Journal of Statistical Software

January 2015, Volume 63, Issue 6. *http://www.jstatsoft.org/*

ads Package for R: A Fast Unbiased Implementation of the K-function Family for Studying Spatial Point Patterns in Irregular-Shaped Sampling Windows

Raphaël Pélissier
UMR AMAP

François Goreaud
IRSTEA

Abstract

ads is an R package that performs multi-scale spatial point pattern analyses through methods derived from Ripley's K-function. These methods apply to univariate, multivariate or marked point data mapped in a rectangular, circular or irregular-shaped sampling window. Specific tests of statistical significance based on Monte Carlo simulations are associated to these methods. The main features of **ads** is to call fast C subroutines for computing Ripley's unbiased local correction of edge effects for various sampling window configurations and for performing Monte Carlo simulations. It thus allows one to analyze large datasets and to compute robust confidence envelopes. This paper is an introduction to **ads** version 1.5, focusing on its complementarity with the other R packages for spatial point pattern analysis, and on recent original developments towards the introduction of multivariate functions for analyzing spatial pattern of species diversity.

Keywords: bivariate function, intertype function, isotropic edge effects correction, local-density function, mark-correlation function, multivariate function, pair-correlation function, Rao's diversity, second-order neighborhood analysis, spatial point pattern analysis, Shimatani's alpha, Shimatani's beta, Simpson's diversity.

1. Introduction

ads is an R package for spatial analysis of mapped data based on second-order multi-scale analyses of spatial point patterns derived from Ripley's (1977) K-function (the K-function family). Initially implemented for forest ecologists who often need to analyze large naturally heterogeneous tree maps, it is based on C subroutines that include geometrical functions to perform unbiased isotropic correction of edge effects for sampling windows of any shape (i.e., rectangular, circular, polygonal or with holes), as well as procedures to fast computation of

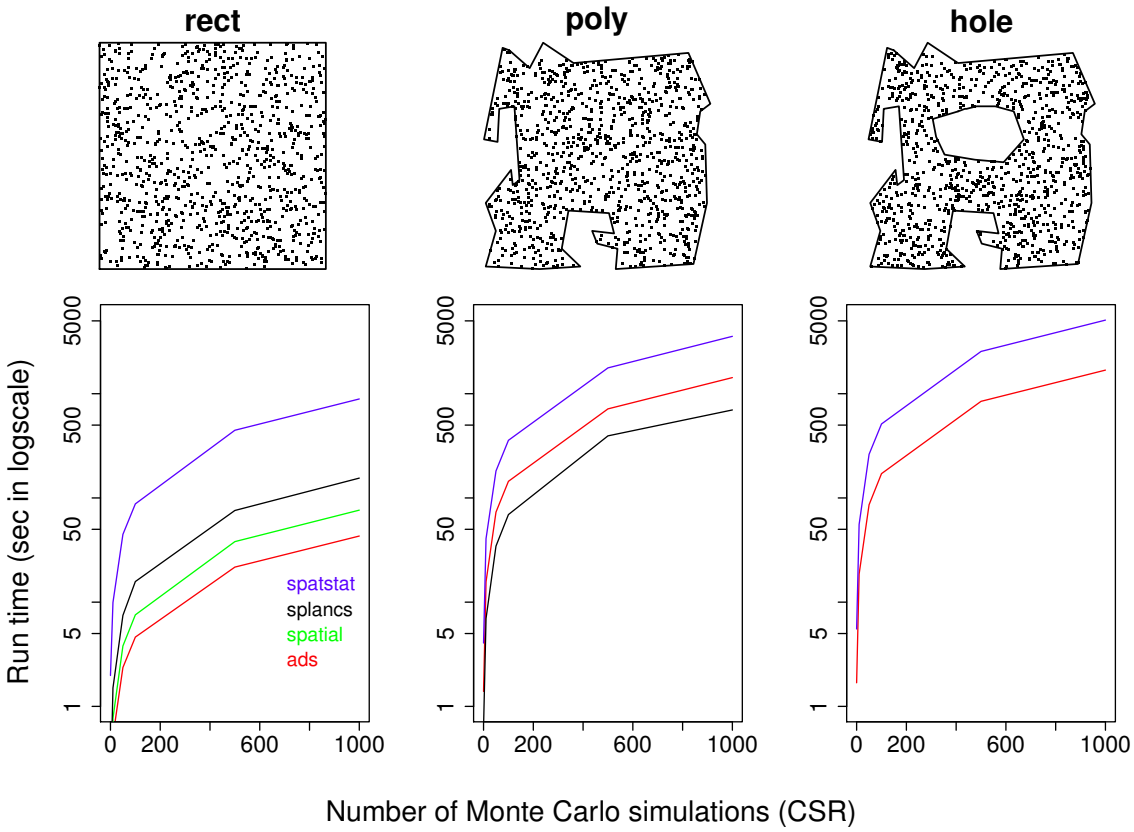

Figure 1: Computational run time[1] of four R packages to estimate Ripley's $K(r)$ for a simulated point pattern of 1000 points in three different sampling windows. Monte Carlo simulations are here used for testing spatial patterns against Complete Spatial Randomness (CSR). Though **spatial** only allows a rectangular sampling window and **splancs** doesn't allow a polygon with holes, the four packages return, in all comparable cases, exactly the same numerical results for $K(r)$ (see replication file). Similar differences in performance were obtained for the other K-functions listed in Table 1.

robust confidence envelopes of basic spatial null hypotheses from a large number of Monte Carlo simulations. It therefore combines efficiency of historical packages for point pattern analysis that interface foreign languages, i.e., **spatial** (Venables and Ripley 2002) or **splancs** (Rowlingson and Diggle 1993), with the flexibility of the more recent **spatstat** (Baddeley and Turner 2005), especially for data manipulation, point pattern modeling and simulation. As an illustration, Figure 1 shows that in terms of run time[1], **ads** largely outperforms the other R packages, except **splancs** for a single polygonal sampling window, probably because of two different implementations of the isotropic edge effects correction (see Agterberg 1994; Goreaud and Pélissier 1999).

In **ads**, a point pattern is defined as a set of point coordinates in a finite sampling window of rectangular, circular or polygonal type, optionally with hole(s). The points may carry individual attributes such as categorical labels (e.g., species names) or real values of a continuous

[1]As returned in variable `elapsed` by function `system.time` on x86_apple-darwin10.8.0 platform (iMac10.1 Intel Core 2 Duo 64-bit @ 3.06 GHz) with R version 3.0.0.

Package name		**ads**	**spatstat**	**spatial**	**splancs**
Allowed shapes for edge effects correction[1]		rect, circ, poly, hole	rect, poly, hole	rect	poly
Univariate point pattern analysis					
Summary statistics					
$K(r)$	Ripley (1977)	`kfun`	`Kest`	`Kfn`	`khat`
$L(r)$	Besag (1977)	—''—	`Lest`		
$g(r)$	Stoyan, Kendall, and Mecke (1987)	—''—	`pcf`		
$n(r)$	Goreaud (2000)	—''—			
Monte Carlo simulations for summary statistics					
CSR[2]	Diggle (1983)	`kfun`	`envelope`	`Kenvl`	`Kenv.csr`
other			—''—	—''—	`Kenv.pcp`[3]
Local statistics					
$n(x, y)$	Pélissier and Goreaud (2001)	`dval`[4]			
$L_i(r)$	Getis and Franklin (1987)	`kval`	`localL`		
$K_i(r), g_i(r), n_i(r)$		—''—	`localK, localpcf`		`khat`[5]
Bivariate point pattern analysis					
Summary statistics					
$K_{12}(r)$	Lotwick and Silverman (1982)	`k12fun`	`Kcross`		`k12hat`
$L_{12}(r)$	Duncan (1991)	—''—	`Lcross`		
$g_{12}(r)$	Stoyan and Stoyan (1994)	—''—	`pcfcross`		
$n_{12}(r)$	Goreaud (2000)	—''—			
Monte Carlo simulations for summary statistics					
RL[6]	Diggle (1983)	`k12fun`	`envelope`		`Kenv.label`
PI[7]	Lotwick and Silverman (1982)	—''—	—''—		`Kenv.tor`
other			—''—		
Local statistics					
$K_{i2}(r), L_{i2}(r), g_{i2}(r), n_{i2}(r)$		`k12val`			
Marked point pattern analysis					
Summary statistics					
$K_m(r), g_m(r)$	Goreaud (2000)	`kmfun`	`markcorr`[8]		
Monte Carlo simulations for summary statistics					
IM[9]	Penttinen, Stoyan, and Henttonen (1992)	`kmfun`	`envelope`		
other			—''—		
Multivariate point pattern analysis					
Array of summary statistics					
$K_{pq}(r)$		`kpqfun`	`Kmulti`[10]		
$L_{pq}(r), g_{pq}(r), n_{pq}(r)$		—''—			
$K_{p.}(r)$		`kp.fun`	`Kdot`[11]		
$L_{p.}(r), g_{p.}(r), n_{p.}(r)$		—''—			
Species diversity statistics					
$K_S(r), g_S(r)$	Shimatani (2001)	`ksfun`			
$K_R(r), g_R(r)$		`krfun`			
$K_d(r), g_d(r)$	Shen, Wiegand, Mi, and He (2013)	`kdfun`			
Monte Carlo simulations for species diversity statistics					
RL[6]	Diggle (1983)	`ksfun, krfun`			
RP[12]	Shen *et al.* (2009)	`ksfun, krfun`			
SE[13]	Shen *et al.* (2013)	`kdfun`			

Table 1: Computational functions of point pattern analysis implemented in **ads** with equivalent functions in other R packages. [1] rectangular (rect), circular (circ), polygonal (poly), polygonal with holes (hole); [2] Complete Spatial Randomness hypothesis; [3] limited to Poisson Cluster Process (Diggle 1983); [4] 1^{st}-order local density function with isotropic correction of edge effects; [5] provides only $K_i(r)$; [6] Random Labeling hypothesis; [7] Population Independence hypothesis; [8] provides the mark correlation function $K_{mm}(r)$ of Penttinen *et al.* (1992), which is slightly different from $K_m(r)$, but also Stoyan's generalized mark correlation function, $K_F(r)$ that allows customizing the test function, F; [9] Independent Marking hypothesis; [10] while `kpqfun` computes an array of $K(r)$ and $K_{12}(r)$ functions for each couple (p, q) of marks in a multivariate point pattern, `Kmulti` computes only $K_{12}(r)$ for a selected couple of marks; [11] while `kp.fun` computes an array of $K_{12}(r)$ functions for each mark p against all other marks grouped together in a multivariate point pattern, `Kdot` computes $K_{12}(r)$ for a selected mark p against all other marks grouped together; [12] Random Placement hypothesis; [13] Species Equivalence hypothesis.

variable (e.g., tree stem diameter), so defining various types of point patterns (i.e., univariate, multivariate or marked). The computational functions automatically recognize these objects and their types, and call accordingly the appropriate C subroutines. Version 1.5 of **ads** computes the most classical members of the K-function family, as in particular does **spatstat**, but also more specific functions for the spatial analysis of species diversity from multivariate point patterns, which, from our knowledge, do not exist in any other R package (Table 1).

2. Data preparation

In **ads**, a spatial point pattern is defined as a set of point coordinates in a finite rectangular, circular or irregular-shaped two-dimensional study domain. The study domain is represented by an object of the class `"swin"` (sampling window), to which point coordinates are attached along with optional attributes (marks) to create an object of the class `"spp"` (spatial point pattern). All computational functions in **ads** use an `"spp"` object as mandatory input argument. Specific `print`, `summary` and `plot` methods apply to `"swin"` and `"spp"` objects. A special function `owin2swin` allows converting **spatstat** observational windows (`"owin"` objects) into `"swin"` objects, so that the extended facilities of **spatstat** for data preparation are also available to **ads** users.

2.1. Sampling window definition

Function `swin` creates an object of the class `"swin"`, which is of "simple" or "complex" type. A simple sampling window is either:

- *Rectangular*, defined as a vector of length 4 giving coordinates $(xmin, ymin)$ and $(xmax, ymax)$ of the origin and the opposite corner of the sampling window, e.g., for a square of 110×90:

```
R> rect <- swin(window = c(0, 0, 110, 90))
```

- *Circular*, defined as a vector of length 3 giving coordinates $(x0, y0)$ of the centre and the radius $r0$ of the sampling window, e.g., for a disc of radius 30 centered at $(50, 50)$:

```
R> circ <- swin(window = c(50, 50, 30))
```

To comply with the method of edge effects correction implemented in **ads** (see Goreaud and Pélissier 1999), a complex sampling window type is defined by removing triangular surfaces from an initial sampling window of simple shape (rectangular or circular). The triangles may be removed near the boundary of a rectangular window in order to design a polygonal sampling window or as "holes" when they do not connect with the outer sampling window boundary. Triangles are passed to function `swin` as an optional list of vectors of length 6, giving for each triangle the coordinates (ax, ay, bx, by, cx, cy) of its vertices, e.g., for a polygonal sampling window defined by 3 triangles within an initial rectangle of 110×90:

```
R> tri <- data.frame(ax = c(0, 0, 80), ay = c(0, 40, 0), bx = c(0, 0, 110),
+         by = c(40, 90, 0), cx = c(50, 60, 110), cy = c(0, 90, 50))
R> poly <- swin(window = c(0, 0, 110, 90), triangles = tri)
```

Combinations of simple or polygonal shapes with holes are possible. In order to help users define non-overlapping triangles, the utility function `triangulate` divides a simple polygon (optionally with holes) into a collection of contiguous triangles using a method based on a fast triangulation algorithm (Nakhede and Manocha 1995). Function `area.swin`, called by `summary.swin`, computes the total area of the resulting study domain.

2.2. Point pattern definition

Function `spp` creates an object of the class `"spp"`, defined from an `"swin"` object to which are attached point coordinates along with some optional attributes, such as categorical or numerical marks, so representing either:

- *Univariate* point pattern:

```
R> x <- runif(100, 0, 110)
R> y <- runif(100, 0, 90)
R> unispp <- spp(x = x, y = y, window = rect)
```

- *Multivariate* point pattern, when a categorical mark is attached to each point of the pattern through optional argument `marks`:

```
R> sp <- as.factor(c(rep("sp1", 50), rep("sp2", 30), rep("sp3", 20)))
R> multispp <- spp(x = x, y = y, window = rect, marks = sp)
```

- *Marked* point pattern, when a numerical mark is attached to each point of the pattern through optional argument `marks`:

```
R> var <- sample(100)/10
R> markspp <- spp(x = x, y = y, window = poly, marks = var)
```

Function `spp` automatically segregates points located outside the limits of the sampling window so that they are ignored in computational functions. For instance, when the sampling window is of "complex" type, `spp` calls function `inside.swin` to segregate the points located inside the triangles. Similarly, attaching the same set of points to a resized sampling window creates a new `"spp"` object (Figure 2). A specific function `ppp2spp` converts `"ppp"` planar point pattern objects from **spatstat** to `"spp"` objects.

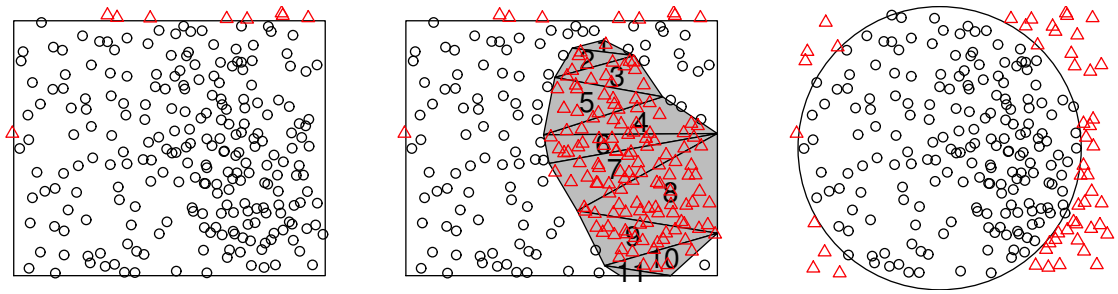

Figure 2: Plots of `"spp"` objects from dataset `BPoirier` with resized sampling windows (`"swin"` objects) attached to the same set of point coordinates. Points located outside the sampling window (in red color) are ignored by **ads** computational functions.

3. Data analysis

Package **ads** basically performs second-order neighborhood analyses derived from Ripley's (1977) K-function for univariate, multivariate or marked point patterns (the K-function family). These functions use `"spp"` objects as input argument and return distance-dependent summary statistics stored as objects of the class `"fads"`. Ancillary functions also compute local values of first- and second-order densities stored as objects of the class `"vads"`. There exist `print` and `plot` methods for both these classes of objects.

3.1. Univariate functions

Implementation principles

In **ads** all univariate functions of second-order neighborhood analysis are estimated following the same principle: the expected number of neighbors within a distance r from an arbitrary point of the pattern is computed as its mean over all points by $\hat{N}(r) = \sum_{i=1}^{N} \sum_{j \neq i}^{N} k_{ij}/N$, where N is the number of points of the pattern, $k_{ij} = 0$ if $d(i,j) > r$, else $k_{ij} = 2\pi r/P_{ij}$, with P_{ij} the perimeter of the circle centered on i and passing through j, which is inside the sampling window. It follows that when $d(i,j) \leq r$, $k_{ij} = 1$ if point i is at a distance smaller than r to the sampling window boundary, and $k_{ij} > 1$ otherwise. This corresponds to Ripley's (1977) unbiased correction of edge effects (also called isotropic correction) for the computation of which explicit geometrical formulas are implemented in **ads** to apply to simple or complex sampling windows as defined above (Goreaud and Pélissier 1999). For computational efficiency, values of $\hat{N}(r)$ are computed for t successive values of r equally spaced by a distance interval of d, with t an integer in $[1, tmax]$, so that $r = td$. Function `kfun` computes various common summary statistics derived from this basic computation:

- The local neighbor density function $\hat{n}(r) = \hat{N}(r)/\pi r^2$ is an area standardization with an expectation of $\hat{\lambda} = N/A$ for all r under the hypothesis of Complete Spatial Randomness (CSR; Diggle 1983), with A the sampling window area.
- Ripley's (1977) function $\hat{K}(r) = \hat{N}(r)/\hat{\lambda}$ with an expectation of πr^2 under CSR.
- The linearized version $\hat{L}(r) = \sqrt{\hat{K}(r)/\pi} - r$ with an expectation of 0 under CSR (Besag 1977).
- The derivative $\hat{g}(r) = [\hat{K}(td) - \hat{K}([t-1]d)]/[\pi(td)^2 - \pi([t-1]d)^2]$ with an expectation of 1 under CSR, is an estimator of the pair density function (Stoyan *et al.* 1987) giving the expected number of neighbors within an annulus between r and $r - d$, so that an estimator of the O-ring statistics (Wiegand and Moloney 2004) is $\hat{O}(r) = \hat{\lambda}\hat{g}(r)$.

Monte Carlo simulations

As the theoretical distributions of these estimators are unknown, Monte Carlo simulations are used to build test statistics and confidence envelopes of specified null hypotheses (Besag and Diggle 1977). The method consists of simulating a large number ($Nsim$) of realizations of a spatial point process representing the null hypothesis, for instance a homogeneous Poisson process for representing CSR hypothesis (the only null hypothesis currently allowed in `kfun`).

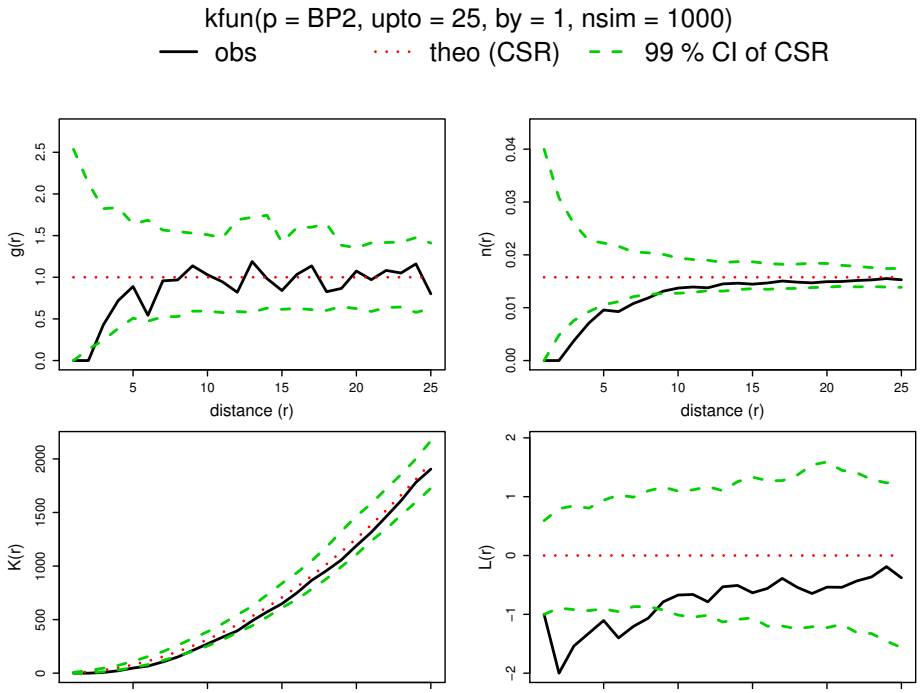

Figure 3: Univariate functions estimated for `BPoirier` dataset with `kfun` function computed from 1 to 25m by 1-m distance steps and a 99% local confidence envelope obtained from 1000 Monte Carlo simulations of CSR hypothesis (run time[1] = 1.686 sec). `BP2` is an `"spp"` object as defined in the replication file and corresponding to the central pattern in Figure 2. $L(r)$ shows a significant negative deviation within a neighborhood distance of 1 to 7m, indicating regularity of the pattern.

Because of edge effects, particularly in complex sampling windows, the process is simulated in the same sampling window and with the same number of points as the observed point pattern. The K-functions are then estimated for each realization of the simulated process. Strictly speaking, simulations should be independent at each distance value, but in practice the bias introduced when the functions are computed over all the range of r distances at each simulation is negligible, at least when the number of simulations is large enough (Goreaud 2000).

At a given r, a local test of the null hypothesis is built by comparing the absolute deviation of estimated values to the expected theoretical values, obtained for the observed point pattern on the one hand, and for the different realizations of the null hypothesis on the other hand (Barot and Gignoux 1999). For the univariate estimators above, this absolute deviation is: $\delta n(r) = |\hat{n}(r) - \hat{\lambda}]$; $\delta K(r) = |\hat{K}(r) - \pi r^2|$; $\delta L(r) = |\hat{L}(r)|$; $\delta g(r) = |\hat{g}(r) - 1|$.

At each r, the rate of simulations that lead to an observed deviation lower than or equal to the simulated one (rejection of the null hypothesis) gives the local significance level of the test (P-value). One can also build a local bilateral confidence envelope for a given type 1 error risk of 2α, by taking as lower and upper limits at each r, the α^{th} and $(100-\alpha)^{th}$ percentiles of the distribution of the simulated values, respectively. When the observed value of a function at a given r is outside the confidence envelope limits, the spatial pattern is significantly different at the 2α risk level from the one expected under the null hypothesis (Figure 3).

3.2. Bivariate functions

The bivariate K_{12}-function proposed by Lotwick and Silverman (1982) is an extension of Ripley's K-function to analyze spatial patterns of points bearing a two-level categorical mark. It is also known as the intertype function (Diggle 1983), which characterizes the spatial interaction between points of two different types (such as trees of two different species) located in the same study domain. In **ads**, bivariate functions are computed from an estimate of the expected number of type 2 points located within a distance r of an arbitrary type 1 point of the pattern, by $\hat{N}_{12}(r) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} k_{ij}/N_1$, where N_1 and N_2 are the number of type 1 and type 2 points of the pattern, respectively, and k_{ij} is as in the univariate case, including the isotropic correction of edge effects, but with i and j representing the points of type 1 and 2, respectively. Function `k12fun` computes the following:

- $\hat{n}_{12}(r) = \hat{N}_{12}(r)/\pi r^2$;
- $\hat{K}_{12}(r) = \hat{N}_{12}(r)/\hat{\lambda}_2$ where $\hat{\lambda}_2 = N_2/A$, and A is the sampling window area;
- $\hat{L}_{12}(r) = \sqrt{\hat{K}_{12}(r)/\pi} - r$;
- $\hat{g}_{12}(r) = [\hat{K}_{12}(td) - \hat{K}_{12}([t-1]d)]/[\pi(td)^2 - \pi([t-1]d)^2]$.

Here also, P-values and confidence envelope limits are computed at each r from observed and simulated absolute deviations between the empirical and theoretical values expected under a given null hypothesis. For the bivariate case, the null hypothesis of no interaction between the two types of points correspond either to (Goreaud and Pélissier 2003):

- *A hypothesis of random labeling* (RL) when the spatial pattern of type 1 and type 2 points result from a single completely random process affecting the points *a posteriori* (for instance a disease attack within a tree population). In this case, expectations of the bivariate functions are: $n_{12}(r) = \lambda$; $K_{12}(r) = K(r)$; $L_{12}(r) = L(r)$; $g_{12}(r) = g(r)$. This hypothesis is tested conditionally to the spatial pattern of the whole population (i.e., without type distinction) by reallocating the types at random amongst the points of the pattern whose location is kept unchanged.
- *A hypothesis of population independence* (PI) when the spatial patterns of type 1 and type 2 points result from different processes (for instance two tree species with their own dispersal process), with expectations: $n_{12}(r) = \lambda_2$; $K_{12}(r) = \pi r^2$; $L_{12}(r) = 0$; $g_{12}(r) = 1$. This hypothesis is tested conditionally to the spatial pattern of each population, classically by shifting type 1 points by a random vector over a torus, while the pattern of type 2 points is kept unchanged (PI-tor), as proposed by Lotwick and Silverman (1982). In case of a non-rectangular sampling window, the torus connects the top and bottom edges of the rectangular frame enclosing the sampling window so that some of the type 1 points can be shifted outside the limits of the actual sampling window and will be thus ignored in computation. This however is not expected to bias $K_{12}(r)$ estimations, since type 1 points are the focal points for which we estimate the number of type 2 neighbors. Problems may however arise in extreme cases when the number of type 1 points ignored is too large (e.g., when the enclosing rectangle is too large with respect to the actual polygonal window), so that the remainder doesn't allow a correct estimation of the mean number of type 2 neighbors (Figure 4). As an alternative, the

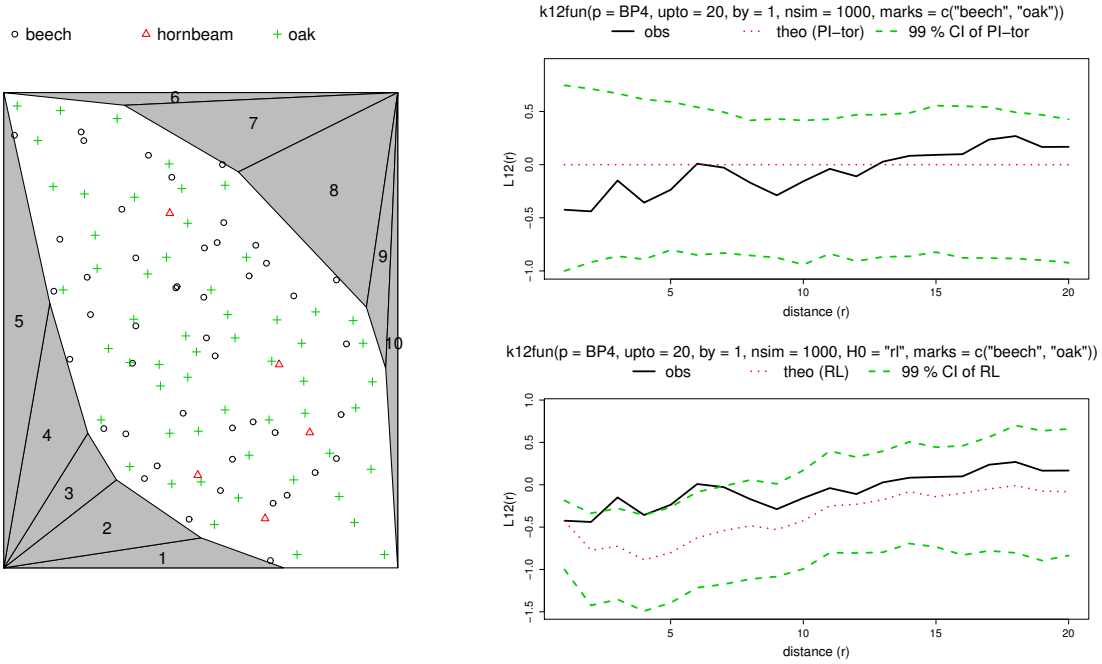

Figure 4: Bivariate functions estimated for "beech" and "oak" species from dataset `BPoirier` (left) using `k12fun` with a 99% local confidence envelope obtained from 1000 Monte Carlo simulations of population independence hypothesis (PI-tor; top panel; run time[1] = 0.610 sec) or random labeling hypothesis (RL; bottom panel; run time= 0.993 sec). Note that in this example (2 different species populations) RL is probably not an ecologically sound null hypothesis (dataset `Allogny` provides a more appropriate example). It however illustrates the two different theoretical expectations (red lines) and confidence envelopes (green lines) under PI and RL null hypotheses. `BP4` is an `"spp"` object as defined in the replication file.

mimetic point process (Goreaud, Loussier, Ngo Bieng, and Allain 2004) introduced in Section 3.6 below can be used to replicate the pattern of type 1 points, keeping here also the pattern of type 2 points unchanged (PI-mim). Here also it is recommended for efficiency to avoid defining a too large enclosing rectangle because the points are first generated within the rectangle and then retained only if they do not fall within a triangle. This procedure, though more robust, is however computationally intensive.

3.3. Multivariate functions

Multivariate functions apply to patterns of points bearing more than two categorical marks, such as, for instance, trees in a mixed-species forest stand. They are basically of two different types in **ads**: simple wrapper functions to display bivariate functions in a practical matrix-like format, or more advanced functions to analyze and test spatial pattern of species diversity.

Wrapper functions for multivariate point patterns

Function `kp.fun` computes a set of K_{12}-functions between all possible marks p and the other marks grouped together in a multivariate spatial pattern, while function `kpqfun` computes,

for all possible pairs of marks (p, q), a set of K- and K_{12}-functions, when $p = q$ and $p \neq q$, respectively. It is however not possible to compute confidence envelopes directly from these functions. Specific `plot` methods display the results in a matrix-like format.

Spatial pattern of species diversity

The Simpson's index (Simpson 1949) of species diversity gives the probability that a randomly selected pair of individuals in a sample has two different species. Its unbiased estimator is: $\hat{D} = 1 - \sum_{p=1}^{S} N_p(N_p - 1)/N(N - 1)$, where S is the total number of species, N_p the number of individuals of species p and $N = \sum_{p=1}^{S} N_p$. Shimatani (2001) proposed functions of multivariate spatial point pattern analysis as distance-dependent extensions of Simpson's index that he demonstrated to be estimable by combining univariate K-function estimators (see also Eckel, Fleischer, Grabarnik, and Schmidt 2008):

- $\hat{K}_S(r) = 1 - \sum_{p=1}^{S} \hat{\lambda}_p^2 \hat{K}_p(r)/\hat{\lambda}^2 \hat{K}(r)$, which is an estimator of $\alpha(r)$ of Shimatani (2001).
- $\hat{g}_S(r) = 1 - \sum_{p=1}^{S} \hat{\lambda}_p^2 \hat{g}_p(r)/\hat{\lambda}^2 \hat{g}(r)$, which is an estimator of $\beta(r)$ of Shimatani (2001).

In the above equations $K(r)$ and $K_p(r)$ — respectively $g(r)$ and $g_p(r)$ — are simply the univariate K-functions — respectively g-functions — computed from the spatial pattern of all species pooled together (no type distinction) and of species p, respectively. $\hat{K}_S(r)$ is the mean over all points of the pattern of Simpson's diversity within a neighborhood distance r, which conceptually corresponds to the notion of a distance-dependent measure of α-diversity *sensu* Whittaker (1972). However, $g_S(r)$ is hard to conceptualize as proposed by Shimatani (2001), as a distance-dependent version of a β-diversity measure *sensu* Whittaker (1972) (i.e., as a dissimilarity between two samples), but more simply gives an estimate of Simpson's diversity in an annulus between r and $r - d$. It follows that under the null hypothesis of a perfectly-mixed species assemblage $\hat{K}_S(r) = \hat{g}_S(r) = \hat{D}$ or equivalently $\hat{K}_S(r)/\hat{D} = \hat{g}_S(r)/\hat{D} = 1$. In **ads**, function `ksfun` computes both functions as wrapper functions of `kfun` as introduced in Section 3.1. The null hypothesis can be tested either by a generalization of the random labeling procedure when the marking process is considered to have affected the points *a posteriori* to the spatial location process (see Section 3.2), or by a random placement procedure (RP) simulating random spatial patterns for each species independently (Shen *et al.* 2009), an approach more compatible with the biological hypothesis of a species assemblage in absence of any inter- and intra-specific interactions.

A more general diversity index that measures the expected difference between individuals of a randomly selected pair in a sample is Rao's (1982) quadratic entropy, for which an unbiased estimator is: $\hat{H}_D = \sum_{p=1}^{S} \sum_{q=1}^{S} d_{pq} N_p N_q / N(N - 1)$, where N_p and N_q are the numbers of individuals of species p and q in a sample of size N, respectively, and d_{pq} measures a phylogenetic or functional Euclidean distance between these species (Pavoine, Ollier, and Pontier 2005). It is noteworthy that when $d_{pq} = 1$ for all $p \neq q$ and 0 otherwise, H_D reduces to Simpson's D (see above). As this index requires considering all pairs of non-conspecific points, it can be envisioned as a combination of univariate and bivariate K-function estimators that *de facto* integrates the isotropic correction of edge effects:

$$\hat{K}_R(r) = \sum_{p=1}^{S} \sum_{q=1}^{S} d_{pq} \hat{\lambda}_p \hat{\lambda}_q \hat{K}_{pq}(r)/\hat{\lambda}^2 \hat{K}(r)$$

$\hat{K}_R(r)$ is the mean over all points of the pattern of the quadratic entropy within a neighborhood distance r. We can also define $g_R(r)$, the quadratic entropy in an annulus between r and $r - d$ as:

$$\hat{g}_R(r) = \sum_{p=1}^{S} \sum_{q=1}^{S} d_{pq} \hat{\lambda}_p \hat{\lambda}_q \hat{g}_{pq}(r) / \hat{\lambda}^2 \hat{g}(r)$$

Under the null hypothesis of a perfectly-mixed species assemblage $\hat{K}_R(r) = \hat{g}_R(r) = \hat{H}_D$ or equivalently $\hat{K}_R(r)/\hat{H}_D = \hat{g}_R(r)/\hat{H}_D = 1$, tested, as for $K_S(r)$, by either a general random labeling or a random placement procedure. In **ads**, function `krfun` computes both functions as wrapper functions of `kfun` and `k12fun` introduced above. When $d_{pq} = 1$ for all $p \neq q$ and 0 otherwise, $K_R(r)$ and $g_R(r)$ are alternative implementations of $K_S(r)$ and $g_S(r)$, respectively. This thus provides an opportunity for testing the highly meaningful ecological hypothesis of species equivalence (SE) based on d_{pq} (Shen *et al.* 2013).

Under SE hypothesis we expect $K_R(r)/H_D = K_S(r)/D$ — respectively $g_R(r)/H_D = g_S(r)/D$ — , which can be tested using a Monte Carlo procedure shuffling the between-species distances by permuting simultaneously the rows and columns in d_{pq}. When d_{pq} is a cophenetic distance matrix, this is equivalent to permuting the tips of a hierarchical (for instance, phylogenetic) tree (Hardy 2008). Shen *et al.* (2013) demonstrated powerfulness of this approach for testing ecological hypotheses of phylogenetic spatial structure based on a function they called $k_d(r)$ and which is actually very close to the ratio of $g_R(r)/H_D$ on $g_S(r)/D$. Thus, Shen's et al. function should have been better quoted as $g_d(r)$ in order to keep, for consistency, the notation $K_d(r)$ for the cumulative version corresponding to the ratio of $K_R(r)/H_D$ on $K_S(r)/D$. In **ads**, function `kdfun` computes:

- $\hat{K}_d(r) = \hat{D} * \sum_{p=1}^{S} \sum_{q=1}^{S} d_{pq} \hat{\lambda}_p \hat{\lambda}_q \hat{K}_{pq}(r) / \hat{H}_D * \sum_{p=1}^{S} \sum_{q=1}^{S} \hat{\lambda}_p \hat{\lambda}_q \hat{K}_{pq}(r)$.
- $\hat{g}_d(r) = \hat{D} * \sum_{p=1}^{S} \sum_{q=1}^{S} d_{pq} \hat{\lambda}_p \hat{\lambda}_q \hat{g}_{pq}(r) / \hat{H}_D * \sum_{p=1}^{S} \sum_{q=1,}^{S} \hat{\lambda}_p \hat{\lambda}_q \hat{g}_{pq}(r)$.

The advantage here is that edge effects cancel out between numerator and denominator, so that an efficient implementation for a sampling window of any shape doesn't require correcting for edge effects (i.e., $k_{ij} = 1$ in $g_{pq}(r)$ and $K_{pq}(r)$ whatever the relative locations of point i of species p and point j of species q with respect to the sampling window boundary; see Section 3.2). Both functions quantify the phylogenetic/functional spatial structure of a community conditionally to the multi-species spatial pattern (see Shen *et al.* 2013), with in both cases, a theoretical expected value of 1 at all r under the species equivalence hypothesis.

Figure 5 gives an example of analysis of the pattern of species diversity for a tropical forest plot of 250 × 250 m with 4128 trees of 332 different species (dataset `Paracou15`; Gourlet-Fleury, Guehl, and Laroussinie 2004).

3.4. Marked functions

The K_m-function analyses the spatial structure of correlations between values of a quantitative variable borne by points of a marked pattern. The version implemented in **ads** is simply a correlogram corrected for edge effects using Ripley's istropic method: $K_m(r) = Cov(X_i, X_j) k_{ij} / Var(X) N(r)$, where X is a random variable defining the marks at points i and j and $N(r)$ is the mean number of pairs of points that are neighbors within distance r.

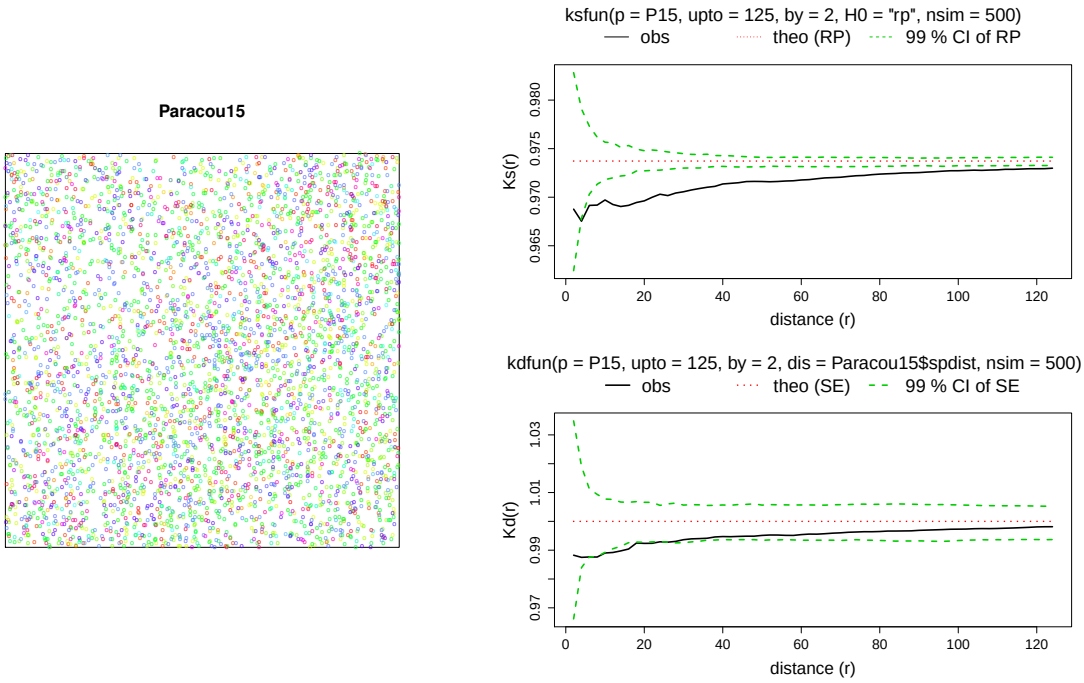

Figure 5: Multivariate functions of species diversity estimated for dataset `Paracou15` (left) using `ksfun` (top panel) and `kdfun` (bottom panel) with a 99% local confidence envelope obtained from 500 Monte Carlo simulations of random placement hypothesis (RP; top panel; run time[1] = 400.735 sec) or species equivalence hypothesis (SE; bottom panel; run time= 151.145 sec). `P15` is an `"spp"` object as defined in replication file, with a between-species distance matrix corresponding to a cophenetic distance based on APG III phylogeny (Chase and Reveal 2009). This example shows a significant negative deviation of Simpson's diversity from the RP hypothesis from about 5 m (more conspecific neighbors than expected for the entire plot) and of Rao's diversity from the SE hypothesis (more closely related neighbors than expected) in the range 5–20m.

An unbiased estimator of this function is (Goreaud 2000):

$$\hat{K}_m(r) = N \sum_{i=1}^{N} \sum_{j \neq i} (x_i - \bar{x})(x_j - \bar{x}) k_{ij} / (\sum_{i=1}^{N} (x_i - \bar{x})^2 \sum_{i=1}^{N} \sum_{j \neq i} k_{ij})$$

where x_i and x_j are the values of X borne by the neighboring points i and j, k_{ij} is as in the univariate function, and $\bar{x}$ is the mean of X over the N points of the pattern. $K_m(r)$ varies between -1 and 1 and has an expectation of 0 when the marks are totally uncorrelated. In **ads**, function `kmfun` also computes the derivative of $K_m(r)$ (or pair correlation function; Penttinen *et al.* (1992)) that gives the correlation of marks within an annulus between r and $r - d$:

$$\hat{g}_m(r) = [\hat{K}_m(td) - \hat{K}_m([t-1]d)]/[\pi(td)^2 - \pi([t-1]d)^2]$$

Like $K_m(r)$, $g_m(r)$ has an expectation of 0 in absence of mark correlation. For both functions local confidence envelops and P-values of departure from the null hypothesis of an absence of correlation between marks is tested by reallocating at random the values of X over all points of the pattern (Independent Marking), keeping the spatial location of points unchanged

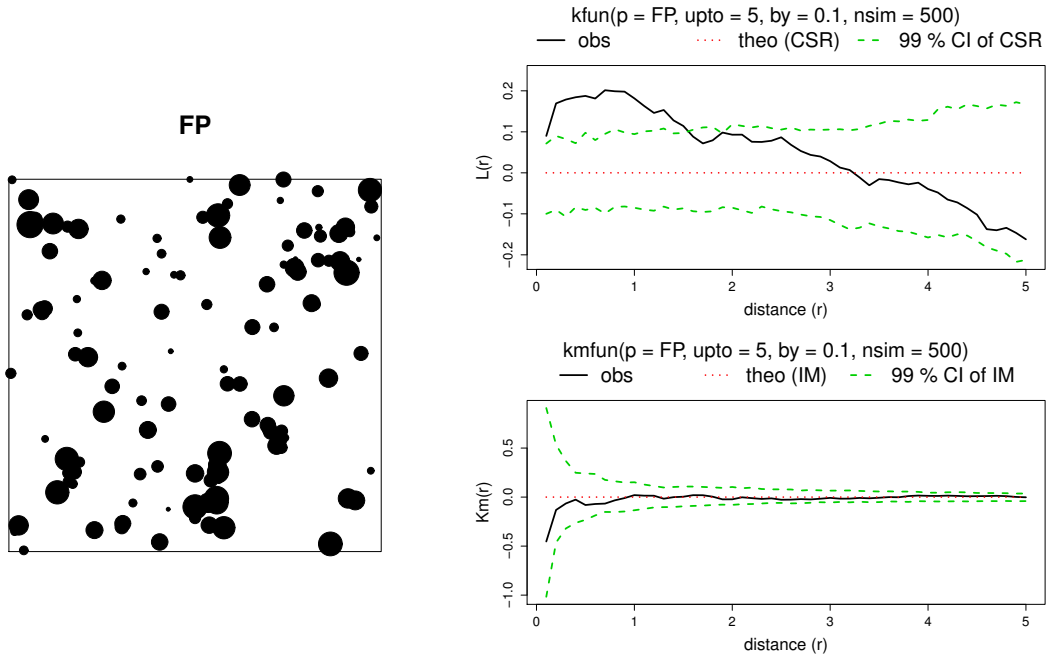

Figure 6: Spatial pattern of tree height from dataset `finpines` (left) along with the univariate L-function with confidence envelope of CSR (top panel) and the marked K_m-function with confidence envelope of independent marking hypothesis (IM) obtained from 500 Monte Carlo simulations (bottom panel; run time[1] = 0.367 sec). This example shows that while saplings are significantly clustered in space, their height is spatially uncorrelated.

(Figure 6). Two specific applications of this function in the context of forest studies can be found in Oddou-Muratorio, Demesure-Musch, Pélissier, and Gouyon (2004) and Madelaine *et al.* (2007).

3.5. Local density functions

Ancillary functions computing first- and second-order density values for a range of neighborhood distances are also available in **ads** (Pélissier and Goreaud 2001). The second-order local density function, `kval`, computes a local version of Ripley's univariate functions as proposed in Getis and Franklin (1987). Basically the function computes, for each point i of the pattern, $n_i(r) = N_i(r)/\pi r^2$, where $N_i(r)$ is the number of neighbors within distance r of point i corrected for edge effects, which has an expectation of $(N-1)/A$ under CSR, with N the total number of points of the pattern and A the sampling window area. It also returns the classical univariate transformations, i.e., $K_i(r)$, $L_i(r)$ and $g_i(r)$ (see Section 3.1). Function `k12val` is the bivariate extension, which gives $n_{i2}(r)$, the local density of neighbors of type 2 within a distance r of each i point of type 1. Consistently with this approach, we also introduced the first-order local density function, `dval`, which computes values of $\hat{n}(x, y) = N(r)/\pi r^2$ at each node (x, y) of a systematic grid covering the sampling window, where $N(r)$ is the number of points of the pattern within a distance r of a node, to which Ripley's isotropic correction of edge effects is applied. Specific `summary` and `plot` methods exist for these functions that are useful for exploratory data analysis.

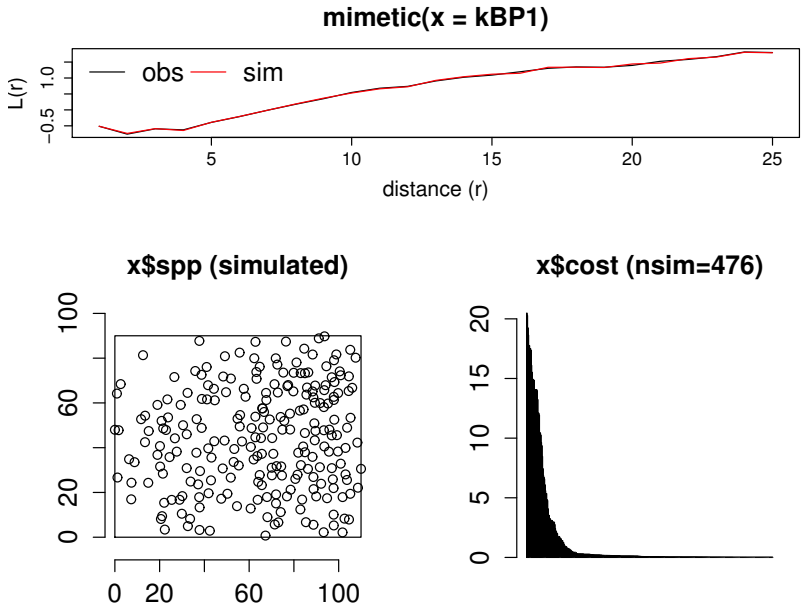

Figure 7: Diagnostic plot of the mimetic point process that replicates the observed `BPoirier` univariate point pattern (left panel in Figure 2) using a depletion-replacement procedure minimizing a global cost based on the difference between the observed and simulated L-functions. `kBP1` is a `"fads"` object returned by function `kfun`. In this example, 476 depletion-replacement events were needed to reach minimum global cost (run time[1] = 1.432 sec).

3.6. Spatial point pattern replication

Spatial point patterns are realizations of spatial point processes. Goreaud *et al.* (2004) developed a stepwise depletion-replacement procedure to generate different realizations of an observed point pattern using a mimetic point process that minimizes the following global cost: $\sum_{i=1}^{N} \|L_{obs}(r) - L_{sim}(r)\|^2$, where L_{obs} is the univariate L-function computed for an observed point pattern and L_{sim} is the L-function computed at each step of the depletion-replacement process. Function `mimetic` implemented in **ads** initially simulates a completely random spatial pattern with same intensity as the observed pattern. The procedure makes the global cost function slowly decrease and the process stops when it converges to a minimum, i.e., when a given number of depletion-replacement events (50 by default) do not make the cost function significantly decrease (Figure 7). The function works with a sampling window of any shape.

3.7. Conclusion and perspectives

Package **ads** has been developed in the late 90's for the requirements of forest ecologists, initially as a standalone C application within a special section of **ADE-4** software (Thioulouse, Chessel, Dolédec, and Olivier 1997) — now package **ade4** for R (Dray and Dufour 2007). The core C code used for the implementation of the geometrical functions of edge effects correction was written by Goreaud (2000) in the framework of his PhD thesis, based on triangle geometry (Goreaud and Pélissier 1999). The original C code was ported to R in 2007. At this time the paper by Agterberg (1994), that presents the geometrical principles behind the method

of edge effects correction used in **splancs**, and which proved faster for polygonal sampling windows than the one used in **ads** (see Figure 1), was unknown by us. There is consequently an avenue of improvement of **ads** by integrating the two approaches on the basis of **splancs** fortran code ported to R by Bivand and Gebhardt (2000). For similar historical reasons, the sampling window (`"swin"`) and spatial point pattern (`"spp"`) objects used in **ads** are similar but not exactly the same to the ones used in **spatstat**. Though we developed special functions, `owin2swin` and `ppp2spp`, for compatibility with **spatstat** objects, much remain to be done to allow users to shift more easily from one package to the other. On the one hand **ads** proved to be always much faster than **spatstat**, a useful property for one to deal with large datasets and robust confidence envelopes (i.e., based on a large number of Monte Carlo simulations), while on the other hand, **spatstat** is more flexible for data manipulation and offers a larger panel of different spatial functions, not only limited to the K-function family. Finally the recent developments introduced in **ads** for the spatial analysis of species diversity, especially the ones based on Rao's (1982) quadratic entropy, do not seem to exist in any other R package. There is thus here an opportunity for the development of new such approaches integrating species diversity and spatial analyses, for instance to study the spatial structure of functional or phylogenetic species diversity in the well-mastered framework of the spatial point processes (see Shen *et al.* 2013). When the distance between species is based on a phylogenetic tree, various null hypotheses can be tested thanks to specific, more or less restricted, randomization procedures, whose implementation could be quite challenging (see Hardy 2008). Perspectives also exists to extend the mimetic point process for replicating multivariate point patterns, as introduced in Goreaud *et al.* (2004).

Acknowledgments

Package **ads** was ported to R with the help of Cyril Picard and Karine Pousseu-Tankeu, thanks to financial support from IRD-Spirales project 2005. Version 1.3 was developed with the help of Denis Redondo and Joris Moyano. We warmly thank Edzer Pedesma acting as Editor of the special issue of JSS on Spatial Statistics, as well as two anonymous reviewers whose comments greatly helped to improve the manuscript.

References

Agterberg FP (1994). "Fortran Program for the Analysis of Point Patterns with Correction for Edge Effects." *Computers & Geosciences*, **20**, 229–245.

Baddeley A, Turner R (2005). "**spatstat**: An R Package for Analyzing Spatial Point Patterns." *Journal of Statistical Software*, **12**(6), 1–42. URL http://www.jstatsoft.org/v12/i06/.

Barot S, Gignoux J (1999). "Population Structure and Life Cycle of *Borassus aethiopum* Mart.: Evidence of Early Senescence in a Palm Tree." *Biotropica*, **31**, 439–448.

Besag JE (1977). "Discussion on Dr. Ripley's Paper." *Journal of the Royal Statistical Society B*, **39**, 193–195.

Besag JE, Diggle PJ (1977). "Simple Monte Carlo Tests for Spatial Pattern." *Applied Statistics*, **26**, 327–333.

Bivand R, Gebhardt A (2000). "Implementing Functions for Spatial Statistical Analysis Using the R Language." *Journal of Geographical Systems*, **2**, 307–317.

Chase MW, Reveal JL (2009). "A Phylogenetic Classification of the Land Plants to Accompany APG III." *Botanical Journal of the Linnean Society*, **161**, 122–127.

Diggle PJ (1983). *Statistical Analysis of Spatial Point Patterns*. Academic Press, London.

Dray S, Dufour AB (2007). "The **ade4** Package: Implementing the Duality Diagram for Ecologists." *Journal of Statistical Software*, **22**(4), 1–20. URL http://www.jstatsoft.org/v22/i04/.

Duncan RP (1991). "Competition and the Coexistence of Species in a Mixed Podocarp Stand." *Journal of Ecology*, **79**, 1073–1084.

Eckel S, Fleischer F, Grabarnik P, Schmidt V (2008). "An Investigation of the Spatial Correlations for Relative Purchasing Power in Baden-Württemberg." *AStA - Advances in Statistical Analysis*, **92**, 135–152.

Getis A, Franklin J (1987). "Second-Order Neighborhood Analysis of Mapped Point Patterns." *Ecology*, **68**, 473–477.

Goreaud F (2000). *Apports de l'Analyse de la Structure Spatiale en Forêt Tempérée à l'Etude et la Modélisation des Peuplements Complexes*. Ph.D. thesis, ENGREF, Nancy, France.

Goreaud F, Loussier B, Ngo Bieng MA, Allain R (2004). "Simulating Realistic Spatial Structure for Forest Stands: A Mimetic Point Process." In *Interdisciplinary Spatial Statistics Workshop*. Paris.

Goreaud F, Pélissier R (1999). "On Explicit Formula of Edge Effect Correction for Ripley's K-Function." *Journal of Vegetation Science*, **10**, 433–438.

Goreaud F, Pélissier R (2003). "Avoiding Misinterpretation of Biotic Interactions with the Intertype K_{12}-Function: Population Independence vs. Random Labelling Hypotheses." *Journal of Vegetation Science*, **14**, 681–692.

Gourlet-Fleury S, Guehl JM, Laroussinie O (2004). *Ecology and Management of a Neotropical Rainforest. Lessons Drawn from Paracou, a Long-Term Experimental Research Site in French Guiana*. Elsevier, France.

Hardy OJ (2008). "Testing the Spatial Phylogenetic Structure of Local Communities: Statistical Performances of Different Null Models and Test Statistics on a Locally Neutral Community." *Journal of Ecology*, **96**, 914–926.

Lotwick HW, Silverman BW (1982). "Methods for Analysing Spatial Processes of Several Types of Points." *Journal of the Royal Statistical Society B*, **44**, 403–413.

Madelaine C, Pélissier R, Vincent G, Molino JF, Sabatier D, Prévost MF, de Namur C (2007). "Mortality and Recruitment in a Lowland Tropical Rain Forest of French Guiana: Effects of Soil Type and Species Guild." *Journal of Tropical Ecology*, **23**, 277–287.

Nakhede A, Manocha D (1995). "Fast Polygon Triangulation Based on Seidel's Algorithm." In AW Paeth (ed.), *Graphics Gems V*, pp. 394–397. Academic Press.

Oddou-Muratorio S, Demesure-Musch B, Pélissier R, Gouyon PH (2004). "Impacts of Gene Flow and Logging History on the Local Genetic Structure of a Scattered Tree Species, *Sorbus torminalis* L." *Molecular Ecology*, **13**, 3689–3702.

Pavoine S, Ollier S, Pontier D (2005). "Measuring Diversity from Dissimilarities with Rao's Quadratic Entropy: Are Any Dissimilarities Suitable?" *Theoretical Population Biology*, **67**, 231–239.

Pélissier R, Goreaud F (2001). "A Practical Approach to the Study of Spatial Structure in Simple Cases of Heterogeneous Vegetation." *Journal of Vegetation Science*, **12**, 99–108.

Penttinen A, Stoyan D, Henttonen HM (1992). "Marked Point Processes in Forest Statistics." *Forest Science*, **38**, 806–824.

Rao CR (1982). "Diversity and Dissimilarity Coefficients: A Unified Approach." *Theoretical Population Biology*, **21**, 24–43.

Ripley BD (1977). "Modelling Spatial Patterns." *Journal of the Royal Statistical Society B*, **39**, 172–212.

Rowlingson B, Diggle PJ (1993). "**splancs**: Spatial Point Pattern Analysis Code in S-PLUS." *Computers & Geosciences*, **19**, 627–655.

Shen G, Wiegand T, Mi X, He F (2013). "Quantifying Spatial Phylogenetic Structures of Fully-Mapped Plant Communities." *Methods in Ecology and Evolution*, **4**, 1132–1141.

Shen G, Yu M, Hu X, Mi X, Ren H, Sun I, Ma K (2009). "Species-Area Relationships Explained by the Joint Effects of Dispersal Limitation and Habitat Heterogeneity." *Ecology*, **90**, 3033–3041.

Shimatani K (2001). "Multivariate Point Processes and Spatial Variation of Species Diversity." *Forest Ecology and Management*, **142**, 215–229.

Simpson EH (1949). "Measurement of Diversity." *Nature*, **688**, 163.

Stoyan D, Kendall WS, Mecke J (1987). *Stochastic Geometry and Its Applications*. John Wiley & Sons.

Stoyan D, Stoyan H (1994). *Fractals, Random Shapes and Point Fields*. John Wiley & Sons, Chichester.

Thioulouse J, Chessel D, Dolédec S, Olivier J (1997). "**ADE-4**: A Multivariate Analysis and Graphical Display Software." *Statistics and Computing*, **7**, 75–83.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with R*. 4th edition. Springer-Verlag, New York.

Whittaker RH (1972). "Evolution and Measurement of Species Diversity." *Taxon*, **21**, 213–251.

Wiegand T, Moloney KA (2004). "Rings, Circles, and Null-Models for Point Pattern Analysis in Ecology." *Oikos*, **104**, 209–229.

Affiliation:

Raphaël Pélissier
IRD, UMR AMAP
TA A51/PS2
34398 Montpellier cedex 05, France
E-mail: Raphael.Pelissier@ird.fr
URL: http://pelissier.free.fr/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 63, Issue 6 *Submitted:* 2013-06-01
January 2015 *Accepted:* 2014-08-18