Exploratory Analysis of the Habitat Selection by the Wildlife in R: the `adehabitatHS` Package

Clement Calenge,
Office national de la chasse et de la faune sauvage
Saint Benoist – 78610 Auffargis – France.

Feb 2011

Contents

1 History of the package adehabitatHS

The package `adehabitatHS` contains functions dealing with the analysis of habitat selection by the wildlife that were originally available in the package `adehabitat` (Calenge, 2006). The data used for such analysis are generally relocation data collected on animals monitored using VHF or GPS collars, as well as habitat data (available as maps).

I developped the package `adehabitat` during my PhD (Calenge, 2005) to make easier the analysis of habitat selection by animals. The package `adehabitat` was designed to extend the capabilities of the package `ade4` concerning studies of habitat selection by wildlife.

Since its first submission to CRAN in September 2004, a lot of work has been done on the management and analysis of spatial data in R, and especially with the release of the package `sp` (Pebesma and Bivand, 2005). The package `sp` provides classes of data that are really useful to deal with spatial data...

In addition, with the increase of both the number (more than 250 functions in Oct. 2008) and the diversity of the functions in the package `adehabitat`, it soon became apparent that a reshaping of the package was needed, to make its content clearer to the users. I decided to "split" the package `adehabitat` into four packages:

- `adehabitatHR` package provides classes and methods for dealing with home range analysis in R.
- `adehabitatHS` package provides classes and methods for dealing with habitat selection analysis in R.
- `adehabitatLT` package provides classes and methods for dealing with animals trajectory analysis in R.
- `adehabitatMA` package provides classes and methods for dealing with maps in R.

We consider in this document the use of the package `adehabitatHS` to deal with habitat selection analysis. All the methods available in `adehabitat` are also available in `adehabitatHS`. Note that the classes of data returned by the functions of `adehabitatHS` are identical to the classes returned by the same functions in `adehabitat`.

Package `adehabitatHS` is loaded by

```
> library(adehabitatHS)
```

2 Basic concepts

2.1 Use and availability

The package `adehabitatHS` aims at making easier the exploration of habitat selection by the wildlife.

According to a common definition, habitat corresponds to the resources and conditions present in an area that produce occupancy – including survival and reproduction – by a given organism (Hall et al. 1997). The aim of habitat selection studies is often to identify the environmental characteristics (e.g., biomass, slope) that make a place suitable for a species.

Theoretically, the distinction between habitat and non-habitat implies the comparison of the environmental composition of sites where the species is present with the environmental composition of sites where the species is absent. However, sites where the species is absent may be practically difficult to identify. Indeed, a species may not appear in a site if sampling failed to identify it (e.g. detection probability lower than 1), if the species is absent from the site or historical reasons, or if the environmental characteristics do not define a habitat. Therefore, the analysis of habitat selection often relies on the comparison of a sample of used sites (sites where the species is present) with a sample of available sites (sites where the species presence is uncertain, but where we consider that it could be present). We describe several tools in this vignette to deal with such data in this context.

We will consider that the study area can be discretized into *resource units* (RU, which may correspond to pixels of a raster map, or to patches of a vector map, see Manly et al., 2002, for a deeper discussion on this concept). Each RU is characterized by several environmental variables (elevation, slope, biomass, etc.). **In habitat selection studies, the environmental variables should be carefully chosen according to the biological issue at hand**.

We will suppose that we have either censused, or collected a sample of, resource units available to the species on the study area. An important point is that *we consider* that these RUs are available. In other words, the definition

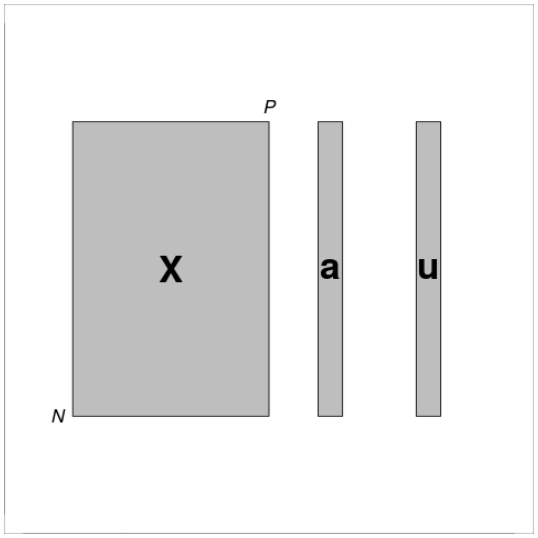
of the availability is necessarily subjective and depends on the issue at hand (it is not a property of the studied system). Each available RU may be characterized by an availability weight describing how the RU is available to the species. For example, these weights may be useful when the RUs correspond to habitat patches and that all patches do not cover the same surface area (this area is then used as an availability weight).

Moreover, we have measured the use of the available RUs by the species. This use may be measured by many different ways (see the concept of "currency of use" by Bingham and Brennan, 2004). For example, it may correspond to the number of animals detected in each pixel of a raster map. We suppose a uniform sampling effort and a uniform detection probability.

2.2 Three types of designs

Thomas and Taylor (1993) distinguished three types of design used in the study of habitat selection.

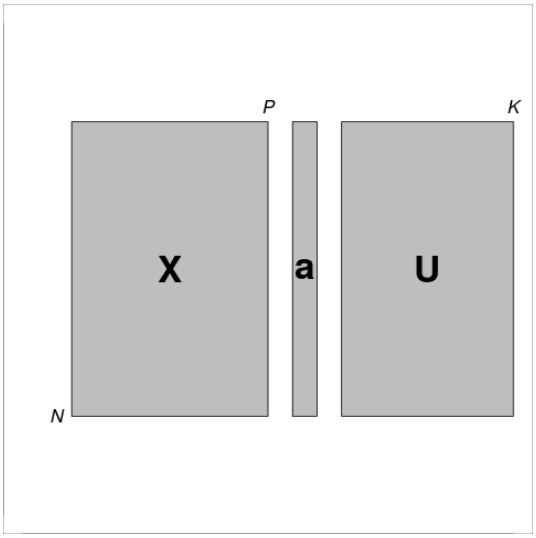
In **design I** studies, the animals are not identified; the habitat use and availability are measured at the scale of the population. For example, a sample of sites (the RUs) is drawn on a given area and each site is investigated to determine whether the species used it (e.g. presence of animals, presence of feces, etc.). That is, the basic data structure for this kind of design is:



where $\mathbf{X}$ is the table containing the value of the P environmental variables for the N RUs, and $\mathbf{a}$ and $\mathbf{u}$ are vectors containing respectively the availability

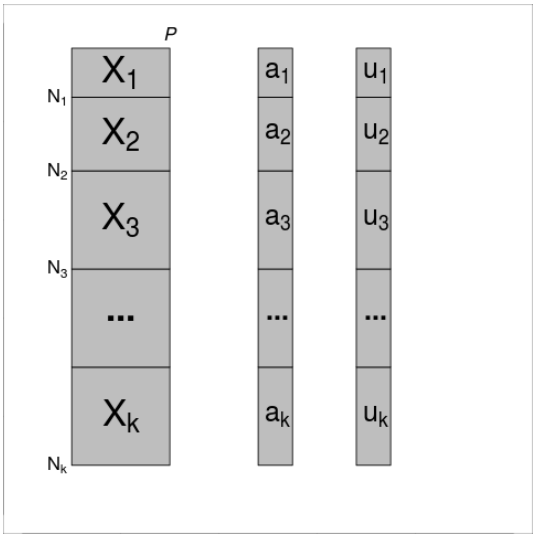
weights and the utilization weights of the N RUs.

In **design II** studies, the animals are identified and the use is measured for each one. For example, a sample of animals is captured on an area, and each one is fitted with a radio-collar. The animals are then monitored using radio-tracking, so that it is possible to estimate the habitat use for each one (e.g. by the proportion of relocations falling in each RU). However, the availability is measured at the scale of the population (each RU is supposed equally available to all monitored animals). The data structure is therefore the following:



where $\mathbf{X}$ is the table containing the value of the P environmental variables for the N RUs, $\mathbf{a}$ is the vector containing the availability weights associated to the n RUs, and $\mathbf{U}$ is the matrix containing the utilization weights of each RU (rows) by the K animals (columns).

In **design III** studies, the animals are identified and both the use and the availability are measured for each one. That is, all the RUs are not supposed to be equally available to all animals, so that the availability weights vary from one animal to the other. For example, a sample of animals is captured on an area, and each one is fitted with a radio-collar. The animals are then monitored using radio-tracking. For each animal, the available RUs correspond to the pixels falling inside the limits of the minimum convex polygon enclosing all its relocations (these available RUs are therefore specific to each animal). The used RUs correspond to the pixels containing at least one relocation, and the utilization weights correspond to the the proportion of relocations falling in each RU:

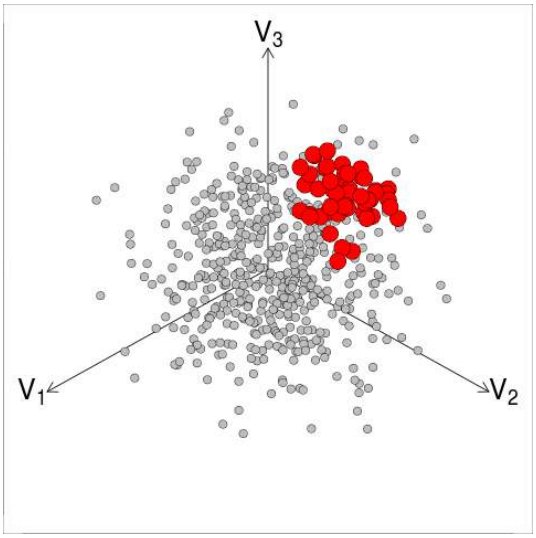


where $\mathbf{X_i}$ contains the value of the P environment variables in the N_i RUs available to the animal i, $\mathbf{a}_i$ contains the availability weights for the animal i, and $\mathbf{u}_i$ contains the utilization weights for the animal i.

2.3 The concept of ecological niche

The ecological niche is a useful model to understand the methods available in `adehabitatHS`. The seminal paper of Hutchinson (1957) defined it as the hypervolume, in the multidimensional space defined by the environmental variables, where the species can potentially maintain a viable population. Although this definition has been debated and precised by many authors (e.g. Chase and Leibhold, 2003), the original definition of the niche is useful for us as most ecologists are already familiar with it. Therefore, we use this definition beyond the limits of the original conceptual framework in which it was developed, i.e. to study habitat selection.

For example, imagine that we are studying habitat selection of a species on a given area. Moreover, P environmental variables have been mapped on this area. Suppose that these maps are raster maps, so that the N RUs correspond to the pixels of the maps. Suppose that we have prospected the study area so that we know the position of all the individuals belonging to a given species. The P environmental variables define a P-dimensional space. We will refer to this space as the *ecological space*. Because each RU is characterized by a measure on each environmental variable, it follows that each RU corresponds to a point in this space:



On this figure, the ecological space is defined by only three variables (V_1, V_2, V_3), but the mathematical concept is still valid for a larger number of variables. The grey points correspond to the RUs available on the study area. A subset of RUs have been used by the species (i.e. the species is present in it; red points on the figure). We will refer to the corresponding subset of points in the ecological space as the "niche" of the species on this area.

The concept of niche, as defined here, is a useful model to tackle the study of habitat selection. Indeed, many tools available in the package `adehabitatHS` are implementations of methods allowing to compare the shape of the niche with the shape of the distribution of available points. Although the methods allow to take into account unequal utilization weights and unequal availability weights, thinking in terms of "niche" as defined here (same utilization weights for all used RUs and same availability weights for all available RUs) is a very useful way to understand the methods provided by `adehabitatHS`.

2.4 Marginality and specialization

Two concepts are useful for the study of the niche: the marginality and the specialization. Consider the niche model described in the previous section:

- The **marginality vector** correspond to the vector connecting the centroid (i.e., the mean) of the distribution of availability weights to the centroid of the distribution of utilization weights. The squared length of this vector is the **marginality** *per se*. For example, consider the data structure corresponding to design I studies (section 2.2). The marginality vector is calculated by $\mathbf{m} = \mathbf{X}^t\mathbf{u} - \mathbf{X}^t\mathbf{a}$. And the marginality is equal to

$m^2 = ||\mathbf{m}||^2$.

- The **specialization** is a measure of habitat selection on a particular direction of the ecological space. For example, consider the variable y_i giving the values of an environmental variable in the RU i. The specialization is defined as:

$$S^2 = \frac{\sum_{i=1}^{N} a_i (y_i - \sum_{j=1}^{N} a_i y_i)^2}{\sum_{i=1}^{N} u_i (y_i - \sum_{j=1}^{N} u_i y_i)^2}$$

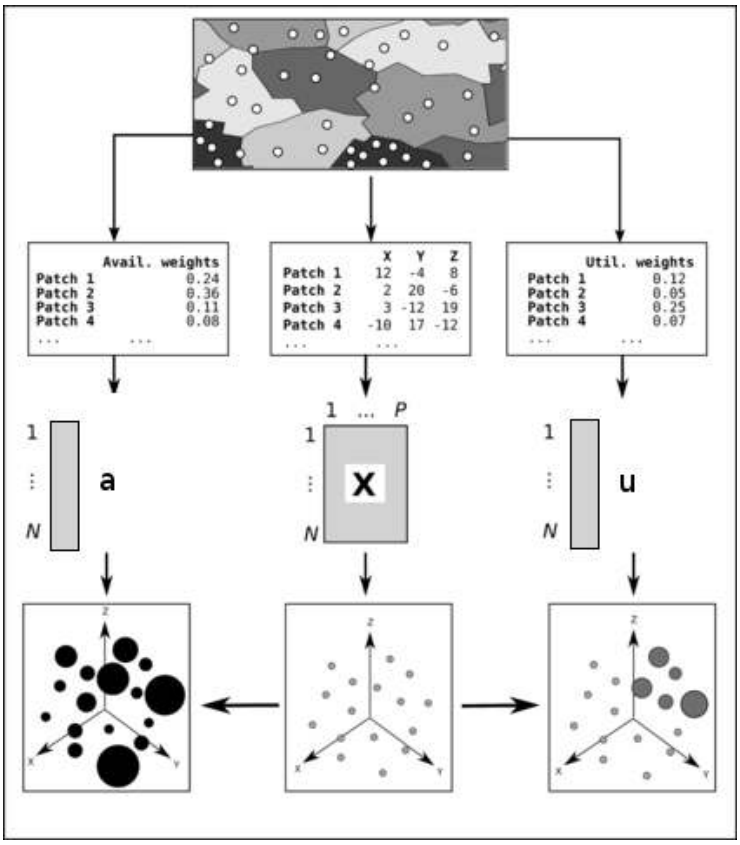
  where a_i and u_i are respectively the availability weight and the utilization weight of the RU i. In other words, the specialization is the ratio of the variance of availability weights divided by the variance of the utilization weights. A strong specialization implies that the variance of available RUs is large in comparison with the variance of used RUs. Or, in other words, that the niche is narrow in comparison to what is available to the species.

We will make use of these two concepts in this vignette.

3 Design I studies

3.1 Basic approach

The aim of `adehabitatHS` is to provide tools for the exploration of habitat selection. We have seen in section 2.2 that the basic data structure for design I studies is built by a table $\mathbf{X}$ containing the value of the environmental variables in each available RU, and vectors $\mathbf{a}$ and $\mathbf{u}$ containing respectively the availability weights and the utilization weights for each RU. That is, our data structure is the following:



Each RU defines a point in the multidimensional space defined by the environmental variables (ecological space). To each point is associated an available weight (in the vector **a**) and an utilization weight (in the vector **u**). The set of points for which the utilization weights are greater than 0 define the niche of the species (as defined in section 2.3). Our aim is to identify the differences between the two distributions of weights, and to relate these differences with particular directions of the ecological space.

To identify these differences, we will use graphical methods. Indeed, as noted by Cleveland (1993), "*Visualisation is critical to data analysis. It provides a front line of attack, revealing intricate structure in data that cannot be absorbed in any other way. We discover unimagined effects, and we challenge imagined ones*". However, because the ecological space is generally highly multidimensional, it is hard to explore it with classical graphical methods. For this reason, the package `adehabitatHS` provides several tools relying on factor analysis, to find the most "interesting" directions on which most of the differences between the two distributions of weights are expressed.

3.2 The general framework for the statistical exploration of the niche

3.2.1 Presentation of the GNESFA

All the factor analyses available in `adehabitatHS` to explore the differences between the habitat use and habitat availability in design I studies can be viewed as particular cases of a general framework named *General Niche-Environment System Factor Analysis* (GNESFA). This framework is described in detail in Calenge and Basille (2008).

The basic principle of this analysis consists in the choice of one of these two distributions, either the utilization weights or the availability weights, as the Reference distribution, and the other, as the Focus distribution.

One will then "distort" the cloud of points, so that the Reference distribution takes a standard spherical shape in the multidimensional space (this is sometimes named "sphering" the data). Then, the GNESFA searches for the directions where the Focus distribution differs the most from this standard spherical shape (by performing a noncentred principal component analysis of the "sphered" data, using the Focus distribution as row weight). This analysis finds the direction on which the following criterion in maximized:

$$\gamma_1 = \frac{\sum_{i=1}^{N} f_i (y_i - \bar{y}_r)^2}{\sum_{i=1}^{N} r_i (y_i - \bar{y}_r)^2}$$

f_i and r_i are respectively the focus and reference weights (depending on which distribution of weights – utilization or availability – has been chosen as a reference). y_i is the score of the i^{th} RU on the first axis of the GNESFA. And, finally, $\bar{y}_r$ is the "reference mean" of these scores, that is:

$$\bar{y}_r = \sum_{i=1}^{N} r_i y_i$$

In other words, the GNESFA searches the values $b_1, b_2, ...b_p$ such that:

- $y_i = b_1 X_{i1} + b_2 X_{i2} + ... + b_p X_{ip}$
- $\sum_{i=1}^{p} b_i^2 = 1$
- γ_1 is maximized

and such that the successive axes are uncorrelated. The GNESFA provides a very flexible framework to tackle the exploration of habitat selection, and the analysis has properties that depends on the distribution chosen as a reference. Calenge and Basille (2008) describes more completely the interesting properties

of this approach.

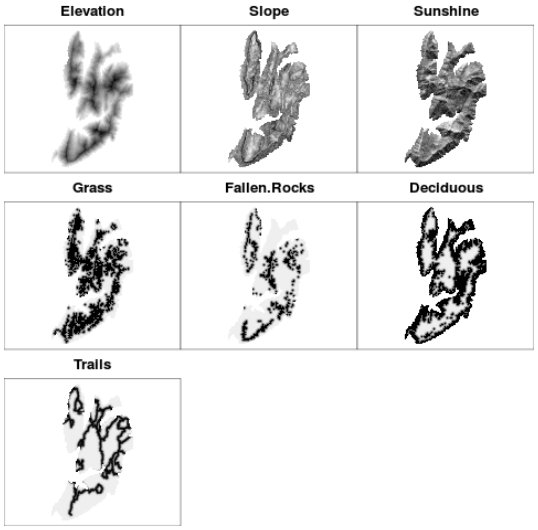
In this vignette, we will illustrate these properties and the practical use of the GNESFA with an example. First load the dataset `bauges` from the package `adehabitatHS`:

```
> data(bauges)
> names(bauges)
```

```
[1] "map"  "locs"
```

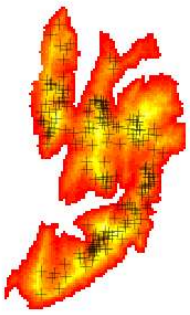
This data set contains the map of 7 environmental variables in the Bauges Mountains (French Alps):

```
> map <- bauges$map
> mimage(map)
```



The maps are stored as an object of class `SpatialPixelsDataFrame` from the package `sp`. Furthermore, this dataset also contains the relocations of 198 chamois groups collected by volunteers and professionals working in various French wildlife and forest management, from 1994 to 2004:

```
> image(map)
> locs <- bauges$locs
> points(locs, pch=3)
```



These relocations are stored as an object of class `SpatialPointsDataFrame` from the package `sp`. We now compute the utilization weights associated to each pixel of the map, by numbering the locations of chamois groups in each pixel of the map. This is done using the function `count.points` from the package `ademaps`:

```
> cp <- count.points(locs, map)
```

Because all the pixels cover the same area, we will consider that they all have the same availability weights (i.e., $\frac{1}{N}$). Now, we "unspatialize" the required elements:

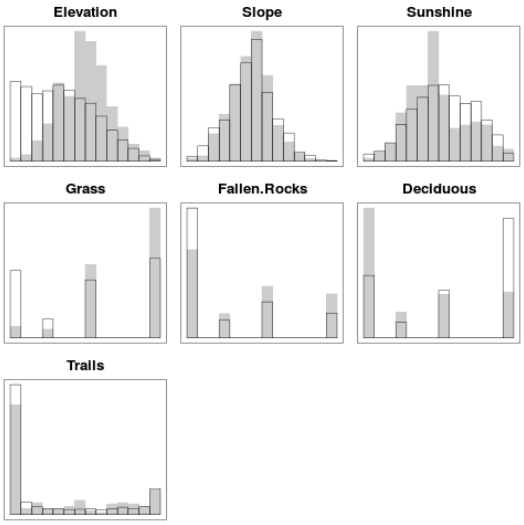
```
> tab <- slot(map, "data")
> pr <- slot(cp, "data")[,1]
```

These two elements are:

- `tab`: a data frame containing the value of the environmental variables in each pixel of the map;
- `pr`: a vector containing the utilization weights associated to each pixel;

First, let us have a look at the distribution of the environmental variables on the area, as well as the distribution of the animals:

```
> histniche(tab, pr)
```



The white histograms show the distributions of available RUs, whereas the grey histograms show the distributions of used RUs. We already highlight a strong selection for high elevation values, high grass cover, high density of fallen rocks and low deciduous cover. Let us perform a GNESFA to summarize these structures.

3.2.2 A preliminary `dudi.*` analysis

We use for this the function `gnesfa`:

```
> args(gnesfa)

function (dudi, Focus, Reference, centering = c("single", "twice"),
    scannf = TRUE, nfFirst = 2, nfLast = 0)
NULL
```

The help page of this function indicates that it takes as main argument an object of class `dudi`. This class is defined in the package `ade4` (Chessel et al. 2004), and is designed to store the results of factor analyses provided by this package. **We use the functions of the package `ade4` as a preliminary step to prepare the data tables for the analysis**. Three main functions are of interest for us:

- `dudi.pca`: performs a principal component analysis of the data frame passed as argument;
- `dudi.acm`: performs a multiple correspondence analysis of the data frame passed as argument (Tenenhaus and Young, 1985);

- `dudi.hillsmith`: performs a Hill-Smith analysis of the data frame passed as argument (Hill and Smith, 1976);

The function `dudi.pca` is to be used when all the variables present in the data.frame are numeric. The function `dudi.acm` is to be used when all the variables present in the data.frame are factors. The function `dudi.hillsmith` (or, equivalently, `dudi.mix`) is to be used when the data.frame contains both types of variables. These functions, used as a preliminary to the GNESFA, are needed to scale the table suitably (so that all the variables have the same mean and the same variance), and to compute the weights of the variables in the analysis. For example, the use of `dudi.hillsmith` on a table containing a numeric variable and a factor with four levels ensures that the factor will have the same weight in the analysis as the numeric variable. Chessel et al. (2004) give a full description of all the possible preliminary `dudi.*` analysis.

The result of the functions `dudi.*` is a list with a component `$tab` containing the table scaled in a suitable way, and a component `$cw` containing the weights of the variables. The function `gnesfa` only uses these components. However, we will see later that other functions of the package also make use of the component `$lw` for the definition of the row weight of the analysis.

Remark: In many cases, it is extremely useful to interpret the results of these preliminary analyses (this allows to identify the main patterns on the study area, even if these patterns are not necessarily related to habitat selection). The `dudi.*` analyses are of interest in themselves. However, their use is already detailed elsewhere (e.g. Chessel et al. 2004), so that we will not describe them in this vignette.

Back to our example: we will use the function `dudi.pca` to prepare the table, because all our variables are numeric (in this particular case, this allows to center and scale the table, like the function `scale`):

```
> pc <- dudi.pca(tab, scannf=FALSE)
```

3.2.3 The FANTER

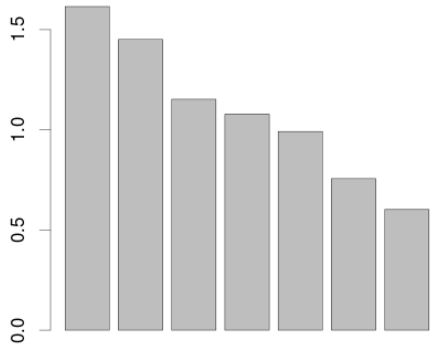
We now perform the GNESFA. We first have to choose one of the two distribution of weights (utilisation weights or availability weights) as the reference distribution and the other as the Focus distribution. Depending on this choice, the result will not be the same. We will first consider the choice:

- Reference = availability
- Focus = utilization

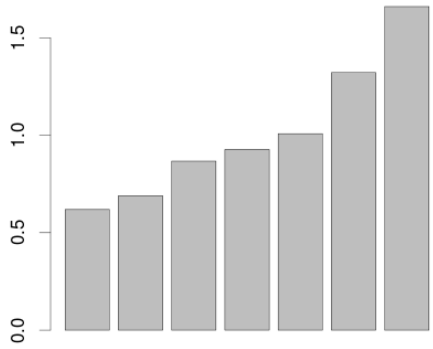
The GNESFA, taking the availability distribution as the reference is also named FANTER (Factor Analysis of the Niche, Taking the Environment as the Reference). We perform the analysis:

```
> gn <- gnesfa(pc, Focus = pr)
Select the first number of axes (>=1): 1
Select the second number of axes (>=0): 1
```

The function asks the user to choose the number of first and last axes to choose. Indeed, when the Reference distribution is the availability weights and the Focus is the utilisation weights, both the first axes and the last axes of the GNESFA may have a meaning. The first axes correspond to the directions where the utilization distribution as a whole is the furthest from the centroid of the availability distribution (these directions often correspond to directions where the *marginality* is strong; i.e. the criterion γ_1 defined previously is maximized). The last axes correspond to the directions where the *width* of the utilization distribution is the smallest relative to the width of the availability distribution (these directions often correspond to directions where the *specialization* is strong, i.e. the criterion γ_1 defined previously is minimized). The first barplot showed by the function correspond to the eigenvalues of the analysis:



Note that it is hard to identify a clear break in the decrease of the eigenvalues. The analysis fails to identify a clear pattern here... The second barplot corresponds to 1/eigenvalues, and allows to see more clearly the possible patterns on the last axes:

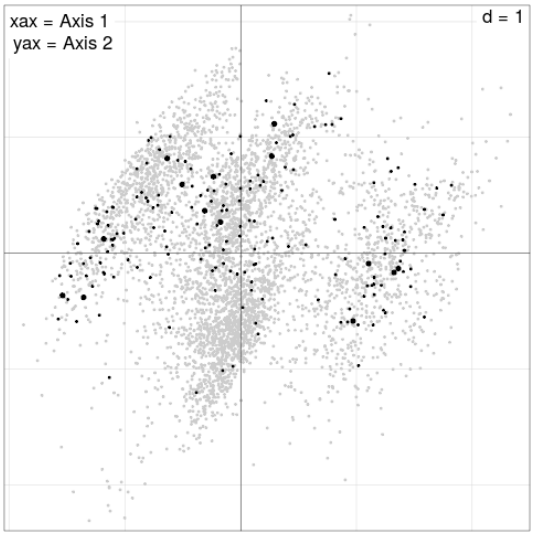


No clear pattern appears on the last axes (no clear break in the increase of 1/eigenvalues)... This analysis does not find any interesting direction in the ecological space. Have a look at the results:

```
> gn

GNESFA
$call: gnesfa(dudi = pc, Focus = pr, scannf = FALSE, nfFirst = 1, nfLast = 1)

$centering: single
eigenvalues: 1.615 1.451 1.153 1.078 0.9917 ...
$nfFirst: 1 first axes saved
$nfLast: 1 last axes saved

  vector     length mode    content
1 $Reference 4628   numeric Weighting matrix of reference distribution
2 $Focus     4628   numeric Weighting matrix of focus distribution
3 $eig       7      numeric eigen values of specialization

  data.frame nrow ncol content
1 $tab       4628 7    modified array
2 $li        4628 2    row coordinates
3 $l1        4628 2    row coordinates (variance weighted by $Reference =1)
4 $co        7    2    column coordinates
5 $cor       7    2    correlation between variables and axes
```

Just for the sake of illustration, we can have a look at the niche on the factorial plane built by the first and last axis of the analysis:

```
> scatterniche(gn$li, pr, pts=TRUE)
```



The grey points show the distribution of the RUs (here the pixels) on the axes found by the analysis. The black points correspond to the RUs used by the chamois. We can see that the used points are distributed over the whole range of the factorial axes found by the analysis. This confirms that the analysis fails to identify any intesting direction.

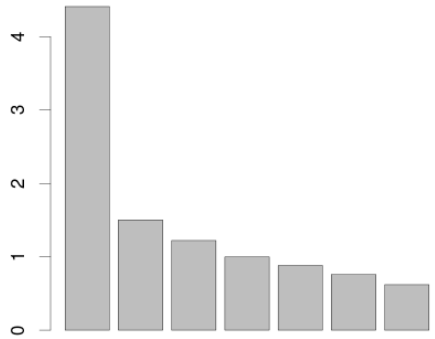
3.2.4 The MADIFA and Mahalanobis distances

But now, consider the opposite point of view, that is:

- Reference = utilization
- Focus = availability

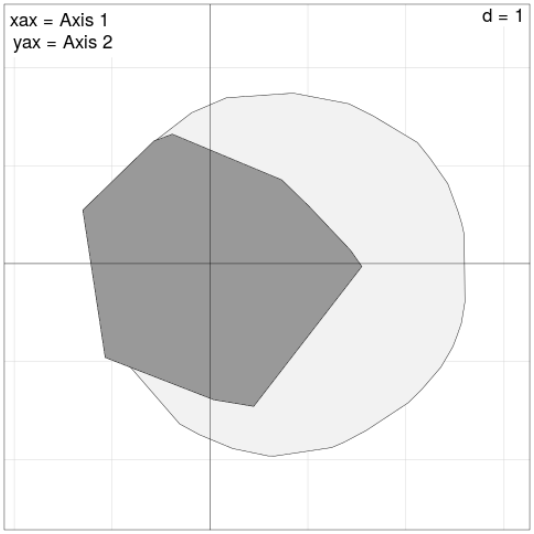
In this case, the last axes of the analysis do not have any biological meaning so that we do not consider them (see Calenge and Basille 2008):

```
> gn2 <- gnesfa(pc, Reference = pr)
Select the first number of axes (>=1): 2
Select the second number of axes (>=0): 0
```



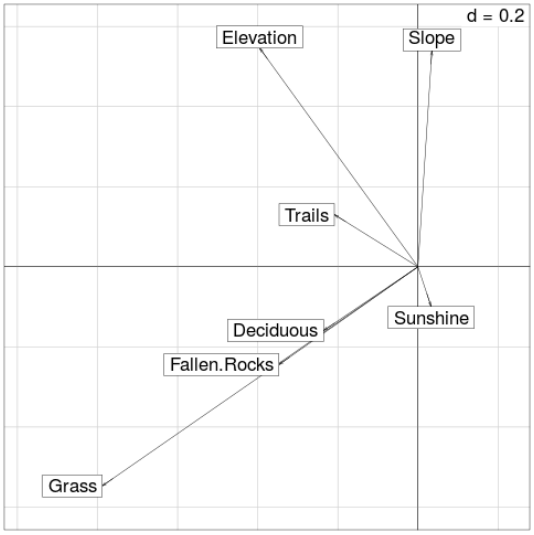
In this case, there is a clear break in the decrease of the eigenvalues. The first axis of the analysis identifies a strong pattern. We can have a look at the niche of the chamois on the first factorial plane found by the analysis:

```
> scatterniche(gn2$li, pr)
```



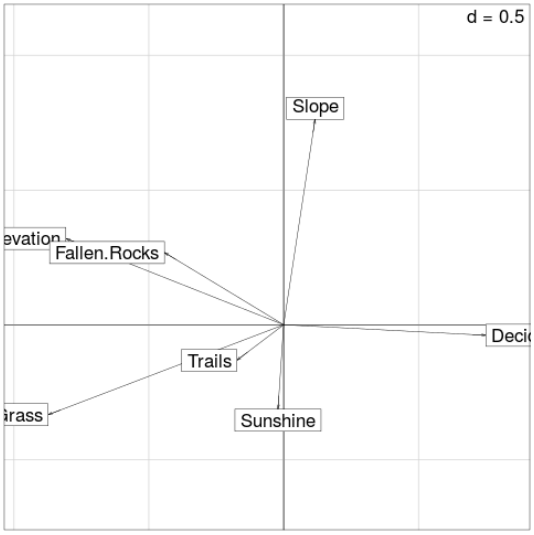
Whereas the utilization distribution is centred on the origin of the space (as we defined it as the reference), the availability distribution is strongly skewed toward the positive values of the first axis. We may have a look at the coefficients of the environmental variables in the definition of this axis to give a meaning to it:

```
> s.arrow(gn2$co)
```



The grass cover seems to be the main variable affecting the distribution of the chamois, and to a lesser extent, the elevation and presence of Fallen Rocks. However, these variables are correlated on the study area (grass and fallen rocks occur at high elevation, whereas deciduous cover occurs at low elevation). This may render the interpretation of the coefficients difficult in an exploratory context. For this reason, we prefer to give a meaning to the axes of the GNESFA with the help of the correlations between the environmental variables and the axes of the GNESFA:

```
> s.arrow(gn2$cor)
```



We now see why interpreting the results with the coefficient may pose problems: the first axis of the GNESFA is negatively correlated with the grass cover, the elevation, and positively correlated with the deciduous cover. This did not appeared with the graph showing the columns coefficients (on which deciduous even had a negative coefficient!).

Therefore, positive values of the first axis correspond to areas at low elevation with a high deciduous cover and a low grass cover, and, to a lesser extent, with a low density of fallen rocks. These areas are rarely used by the chamois, which prefers area at high elevation, with high grass cover and close to the fallen rocks.

Remark: in this case, the results are not great biological discoveries (we demonstrate that the chamois lives in the mountains!). Indeed, the dataset used to illustrate the methods has been slightly destroyed to preserve copyright (for the original analysis of this dataset, see Calenge et al. 2008). However, this dataset has an interesting property: choosing the availability or the utilization as the Reference distribution does not return the same results.

The GNESFA applied with the availability distribution as the focus is also named MADIFA (Mahalanobis Distances factor analysis, Calenge et al. 2008). Note that this function can also be applied using the function `madifa` (see the help page of this function):

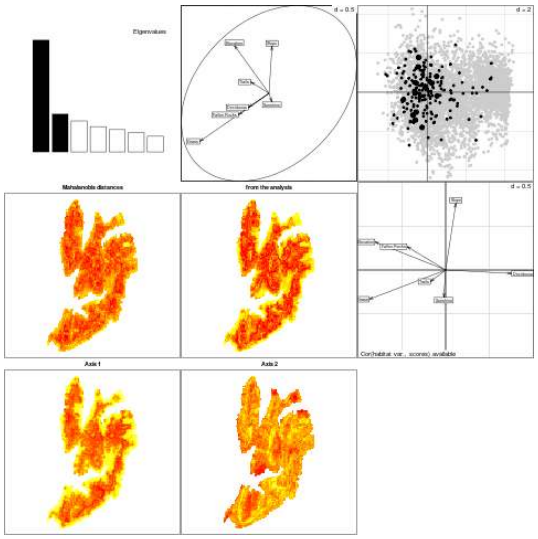
```
> (mad <- madifa(pc, pr, scan=FALSE))

MADIFA
$call: madifa(dudi = pc, pr = pr, scannf = FALSE)

eigen values: 4.411 1.5 1.221 0.9983 0.8804 ...
$nf: 2 axes saved

  vector  length mode    content
1 $pr     4628   numeric vector of presence
2 $mahasu 4628   numeric squared Mahalanobis distances
3 $lw     4628   numeric row weights
4 $cw     4628   numeric column weights
5 $eig    7      numeric eigen values

  data.frame nrow ncol content
1 $tab       4628 7    modified array
2 $li        4628 2    row coordinates
3 $l1        4628 2    row normed scores (variance weighted by $pr = 1)
4 $co        7    2    column coordinates
5 $cor       7    2    cor(habitat var., scores) for available points
```

Note that the results returned by this function are identical to the results of the function `gnesfa`:

```
> mad$eig

[1] 4.4105514 1.5003687 1.2214409 0.9983498 0.8804264 0.7608612 0.6191104

> gn2$eig

[1] 4.4105514 1.5003687 1.2214409 0.9983498 0.8804264 0.7608612 0.6191104
```

However, more methods are provided by `adehabitatHS` to deal with the results of the function `madifa`. For example, a full summary of the analysis can be obtained with:

```
> plot(mad, map)
```



This figure presents:

- the eigenvalue diagram of the analysis;
- the coefficients of the variables in the definition of the axis;
- the plot of the niche on the factorial axes;
- the map of the Mahalanobis distances (see below);
- the map of the approximate Mahalanobis distances computed from the axes of the analysis;
- the correlation between the environmental variables and the axes of the analysis;

- the maps of the scores of the pixels of the map on the axes of the analysis.

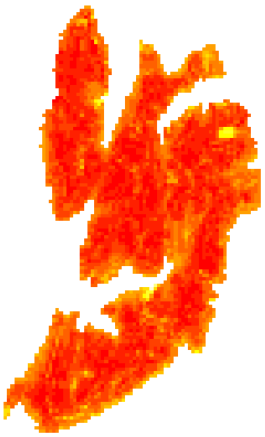
Remark: by default, the MADIFA is carried out with equal availability weights for all RUs when the function `madifa` is used. However, unequal availability weights can also be defined with this function. To proceed, the user should pass the vector of availability weights as row weights to the function `dudi.*` performed as a preliminary step (see previous section). In our example, if we had a vector named `av.w` containing the availability weights of the RUs, the preliminary PCA could have been performed with the following code:

```
> pc <- dudi.pca(tab, row.w = av.w, scannf=FALSE)
```

(not executed here).

The MADIFA, as its names indicates, is closely related to the Mahalanobis distances methodology introduced by Clark et al. (1993) in Ecology. They proposed to use the Mahalanobis distance between a pixel and the distribution of utilization weights in the ecological space as a measure of habitat suitability of this pixel for the species. A map of these Mahalanobis distances can be computed by:

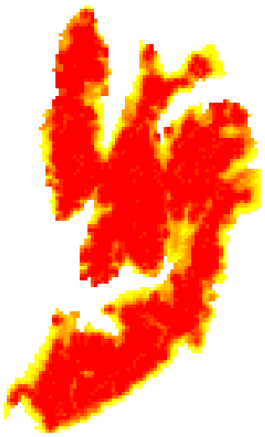
```
> MD <- mahasuhab(map, locs)
> image(MD)
```



As we can see, this map is very noisy. The MADIFA provides a way to remove a part of this noise. Actually, the MADIFA finds the direction in the ecological space where the average squared Mahalanobis distances between the available RUs and the distribution of utilization weights is the largest (this is

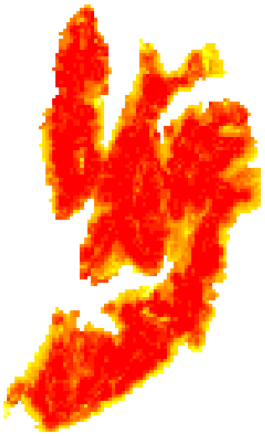
the meaning of the criterion γ_1 when the utilization weights are chosen as the reference). It is then possible to compute a reduced-rank Mahalanobis distance between pixels and this utilization from the results of the analysis. Practically, this is done by summing the squared scores of the pixels on the successive axes of the analysis. For example, for a map based on only one axis:

```
> RMD1 <- data.frame(RMD1=mad$li[,1]^2)
> coordinates(RMD1) <- coordinates(map)
> gridded(RMD1) <- TRUE
> image(RMD1)
```



This map is much clearer than the previous one. For the sake of the illustration, to compute reduced rank Mahalanobis distances based on two axes:

```
> RMD2 <- data.frame(RMD2 = apply(mad$li[,1:2], 1, function(x) sum(x^2)))
> coordinates(RMD2) <- coordinates(map)
> gridded(RMD2) <- TRUE
> image(RMD2)
```



This second map is more noisy than the first. We would keep the first one to describe habitat selection.

More details about the MADIFA and the GNESFA can be found on the help pages of the functions `madifa` and `gnesfa`.

3.2.5 The ENFA

The Ecological niche factor analysis has been proposed by Hirzel et al. (2002) for habitat suitability mapping purposes. Basille et al. (2008) have showed that this analysis could be used efficiently to explore the ecological niche. The basic principle of this analysis is the following:

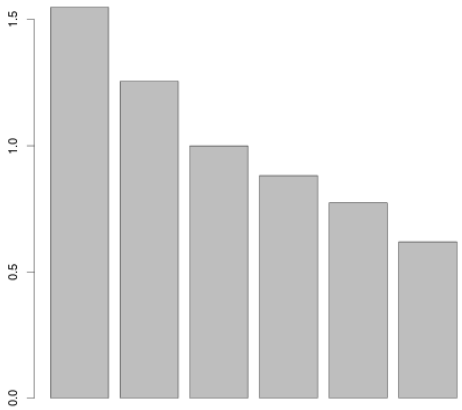
- First compute the marginality vector (connecting the centroid of the distribution of availability weights to the centroid of the distribution of utilization weights);
- Then project the cloud of RUs on the hyperplane orthogonal to the marginality vector;
- Find the directions in this subspace on which the specialization (ratio variance of the distribution of availability weights / variance of the distribution of utilization weights) is the largest. These axes are named "specialization axes"

Basille et al. (2008) have showed that the ENFA can also be viewed as a particular case of the GNESFA. Actually, once the data have been projected on the hyperplane orthogonal to the marginality vector, the GNESFA of the data table is identical to the ENFA, whatever the distribution chosen as a Reference (in the case where the availability is chosen as the reference, the last axes maximize the specialization; in the case where the utilization is chosen as the reference, the first axes maximize the specialization). The ENFA can be carried out with the function `enfa` or the function `gnesfa`:

```
> en1 <- enfa(pc, pr, scan=FALSE)
> gn3 <- gnesfa(pc, Reference=pr, scan=FALSE, centering="twice")
> en1$s

[1] 1.5488351 1.2557293 0.9996892 0.8818026 0.7738212 0.6191188

> gn3$eig

[1] 1.5488351 1.2557293 0.9996892 0.8818026 0.7738212 0.6191188
```

The eigenvalues diagram is presented below:

```
> barplot(en1$s)
```

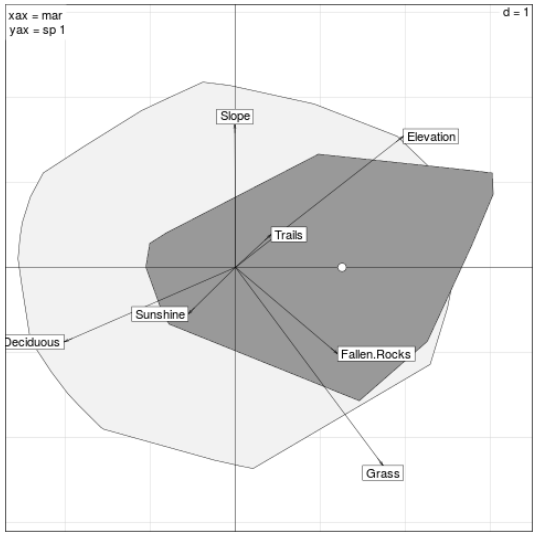


The ENFA fails to identify any specialization pattern in the data. However, remember that the ENFA is an analysis of the table projected on the hyperplane orthogonal to the marginality vector. There may be an interesting pattern in the direction of the marginality vector, which is therefore not expressed in the directions found by the ENFA. For this reason, it is essential to consider the projection of the data table on the marginality vector as well as the projections of the data table on the directions of the specialization axes (see Basille et al.

2008).

Basille et al. (2008) showed that a biplot (Gabriel, 1971) can be used to show both the variables and the RU scores on the plane formed by the marginality vector (X direction) and any specialization axis (Y direction). This biplot can be drawn by:

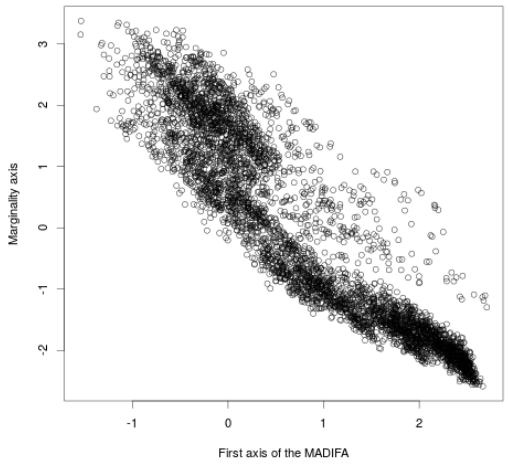
```
> scatter(en1)
```



This biplot is the main result of the ENFA. The dark grey polygon shows the position of the distribution of used RUs, whereas the light grey polygon displays the position of the distribution of available RUs. The abscissa is the marginality axis (the direction where the centroid of the distribution of utilization weights – displayed by a dot – is the furthest from the centroid of the distribution of available weights – the origin of the axes), and the ordinate is the first specialization axis (the direction where the variance of the utilization distribution is the smallest relative to the variance of the availability distribution).

The eigenvalue diagram indicated that the ENFA failed to identify a direction where the specialization is high (it is clear on the previous figure that the variance of the niche on the y-direction is not much smaller than the variance of the distribution of available RUs) . However, we can see that the main pattern in the dataset (identified by the MADIFA) is expressed by the marginality axis. The results of the ENFA are therefore consistent with the results returned by the MADIFA. Indeed the marginality axis is strongly correlated with the first axis of the MADIFA:

```
> plot(mad$li[,1], en1$li[,1], xlab="First axis of the MADIFA",
+      ylab="Marginality axis")
```



Remark: by default, the ENFA is carried out with equal availability weights for all RUs when the function `enfa` is used. However, unequal availability weights can also be defined with this function. To proceed, the user should pass the vector of availability weights as row weights to the function `dudi.*` performed as a preliminary step (see section 3.2.2). In our example, if we had a vector named `av.w` containing the availability weights for each RU, the preliminary PCA could have been performed with the following code:

```
> pc <- dudi.pca(tab, row.w = av.w, scannf=FALSE)
```

(not executed here).

3.2.6 Conclusions regarding the GNESFA

In practice, the three analyses taking place within the framework of the GNESFA often return the same results. However, this is not always the case, as demonstrated by our example. Taking the utilization weights as the Focus distribution (the FANTER) does not allow to identify any pattern in the data here. Actually, even if the FANTER tends to maximize the marginality on the first axes and the specialization on the last axes, it does not explicitly make a distinction between the marginality and the specialization (for more formal details, see Calenge and Basille, 2008).

The MADIFA identifies a strong pattern in the data, and the ENFA indicates that this pattern is essentially a marginality axis. So, we may wonder why it does not appear on the first axis of the FANTER... Actually, on the first axis of the MADIFA, both the marginality and the specialization are strong. That is, on the marginality axis, the ratio (used variance)/(available variance)

is very low. Therefore, because both parameters (specialization and marginality) are strong, the FANTER fails to disentangle the information carried by two measures... and does not identify any pattern. By forcing the extraction of the marginality axis, the ENFA identifies this direction, and by combining marginality and specialization into a single measure (the average Mahalanobis distances), the MADIFA identifies this direction.

We should not conclude from this example that the FANTER is useless. Each method has particular properties that are not shared by the other methods. Thus, the FANTER is the only analysis of this framework allowing to identify bimodal niches (as demonstrated by Calenge and Basille 2008). The MADIFA is the only analysis combining the marginality and the specialization into a unique measure of habitat selection. The ENFA is the only analysis distinguishing formally the marginality and the specialization. The three methods provide three complementary points of view on the ecological space.

3.2.7 An alternative analysis proposed by James Dunn

During an e-mail discussion related to the MADIFA, James Dunn (formerly University of Arkansas) proposed an alternative analysis with several interesting properties. Because this analysis is unpublished, we reproduce the derivation of the analysis in the appendix (with the authorization of Pr. Dunn). This analysis finds the direction where the specialization is maximized, *without projecting the data orthogonally to the marginality vector*. Thus, the first axis of the analysis maximizes the ratio:

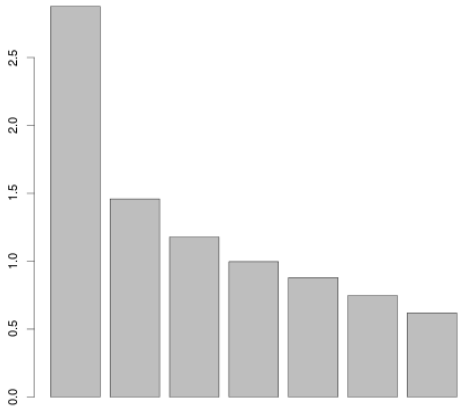
$$\frac{\text{Variance of the availability distribution}}{\text{Variance of the utilization distribution}}$$

and does not consider the marginality as a separate parameter (actually, it does not consider the marginality at all!). Thus, if both the marginality and the specialization are strong in a direction, this direction will be found by this analysis. We have programed this approach in the function `dunnfa` of the package `adehabitatHS`. Note that this function does not (yet) allow to define unequal availability weights for the RUs. We can try this approach on the `bauges` dataset:

```
> dun <- dunnfa(pc, pr, scann=FALSE)
```

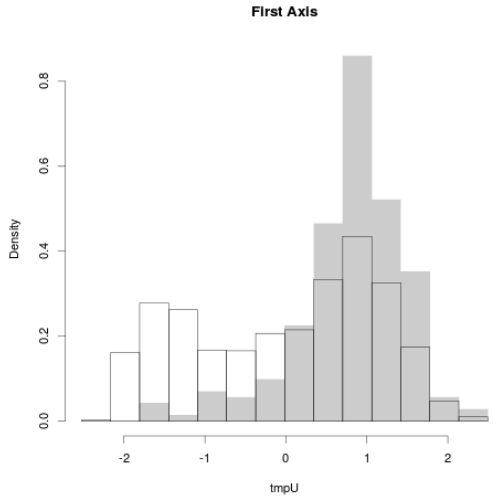
Have a look at the eigenvalues:

```
> barplot(dun$eig)
```



There is a clear break in the decrease of the eigenvalues after the first one. The first axis therefore expresses a clear pattern. We can have a look at the niche of the species on this first axis:

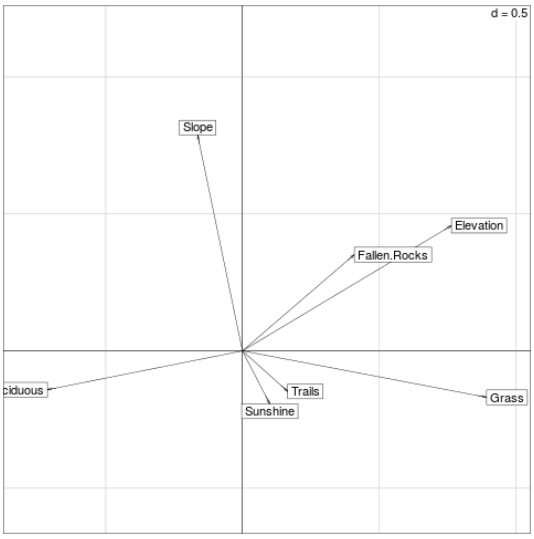
```
> histniche(data.frame(dun$liA[,1]), pr, main="First Axis")
```



The white histogram shows the distribution of the available RUs, whereas the grey histogram corresponds to the distribution of the used RUs. We can see that the niche of the species is positively skewed on the first axis: the chamois searches for high values of the first axis.

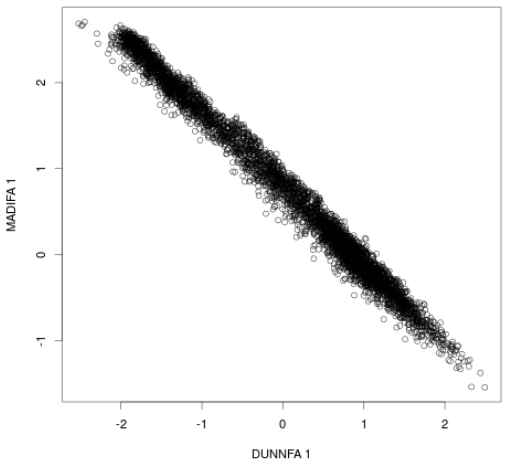
Similarly to the GNESFA, we can interpret the meaning of the axis with the help of the coefficients of the variables or with the correlations between the variables and the scores of the available RUs on the first axis. We choose the latter:

```
> s.arrow(dun$cor)
```



We find again the structure highlighted on the first axis of the MADIFA and the marginality axis of the ENFA. There is indeed a very strong correlation between the first axis of the MADIFA and the first axis of this analysis:

```
> plot(dun$liA[,1], mad$li[,1], xlab="DUNNFA 1", ylab="MADIFA 1")
```



Note that the DUNNFA also allows to compute reduced ranks Mahalanobis distances (see appendix). Presently, the DUNNFA has been implemented only for the data.frames containing only numeric variables (the preliminary `dudi.*` analysis should be `dudi.pca`).

3.3 One word about habitat suitability maps

In the previous sections, we already have seen two methods designed to map habitat suitability for a species: the Mahalanobis distances (Clark et al. 1993, implemented in `mahasuhab`) and the reduced ranks Mahalanobis distances (computed with the help of the MADIFA or the DUNNFA).

We now stress that the package adehabitatHS is not designed for predictive purpose, but rather for exploratory purposes. There are many packages available for habitat prediction in R (see the CRAN Task view: `http://cran.rproject.org/web/views/Environmetrics.html`). The methods in `adehabitatHS` allowing habitat suitability mapping are there as a recognition that visual exploration of such maps may bring insight into the processes at work on the area. Except for the methods related to the Mahalanobis distances, the package `adehabitatHS` provides only the DOMAIN algorithm for such mapping (Carpenter et al. 1993, implemented in the function `domain`).

3.4 When habitat is defined by several categories

A very common approach to the study of habitat selection consists in defining several habitat categories on the study area and comparing the use and availability of each habitat category by the species. For example, let us consider the elevation map on the study area. Define four classes of elevation (using the R function `cut`):

```
> elev <- map[,1]
> av <- factor(cut(slot(elev, "data")[,1], 4),
+              labels=c("Low","Medium","High","Very High"))
```

The number of pixels in each class is:

```
> (tav <- table(av))

av
      Low    Medium      High Very High
     1775      1657       995       201
```

Let us map this variable:

```
> slot(elev, "data")[,1] <- as.numeric(av)
> image(elev)
```

Now, compute the percentage of use of each habitat class by the chamois. That is, we compute the number of chamois detections in each habitat class:

```
> us <- join(locs, elev)
> tus <- table(us)
> names(tus) <- names(tav)
> tus

      Low    Medium      High Very High
       11        89        84        14
```

To study the habitat selection of the chamois, we have to compare the use and availability of each habitat class. Manly et al. (2002) provide a methodology for this kind of design, relying on the calculation of selection ratios:

$$w_j = \frac{u_j}{a_j}$$

where u_j is the proportion of use of the habitat class j and a_j is the proportion of availability of this habitat class j. Manly et al. (2002) present the use of these selection ratios in an inferential context, but such ratios are also useful in exploratory contexts. Note that these ratios may be scaled so that their sum is equal to 1, that is:

$$B_j = \frac{w_j}{\sum_i w_i}$$

The function `widesI` implements the approach of Manly et al. (2002; more details are presented in the example section of the help page of this function):

```
> class(tus) <- NULL
> tav <- tav/sum(tav)
> class(tav) <- NULL
> (Wi <- widesI(tus, tav))

************** Manly's Selection ratios for design I ********

Significance of habitat selection:
  Khi2L      df  pvalue
125.821   3.000   0.000


Table of ratios (p-values should be
 compared with Bonferroni level= 0.0125 )
           used avail    Wi SE.Wi     P    Bi
Low       0.056 0.384 0.145 0.042 0.000 0.029
Medium    0.449 0.358 1.255 0.099 0.010 0.251
High      0.424 0.215 1.973 0.163 0.000 0.395
Very High 0.071 0.043 1.628 0.419 0.134 0.326


Bonferroni classement
Based on 95 % confidence intervals on the differences of Wi :

habitat    High  Very High  Medium  Low
High      -----------------
Very High -------------------------
Medium          -------------------
Low                                -----
```

The "table of ratios" presents the selection ratios, together with the proportion of use and availability (and other measures fully described in Manly et al. 2002). This table illustrates that higher elevations are selected by the chamois. Note that a graphical summary of the results can be obtained by typing:

```
> plot(Wi)
```

(not executed in this report). Selection ratios are a useful approach to the study of habitat selection, especially in the context of designs II and III.

Remark: There are close connections between the theory of selection ratios and the GNESFA presented in the previous sections. Indeed, let us consider the table $\mathbf{X}$ containing binary data, indicating whether each pixel (in row) contains (1) or not (0) each habitat type (in column). Let $\mathbf{u}$ be the vector containing the number of chamois detections in each pixel. It can be shown that when the distribution of utilization weights $\mathbf{u}$ is chosen as the Focus distribution, the analysis finds the direction where the selection ratios are the largest. Indeed, the

sum of the eigenvalues of the GNESFA (with utilization weights corresponding to the utilization distribution) corresponds to the sum of the selection ratios minus one. We chan check it on our example:

```
> dis <- acm.disjonctif(slot(elev, "data"))
> pc2 <- dudi.pca(dis, scan=FALSE)
> gnf <- gnesfa(pc2, pr, scan=FALSE)
```

We first used the function `acm.disjonctif` from the package ade4 to convert the factor variable into a complete disjonctive table. Then, after a preliminary PCA on this table (see section 3.2.2), we performed the GNESFA on this table, using the vector `pr` created in previous sections. Now check that the sum of the eigenvalues of the GNESFA corresponds to the sum of the selection ratios plus one:

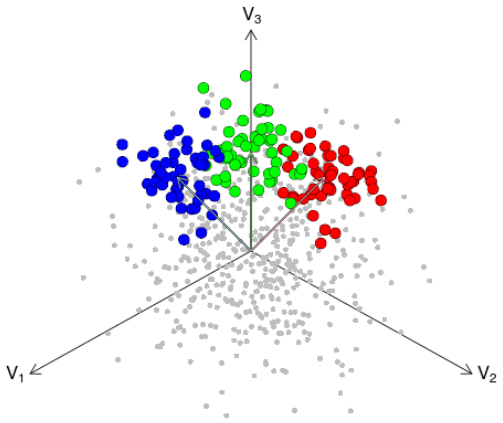
```
> sum(gnf$eig)

[1] 4.001572

> sum(Wi$wi)

[1] 5.001572
```

Thus, in the particular case where there is only one factor variable describing the habitat, the selection ratios and the GNESFA are closely connected. Although the former are easier to understand in this case, when the number of habitat variables increase, the GNESFA provides a consistent way to explore habitat selection in design I studies.

4 Design II studies

4.1 Basic data structure

As explained in section 2.2, design II studies correspond to studies for which animals are identified (e.g. using radio-tracking) and habitat use is measured for each one, but availability is considered to be the same for all animals of the population. Again, the model of the ecological niche can be very useful in this context:



Each RU is characterized by a value for all the environmental variables, so that to each RU is associated a point in the ecological space (grey points on the figure). For each animal, the set of used RUs defines a "niche" in the ecological space (as defined in section 2.3). So that there are as many niches in the ecological space as there are animals. Note that a given RU may possibly be used by several animals. Although both the availability weights and utilization weights may vary from one RU to the other, this model is useful to understand the methods that can be used to study habitat selection with such designs.

Our aim is to identify the directions in the ecological space on which (i) the niches are the most different from the distribution of available points, (ii) these differences between the distribution of available points and the niches are the most similar.

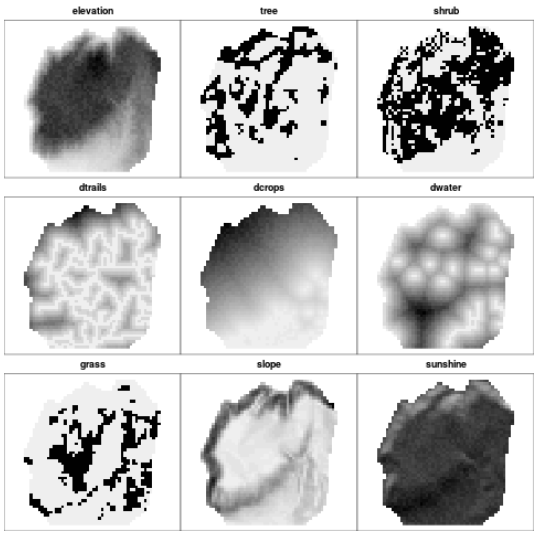
The functions of `adehabitatHS` make an extensive use of the marginality vectors in this context. These vectors connect the mean of the distribution of available points (here, at the origin of the ecological space) to the mean of the niches. These vectors are a rough summary of the selection (they measure the distance between what is available in average and what is used in average by an animal). These vectors are represented by arrows on the figure.

In this section, we will use as an example a dataset collected by Daniel Maillard (Office national de la chasse et de la faune sauvage) on the wild boar. First load the data:

```
> data(puech)
> names(puech)

[1] "relocations" "maps"
```

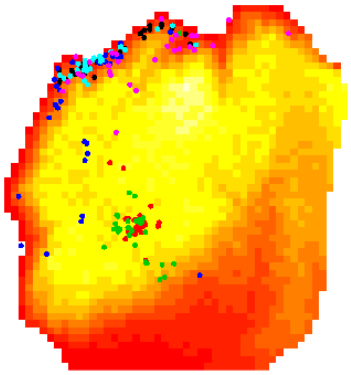
This data set has two components: the component `maps` describes the values of 9 environmental variables on the study area:

```
> maps <- puech$maps
> mimage(maps)
```



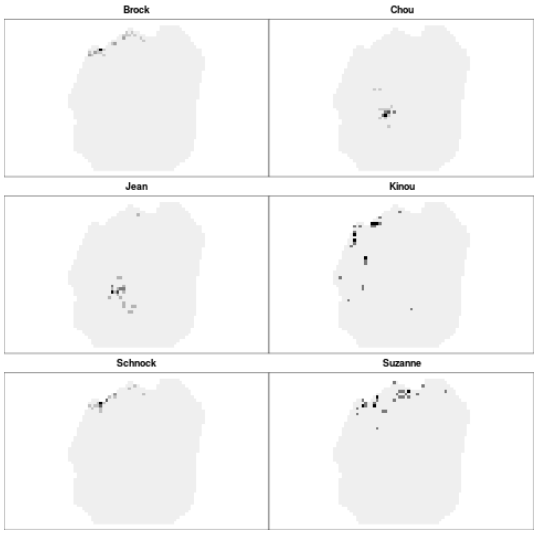
and the component `relocations` contains the relocations of 6 wild boar on the study area:

```
> locs <- puech$relocations
> image(maps)
> points(locs, col=as.numeric(slot(locs, "data")[,1]), pch=16)
```



We now count the number of relocations in each pixel of the map, for each animal (see the help page of the function `count.points`):

```
> cp <- count.points(locs, maps)
> mimage(cp)
```



We can now derive the elements required for the analysis of habitat selection:

```
> X <- slot(maps, "data")
> U <- slot(cp, "data")
```

where **X** and **U** correspond respectively to the table containing the values of the environmental variables and to the utilization weights of each RU and for each animal (see section 2.2).

4.2 The OMI analysis

The Outlying Mean Index (OMI) analysis is one possible approach to the study of habitat selection (Doledec et al. 2000). First the table **X** is centred for the availability weights, so that the origin of the space corresponds to what is available in average to the animals. Then, one performs a noncentred principal component analysis of the coordinates of the marginality vectors in this space. This allows to find the directions where the marginality is in average the largest. More formally, this analysis finds the vector $\mathbf{b} = (b_1, b_2, ..., b_p)$ such that:

- $y = b_1 x_1 + ... + b_p x_p$
- $\sum_{i=1}^{K} \bar{y}_{ui}^2 = \sum_{i=1}^{K} (\bar{y}_{ui} - \bar{y}_{ai})^2$ is maximum on the first axis (where $\bar{y}_{ui}$ is the mean of the utilization distribution of the i^{th} animal and $\bar{y}_{ai}$ is the

mean of the availability distribution for this animal)

The OMI analysis is implemented in the function `niche` of the package `ade4`. Let us try it on our example. As for the GNESFA, it is required that a `dudi.*` analysis is performed as a preliminary step (see section 3.2.2). Because all our variables are numeric, we first perform a principal component analysis:

```
> pc <- dudi.pca(X, scannf=FALSE)
```

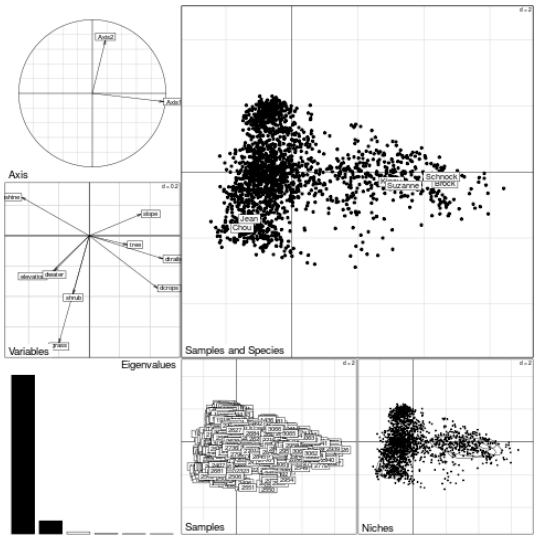
Then, we use the function `niche`:

```
> (ni <- niche(pc, U, scannf=FALSE))
```

```
Niche analysis
call: niche(dudiX = pc, Y = U, scannf = FALSE)
class: niche dudi

$rank (rank)     : 6
$nf (axis saved) : 2
$RV (RV coeff)   :

eigen values: 10.89 0.9072 0.151 0.04964 0.01842 ...

  vector length mode    content
1 $eig   6      numeric eigen values
2 $lw    6      numeric row weigths (crossed array)
3 $cw    9      numeric col weigths (crossed array)

  data.frame nrow ncol content
1 $tab       6    9    crossed array (averaging species/sites)
2 $li        6    2    species coordinates
3 $l1        6    2    species normed scores
4 $co        9    2    variables coordinates
5 $c1        9    2    variables normed scores
6 $ls        1962 2    sites coordinates
7 $as        2    2    axis upon niche axis
```

A summary of the analysis is obtained by typing:

```
> plot(ni)
```



This figure contains a full summary of the analysis:

The **eigenvalues diagram** shows the amount of marginality accounted for by each axis. It is very clear that one axis accounts for most marginality present in the dataset. Actually, the first axis accounts for:

```
> ni$eig[1]/sum(ni$eig)

[1] 0.9052399
```

more than 90% of the marginality in the dataset!! However, the second axis also expresses a notable amount of the marginality: there is a clear break in the decrease of the eigenvalues after the second one. The second axis accounts for:

```
> ni$eig[2]/sum(ni$eig)

[1] 0.07538909
```

7.5% of the marginality. Together, the two axes account for 98% of the marginality in the dataset.

The main graph (**Samples and species**) shows the projection of the RUs (named "samples" on the graph) on the first factorial plane, as well as the position of the mean of the distribution of utilization weights for each animal (named "species" on the graph – this analysis was originally developed for the analysis of multiple species distribution). Four animals out of 6 are characterized by a strong selection of the positive values of the first axis. Note that two animals are characterized by strong negative values of the second axis.

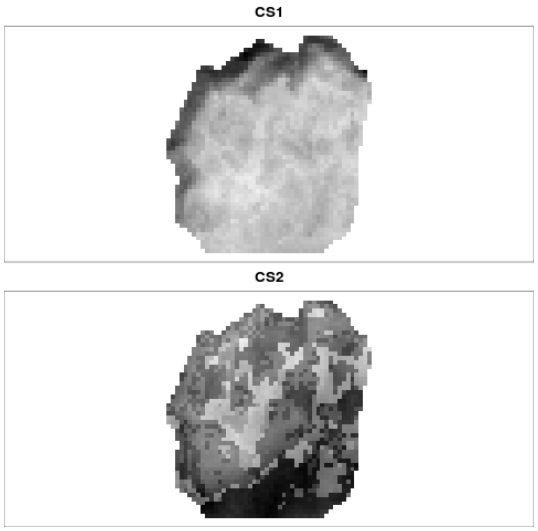
To give a meaning to these axes, have a look at the graph labelled "**Variables**", which presents the scores of the variables on the axes of the analysis (i.e., the values of the coefficients b_i found by the analysis). The positive values of the first axis (strongly selected by four animals) correspond to areas characterized by steep slopes (difficult access for human), far from trails and crops (idem) and low sunshine (this area is Mediterranean, and therefore very warm [often $> 40^{\circ}$ C] and the animals are relocated during the day in summer). The strong selection for the positive values of the first axis are therefore easily interpretable biologically. The negative values of the second axis (strongly selected by two animals) correspond to areas with high grass cover (it is easier for the animals to hide there).

The graph labelled "**Axis**" is often of interest. It represents the correlation between the first two axes of the PCA of the available RUs (i.e. the preliminary `dudi.*` analysis) and the axes of the OMI analysis. The first axis of the preliminary PCA identifies the direction where the variance of the availability distribution is the largest. Thus, the first two axes of the PCA identifies the main directions structuring the study area. The strong correlation between the scores of the available RUs on the first axis of the PCA and the scores of the available RUs on first axis of the OMI analysis indicates that the direction where the habitat selection appears the strongest is also the direction where the environment is the most variable. In some studies, this observation may be of biological interest (and we will see in the next chapter that this may also be problematic in some studies).

The other graphs are not of interest for us.

Another useful display consists in mapping, in the geographical space, the scores of the RUs on the axes. This is done simply by:

```
> ls <- ni$ls
> coordinates(ls) <- coordinates(maps)
> gridded(ls) <- TRUE
> mimage(ls)
```



This kind of maps is also very useful to give a biological meaning to the axes found by the analysis. In this case, it is very clear that the four animals selecting the positive values of the first axis are located at the extreme north of the study area (this area correspond to the gorges of the Herault river).

4.3 The canonical OMI analysis

We noted that the strong correlation between the first axis of the PCA and the first axis of the OMI analysis may sometimes indicate a problem. We now explain why. Dray et al. (2003) noted that the OMI analysis is a co-inertia analysis. A full description of the co-inertia analysis can be found in Dray et al. (2003) and Doledec and Chessel (1994). We give here a rapid description of this method. The co-inertia analysis is a "two-table" factor analysis, like the redundancy analysis or the canonical correlation analysis. Its basic aim is similar to the aim of canonical correlation analysis: to find similarities between two tables $\mathbf{A}$ and $\mathbf{B}$ with the same number of rows. Mathematically, it corresponds to the principal component analysis of the table $\mathbf{T} = \mathbf{A}^t\mathbf{B}$, using uniform row weights and column weights for $\mathbf{T}$. It can be shown that when both $\mathbf{A}$ and $\mathbf{B}$ are centred and scaled, so that their columns all have a mean equal to zero and a standard deviation of 1, then this analysis finds a direction $\mathbf{u}$ in the space defined by the columns of $\mathbf{A}$ and a direction $\mathbf{v}$ in the space defined by the columns of $\mathbf{B}$ so that the covariance between $\mathbf{Au}$ and $\mathbf{Bv}$ is maximized. The main quality of the co-inertia analysis is that there is no mathematical constraint on the number of rows of the tables $\mathbf{A}$ and $\mathbf{B}$ (The number of rows of these tables may even be smaller than the number of columns).

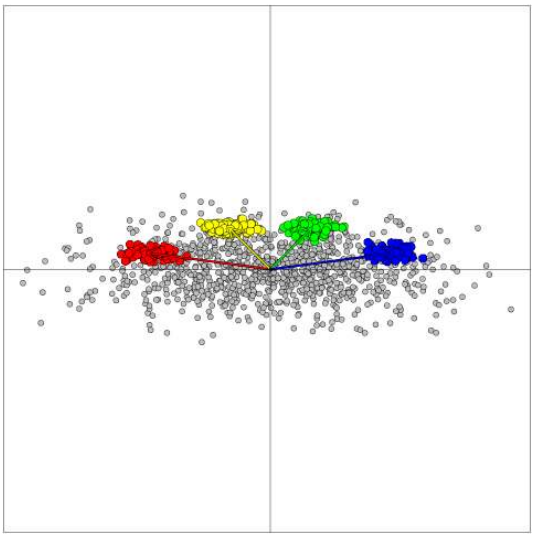
An important point here is that:

$$\text{cov}(\mathbf{Au}, \mathbf{Bv}) = \text{cor}(\mathbf{Au}, \mathbf{Bv})\sqrt{\text{var}(\mathbf{Au})}\sqrt{\text{var}(\mathbf{Bv})}$$

Therefore, if the variance of, say, the table **A** is *very* large in a given direction, there is a risk that the analysis returns this direction, even if the correlation between this direction and any direction in the space defined by the columns of **B** is small. This may or may not be a problem, depending on the context in which the co-inertia analysis is used. However, for the study of habitat selection, it is problematic.

In the case of OMI analysis, the table **A** is centred (it corresponds to the table **X** defined in section 2.2), but the table **B** is not (it corresponds to the table **U** in section 2.2), so that it is not the *covariance* that is maximized by the analysis, but more generally a *co-inertia* (a covariance is a particular type of co-inertia). In particular, as we have already seen, the OMI analysis maximizes the average marginality of the animals. However, the arguments given above still hold. *When there is a strong environmental pattern on the study area, there is a non-negligible risk that the first axes of the OMI analysis identifies this pattern, more than habitat selection.*

This can be illustrated on the following figure, illustrating the habitat selection of four animals on two variables:

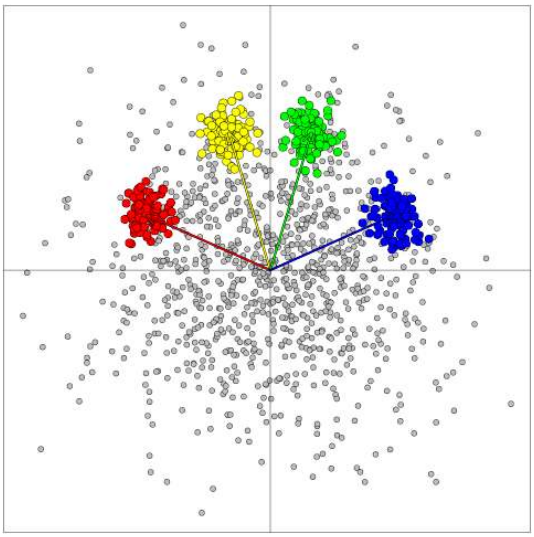


The grey points correspond to the available RUs and the colored points correspond to the used RUs of 4 animals. The marginality vectors are also indicated. It is clear on this figure that the direction that maximizes the marginality is the abscissa (because this is the direction that is in average the "closest" to the marginality vectors). However, in this direction, we find animals using both very positive and very negative values of this axis, so that there is no strong selection of the animals on this direction. On the other hand, it is also clear that the direction on which habitat selection occurs is the ordinate (all the marginality

vectors are characterized by a positive coordinate, no use of the negative values in this direction). Therefore, maximizing the marginality explained by the first axis does not necessarily returns the direction where the habitat selection is the largest. If there is a direction of the ecological space where the variance of the available RUs is much larger than in any other direction, there is a risk that the OMI analysis return this direction, whatever the habitat selection by the species. This may be a problem with OMI analysis.

I already have met this problem, when I was working on the habitat selection of the mouflon in the Bauges mountains (Darmon et al., 2012). In this mountainous area, the elevation strongly structures all the environmental variables. The first axis of the OMI analysis was mainly determined by the elevation pattern, and no common direction for the marginality vectors was identified.

This is where the *canonical OMI analysis* may be useful. This analysis first distort the ecological space so that the distribution of availability weights takes a standard spherical shape:



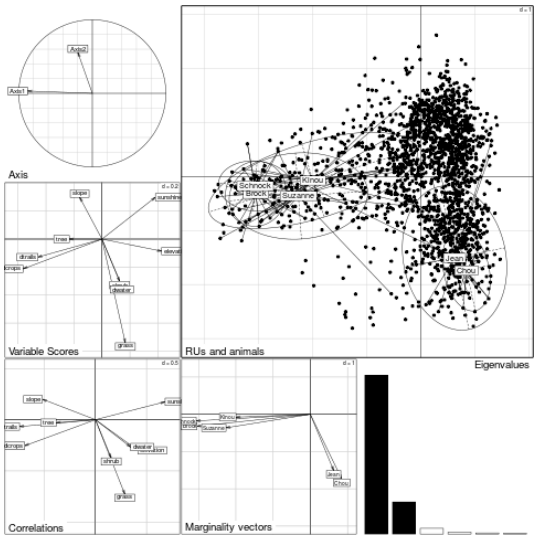
It is clear then that the direction where the marginality is the largest in this distorted space is the direction where habitat selection will be the clearest. However, there is a counterpart: the canonical OMI analysis places mathematical constraint on (i) the number of RUs (it should be larger than the number of environmental variables to allow the matrix inversion required by the analysis) and (ii) the environmental variables should be linearly independent (also to allow this inversion).

Additional mathematical details can be found in Darmon et al. (2012). The canonical OMI analysis is implemented in the function `canomi`. We can carry out the analysis:

```
> com <- canomi(pc, U, scannf=FALSE)
```

A full summary of the results is displayed below:

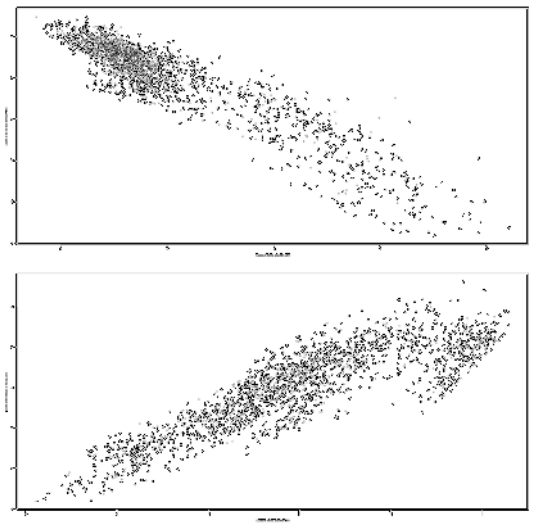
```
> plot(com)
```



The **Eigenvalues** indicates that the first two axes account for most of the marginality of the dataset, once the ecological space has been distorted to sphere the distribution of availability weights (clear break after the second eigenvalue). The main graph (**RUs and animals**) indicates that four animals select strongly for negative values of the first axis, and two animals for negative values of the second axis. This is confirmed by the **Marginality vectors**. Both the **variable scores** and the **correlations** indicate that negative values of the first axis correspond to areas with steep slopes, far from the crops and from recreational trails, with low elevation and sunshine (but the two graphs may sometimes indicate different results (see section 4.3 for further discussion on this aspect).

There is actually a strong correlation between the results of the classical OMI analysis and the results of the canonical OMI analysis:

```
> par(mfrow=c(2,1))
> plot(ni$ls[,1], com$ls[,1],
+      xlab="Scores on the first axis of the OMI",
+      ylab="Scores on the first axis of the can. OMI")
> plot(ni$ls[,2], com$ls[,2],
+      xlab="Scores on the first axis of the OMI",
+      ylab="Scores on the first axis of the can. OMI")
```



The classical and canonical OMI analyses here return the same patterns, but this may not always be the case.

Remark: note that variable availability weights can be defined for both the OMI and canonical OMI analysis, by passing them as the `row.w` argument to the preliminary `dudi.*` analysis.

4.4 When habitat is defined by several categories

Similarly to design I studies, it is a very common approach to study habitat selection by several animals by defining several habitat categories on the study area. For example, let us consider the slope map on the study area. Define four classes of slope (using the R function `cut`):

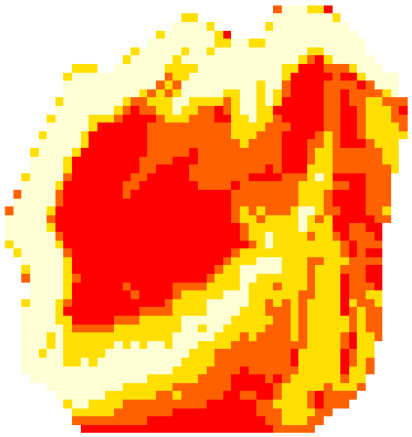
```
> slope <- maps[,8]
> sl <- slot(slope, "data")[,1]
> av <- factor(cut(sl, c(-0.1, 2, 5, 12, 50)),
+              labels=c("Low","Medium","High","Very High"))
>
```

The number of pixels in each class is:

```
> (tav <- table(av))

av
      Low    Medium      High Very High
      545       467       477       473
```

Let us map this variable:

```
> slot(slope, "data")[,1] <- as.numeric(av)
> image(slope)
```



Now, compute the percentage of use of each habitat class by each animal. That is, we compute the number of relocations for each animal in each habitat class:

```
> us <- join(locs, slope)
> tus <- table(slot(locs,"data")[,1],us)
> class(tus) <- NULL
> tus <- as.data.frame(tus)
> colnames(tus) <- names(tav)
> tus
```

```
        Low Medium High Very High
Brock     0      0    1        29
Chou     28      2    0         0
Jean     22      3    2         3
Kinou     6      0    2        21
Schnock   0      0    1        25
Suzanne   1      0    8        21
```

There are several ways to compare use and availability in design II studies when several habitat types have been defined. A common approach is the compositional analysis proposed by Aebischer et al. (1993). This approach is implemented in the function `compana`:

```
> tav2 <- matrix(rep(tav, nrow(tus)), nrow=nrow(tus), byrow=TRUE)
> colnames(tav2) <- names(tav)
> compana(tus, tav2, test = "randomisation",
+         rnv = 0.01, nrep = 500, alpha = 0.1)
```

```
************ Compositional analysis of habitat use ***************

The analysis was carried out with 6 animals and 4 habitat types
1. Test of the habitat selection:
   randomisation test
   Lambda         P
0.1256895 0.0780000

2. Ranking of habitats (profile):

habitat    Very High  High  Low  Medium
Very High -----------      -----
High                -------------------
Low       ------------------------------
Medium              -------------------
```

Here, the compositional analysis does not identify any significant habitat selection. However, this approach relies on the hypothesis that all the animals are selecting habitat in the same way. And we have seen in the previous sections that this is not necessarily the case...

Another approach has been proposed by Manly et al. (2002), relying on selection ratios (see section 3.4). For each animal, a selection ratio can be calculated for each habitat type. Then, after having tested whether habitat selection is the same for all animals, it is possible to average selection ratios over all animals:

```
> tav <- as.vector(tav)
> names(tav) <- names(tus)
> (WiII <- widesII(tus, tav))

************** Manly's Selection ratios for design II ********

1. Test of identical use of habitat by all animals
   (Classical Khi-2 performed on the used matrix):
  Khi2L1       df   pvalue
176.8034  15.0000   0.0000
2. Test of overall habitat selection:
  Khi2L2       df   pvalue
311.5155  18.0000   0.0000
3. Test of hypothesis that animals are on average using resources
   in proportion to availability, irrespective of whether they are
   the same or not (Khi2L2 - Khi2L1):
Khi2L2MinusL1            df        pvalue
     134.7121        3.0000        0.0000
```

```
Table of selection ratios:
          Available       Used        Wi         SE IClower ICupper
Low       0.2777778 0.32571429 1.1725714 0.61211413 -0.3563  2.7015
Medium    0.2380224 0.02857143 0.1200367 0.07714862 -0.0727  0.3127
High      0.2431193 0.08000000 0.3290566 0.16387318 -0.0803  0.7384
Very High 0.2410805 0.56571429 2.3465781 0.72104883  0.5456  4.1475


Bonferroni classement
Based on 95 % confidence intervals on the differences of Wi :

habitat    Very High  Low  High  Medium
Very High ----------------
Low       ------------------------------
High                 -------------------
Medium               -------------------
```

In this case, we can see that habitat selection is not the same from one animal to the other. Therefore, it does not make sense to compute the average selection ratios. Rather, we should investigate what causes these differences.

Again, a factorial analysis will be of help here. The *eigenanalysis of selection ratios* (Calenge and Dufour, 2006) has been developed to explore graphically habitat selection by the wildlife when habitat is defined by several categories. This analysis is closely related to the theoretical context underlying the selection ratios. Indeed, let $\mathbf{W}$ be the table containing the selection ratios for each animal (in row) and each habitat type (in column). The eigenanalysis consists in a noncentred and nonscaled principal component analysis of the table $\mathbf{W} - \mathbf{1}$, using the availability weight of each habitat type as column weight and the number of relocations of each animal as row weight (see Calenge and Dufour, 2006, for more mathematical details). This analysis partition the statistics:

$$S = \sum_{i=1}^{P} \sum_{j=1}^{K} \frac{(u_{ij} - p_i u_{+j})^2}{p_i u_{+j}}$$

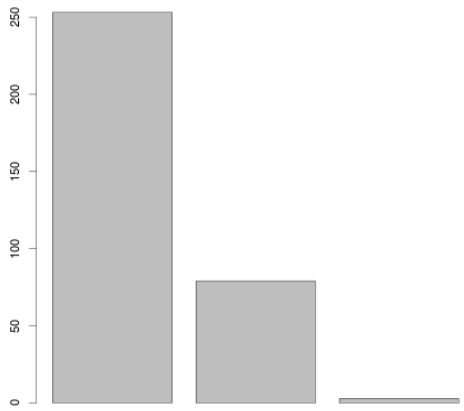
where u_{ij} is the number of relocations of animal j in habitat i, p_i is the proportion of available resource units in habitat i and u_{+j} is the total number of relocations of animal j. This statistic was proposed by White and Garrott (1990) to test habitat selection in design II studies. What is interesting is that this analysis connects two widely used approach for habitat selection studies into a unified framework (selection ratios and White and Garrott (1990) statistic). The direction of the ecological space that maximizes the statistic S – and therefore habitat selection – is found by an analysis of the selection ratios.

We now perform the eigenanalysis of selection ratios on our example dataset, with the function `eisera`:

```
> (eis <- eisera(tus,tav2, scannf=FALSE))

Factorial analysis of selection ratios

$call: eisera(used = tus, available = tav2, scannf = FALSE)

$nf: 2 axis-components saved
$rank: 3
eigen values: 253.2 79.04 2.861
  vector length mode    content
1 $cw    4      numeric column weights
2 $lw    6      numeric row weights
3 $eig   3      numeric eigen values

  data.frame nrow ncol content
1 $tab       6    4    modified array
2 $li        6    2    row coordinates
3 $co        4    2    column coordinates
4 $available 6    4    available proportions
5 $used      6    4    number of relocations
6 $wij       6    4    selection ratios
```

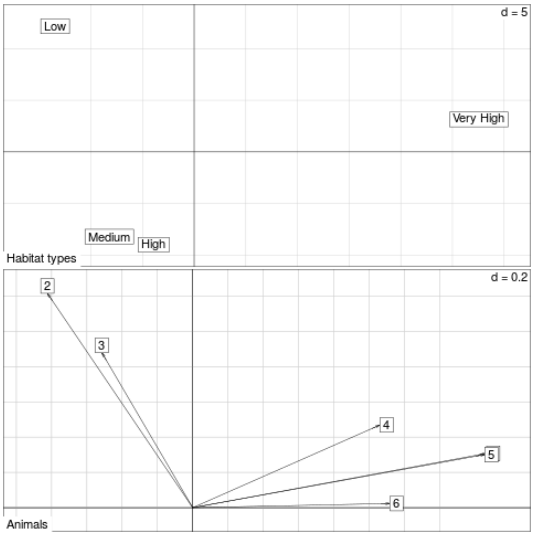
We focus our interpretation on two axes after considering the eigenvalues diagram:

```
> barplot(eis$eig)
```



The results of the analysis are presented below:

```
> scatter(eis)
```



Here, we find similar results as in previous sections (a group of four animals is selecting for steep slopes and a group of three animals is selecting for low slopes).

4.5 Concluding remarks regarding design II analyses

We have seen that the eigenanalysis of selection ratios connects two widely used approaches for habitat selection studies into a unified framework (selection ratios and White and Garrott statistic). Actually, it can be shown that this analysis is a particular case of canonical OMI analysis. Let us perform a canonical OMI analysis on our example dataset, to illustrate this point. First transform the map of slopes into a complete disjonctive table:

```
> avdis <- acm.disjonctif(data.frame(av))
> head(avdis)
```

```
  av.Low av.Medium av.High av.Very High
1      1         0       0            0
2      1         0       0            0
3      1         0       0            0
4      1         0       0            0
5      1         0       0            0
6      1         0       0            0
```

Then, perform a canonical OMI analysis using this dataset:

```
> pc <- dudi.pca(avdis, scannf=FALSE)
> com2 <- canomi(pc, U, scannf=FALSE)
```

Now compare the eigenvalues of the canonical OMI analysis with the eigenvalues of the eigenanalysis of selection ratios:

```
> eis$eig/com2$eig
```

```
[1] 175 175 175
```

The eigenvalues of the eigenanalysis of selection ratios are equal to the eigenvalues of the canonical OMI analysis multiplied by a constant (this constant is equal to $u_{++}-P-1$, where u_{++} is the total number of relocations for all animals and P is the number of habitat types). The absolute value of the coordinates of the marginality vectors are also identical:

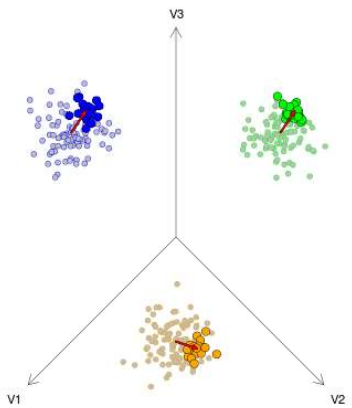
```
> eis$li[,1]/com2$li[,1]
```

```
[1] 1 1 1 1 1 1
```

5 Design III studies

5.1 Basic data structure

As explained in section 2.2, design III studies correspond to studies for which animals are identified (e.g. using radio-tracking) and both habitat use and availability is measured for each one. The key aspect is that the availability varies from one animal to the other. Again, the model of the ecological niche can be very useful in this context:

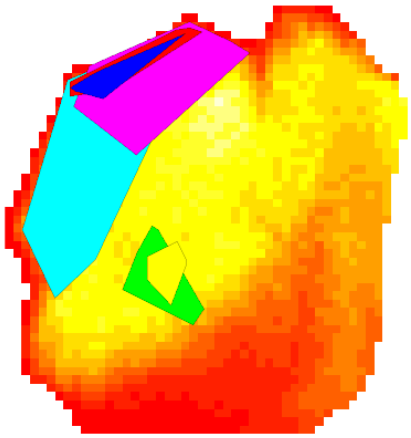


A particular set of RUs is available to each animal, and a subset of it is used by the animal. Therefore, each animal is characterized by a niche defined in an

"available space" particular to it. As in design II studies, we will work with the marginality vectors (that relate what is available to the animal in average to what has been used in average by the animal).

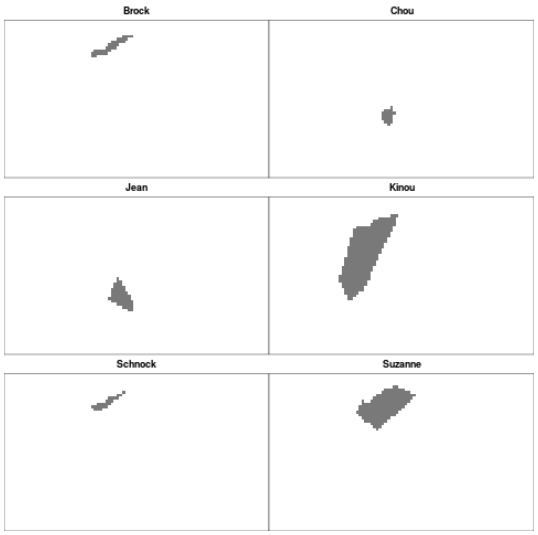
We will use as an example the dataset analyzed in the previous section (see section 4.1 for a description of this data set). Let us estimate the home range of the animals, e.g. using a minimum convex polygon:

```
> pcc <- mcp(locs)
> image(maps)
> plot(pcc, col=rainbow(6), add=TRUE)
```



We rasterize these polygons

```
> hr <- do.call("data.frame", lapply(1:nrow(pcc), function(i) {
+     over(maps,geometry(pcc[i,]))
+ }))
> names(hr) <- slot(pcc, "data")[,1]
> coordinates(hr) <- coordinates(maps)
> gridded(hr) <- TRUE
> mimage(hr)
```



We will study habitat selection inside the home-range (third order habitat selection according to the scale of Johnson, 1980). The black pixels define what is available to each animal, and pixels containing at least one relocation define the use by the animal (stored in the object `cp` built in section 4.1). The key problem is that the available pixels are not the same from one animal to the other.

5.2 The K-select analysis

The K-select analysis is one possible approach to this kind of study (Calenge et al. 2005). This analysis can be seen as an extension of the OMI analysis presented in section 4.2. This analysis focuses on the marginality vectors of the animals. These marginality vectors are recentred so that they have a common origin. Then, a noncentred principal analysis is performed on the table containing the coordinates of these recentred marginality vectors. This analysis is therefore similar to the OMI analysis: it finds the direction in the ecological space where the marginality is the strongest (it is identical to the OMI analysis when the available RUs are the same for all animals). More formal details can be found in Calenge et al. (2005).

Let us perform a K-select analysis on our dataset. First prepare the data for the analysis:

```
> pks <- prepksel(maps, hr, cp)
> names(pks)
```

```
[1] "tab"    "weight" "factor"
```

The component `tab` correspond to the concatenated tables $\mathbf{X_i}$ containing the values of the environmental variables (columns) in each pixel (rows) available to

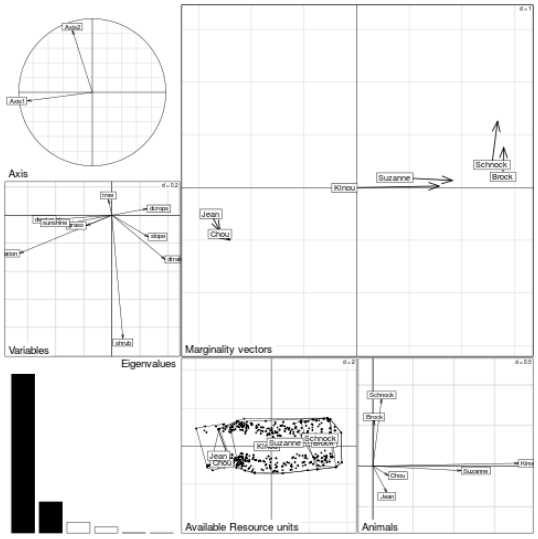
the animal i. The vector `weight` contains the corresponding utilization weights, and the vector `factor` is a factor allowing to define the limits of the $\mathbf{X}_i$ in the component `tab`. We suppose here that the availability weights are the same for all pixels.

We perform a preliminary PCA of the data (the section 3.2.2 explains why this preliminary PCA is required):

```
> pc <- dudi.pca(pks$tab, scannf=FALSE)
```

And we perform the K-select analysis:

```
> ksel <- kselect(pc, pks$factor, pks$weight, scann=FALSE)
> plot(ksel)
```



This graph summarizes the results of the analysis. The **Eigenvalues** diagram indicate that the first axis explains most of the marginality present in the dataset (there is a clear break in the decrease of the eigenvalues after the first one). The graph labelled **Animals** shows the recentred marginality vectors (i.e. marginality vectors shifted so that they have a common origin). This graph indicates that all animals are characterized by a positive score on the first axis of the analysis.

The graph labelled **Variables** gives the scores of the variables. This graph allows to give a biological meaning to the axes. Here, positive values of the axis correspond to the RUs located at low elevation, close to water, with low sunshine, far from trails and crops, with high slopes. This axis therefore opposes the areas located in the gorges of the Herault river and the areas located on the plateau.

The graph labelled **Available resource units** shows the distribution of available RUs on the first two axes of the analysis. It is clear on this graph that Jean and Chou are characterized by available RUs all located on the plateau (with a restricted range of values on the first axis), Schnock and Brock are characterized by available RUs all located in the gorges of the Herault river (also with a restricted range of values on the first axis), and Suzanne and Kinou have the two types of habitat within their home range (large range of values on the first axis).

The main graph, labelled **Marginality vectors** shows the original (i.e. non-recentred) marginality vectors. The origin of the arrow indicates what is available in average to the animal, the end of the arrow indicates what has been used in average by the animal, and the direction and length of the arrows indicate respectively the direction and strength of habitat selection. This graph shows that the strongest selection is showed by the two animals having the largest choice of values of the first axis of the analysis. A possible interpretation would be that the gorges of the Herault river is a preferred habitat that is selected when available.

5.3 When habitat are defined by several categories

As for other designs, studying habitat selection when habitat is defined by several categories is a common approach. We will not illustrate the methods available to deal with such designs here, as the methods available to deal with design III are just extensions of the methods available for design II. We therefore refer the reader to the help pages of the following functions:

- `compana`: the compositional analysis is also a possible approach for the analysis of habitat selection in design III studies (Aebischer et al. 1993);
- `widesIII`: Manly et al. (2002) also extended the theoretical framework underlying the selection ratios to the analysis of habitat selection;
- `eisera`: the eigenanalysis of selection ratios is also a possible approach to the analysis of habitat selection in design III studies (Calenge and Dufour, 2006).

6 Conclusion

I included in the package `adehabitatHS` several functions allowing the exploration of habitat selection by the wildlife. The functions from the other brother packages can be used to explore habitat selection using a wide variety of approaches.

References

Aebischer, N., Robertson, P. and Kenward, R. 1993. Compositional analysis of habitat use from animal radio-tracking data. Ecology, 74, 1313-1325.

Bingham, R. and Brennan, L. 2004. Comparison of type I error rates for statistical analyses of resource selection Journal of Wildlife Management, 68, 206–212.

Basille, M., Calenge, C., Marboutin, E., Andersen, R. and Gaillard, J.M. 2008. Assessing habitat selection using multivariate statistics: Some refinements of the ecological-niche factor analysis. Ecological Modelling, 211, 233-240.

Calenge, C. 2005. Des outils statistiques pour l'analyse des semis de points dans l'espace ecologique. Universite Claude Bernard Lyon 1.

Calenge, C., Dufour, A. and Maillard, D. 2005. K-select analysis: a new method to analyse habitat selection in radio-tracking studies Ecological Modelling, 186, 143-153.

Calenge, C. 2006. The package adehabitat for the R software: a tool for the analysis of space and habitat use by animals. Ecological modelling, 197, 516–519.

Calenge, C. and Dufour, A. 2006. Eigenanalysis of selection ratios from animal radio-tracking data. Ecology, 87, 2349–2355.

Calenge, C. and Basille, M. (2008) A general framework for the statistical exploration of the ecological niche. Journal of Theoretical Biology, 252, 674-685.

Calenge, C., Darmon, G., Basille, M., Loison, A. and Jullien, J. 2008. The factorial decomposition of the Mahalanobis distances in habitat selection studies. Ecology, 89, 555-566.

Carpenter, G., Gillison, A. and Winter, J. 1993. DOMAIN: a flexible modelling procedure for mapping potential distributions of plants and animals. Biodiversity and Conservation, 2, 667-680.

Chase, J. and Leibold, M. 2003. Ecological niches. Linking class and contemporary approaches. The University of Chicago Press.

Chessel, D., Dufour, A. and Thioulouse, J. (2004) The ade4 package. R news.

Clark, J., Dunn, J. and Smith, K. 1993. A multivariate model of female black bear habitat use for a geographic information system. Journal of Wildlife Management, 57, 519-526

Cleveland, W. 1993. Visualizing data. Hobart Press.

Darmon, G., Calenge, C., Loison, A., Jullien, J.M., Maillard, D. and Lopez, J.F. 2012. Spatial distribution and habitat selection in coexisting species of mountain ungulates. Ecography, 35, 44-53.

Doledec, S., Chessel, D. and Gimaret-Carpentier, C. 2000. Niche separation in community analysis: a new method. Ecology, 81, 2914-2927.

Doledec, S. and Chessel, D. 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biology, 31, 277-294.

Dray, S., Chessel, D. and Thioulouse, J. 2003. Co-inertia analysis and the linking of ecological tables. Ecology, 84, 3078-3089.

Gabriel, K. 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika, 58, 453-467.

Hall, L., Krausman, P. and Morrison, M. 1997. The habitat concept and a plea for standard terminology. Wildlife Society Bulletin, 25, 173–182.

Hill, M. and Smith, A. 1976. Principal component analysis of taxonomic data with multi-state discrete characters. Taxon, 25, 249-255.

Hirzel, A., Hausser, J., Chessel, D. and Perrin, N. 2002. Ecological-niche factor analysis: How to compute habitat suitability maps without absence data? Ecology, 83, 2027-2036.

Hutchinson, G. 1957. Concluding remarks. Cold Spring Harbour Symposium, Quantitative Biology, 22, 415–427.

Johnson, D. 1980. The comparison of usage and availability measurements for evaluating resource preference. Ecology, 61, 65-71.

Manly, B., McDonald, L., Thomas, D., MacDonald, T. and Erickson, W. 2002. Resource selection by animals. Statistical design and analysis for field studies. Kluwer Academic Publisher.

Pebesma, E. and Bivand, R.S. 2005. Classes and Methods for Spatial data in R. R News, 5, 9–13.

Tenenhaus, M. and Young, F. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika, 5, 91–119.

Thomas, D. and Taylor, E. 1990. Study designs and tests for comparing resource use and availability. Journal of Wildlife Management, 54, 322–330.

White, G. and Garrott, R. 1990. Analysis of wildlife radio-tracking data Academic press.

7 Appendix: the derivation of a new factor analysis by James Dunn

Let $\mathbf{y}$ be a P-dimensional vector of random variables, with successive realizations being the data vectors corresponding to the z vectors of the manuscript. Let $\mathbf{a}$ denote any P-dimensional vector of length one whose extensions define an axis in P-space. The projection of $\mathbf{y}$ on the axis defined by $\mathbf{a}$ is $x = \mathbf{a}^t\mathbf{y}$ and immediately:

$$\mathrm{Var}(x_u) = \mathbf{a}^t\Sigma_u\mathbf{a}$$

if $\mathbf{y}$ is any "used" pixel, and

$$\mathrm{Var}(x_a) = \mathbf{a}^t\Sigma_a\mathbf{a}$$

if $\mathbf{y}$ is any "available" pixel. Note that Σ_u and Σ_a are correlation matrices if the elements of $\mathbf{y}$ are properly scaled.

The problem: Find $\mathbf{a}$ of length one such that:

$$\phi = \mathrm{var}(x_a)/\mathrm{var}(x_u) = \mathbf{a}^t\Sigma_a\mathbf{a}/\mathbf{a}^t\Sigma_u\mathbf{a}$$

is maximized.

We arrive at the stationary values of ϕ by differentiating with respect to $\mathbf{a}$ and equating the result to a vector of zeros:

$$\frac{\partial\phi}{\partial\mathbf{a}} = \frac{2\Sigma_a\mathbf{a}}{\mathbf{a}^t\Sigma_u\mathbf{a}} - \frac{(2\mathbf{a}^t\Sigma_a\mathbf{a})\Sigma_u\mathbf{a}}{(\mathbf{a}^t\Sigma_u\mathbf{a})^2}$$

Equating this to $\mathbf{0}$ and performing the obvious cancellations yields:

$$\Sigma_a\mathbf{a} - \left(\frac{\mathbf{a}^t\Sigma_a\mathbf{a}}{\mathbf{a}^t\Sigma_u\mathbf{a}}\right)\Sigma_u\mathbf{a} = \mathbf{0}$$

or

$$(\Sigma_a - \phi\Sigma_u)\mathbf{a} = \mathbf{0}$$

by recognizing the form of the criterion function, or

$$(\Sigma_u^{-1}\Sigma_a - \phi\mathbf{I})\mathbf{a} = \mathbf{0}$$

The latter equation identifies the required solution, $\mathbf{a}$, as an eigenvector (scaled to length one) corresponding to the largest eigenvalue of $\Sigma_u^{-1}\Sigma_a$. Clearly smaller stationary values of ϕ correspond to the smaller eigenvalues of the same matrix and occur at coordinates given by their respective associated eigenvectors.

The apparent computational difficulty associated with finding eigenvalues and eigenvectors of non-symmetric $\Sigma_u^{-1}\Sigma_a$ is alleviated by the factorization $\Sigma_a = \mathbf{T}^t\mathbf{T}$, where $\mathbf{T}$ is upper triangular. In terms of the characteristic root (eigenvalue) operator, ch(.), then

$$\phi_1 = \text{Max } ch(\Sigma_u^{-1}\Sigma_a) = \max\ ch(\Sigma_u^{-1}\mathbf{T}^t\mathbf{T}) = \max\ ch(\mathbf{T}\Sigma_u^{-1}\mathbf{T}^t)$$

where $\mathbf{T}\Sigma_u^{-1}\mathbf{T}^t$ is symmetric. Its associated eigenvector $\mathbf{a}_1$ must satisfy:

$$\begin{aligned}
(\Sigma_u^{-1}\Sigma_a - \phi_1\mathbf{I})\mathbf{a}_1 &= (\Sigma_u^{-1}\mathbf{T}^t\mathbf{T} - \phi_1\mathbf{I})\mathbf{a}_1 = \mathbf{0} && (1)\\
(\Sigma_u^{-1}\mathbf{T}^t - \phi_1\mathbf{T}^{-1})\mathbf{T}\mathbf{a}_1 &= \mathbf{0} && (2)\\
(\mathbf{T}\Sigma_u^{-1}\mathbf{T}^t - \phi_1\mathbf{I})\mathbf{T}\mathbf{a}_1 &= (\mathbf{T}\Sigma_u^{-1}\mathbf{T}^t - \phi_1\mathbf{I})\mathbf{b}_1 = \mathbf{0} && (3)
\end{aligned}$$

where $\mathbf{b}_1 = \mathbf{T}\mathbf{a}_1$ is the eigenvector of the symmetric matrix $\mathbf{T}\Sigma_u^{-1}\mathbf{T}^t$ associated with its largest eigenvalue ϕ_1. Clearly then $\mathbf{a}_1 = \mathbf{T}^{-1}\mathbf{b}_1$ is the required solution to the problem as posed. Additional eigenvalue-eigenvector pairs are similarly found. If the $\mathbf{A} = (\mathbf{a}_1, ..., \mathbf{a}_P)$ and $\mathbf{B} = (\mathbf{b}_1, ..., \mathbf{b}_P)$ represent column-wise concatenations of the respective eigenvectors, then $\mathbf{B} = \mathbf{T}\mathbf{A}$, or $\mathbf{A} = \mathbf{T}^{-1}\mathbf{B}$ whose first column defines the axis required to solve the original problem. Additional orthogonal axes are defined by the remaining columns of $\mathbf{A}$ (as in principal components analysis).

The foregoing development relates to Mahalanobis D^2 as follows:

We shall want D^2 to reflect dissimilarity between any pixel, $\mathbf{y}$, and the mean of the "used" pixels, μ_u, using Σ_u as the metric. Our reasoning is that any "available" site has the potential of being a "used" site until proved otherwise by the magnitude of D^2.

$$\begin{aligned}
D^2 &= (\mathbf{y} - \mu_u)^t\Sigma_u^{-1}(\mathbf{y} - \mu_u) && (4)\\
&= (\mathbf{y} - \mu_u)^t\mathbf{T}^{-1}\mathbf{T}\Sigma_u^{-1}\mathbf{T}^t\mathbf{T}^{-t}(\mathbf{y} - \mu_u) \text{ (where } \mathbf{T}^{-t} = (\mathbf{T}^t)^{-1}) && (5)\\
&= (\mathbf{y} - \mu_u)^t\mathbf{T}^{-1}\mathbf{B}\mathbf{D}_\phi\mathbf{B}^t\mathbf{T}^{-t}(\mathbf{y} - \mu_u) \text{ (where } \mathbf{D}_\phi = diag(\phi_1, ..., \phi_P) && (6)\\
&= (\mathbf{y} - \mu_u)^t\mathbf{A}\mathbf{D}_\phi\mathbf{A}^t(\mathbf{y} - \mu_u) && (7)\\
&= \sum_{j=1}^{P}\phi_j((\mathbf{y} - \mu_u)^t\mathbf{a}_j)^2 && (8)\\
&= \sum_{j=1}^{P}\left(\frac{(\mathbf{y} - \mu_u)^t\mathbf{a}_j}{1/\sqrt{\phi_j}}\right)^2 && (9)
\end{aligned}$$

Clearly any R-dimensional subset of these components also could be used to reflect the species-suitability of a pixel, e.g,

$$D_R^2 = \sum_{j=1}^{R} \left(\frac{(\mathbf{y} - \mu_u)^t \mathbf{a}_j}{1/\sqrt{\phi_j}} \right)^2$$

assuming that the eigenvalues are ordered $\phi_j > \phi_{j+1}$. Thus, D^2 is seen to partition into a weighted sum of squares of projections of $\mathbf{y} - \mu_u$ on each successive derived axes.