considérons un modèle booléen dont les grains primaires sont des disques de rayon r constant (voir Figure 2). Dans ce cas, la covariance non centrée, entre deux sites à distance h s'écrit [10]

$$\begin{aligned} C(h) &= \mathbb{P}(\{s, s+h\} \in \Xi) = \mathbb{P}(\{0, h\} \in \Xi) \\ &= 2p \; - \; 1 \; + \; (1-p)^2 \exp(\Lambda V(h)), \end{aligned}$$

où $V(h)$ est l'aire d'intersection de deux disques de rayon r dont les centres sont distants de h : si $R = 2r$ alors

$$V(h) = 2R^2 \arccos(h/R) - hR(1 - h^2/R^2)^{1/2},$$

lorsque $h \leq R$, et 0 sinon.
Le paramètre p, appelé fraction de volume est la moyenne de l'aire occupée par Ξ dans une région d'aire unité :

$$p = \mathbb{E}(\mathcal{A}(\Xi \cap B)), \quad \mathcal{A}(B) = 1.$$

On peut montrer que $p = \mathbb{P}(s \in \Xi) = 1 - \exp^{\Lambda \pi r^2}$, $\forall \; s \in W$, [10]. Ainsi $m = \mathbb{E}[Z_s] = p$ est connu dès lors que l'on connait r et Λ.

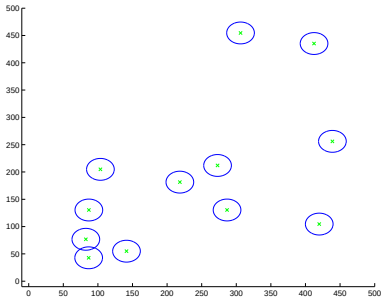


FIG. 2 – Réalisation d'un modèle booléen avec pour grain primaire un disque de rayon constant.

6.2 Analyse sur données simulées

Nous avons effectué, dans le cas du modèle booléen, l'étude des performances des méthodes approchées statique et adaptative, ainsi qu'une comparaison avec des méthodes d'échantillonnages plus classiques : l'échantillonnage systématique et aléatoire. Nous avons supposé que les paramètres r et Λ sont connus. La zone d'étude W est le rectangle $[0; 500] \times [0; 500]$. L'ensemble Δ des points échantillonnable forme une grille régulière. Ils sont espacés de 10 unités dans les deux directions : $\Delta = \{(10k, 10k') \in \mathbb{R}^2 \; / \; k \in \{0, \ldots, 49\}, k' \in \{0, \ldots, 49\}\}$

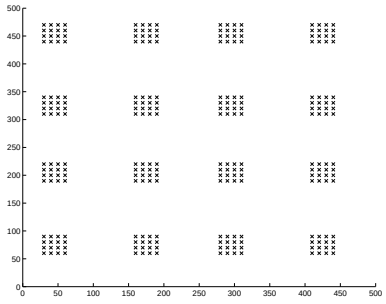


FIG. 3 – Sites explorés lors de l'échantillonnage initial.

($|\Delta| = 2500$). L'échantillon initial est régulièrement réparti, comme sur la Figure 3 et correspond à peu près à 10% des points échantillonnables. On considère que l'échantillonnage systématique revient à placer aléatoirement et sans chevauchement un certain nombre de grilles de points régulièrements espacés de 5 unités. Chaque grille représente 4% des points échantillonnables. Pour un jeu de paramètres donné (Λ, r, θ), 10 cartes d'occurrence (réalisations du champ Z) sont simulées selon le modèle booléen correspondant dans le cas des échantillonnages statique et adaptatif, 6 cartes pour les échantillonnages systématique et aléatoire . Puis, pour une carte z, 10 échantillonnages et reconstructions sont effectués. La procédure d'échantillonnage puis reconstruction dans le cas statique est la suivante :

1. Simulation des observations initiales y^{init} sachant z selon la loi définie par (1) ;
2. Calcul de $\mathbb{P}(Z_s = 1 \mid y^{init})$ par krigeage simple, pour tout $s \in \Delta$ (section 4.3) ;
3. Classement des points de Δ par valeurs décroissantes de $\min\left\{\mathbb{P}(Z_s = 1 \mid y^{init}), 1 - \mathbb{P}(Z_s = 1 \mid y^{init})\right\}$ et sélection des K premiers, $\{\alpha_1, \ldots, \alpha_K\}$;
4. Simulation des observations correspondantes $y^{obs} = \{y_{\alpha_1}, \ldots, y_{\alpha_K}\}$ sachant z, selon la loi définie par (1) ;
5. Calcul des probabilités $\mathbb{P}(Z_s = 1 \mid y^{init}, y^{obs})$, pour tout $s \in \Delta$, par krigeage simple. La reconstruction $\hat{z}$ de la carte z est alors définie par $\hat{z}_s = 1$ si $\mathbb{P}(Z_s = 1 \mid y^{init}, y^{obs}) > \frac{1}{2}$ et 0 sinon. Si s a été exploré et que $y_s = 1$, alors $\hat{z}_s = 1$.

Cette procédure est répétée pour des valeurs croissantes de K, allant de 5% à 100% du nombre de sites échantillonnables.

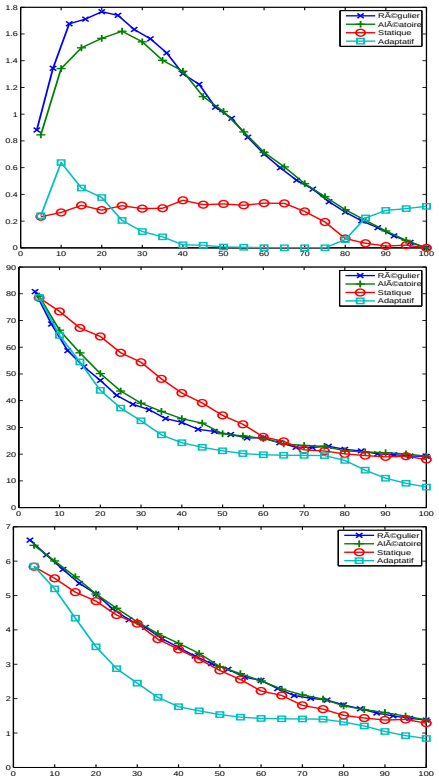
Dans le cas de l'échantillonnage adaptatif, on considère également une taille d'échantillon égale à 5 % de la taille de Δ à chaque étape. La première étape est identique au cas statique. Pour les étapes suivantes, on intègre les observations y^{obs} recueillies aux observations initiales y^{init} avant d'appliquer la même procédure que dans le cas statique. On utilise la méthode de la section 5.3 pour calculer les probabilités conditionnelles de Z_s sachant y^{init} et ordonner les sites de Δ.

Les résultats sont analysés en termes de pourcentage de faux positifs $\frac{card(s \in \Delta \text{ t.q. } z_s=0, \hat{z}_s=1)}{card(s \in \Delta \text{ t.q. } z_s=0)}$, pourcentage de faux négatifs $\frac{card(s \in \Delta \text{ t.q. } z_s=1, \hat{z}_s=0)}{card(s \in \Delta \text{ t.q. } z_s=1)}$et d'erreur globale $\frac{card(s \in \Delta \text{ t.q. } z_s \neq \hat{z}_s)}{card(\Delta)}$. Nous avons considéré 4 jeux de paramètres : dans tous les cas $\theta = 0.8$ et $\Lambda = 0.0001$, ce qui correspond à une moyenne de 25 disques dans une carte z. Le rayon r vaut respectivement 10, 15, 20 et 30, ce qui correspond, sur les simulations réalisées, à une moyenne de 3.4 %, 7.2%, 11.4% et 24.5% de sites où le phénomène est présent. Nous ne reportons ici les taux d'erreur que pour les cas $r = 15$ et $r = 30$ (voir Figure 4), les résultats étant qualitativement similaires pour les 4 jeux de paramètres.

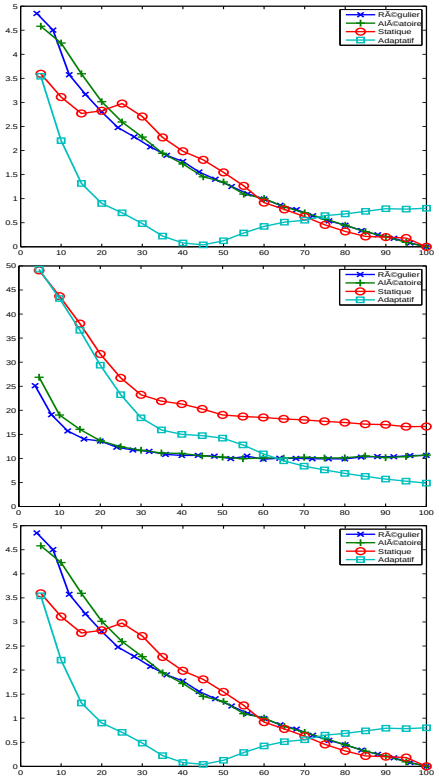
L'erreur de reconstruction est essentiellement due à des faux négatifs. Le pourcentage de faux négatifs et l'erreur totale diminuent lorsque le rayon augmente, alors que le pourcentage de faux positifs à l'inverse, augmente. Cela est dû au fait que le krigeage conduit à une surestimation du nombre de site où le phénomène est reconstruit présent et cette surestimation croit avec le rayon. L'erreur globale est très raisonnable, même pour de petites valeurs de la taille de l'échantillon (K) et diminue rapidement lorsque K augmente. La méthode d'échantillonnage statique va conduire à explorer les sites les plus éloignés des sites visités lors de l'échantillonnage initial lorsque le phénomène n'a pas été observé, alors que la méthode adaptative conduit à une couverture moins "régulière", car guidée par les observations intermédiaires (Figure 5). On observe l'interêt de cette dernière stratégie, qui utilise l'information des étapes précédentes et autorise à retourner voir un site où l'incertitude reste élevée : l'erreur globale est toujours inférieure à celle obtenue avec un échantillonnage statique et le gain peut atteindre 30%. En revanche lorsque la zone d'étude commence à être recouverte de facon régulière, retourner voir des sites déjà visités conduit à une augmentation du pourcentage de faux positifs.

7 Conclusion

Cet article présente une méthode originale basée sur le krigeage pour le choix d'un échantillon spatial dans le but de reconstruction d'une carte d'occurence. Deux stratégies sont étudiées : l'échantillonnage statique et l'échantillonnage adaptatif. La résolution du problème d'optimisation défini par la méthode exacte n'étant pas accessible, nous présentons une méthode approchée, ainsi que les hypothèses simplificatrices dont elle découle. Une analyse dans le cas de données simulées et à paramètres connus montre que, malgré cette approximation, les reconstructions obtenues par échantillonnage statique ou adaptatif présentent des taux d'erreur très raisonnables. Le gain obtenu avec la stratégie adpatative, lié au fait que l'on intègre les observations obtenues aux étapes d'échantillonnage intermédiaires pour sélectionner l'échantillon suivant, est significatif. Même face aux méthodes d'échantillonnages aléatoire et systématique, couramment utilisées.



(a)



(b)

FIG. 4 – Erreurs de reconstruction pour les échantillonnages statique et adaptatif. De haut en bas : pourcentage de faux positifs, faux négatifs, et erreur globale ; (a) $r = 15$, (b) $r = 30$.

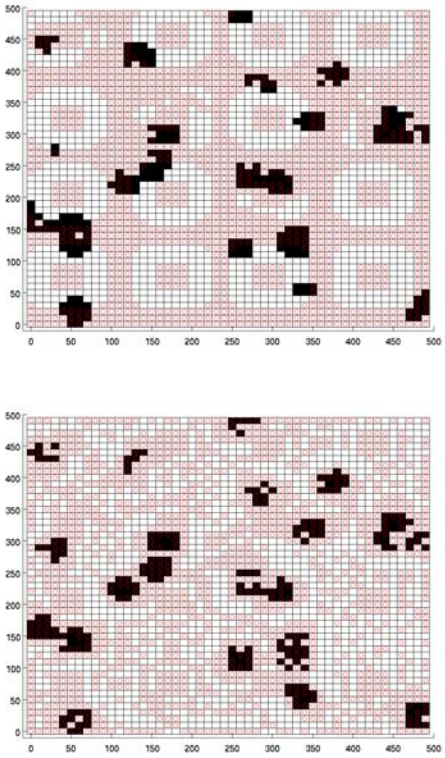


FIG. 5 – Exploration de la zone d'étude pour les échantillonnages statique (haut) et adaptatif (bas). Une croix rouge indique les sites qui ont été visités, un fond blanc indique un site pour lequel le phénomène a été reconstruit comme absent, un fond noir indique un site pour lequel le phénomène a été reconstruit comme présent ou un site où le phénomène a été observé.

Une analyse plus poussée des performances de ces méthodes reste néanmoins nécessaire. Il s'agit, entre autres, de tester la qualité de la reconstruction dans le cas de cartes d'occurrence réelles. Cela demande d'estimer les paramètres du modèle booléen (ou, plus généralement, d'estimer un variogramme et une espérance [5]) à partir de données incomplètes et bruitées, puisque nous nous plaçons dans le cas où la seule information disponible est celle fournie par l'échantillonnage initial. Nous avons mis en œuvre la méthode des moindres carrés pondérés [5] mais elle n'a pas conduit à des résultats satisfaisants, du fait de la faible taille de l'échantillon initial. D'autres approches doivent être explorées ou développées.

Par ailleurs nous avons développé une méthode pour l'échantillonnage spatial inspirée des approches utilisées en analyse d'image [8]. Cette méthode repose sur une modélisation par champ de Markov caché et le critère du Maximum Posterior Marginal (MPM). Une étude comparative des deux approches est en cours.

Références

[1] M. Buesco, J. Angulo, and Alonso F. A state-space model approach to optimum spatial sampling design based on entropy. *Environmental and Ecological Statistics*, (5) :29–44, 1998.

[2] C.T. Chao and S.K. Thompson. Optimal adaptative selection of sampling sites. *Environmetrics*, 12 :517–538, 2001.

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

[4] D.D. Cox, L.H. Cox, and K.B. Ensor. Spatial sampling and the environment : some issues and directions. *Environmental and Ecological Statistics*, 4 :219–233, 1997.

[5] N. Cressie. *Statistics for spatial data*. Wiley series in probability and statistics, 1993.

[6] J. de Gruijter, D. Brus, M. Bierkens, and M. Knotters. *Sampling for natural resource monitoring*. Springer, 2006.

[7] M. Fuentes, A. Chaudhuri, and D. Holland. Bayesian entropy for spatial sampling design of environmental data. *Environmental and ecological Statistics*, (14) :323–340, 2007.

[8] N. Peyrard, R. Sabbadin, D. Spring, R. Mac Nally, and B. Brook. Model-based adaptive spatial sampling for fire ants invasion map construction. *Rapport de Recherche UBIAT-INRA*, 2009.

[9] A. Siefi and M.J. Karimifar. Entropy based spatial design : A genetic algorithm approach (case study). *World academy of science, engineering and technology*, 45 :637, 2008.

[10] D. Stoyan, W. Kendall, and J.Mecke. *Stochastic geometry and its applications*. Wiley series in probability and statistics, 1995.

[11] J.W. van Groenigen, W. Siderius, and A. Stein. Constrained optimisation of soil sampling for minimisation of the kriging variance. *Geoderma*, 87 :239–259, 1999.

Annexe

Montrons que la méthode approchée statique découle du problème exact sous 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 $k \in \{1, \ldots, K\}$:

$$\mathbb{E}[Z_s Z_{\alpha_k} \mid y^{init}] = \mathbb{E}[Z_s \mid y^{init}]\, \mathbb{E}[Z_{\alpha_k} \mid y^{init}]$$

Le critère U^{kri} peut se réécrire comme une somme de variances du krigeage conditionnel en tout point s de Δ. Cette variance s'exprime simplement en fonction des espérances et covariances conditionnelles, comme dans le cas du krigeage simple :

$$\begin{aligned} & U^{kri}(\alpha_1, \ldots, \alpha_K) \\ = & -\sum_{s\in\Delta} \mathbb{E}[(p^*(Y^{obs}, y^{init}, s) - Z_s)^2 \mid y^{init}] \\ = & -\sum_{s\in\Delta} \mathrm{Var}[(p^*(Y^{obs}, Y^{init}, s) - Z_s)^2 \mid y^{init}] \\ = & -\sum_{s\in\Delta} \left(m'_s(1-m'_s) - \sum_{k=1}^{K} \lambda^*_{\alpha_k} \sigma'_{s\alpha_k}\right) \end{aligned}$$

Si $s \in \{\alpha_1, \ldots, \alpha_K\}$, par exemple $s = \alpha_1$, alors sous la première hypothèse la prédiction de Z_s par $p^*(Y^{obs}, y^{init}, s)$ est exacte. En effet, dans ce cas

$$p^*(Y^{obs}, y^{init}, s) \quad = \quad m'_s - \sum_{k=1}^{K} \lambda^*_k m'_{\alpha_k} + \sum_{k=1}^{K} \lambda^*_k Z_{\alpha_k}$$

Dans ce cas, une solution triviale du système du krigeage conditionnel en α_k est $\lambda^*_k = 1$ si $k = 1$ et 0 sinon. Donc

$$p^*(Y^{obs}, y^{init}, \alpha_1) = m'_{\alpha_1} - m'_{\alpha_1} + Z_{\alpha_1} = Z_{\alpha_1}$$

Ainsi les termes portant sur les sites $s \in \{\alpha_1, \ldots, \alpha_K\}$ dans U^{kri} disparaissent :

$$\begin{aligned} & U^{kri}(\alpha_1, \ldots, \alpha_K) \\ = & -\sum_{s\in\Delta\setminus\{\alpha_1,\ldots,\alpha_K\}} \left(m'_s(1-m'_s) - \sum_{k=1}^{K} \lambda^*_{\alpha_k} \sigma'_{s\alpha_k}\right) \end{aligned}$$

Par ailleurs, sous les deux hypothèses simplificatrices, $\sigma'_{s\alpha_k} = 0$ puisque

$$\sigma'_{s\alpha_k} = \mathbb{E}[Z_s Z_{\alpha_k} \mid y^{init}] - \mathbb{E}[Z_s \mid y^{init}]\mathbb{E}[Z_{\alpha_k} \mid y^{init}]$$

On obtient alors

$$U^{kri}(\alpha_1, \ldots, \alpha_K) = -\sum_{s\in\Delta\setminus\{\alpha_1,\ldots,\alpha_K\}} m'_s(1-m'_s)$$

Enfin

$$\begin{aligned} & \max_{\{\alpha_1,\ldots,\alpha_K\}\subseteq\Delta} U^{kri}(\alpha_1, \ldots, \alpha_K) \\ = & \min_{\{\alpha_1,\ldots,\alpha_K\}\subseteq\Delta} \sum_{s\in\Delta\setminus\{\alpha_1,\ldots,\alpha_K\}} m'_s(1-m'_s) \\ = & \max_{\{\alpha_1,\ldots,\alpha_K\}\subseteq\Delta} \sum_{s\in\{\alpha_1,\ldots,\alpha_K\}} m'_s(1-m'_s) \\ = & \max_{\{\alpha_1,\ldots,\alpha_K\}\subseteq\Delta} \sum_{k=1}^{K} \mathbb{P}(Z_{\alpha_k} = 1 \mid y^{init}) \times \mathbb{P}(Z_{\alpha_k} = 0 \mid y^{init}) \end{aligned}$$

Comme les fonctions $p(1-p)$ et $\min(p, 1-p)$ atteignent leur maximum pour la même valeur de p, trouver $\{\alpha_1, \ldots, \alpha_K\}$ qui maximise l'expression ci-dessus est équivalent à trouver $\{\alpha_1, \ldots, \alpha_K\}$ qui maximise $\tilde{U}^{kri}(\alpha_1, \ldots, \alpha_K)$.