Computational Statistics and Data Analysis 53 (2009) 3082–3093

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

CART algorithm for spatial data: Application to environmental and ecological data

L. Bel [a,*], D. Allard [b], J.M. Laurent [c], R. Cheddadi [c], A. Bar-Hen [d]

[a] UMR 518 AgroParisTech/INRA, 16 rue Claude Bernard, 75231 PARIS Cedex 05, France

[b] Unité Biostatistique & Processus Spatiaux, INRA, 84914 Avignon, France

[c] Institut des Sciences de l'Évolution, CP 61, CNRS UMR 5554, 34095 Montpellier, France

[d] Université Paris Descartes, Paris 5, UMR CNRS 8145 MAP5, 45 rue des Saints Pères, 75270 Paris cedex 06, France

ARTICLE INFO

Article history:
Available online 18 September 2008

ABSTRACT

Most statistical learning techniques such as Classification And Regression Trees (CART) assume independent samples to compute classification rules. This assumption is very practical for estimating quantities involved in the algorithm and for assessing asymptotic properties of estimators. In many environmental or ecological applications, the data under study are a sample of some regionalized variables, which can be modeled as random fields with spatial dependence. When the sampling scheme is very irregular, a direct application of supervised classification algorithms leads to biased discriminant rules due, for example, to the possible oversampling of some areas. The CART algorithm is adapted to the case of spatially dependent samples, focusing on environmental and ecological applications. 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. The second one uses spatial estimates of the quantities involved in the construction of the discriminant rule at each step of the algorithm. These methods are tested on simulations and on a classical dataset to highlight their advantages and drawbacks. They are then applied on an ecological data set to explore the relationship between pollen data and presence/absence of tree species, which is an important question for climate reconstruction based on paleoecological data.

1. Introduction

Most statistical learning techniques assume independent samples to compute classification rules. This assumption is useful for computing the class proportions, estimating other quantities involved in the algorithm and assessing asymptotic properties of estimators. When dealing with environmental and ecological questions, samples often consist of spatial measurement of variables within a domain. For practical reasons, it is often difficult to sample data over the whole studied area. Sampling units are thus often clustered. Because close observations are more similar than remote ones in the presence of spatial dependence, the spatial pattern of the sampling units is a critical issue. A direct application of supervised classification algorithms leads to biased discriminant rules because the same weight is given to every record and thus regions with high sampling density are overweighted.

* Corresponding author. Tel.: +33 144081670; fax: +33 144081666.
E-mail addresses: Liliane.Bel@agroparistech.fr (L. Bel), Denis.Allard@avignon.inra.fr (D. Allard), cheddadi@isem.univ-montp2.fr (R. Cheddadi), avner@math-info.univ-paris5.fr (A. Bar-Hen).

0167-9473/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.09.012

There are many ways to construct discrimination rules. In this paper we focus on Classification And Regression Trees (CART; Breiman et al. (1984), see Section 2). CART procedures have proven to be very useful in ecological and environmental 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 data sets 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. Various extensions of CART are proposed in order to improve the robustness (Gey and Poggi, 2006). We focus on CART but the main idea is to obtain a spatial estimate of the parameters involved in the classification rule. In a related context Hennig and Hausdorf (2004) tries to find clusters from a presence–absence matrix. Dray et al. (2002) focuses on the matching of two spatial samplings. A first attempt to account for spatial redundancy was proposed in Bel et al. (2005). The idea was to "decluster" the data: since kriging is a classical approach for spatial data, observations were weighted with the weights obtained from a kriging of the spatial mean.

Two approaches for dealing with spatial dependencies in samples are considered successively. The first one, presented in Section 3, accounts for the irregularity of the sampling by weighting the data according to their spatial pattern. Existing geometric methods based on Voronoï tessellation and regular grid are first considered, as well as the algorithm presented in Bel et al. (2005). These three weighting methods will be compared with standard CART and a second approach, which is our main proposal. In this second approach, presented in Section 4, we will use spatial estimates based on the spatial dependence on all quantities involved in the construction on the discriminant rule, class proportions and empirical risk. The mathematical models used in geostatistics allow us to model the dependence between the observations and provide unbiased estimates of the parameters of the classification rule. This approach fully takes into account the spatial structure of the data and can be adapted to other classification techniques.

To compare the proposed methods and to highlight their advantages and drawbacks, we tested them in Section 5 on simulations of spatially dependent samples and on the well known Swiss Jura data-set (Atteia et al., 1994; Goovaerts, 1997). We then proceed with our motivating example in Section 6. Climate reconstruction models are based on spatial distribution of plant species. Since plants are not readily preserved in the fossil record, the plant species distribution is interpolated from fossil pollen data. It is therefore assumed that presence/absence of plants on a given place can be predicted from pollen percentage. The validity of this assumption is tested with pollen samples irregularly collected in France and a map of vegetation. It is a typical supervised classification problem with one explanatory variable (pollen percentage) to predict class affection (presence/absence of plant from a map of vegetation). Finally results are summarized and discussed 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 a more detailed presentation see Breiman et al. (1984) or Ripley (1996), Chap. 7). The data are considered as independent samples of random variables $(X^1, \ldots, X^p, Y)$, where the X^ks 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 a 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 consider a decision tree T with one of its leaves t. Mathematically speaking, T is a mapping that assigns a leaf t to each sample $(X_i^1, \ldots, X_i^p)$, where i is an index for the samples. T can be viewed as a mapping to assign a value $\widehat{Y}_i = T(X_i^1, \ldots, X_i^p)$ to each sample. Let $p(j \mid t)$ 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

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

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

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

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 leaf 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 assigned to the most present class (the conditional mode). In general the final tree overfits the available data and the prediction error $R(T) = P\{T(X^1, \ldots, X^p) \neq Y\}$ 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 risk is inferior to $R(T)$. Since the distributions of Y and $X^1, \ldots, X^p$ 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 proportions $p(j \mid t)$ are estimated by $\widehat{p}(j \mid t) = n_{jt}/n_t$, 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 $D_t - \left(n_{t\_} D_{t\_} + n_{t+} D_{t+}\right)/n_t$ for the Gini index, and similarly for the entropy index. The empirical risk is

$$\widehat{R}(T) = \frac{1}{n} \sum_{\alpha=1}^{n} \mathbb{I}\{T(X_\alpha^1, \ldots, X_\alpha^p) \neq Y_\alpha\}, \tag{2}$$

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 $\{X^1(s_\alpha), \ldots, X^p(s_\alpha), Y(s_\alpha)\}$, $\alpha = 1, \ldots, n$ are originated from random fields $\{X^1(\cdot), \ldots, X^p(\cdot), Y(\cdot)\}$ 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 ones. 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 $(X^1, \ldots, X^p)$. Specifically, we will consider that

$$\widehat{p}(i \mid t) = \frac{1}{\sum_{\alpha \in t} w_\alpha} \sum_{\alpha \in t} w_\alpha \mathbb{I}\{Y(s_\alpha) = i\}, \tag{3}$$

and

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

with the condition $\sum_\alpha w_\alpha = 1$.

There are obviously many ways to weight the data. We will consider three different methods for determining these weights. The first two methods, with very different behavior are non parametric in essence and are only based on the geometry of the sampling design: in short, the weight associated with each observation is related to the local density of samples: the higher the local density, the lower the weights. The third method, presented in Bel et al. (2005) relies on kriging weights.

3.1. Weights derived from Voronoï tesselation

A first method is to use the Voronoï tessellation generated by the sample locations $(s_1, \ldots, s_n)$. A Voronoï cell around a sample location s_α is the set of points of $\mathcal{D}$ closer to s_α than to any other sample location (Okabe et al., 2000). Clustered observations produce smaller cells while sparse data produce larger ones. The inverse of their area is an estimator of the local density of the sample design, a property that has been used in spatial statistics for the clustering of spatial point processes (Allard and Fraley, 1997) or for the estimation of boundaries (Picard and Bar-Hen, 2000). The weight of a sample at a site s_α is thus set to be proportional to the area of its Voronoï cell. This approach is attractive and easy to implement but leads to undesirable boundary effects: data at the border of the domain have larger weights than data within the domain and weights of samples near the border depend strongly on the limit of the domain.

3.2. Weights derived from a regular grid

A related method is the declustering technique proposed in Isaaks and Srivastava (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 of the data must not be contained in just a handful of cells) nor too small (most cells must contain more than 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 covered by the grid. Note that for these two methods, the weights are proportional to the inverse of a non parametric estimate of the local density of samples.

3.3. Weights derived from geostatistics

Based on Bel et al. (2005), we now propose a third method for deriving the weights that takes into account spatial dependencies on the data under study. This method is based on a covariance function modeling the spatial dependence of the data. If the class variable is binary or ordinal, its covariance function can be easily estimated. If the class variable is

nominal, the covariance function of one of the explanatory variables or from a linear combination of the variables (e.g. first principal component) can be used.

We first recall briefly some notions of geostatistics that will be used in this section and in Section 4. Further details can be found e.g. in Wackernagel (2003) or Chilès and Delfiner (1999). 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 $\mathcal{D} \to \mathbf{R}^2$). 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 $\mathrm{Var}(\hat{Z}_{\mathcal{D}} - Z_{\mathcal{D}})$ is minimum. It can be shown that the vector $W = (w_1, \ldots, w_n)^T$ is the solution of the system

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

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

$$C(s_\alpha, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} C(s_\alpha, s)\mathrm{d}s, \tag{5}$$

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

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

The kriging weights w_α 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 from other ones are 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 using the kriging weights computed from the covariance function estimated on the variable Y. In most situations the kriging weights are all positive. In some situations however some of the weights can be negative. Since the partitioning algorithm needs positive weights to compute the heterogeneity indices, we impose a positiveness condition on the w_α. The weights W will thus be the solution of the adapted kriging system:

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

where $Y = (Y_1, \ldots, Y_n)^T$. For the two classes case, extensive simulations can be found in Bel et al. (2005).

4. Spatial CART

4.1. General framework

Instead of simply introducing weights, a second approach consists in deriving spatial estimates for all quantities involved in the algorithm: proportions in leaves, Gini index and empirical risks. Taking into account the spatial dependencies in the data requires us to specify a statistical model. We will therefore consider a model approach, also called "population approach" of the CART algorithm (see e.g. Ripley (1996), Chap. 7, and references therein), in which a tree is a probability model for the training set.

We start from the general notation introduced in Section 2, in which a tree T is a function with a finite number of modalities. We will consider that the observed categorical variable is a non linear function of the expectation of the explanatory variables, $X^i(\cdot)$. These are modeled as the sum of a fixed effect $\beta_i(\cdot)$ which can vary in space and a random effect modeled as a stationary random field $\epsilon_i(\cdot)$ with null expectation,

$$X^i(s) = \beta_i(s) + \epsilon_i(s).$$

Hence we write

$$Y(s) = f(\beta^1(s), \ldots, \beta^p(s)),$$

where f only takes a finite number of values. This is the model we wish to fit. In this setting a tree T is an estimate of $f(\cdot)$.

4.2. Estimation of the proportions in a leaf t

Consider a leaf t of the tree T. The theoretical proportion $p(j \mid t)$ of class j in t is the conditional probability $P(Y = j \mid X \in B_t) = E\{\mathbb{I}(Y = j \mid X \in B_t)\}$ where B_t is the subdomain of $\mathbf{R}^p$ corresponding to the leaf t. It can thus be estimated by kriging the spatial average of the variable $I_j(s) = \mathbb{I}\{Y(s) = j \mid X(s) \in B_t\}$ over the domain $\mathcal{D}_t$ defined as the set of locations $s \in \mathcal{D}$ such that $(X^1(s), \ldots, X^p(s)) \in B_t$. Note that the domain $\mathcal{D}_t$ is not necessarily spatially connected. Applying

the kriging approach on the estimation of $p(j \mid t)$ leads to

$$\hat{p}(j \mid t) = \sum_{\alpha: X(s_\alpha) \in B_t} \lambda_\alpha I_j(s_\alpha),$$

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

$$\sum_{\beta: X(s_\beta) \in B_t} \lambda_\beta C_j(s_\alpha, s_\beta) = \frac{1}{|\mathcal{D}_t|} \int_{\mathcal{D}_t} C_j(s_\alpha, s)\, \mathrm{d}s, \tag{6}$$

under the constraint $\sum_\alpha \lambda_\alpha = 1$. In the above equations, $C_j(s, s')$ is the covariance function of $I_j(s)$ for $s \in \mathcal{D}_t$, which can be assumed (i) second order stationary because the ϵ_is are second order stationary; (ii) ergodic, *i.e.* $\frac{1}{|\mathcal{D}|} \int_{\mathcal{D}} C_j(0, s)\mathrm{d}s \to 0$ as $\mathcal{D} \to \mathbb{R}^2$.

The Gini index is then computed from the estimated proportions:

$$\hat{D}_t = 1 - \sum_i \hat{p}(i \mid t)^2. \tag{7}$$

For the sake of lighter notation, we drop the subscript referring to leaf t and consider that the proportion is estimated on some domain $\mathcal{D}$.

Proposition 1. *For the model described above, let us denote D the population Gini index in $\mathcal{D}$, and $\hat{D}$ its estimate computed from (7). Let us further denote $n(\mathcal{D})$ the number of samples in $\mathcal{D}$. Assume that the density of samples tends to a strictly positive quantity as the domain increases: $n(\mathcal{D})/|\mathcal{D}| \to \lambda > 0$ as $\mathcal{D} \to \mathbb{R}^2$. Then, as $\mathcal{D} \to \mathbb{R}^2$,*

$$E[\hat{D}] \to D.$$

Proof. The estimated proportions $\hat{p}_j$ are obtained from the solution of the kriging equation (6):

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

where $\mathbf{C}_j$ and $C_{j,\mathcal{D}}$ are the matrix (resp. vector) corresponding to the left-hand side (resp. right-hand side) of (6). It is easy to show that $E(\hat{p}_j) = p_j$ and $E(\hat{p}_j^2) = \Lambda_j^T \mathbf{C}_j \Lambda_j + p_j^2$. Hence, we have

$$E(\hat{D}) = D - 2 \sum_j \Lambda_j^T \mathbf{C}_j \Lambda_j. \tag{8}$$

After straightforward developments, each term of the sum in (8) is seen to be equal to twice $C_{\mathcal{D},j}^T \mathbf{C}_j^{-1} C_{\mathcal{D},j} + \{1 - (\mathbf{1}^T \mathbf{C}_j^{-1} C_{\mathcal{D},j})^2\} / \mathbf{1}^T \mathbf{C}_j^{-1} \mathbf{1}$. But, as $\mathcal{D} \to \mathbb{R}^2$, each element of the vector $C_{\mathcal{D},j} \to 0$ by the ergodic assumption; and $\mathbf{1}^T \mathbf{C}_j^{-1} \mathbf{1} \to \infty$ as $\mathcal{D} \to \mathbb{R}^2$ because the number of samples in $\mathcal{D}$ tends to infinity. Hence $E[\hat{D}] \to D$ as $\mathcal{D} \to \mathbb{R}^2$. □

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 $P(X \in B_{t_-} \mid X \in B_t) = E\{\mathbb{I}(X \in B_{t_-} \mid X \in B_t)\}$. For spatially dependent data it can be estimated using the kriging of the spatial average of the variable $\mathbb{I}(X(\cdot) \in B_{t_-} \mid X(\cdot) \in B_t)$ 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 $\mathcal{D}$. Since the domain $\mathcal{D}_t$ on which the right hand side of (6) 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 within that convex hull. Kriging an indicator variable has some drawbacks. For example, estimates can be outside the interval [0, 1]. Constraining the weights to be positive is prohibitive for computing time, therefore the estimates are shrunk to 0 or 1 when necessary.

5. Simulations and example

5.1. Simulations

In this section we use simulations to compare standard CART (ID) to the three weighted CART methods, namely Kriging of the Mean (KM), Voronoï Cells (VC), Regular Grid (RG) and to 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 square $\mathcal{D} = [-2, 2] \times [-2, 2]$ a Neyman-Scott point process (Diggle, 1983) with a number of parents distributed as a Poisson random variable with parameter λ and nc children per parent within a circle of radius ρ. 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 $200 - n_1$. Four classes are determined according to Fig. 1. For each point $s = (x, y)$ in $\mathcal{D}$, the measured

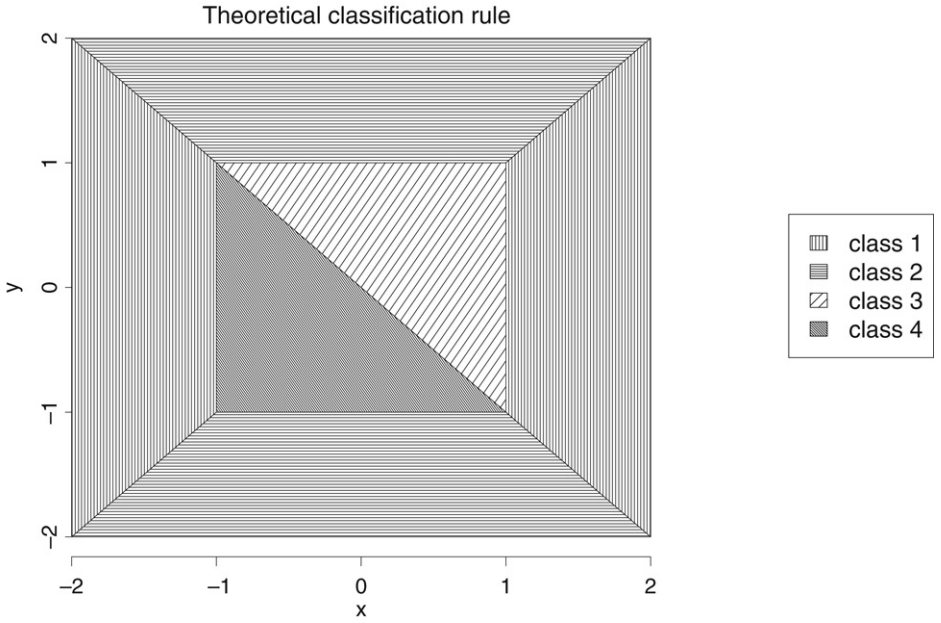


Fig. 1. Definition of the classes for simulations: $X^1 = y - x, X^2 = y + x, X^3 = x, X^4 = x$, x and y the coordinates.

Table 1
Simulations; average proportions of misclassified points, according to the number of clusters for the five methods (EP: Estimated proportion, ID: standard CART, KM: kriging of the mean, VC: Voronoi Cells, RG: Regular Grid)

Number of clusters (Number of simulations)	EP	ID	KM	VC	RG
0 (15)	0.16	0.20	0.18	0.20	0.18
1 (29)	0.18	0.20	0.19	0.19	0.20
2 (24)	0.17	0.19	0.18	0.18	0.19
3 (22)	0.18	0.22	0.18	0.19	0.19
4 (10)	0.15	0.18	0.16	0.16	0.16
All (100)	0.17	0.20	0.18	0.19	0.19

explicative variables are $X^1 = y - x + \varepsilon_1, X^2 = y + x + \varepsilon_2, X^3 = x + \varepsilon_3, X^4 = y + \varepsilon_4$. The perturbations $\varepsilon_i(s)$ are independent Gaussian random functions with an exponential correlation function with range parameter r and variance σ^2. For each method some parameters needed to be set. For KM and EP an exponential variogram model, chosen for its flexibility, is fitted. The integrals (5) are computed on a 101×101 grid, which insures a good estimate of the integrals. For VC and RG the domain boundary is $\mathcal{D}$. For RG the square is divided into 10×10 cells. The size of the grid is set empirically, as discussed in Section 3.2. For all methods the minimum number of samples for splitting a leaf is 10 in order to stabilize the estimates. For weighted and standard CART the penalization parameter is the default value (equal to 0.01) of the `rpart` function in the R software. This penalty avoids splittings with too small a decrease of heterogeneity.

The performance of the classification rules is assessed in the following way. On the 1600 points of a regular grid on $\mathcal{D}$, the classification rule is applied with the uncorrupted variables $(y - x, y + x, x, y)$ and compared to the theoretical value depicted Fig. 1. For each classification rule, the misclassification rate, R_m, is then computed. Applying the classification rule on the variables (X^1, X^2, X^3, X^4) instead of the uncorrupted variables would not allow us to separate the errors due to the classification rule from those due to random corruption of the variables. We performed 100 simulations. Table 1 shows the average proportion of misclassified points according to the number of clusters. A map of the pointwise misclassification rate is represented in Fig. 2.

These results show that the standard method (ID) ranks last with the highest average wrong allocation, whatever the number of clusters is. Among the methods taking into account the spatial correlation, spatial CART (EP) always provides the lowest misclassification rate. Weighted CART methods are intermediate between ID and EP, and not very different one from the other. As the number of clusters increases, the difference in performance between ID and the other methods increases as well, clearly showing the necessity of taking the correlation into account when data are clustered.

Fig. 2 shows that for all methods, the boundaries between classes are often misclassified. Standard CART is the most affected by the noise. This is particularly striking in the center and in the four corners of the domain, where the misclassification rate is higher than for the other methods. To investigate the sensitivity to the simulations parameters of these results, simulations for various values have been performed. Table 2 shows that, except in one case, EP always has the lowest misclassification rate R_m while ID has the highest. Other methods have intermediate results. The parameter

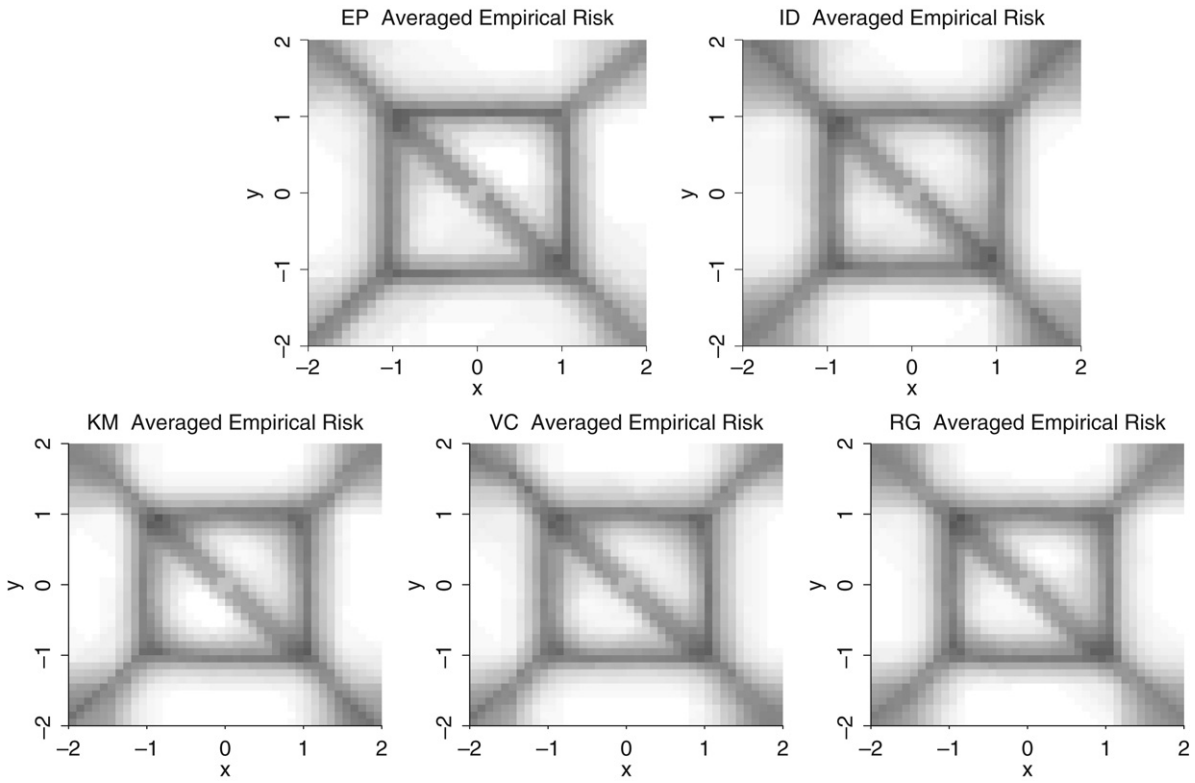


Fig. 2. Average of the pointwise misclassification rate on 100 simulations for the five methods (EP: Estimated proportion; ID: standard CART; KM: kriging of the mean; VC: Voronoi Cells; RG: Regular Grid). A dark color indicates a high misclassification rate: black for rate = 1 and white for rate = 0. Class definition is given in Fig. 1 and perturbations are Gaussian random functions with an exponential correlation function.

Table 2
Simulations; misclassification rate for different values of the simulation parameters, λ: Poisson parameter, ρ: radius, nc: number of children, r: range parameter, σ^2: perturbation ε variance, for the five methods (EP: Estimated proportion, ID: standard CART, KM: kriging of the mean, VC: Voronoi Cells, RG: Regular Grid)

λ	ρ	nc	r	σ^2	EP	ID	KM	VC	RG
2	0.01	10	0.1	0.1	0.17	0.20	0.18	0.19	0.19
5	0.01	10	0.1	0.1	0.19	0.22	0.20	0.20	0.21
2	0.05	10	0.1	0.1	0.17	0.20	0.19	0.19	0.19
2	0.01	5	0.1	0.1	0.18	0.20	0.19	0.19	0.20
2	0.01	20	0.1	0.1	0.18	0.21	0.20	0.19	0.20
2	0.01	10	0.2	0.1	0.19	0.21	0.20	0.20	0.20
2	0.01	10	0.3	0.1	0.20	0.23	0.22	0.21	0.23
2	0.01	10	0.1	0.2	0.23	0.24	0.23	0.22	0.23
2	0.01	10	0.1	0.3	0.26	0.27	0.26	0.26	0.27

that affects R_m the most is the noise variance, σ^2: as σ^2 increases, R_m increases for all methods with only non significant differences between them. Indeed, as σ^2 increases, the amount of spatial auto-correlation decreases and all methods are essentially equivalent.

5.2. Example: The Swiss Jura data-set

We now illustrate the methods on a set of soil data in a 14.5 km^2 region of the Swiss Jura. These data were obtained and first analyzed by Atteia et al. (1994). They are also presented and analyzed in Goovaerts (1997). The concentrations of seven heavy metal pollutants (cadmium, cobalt, chromium, copper, nickel, lead and zinc) were measured in the topsoil at 359 locations. The region is composed of five Jurassic limestone formations: Argovian, Kimmeridgian, Sequanian, Portlandian, Quaternary. Using a one-way analysis of variance, Atteia et al. (1994) shows that geology, and in particular the difference between Argovian and the other formations is the factor most strongly related to heavy metal concentrations.

Fig. 3 shows the boxplots computed for each heavy metal concentration and each geological formation. Class Portlandian has very few samples. Sequanian, Portlandian and Quaternary are hardly distinguishable. These three classes are thus

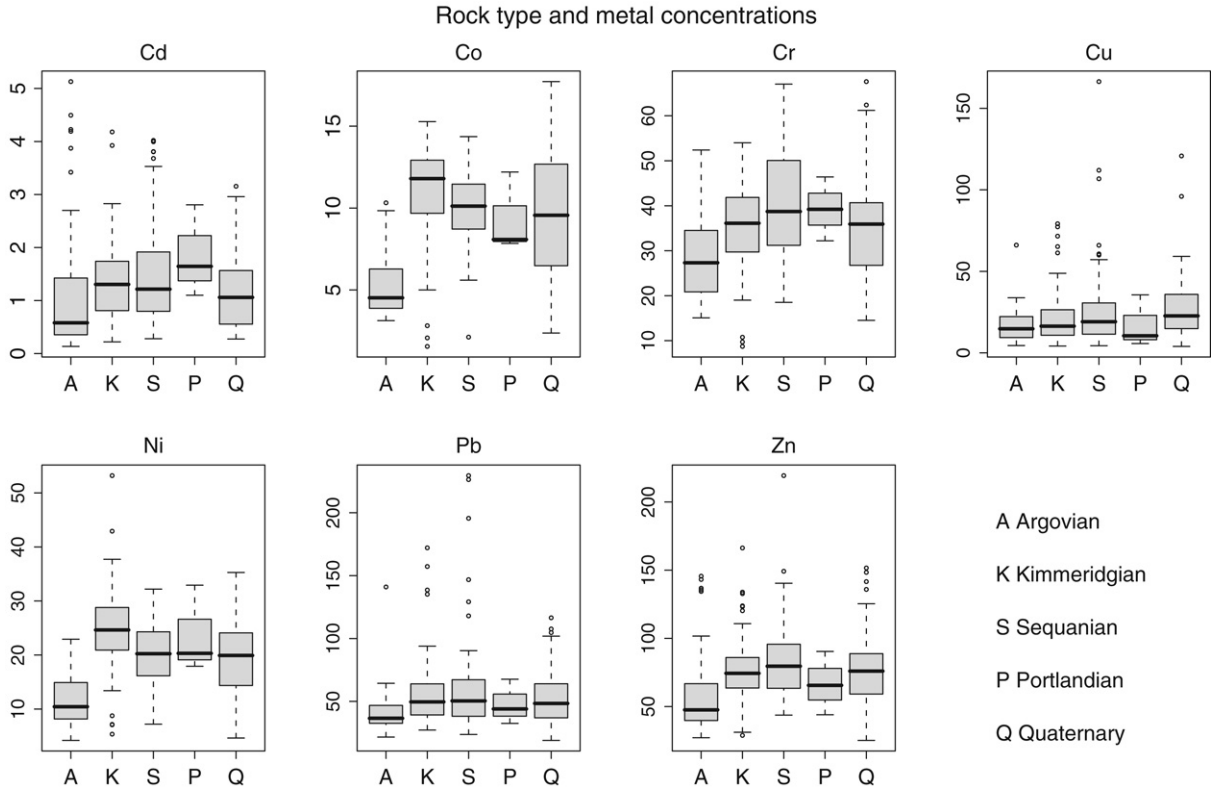


Fig. 3. Swiss Jura data-set; boxplot of heavy metal concentrations for cadmium (Cd), cobalt (Co), chromium (Cr), copper (Cu), nickel (Ni), lead (Pb), zinc (Zn), by geological class (Argovian, Kimmeridgian, Sequanian, Portlandian, Quaterny).

gathered into a single one. The geology is finally coded as follows: 1 for Argovian, 2 for Kimmeridgian and 3 for the others. To evaluate the performance of each method, 100 random samples from the data-set are set aside to form a validation data set. The remaining 259 samples are used for establishing the classification rule. All methods have been run with the same set of parameters as those used for the simulation study.

For the standard method (ID), a constant weight $1/259 = 0.0386$ is associated with all samples. For the weighted methods, isolated locations have higher weights while those of clustered samples are lower than 1/259. This is particularly striking for the Voronoï weights that can be very large for sample locations at the edge of the sampled area. Fig. 4 shows the resultant trees for all methods, with the empirical misclassification rate calculated on the validation set.

The empirical risk (i.e. the misclassification rate estimated on the learning set) is highest for standard CART (41%). It is lowest for EP (34%) and intermediate (from 37% to 40%) for the methods with weights, thus confirming the results obtained on simulations. Trees obtained with KM and VC are very similar. All five methods give the first split on the Cobalt (Co) variable, generally at threshold 5.9 (except 6.66 for the Voronoï method). This split discriminates the Argovian rock type (class 1) from the others. The five methods present globally similar misclassification rates for the Argovian type rock (Table 3), but some differences are worth mentioning. KM and RG make an additional split ($Co > 3.166$) to distinguish classes 1 and 2. This split is related to 3 outliers (see also the boxplots in Fig. 3) that have higher weights for these methods, but it increases the misclassification rate on the test sample and is therefore not relevant. EP creates a split based on Chromium (Cr), for a single sample only.

Differentiating class 2 and 3 is more difficult, and the results are quite different from one method to the other (see Table 3 and Fig. 5). Generally speaking, while class 2 is well predicted (from 27 to 33 samples out of 39), this is not the case for class 3 (from 17 to 21 out of 39). EP ranks first for class 2, and second for class 3. All five methods select the variables nickel (Ni), copper (Cu) and chromium (Cr) to design the classification rule but not in the same order. EP and RG also use the lead variable (Pb). The maps (Fig. 5) show some differences, but they coincide globally on the location of the Argovian rock type. EP provides the most regular classification map among all methods.

6. Application

Future climate change will strongly affect vegetation distribution (Beerling et al., 1997). Reconstructing modern and past plant cover is essential for understanding vegetation dynamic and predicting their future ranges under changing climate (Intergovernmental Panel on Climate Change, 2001). Pollen data are one of the most appropriate proxies for reconstructing

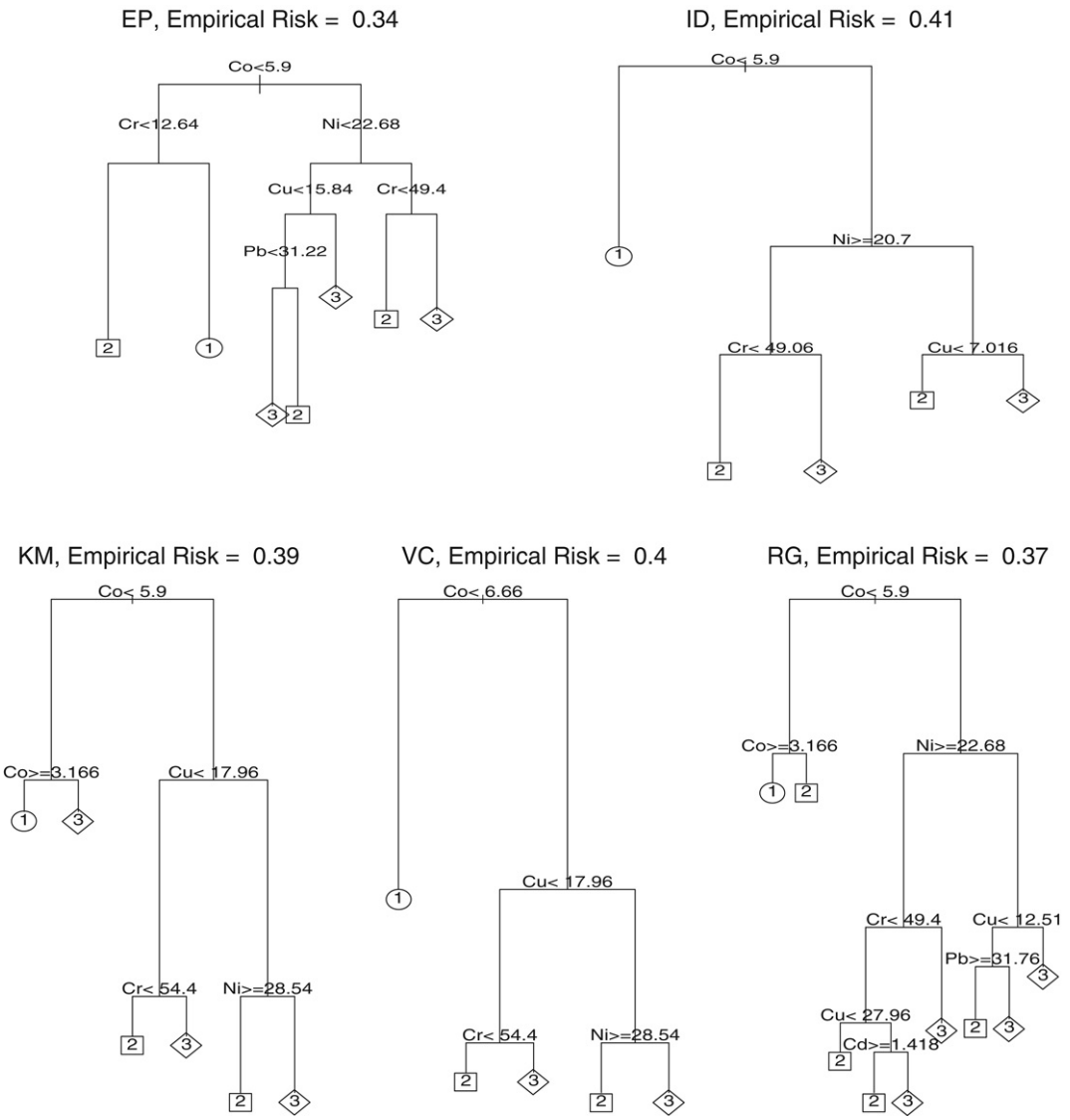


Fig. 4. Swiss Jura data-set; classification trees for rock type according to heavy metal variables cadmium (Cd), cobalt (Co), chromium (Cr), copper (Cu), nickel (Ni), lead (Pb), zinc (Zn) with the five methods (EP: Estimated proportion, ID: standard CART, KM: kriging of the mean, VC: Voronoi Cells, RG: Regular Grid). Circles: class 1 (Argovian); squares: class 2 (Kimmeridgian); diamonds: class 3 (Sequanian + Portlandian + Quarternary).

Table 3
Swiss Jura data-set; number of well classified locations of the validation set with the five methods (EP: Estimated proportion, ID: standard CART, KM: kriging of the mean, VC: Voronoi Cells, RG: Regular Grid)

	Class 1	Class 2	Class 3	Total
Number of samples	23	39	38	100
EP	14	33	19	66
ID	14	27	18	59
KM	13	30	18	61
VC	15	28	17	60
RG	13	29	21	63

Class 1 (Argovian), class 2 (Kimmeridgian), class 3 (Sequanian + Portlandian + Quarternary).

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).

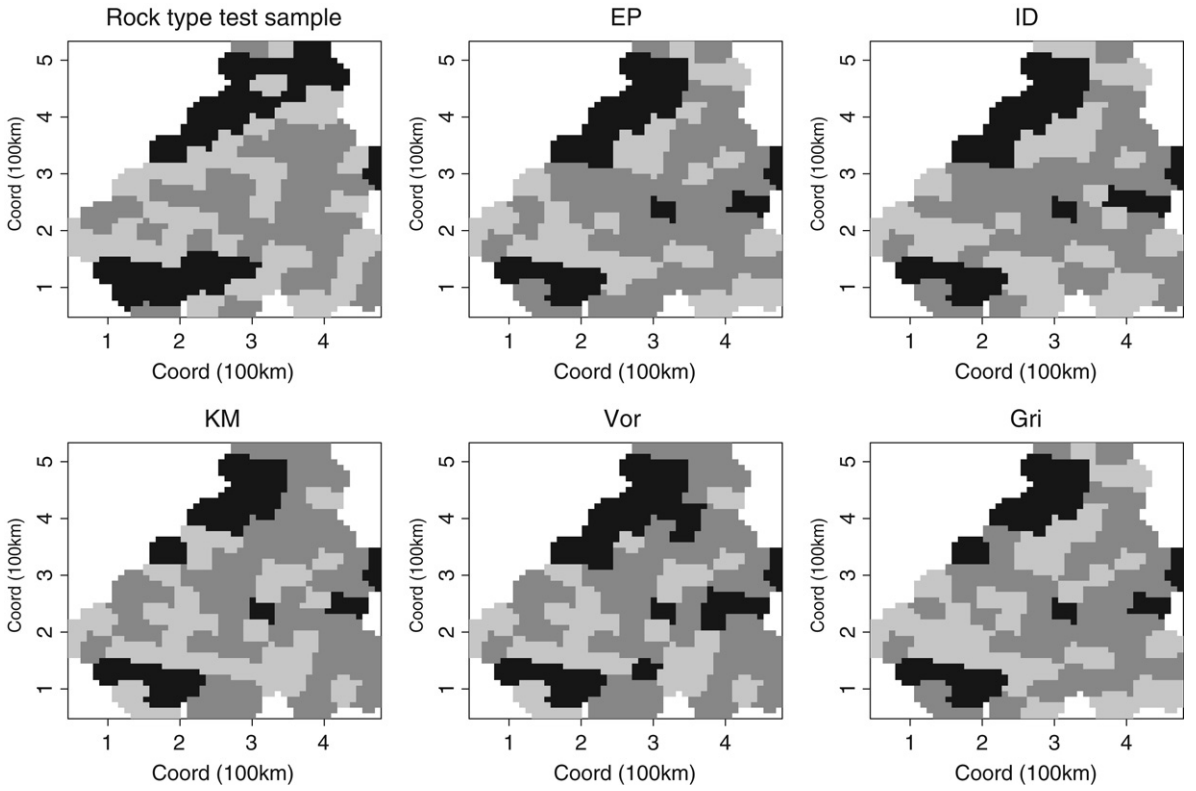


Fig. 5. Swiss Jura data-set; observed rock type and classification results for the test sample with the five methods (EP: Estimated proportion, ID: standard CART, KM: kriging of the mean, VC: Voronoi Cells, RG: Regular Grid). Black: Argovian; grey: Kimmeridgian; light grey: Sequanian+Portlandian+Quaternary.

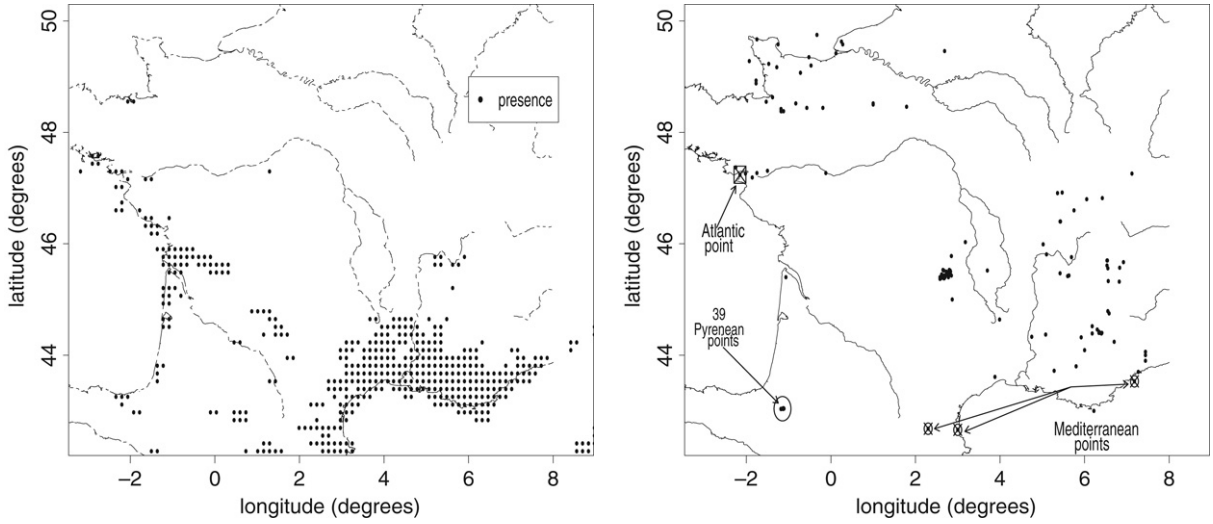
To evaluate the link between amount of pollen and presence of species in a site, we wish to elaborate a rule based on present data. The SOPHY database (http://sophy.u-3mrs.fr/sommaire.html) georeferences the distribution of taxa (families, genera or species). For each point of a grid, a binary variable indicates the presence or absence for each species. Palynological species are gathered into functional groups of plants, the Bioclimatic Affinity Groups of plants (BAGs) (Laurent et al., 2004). These BAGs are characterized by different geographical ranges and climatic tolerances and requirements. In this work we focus on BAG 8, which is principally located around the Mediterranean Sea and near the Atlantic coast of western France (Fig. 6a). The class variable is the presence/absence variable of plant BAG 8 for each point of the grid. We will call class 1 (resp. class 0) the class corresponding to the presence (resp. absence) of BAG 8.

A total of 356 pollen samples are irregularly collected in France. Pollen counts of samples collected at the same place are averaged. The resulting 154 samples provided pollen percentages of BAGs thus defining the explained variable. The pollen sample sites do not coincide with the BAG grid, and we assign to each sample site the nearest point class value.

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 also the boxplot in Fig. 7).

To perform the analysis the parameters are set to the same values as in the simulations. Fig. 7 shows the comparison between the five methods. The highest pollen percentage in class 0 is 9.99%. When the pollen percentage is larger than 11.24% 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 provide quite different results. For example, the Atlantic site (Fig. 6b) has a low pollen percentage (0.44%) but is classified as present. This site is located in an area strongly impacted by humans which may explain the presence of BAG 8. It may also be underlined that this site is located at the border of a small cluster in Western 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. Consequently, three sites with comparable pollen percentage but located in the central part of different clusters of data (therefore with very low weight) are misclassified.

Three sites in class 1, located near the Mediterranean Sea (Fig. 6b), 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 in the lower limit, all in the Pyrenean cluster. This example typically highlights the difference



(a) Presence/absence of plant BAG8.

(b) Pollen sample sites.

Fig. 6. (a) Dots indicate presence of plant BAG 8; (b) 154 sample sites of pollen records: ⊠: Atlantic site; ⊗: Mediterranean sites.

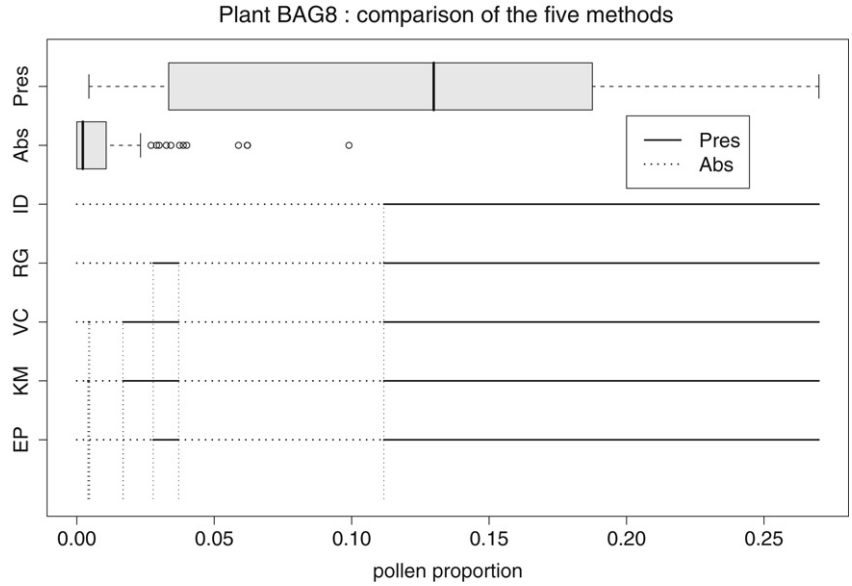


Fig. 7. Top line: boxplot of pollen frequencies for samples where BAG 8 is observed, second line: boxplot of pollen frequencies for samples where BAG 8 is not observed; lines 3–7: classification rule on plant BAG 8 for the five methods (EP: Estimated proportion, ID: standard CART, KM: kriging of the mean, VC: Voronoi Cells, RG: Regular Grid). Solid lines represent class 1 assignment (presence of BAG 8), dashed lines represent class 0 assignment (absence of BAG 8).

between the four methods. For VC and KM, the weights of sites within the cluster are almost null and therefore not taken into account in the analysis. For RG, these sites have the same weight as those at the border as soon as the whole cluster is within the same cell. Their weight being sufficiently large they participate in 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 up with non null weights and are considered when splitting. Therefore the lower limit of the cut for EP is the same as RG.

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. This threshold depends on the differential capacity of pollen production of the species.

The threshold generated 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 about the cluster and probably provide the most informative threshold.

7. Conclusion and discussion

We have presented some adaptations to the CART algorithm that are useful when the sampling pattern is very irregular, in particular in the presence of clusters. The simulation study and the analysis of two real data sets show that when the sample locations are not clustered standard CART can be used without restriction. In the presence of clusters, Standard CART (ID) shows systematically higher misclassification and methods taking into account only the spatial organization (VC, RG) or the spatial dependence (KM, EP) should be preferred. Adapted methods are aimed at reducing the bias of the classification tree by taking into account the spatial redundancy of the data. This implies that the equivalent number of data is reduced, hence that the variance of the classification parameters is increased. As expected, ID performs systematically worse than any method taking into account the spatial autocorrelation, when autocorrelation is present. EP leads in general to the lowest misclassification rate, but at the price of being a more time-consuming method. It has proven to be particularly useful when several classes are present in clusters. Weighted methods have in general intermediate misclassification rates. Among them, it must be understood that with VC and KM sample sites at the border of a cluster have systematically very high weights while those in the center have weights close to zero, 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.

As illustrated on the ecological example, in a data exploratory 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 to perform comparisons between observed pollen data and vegetation simulated ecosystems or BAGs. Here we have tested the methods for only one BAG (Mediterranean vegetation). To reach 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

Allard, D., 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.
Atteia, O., Dubois, J.-P., Webster, R., 1994. Geostatistical analysis of soil contamination in the Swiss Jura. Environmental Pollution 86, 315–327.
Beerling, D.J., Woodward, F.I., Lomas, M., 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.
Bel, L., Laurent, J.M., Bar-Hen, A., Allard, D., Cheddadi, R., 2005. A spatial extension of CART: Application to classification of ecological data. In: Renard, P., Demougeot-Renard, H., Froidevaux, R. (Eds.), Geostatistics for Environmental Applications. Springer, pp. 99–109.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth International Group, Belmont.
Chilès, J.P., Delfiner, P., 1999. Geostatistics: Modeling Spatial Uncertainty. Wiley, New-York.
De'ath, G., Fabricius, K.E., 2000. Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3198.
Diggle, P.J., 1983. Statistical Analysis of Spatial Point Patterns. Academic Press, New-York.
Dray, S., Pettorelli, N., Chessel, D., 2002. Matching data sets from two different spatial samplings. Journal of Vegetation Science 13, 867–874.
Gey, S., Poggi, J.M., 2006. Boosting and instability for regression trees. Computational Statistics and Data Analysis 50, 533–550.
Goovaerts, P., 1997. Geostatistics for Natural Resource Evaluation. Oxford University Press, New-York.
Hennig, C., Hausdorf, B., 2004. Distance-based parametric bootstrap tests for clustering of species ranges. Computational Statistics and Data Analysis 45, 875–896.
Intergovernmental Panel on Climate Change, 2001. In: Houghton, J.T., Ding, Y., Griggs, D.J., Noguer, M., van der Linden, P.J., Dai, X., Maskell, K., Johnson, C.A. (Eds.), Climate Change 2001: The Scientific Basis. Cambridge University Press, Cambridge.
Isaaks, E.H., Srivastava, R.M., 1989. Applied Geostatistics. Oxford University Press, New-York.
Laurent, J.M., Bar-Hen, A., François, L., Ghislain, M., Cheddadi, R., 2004. Refining vegetation simulation models: From plant functional types to bioclimatic affinity groups of plants. Journal of Vegetation Science 15, 739–746.
Okabe, A., Boots, B., Sugihara, K., Chiu, S.N., 2000. Spatial Tessellations: Concepts and Applications of Voronoi Diagrams, second edition. John Wiley & Sons, Ltd., Chichester.
Picard, N., Bar-Hen, A., 2000. Estimation of the envelope of a point set with loose boundaries. Applied Mathematics Letters 13, 13–18.
Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.
Wackernagel, H., 2003. Multivariate Geostatistics, 3rd edition. Springer-Verlag, Berlin.