



USING SPATIAL ESTIMATES IN THE CART  
ALGORITHM:  
APPLICATION TO ECOLOGICAL DATA

BY

LILIANE BEL  
DENIS ALLARD  
AVNER BAR-HEN  
JEANNE-MARIE LAURENT  
RACHID CHEDDADI

RESEARCH REPORT NO. 16  
DECEMBER 2005

Unité de Biométrie  
Institut National de la Recherche Agronomique  
Avignon, France  
<http://www.avignon.inra.fr/biometrie>

# Using spatial estimates in the CART algorithm: application to ecological data

L. Bel<sup>1</sup>      D. Allard<sup>2</sup>      A. Bar-Hen<sup>3</sup>      J.M. Laurent<sup>4</sup>  
R. Cheddadi<sup>4</sup>

<sup>1</sup>`Liliane.Bel@math.u-psud.fr`

Laboratoire de Mathématique, Université Paris-Sud, 91405 Orsay Cedex, France

<sup>2</sup>`Denis.Allard@avignon.inra.fr`

Unité de Biométrie, Institut National de la Recherche Agronomique, 84914 Avignon, France

<sup>3</sup>`avner@inapg.fr`

Institut National d'Agronomie, 16 rue Claude Bernard, 75005 Paris, France

<sup>4</sup>`laurent@isem.univ-montp2.fr`

Institut des Sciences de l'Evolution, CP 61, CNRS UMR 5554, 34095 Montpellier, France

## Abstract

Most of statistical learning techniques such as Classification And Regression Trees (CART) assume independent samples to compute classification rules. This assumption is very practical to estimate the empirical risk and asymptotic properties of estimators. In many environmental or ecological applications, the data under study are a sample of some regionalized variable for which the implicit assumption of independence is not acceptable. When the sampling scheme is very irregular, a direct application of supervised classification algorithms leads to biased discriminant rules due to the possible oversampling of some areas. We propose to adapt the CART algorithm to the case of spatially dependent samples. Two approaches are considered. The first one takes into account the irregularity of the sampling by weighting the data according to their spatial pattern using two existing methods based on Voronoï tessellation and regular grid, and one original method based on kriging. We also propose a second approach that uses spatial estimates of the

quantities involved in the construction of the discriminant rule at each step of the algorithm. These methods are compared on a simulation study and illustrated on paleoecological data, showing clearly the necessity of taking into account the spatial pattern of the sampling design.

**Keywords:** CART, ecology, grid, kriging, spatial estimation, Voronoi

## 1 Introduction

Paleoecology is the science of reconstructing past environments using fossil materials of plants, animals, or other indicators. These studies are useful for understanding the dynamics of ecosystem changes and thus for predicting their future evolution. They also provide tools to reconstruct conditions that existed before the impacts of industrialized societies on natural ecosystems. As far as vegetation is concerned, the basic idea is to consider that plants geographical distribution are in a dynamic equilibrium with climate (Woodward, 1987).

Harrison and Prentice (2003) and many others used pollen samples to evaluate simulated modern and/or past biome ranges. The geographic distribution of pollen frequencies is supposed to reproduce more or less properly plant range. However, such relationship between plants and their representation is often biased by different pollen production rate. Some of them are: different pollen production rate between taxa and unequal transport of pollen grains depending on their shapes and densities. One approach is to combine taxa with similar climatic environmental envelopes, thus defining functional group of plants (BAGs: Bioclimatic Affinity Groups) (Laurent et al., 2004). The validation of vegetation models requires more precise evaluation of the vegetation reconstructions, with quantification of the efficiency of each pollen group to represent surrounding vegetation at the simulated scales.

Pollen data are continuous variables while actual vegetation is a categorical variable. BAGs are the levels of this variable. From a statistical point of view, the question is how to predict a class of vegetation with a continuous variable. This situation

is characteristic of discriminant analysis (also known as supervised classification). There are many ways to construct discrimination rules but we focus on Classification And Regression trees (CART). CART procedures have proven to be very useful in ecological contexts because both continuous and discrete predictive variables can be used in the models and the outputs are easily understood (De'ath and Fabricius, 2000). Because they are nonparametric and divide datasets into distinct groups, CART models have several additional advantages over other techniques: input data do not need to be normally distributed; it is not necessary for predictor variables to be independent; and non linear relationships between predictor variables and observed data can be modeled. CART is presented in Section 2.

Most of statistical learning techniques such as CART assume independent samples to compute classification rules. This assumption is very practical to estimate the empirical risk and asymptotic properties of estimators. When dealing with environmental and ecological questions, samples often consist of measurement of variables within a domain. These data generally exhibit strong dependence due to their possible proximity. When the sampling scheme is very irregular, a direct application of supervised classification algorithms leads to biased discriminant rules because the same weight to every record is given and thus regions with high sampling density are overweighted.

We propose to adapt CART to the case of spatially dependent samples. Two approaches are considered. The first one, presented in section 3, considers the irregularity of the sampling by weighting the data according to their spatial pattern. The idea is to "decluster" the data: two existing methods based on Voronoï tessellation and regular grid are first considered, and we propose a new method based on kriging. In section 4 we also propose a second approach that uses spatial estimates of the quantities involved in the construction of the discriminant rule.

In this article we focus on CART but the main idea is to obtain a spatial estimate of the parameters involved in the classification rule, therefore extension of

our results to other discriminant techniques is possible. In a related context Hennig and Hausdorf (2004) try to find clusters from a presence-absence matrix. Dray et al. (2002) focus on the matching of two spatial sampling.

To compare the methods and to highlight the advantages and drawbacks of the various methods, we performed simulations. Results are presented in section 5. We present the analysis of our paleocological motivating example in section 6. We then conclude in section 7.

## 2 CART

We first recall some general background on Classification And Regression Trees (CART) in its usual *i.i.d* setting. For more detailed presentation see Breiman et al. (1984) or Ripley (1996, chap. 7). The data are considered as independent samples of random variables  $(X^1, \dots, X^p, Y)$ , where the  $X^k$ s are the explanatory variables and  $Y$  is the categorical variable to be explained. CART is a rule based method that generates a binary tree through binary recursive partitioning that splits a subset (called a leaf) of the data set into two subsets (called sub-leaves) according to the minimization of some heterogeneity criterion computed on the resulting sub-leaves. Each split is based on a single variable; some variables may be used several times while others may not be used at all. Each sub-leaf is then split further based on independent rules. Let us denote  $T$  a decision tree and  $t$  one of its leaf. Let  $p(j | t)$  be the proportion of a class  $j$  in a leaf  $t$ . The two most popular heterogeneity criteria are the entropy and the Gini index. The entropy index is

$$E_t = \sum_j p(j | t) \log\{p(j | t)\},$$

with, by convention,  $x \log x = 0$  when  $x = 0$ . The Gini index is

$$D_t = \sum_{i \neq j} p(i | t)p(j | t) = 1 - \sum_i p(i | t)^2. \quad (1)$$

Both indices are equal to 0 when there is only one class present in leaf  $t$  and are maximum when all classes are present with equal probabilities. Among all partitions of the explanatory variables at the node  $t$ , the principle of CART is to split a leaf  $t$  into two sub-leaves  $t_-$  and  $t_+$  according to a threshold on one of the variables, such that the difference between the heterogeneity of a leaf and the total resulting heterogeneity within the two sub-leaves is maximized and positive. The procedure is finished when there is no more admissible splitting. Each leaf is affected to the most present class (the conditional mode). In general the final tree overfits the available data and the prediction error  $R(T) = P\{T(X^1, \dots, X^p) \neq Y\}$  is typically large. In designing a classification tree, the ultimate goal is to produce from the available data a tree  $T$  whose probability of prediction error  $R(T)$  is as small as possible. Thus, in a second stage the tree  $T$  is "pruned" to produce a subtree  $T'$  whose expected performance is superior to  $R(T)$ . Since the distributions of  $Y$  and  $X^1, \dots, X^p$  are generally unknown, the pruning is based on the empirical risk  $\hat{R}(T)$  computed on cross-validation. The CART pruning algorithm seeks to balance optimistic estimates of empirical risk by adding a complexity term that penalizes larger subtrees.

When the data are independent samples, the proportion  $p(j | t)$  is estimated by  $\hat{p}(j | t) = n_{jt}/n_t$ , where  $n_{jt}$  is the number of samples in leaf  $t$  that are in class  $j$ , and  $n_t$  is the total number of samples in leaf  $t$ . The criterion to minimize is then  $D_t - (n_{t_-} D_{t_-} + n_{t_+} D_{t_+}) / n_t$  for the Gini index, and similarly for the entropy index. The empirical risk is

$$\hat{R}(T) = \frac{1}{n} \sum_{\alpha=1}^n \mathbb{I}\{T(X_{\alpha}^1, \dots, X_{\alpha}^p) \neq Y_{\alpha}\} \quad (2)$$

where  $\mathbb{I}(\cdot)$  is the indicator function and  $n$  the total number of samples.

In many environmental or ecological applications, the data under study are a sample of some regionalized variable for which the implicit assumption of independence in (1) and (2) is not acceptable. We must therefore consider that the samples  $\{X^1(s_{\alpha}), \dots, X^p(s_{\alpha}), Y(s_{\alpha})\}$ ,  $\alpha = 1, \dots, n$  are originated from random fields

$\{X^1(\cdot), \dots, X^p(\cdot), Y(\cdot)\}$  on some domain  $\mathcal{D} \in \mathbb{R}^2$  and explicitly take into account the dependence structure on these fields. In the next two sections we propose two approaches to adapt the CART algorithm to spatial data.

### 3 Weighted CART

The first approach is to weight the samples such that clustered data have less weight than sparse data. This idea stems from the fact that when there is a dependence structure in the random fields, data that are close to each other are likely to have similar values and will therefore carry somehow redundant information about the relationship between  $Y$  and  $(X^1, \dots, X^p)$ . Specifically, we will consider that

$$\hat{p}(i | t) = \sum_{\alpha \in t} w_{\alpha} \mathbb{I}\{Y(s_{\alpha}) = i\} \quad (3)$$

and

$$\hat{R}(T) = \sum_{\alpha=1}^n w_{\alpha} \mathbb{I}[T\{X^1(s_{\alpha}), \dots, X^p(s_{\alpha})\} \neq Y(s_{\alpha})],$$

with the condition  $\sum_{\alpha} w_{\alpha} = 1$ . We will consider three different methods for determining these weights. The two first methods are only based on the geometry of the sampling design: in short, the weight associated with each data is a measure of its area of influence. The last one is original and relies on the modeling of the spatial dependence on the response variable.

First method is to use the Voronoï tessellation generated by the sample locations  $(s_1, \dots, s_n)$ . A Voronoï cell around a sample location  $s_{\alpha}$  is the set of points of  $\mathcal{D}$  closer to  $s_{\alpha}$  than to any other sample locations (Okabe et al., 1992). Clustered observations produce smaller cells while sparse data produce larger ones. The inverse of their area is thus an estimator of the local density of the sample design, a property that has been used in spatial statistics to the clustering of spatial point processes (Allard and Fraley, 1997) or to the estimation of boundaries (Picard and Bar-Hen, 2000). The weight of a sample at a site  $s_{\alpha}$  is thus set to be proportional to the area of its Voronoï cell. This approach is attracting and easy to implement but leads to

undesirable boundary effects: first, data at the border of the domain have larger weights than data within the domain; second, weights of samples near the border depend strongly on the precise limit of the domain.

A related method is the declustering technique proposed in Isaaks and Shrivastava (1989). A regular grid is superimposed on the sampling region. A total weight of  $1/c$  is assigned to each cell  $a_k$ , where  $c$  is the number of occupied cells. Each observation falling in cell  $a_k$  is weighted by  $w_\alpha = (n_k c)^{-1}$ , where  $n_k$  is the number of samples in  $a_k$ . The weights obviously depend on the cell size and on the origin of the grid network. The cell size should neither be too large (most data must not be contained in just a handful of cells) nor too small (all cells must contain at least only one observation). Compared to the Voronoï tessellation, this method does not necessitate a definition of the domain  $\mathcal{D}$ . It suffices that the data are recovered by the grid.

We now propose a third method related to geostatistics. If the covariance function  $C(\cdot)$  of a random field  $Z(\cdot)$  is known, the best linear unbiased predictor of a regional average on a 2d domain  $\mathcal{D}$  is the so called kriging of a regional average (or kriging of the mean if  $\mathcal{D} \rightarrow \mathbf{R}^2$ ). Let us denote  $Z_{\mathcal{D}}$  the average of  $Z(\cdot)$  over  $\mathcal{D}$ . The kriging of  $Z_{\mathcal{D}}$  is the quantity  $\hat{Z}_{\mathcal{D}} = \sum_{\alpha} w_{\alpha} Z(s_{\alpha})$  such that  $E(\hat{Z}_{\mathcal{D}} - Z_{\mathcal{D}}) = 0$  and  $\text{var}(\hat{Z}_{\mathcal{D}} - Z_{\mathcal{D}})$  is minimum. It can be shown (Wackernagel,2003) that the vector  $W = (w_1, \dots, w_n)^T$  is solution of the system

$$\begin{pmatrix} \mathbf{C} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{pmatrix} = \begin{pmatrix} W \\ \nu \end{pmatrix} \begin{pmatrix} C_{\mathcal{D}} \\ 1 \end{pmatrix}, \quad (4)$$

where  $\mathbf{C}$  is the matrix whose  $\alpha, \beta$  element is  $C(s_{\alpha}, s_{\beta})$ ,  $C_{\mathcal{D}}$  is the vector with elements

$$C(s_{\alpha}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} C(s_{\alpha}, s) ds, \quad (5)$$

$\mathbf{1}$  is a vector of ones of length  $n$  and  $\nu$  is the Lagrange parameter associated with the unbiasedness condition. To evaluate the integral in (5) we define a grid  $G$  on  $\mathcal{D}$

and use the following approximation

$$\frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} C(s_{\alpha}, s) ds \simeq \frac{1}{n_G} \sum_{s_{\beta} \in G} C(s_{\alpha}, s_{\beta}).$$

For Gaussian random fields, when the covariance function is known, kriging of the mean coincides with the maximum likelihood estimator of the expectation. The kriging weights  $w_{\alpha}$  only depend on the covariance function and the relative position of the data. Kriging weights of clustered samples tend to be small or even negative while kriging weights of isolated samples sufficiently remote to other samples is nearly equal to the inverse of the equivalent number of independent observations. Kriging of the mean can thus be seen as a "natural declustering" of the data. Our third method consists in applying the kriging paradigm to the regional average of  $Y(\cdot)$  and use the resulting kriging weights in the CART algorithm. Because the partitioning algorithm needs positive weights to compute the heterogeneity indices, we impose a positiveness condition on the  $w_{\alpha}$ . The weights  $W$  will thus be the solution of the adapted kriging system:

$$\min_W \text{var}(W^T Y - Y_{\mathcal{D}}), \quad \text{with } \mathbf{1}^T W = 1 \quad \text{and} \quad w_{\alpha} \geq 0,$$

where  $Y = (Y_1, \dots, Y_n)^T$ . We have chosen to present this third method on the variable  $Y$  because it is the most natural choice. But it is also possible to compute these weights on one of the explanatory variables  $X^k$ .

Weighted methods are aimed at reducing the bias of the regression tree by taking into account the spatial redundancy of the data. This implies that the equivalent number of independent data is reduced, hence that the variance of the classification and regression parameters are increased.

## 4 Spatial CART

### 4.1 Gini index for spatial data

For independent samples, the Gini index is by definition  $D = P(Y_\alpha \neq Y_\beta) = \sum_i \sum_{j \neq i} p_i p_j = 1 - \sum_i p_i^2$ . In the case of two classes, only one parameter describes the proportion of each class and the variance of the indicator of, say class 1, and hence

$$D = 2p(1 - p) = 2\sigma^2.$$

There are several possibilities for extending the definition of the Gini index to spatial data.

One can write  $G_1 = P\{Y(S) \neq Y(S')\}$  for two uniform random points  $S$  and  $S'$  of  $\mathcal{D}$ . Then,

$$\begin{aligned} G_1 &= P\{Y(S) \neq Y(S')\} \\ &= \frac{1}{|\mathcal{D}|^2} \int_{\mathcal{D}} \int_{\mathcal{D}} P\{Y(s) \neq Y(s')\} ds ds' \\ &= \frac{1}{|\mathcal{D}|^2} \int_{\mathcal{D}} \int_{\mathcal{D}} E[\mathbb{I}\{(Y(s) - Y(s'))^2 = 1\}] ds ds' \\ &= \frac{2}{|\mathcal{D}|^2} \int_{\mathcal{D}} \int_{\mathcal{D}} \gamma(s - s') ds ds' \\ &= 2\bar{\gamma}(\mathcal{D}, \mathcal{D}) \end{aligned}$$

where  $\gamma(h) = p(1 - p) - C(h)$  is the variogram of the random function  $Y(\cdot)$ , and  $\bar{\gamma}(\mathcal{D}, \mathcal{D})$ , the average of  $\gamma(s - s')$  for  $s$  and  $s'$  in  $\mathcal{D}$ , can be computed from the fitted variogram.

One can alternatively start from  $G_2 = 2\text{Var}\{Y(\cdot)\}$ , in which case  $G_2 = 2 \lim_{\|h\| \rightarrow \infty} \gamma(h)$  is directly given by the estimated variogram. Note that  $G_1$  is always smaller than  $G_2$  since it is an average that includes values at small distances for which the variogram is smaller than its sill.

Finally, instead of estimating the variance from the variogram function, one can re-compute the Gini index from the estimated proportion:  $G_3 = 2\hat{p}(1 - \hat{p})$ , where  $\hat{p}$

is the estimated proportion of class 1 using the kriging equations (4), for which

$$\hat{p} = Y^T \Lambda, \quad \Lambda = \mathbf{C}^{-1} \left( C_{\mathcal{D}} + \frac{1 - \mathbf{1}^T \mathbf{C}^{-1} C_{\mathcal{D}}}{\mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}} \mathbf{1} \right).$$

Then, it is easy to show that  $E(\hat{p}) = p$  and  $E(\hat{p}^2) = \Lambda^T \mathbf{C} \Lambda + p^2$ . Since  $G_3 = 2\hat{p}(1-\hat{p})$ , we have

$$E(G_3) = 2p(1-p) - 2\Lambda^T \mathbf{C} \Lambda. \quad (6)$$

After straightforward developments, the last term in (6) is seen to be equal to twice  $C_{\mathcal{D}}^T \mathbf{C}^{-1} C_{\mathcal{D}} + \{1 - (\mathbf{1}^T \mathbf{C}^{-1} C_{\mathcal{D}})^2\} / \mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}$ . This term is in general small, in particular when  $\mathcal{D}$  is large compared to the range parameter of the covariance function. Notice that  $C_{\mathcal{D}} \rightarrow 0$  as  $\mathcal{D} \rightarrow \mathbf{R}^2$ . Hence, as  $\mathcal{D} \rightarrow \mathbf{R}^2$ ,  $G_1 \rightarrow 2\sigma^2 = 2p(1-p) = G_2$  and  $E(G_3) \rightarrow 2\sigma^2(1 - 1/n^*)$ , where

$$n^* = \sigma^2 \mathbf{1}^T \mathbf{C}^{-1} \mathbf{1}$$

is the equivalent number of independent data of  $Y$ .

The indices  $G_1$  and  $G_2$  only rely on the estimated covariance function, while  $G_3 = 2Y^T \Lambda(1 - \Lambda^T Y)$  depends on both the covariance function and the data.  $G_3$  is thus a more robust estimator of the Gini index than  $G_1$  and  $G_2$ . Furthermore,  $G_3$  can readily be generalized to any number of classes, which is not true for  $G_1$  and  $G_2$ . We will thus use  $G_3$  in the rest of this paper.

## 4.2 Estimation of the proportions in a leaf $t$

Consider a leaf  $t$  of the tree. The theoretical proportion  $p(j | t)$  of class  $j$  in  $t$  is the conditional probability  $P(Y = j | X \in B_t) = E\{\mathbb{I}(Y = j | X \in B_t)\}$  where  $B_t$  is the subdomain of  $\mathbf{R}^p$  defining the leaf  $t$ . It can thus be estimated by kriging the spatial average of the variable  $\mathbb{I}\{Y(\cdot) = j | X(\cdot) \in B_t\}$  over the domain  $\mathcal{D}_t$  where  $\mathcal{D}_t = \{s \in \mathcal{D} : (X^1(s), \dots, X^p(s)) \in B_t\}$ , i.e.

$$\{X^1(s), \dots, X^p(s)\} \in B_t \Leftrightarrow s \in \mathcal{D}_t.$$

Applying the kriging approach on the estimation of  $p(j | t)$  leads to

$$\hat{p}(j | t) = \sum_{\alpha: X(s_\alpha) \in B_t} \lambda_\alpha \mathbb{I}\{Y(s_\alpha) = j\}$$

where  $(\lambda_\alpha)$  is the solution of the system of  $n_t$  equations with  $\alpha : X(s_\alpha) \in B_t$ :

$$\sum_{\beta: X(s_\beta) \in B_t} \lambda_\beta C_j(s_\alpha, s_\beta) = \frac{1}{|\mathcal{D}_t|} \int_{\mathcal{D}_t} C_j(s_\alpha, s) ds, \quad (7)$$

under the constraint  $\sum_\alpha \lambda_\alpha = 1$ . In the above equations,  $C_j(s, s')$  is the covariance function of  $\mathbb{I}\{Y(s) = j | X(s) \in B_t\}$ . The Gini index is then computed from the estimated proportions:

$$D_t = 1 - \sum_i \hat{p}(i | t)^2.$$

Similarly, in the case of independent data, when the leaf  $t$  is split into 2 sub-leaves the quantity  $n_{t_-}/n_t$  is an estimate of the probability  $P(X \in B_{t_-} | X \in B_t) = E\{\mathbb{I}(X \in B_{t_-} | X \in B_t)\}$ . For spatially dependent data it can be estimated using the kriging of the spatial average of the variable  $\mathbb{I}(X(\cdot) \in B_{t_-} | X(\cdot) \in B_t)$  on  $\mathcal{D}_t$ . In this setting, the empirical risk is also estimated by kriging the spatial average of the variable  $\mathbb{I}[T\{X(\cdot)\} \neq Y(\cdot)]$  on  $\mathcal{D}$ .

This method seems appealing from a theoretical point of view, but it suffers from some practical difficulties. First, it is very time consuming because variogram fitting and kriging must be done in each leaf for each variable. Second, it is known in the geostatistics literature (Rivoirard, 1993) that kriging an indicator variable is not very efficient and leads to estimates that can be outside the interval  $[0, 1]$ . Because constraining the weights to be positive is even more time consuming, we prefer to shrink the estimates to 0 or 1 when necessary. Third, the domain  $\mathcal{D}_t$  on which the right hand side of (7) is computed is not known. It is approximated by the convex hull of the sample points lying in  $\mathcal{D}_t$  and the integral is computed on the points of the grid  $G$  falling in that convex hull.

## 5 Simulations

In this section we compare on simulations standard CART (ID), the three weighted CART methods, namely Kriging of the mean (KM), Voronoï Cells (VC), Regular Grid (RG) and the spatial CART with estimated proportions (EP). To mimic a situation where the observations have strong spatial dependence and clustered samples we simulate in the unit square  $W$  a Neyman-Scott point process (Diggle, 1983) with a number of parents distributed as a Poisson random variable with parameter 2 and 10 children per parent within a circle of radius 0.01. We thus obtain a number  $n_1$  of clustered sample locations. Additional  $n_2$  sample locations are then simulated according to a homogeneous Poisson process with intensity  $100 - n_1$ .

Each point  $s = (x, y)$  in  $W$  is said to belong to class 1 if  $(x - 0.5)(y - 0.5) > 0$ , to class 0 otherwise. This rule generates four subsquares: the upper right and the lower left subsquares are in class 1 while the two remaining ones are in class 0. Following this scheme, the variable  $X(s)$  used to classify the data is  $X(s) = \beta(s) + \varepsilon(s)$ , where the signal is  $\beta(s) = \sum_{i=1}^4 \beta_i \mathbb{I}(s \in W_i)$ , with  $W_1 = \{x < 0.5, y < 0.5\}$ ,  $W_2 = \{x < 0.5, y > 0.5\}$ ,  $W_3 = \{x > 0.5, y > 0.5\}$ ,  $W_4 = \{x > 0.5, y < 0.5\}$  and  $(\beta_1, \beta_2, \beta_3, \beta_4) = (-3, -1, 1, 3)$ . The perturbation  $\varepsilon(s)$  is a Gaussian random function with an exponential correlation function with range parameter equal to 0.1. Because our goal is to evaluate the ability of the methods to deal with clusters of samples with values that strongly perturb the signal we condition the random function  $\varepsilon(s)$  to  $-2$  at the first point of the sample corresponding to the first cluster of samples (when such a cluster is present). Then, by the Cholesky decomposition method (Chilès and Delfiner, 1999; p. 465) we get a  $n$ -sample of a Gaussian random field.

Since  $E[\varepsilon(\cdot)] = 0$ , we expect the classification rule to allocate points for which  $X(\cdot) \in ]-\infty; -2] \cup [0; 2]$  to class 1 and other points to class 0. The cluster with a noise value around  $-2$  (when it is present) is here to perturb seriously the classification

rule. We expect the methods taking into account the spatial structure of the data to overcome the difficulty.

For each method some parameters were to be set. KM and EP: the integrals (5) are computed on a regular  $101 \times 101$  grid; an exponential variogram model is fitted. VC: the domain boundary is the unit square; if an area associated with a point cannot be computed, it is set to 0. RG: since  $\sqrt{n} \simeq 10$ , the square is divided into  $10 \times 10$  cells. For all methods the minimum number of samples for splitting a leaf is 10. For weighted and standard CART the penalization parameter is the default value (equal to 0.01) of the `rpart` function in the R software.

To validate the classification rules, on each simulation we simultaneously simulate a validation set of 100 uniform locations in  $W$  from the same Gaussian field. We performed 100 simulations: 9 simulations have no cluster, 26 have 1 cluster, 36 have 2 clusters, 29 have 3 clusters or more. Points in the validation set can be misclassified for two different reasons: i) the splitting rule is wrong or, ii) although the splitting value is correct, the explained variable is locally perturbed by the noise.

Table 1 shows the average number of misclassified points according to the number of clusters. The first number is the average total number of misclassified points; the second one is the average number of points that are within the correct intervals, but misclassified when applying the estimated classification rule. Table 2 presents the number of leaves.

Table 1 here

Table 2 here

To better evaluate the effect of an estimated classification rule, we also calculate the measure of wrong allocation in the interval  $[\min(X), \max(X)]$ . For example if

the classification rule is: class 1 corresponds to  $[\min(X), -1.5) \cup [-0.1, 3)$  and class 0 corresponds to  $[-1.5, -0.1) \cup [3, \max(X)]$  with  $\min(X) = -4$  and  $\max(X) = 4$ , the proportion of wrong allocation is  $(0.5 + 0.1 + 1)/\{\max(X) - \min(X)\} = 0.2$ . Figure 1 shows the boxplot of the measure of wrong allocation for each method and for each number of clusters.

Figure 1 here

These results show that when there is no cluster, the Standard method (ID) generally performs better than the other ones. When the sample locations are not clustered, the spatial dependence does not affect too much the classification rule. Taking into account the spatial dependence actually reduces the number of equivalent independent data and disadvantages weighted and spatial methods.

As soon as there is at least one cluster, ID ranks last and wrong allocation is higher on average. Among the methods taking into account the spatial correlation, weighted methods provide the most significative improvement. Table 2 shows that EP favors less leaves than the other methods, thereby leading to an increasing misclassification rate. Let us now consider more closely two particular cases. Remember that on these simulations the lower-left and upper-right subsquares are in class 1, corresponding to  $X \in ]-\infty, -2] \cup [0, 2]$ .

*Simulation with no cluster*

Figure 2 here

We first consider the case of a simulation with no cluster of sample locations. On this simulation, some points with an  $X$  value not in the expected interval are isolated and hence have important weight. The weighted methods thus create more intervals than the standard method in an attempt to adapt to these data and to class them correctly. With less leaves, ID performs better than all other methods.

Figure 3 here

In this simulation there are 16 points clustered in the upper-right subsquare with values around  $-1$  (expected value in  $[0, 2]$ ) and 12 points clustered in the upper-left subsquare with values close to  $0.5$  (expected value in  $[-2, 0]$ ). This is a very unfavorable case for ID because all samples in the clusters have the same weight, equal to  $1/n$ , thus leading to large global weights for the clusters. It yields to a single splitting value at  $-0.5$  with 40% wrong allocation and 32% misclassification. The other methods yield to comparable solutions with four leaves, wrong allocation between 17% and 25% and misclassification between 21% and 25%.

## 6 Example

Future climate change will strongly affect vegetation distribution (Beerling et al., 1997). Reconstructing modern and past plant cover is essential to understand vegetation dynamic and to predict their future ranges under changing climate (IPCC, 2001). Pollen data are one of the most appropriate proxies to reconstruct modern and past vegetation. They are abundant in fossil records but they give a biased image of surrounding vegetation. Pollen records depend on: distance from the population to sampling site, population density, pollen production rates (rates are different between species, individuals and even between years), transport (depending on pollen morphology and density) and preservation (more or less resistant according to the thickness of their envelope). Palynological species were gathered into functional groups of plants, the Bioclimatic Affinity Groups of plants (BAGs) (Laurent et al., 2005). Laurent et al. (2005) georeferenced geographic ranges of 320 European species of plants and gathered these data following palynological taxonomy. Combination of taxa ranges with climate variables (New et al. 1999) provided potential

climate ranges for each taxon. Taking into account these climatic envelopes, Laurent et al. (2004) have created 25 BAGs using hierarchical cluster analyses. These BAGs are characterized by different geographical ranges and climatic tolerances and requirements. The distribution of taxa (families, genera or species) is georeferenced from the database SOPHY (<http://sophy.u-3mrs.fr/sommaire.html>). For each point of the grid, a binary variable indicates the presence or absence for each species. These binary values for plants belonging to the same BAG were averaged to create one map for each group. A total of 356 pollen samples were evenly collected in France. We averaged pollen counts of samples collected at the same place. The resultant 154 samples provided pollen percentages of BAGs. In this work we focus on BAG 8, which is principally located around the Mediterranean sea and near the Atlantic coast of south-western France (Figure 4a).

Figure 4 here

We assign each sample site to the nearest point class value. Then we estimate the link between pollen percentage and the presence/absence of BAG 8. Only 11 sites are in class 1, in which the pollen mean rate is 0.126 with a standard deviation of 0.09. The mean for class 0 is 0.008 with a standard deviation of 0.01. Hence class 0 implies low pollen percentage but the opposite is not true (see boxplot of Figure 5).

Some parameters have to be set to perform the analysis :

- only leaves with more than ten individuals can be split,
- an exponential variogram for KM and EP is fitted,
- the complexity parameter for ID, KM, EP and RG is 0.01 (the default value of library `rpart` in R).

Figure 5 shows the comparison between the five methods. The highest pollen percentage in class 0 is 9.99%, and over 12.44% all sites are classified as present.

This explains that all methods give the presence class over 11.24%. In the lower pollen percentage, the five methods provided quite different results. For example, the Atlantic site (Figure 4b) has low pollen percentage (0.44%) but is classified as present. This site is located in an area strongly impacted by Man which may explain the presence of BAG 8. One may stress also that this site is located at the border of a small cluster in the West of France. For VC and KM the associated weight is large (above 1%) because it is a border site and thus a class 1 leaf ( $[0.43\%, 0.47\%]$  for VC,  $[0.40\%, 0.46\%]$  for KM) is created. As a consequence, three sites with comparable pollen percentage but located in the center part of different clusters (therefore with very low weight) are misclassified.

Three sites in class 1, located near the Mediterranean sea (Figure 4b), have their pollen percentage between 2.8% and 3.6%. They are isolated and have an important weight (from 1.5 % to 3%) in the weighted methods. Four Pyrenean points have their pollen percentage in the same interval, but since they are located in the center of a cluster, they have very low or null weight. VC, EP, RG and KM thus generate a new cut to handle the three sites of class 1 with low pollen percentage, whereas ID favors the Pyrenean absence sites. The upper limit is the same for all methods but the lower limit differs. Thirteen sites are concerned by the change of the lower limit, all in the Pyrenean cluster. This example typically highlights the difference between the four methods. For VC and KM, the weights of sites within the cluster is almost null and therefore not taken into account in the analysis. For RG, these sites have the same weight than those at the border as soon as the whole cluster is within the same cell. Their weight being sufficiently large they participate to the decision rule and make the lower limit of the cut higher : 2,77% instead of 1,69%. For EP, the weights are computed iteratively at each split of a leaf. At the beginning of the procedure the weights of the sites within the cluster are almost null but, since the percentage of pollen of Pyrenean sites is highly variable, weights increase when sites of the cluster are assigned to different leaves. Finally the sites within the cluster

end with non null weights and are considered when splitting. Therefore the lower limit of the cut for EP is the same as RG.

Figure 5 here

The classification obtained for the five methods differ substantially for low pollen percentage. A key point in past vegetation reconstructions is the threshold of pollen percentage beyond which a plant species is considered present in the area where the site is located. Although, that threshold depends on the differential capacity of pollen production of the species, the generated threshold by ID (11%) is uninformative because it is too high. It is well known that such a level of pollen leads to the presence of the species. The most informative information is the upper limit of the cut at 2.7% . VC and KM produce the lowest limit but they ignore the sites within the cluster. It is a crude way to "decluster" the data. RG and EP tend to summarize the information of the cluster and gives probably the most informative threshold.

## 7 Conclusion and discussion

We have presented some adaptations to the CART algorithm that have been shown useful when the sampling pattern is very irregular, in particular in the presence of clusters. Standard CART (ID) shows systematically higher misclassification rate on validation dataset than adapted methods when clusters of samples are present. Taking into account the spatial dependence between data reduces the bias in the regression parameters, at the cost of higher variance. This is exemplified by the EP algorithm for which the misclassification rate is increased despite a higher adaptability. The simulation study and the analysis of real data set show that when the sample locations are not clustered standard CART can be used without restriction. In the presence of clusters, methods taking into account the spatial organization

and/or the spatial dependence should be preferred. In particular the weighted methods have in general lower misclassification rates. Among them, it must be understood that with VC and KM sample sites at the border of a cluster have systematically higher weights than those in the center, a problem not encountered with RG. However, since RG is sensitive to the size of the grid and the location of its origin, it is sometimes advisable to test different grid size and origin and, if necessary, apply a voting system for the classes. On simulations EP leads often to higher misclassification rate, but although being difficult to tune and a more time consuming method, it has proved to be useful when several classes are present in clusters. As illustrated on the example, in a an data exploration stage, it is actually of great interest to analyze the difference between the classification trees provided by the different methods.

Such statistical approaches were developed in order to perform comparisons between observed pollen data and vegetation simulated ecosystems or BAGs. Here we have tested the methods for only one BAG (Mediterranean). In order to reach a more accurate data-model comparisons over Europe we aim to set up thresholds for all other different BAGs (25) developed by Laurent et al. (2004).

## References

- [1] Allard, D. and Fraley, Ch. (1997). Non Parametric Maximum Likelihood Estimation of Features in Spatial Point Processes Using Voronoï Tessellation, *Journal of the American Statistical Association*, **92**, 1485-1493.
- [2] Beerling, D.J., Woodward, F.I., Lomas, M. and Jenkins, A.J. (1997). Testing the responses of a dynamic global vegetation model to environmental change: a comparison of observations and predictions. *Global Ecology and Biogeography Letters*, **6**, 439-450.

- [3] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984). *Classification and regression trees*. Wadsworth International Group, Belmont.
- [4] Chilès, J.P. and Delfiner, P. (1999). *Geostatistics: modeling spatial uncertainty*. Wiley, New-York.
- [5] Cressie, N. (1993). *Statistics for spatial data, Revised Edition*. Wiley, New-York.
- [6] De'ath, G. and Fabricius, K.E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, **81**, 3178-3198.
- [7] Diggle, P.J. (1983). *Statistical analysis of spatial point patterns*. Academic Press, New-York.
- [8] Dray, S., Pettorelli, N. and Chessel, D. (2002). Matching data sets from two different spatial samplings. *Journal of Vegetation Science*, **13**, 867-874.
- [9] Harrison, S.P. and Prentice, I.C. (2003). Climate and CO2 controls on global vegetation distribution at the last glacial maximum: analysis based on palaeovegetation data, biome modelling and palaeoclimate simulations. *Global Change Biol.*, **9**, 983-1004.
- [10] Hennig, C. and Hausdorf, B. (2004). Distance-based parametric bootstrap tests for clustering of species ranges. *Computational Statistics and Data Analysis*, **45**, 875-896.
- [11] Intergovernmental Panel on Climate Change (2001). *Climate change 2001: the scientific basis*, Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Dai X, Maskell K, Johnson CA (eds). Cambridge University Press, Cambridge.
- [12] Isaaks, E.H. and Srivastava, R.M. (1989). *An Introduction to Applied Geostatistics*, Oxford University Press, New York.

- [13] Laurent, J.M., Bar-Hen, A., François, L., Ghislain, M., and Cheddadi, R. (2004): Refining vegetation simulation models: From Plant Functional Types to Bioclimatic Affinity Groups of plants. *Journal of Vegetation Science*, **15**, 739–746.
- [14] New, M., Lister, D., Hulme, M. and Makin, I., 2002. A high-resolution data set of surface climate over global land areas. *Clim. Res.*, **21**, 1-25.
- [15] Okabe, A., Boots, B., Sugihara, K. and Chiu S.N. (2000) *Spatial tessellations: concepts and applications of Voronoi diagrams*. Second edition. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester.
- [16] Picard, N. and Bar-Hen, A. (2000). Estimation of the Envelope of a Point Set with Loose Boundaries. *Applied Mathematics Letters*, **13**, 13-18.
- [17] Ripley, B.D. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.
- [18] Rivoirard, J. (1994) *Introduction to Disjunctive Kriging and non Linear Geostatistics*. Clarendon Press, Oxford .
- [19] Wackernagel, H. (2003). *Multivariate Geostatistics*. Springer-Verlag, 3rd edition, Berlin.

Table 1: Statistics on misclassified points on test samples.

clusters	EP		KM		VC		RG		ID	
	total	rule	total	rule	total	rule	total	rule	total	rule
0	32.8	17.7	29.9	12.7	30.4	13.3	30.3	11.9	29.7	10.8
1	28.3	14.5	27.7	12.1	26.5	10	27.8	12.8	29.8	16.5
2	28.3	14.1	25.9	10.8	26.9	11.7	26.3	10.8	28.9	15.6
$\geq 3$	29.3	14.4	27.6	11.3	27.9	10.7	28.5	11.3	31.9	17.6

Table 2: Number of leaves for each method

	EP	KM	VC	RG	ID
2 leaves	23	0	2	0	10
3 leaves	0	1	0	0	3
4 leaves	70	73	69	68	55
> 4 leaves	7	26	31	32	32

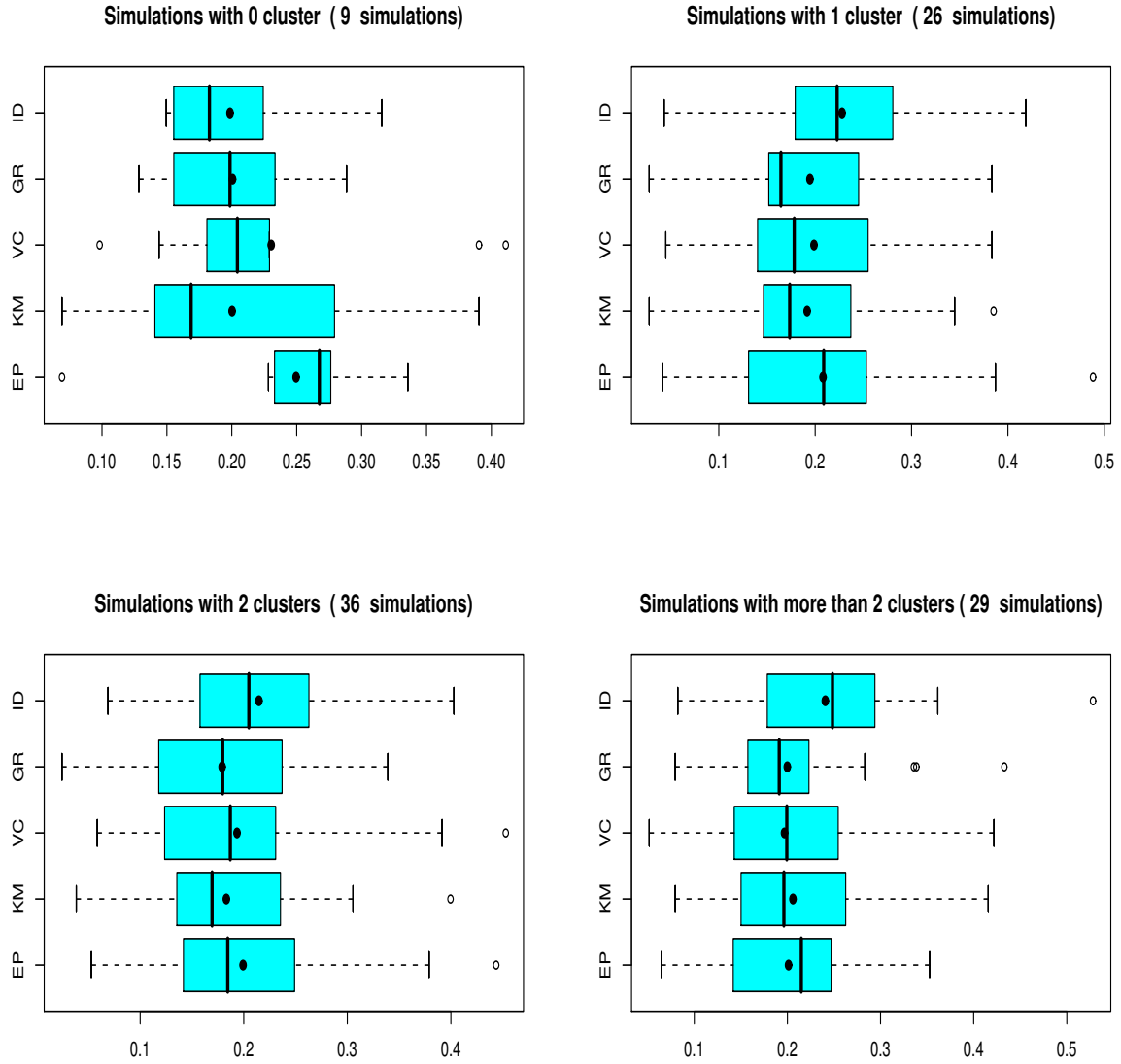


Figure 1: Measure of wrong allocation for the five methods, by number of clusters

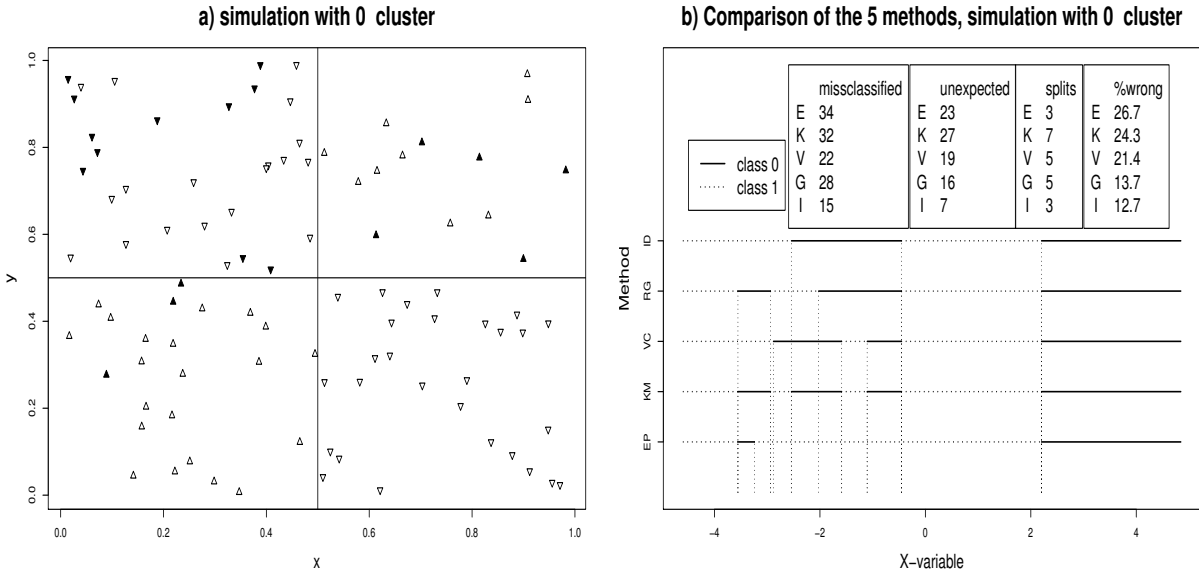


Figure 2: a) Simulation with no cluster. Upward triangles in class 1, downward triangles in class 0; black triangles are points for which  $X(s_\alpha)$  is not in the expected interval. b) Classification rule for the five methods

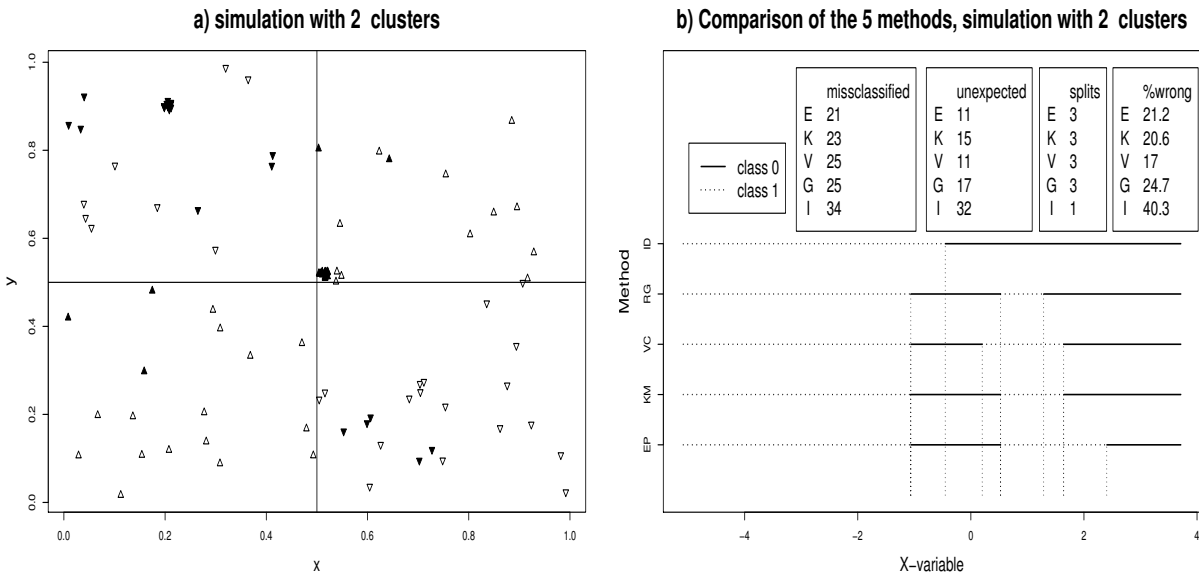


Figure 3: Simulation with two clusters

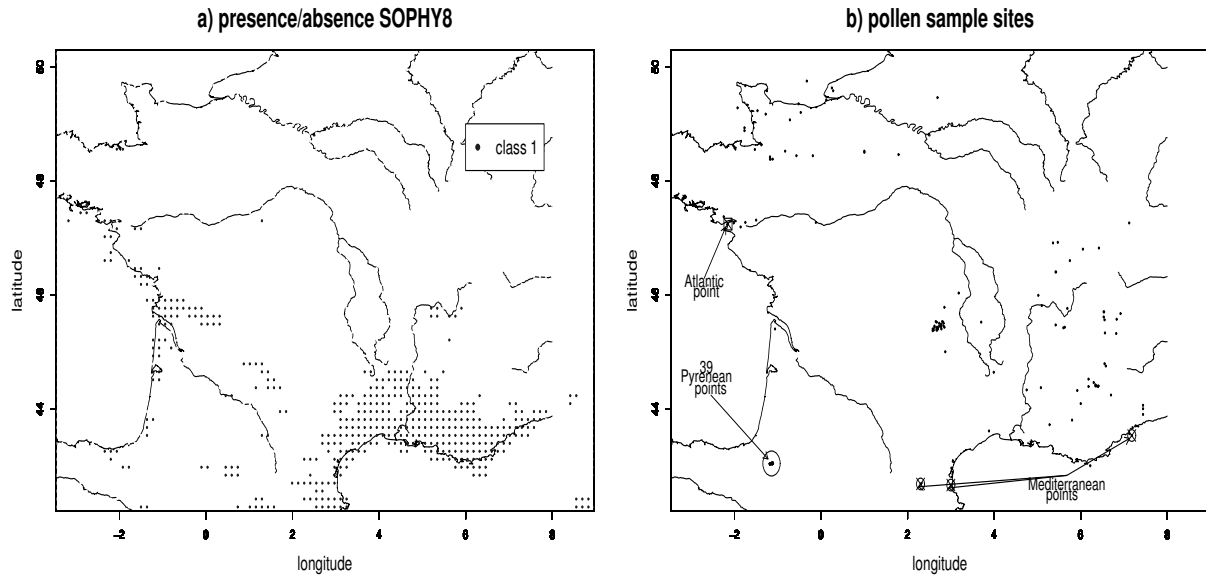


Figure 4: a) Presence and absence of BAG 8; b) sample sites of pollen frequencies,  $\boxtimes$ :Atlantic site,  $\otimes$ : Mediterranean sites

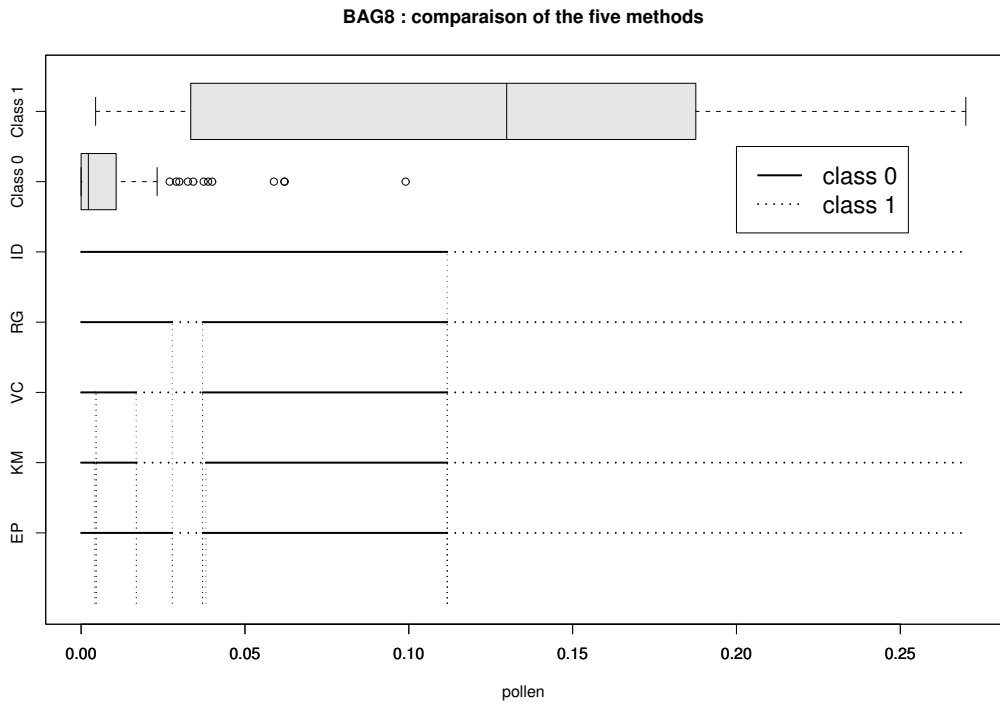


Figure 5: Classification rule for the five methods