Spatial CART Classification Trees

Avner Bar-Hen, Servane Gey, Jean-Michel Poggi

Spatial CART Classification Trees

Avner Bar-Hen[*], Servane Gey[†], Jean-Michel Poggi[‡]

Abstract

Based on links between partitions induced by CART classification trees and marked point processes, we propose a variant of spatial 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 two populations. This index used for the growing step of the CART strategy, leads to a heterogeneity function consistent with the CART original algorithm. Then different pruning strategies, including the classical pruning step using the misclassification rate, are performed. The proposed procedure is implemented, illustrated on classical examples and compared to natural competitors. SpatCART is finally applied to a tropical forest example.

1 Introduction

CART (Classification And Regression Trees) is a statistical method, introduced by Breiman *et al.* [7], and designing tree predictors for both regression and classification. The general principle is to partition recursively the input space using binary splits and then determine an optimal partition for prediction. Let us restrict our attention on the classification case with two populations. Each observation is characterized by some input variables gathered in vector X and a binary label Y which is the output or response variable.

The classical representation of the model relating Y to X is a tree representing the underlying process of construction of the model as a recursive partitioning of the

[*]Cnam, Paris, France
[†]Laboratoire MAP5, Univ. Paris Descartes, France
[‡]Laboratoire de Mathématiques, Univ. Paris Sud, Orsay, France and Univ. Paris Descartes, France

space of explanatory variables. In the special case where the explanatory variables are spatial coordinates, we get a spatial decision tree. Focusing on the observations of a cell of a given partition which is a piecewise constant function, it is possible to consider it as points of the space variables and class assignment as the mark of those points, leading to a natural link with the representation of binary marked point processes.

Even if CART is often limited to the one most commonly used and implemented, presented in the book of Breiman et al. [7], there exist several ways to build CART type decision trees, by changing the family of admissible splits, the cost function or the stopping rule. A classical assumption is to consider a sample of i.i.d. observations, our variant take into account some spatial dependence through the quantification of the interaction between the two considered populations, *i.e.* the link between the labels of the points.

In Section 2, we recall some basics about CART decision trees in the classification case and some variants and extensions. In Section 3, we review some conventional ways to quantify the interaction between two point processes and define the heterogeneity function associated with the Ripley's intertype K-function. Then in Section 4 we propose a variant of spatial CART dealing with point processes and describe how the variants are implemented and illustrates its use on classical examples. Section 5 finally addresses an application to the spatial distribution of two species of tropical forest.

2 CART method

Let us briefly recall, following Bel *et al.* [4], some general background on classical settings about Classification And Regression Trees (CART). The data are considered as an independent sample of the random variables $(X^1, \ldots, X^p, Y)$, where the X^ks are the explanatory variables (supposed to be numerical in this article) and Y is the categorical variable to be explained. CART is a rule-based method that generates a binary tree through recursive partitioning that splits a subset (called a node) of the data set into two subsets (called sub-nodes) according to the minimization of a heterogeneity criterion computed on the resulting sub-nodes. Each split involves a single variable. Some variables may be used several times while others may not be used at all.

2.1 CART classification trees

Let us consider a decision tree T. When Y is a categorical variable a class label is assigned to each terminal node (or leaf) of T. Hence T can be viewed as a mapping

to assign a value $\widehat{Y}_i = T(X_i^1, \ldots, X_i^p)$ to each observation. The growing step leading to a deep maximal tree is obtained by recursive partitioning of the training sample by selecting the best split at each node according to some heterogeneity index, such that it is equal to 0 when there is only one class represented in the node to be split, and is maximum when all classes are equally frequent. The two most popular heterogeneity criteria are the Shannon entropy and the Gini index. Among all binary partitions of each set of values of the explanatory variables at a node t, the principle of CART is to split t into two sub-nodes t_L and t_R according to a threshold on one of the variables (or a subset of the labels for categorical variables), such that the reduction of heterogeneity between a node and the two sub-nodes is maximized. The growing procedure is stopped when there is no more admissible splitting. Each leaf is assigned to the most frequent class of its observations. Of course, such a maximal tree (denoted by T_{max}) generally overfits the training data and the associated prediction error $R(T_{max})$, with

$$R(T) = \mathbb{P}(T(X^1, \ldots, X^p) \neq Y), \tag{1}$$

is typically large. Since the goal is to build from the available data a tree T whose prediction error is as small as possible, in a second stage the tree T_{max} is pruned to produce a subtree T' whose expected performance is close to the minimum of $R(T')$ over all binary subtrees T' of T_{max}. Since the joint distribution $\mathbb{P}$ of $(X^1, \ldots, X^p, Y)$ is unknown, the pruning is based on the penalized empirical risk $\hat{R}_{pen}(T)$ to balance optimistic estimates of empirical risk by adding a complexity term that penalizes larger subtrees. More precisely the empirical risk is penalized by a complexity term, which is linear in the number of leaves of the tree:

$$\hat{R}_{pen}(T) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{T(X_i^1, \ldots, X_i^p) \neq Y_i} + \alpha |T| \tag{2}$$

where $\mathbb{1}$ is the indicator function, n the total number of observations, α a positive penalty constant, $|T|$ denotes the number of leaves of the tree T and Y_i is the ith random realization of Y.

2.2 Variants of CART trees

Let us start this section by a direct variant of CART designed to predict probabilities.

2.2.1 CART class probability trees

In the case of class probability trees, one aims at predicting probabilities $p(j|x) = P(Y = m_j \mid X = x)$ rather than labels $Y \in \{m_1; \ldots; m_J\}$. To construct such trees,

the growing step is the same as the one used for classification trees, using either the Shannon entropy or the Gini index. The only difference relies on the pruning step, derived from the mean square error risk constructed as follows: let $(Z_1, \ldots, Z_J)$ be defined by $Z_j = \mathbb{1}_{Y=m_j}$, and let $p(j|x) = P(Y = m_j \mid X = x)$, $j = 1; \ldots; J$. Then $E(Z_j \mid X = x) = p(j|x)$ and the mean square error of the class probability tree $T = (T_1, \ldots, T_J)$ is given by

$$R_G(T) = E\left[\sum_{j=1}^{J} \left(Z_j - T_j(X^1, \ldots, X^p)\right)^2\right]. \tag{3}$$

R_G is minimum for the Bayes estimator $(P(Y = m_j \mid X = x))_{j=1;\ldots;J}$, and it is proved in [7], section 4.6, that the empirical version of the risk R_G can be written using Gini index, leading to the following penalized criterion:

$$\hat{R}_{pen_G}(T) = \frac{1}{n} \sum_{t \in \widetilde{T}} n_t \left(1 - \sum_{j=1}^{J} \hat{p}(j|t)^2\right) + \alpha|T|, \tag{4}$$

where $\widetilde{T}$ denotes the set of the leaves of T, n_t the number of observations falling in note t, and $\hat{p}(j|t)$ the proportion of observations with mark j falling in node t.

2.2.2 Other variants

As mentioned in the introduction, variants of CART trees have been proposed for various purposes.

Let us first mention an extension to spatial data in [4] in the regression context, for environmental data. It considers the observations as a sample of regionalized variables, which can be modeled as random fields with spatial dependence. It proposes various reweighing schemes to equalize the contribution of each surface unit for the choices of splits. It is not adapted to the classification case.

To determine optimal splits, chosen among a family of deterministic admissible splits, eliminating the dependency of the family of models with respect to data, which is one of the difficulties to derive theoretical results. In [12] for applications in image processing, the considered admissible splits are dyadic. Indeed, the initial dyadic splitting of a rectangle map is deterministic and complete until the resolution of the pixel, it is the analogue of the maximal tree and pruning is performed with the algorithm for the choice of the best basis for wavelet packets. Similar ideas have been extended to the general settings by introducing the so-called dyadic CART, see [5].

There are also extensions, in the regression case, which build smoother predictors as usual trees that are piecewise constant, for example, the algorithm MARS introduced by [14]. We should also mention one of the most used extensions: CART

methods for survival data, for example [18] and [21], as well as the most recent review article [6]). The recent book [27] presenting more widely the methods based on recursive partitioning and also variants of CART for longitudinal data or functional data. In this line, note the use of CART in chemometrics in [25].

3 Quantifying the link between two point patterns

3.1 Basics on point processes

A point process is a random variable that gives the localization of events in a set $W \subset \mathbf{R}^d$. Another way to define a given point process is to consider, for each $B \subset W$, the number of events $\phi(B)$ occurring in B, where ϕ is the distribution of the occurrences of the point process.
Since characterization of a spatial repartition is strongly dependent on the scale of observation, the repartition has to be characterized for each scale.

There are classical assumptions about point processes. At first we consider that the probability to observe two points at the same place is null. Up to misrecording, this hypothesis is not very restrictive. Two extra common hypotheses are stationarity and isotropy. Intuitively, it means that the distribution of the occurrences of the point process ϕ is not affected by translation or rotation around the origin. Moreover, this means that the characteristics of the point process are the same for the whole area under study.
A marked point process is a point process such that a random mark is associated with each localization. In this article, we only consider bivariate point processes, *i.e.* the mark is a qualitative random variable with two possible issues. Equivalently, the bivariate point process can be viewed as the realization of two point processes (one par level of the mark).
There are several ways to consider the relationships between two clouds of points, mainly related to three aspects: independence, association and random labelling (see [3] for example). It ends up that relationships between two clouds of points can be described in various ways and therefore many indices can be defined. Each index will give a specific information about these relationships and will greatly depends on the point process that leads to the observed repartition. For bivariate point processes, many tools based, using first-order characteristics of the point processes, may be used to quantify departure from independence (see [9] for example).

3.2 Intertype K-function

3.2.1 Definition

Under the assumptions of stationarity and isotropy, the intertype K_{ij}-function is a bivariate extension of Ripley's K-function, proposed in [20], and defined as

$$K_{ij}(r) = \lambda_j^{-1}\mathbb{E}(\text{number of points of type } j \text{ within distance } r \text{ of a randomly chosen point of type } i) \tag{5}$$

where the intensity parameters λ_i and λ_j correspond to the expected numbers of type i and type j points per unit area, respectively.

While Ripley's K-function characterizes the spatial structure of a univariate pattern at various scales, the intertype K_{ij}-function characterizes the spatial structure of a bivariate pattern, and more precisely the spatial relationship between two types of points located in the same study area. The intertype K_{ij}-function is defined so that $\lambda_j K_{ij}(r)$ is the expected number of type j points in a circle of radius r centered on an arbitrary type i point of the pattern. Symmetrically, we can define an intertype K_{ji}-function so that $\lambda_i K_{ji}(r)$ is the expected number of type i points in a circle of radius r centered on an arbitrary type j point.
If the bivariate spatial process is stationary and homogeneous then $K_{ij}(r) = K_{ji}(r)$. Under independence, the intertype K function is $K_{ij}(r) = \pi r^2$, regardless of the individual univariate spatial patterns of the two types of events. Note that it is easier to work with the corresponding $L_{ij}(r) = \sqrt{K_{ij}(r)/\pi}$ function, because the variance of $L_{ij}(r)$ is approximately constant. Under independence, $L_{ij}(r) = r$. Positive values of $L_{ij}(r) - r$ indicate attraction between the two processes at distance r while negative values indicate repulsion. Because of its definition, $K_{ij}(r)$ characterizes the spatial interaction between points of type i and points of type j. In order to interpret the observed values of $K_{ij}(r)$, we need to compare them to the theoretical values obtained for simple cases of bivariate patterns, and especially to πr^2 that corresponds to a null hypothesis of absence of interaction between the two types of points. However, depending on the context of the study, this appropriate null hypothesis can correspond to at least two different statistical hypotheses (see [10]): independence or random labelling. Hypothesis of random labelling means that the probability that one event occurs is the same for all points and does not depend on neighbors.

3.2.2 Estimation

The estimator of the intertype K_{ij}-function can be defined by:

$$\hat{K}_{ij}(r) = (\hat{\lambda}_i \hat{\lambda}_j A)^{-1} \sum_{k,l} \mathbb{1}_{d_{i_k,j_l} < r} \tag{6}$$

where d_{i_k,j_l} is the distance between the kth location of type i point and the lth location of type j point, A is the area of the region of interest and where $\hat{\lambda_i}$ and $\hat{\lambda_j}$ are the estimated intensities.
This estimator characterizes the relationship between the marginal patterns at all scales. $\hat{K}_{ij}(r)$ is a function of the distance between points. It can be integrated to provide single valued indices (see [11] for example).
As the theoretical distributions of the estimators are unknown, confidence intervals are commonly estimated through Monte Carlo simulations of a specified null hypothesis. To test independence, a classical method is to keep the patterns of both point processes unchanged, but to randomize their relative position at each Monte Carlo simulation while to test random labelling a classical approach is to simulate realizations of a point process with the same spatial structure as the overall observed pattern (i.e. without type distinction), and a random attribution of marks.
Various edge corrections have been suggested; one common example is the extension of Ripley's estimator, leading to:

$$\hat{K}_{ij}(r) = (\hat{\lambda_i}\hat{\lambda_j}A)^{-1} \sum_{k,l} w(i_k, j_l) \mathbb{1}_{d_{i_k,j_l}<r}$$

The coefficient $w(i_k, j_l)$ is the inverse of the proportion of the perimeter of the circle centered at the kth location of type i point with radius d_{i_k,j_l} that lies inside the study area. Basically, this corresponds to an estimate of the number of points at the same distance that would be outside the study area. Ripley ([23]) shows that this corrected estimator is unbiased. Unfortunately this correction leads to two problems. At first, when edge corrections are used, then $\hat{K}_{ij}(r)$ and $\hat{K}_{ji}(r)$ are positively correlated and no more equal. Moreover, our aim is to focus on points within the study area and it is out of sense to have an estimate of points outside the window. In the sequel we use a Ripley's intertype function without edge correction.

4 Spatial CART

4.1 Impurity loss based on K_{ij}

The key idea is to take into account in the splitting strategy, the spatial dependency of the data. It is done by modifying the original impurity loss, which is usually the entropy index. We introduce a dissimilarity index based on Ripley's intertype K-function quantifying the interaction between two populations.
Let focus on the impurity loss associated with $\hat{K}_{ij}$. For a node t and a split s splitting t into two child nodes t_L and t_R, we define

$$\begin{aligned}\hat{K}_{ij}^{t}(r) &= \frac{A^{t_L}}{A^t}\frac{\hat{\lambda}_i^{t_L}\hat{\lambda}_j^{t_L}}{\hat{\lambda}_i^t\hat{\lambda}_j^t}\hat{K}_{ij}^{t_L}(r) + \frac{A^{t_R}}{A^t}\frac{\hat{\lambda}_i^{t_R}\hat{\lambda}_j^{t_R}}{\hat{\lambda}_i^t\hat{\lambda}_j^t}\hat{K}_{ij}^{t_R}(r) \\ &\quad + \frac{1}{\hat{\lambda}_i^t\hat{\lambda}_j^t A^t}\left[\sum_{i_k\in t_R}\sum_{j_l\in t_L}\mathbb{1}_{d_{i_k,j_l}<r} + \sum_{i_k\in t_L}\sum_{j_l\in t_R}\mathbb{1}_{d_{i_k,j_l}<r}\right]\end{aligned}$$

where one has:

A^t	the area of node t
$\hat{\lambda}_*^t,\ *=i,j,$	the estimation of the density of mark $*$ in node t
$d_{i_k j_l}$	the euclidean distance between i-marked individual i_k and j-marked individual j_l
$\hat{K}_{ij}^t$	the estimation of the Ripley's intertype K function restricted to node t

This leads to the natural definition of $\Delta I_{ij}(s,t,r)$ as the variation of heterogeneity coming from the split of node t using s, at radius r:

$$\Delta I_{ij}(s,t,r) := \hat{K}_{ij}^t(r) - \alpha_s\frac{\hat{\lambda}_i^{t_L}\hat{\lambda}_j^{t_L}}{\hat{\lambda}_i^t\hat{\lambda}_j^t}\hat{K}_{ij}^{t_L}(r) - (1-\alpha_s)\frac{\hat{\lambda}_i^{t_R}\hat{\lambda}_j^{t_R}}{\hat{\lambda}_i^t\hat{\lambda}_j^t}\hat{K}_{ij}^{t_R}(r) \tag{7}$$

where $\alpha_s = \frac{A^{t_L}}{A^t}$. The area factor α_s is natural when dealing with spatial data since it leads to reweight properly the impurity of the two nodes t_R and t_L.

The idea of intertype K_{ij} function is to characterize interaction between two point processes, *i.e.* second order characteristic of the bivariate point process. Hypothesis of stationarity is quite strong and can be relaxed using local estimation of the intensity [2] but this lead to instability of the estimator of the intertype K_{ij} function. To robustify the algorithm we assume the hypothesis of stationarity.

This choice of impurity loss is natural since equivalently:

$$\Delta I_{ij}(s,t,r) = \frac{1}{\hat{\lambda}_i^t\hat{\lambda}_j^t A^t}\left[\sum_{i_k\in t_R}\sum_{j_l\in t_L}\mathbb{1}_{d_{i_k,j_l}<r} + \sum_{i_k\in t_L}\sum_{j_l\in t_R}\mathbb{1}_{d_{i_k,j_l}<r}\right] \tag{8}$$

Note that $\Delta I_{ij}(s,t,r)$ is positive, which is necessary to define it as impurity loss. In addition, $\Delta I_{ij}(s,t,r)$ is null if and only if the children nodes t_L and t_R are pure

at distance r along split s, that is

$$\forall i_k \in t_L, j_l \in t_R \text{ and } \forall i_k \in t_R, j_l \in t_L \quad d_{i_k,j_l} \geqslant r,$$

highlighting splits that do not discriminate labels at all.

Hence maximizing $\Delta I_{ij}(s,t,r)$ leads to increasing spatial purity at fixed scale r. In addition to the positivity of $s \mapsto \Delta I_{ij}(s,t,r)$, which is mandatory, a desirable property is the strict concavity, which would ensure that the best split is unique, and then avoids ties. This is an ingredient of the original CART algorithm but this is not the case here as in most extensions of CART. It should be noted that nevertheless, the growing part of the algorithm still work, that is splitting always purifies nodes even if from an algorithmic point of view, the choice of the split could be arbitrary, without any statistical drawback.

4.2 Description of the algorithm

While the usual penalized misclassification error rate (equation 2) used in CART classification trees, is natural to predict marks, it is no more convenient in the context of intensity estimation. Our focus is either to classify conveniently the point process in terms of its marks, or to estimate the marks' intensity. In the latter case, to be able to choose a convenient definition for the risk, we consider the following property of the intensity of the marks: in the case where the marked spatial point process (X, M) is stationary, then the intensity of points with mark $j \in \{1; 2\}$ inside surface $\mathcal{A}$ can be written as

$$\begin{aligned} \Lambda(\mathcal{A}, j) &= P(X \in \mathcal{A}, M = j) \\ &= P(M = j \mid X \in \mathcal{A}) P(X \in \mathcal{A}) \\ &= P(M = j \mid X \in \mathcal{A}) \lambda A \end{aligned}$$

where $\lambda > 0$ is a constant depending only on the marginal distribution of X, and A is the area of $\mathcal{A}$. Then estimating $\Lambda(\mathcal{A}, j)$ is equivalent to estimate $P(M = j \mid X \in \mathcal{A})$, since the constant λ can be estimated directly from the point process. Thus the penalized criterion (4) used in CART class probability trees is natural to estimate the intensities of the marks: each intensity will then be locally estimated on a tessellation of the plane.

We propose an algorithm using impurity loss ΔI_{ij} defined by (7) to develop the maximal tree T_{max}. The estimator $\hat{K}_{ij}$ of the intertype K-function K_{ij} is computed at each node t, the value of r is fixed as the one for which the estimated intertype K-function is the farthest from the one of random labelling.

A maximal tree T_{max} is computed via recursive partitioning, and then pruned with the penalized criterion based on misclassification rate defined in (2), or based on Gini index defined in (4). This produces a decreasing sequence of K subtrees pruned each one from another, denoted by $T_1 \succ \ldots \succ T_K = \{t_1\}$, and associated with an increasing sequence of K complexities, denoted by $0 = \alpha_1 < \ldots < \alpha_K$.

Rephrasing the algorithm in spatial terms, we could say that starting from the whole original region, for which we compute r_0 which can be considered as the characteristic scale at this resolution. Then before splitting, we consider for $r = r_0$ the quantity $\Delta I_{ij}(s, t, r)$ and we seek to the best split. After splitting we seek to the best $r \leq r_0$ maximizing $\Delta I_{ij}(\hat{s}, t, r)$. Then after splitting, the two child are considered in parallel in the same way, recomputing $r_{0,L}$ and $r_{0,R}$. It turns out that $\Delta I_{ij}(s, t, r)$ is decreasing along any branch of the tree. So the reordered sequence of r_i can be used to define a sequence of nested subtrees. Let us remark that, in the algorithm, the value of r_t maximizing the impurity criterion for the best split is set to 0 if t is a leaf, what corresponds to the smallest possible value of r that is the best resolution.

The final step is to choose a convenient tree among the sequence $T_1 \succ \ldots \succ T_K = \{t_1\}$. This choice can be made via cross-validation Nevertheless, to avoid randomness due to the choice of subsamples, we use a heuristic method proposed in [15] and based on the behavior of the number of leaves with respect to the complexity. The general idea is that, if a tree is a good predictor, then high energy is necessary to prune it. Hence, the penalty to be chosen in the penalized criterion shall increase a lot before the tree is pruned. An artificial example of such behavior is given in Figure 1. We can see a plateau in the right hand side of the graph, corresponding to the focused complexity. Hence, we select the tree associated with the plateau, here on the figure the tree having 3 leaves.

The algorithm used to obtain the final tree is detailed in Table 1. Let us emphasize that the sequences of pruned subtrees and complexities can also be kept to allow the user to choose another tree if needed (especially if the best subtree is the root).

5 CART and SpatCART in action: illustration by simulations

We use the following `R` packages to implement the SpatCART algorithm and to display the results:

- `spatstat` to deal with point processes, and in particular to compute ΔI_{ij} in the construction of the maximal tree,

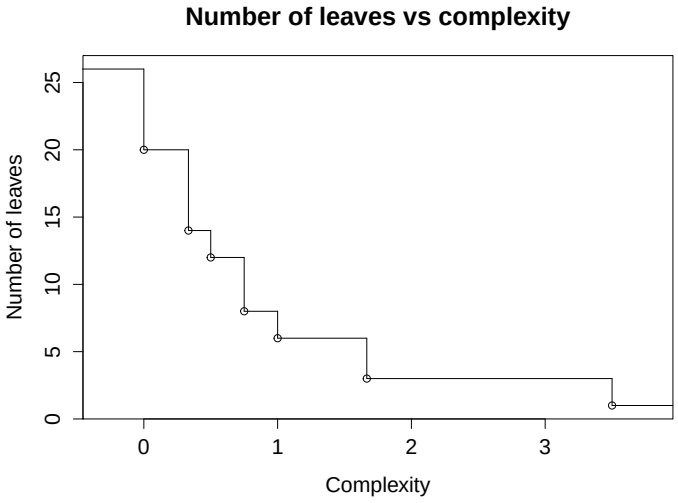

Figure 1: Typical example of the behavior of number of leaves with respect to complexity.

– `tree` to deal with tree structures.

Two methods are to be compared: Spatial CART (SpatCART, see Table 1), and CART with Gini splitting criterion. Class probability trees and Classification trees are constructed using either SpatCART or CART.

We expect the following results:

– Since the pruning strategy is the same for the two methods, based either on the minimization of penalized misclassification rate or on the minimization of penalized Gini index rate, the major differences between our proposal (Spatial CART) and the existing schemes must be concentrated on the growing step. In addition to the final trees, we also have to look at the comparison of the two maximal trees.

– Since the new splitting criterion reduces more or less to the classical one in the case of spatial homogeneity, we have to consider some artificial homogeneous examples to illustrate this and, what is more crucial, to define some artificial inhomogeneous examples to capture the interest of our proposition. This second point must be considered because the spatial nature of the data must be taken into account, and we can refer to [17] and [1].

	Spatial Classification Trees
Input	Marked point process, scale $r \in]0;1[$, minimal number of observations in node to split `minsplit`, minimal value of Gini deviance in node to split `mindev`.

Maximal tree	`Initialize`,
	node $t = t_1$ the root of the tree containing all observations,
	$n_t = n_{t_1}$ the number of observations in node t,
	$r_t = r$ the scale value at node t,
	$R(t) = R(t_1)$ the value of entropy in node t,
	$\text{argmax}\{\lambda_1^t, \lambda_2^t\}$ the label of node t.
	`While` $n_t >$ `minsplit and` $R(t) >$ `mindev`,
	`Compute`
	$i_0 = \text{argmax}_{i\in\{1;2\}}\lambda_i^t$, $j_0 = \text{argmin}_{i\in\{1;2\}}\lambda_i^t$,
	$\hat{s} = \text{argmax}_s \Delta I_{i_0 j_0}(s, t, r_t)$,
	`Set`
	$t_L = \{\text{points in } t \mid \text{answer "yes" to } \hat{s}\}$,
	$t_R = \{\text{points in } t \mid \text{answer "no" to } \hat{s}\}$.
	`Recursion`
	$r_t = \text{argmax}_r \Delta I_{i_0 j_0}(\hat{s}, t, r)$,
	`left`: $t = t_L$,
	`right`: $t = t_R$.
	`Output`
	Maximal tree T_{max}.

Pruning	Sequence $(T_{\alpha_k})_{1\leqslant k\leqslant K}$ of subtrees pruned from T_{max}, and complexities $(\alpha_k)_{1\leqslant k\leqslant K}$ (see Table 2).
Selection	`Set` $\hat{k} = \max\left\{k \mid k = \text{argmax}_{0\leqslant j\leqslant K-1}(\alpha_{j+1} - \alpha_j)\right\}$.
Output	Tree $T_{\hat{k}}$, sequences $T_1 \succ \ldots \succ T_K = \{t_1\}$ and $0 = \alpha_1 < \ldots < \alpha_K$.

Table 1: Spatial Classification Trees (`SpatCART`).

5.1 Illustrative examples

The first example is the `Chess` bivariate Poisson point process on the unit square, with marks simulated from a blue and red chessboard with 9 squares, represented in Figure 2. It is an example of spatial homogeneity for which CART and SpatCART should lead to similar results.

	Pruning Algorithm
Input	Maximal tree T_{max}, number of observations n, risk ρ from misclassification error or Gini index.
Initialization	$k = 1$ and $\alpha_1 = 0$, $T = T_1$ the smallest subtree pruned from T_{max} at complexity α_1, n_f the number of leaves of T,
Compute	for each node t of T n_t the number of observations in t, $\rho(t)$ the local risk of t: $\rho(t) = n^{-1} \sum_{(x_i,m_i)\in t} \mathbb{1}_{\hat{m}_t \neq m_i}$ for misclassification rate $\rho(t) = (1 - \sum_{j=1}^{2} p(j\|t)^2) n_t/n$ for Gini index $\rho(T_t) = \sum_{\{f \text{ leaf of } T_t\}} \rho(f)$ the risk of the branch T_t issued from t, n_{T_t} the number of leaves of branch T_t.
While	While $n_f > 1$, `compute` $\alpha_{k+1} = \min_{\{t \text{ internal node of } T\}} \frac{\rho(t) - \rho(T_t)}{n_{T_t} - 1}$. `Prune` all branches T_t of T verifying $\rho(T_t) + \alpha_{k+1} n_{T_t} = \rho(t) + \alpha_{k+1}$ `Set` T_{k+1} the pruned subtree obtained in that way. `Set` $T = T_{k+1}$ and $k = k + 1$.
Output	Sequence $(T_{\alpha_k}, \alpha_k)_{1 \leqslant k \leqslant K}$.

Table 2: Pruning Algorithm.

The second example is the `Locally repulsive` bivariate Poisson point process on the unit square, with a majority red mark, and a minority blue one. The marked point process is designed by splitting vertically the unit square into two parts (see Figure 3): on the left part, the minority blue marked points repulse red ones, with a constant repulsion radius equal to $r = 0.05$; on the right part, blue and red marked points are independently distributed.
In this last example, a spatial inhomogeneity is introduced, then CART and SpatCART should lead to different results: CART should miss the difference between the two parts and SpatCART should highlight it. With these two different scenarios

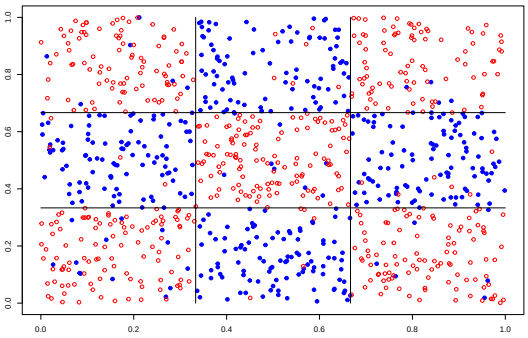

Figure 2: `Chess` simulated data set.

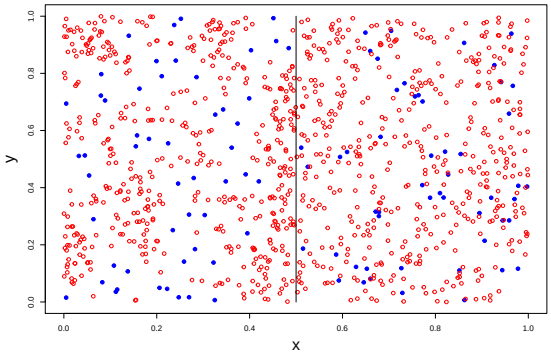

Figure 3: `Locally repulsive` simulated data set.

in mind, it remains to analyze the results. But before that, we propose to focus on the kernel of our proposal: the new splitting criterion convenient for detecting a certain kind of spatial heterogeneity and the initial spatial resolution to be selected.

5.2 Intertype K-function and splitting criterion

The key tool for defining the splitting criterion is the intertype K-function. More precisely the difference between the estimated one and its theoretical value in the homogenous situation is evaluated in order to select the split. In Figure 4 in the homogeneous situation `Chess` simulated example, one can find, on the left, the estimated intertype K function (as a function of r) and its theoretical counterpart and, on the right, the difference between these two functions. In this typical situation,

this difference is strictly deceasing and the naturel choice is to take r as large as possible.

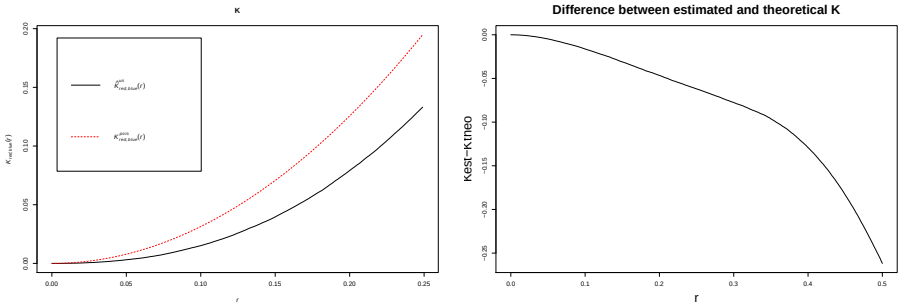

Figure 4: Example of intertype K function and its theoretical value in the homogeneous situation for `Chess` simulated example. *Left* : Intertype K functions. *Right* : Difference between estimated and theoretical intertype K functions.

The next object of interest is the impurity function directly connected with the splitting criterion. Figure 5 illustrates the behavior of the impurity function with respect to first split for `Chess` example, which appears to be clear and expected.

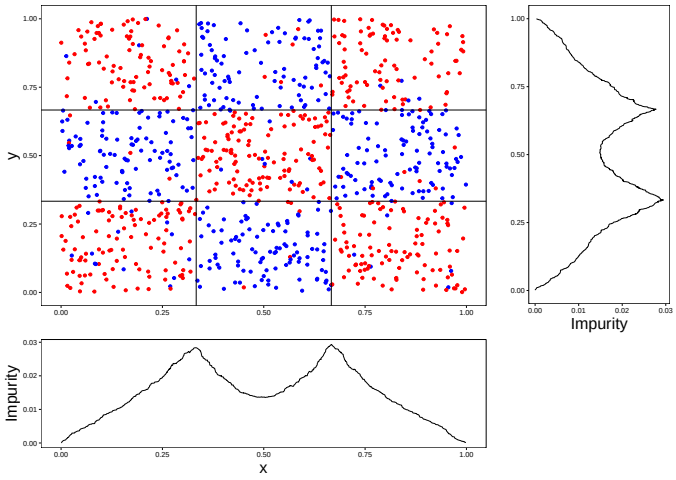

Figure 5: Behavior of impurity with respect to first split for `Chess` example.

5.3 The initial resolution r

The other parameter is the initial resolution r which is crucial to define the first split (the most important one). One first idea to provide a default value is to take

as initial value for r the one corresponding to maximum of the difference between the estimated one and its theoretical value, that the one the one for which the dissimilarity with the homogeneous case is maximal. Depending on the criterion configurations, it appears that critical values r can be far from this default value. So the advice is to let the user define this value.

Let us consider the same difference for 2 simulated examples (see Figure 6). Except for the `Locally repulsive` bivariate Poisson point process, the behavior is the same as previously. But in this last, the function seems to highlight a bump and we should select the value r leading to this bump instead of the larger one.

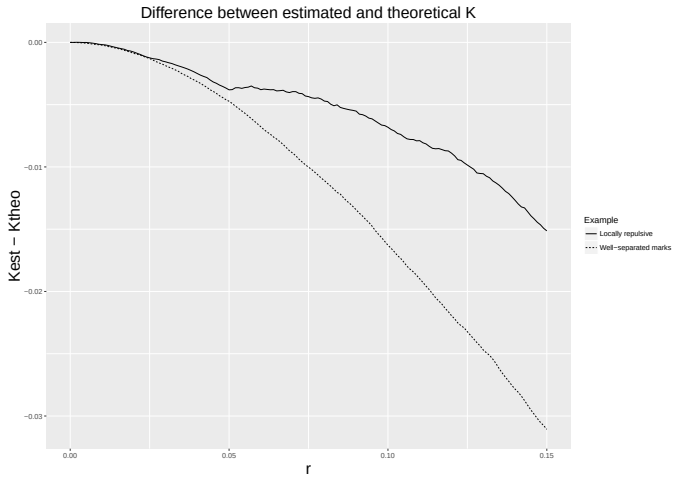

Figure 6: Difference between estimated intertype K function and its theoretical value for 2 typical simulated examples.

5.4 Maximal and optimal partitions

With two different situations depicted by the Figures 2 and 3, we compare the results of the usual CART strategy (for growing and pruning) and those obtained applying our new proposal.

5.4.1 The homogeneous case

The first global difference is to examine the respective performance indices and illustrate its stability. Figure 7 gives the percentage of the number of times where maximal and minimal optimal pruned subtrees classify differently from Bayes classifier over 1000 simulations of `Chess` data set. The two algorithms lead globally to accurate classifiers (less than 4% of the observations are classified differently by

the Bayes classifier), and it appears clearly that SpatCART exhibits slightly lower accuracy, but higher stability. One can also see that the maximal plateau method generally lead to less accurate classifiers, what illustrates its penchant to overpenalize the misclassification rate.

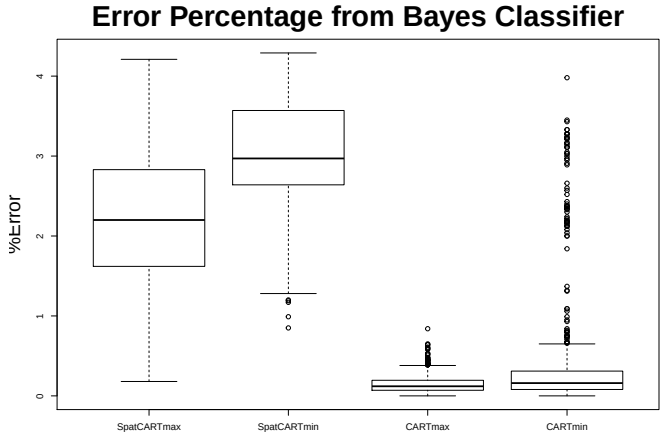

Figure 7: Percentage of number of times where maximal and minimal optimal pruned subtrees classify differently from Bayes classifier over 1000 simulations of `Chess` data set.

Comparing the maximal trees in Figure 8 for SpatCART (top) and CART (bottom) on `Chess` data set, leads to the conclusion that the partitions are quite similar. More precisely they are the same for the true underlying frontiers and the only differences appear inside the blocks to be recovered.

So, in this homogeneous case, as expected, CART and SpatCART lead to similar results. Examining the corresponding behaviors of the number of leaves as a function of complexity for SpatCART and CART (see Figure 9), reinforce this similarity.

It remains to inspect the maximal trees (see Figure 10) for SpatCART (top) and two variants of the usual CART (bottom). The two variants of CART differ only in the selection step, the *min* variant corresponds to the largest plateau while the *max* variant corresponds to the largest jump. In that case all the variants lead to essentially the same result. Therefore we decide to keep the *min* variant which leads to the most stable results.

So to summarize, as expected, when there is nothing to discover from the spatial viewpoint, it seems that SpatCART behaves as the usual CART, without any degradation.

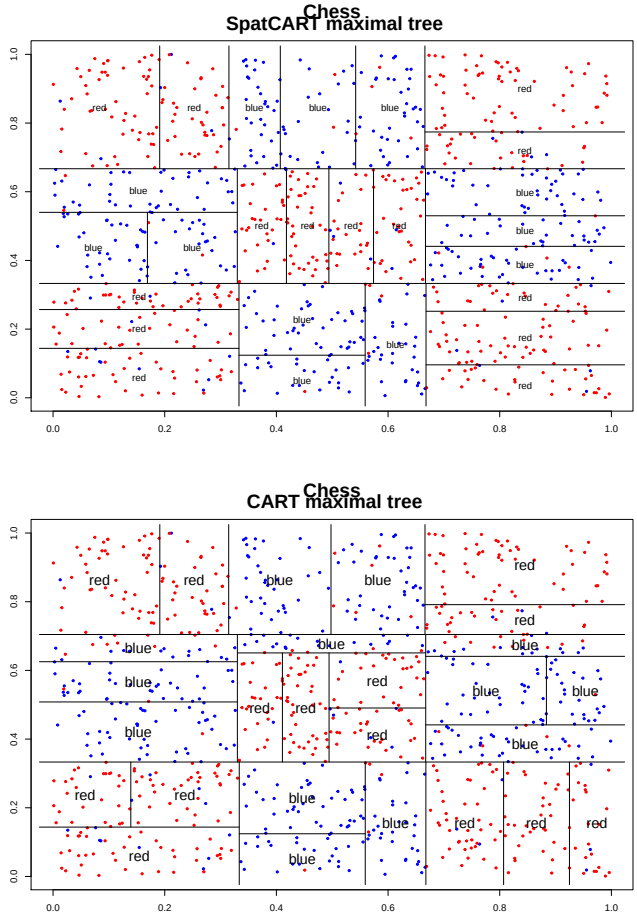

Figure 8: Maximal trees for SpatCART and CART on `Chess` data set.

5.4.2 The inhomogeneous case

Using roughly the same sequence of plots, let us examine now a more interesting aspect, crucial to illustrate what is the specific contribution of SpatCART.

The first step is to compare the maximal trees. We start with an initial resolution r_0 smaller than the locally repulsion scale $r = 0.05$. Figure 11 contains the SpatCART results (top) and the CART (bottom) ones on `Locally repulsive` data set. The partitions are extremely different. More precisely, since r_0 is small enough SpatCART does not split the left half of the square, and the right half is split iteratively according to the direction of the first split. This is expected since the precision of the estimation of the intensities is directly related to the number

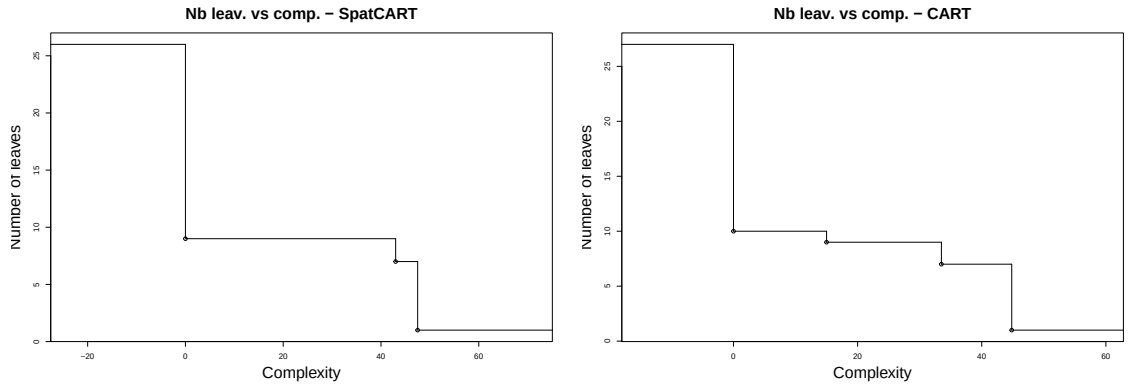

Figure 9: Behavior of number of leaves versus complexity for SpatCART and CART on `Chess` data set.

of points and the intensity of the rarer is too low to provide a reliable estimate. A solution could be to impose a minimum number of points for each mark to allow splitting. We haven't implement it since pruning will remove these artificial splits. Note that If the gain of impurity is the same for both direction the x direction is chosen. On the other hand, CART splits according to the empirical distribution of the blue points leading a very different partition. So the generated maximal partitions are, as expected, different in the inhomogeneous situation.

The corresponding sequences of complexities are given by Figure 12. A second difference appears: SpatCART seems to generate less splits to provide partitions similar to CART in terms of spatial repartition, generating less false alarms.

To end this analysis, Figure 13 shows the optimal pruned subtrees (represented as trees and not through the corresponding partitions as previously) for this inhomogeneous case. It should be noted that in this case, results using the pruning strategy are convenient for Class Probability Trees: the tree based on misclassification rate leads to a tree without leaves. This is logical since the percentage of blue marks is too low to be recovered. Here, the difference is clear: since there is no heterogeneity in the marginal structures of the response distribution nothing is inferred using CART. On th other hand SpatCART recovers the spatial structure.

To summarize, it seems that:

- SpatCART seems to provide better partitions in terms of marks spatial repartition (see the maximal trees figures),
- SpatCART needs less splits to provide partitions similar to CART in terms of spatial repartition (see figures and number of leaves of the pruned subtrees), suggesting that it avoids false alarms in a better way.

6 Example

6.1 Data

We applied these methods to a tropical rain-forest located at Paracou, 40 km west from Kourou in French Guiana (5°15'N, 52°55'W). It is an experimental site that is devoted to studying the effects of logging damage on stock recovery. Twelve plots (6.25 hectares each) of undisturbed forest were settled in 1984. On each plot, the circumference of every tree with a DBH (diameter at breast height) greater than 10 cm was measured with a precision of 0.5 cm. Its spatial coordinates ($\pm$ 50 cm) and its specie were noted too. Measurements have been carried annually from 1984 to 1995, and once every two years since. A more precise description of the Paracou plots may be found in [16].

We focus on two species:

- *Vouacapoua americana* is a hermaphroditic shade tolerant tree specie of mature tropical rainforests and whose distribution spans the eastern part of the Guiana shield. Its local density averages around 10 individuals per hectare for trees with greater than 10 cm DBH, but this varies greatly because of spatial clustering. Individuals are clustered in large patches of a few hectares that are mainly located on hill tops and slopes [26].
- *Oxandra asbeckii* is a specie of the understorey, the largest individuals of which measure less than 15 cm DBH. It is a shade tolerant specie [13]. It shows animal dispersal: the seeds are dispersed after passage through the vertebrate gut. *Oxandra asbeckii* is located on hill tops and slopes

The two species *Vouacapoua americanaa* and *Oxandra asbeckii* were selected at Paracou because their spatial distribution is linked to the relief: they are both located on hill tops and slopes. Elevation is the environmental factor that drives their spatial distribution and this creates a strong interaction between both repartitions. We focus on one plot represented in Figure 15. The figure also gives the contour lines in order to be able to determine hill tops and slopes of the plot. Seventy trees of *Vouacapoua americana* and eighty trees of *Oxandra asbeckii* were referenced fort his plot.

The data consists of seventy lines (one par tree) and four columns: the 3-D coordinates (longitude, latitude and elevation) as well as the specie indication.

6.2 Results and discussion

From the representation of the estimated and theoretical intertype $K-$functions and their difference evaluated on the Paracou data set (see Figure 14), we can see

three regimes: one before the scale value $r = 6$ (blue dashed line on the right of Figure 14), where there is no interactions between species; one between scale values $r = 6$ and $r = 24$ (red dashed line on the right of Figure 14) where species begin to interact; and one after the scale value $r = 24$ where the interaction between species increases rapidly. Hence we choose the initial median scale value $r_0 = 15$ to proceed with SpatCART, in order to be sure to catch a sufficiently large scale to catch interaction, and not too large to avoid deeper maximal trees.
Figure 15 highlights the presence of *Oxandra asbeckii* at the hill of left top of the plot as well as the competition between both species for the hill at the bottom of the plot. Even if class probability tree has more leaves than classification tree, both trees are coherent and lead to the same conclusions.

A contrario, CART results are really poor with only two leaves. Basically it separates the hill at the bottom of the plot from the rest but cannot catch the mixed structure of species with this hill or the hill at the top left of the plot. The spatial structure as well the ecology of the two species on this plot cannot be inferred from CART results.

References

[1] L Anselin and A Getis Spatial statistical analysis and geographic information systems, *in Perspectives on spatial data analysis, Springer* 35–47, 2010

[2] A Baddeley, J Moller and R Waagepetersen Non- and semiparametric estimation of interaction in inhomogeneous point patterns, *Statistica Neerlandica* 54, 329-350, 2000

[3] A Bar-Hen and N Picard. Simulation study of dissimilarity between point process *Computational Statistics*, 21(3-4):487-507, 2006

[4] L Bel, D Allard, JM Laurent, R Cheddadi and A Bar-Hen. CART algorithm for spatial data: application to environmental and ecological data *Computat. Stat. and Data Anal.*, 53(8):3082-3093, 2009

[5] G Blanchard, C Schäfer, Y Rozenholc, and K-R Müller. Optimal dyadic decision trees. *Machine Learning*, 66(2-3):209–241, 2007

[6] I Bou-Hamad, D Larocque, H Ben-Ameur. A review of survival trees. *Statistics Surveys*, 5:44–71, 2011

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

[8] DR Brillinger Measuring the association of point processes: a case history *American Mathematical Monthly* 83:16–22, 1976

[9] N Cressie *Statistics for Spatial Data*, John Wiley & Sons, New York, 1991

[10] PJ Diggle and AG Chetwynd Second-order analysis of spatial clustering for inhomogeneous populations, *Biometrics* 47:1155–1163, 1991

[11] PJ Diggle and RK Milne Bivariate Cox processes: some models for bivariate spatial point patterns *Journal of the Royal Statistical Society, Series B* 45:11–21, 1983

[12] DL Donoho Cart and best-ortho-basis: a connection. *The Annals of Statistics*, 25(5):1870–1911, 1997

[13] V Favrichon. Classification des espèces arborées en groupes fonctionnels en vue de la réalisation d'un modèle de dynamique de peuplement en forêt guyanaise. Rev. Ecol., 49, 379–403, 1994

[14] J H Friedman. Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67, 1991

[15] S Gey and E Lebarbier Using CART to detect Multiple Change Points in the Mean, *Preprint in Statistics and System Biology* 12, HAL 00327146.

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

[17] R Haining Bivariate correlation with spatial data, *Geographical Analysis* 23(3):210–227, 1991

[18] M LeBlanc and C Crowley Survival trees by goodness of split, *Journal of the American Statistical Association*, 88(422):457–467, 1993

[19] HW Lotwick Some models for multitype spatial point processes, with remarks on analysing multitype patterns, *Journal of Applied Probability* 21:575–582, 1984

[20] 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

[21] A Molinaro, S Dudoit, and MJ Van der Laan. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(1):154–177, 2004

[22] J Møller and RP Waagepetersen *Statistical Inference and Simulation for Spatial Point Processes*, number 100 *in* Monographs on Statistics and Applied Probability, Chapman & Hall/CRC, Boca Raton, 2004

[23] BD Ripley. Modelling spatial patterns. Journal of the Royal Statistical Society. Series B (Methodological), 172-212, 1977.

[24] G Saporta and G Youness Comparing partitions of two sets of units based on the same variables, *Advances in Data Analysis and Classification* 4(1):53–64, 2010

[25] F Questier, R Put, D Coomans, B Walczak, and Y Vander Heyden. The use of cart and multivariate regression trees for supervised and unsupervised feature selection. *Chemometrics and Intelligent Laboratory Systems*, 76(1):45–54, 2005

[26] S Traissac. Dynamique spatiale de Vouacapoua americana (Aublet), arbre de forêt tropicale humide à répartition agrégée. PhD Thesis. Université Claude Bernard-Lyon 1, Lyon, 2003

[27] H Zhang and B Singer. *Recursive partitioning in the health sciences*. Springer Science & Business Media, 2013

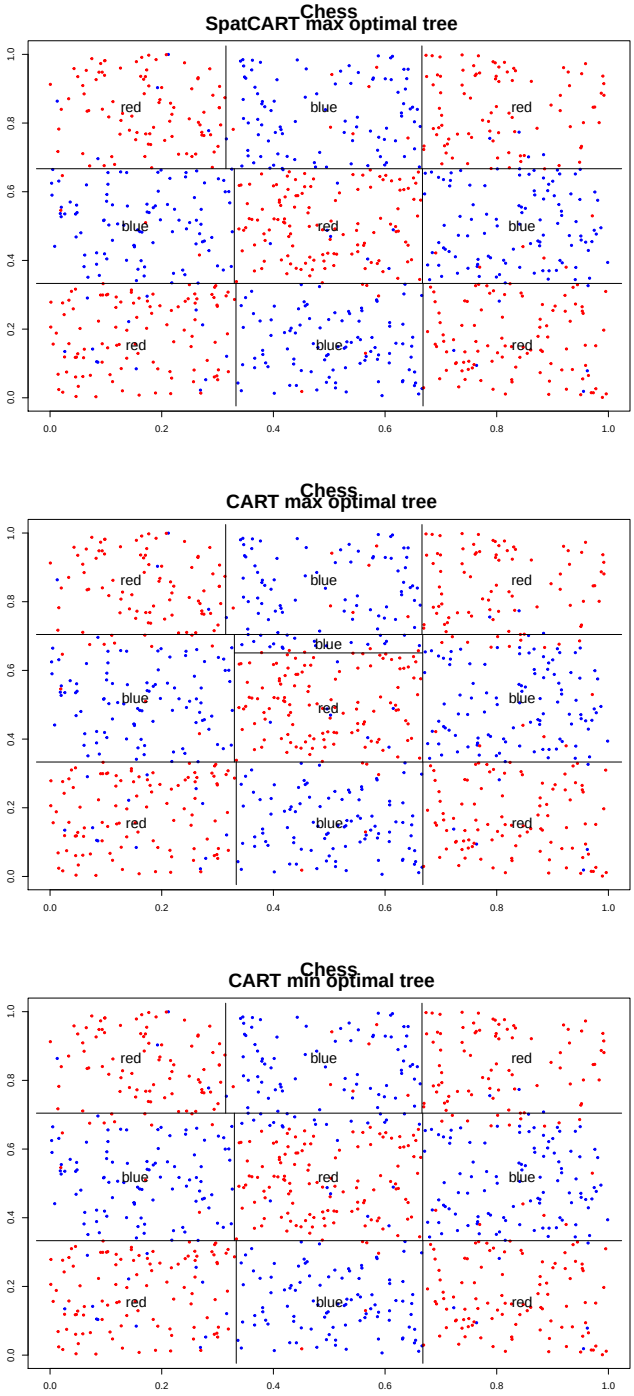

Figure 10: Optimal pruned subtrees for SpatCART and CART on `Chess` data set.

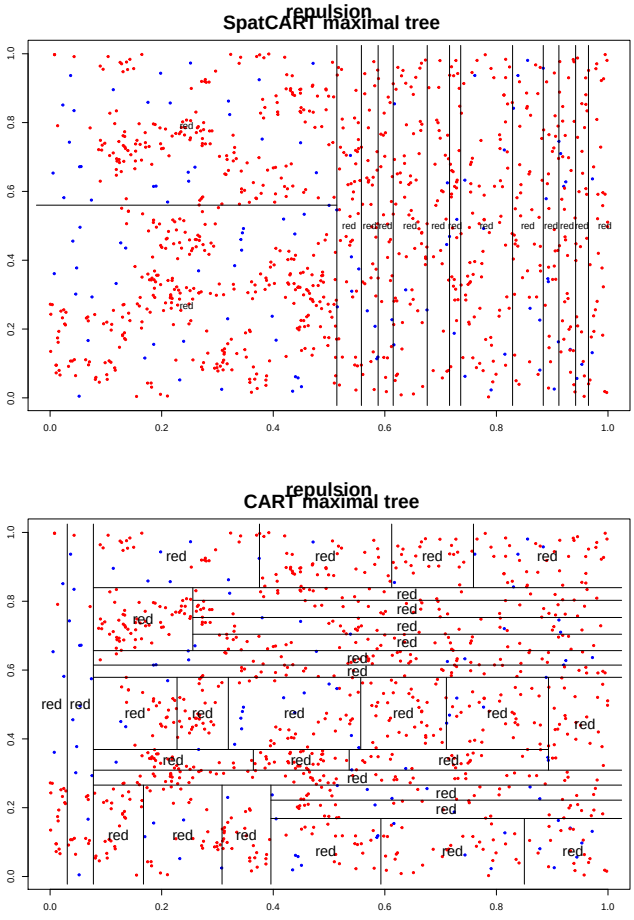

Figure 11: Maximal trees for SpatCART and CART on `Locally repulsive` data set, with initial resolution r_0 smaller than the locally repulsion scale $r = 0.05$.

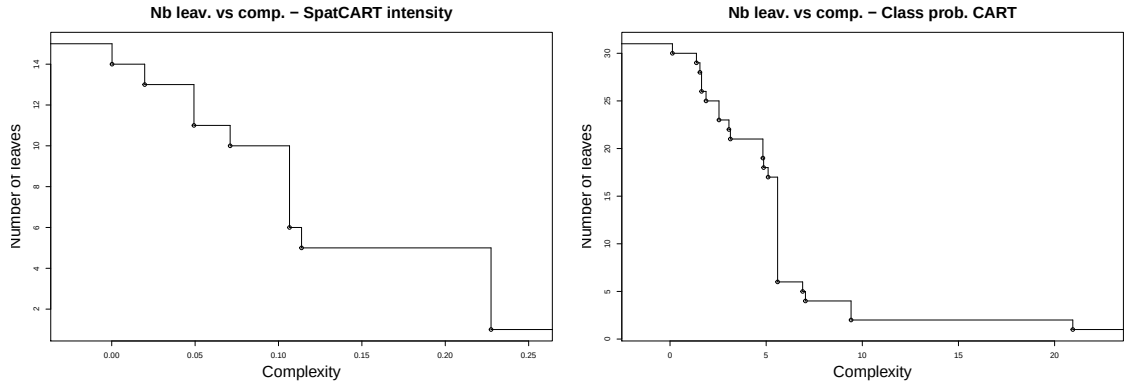

Figure 12: Behavior of number of leaves versus complexity for SpatCART and CART on `Locally repulsive` data set, with initial resolution r_0 smaller than the locally repulsion scale $r = 0.05$.

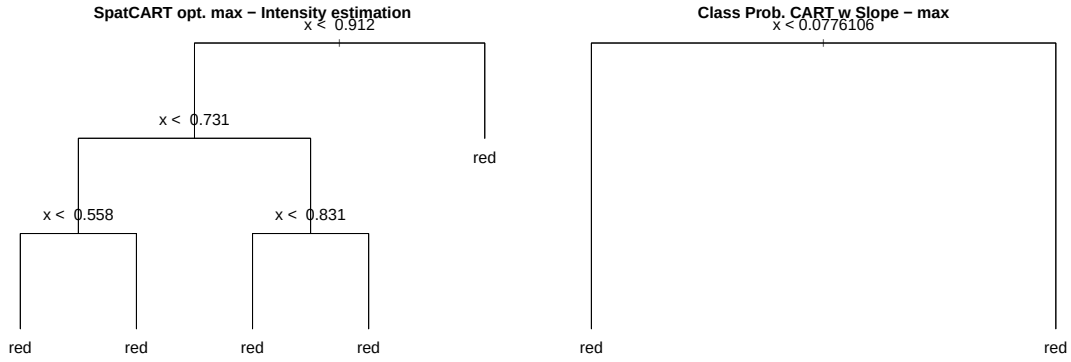

Figure 13: Optimal pruned subtrees for SpatCART and CART on `Locally repulsive` data set, with initial scale r_0 smaller than the locally repulsion scale $r = 0.05$.

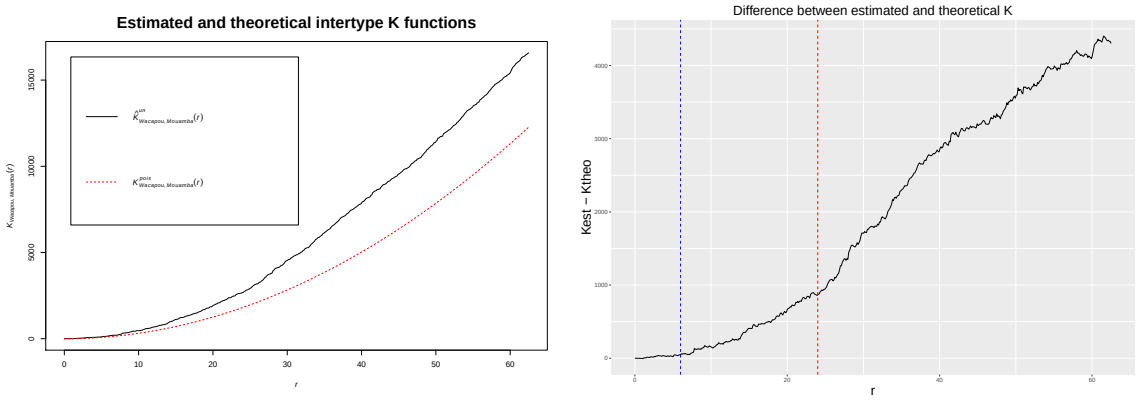

Figure 14: Intertype K function and its theoretical value in the homogeneous situation for the Paracou data set. *Left* : Intertype K functions. *Right* : Difference between estimated and theoretical intertype K functions; `blue`: scale $r = 6$, `red`: scale $r = 24$.

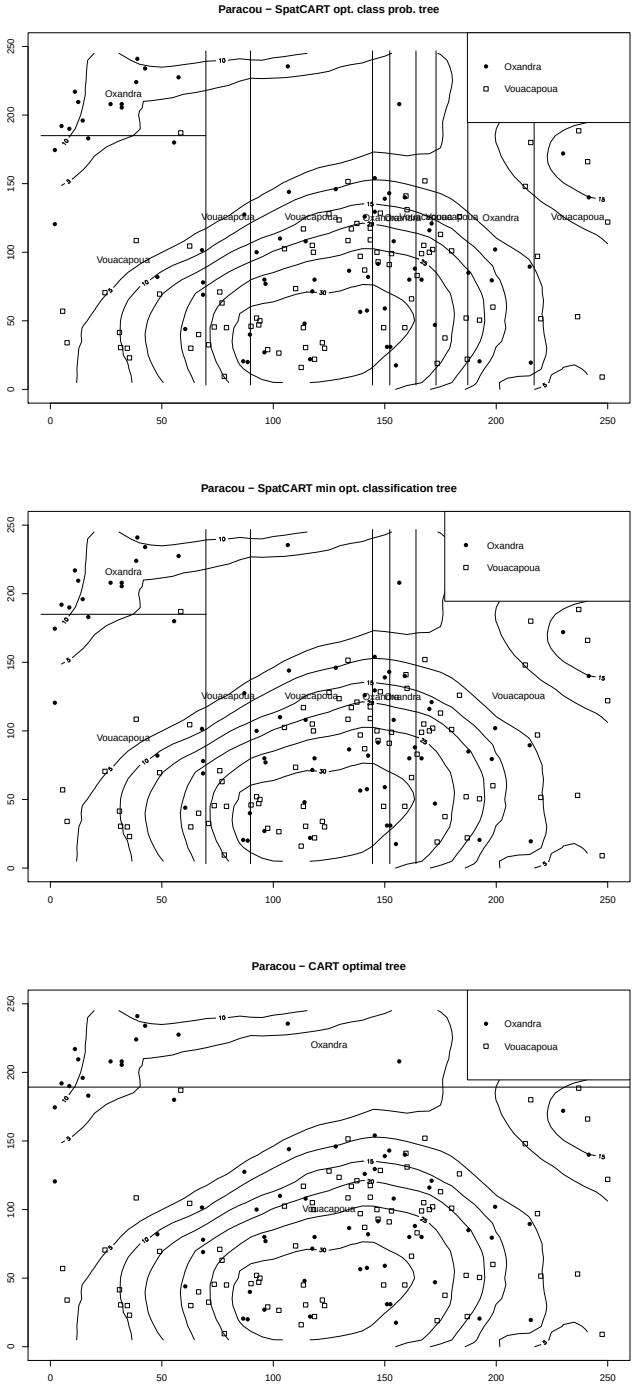

Figure 15: SpatCART and CART optimal trees on the Paracou data set.

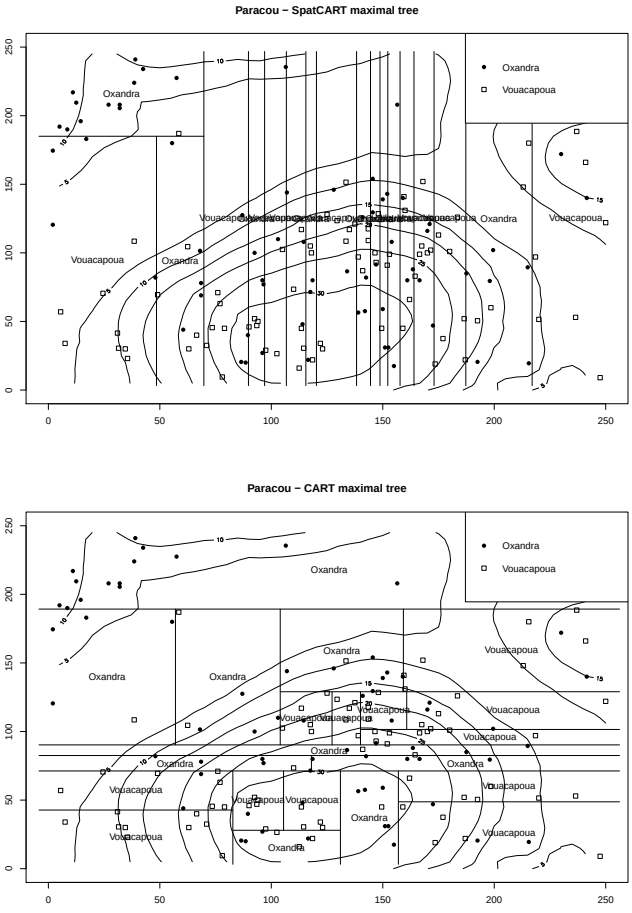

Figure 16: SpatCART and CART maximal trees on the Paracou data set.