SPODT: An R Package to Perform Spatial Partitioning

Jean Gaudart, Nathalie Graffeo, Guillaume Barbet, Stanilas Rebaudet, Nadine Dessay, Ogobara Doumbo, Roch Giorgi

Journal of Statistical Software

January 2015, Volume 63, Issue 16. http://www.jstatsoft.org/

SPODT: An R Package to Perform Spatial Partitioning

Jean Gaudart
Aix-Marseille University

Nathalie Graffeo
Aix-Marseille University

Drissa Coulibaly
MRTC, USTT Bamako

Guillaume Barbet
Aix-Marseille University

Stanilas Rebaudet
Aix-Marseille University

Nadine Dessay
IRD

Ogobara K. Doumbo
MRTC, USTT Bamako

Roch Giorgi
Aix-Marseille University

Abstract

Spatial cluster detection is a classical question in epidemiology: Are cases located near other cases? In order to classify a study area into zones of different risks and determine their boundaries, we have developed a spatial partitioning method based on oblique decision trees, which is called spatial oblique decision tree (SpODT). This non-parametric method is based on the classification and regression tree (CART) approach introduced by Leo Breiman. Applied to epidemiological spatial data, the algorithm recursively searches among the coordinates for a threshold or a boundary between zones, so that the risks estimated in these zones are as different as possible. While the CART algorithm leads to rectangular zones, providing perpendicular splits of longitudes and latitudes, the SpODT algorithm provides oblique splitting of the study area, which is more appropriate and accurate for spatial epidemiology. Oblique decision trees can be considered as non-parametric regression models. Beyond the basic function, we have developed a set of functions that enable extended analyses of spatial data, providing: inference, graphical representations, spatio-temporal analysis, adjustments on covariates, spatial weighted partition, and the gathering of similar adjacent final classes. In this paper, we propose a new R package, **SPODT**, which provides an extensible set of functions for partitioning spatial and spatio-temporal data. The implementation and extensions of the algorithm are described. Function usage examples are proposed, looking for clustering malaria episodes in Bandiagara, Mali, and samples showing three different cluster shapes.

Keywords: spatial, partitionning, malaria, oblique decision tree, R package.

1. Introduction

Spatial cluster detection is a classical question in epidemiology: are cases located near other cases? Among various approaches, general methods allow us to detect high risk zones of unspecified locations within a study area, without specifying any *a priori* point source (Colonna, Esteve, and Menegoz 1993; Elliott, Martuzzi, and Shaddick 1995; Wakefield, Quinn, and Rabb 2001; Waller and Gotway 2004; Chirpaz, Colonna, and Viel 2004; Gaudart, Ramatriravo, and Giusiano 2006b). Global detection methods, such as Moran's or Tango's ones (Tiefeldorf 2002; Tango 2002), test a statistic estimated over the entire study area, whereas local detection methods, such as Anselin's or Kulldorff's ones (Anselin 1995; Kulldorff 1997), test several statistics estimated over distinct zones within the study area. By scanning the study region with a circular or elliptic window, the SaTScan algorithm (Kulldorff 1997) compares observed and expected cases, inside and outside each potential cluster. It has the advantage of not depending on the underlying spatial architecture, although the choice of windowing is often critical and sensitive to edge effects (Gregorio, Samociuk, DeChello, and Swede 2006). These methods are also sensitive to geographical constraints, such as rivers, mountains, seas, or walls and corridors for outbreaks in buildings (e.g., healthcare-associated infections, or legionellosis).

We have introduced a spatial partitioning method based on oblique decision trees, called spatial oblique decision tree (SpODT), in order to classify a study area into zones of different risks and determine their boundaries, while being less sensitive to edge effects (Gaudart *et al.* 2006b). This non-parametric method is based on the classification and regression tree (CART) approach introduced by L. Breiman (Breiman, Friedman, Olshen, and Stone 1993). Beyond the basic function, we have developed a set of functions for an extended analysis of spatial data, providing: inference, graphical representations, spatio-temporal analysis, adjustments on quantitative covariates, spatial weighted partition, and the gathering of similar adjacent final classes.

In this paper, we propose a new package **SPODT** for R (R Core Team 2014) which provides an extensible set of functions for partitioning spatial and spatio-temporal data. The implementation and extensions of the algorithm are described and function usage are proposed based on a field observation datafile (malaria episodes in Mali) (Coulibaly *et al.* 2013) and samples showing three different cluster shapes. The results are compared to the CART approach using the **tree** package (Ripley 2014). All results were obtained using R 3.1.0 (Windows 7, Intel Core i7, CPU Q820 @1.73GHz, 64-bit). The **SPODT** package is freely available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=SPODT`.

2. Method

2.1. Basic algorithm

This non-parametric method is based on classification and regression tree (CART) (Breiman *et al.* 1993; Crichton, Hinde, and Marchini 1997; Gaudart, Poudiougou, Ranque, and Doumbo 2005). For each covariate, the CART algorithm searches for the threshold to split the covariate space into two classes, which optimizes a defined criteria (such as interclass variance). Then, the CART algorithm pursues recursively the binary partition of the covariate space, reaching

stopping rules. Applied to epidemiological spatial data, the CART algoritm searches among the planar coordinates $\{x_i, y_i\}$ (of each location M_i) for a threshold or boundary between two spatial classes (two geographic zones), so that the risks estimated in these two classes are as different as possible (maximum interclass variance or sum of squared errors SSE_{inter}). The algorithm then continues splitting recursively each of these two classes, and stops when reaching stopping rules. The root of the resulting regression tree is the entire study area. The final classes are sub-classes splitting the whole study area. Regression trees estimate changing lines of a constant function in each class of $\mathbb{R}^2$ (Gey 2002), interpreted as boundaries between zones (spatial classes) of different risks. However, the CART algorithm leads to rectangular classes (Murthy, Kasif, and Salzberg 1994; Cantu-Paz and Kamath 2003), providing perpendicular splits of the projected longitudes and latitudes. The SpODT (spatial oblique decision tree) algorithm (Gaudart *et al.* 2005) is a modification of the CART algorithm providing oblique splitting of the study area, which is more appropriate and accurate for spatial epidemiology. Oblique decision trees can be considered as non-parametric regression models. The functional form can be written as:

$$z_i = f(x_i, y_i) + \varepsilon_i,$$

where $\{x_i, y_i\}$ are the planar coordinates of each point location M_i, $i = 1...N$, and $\varepsilon_i \in \mathbb{R}$. These coordinates have to be euclidean coordinates in case of small area (e.g., hospital wards, rooms within buildings) or projections of geographical coordinates. Note that the use of non projected geographical coordinates may lead to erroneous results. The function $f(x_i, y_i)$ can be written as:

$$f(x_i, y_i) = \sum_{j=1}^{P} \bar{z}_j \, \mathbb{I}_{\{M_i(x_i,y_i) \in class_j\}}$$

where $class_j$, for $j = 1, ...P$, are the final P classes after splitting the whole study area; $\bar{z}_j = \frac{1}{N_j} \sum_{M_i \in class_j} z_i$ is the mean of observed values at N_j locations $M_i \in class\, j$. In other words, for each point location M_i belonging to a class j, the predicted risk will be $z_i = \bar{z}_j \pm \varepsilon_i$. The main problem is to determine the class set $\{class_j,\ j = 1, ..., P\}$. Boundaries between classes are linear functions $s_j(x_i, y_i)$ of the planar coordinates ($ax_i + by_i + c = 0$). These boundaries, or splitting directions, are recursively determined for each location sample, also called node ξ, corresponding to the whole study area at the beginning of the algorithm, or corresponding to a zone (geographical class) issued from a previous split. This node ξ is split into two classes by the partition direction $s_j(x_i, y_i)$. If $s_j(x_i, y_i) < 0$, then the location M_i will belong to the left "child" class jl of the tree. If not, the location M_i will belong to the right "child" class jr. For each node ξ constituted by a set of $n(\xi)$ locations, the algorithm searches, among the S set of every linear functions of (x_i, y_i), for the function $s_j(x_i, y_i)$ such as:

$$SSE_{inter}(s_j, \xi) = \max_{s \in S} \{SSE_{inter}\}.$$

We have shown (Gaudart *et al.* 2005; Fichet, Gaudart, and Giusiano 2006) that S, the set of every linear functions splitting a finite set of points in $\mathbb{R}^2$, is a finite set. There are an infinite number of lines splitting a set of points into two sub-sets. However, several lines lead to the same classification, splitting the point set identically. Therefore, the algorithm has to identify the possible lines to analyze only once each separate partition. For that purpose, the algorithm uses properties related to the order of abscissas of the points to be split, after rotation of the x-axis. Then, the algorithm performs vertical splitting of images of the x-axis

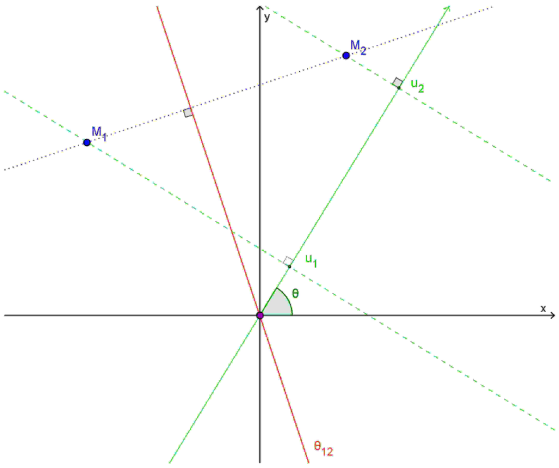

Figure 1: Determination of the critical angle θ_{12} and line (red) associated to pairs (M_1, M_2). In green: image of the x-axis before rotation and projections of point M_1 (u_1) and M_2 (u_2), before rotation.

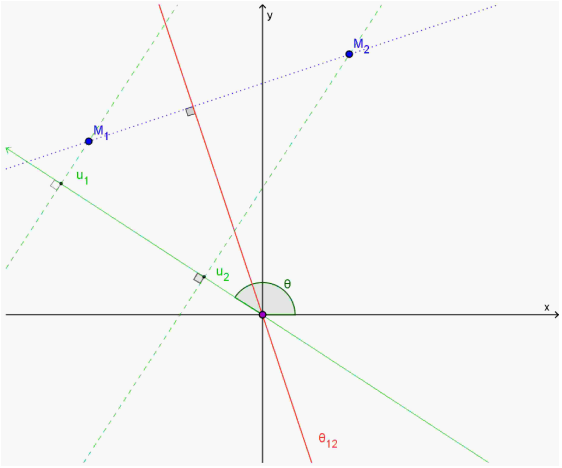

Figure 2: Determination of the critical angle θ_{12} and line (red) associated to pairs (M_1, M_2). In green: image of the x-axis after rotation and projections of point M_1 (u_1) and M_2 (u_2), after rotation.

for each rotation. To determine the angles of these rotations, critical angles associated to each pair of points are defined. They allow to define angular sectors within which the image of the axis preserves the order of the point abscissas. Indeed, during a rotation center O of the x-axis, the order of the point abscissas can be changed. For two points M_1 and M_2 , the critical angle θ_{12}, associated with the pair (M_1, M_2), defines the minimum angle of rotation to be applied to the x-axis so that points M_1 and M_2 have their abscissas u_1 and u_2 permuted (Figure 1 and Figure 2). During the passage of the x-axis image from an angular sector to the next, only the points associated to the critical angle, formed by the line delimiting the two angular sectors, have their abscissa order changed. The algorithm splits the plane perpendicularly to x-axis and x-axis images after rotations. Thus, permutations in the abscissa order scan the interval $[0, \pi[$, and characterize distinct splits that will be tested to maximize the interclass variance of the generated classes.

After splitting the initial set into 2 classes, the algorithm continues recursively. The number P of final classes (or zones) is recursively defined by the number of terminal nodes of the regression tree, after reaching stopping criteria.

A node ξ is a terminal node if one of the following criteria is reached:

1. $SSE_{inter}(s_j, \xi) \leq R_c^2 \times SSE_{total}(\xi) \iff R_\xi^2 \leq R_c^2$, i.e., a new partition will not explain enough variance; where R_ξ^2 is the explained variance calculated over the split of a node ξ and R_c^2 is the minimal explained variance (fixed by the user).

2. $n(\xi) \leq n_{c1}$, where $n(\xi)$ is the size of node ξ and n_{c1} is the fixed minimal size of a node below which the splitting algorithm is stopped (fixed by the user).

3. $(n(class_{jl}) \leq n_{c2}) \vee (n(class_{jr}) \leq n_{c2})$, where $class_{jl}$ and $class_{jr}$ are the two children classes issued from the split of node ξ. The fixed value n_{c2} is the minimal size of children classes below which the split is rejected (fixed by the user).

4. The maximal number of tree levels (fixed by the user).

Once the oblique regression tree is obtained (partition of the entire area into spatial classes of different risks), the main feature of this model is the overall variance explained in the dependent variable by the terminal classification, R^2_{global}, defined as the ratio of the sum of squared deviations between classes (calculated on the overall terminal classes) to the total sum of squares.

This approach, defined as a general method detecting spatial clusters, can be interpreted either as a global assessment of a spatial structure, or as a local analysis producing a map of the response variable.

2.2. Program developments

We have developed different R functions for a complete analysis of spatial data, according to our method. On the basis of the basic algorithm, several extensions have been developed:

- *Spatio-temporal analysis*: Integration of splits of a time covariate. The statistical unit is then defined by planar coordinates and a date. On an unique location different values can be observed at different dates. As CART algorithm, SpODT algorithm can thus provide a spatial splitting or a temporal splitting.

- *Adjustments*: Following the same procedure, the SpODT algorithm can provide a classification of different quantitative covariates. For these covariates, the standard CART algorithm is applied (i.e., no oblique split is performed).

- *Gathering similar adjacent final classes*: This option makes possible to gather similar adjacent classes at the end of the recursive splitting algorithm. Indeed, because of the recursiveness of the algorithm, the left branch of the tree ignores the right branch and conversely. This can lead to a final classification with similar adjacent classes, only separated because of the recursion. In this approach, the global R^2_{global} is calculated after grafting these two adjacent classes, and this grafted new classification is kept if this new global R^2_{global} is not sufficiently different from the previous one (without grafting classes).

- *Weighting the classification criterion*: In the basic SpODT algorithm, the calculation of the interclass variance doesn't take into account the child class sizes nor the spatial distribution of the locations within each child class. However, a class is all the less important in the analysis as its size is small and its locations are dispersed. We have then developed a weighted sum of squared error

$$SSE_{j\alpha} = \sum_{j=1}^{2} \alpha_j \, n(class_j) \, (\bar{z}_j - \bar{z})^2.$$

  The weight function α_j has to be a continuous non-decreasing bounded function of the size $n(class_j)$ (size of the class $j \in 1, 2$) and the spatial dispersion δ_j. The weight function actually proposed is

$$\alpha_j = \frac{\exp\{\frac{n(class_j)}{n(class_j)+\delta_j}\}}{1 + \exp\{\frac{n(class_j)}{n(class_j)+\delta_j}\}}$$

where $\delta_j = \det(\mathcal{V}_j)$ and $\mathcal{V}_j$ is the variance-covariance matrix for each class j.

- *Inference*: A "test" function has been developed in order to test the final SpODT classification using a Monte-Carlo approach. This test function simulates a specified number of data sets under a specified null hypothesis conditionally to the location, and the `spodt` function provides a classification tree for each of the simulated data set. The empirical distribution of the global R^2_{global} under the null hypothesis is obtained and, then, the test function provides a p value.

3. Overview of the functions

3.1. Basic function

The `spodt` function performs the classification of the data set.

```
spodt(z ~ 1, data, weight = FALSE, graft = 0, level.max = 5,
  min.parent = 10, min.child = 5, rtwo.min = 0.001)
```

Arguments:

- `z ~ 1`: A formula, using the `formula` function, with a response but no interaction terms. The left hand side has to contain a quantitative response variable (numeric). The right hand side should contain the quantitative and qualitative variables to be split according to a non oblique algorithm (e.g., `z ~ V1 + V2`). For single spatial analysis (with no cofactor) the right hand side should be `z ~ 1`.
- `data`: A `SpatialPointsDataFrame` containing the coordinates and the variables. SpODT functions need planar coordinates. Geographic coordinates have to be projected. Otherwise, euclidian coordinates can be used (for small area analysis such as rooms within buildings).
- `weight`: A logical value indicating whether the interclass variances should be weighted or not.
- `graft`: A numerical value between 0 and 1 indicating the minimal modification of R^2_{global} required to graft the final classes. If `graft = 0` the algorithm will not graft any adjacent classes.
- `level.max`: The maximal level of the regression tree above which the splitting algorithm is stopped.
- `min.parent`: The minimal size of a node below which the splitting algorithm is stopped (n_{c1}).
- `min.child`: The minimal size of the children classes below which the split is refused and the algorithm is stopped (n_{c2}).
- `rtwo.min`: R^2_c, the minimal value of R^2_ξ above which the node split is refused and the algorithm is stopped. Specified as a numerical value between 0 and 1.

Value: The `spodt` function computes an object of class `spodt` with the different components of the classification tree, i.e., *i*) at each step: the point locations within each class, R^2_ξ, coefficients of the splitting line; ii) global results: the global R^2_{global} (`object@R2`), the final partition (`object@partition`) including the graft results.

3.2. Tree and spatial lines

```
spodt.tree(object)
```

This graphical function provides the tree issued from the `spodt` function. Each step of the classification is presented with main statistics. `object` is an object of class `spodt`, usually a result of a call to `spodt`. For graphical convenience, grafted classes are not presented but only indicated by their `id` number.

```
spodtSpatialLines(object, data)
```

This function provides the `SpatialLines` object (see the package **sp**, Bivand, Pebesma, and Gómez-Rubio 2013) that contains the boundaries of the spatial classification issued from the `spodt` function. `object` is an object of class `spodt`, usually a result of a call to `spodt`. `data` is the initial `SpatialPointsDataFrame` containing the planar coordinates and the variables. The `SpatialLines` object obtained can be used, for example to obtain maps.

3.3. Hypothesis testing

The `test.spodt` function provides a Monte Carlo hypothesis test of the final classification issued from the `spodt` function. This function performs simulations of the specified null hypothesis and the classification of each simulated data set, using the same rules as the observed data set classification.

```
test.spodt(z ~ 1, data, obs.R2, rdist, par.rdist, nb.sim, weight, graft,
  level.max, min.parent, min.child, rtwo.min)
```

Arguments:

- `z ~ 1`: A formula, such as in the `spodt` function, with a response but no interaction terms. The left hand side has to contain a quantitative response variable (numeric). The right hand side should contain the quantitative and qualitative variables to be split according to a non oblique algorithm (e.g., `z ~ V1 + V2`). For single spatial analysis (with no cofactor) the right hand side should be `z ~ 1`.
- `data`: A `SpatialPointsDataFrame` containing the coordinates and the variables. SpODT functions need planar coordinates. Geographic coordinates have to be projected. Otherwise, euclidian coordinates can be used (for small area analysis such as rooms within buildings).
- `obs.R2`: The global R^2_{global} issued from the previous `spodt` final classification of the observed data set. Specified as a numerical value between 0 and 1.

Graft number	id $class_1$	id $class_2$	id $class_{12}$
1	55	105	111
2	104	111	113
3	12	113	115
4	108	115	117
5	53	117	119
6	7	119	121

Table 1: Grafting classes, malaria episodes ($class_{12} = class_1 \cup class_2$).

- `rdist`: A description of the distribution of the dependent variable under the null hypothesis. This can be a character string naming a random generation of a specified distribution, such as `"rnorm"` (Gaussian distribution), `"rpois"` (Poisson distribution), `"rbinom"` (binomial distribution), `"runif"` (uniform distribution), ...
- `par.rdist`: A list of the parameters needed for the random generation, depending on the null hypothesis distribution, such as `c(n, mean, sd)` (Gaussian distribution), `c(n, lambda)` (Poisson distribution), `c(n, size, prob)` (binomial distribution), `c(n, min, max)` (uniform distribution), ...
- `nb.sim`: The number of simulations, specified as a positive integer.
- `weight`, `graft`, `level.max`, `min.parent`, `min.child`, `rtwo.min`: These arguments have to be specified, similarly to the previous `spodt` classification of the observed data set.

Value: The `test.spodt` function computes classification trees for the simulated data sets. It provides the global $R^2_{global/H0}$ empirical distribution under the null hypothesis, compared to the observed global R^2_{global}, and a p value.

4. Data examples

4.1. Clustering malaria episodes (Bandiagara, Mali)

Malaria parasite transmission and clinical disease are characterized by important microgeographic variations, often between adjacent villages, households or families (Greenwood 1989; Carter, Mendis, and Roberts 2000; Gaudart, Poudiougou, Dicko, Ranque, Sagara, Diallo, Diawara, Ouattara, Diakite, and Doumbo 2006a). This local heterogeneity is driven by a variety of factors including distance to breeding sites, housing constructions and socio-behavioral characteristics (Koram, Bennett, Adiamah, and Greenwood 1995; Coleman, Mabuza, Kok, Coetzee, and Durrheim 2009; Ernst *et al.* 2009). The study was conducted in Bandiagara, Mali, following a cohort of 300 children, at 168 locations. The household of each child was geo-located (decimal degrees). Approval from Institutional review boards at the Faculty of Medicine, Pharmacy and Dentistry of the University of Mali, community approval and written informed consents from parents were obtained before inclusion (see Coulibaly *et al.* 2013, for further details). We applied **SPODT** functions to classify the entire area into different risk zones with homogeneous number of malaria episodes per child at each household, from

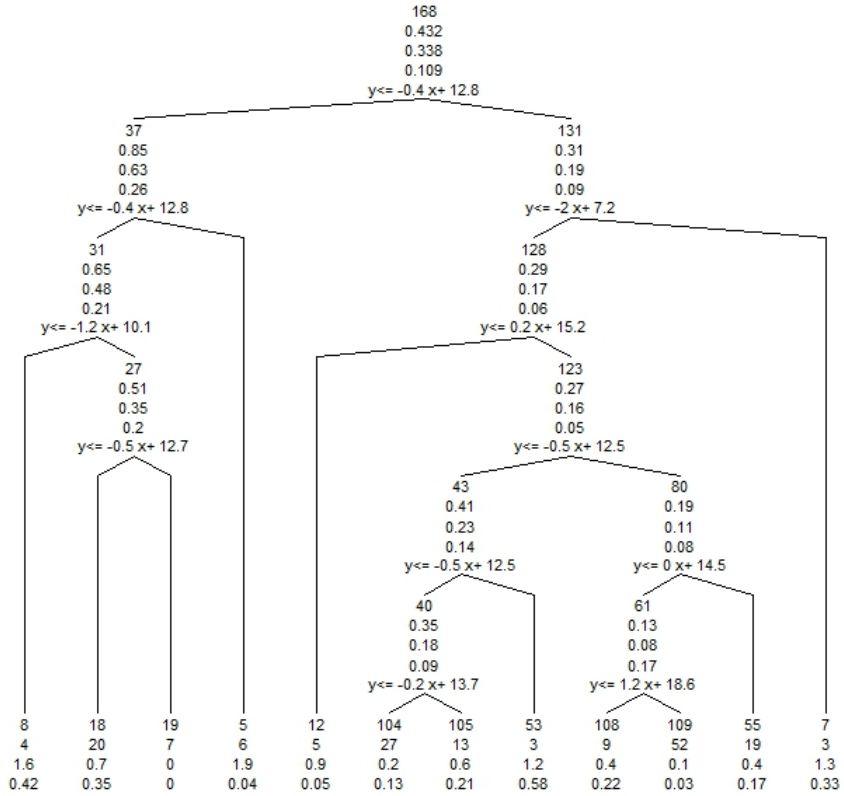

Figure 3: Classification tree (`spodt.tree(object)`) of malaria episodes in Bandiagara, Mali. This classification was obtained by using the **SPODT** package. Each node (excluding terminal nodes) is presented with its `id` number, mean, variance and local R^2_ξ after splitting, as well as the function of the splitting line. Each terminal node is presented with its `id` number, number of locations, mean and variance.

November to December 2009. We used the `spodt` function to provide a spatial classification of the study site, with no covariates, with a weighted classification criterion, 7 tree levels, a minimal parent size of $n_{c1} = 25$, a minimal child size of $n_{c2} = 2$, and with a minimal $R^2_c = 0.01$. We also used the grafting option (minimal R^2_{global} improvement of 0.13). After projection, the function can be written as follows, and results were obtained in 0.53 seconds:

```
R> data("dataMALARIA")
R> coordinates(dataMALARIA) <- c("x", "y")
R> proj4string(dataMALARIA) <- "+proj=longlat +datum=WGS84 +ellps=WGS84"
R> dataMALARIA <- spTransform(dataMALARIA,
+    CRS("+proj=merc +datum=WGS84 +ellps=WGS84"))
R> spodt.results <- spodt(z ~ 1, data = dataMALARIA, graft = 0.13,
+    level.max = 7, min.parent = 25, min.child = 2, rtwo.min = 0.01)
```

The tree (Figure 3) and the map (Figure 4) were obtained by the following R codes:

```
R> spodt.tree(spodt.results)
R> SSL.result <- spodtSpatialLines(spodt.results, dataMALARIA)
R> plot(SSL.result)
R> points(dataMALARIA, cex = log(dataMALARIA@data$z * 10))
```

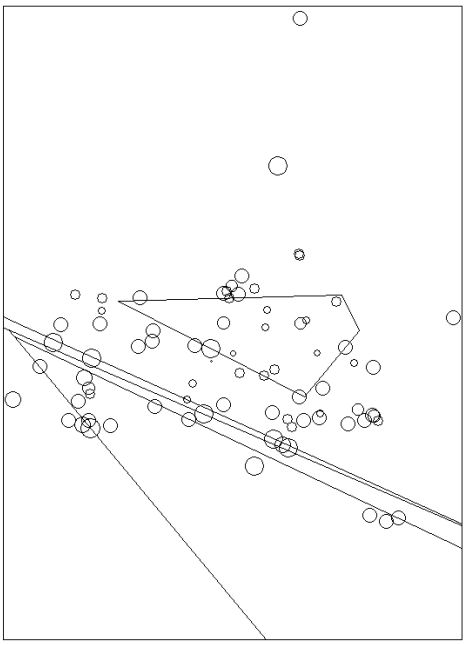

Figure 4: Mapping of the SpODT classification (`spodtSpatialLines(object, data)`). Each location (circles) is presented according to its projected coordinates.The lines represent the borders of each class. The circle size represents the mean number of malaria episodes at each location.

The non-grafted tree (Figure 3) showed 12 final classes with different risks before grafting (Figure 6). Adjacent classes were grafted according to the graft criteria described in Table 1, which finally provides 6 classes, with $R^2_{global} = 0.49$ (given by `spodt.results@R2`). This result shows that spatial variations can explain an important part of the malaria risk variability, although other factors remain such as behaviors, genetic, personal medical history, household characteristics etc. The spatial classification (Figure 4) highlighted a central low risk cluster (class id 109) with a mean malaria episode of 0.08 per child (95% confidence interval, CI[0.04-0.11]) (Table 2), with a polygonal and asymmetric shape. Around this low risk cluster, the mean malaria episodes per child was higher (0.47 [0.39-0.55]). Note that there is a pond in the north of the city and a river in the south, which are breeding sites for malaria transmission mosquitoes (Coulibaly *et al.* 2013). The remaining zone showed an alternation of high and low risk clusters.

The test of the tree algorithm was performed using 99 simulated samples following a Poisson distribution and with the same criteria as previously, such as follows (results were obtained in 28.46 seconds):

```
R> test.spodt(z ~ 1, data = dataMALARIA, spodt.results@R2, "rpois",
+    c(length(dataMALARIA@data$loc), mean(dataMALARIA@data$z)), 99,
+      weight = TRUE, graft = 0.13, level.max = 7, min.parent = 25,
+      min.child = 2, rtwo.min = 0.01)
```

With a p value of 0.01, the classification obtained by the `spodt` function was significantly different from a homogeneous spatial distribution of malaria episodes (Figure 5).

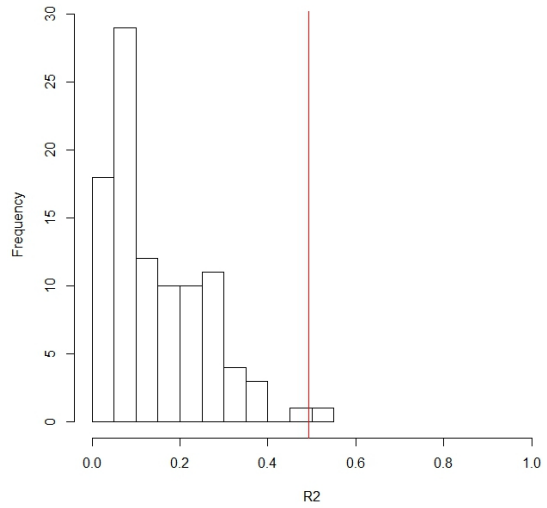

Figure 5: Testing of the classification (`test.spodt(object)`). The histogram of the $R^2_{global/H0}$ obtained after 99 simulations, is presented together with the observed R^2_{global} (red line).

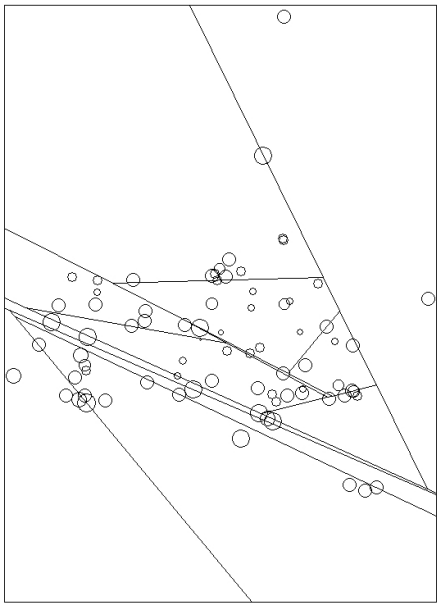

Figure 6: Mapping of the SpODT classification (`spodtSpatialLines(object, data)`) with no `graft` option.

Among the different tuning parameters of the `spodt` function, `level.max`, `min.parent`, `min.child` and `rtwo.min` are similar to those of the **tree** package, and have to be chosen similarly to CART approaches. In the **SPODT** package, as we have introduced a gathering option, a `graft` tuning parameter has been added. In order to assess the sensitivity of the SpODT algorithm to this option, we ran it with different values of `graft` ranging from 0.0 to 1 (with a step of 0.001), the other tuning parameters being fixed as previously. We also assessed the sensitivity of the SpODT algorithm to `rtwo.min` values, running the algorithm

Final class `id`	Location count	Mean [95% confidence intervals]
5	6	1.92 [1.77–2.07]
8	4	1.58 [1.14–2.03]
18	20	0.69 [0.49–0.89]
19	7	0
109	52	0.08 [0.04–0.11]
121	79	0.47 [0.39–0.55]
Global	168	0.43 [0.37–0.50]

Table 2: Mean malaria episodes per child, SpODT classification.

`graft`	R^2_{global}	Number of classes
0.000–0.047	0.601	12
0.047–0.062	0.596	11
0.062–0.103	0.585	10
0.103–0.105	0.565	8
0.105–0.123	0.535	7
0.123–0.154	0.494	6
0.154–0.190	0.489	5
0.190–0.270	0.403	4
0.270–0.426	0.359	3
0.426–1.000	0.003	2

Table 3: Tuning parameter of the `spodt` function: the `graft` option. `level.max = 7`; `min.parent = 25`; `min.child = 7`; `rtwo.min = 0.01`.

`rtwo.min`	R^2_{global}	Number of classes
0.000–0.068	0.494	6
0.068–0.149	0.408	5
0.149–1.000	0.000	1

Table 4: Tuning parameter of the `spodt` function: the `rtwo.min` option. `level.max = 7`; `min.parent = 25`; `min.child = 7`; `graft = 0.13`.

with values ranging from 0.0 to 1 (with a step of 0.001), the other tuning parameters being fixed as previously (`graft` = 0.13).

The R^2_{global} obtained ranged from 0.6 (12 final classes) to 0.003 (2 final classes), showing a step decrease of the number of classes (Table 3) when `graft` increased. When `rtwo.min` increased, the algorithm stopped rapidly with no classification (Table 4). Choice of the tuning parameters has thus to be made between no classes and too many classes, such as for CART approaches. From a practical point of view, together with field knowledge, the number of final classes, the R^2_{global} and the test procedure provided by this package can guide the user in this choice. Note that the choice of a deep tree will be corrected by the `graft` parameter.

The results were compared to the CART approach, using the **tree** package, tuning parameters being set as follows: `mincut = 5`, `minsize = 10`, `mindev = 0.01`. The CART approach showed a less accurate classification with 16 final classes (Table 5 and Figure 7). A central low risk cluster was also detected (class id 27) as well as the alternation of high and low risk

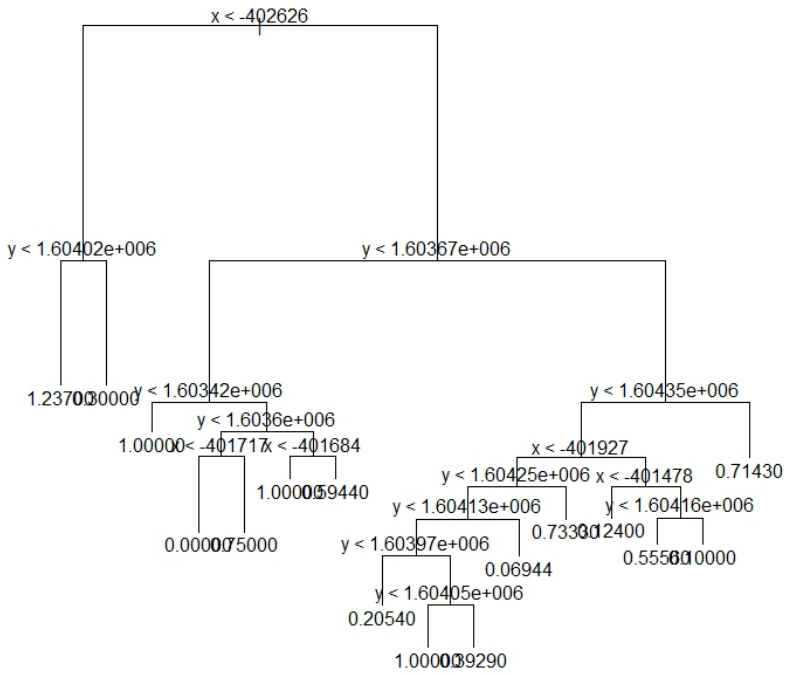

Figure 7: Classification tree (`plot.tree(object)`) of malaria episodes in Bandiagara, Mali. This classification was obtained by using the **tree** package.

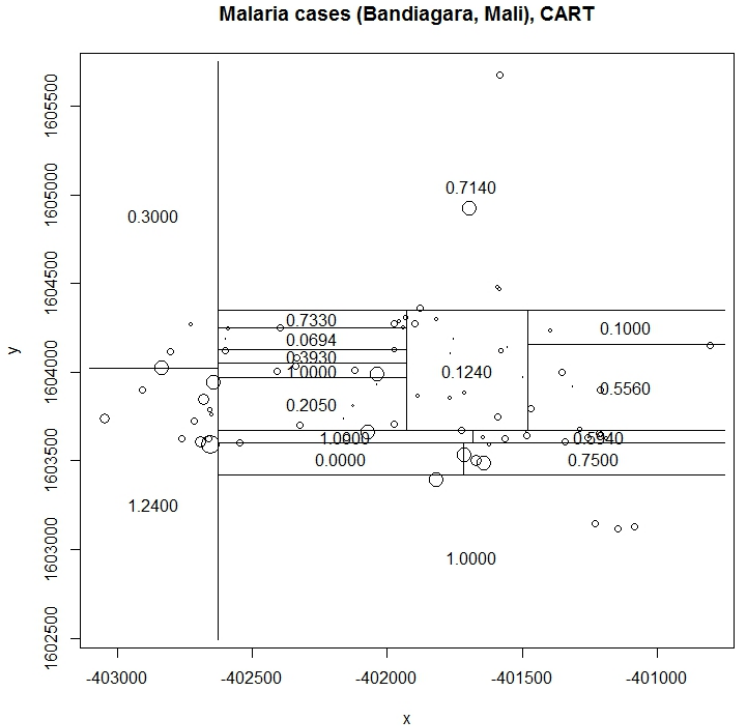

Figure 8: Mapping of the CART classification (`partition.tree(object)`). Each location is presented according to its coordinates and the circle size represents the mean number of malaria episodes. The lines represent the borders of each class.

Final class `id`	Location count	Mean	95% confidence intervals
3	13	1.24	0.87–1.60
4	5	0.30	0.00–0.69
7	5	1.00	0.38–1.62
10	6	0.00	
11	8	0.75	0.11–1.39
13	5	1.00	0.38–1.62
14	12	0.59	0.35–0.84
20	14	0.21	0.02–0.40
22	5	1.00	0.38–1.62
23	7	0.39	0.03–0.76
24	12	0.07	0.00–0.16
25	5	0.73	0.51–0.96
27	50	0.12	0.05–0.19
29	9	0.56	0.25–0.86
30	5	0.10	0.00–0.30
31	7	0.71	0.20–1.23
Global	168	0.43	0.37–0.50

Table 5: Mean malaria episodes per child, CART classification.

clusters in the South, but this approach failed to detect the polygonal shape and to gather similar adjacent classes (Figure 8). From an epidemiological point of view, numerous small classes is not very useful in this context. Note that changes in the tuning parameters did not change the global interpretation of the results. In the case of a greater `mindev` value (e.g., $>$ 0.0134), the central low risk cluster was not detected (data not shown).

4.2. Different cluster shapes and levels

We assessed the **SPODT** functions analyzing three different situations, and in comparison to the CART algorithm (**tree** package). The following situations have been studied:

- Clustered data with a high level within a centered rotated square, and a low level outside.
- Clustered data with a low level inside a centered ball shape, and a high level outside.
- Clustered data with a high level under a "V" shape border, and a low level above.

For each situation, samples ($n = 300$) were provided:

- Planar coordinates following a uniform distribution $(1, -1)$.
- A dependent variable following a Gaussian distribution with a constant variance (0.09) and a constant mean for the two level zones: $\mu_1 = 1$ for the low level zone, $\mu_2 = 1 + \beta$ for the high level zone. For each situation, we used four samples: $\beta = 0$ (no cluster), $\beta = 0.5$, $\beta = 1.5$ and $\beta = 2$.

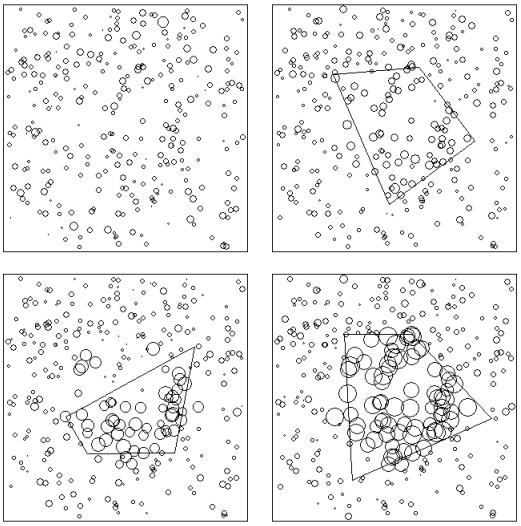

Figure 9: Rotated square situation: Mapping of the SpODT classification.

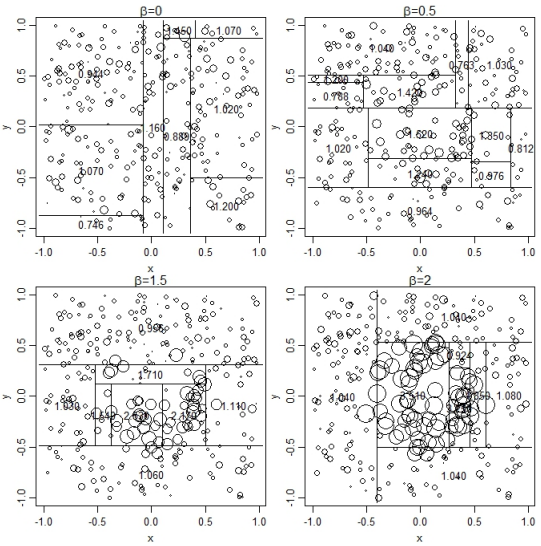

Figure 10: Rotated square situation: Mapping of the CART classification.

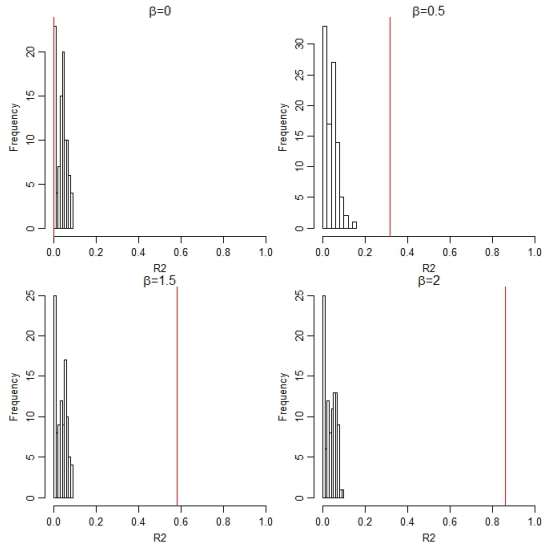

Figure 11: Rotated square situation: Testing of the SpODT classification.

As planar coordinates were used, no projection were applyed to the `SpatialPointsDataFrame`. This provides a warning message when using `spodt` and `test.spodt` functions.

For both SpODT and CART approaches, default tuning parameters were used, except for `graft` = 0.2 (SpODT algorithm). Changing these parameters did not greatly change the interpretation of the comparisons.

Whatever the shape was, the SpODT algorithm did not show any significant cluster for $\beta = 0$ (Figures 9, 12, 15, $\beta = 0$ panels). *A contrario*, the CART algorithm split the spatial area even with no cluster (Figures 10, 13, 16, $\beta = 0$ panels). According to the `spodt` test procedure,

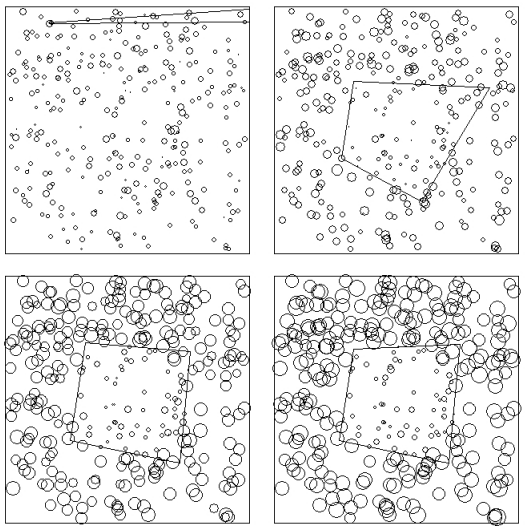

Figure 12: Ball shape situation: Mapping of the SpODT classification.

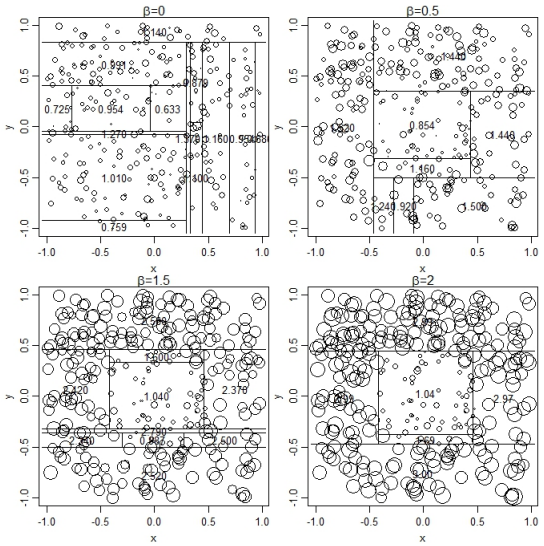

Figure 13: Ball shape situation: Mapping of the CART classification.

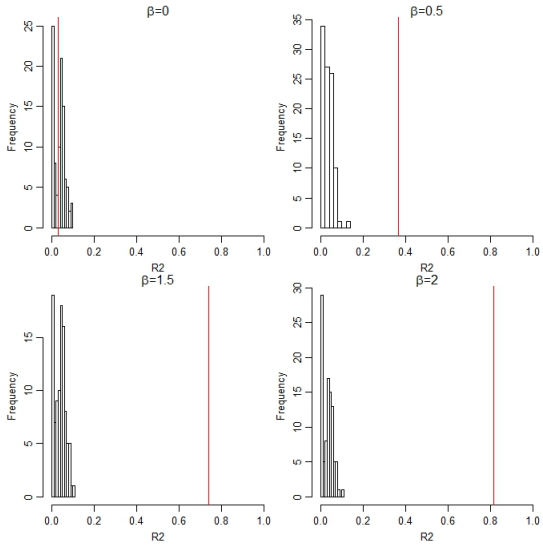

Figure 14: Ball shape situation: Testing of the SpODT classification.

SpODT classes showed no significant classification for $\beta = 0$, and then progressively significant results as β increased (Figures 11, 14, 17).

- *Rotated square shape situation*: The SpODT algorithm did show the central cluster even for low values in the high level cluster (Figure 9, $\beta = 0.5$, $\beta = 1.5$ and $\beta = 2$ panels). But the obtained shape was only approximatively a rotated square. *A contrario*, the shape obtained with the CART algorithm was accurate only for higher values ($\beta = 2$), but showed no rotated square (Figure 10).

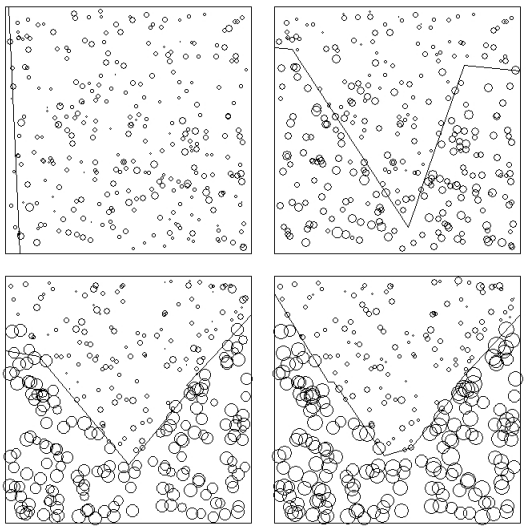

Figure 15: "V" shape border situation: Mapping of the SpODT classification.

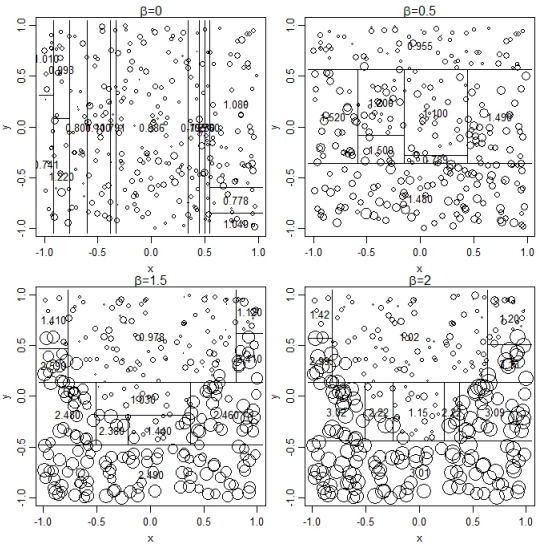

Figure 16: "V" shape border situation: Mapping of the CART classification.

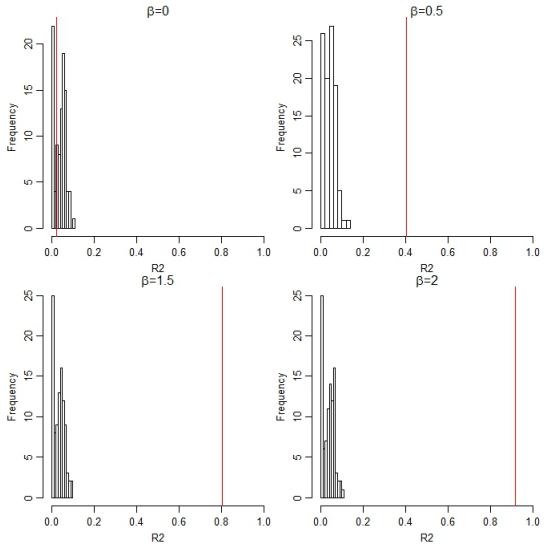

Figure 17: "V" shape border situation: Testing of the SpODT classification.

- *Ball shape situation*: The SpODT and the CART algorithms failed to precisely detect this particular form, but precisely located square clusters (Figures 12, 13, $\beta = 0.5$, $\beta = 1.5$ and $\beta = 2$ panels). Again, CART failed to detect only two levels: it detected few classes in the high level zone only for $\beta = 2$.

- *"V" shape border*: The SpODT algorithm detected a very accurate border even for low values in the high level zone ($\beta = 0.5$). The CART algorithm failed to detect such a particular shape. Nevertheless, it showed lower values in the north, higher values in the south, and a mitigate central band (with numerous different classes).

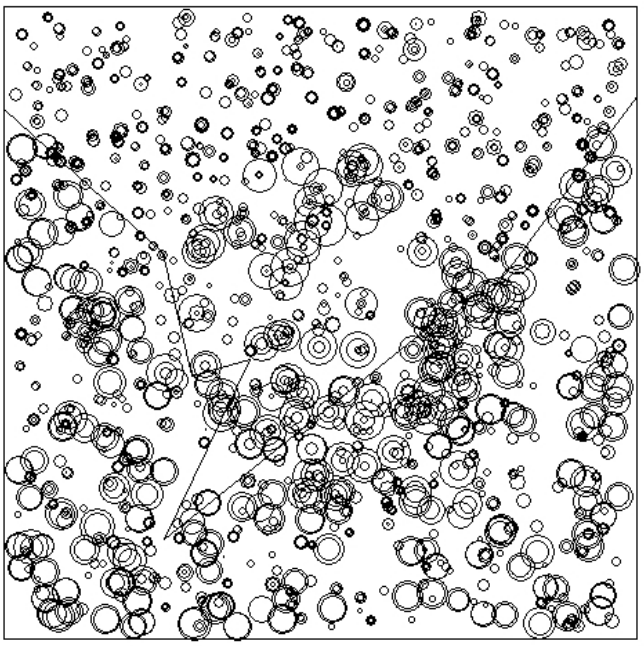

Figure 18: Space-time situation: Mapping of the grafted-SpODT classification.

4.3. Spatial partition with a time covariate

A sample was build that concatenates 6 different situations: 2 rotated square situations ($\beta = 2$ and $\beta = 1.5$), 2 "no cluster" situations ($\beta = 0$), and two "V" shape situations ($\beta = 2$ and $\beta = 1.5$), which thus form a numeric time covariate (1 unit of time up to 6). The `spodt` function was used to provide a classification of the area, including this time covariate, with a weighted classification criteria, a maximum of 5 tree levels, a minimal parent size of $n_{c1} = 10$, a minimal child size of $n_{c2} = 5$, a minimal $R_c^2 = 0.001$, and a grafting option of $graft = 0.2$. The function can be written as follows:

```
R> data("dataCOV")
R> coordinates(dataCOV) <- c("x", "y")
R> spodt.results.cov <- spodt(z ~ V1, data = dataCOV, weight = TRUE,
+    graft = 0.2, level.max = 5, min.parent = 10, min.child = 5,
+    rtwo.min = 0.001)
```

The non-grafted tree (Figure 20), provided by the SpODT algorithm, showed 16 final classes, with 2 time splits: less than 2 and less than 5. These 3 time periods was related to 3 situations: rotated square with high values ($\beta = 2$), "no cluster" or rotated square with medium values($\beta = 1.5$), and "V" shape situation ($\beta = 1.5$ and $\beta = 2$). The graft option led to two main classes (Figure 18), a high level zone in the South and a low level zone in the north, which highlight the impact of the "V" shape situation in this exemple (more locations showing high values in this part of the area at this period). The CART algorithm provided a similar tree with the same time splits (Figure 19), but 15 different spatial classes.

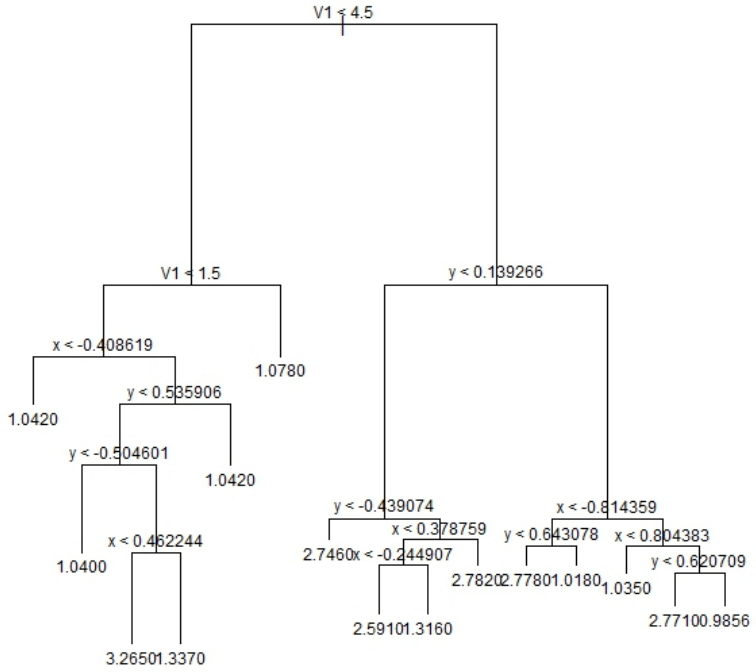

Figure 19: Space-time situation: CART classification tree.

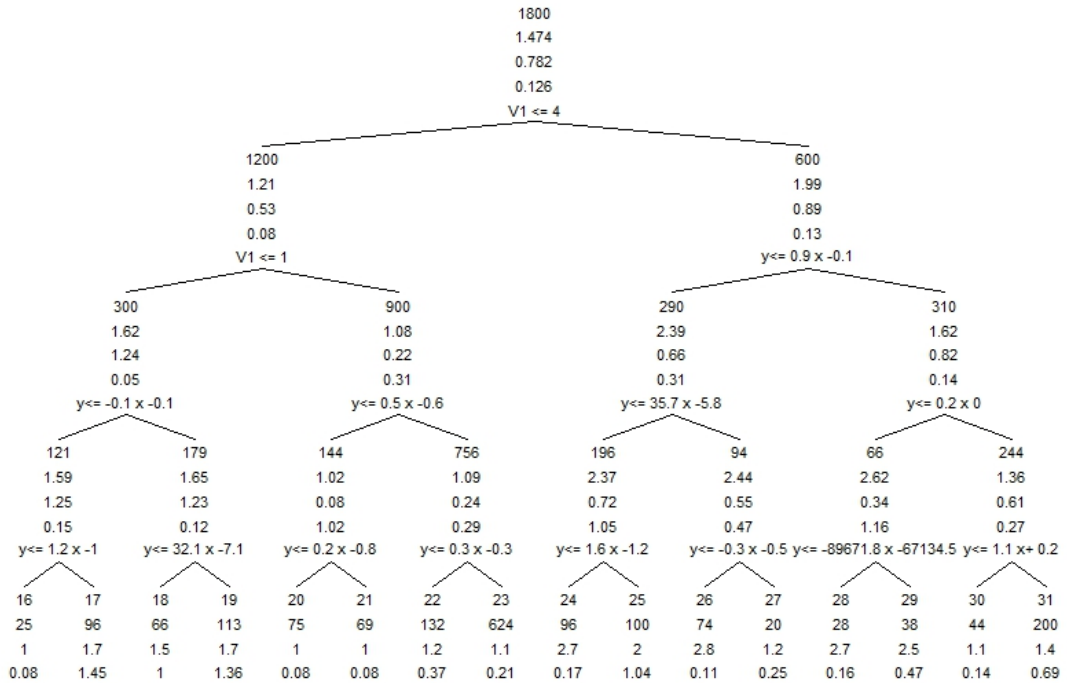

Figure 20: Space-time situation: SpODT classification tree.

5. Conclusion

Among the different tools used dedicated to spatial classification (e.g., Assuncao, Neves, Camara, and da Costa Freitas 2006; Oden, Sokal, Fortin, and Goebl 1993), the proposed **SPODT** package provides a classification of a spatial area based on the spatial variability of a dependant variable. Space splitting can be oblique and this classification can be adjusted on covariates and gather similar adjacent classes. Associated functions (`spodt.tree` and `spodtSpatialLines`) are useful for graphical representations of the classification, and the `spodt.test` function provides a test of the oblique decision tree algorithm. **SPODT** package is provided with a real example set of malaria cases observed in Mali. Using this set and others, SpODT detected spatial and spatio-temporal clusters more accurately than the CART algorithm in all performed comparisons.

Acknowledgments

The authors thank Dr. Bernard Fichet for many valuable discussions, the reviewers and the editor for their helpful comments. This work was supported by the AMMA consortium (African Monsoon Multidisciplinary Analysis). Dr. Jean Gaudart was also supported by the ADEREM association for biological and medical research development (Association pour le Developpement des Recherches biologiques et Medicales). The Malaria incidence field study (Bandiagara, Mali) was coordinated by the Malaria Research and Training Center (MRTC, Bamako, Mali), supported by cooperative agreement 5U01AI065683 from the National Institute of Allergy and Infectious Diseases and the grant D43TW001589 from the Fogarty International Center, National Institutes of Health.

References

Anselin L (1995). "Local Indicators of Spatial Association: LISA." *Geographical Analysis*, **27**, 93–116.

Assuncao RM, Neves MC, Camara G, da Costa Freitas C (2006). "Efficient Regionalization Techniques for Socio-Economic Geographical Units Using Minimum Spanning Trees." *International Journal of Geographical Information Science*, **20**(7), 797–811.

Bivand RS, Pebesma E, Gómez-Rubio V (2013). *Applied Spatial Data Analysis with R*. 2nd edition. Springer-Verlag, New York.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1993). *Classification and Regression Trees*. Chapman and Hall.

Cantu-Paz E, Kamath C (2003). "Inducing Oblique Decision Trees with Evolutionary Algorithms." *IEEE Transactions on Evolutionary Computing*, **7**, 54–68.

Carter R, Mendis KN, Roberts D (2000). "Spatial Targeting of Interventions Against Malaria." *Bulletin of the World Health Organisation*, **78**, 1401.

Chirpaz E, Colonna M, Viel JF (2004). "Cluster Analysis in Geographical Epidemiology : The Use of Several Statistical Methods and Comparison of Their Results." *Revue de Epidemiologie et de Sante Publique*, **52**, 139–149.

Coleman M, Mabuza AM, Kok G, Coetzee M, Durrheim DN (2009). "Using the SaTScan Method to Detect Local Malaria Clusters for Guiding Malaria Control Programs." *Malaria Journal*, **8**, 68.

Colonna M, Esteve J, Menegoz F (1993). "Detection of Spatial Autocorrelation in Cancer Hazard with Hetergogeneous Population." *Revue de Epidemiologie et de Sante Publique*, **41**, 235–240.

Coulibaly D, Travassos MA, Rebaudet S, Laurens MB, Tolo Y, Kone AK, Traore K, Guindo AB, Diarra I, Niangaly A, Daou M, Dembele A, Cissoko M, Kouriba B, Dessay N, Gaudart J, Thera MA, Piarroux R, Plowe CV, Doumbo OK (2013). "Spatial and Temporal Patterns of Malaria Incidence in Bandiagara, Mali." *Malaria Journal*, **12**, 82.

Crichton NJ, Hinde JP, Marchini J (1997). "Models for Diagnosing Chest Pain: Is CART Helpful?" *Statistics in Medicine*, **16**, 717–727.

Elliott P, Martuzzi M, Shaddick G (1995). "Spatial Statistical Methods in Environmental Epidemiology: A Critique." *Statistical Methods in Medical Research*, **4**, 13759.

Ernst KC, Lindblade KA, Koech D, Sumba PO, Kuwuor DO, John CC, Wilson ML (2009). "Environmental, Socio-Demographic and Behavioural Determinants of Malaria Risk in the Western Kenyan Highlands: A Case-Control Study." *Tropical Medicine & International Health*, **14**, 1258–1265.

Fichet B, Gaudart J, Giusiano B (2006). "Bivariate CART with Oblique Regression Trees." In *International Conference of Data Science and Classification*. International Federation of Classification Societies, Ljubljana, Slovenia.

Gaudart J, Poudiougou B, Dicko A, Ranque S, Sagara I, Diallo M, Diawara S, Ouattara A, Diakite M, Doumbo OK (2006a). "Space-Time Clustering of Childhood Malaria at the Household Level: A Dynamic Cohort." *BMC Public Health*, **6**, 286.

Gaudart J, Poudiougou B, Ranque S, Doumbo OK (2005). "Oblique Decision Trees for Spatial Pattern Detection: Optimal Algorithm and Application to Malaria Risk." *BMC Medical Research Methodology*, **5**, 22.

Gaudart J, Ramatriravo NO, Giusiano B (2006b). "Spatial Pattern Detection: Power Evaluation of Scan Methods and Regression Trees." *Revue du Epidemiologie et de Sante Publique*, **54**(HS2), 31.

Gey S (2002). *Bornes de Risque, Detection de Ruptures Boosting: Trois Themes Statistiques Autour de CART en Regression*. Ph.D. thesis, University of Paris XI, Paris, France.

Greenwood BM (1989). "The Microepidemiology of Malaria and Its Importance to Malaria Control." *Transactions of the Royal Society of Tropical Medicine & Hygiene*, **83**, 25–29.

Gregorio DI, Samociuk H, DeChello L, Swede H (2006). "Effects of Study Area Size on Geographic Characterizations of Health Events: Prostate Cancer Incidence in Southern New England, USA, 1994–1998." *International Journal of Health Geography*, **5**, 8.

Koram KA, Bennett S, Adiamah JH, Greenwood BM (1995). "Socio-Economic Risk Factors for Malaria in a Peri-Urban Area of The Gambia." *Transactions of the Royal Society of Tropical Medicine & Hygiene*, **89**, 146–150.

Kulldorff M (1997). "A Spatial Scan Statistic." *Communications in Statistics: Theory and Methods*, **26**, 1481–1496.

Murthy SK, Kasif M, Salzberg S (1994). "A System for Induction of Oblique Decision Trees." *Journal of Artifical Intelligence Research*, **2**(1-32).

Oden N, Sokal R, Fortin M, Goebl H (1993). "Categorical Wombling: Detecting Regions of Significant Change in Spatially Located Categorical Variables." *Geographical Analysis*, **25**(4), 315–336.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ripley B (2014). ***tree**: Classification and Regression Trees*. R package version 1.0-35, URL http://CRAN.R-project.org/package=tree.

Tango T (2002). "Score Tests for Detecting Excess Risks around Putative Sources." *Statistical Medicine*, **21**, 497–514.

Tiefeldorf M (2002). "The Saddlepoint Approximation of Moran's I and Mocal Moran's I's Reference Distribution and Their Numerical Evaluation." *Geographical Analysis*, **34**, 187–206.

Wakefield J, Quinn M, Rabb G (2001). "Disease Clusters and Ecological Studies." *Journal of the Royal Statistical Society A*, **164**, 1–2.

Waller LA, Gotway CA (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley & Sons, Hoboken.

Affiliation:

Jean Gaudart
Aix-Marseille University
UMR912 SESSTIM (INSERM IRD AMU)
Faculty of Medicine

27 Bd Jean Moulin
13005 Marseille, France
E-mail: jean.gaudart@univ-amu.fr
URL: http://www.sesstim-orspaca.org/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 63, Issue 16 *Submitted:* 2013-06-02
January 2015 *Accepted:* 2014-10-14