Prédire l'intensité locale d'un processus ponctuel partiellement observé

Edith Gabriel [1,2] & Florent Bonneu [1] & Pascal Monestiez [2,3] & Joël Chadœuf [4]

[1] *Université d'Avignon et des Pays de Vaucluse, LMA EA 2151, 84000 Avignon,* edith.gabriel@univ-avignon.fr *et* florent.bonneu@univ-avignon.fr
[2] *INRA, BioSP, UR 546, 84000 Avignon,* monestiez@avignon.inra.fr
[3] *INRA, USC1339, CNRS, UMR 7372, CEBC, 79360 Villiers en bois*
[4] *INRA, Statistics GAFL, UR 1052, 84000 Avignon,* joel@paca.inra.fr

Résumé. Nous considérons un semis de points observé dans une grande fenêtre. Nous supposons que le processus sous-jacent est stationnaire, isotrope et obtenu par un processus à faible dépendance dont un paramètre est dirigé par un champ aléatoire stationnaire à une échelle supérieure. Dans un objectif de prédire l'intensité locale du processus ponctuel dans des zones non-échantillonnées, notre approche consiste à définir les caractéristiques du champ aléatoire à partir de celles du processus ponctuel, puis à interpoler l'intensité locale par un krigeage ordinaire revisité.

Mots-clés. Estimation d'intensité, Prédiction, Processus ponctuel spatial

Abstract. We consider a spatial point pattern observed within a large window. We assume that the underlying process is stationary and isotropic, and that it is obtained by a weak dependent process with a parameter driven by a stationary random field at a larger scale. In order to predict the local intensity, we propose to define the first- and second-order characteristics of the random field from the ones of the point process and to interpolate the local intensity by using a revisited ordinary kriging.

Keywords. Intensity estimation, Prediction, Spatial point process

1 Introduction

Nous parlons d'estimation de l'intensité d'un processus ponctuel lorsque nous avons une observation complète du processus dans une fenêtre et que nous nous intéressons à ses variations locales sur une grille donnée. Cette question, déjà bien connue, a pu être traitée par des approches non-paramétriques de type noyau, voir par exemple Silverman (1986) ou Guan (2008) en présence de covariables, ou des approches paramétriques, voir Illian et al (2008) pour une synthèse sur ces méthodes. Cependant quelle que soit l'approche considérée, une question récurrente concerne le choix de la fenêtre de lissage ou de la maille de la grille d'interpolation. On trouve parmi les solutions existantes la validation croisée (Härdle, 1991) ou le double noyau (Devroye, 1989).

Contrairement aux méthodes précédentes qui regardent les variations de l'intensité locale dans la fenêtre d'observation, notre objectif est aussi de prédire l'intensité locale hors de cette fenêtre. Cette question se pose fréquemment lorsque la fenêtre d'échantillonnage n'est pas connexe, ce qui par exemple le cas en écologie des plantes où l'on va chercher des informations locales tout en gardant une surface d'intérêt très large.

Il existe des méthodes de reconstruction (Tscheschel and Stoyan, 2006) basées sur les caractéristiques d'ordre 1 et d'ordre 2 du processus ponctuel. Une fois le processus prédit dans une zone donnée, son intensité locale peut être estimée par noyau. Cette approche est cependant longue et lourde à mettre en œuvre, car basée sur des simulations, en particulier lorsque la zone de prédiction est grande ou lorsque le processus ponctuel est complexe.

Certains auteurs modélisent le semis de points par un processus ponctuel dont l'intensité est dirigée par un champ aléatoire stationnaire. Dans Diggle et al (2007) et Diggle et al (2013), la procédure repose sur une modélisation complète du processus. Ils considèrent un modèle de Cox log-Gaussien et l'estimation des paramètres, l'estimation de l'intensité et la prédiction hors de la zone d'échantillonnage sont faites dans un cadre bayésien. Ces méthodes nécessitent de disposer d'un modèle complet pour lequel la prédiction doit être possible. L'approche proposée dans Monestiez et al (2006) et Bellier et al (2013) est proche de la géostatistique classique, dans le sens où elle consiste à dénombrer les points dans les cellules d'une grille donnée, calculer le variogramme empirique et le relier théoriquement à celui obtenu à partir du champ aléatoire dirigeant l'intensité. Le variogramme est ensuite ajusté et le krigeage est utilisé pour prédire l'intensité. L'avantage de cette approche sur la précédente est que les caractéristiques d'ordre 1 et d'ordre 2 suffisent à décrire le processus ponctuel, il n'est pas nécessaire d'avoir une modélisation plus fine. Mais bien que cette approche nécessite peu d'hypothèses, le modèle est contraint dans une classe (Cox) et la taille de la grille d'interpolation est arbitraire.

Nous proposons (Gabriel et al, 2014) une approche assez similaire à la dernière, adaptée à une plus large gamme de processus ponctuels et qui permet d'optimiser la taille de la grille d'interpolation. Nous supposons que le processus ponctuel est stationnaire et isotrope, obtenu par un processus à faible dépendance (*e.g.* Poisson, Thomas, Markov) dont un paramètre est dirigé par un champ aléatoire stationnaire à une échelle supérieure (*e.g.* un processus de Cox par rapport à un processus de Poisson).
L'intensité locale s'écrit $\Lambda(x) = \lambda + Y(x)$, où λ représente la moyenne du champ aléatoire et $Y(x)$ est un champ aléatoire centré. Le nombre de points dans un borélien B est donné par

$$N(B) = \Phi(B) = \lambda\nu(B) + \int_B Y(x)\,\mathrm{d}x + \eta, \tag{1}$$

i.e. la somme de la moyenne globale, des variations de l'intensité locale et d'une erreur liée à la différence entre les observations et l'intensité locale.

L'équation (1) étant semblable à la décomposition géostatistique, nous proposons de définir les caractéristiques du champ aléatoire à partir de celles du processus ponctuel.

L'intensité locale peut ensuite être prédite par un krigeage dont les poids dépendent de la structure du processus ponctuel. Cette approche permet d'utiliser l'ensemble des points pour prédire localement, ce qui n'est pas le cas des méthodes à noyau, et utilise l'information à fine échelle du processus ponctuel, ce qui n'est pas le cas des approches classiques de la géostatistique. Nous la détaillons dans les sections suivantes.

2 Lier les caractéristiques d'un processus ponctuel à celles d'un champ aléatoire

Nos données sont un semis de points, *i.e.* la réalisation d'un processus ponctuel Φ, alors que les techniques de géostatistique (comme le krigeage) opèrent sur les valeurs d'un champ aléatoire Z observé en divers points d'échantillonnage, par exemple des centres de cellules d'une grille. Afin de lier les caractéristiques issues de la théorie des processus ponctuels à celles de la géostatistique, nous devons régulariser le processus sur un compact. Cela consiste à définir $Z(x)$ par le comptage $\Phi(B)$ du processus ponctuel sur la cellule B centrée en x, *i.e.* $Z(x) = \Phi(x \oplus B)$.

Géostatistique Soit Z un champ aléatoire à valeurs réelles. Son moment d'ordre 1 correspond à la fonction moyenne $\mathbb{E}[Z(x)] = m(x)$, et son moment d'ordre 2 est classiquement décrit en géostatistique par le (semi)-variogramme qui correspond à l'écart quadratique moyen à distance h : $\gamma(h) = \frac{1}{2}\mathbb{E}\left[(Z(x) - Z(x+h))^2\right]$. Nous supposons que le champ aléatoire est stationnaire et isotrope, aussi nous avons les relations suivantes

$$\begin{aligned} \mathbb{E}[Z(x)] &= m, \\ \gamma(h) &= \sigma^2 - \mathbb{C}\mathrm{ov}(Z(x), Z(x+h)), \end{aligned}$$

où σ^2 désigne la variance du champ.

Processus ponctuels Soit Φ un processus ponctuel dans $\mathbb{R}^2$ observé dans une fenêtre W_S. Les caractéristiques d'ordre 1 et 2 de Φ sont données par l'intensité d'ordre 1, λ, et l'intensité d'ordre 2, λ_2, ou de façon équivalente par la fonction K de Ripley ou la fonction de corrélation de paire (g) :

$$\begin{aligned} \lambda &= \frac{\mathbb{E}[\Phi(W_S)]}{\nu(W_S)}, \\ K^*(r) &= \frac{1}{\lambda}\mathbb{E}\left[\Phi(b(0,r)) - 1 | 0 \in \Phi\right], \\ g(r) &= \frac{1}{2\pi r}\frac{\partial K^*(r)}{\partial r}, \end{aligned}$$

où K^* désigne la fonction K de Ripley privée du point 0 et $b(0,r)$ est le disque centré en 0 et de rayon r.

Liens La proposition suivante donne les expressions des caractéristiques d'ordre 1 et 2 du champ aléatoire régularisé $Z(x) = \Phi(x \oplus B)$ en fonction des caractéristiques du processus ponctuel Φ. Ayant supposé la stationnarité du champ aléatoire, nous pouvons indifféremment utiliser la fonction de covariance ou le variogramme du champ aléatoire.

Proposition 1 *Pour $Z(x) = \Phi(x \oplus B)$ où B est un borélien, nous avons*

1. $m = \lambda\nu(B)$,
2. *Pour B et D deux blocs de régularisation, $B_D = B\backslash D$, $D_B = D\backslash B$,*

$$\begin{aligned} 2\gamma(B,D) &= \lambda\left(\nu(B_D)+\nu(D_B)\right) + \lambda^2\left(\int_{B_D\times B_D} g(x-y)\,dx\,dy \right. \\ &\quad \left. + \int_{D_B\times D_B} g(x-y)\,dx\,dy - 2\int_{B_D\times D_B} g(x-y)\,dx\,dy\right). \end{aligned}$$

3. *Si B et D sont deux surfaces élémentaires centrées en des points distants de r, alors pour $\nu(B) = \nu(D) \to 0$*

$$\mathbb{C}\mathrm{ov}\left(\Phi(B),\Phi(D)\right) \simeq \lambda\nu(B)\Big(\mathbb{I}_{\{B=D\}} + \lambda\nu(B)\big(g(r)-1\big)\Big).$$

3 Définir un krigeage pour les processus ponctuels

Nous considérons une grille régulière de maille carrée et notons $x = \cup_{i=1}\{x_i\}$ l'ensemble des centres de ces cellules. Soit B un carré élémentaire centré en 0, $B_i = x_i \oplus B$ le carré élémentaire centré en x_i tel que $B_i \cap B_j = \emptyset$. On note $S = \cup_{i=1}^n B_i$ la zone d'intérêt, *i.e.* l'union de la fenêtre d'observation, $W_S = \cup_{j=1}^n B_j$, et de la zone de prédiction, ainsi $n = \frac{\nu(W_S)}{\nu(B)} \leq \frac{\nu(S)}{\nu(B)} = n_S$.

Définir l'interpolateur Notre objectif est d'interpoler l'intensité locale conditionnelle $\lambda(x|\Phi(W_S))$. L'interpolateur de krigeage ordinaire en x_o devrait être $\lambda(x_o|\Phi(W_S)) = \mu^T\Lambda$ pour des poids μ et des observations $\Lambda = \big(\lambda(x_1|\Phi(W_S)),\ldots,\lambda(x_n|\Phi(W_S))\big)^T$. Cependant, dans notre cas nous n'observons pas l'intensité locale en x_i et l'estimons par

$$\tilde{\lambda}(x_i|\Phi(W_S)) = \tfrac{\Phi(B_i)}{\nu(B)}.$$

Proposition 2 *L'interpolateur de l'intensité locale en x_o défini par*

$$\widehat{\lambda}(x_o|\Phi(W_S)) = \sum_{x_i\in W_S} \mu_i \frac{\Phi(B_i)}{\nu(B)}, \tag{2}$$

où $\mu = (\mu_1,\ldots,\mu_n) = C^{-1}C_o + \dfrac{1-\mathbf{1}^TC^{-1}C_o}{\mathbf{1}^TC^{-1}\mathbf{1}}C^{-1}\mathbf{1}$, est le meilleur estimateur linéaire sans biais (best linear unbiased predictor - BLUP). Les poids dépendent de

– *la matrice de covariance* $C = \lambda\nu(B)\,[\boldsymbol{I} + \lambda\nu(B)(G-1)]$,
où $G = \{g_{ij}\}_{i,j=1,\dots,n}$, *avec* $g_{ij} = \frac{1}{\nu^2(B)} \int_{B\times B} g(x_i - x_j + u - v)\,du\,dv$, *et* $\boldsymbol{I}$ *est la* $n \times n$*-matrice identité,*
– *le vecteur de covariance* $C_o = \lambda\nu(B)\mathbb{I}_{x_o} + \lambda^2\nu^2(B)(G_o - 1)$,
où $G_o = \{g_{io}\}_{i=1,\dots,n}$, *et* $\mathbb{I}_{x_o}$ *est le* n*-vecteur nul ayant un terme égal à un en* $x_o = x_i$ *(ce qui n'est le cas qu'en estimation).*

Propriétés de l'interpolateur La proposition suivante donne la variance de l'interpolateur défini dans l'équation (2).

Proposition 3 *Si* $\nu(B)$ *tend vers 0 et* S *est suffisamment grande, alors*
– *pour* x_o *appartenant à la fenêtre d'observation,* W_S,

$$\mathbb{V}\mathrm{ar}\left(\widehat{\lambda}(x_o|\Phi(W_S))\right) \approx \frac{\lambda}{\nu(B)} \tag{3}$$

– *pour* x_o *hors de la fenêtre d'observation,* $S \backslash W_S$,

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\widehat{\lambda}(x_o|\Phi(W_S))\right) &= \lambda^3\nu^2(B)(G_o-1)^T(G_o-1) \\ &\quad +\lambda^4\nu^3(B)(G_o-1)^T J_\lambda (G_o-1) \\ &\quad +\frac{1-\left[\lambda\nu(B)\mathbf{1}^T(G_o-1)+\lambda^2\nu^2(B)\mathbf{1}^T J_\lambda(G_o-1)\right]^2}{\frac{\nu(S)}{\lambda}+\nu^2(B)\mathbf{1}^T J_\lambda \mathbf{1}}\end{aligned} \tag{4}$$

où $J_\lambda = \sum_{k=1}^{\infty}(-1)^k\lambda^{k-1}H^k$ *et* $H^k = \int_{{W_S}^{k+1}} \prod_{m=1}^{k}(g(x_m, x_{m+1}) - 1)\ dx_1 \dots\ dx_{k+1}$.

Définir la surface optimale de la maille La surface optimale de la maille de la grille d'interpolation est déterminée par minimisation de l'erreur quadratique moyenne intégrée (Integrated Mean Squared Error - IMSE) de $\widehat{\lambda}(x|\Phi(W_S))$:

$$\begin{aligned}IMSE\left(\widehat{\lambda}(x|\Phi(W_S))\right) &= \int_S \left[\left(\lambda(x|\Phi(W_S)) - \mathbb{E}[\widehat{\lambda}(x|\Phi(W_S))]\right)^2 + \mathbb{V}\mathrm{ar}\left(\widehat{\lambda}(x|\Phi(W_S))\right)\right] \mathrm{d}x \\ &\approx \frac{\sqrt{\nu(B)}}{12}\int_S \|\mathrm{grad}(\lambda(x|\Phi(W_S)))\|^2\,\mathrm{d}x + \frac{\lambda\nu(S)}{\nu(B)}.\end{aligned}$$

Ainsi, pour une grille de maille carrée, la surface optimale de la maille est donnée par :

$$\nu_{opt}(B) = \left(\frac{24\lambda\nu(S)}{\int_S \|\mathrm{grad}(\lambda(x|\Phi(W_S)))\|^2\,\mathrm{d}x}\right)^{2/3}.$$

La surface optimale dépend de l'inverse du gradient de l'intensité locale, aussi elle augmente pour des semis de points agrégés et diminue pour des semis de points réguliers (ce à quoi nous pouvions nous attendre, mais il est intéressant de voir à quelle vitesse se fait cette régulation).

4 Résultats sur données simulées et réelles

Cette partie sera présentée durant l'exposé.

Références

Bellier E, Monestiez P, Certain G, Chadœuf J, Bretagnolle V (2013) Reducing the uncertainty of wildlife population abundance : model-based versus design-based estimates. Environmetrics 24(7) :476–488

Devroye L (1989) The double kernel method in density estimation. Les Annales de l'IHP, section B 25(4) :533–12

Diggle P, Gómez-Rubio V, Brown P, Chetwynd A, Gooding S (2007) Second-order analysis of inhomogeneous spatial point processes using case-control data. Biometrics 63(2) :550–557

Diggle P, Moraga P, Rowlingson B, Taylor B (2013) Spatial and spatio-temporal log-gaussian cox processes : Extending the geostatistical paradigm. Statistical Science 28(4) :542–563

Gabriel E, Bonneu F, Monestiez P, Chadoeuf J (2014) Predicting the local intensity of partially observed data from a revisited kriging for point processes. arcXiv 1409.6441, URL `http://arxiv.org/abs/1409.6441`

Guan Y (2008) On consistent nonparametric intensity estimation for inhomogeneous spatial point processes. Journal of the American Statistical Association 103(483) :1238–1247

Härdle W (1991) Smoothing techniques, with implementation in S. Springer & Verlag, New York

Illian J, Penttinen A, Stoyan H, Stoyan D (2008) Statistical Analysis and Modelling of Spatial Point Patterns. John Wiley and Sons, London

Monestiez P, Dubroca L, Bonnin E, Durbec J, Guinet C (2006) Geostatistical modelling of spatial distribution of balaenoptera physalus in the northwestern mediterranean sea from sparse count data and heterogeneous observation efforts. Ecological Modelling 193 :615–628

Silverman B (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall/CRC, London

Tscheschel A, Stoyan D (2006) Statistical reconstruction of random point patterns. Computational Statistics and Data Analysis 51 :859–871