Arbres CART pour données spatiales

Jean-Michel Poggi [1] & Avner Bar-Hen [2] & Servane Gey [3]

[1] *LMO, Univ. Paris Sud, Orsay, France et Univ. Paris Descartes, Paris, France*
Jean-Michel.Poggi@math.u-psud.fr
[2] *CNAM, Paris, France Avner@cnam.fr*
[3] *Laboratoire MAP5, Univ. Paris Descartes, Paris, France*
Servane.Gey@parisdescartes.fr

Résumé. En reliant les partitions induites par les arbres de classification CART aux processus ponctuels marqués, nous proposons une variante spatiale de la méthode CART, SpatCART, qui se concentre sur le cas de deux populations. Alors que l'arbre CART habituel ne tient compte que de la distribution marginale de la variable réponse en chaque nœud, nous proposons de prendre en compte la position spatiale des observations. Nous introduisons un indice de dissimilarité basé sur la fonction intertype K de Ripley qui quantifie l'interaction entre deux populations. Cet indice utilisé pour l'étape de construction de l'arbre maximal de la stratégie CART, conduit à une fonction d'hétérogénéité cohérente avec l'algorithme CART original. La procédure proposée est mise en œuvre, illustrée par des exemples classiques et comparée aux concurrents directs. SpatCART est enfin appliquée à un exemple de l'analyse de deux espèces d'arbres dans une forêt tropicale.

Mots-clés. Arbre de décision, Classification, Processus ponctuel, Données spatiales

Abstract. Based on the links between partitions induced by CART classification trees and marked point processes, we propose a spatial variant of the CART method, SpatCART, focusing on the two populations case. While usual CART tree considers marginal distribution of the response variable at each node, we propose to take into account the spatial location of the observations. We introduce a dissimilarity index based on Ripley's intertype K-function quantifying the interaction between the two populations. This index used for the growing step of the CART strategy, leads to a heterogeneity function consistent with the original CART algorithm. The proposed procedure is implemented, illustrated on classical examples and compared to direct competitors. SpatCART is finally applied to a tropical forest example.

Keywords. Decision Tree, Classification, Point process, Spatial data

1 Introduction

L'algorithme CART (Classification and Regression Trees), proposé par Breiman *et al.* [2] en 1984, construit des classifieurs constants par morceaux par découpe dyadique récursive

de l'espace des variables explicatives $\mathcal{X}$. Ces classifieurs sont représentés sous forme d'arbres binaires, rendant leur interprétation à la fois simple et rapide. La représentation classique par arbres binaires de la relation existant entre les covariables X et la variable expliquée Y traduit le processus de construction du modèle CART comme une partition récursive de l'espace des covariables. Lorsque cet espace est planaire, on obtient alors des arbres spatiaux : les points à l'intérieur des cellules définies par la partition sont une réalisation d'un processus spatial, marqué par les étiquettes affectées à chaque cellule. Suivant cette idée, la variante de l'algorithme CART que nous proposons prend en compte la dimension spatiale des processus marqués, et permet ainsi de fournir à l'usager un premier pavage du plan définissant des zones homogènes d'interaction entre les marques.

2 Algorithme CART spatial : SpatCART

La variante que nous proposons repose sur la fonction intertype K, dérivée de la fonction K de Ripley (proposée dans [4]), et qui mesure l'interaction entre les marques i et j d'un processus spatial. Alors que la fonction K de Ripley caractérise la structure spatiale d'un processus ponctuel univarié, la fonction intertype K_{ij} caractérise la structure spatiale d'un processus bivarié, et plus précisément la relation spatiale entre deux types de points localisés dans la même aire d'étude. De fait, K_{ij} est définie de telle sorte que $\lambda_j K_{ij}(r)$ est l'espérance du nombre de points de type j dans un disque de rayon r et de centre un point de type i pris aléatoirement, où λ_j est l'espérance du nombre de points de marque j par unité d'aire. De manière symétrique, on définit $\lambda_i K_{ji}(r)$, et, si le processus est stationnaire et homogène, alors on a l'identité $K_{ij} = K_{ji}$.

L'idée ici est d'utiliser la fonction intertype K_{ij} comme critère de découpe dans CART : au lieu de se faire uniquement sur les proportions des marques de manière à les séparer au mieux, les découpes se feront selon la force d'interaction entre les deux populations de manière à définir des zones d'interaction homogènes. Pour ce faire, on maximise sur toutes les découpes possibles la fonction d'impureté définie pour un nœud t de l'arbre, une découpe s de t en deux nœuds fils t_L et t_R, et une échelle $r > 0$, par :

$$\Delta I_{ij}(s,t,r) := \hat{K}_{ij}^{t}(r) - \alpha_s \frac{\hat{\lambda}_i^{t_L}\hat{\lambda}_j^{t_L}}{\hat{\lambda}_i^{t}\hat{\lambda}_j^{t}} \hat{K}_{ij}^{t_L}(r) - (1-\alpha_s) \frac{\hat{\lambda}_i^{t_R}\hat{\lambda}_j^{t_R}}{\hat{\lambda}_i^{t}\hat{\lambda}_j^{t}} \hat{K}_{ij}^{t_R}(r) \tag{1}$$

avec

A^t	aire du nœud t
$\alpha_s = \frac{A^{t_L}}{A^t}$	proportion d'aire occupée par $t_L \subset t$
$\hat{\lambda}^t_*,\ * = i, j$	estimation de la densité de la marque $*$ dans le nœud t
$d_{i_k j_l}$	distance euclidienne entre les individus i_k et j_l
$\hat{K}^t_{ij}(r) = (\hat{\lambda}^t_i \hat{\lambda}^t_j A^t)^{-1} \sum_{\{i_k, j_l \in t\}} 1_{d_{i_k,j_l} < r}$	estimation de la fonction intertype restreinte aux points i_k de marque i et j_l de marque j appartenant au nœud t

Nous considérons ici la version non-pondérée de l'estimateur de K_{ij}, qui permet notamment de conserver la symétrie entre les marques. Par ailleurs, cet estimateur permet d'assurer que le critère d'impureté Δ_{ij} est positif, condition nécessaire pour obtenir la décroissance de l'impureté des nœuds le long des branches de l'arbre.
La fonction d'impureté traduit la variation d'hétérogénéité venant de la découpe du nœud t en utilisant s. Sa forme est classique, à la proportion α_s près, qui est naturelle dans le cas de données spatiales, puisqu'elle pondère proprement les aires des deux nœuds fils. Dans toute la suite, nous supposerons donc que le processus non-marqué est stationnaire afin d'assurer une certaine robustesse dans l'algorithme.
Le paramètre r intervenant dans le critère ΔI_{ij} donné par (1) est un paramètre d'échelle permettant de zoomer sur les points d'intérêt du processus spatial : pour chaque point i_k du processus considéré, seuls les points à distance au plus r de i_k sont pris en compte. Le paramètre est initialisé à une valeur r_0, choisie par l'utilisateur en fonction des données, et qui fixe l'échelle initiale dans la fenêtre définissant la racine de l'arbre. Cette échelle est ensuite recalculée automatiquement à chaque nœud, lui associant ainsi une échelle locale, qui décroît naturellement avec la profondeur de l'arbre. Lorsque l'échelle initiale r_0 est bien choisie, ces changements de focale successifs permettent d'obtenir des arbres plus équilibrés, tout en évitant le sur-apprentissage dans les découpes les plus profondes.

Une fois l'arbre maximal construit, deux variantes peuvent être utilisées pour la phase d'élagage : (a) via le critère classique : le taux de mal classés pénalisé par le nombre de feuilles ou (b) via le critère de Gini pénalisé, proposé par Breiman *et. al.* pour les arbres de probabilité des classes

$$\mathrm{crit}^G_\alpha(T) = \frac{1}{n} \sum_{t \in \widetilde{T}} n_t \left(1 - \sum_{*=i,j} \hat{p}(*|t)^2 \right) + \alpha \frac{|\widetilde{T}|}{n}, \tag{2}$$

où n désigne le nombre total de points dans la fenêtre considérée, $\widetilde{T}$ l'ensemble des feuilles de l'arbre T, n_t le nombre d'individus présents dans le nœud t, $\hat{p}(*|t)$ la proportion de points de marque $*$ présents dans le nœud t, et $|\widetilde{T}|$ le nombre de feuilles de T.

La première variante sera utilisée dans le cas où l'on souhaite simplement séparer les marques, alors que la seconde sera plutôt utilisée dans le cas où l'on souhaite estimer les intensités locales des marques. En effet, les arbres de probabilité des classes sont conçus pour estimer les probabilités des marques dans chaque zone du plan. Lorsque le processus non-marqué est stationnaire, ces probabilités conditionnelles à une partie du plan sont directement proportionnelles aux intensités locales des marques sur cette partie du plan.

Enfin, l'arbre optimal est choisi dans la suite issue de l'élagage par la méthode du plus grand plateau de complexité, dérivé de l'heuristique de pente (voir [1]). Celle-ci a l'avantage de sélectionner un arbre sur la base de tous les points, contrairement aux méthodes classiques qui nécessitent de découper l'échantillon en un ou plusieurs sous-échantillons. Vues la spatialité des données et du critère de découpe, retirer des points introduirait un biais trop important dans l'estimation.

3 CART et SpatCART en action sur des simulations

Étant donné que SpatCART diffère des variantes classiques de CART uniquement dans le critère de découpe, la comparaison par simulations se focalise essentiellement sur les partitions obtenues par les deux algorithmes, et non sur la qualité des classifieurs obtenus. L'idée ici est de contrôler que les deux algorithmes se comportent comme attendu sur deux exemples typiques :
- (a) un exemple témoin où le classifieur de Bayes est un arbre de classification. Dans ce cas, CART et SpatCART doivent tous les deux retrouver une partition proche de la partition définie par le modèle sous-jacent.
- (b) un exemple où le problème de classification est trivial, et où l'interaction entre les marques joue un rôle important. Dans ce cas, SpatCART doit produire des découpes sensiblement différentes de celles produites par CART.

L'exemple (a) est un processus de Poisson marqué bivarié sur la fenêtre unité, dont les marques sont réparties suivant un damier à 9 cases, avec un bruit de classification uniforme, fixé à 5%. Une étude de simulations a été effectuée sur ce modèle et valide le comportement du critère de découpe de SpatCART dans ce cas témoin : les arbres produits par CART et SpatCART diffèrent du classifieur de Bayes sur moins de 4% des points. Par ailleurs, si CART obtient en moyenne de meilleurs résultats en terme de qualité de classification, il apparaît que SpatCART est beaucoup plus stable.

L'exemple (b) est un processus de Poisson marqué bivarié sur la fenêtre unité, avec une marque majoritaire (95% des points). La fenêtre est séparée en deux sous-fenêtres : une partie (à gauche) où les points de la marque minoritaire repoussent ceux de la marque majoritaire à au moins $0,05$ unités, et une partie (à droite) où les deux marques sont

réparties de manière indépendante. On obtient ainsi un processus inhomogène avec deux sous-régimes d'interaction entre les marques. En terme de classification, ce modèle est trivial, étant donné que le classifieur de Bayes est constant égal à la marque majoritaire. Les classifieurs obtenus par CART et SpatCART sont ainsi réduits au vote de la classe majoritaire. L'apport spatial de SpatCART s'illustre dans les découpes successives pour la construction de l'arbre maximal : CART échoue à repérer les deux zones où les régimes d'interaction diffèrent, SpatCART y réussit dès que l'échelle initiale r_0 utilisée dans le critère de découpe est plus petite que le paramètre de répulsion de la zone de gauche.

4 Zones homogènes d'interaction entre deux espèces d'arbres à Paracou

CART et SpatCART sont ainsi ensuite appliqués sur une parcelle expérimentale de forêt tropicale localisée à Paracou, à l'Ouest de Kourou en Guyane française. On trouvera la description des données et des cartes dans [3]. Nous nous intéressons ici à l'interaction entre deux espèces d'arbres présentes dans cette parcelle, *Vouacapoua americana* et *Oxandra asbeckii*, toutes deux localisées sur les sommets et flancs de collines. L'altitude est le facteur environnemental qui dirige leurs distributions spatiales et qui crée une forte interaction entre ces deux espèces. Nous nous intéressons ici à une carte des espèces référençant 70 arbres *Vouacapoua americana*, et 80 arbres *Oxandra asbeckii*. Les résultats obtenus sont représentés Figure 2. Sur chaque figure, les lignes de contour permettent de déterminer les sommets et flancs de collines.

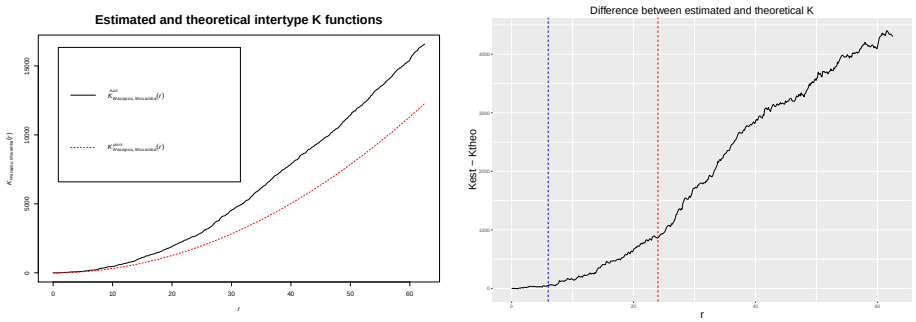

Figure 1: *Gauche* : Fonction intertype K et sa valeur théorique dans le cas indépendant pour les données de Paracou. *Droite* : Différence entre les fonctions intertype K estimées et théoriques ; `bleu`: échelle $r = 6$, `rouge`: échelle $r = 24$.

Les résultats de la Figure 2 montrent que SpatCART (arbre de classification et arbre de probabilité des marques) indique la présence de *Oxandra asbeckii* sur la colline en haut à gauche de la carte, ainsi que la compétition entre les deux espèces pour la colline en bas de la carte. En revanche, les résultats de CART sont très pauvres, avec un arbre à

2 feuilles seulement (les deux variantes de critères pénalisés donnent le même résultat). Celui-ci ne retrouve que la colline en haut de la carte, mais sans détecter la structure mixte de cette colline, ni la colline en haut à gauche. Sur cet exemple, CART ne permet pas de retrouver la structure spatiale ni l'écologie des deux espèces.

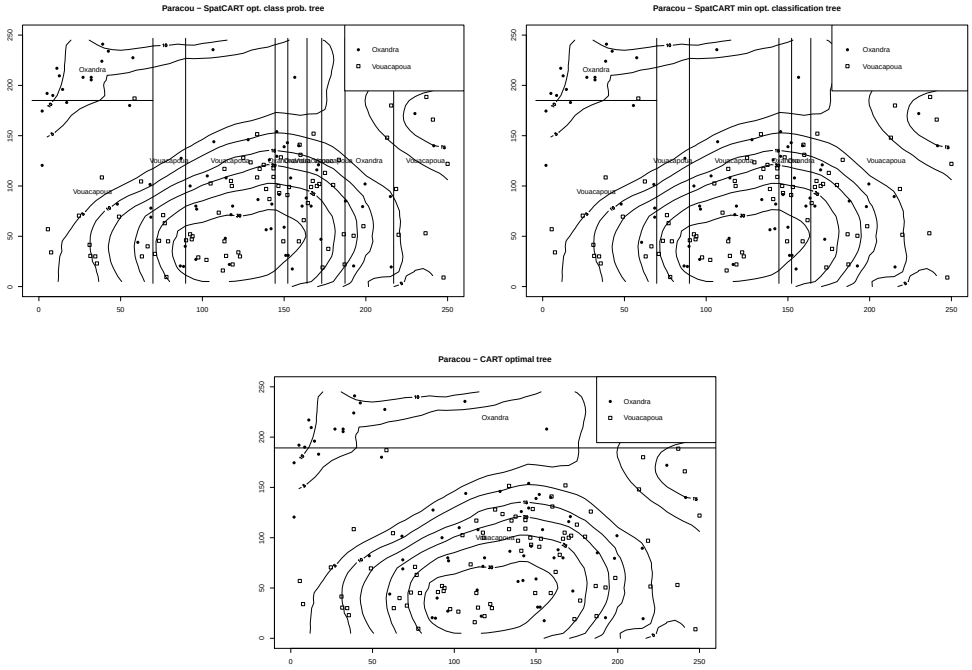

Figure 2: Arbres optimaux obtenus par SpatCART et CART sur les données de Paracou.

References

[1] JP Baudry, C Maugis and B Michel. Slope heuristics: overview and implementation, *Statistics and Computing*, 22(2), 455-470, 2012

[2] L Breiman, JH Friedman, RA Olshen and CJ Stone. *Classification and regression trees*. Chapman & Hall, 1984

[3] S Gourlet-Fleury, JM Guehl and O Laroussinie (eds). Ecology and Management of a Neotropical Rainforest: Lessons Drawn from Paracou, a Long-term Experimental Research Site in French Guiana. Paris: Elsevier, 2004

[4] HW Lotwick and BW Silverman. Methods for analysing spatial processes of several types of points, *Journal of the Royal Statistical Society, Series B*, 44(3):406–413, 1982