Masaryk University Brno
Department of Botany and Zoology

Internship Report

10.04.2017 - 26.05.2017

Analysis & Exploitation of Ecological Data Sets

Markus Neupert

Training Supervisor :	Lubomír Tichý
Academic Advisor :	Matthieu Chauvat

Analysis & Exploitation of Ecological Data Sets

Markus Neupert

10.04.2017 - 26.05.2017

Table of Contents

Acknowledgements

I would like to take the time to thank all of the people that have been involved, in some way or another, with the internship, as well as the present report that stems from it.

Milan Chytrý: For making this whole endeavor possible, by allowing me to join the department of Botany and Zoology for these seven weeks.

Lubomír Tichý: For providing the subject of the internship. For showing me around the facilities, and for introducing me to the outstanding people working in this department. For proofreading this report, and especially for turning this period into a valuable and pleasant experience.

David Zelený: For creating the *ordijuice* package, and for providing very useful information about his code, as well as interesting ideas for the future.

Presentation of Masaryk University

Founded in 1919, Masaryk University Brno counts 9 faculties in the city of Brno, and 35.000 students. Tomáš Garrigue Masaryk, first president of the independent Czechoslovakia, and leader of the movement for a second czech university, has lent its name to the second largest university of the Czech Republic.[1]

Figure 0.1: University Campus Bohunice

Belonging to the Faculty of Science, the Department of Botany and Zoology specializes in the fields of evolutionary biology, biosystematics and ecology. Research focuses on phylogenetic relationships and classification of organisms, as well as ecological relationships of species, populations and communities in entire ecosystems. It presents two centres of excellence - Pladias (Plant Diversity Analysis and Synthesis Centre) and Ecip (European Centre of Ichthyoparasitology) - which have rewarded the department with prestigious research projects. The Department is also in charge of the study programmes Ecological and Evolutionary Biology, and Teaching of Biology at Secondary Schools, which cover Bachelor's, Master's, and Ph.D. degrees. The students, professors and researchers, have at their disposal the department's herbarium, with over 597.000 specimens, along with the zoological collection, containing approximately 1500 vertebrate specimens and about 5000 items of invertebrates.[2]

Department Adress:
University Campus Bohunice
Buildings A31, A32
Kamenice 5
625 00 Brno-Bohunice

[1] Source: Masaryk University Homepage [10]
[2] Source: Dep. of Botany and Zoology Homepage [9]

1. Introduction

1.1 Context of the subject

For a long time, community ecology has mostly been a descriptive science. However, since the 1950's, it has become a field that can be analysed statistically, and that is therefore suitable for hypothesis testing. Today, various numerical methods exist, aiming at answering ecological questions, identifying and describing ecological factors, and providing an overall methodology for the use of ecological data. LEGENDRE, P. and LEGENDRE, L. produced an extensive inventory of the available numerical methods, which can be found in the third edition of their work *Numerical Ecology* [3].

Ecological data may regroup information about species, samples/sites, and environmental variables. Such data set is accordingly large, and applying mathematical procedures to it can prove to be a tricky task. The help of computers is therefore very welcome, in regards of management and computing power.

Since 1998, the Department of Botany and Zoology at Masaryk University Brno has been developing its own software package, designed to be used with phytosociological data, and called *JUICE*. The data-sets that can be manipulated thanks to this program, take the form of a community matrix, gathering relevés and the associated species data (mostly abundance-related variables, such as cover or frequency). In addition, so called header-data may contain information about environmental factors, e.g. pH, layer, slope, humidity etc.[1]

Besides managing community data, *JUICE* provides powerful tools for vegetation data analysis, among others classification methods, calculation of associations and ordinations. The latter can mainly be achieved thanks to the connection of *JUICE* with R[2]. In fact, R makes use of a specific package, called *ordijuice*, regrouping various functions enabling R to import data from *JUICE*, and to calculate/plot various ordinations. This particular package was created in 2006 by David Zelený, and updated by him ever since.

[1] TICHÝ, L. - *JUICE, software for vegetation classification* [5]. Additional information can be found on the JUICE Homepage [13]

[2] See R-Project Homepage [8]

1.2 Objectives

The focus of this internship is to detail the previously mentioned *ordijuice* package. It contained all the necessary functions to perform DCA (Detrended Correspondance Analysis), PCA (Principal Component Analysis) and NMDS (Non-Metric Dimensional Analysis) on the data-output from *JUICE*. The results of the ordination can then be represented in two-dimensional or three-dimensional space. The objectives of this project, as fixed at its beginning, are the following:

Table 1.1: Main objectives of the internship, ordered from most essential to most optional.

1. Update the code of the *ordijuice* package, so that it will work in the most recent R-version.[a]
2. Add new ordination-methods to the package, as well as other improvements, such as:
   (a) Hellinger transformation for PCA
   (b) PCoA (Principal Coordinate Analysis)
   (c) CA (Correspondance Analysis)
   (d) Various similarity/dissimilarity indices (Bray-Curtis, Manhattan, Jaccard, Simpson, ...) for PCoA and NMDS
3. Create documentation for this package, so that it can be uploaded directly to the CRAN website (as this would greatly improve user-friendliness, and possibly allow automatic updates for future R-versions).[b]
4. Improve the graphical representation of the results obtained by the selected ordination method, e.g. using the R-packages *ggplot2*[c] or *plotly*[d].

[a] Full functionality was previously ensured up until R-2.11.0. During April 2017, R-3.4.0 was published, and became therefore the desirable reference.
[b] Currently, the binary packages for *ordijuice* are hosted on the servers of Masaryk University.
[c] See Wickham, H. *ggplot2 - Elegant Graphics for Data Analysis* [6]
[d] See Plotly Homepage [12]

1.3 Approach

Given the sometimes unpredictable nature of the outcome of code-writing, it was difficult to establish a precise time frame for each of the objectives cited above. Indeed, the code contained in the *ordijuice* package has to comply with the inherent grammar of R, the logic behind *JUICE*, and the user's operating system.

Since time was the limiting factor here, it was decided to tackle each problem in decreasing order of urgency. With the aim of establishing this order, I was given the opportunity to present this project during one of the department's weekly meetings, and to engage the attendants into a discussion/brainstorming session afterwards.

The result of this meeting is the above list of items, the first one being the most essential, and the last the most optional.

In order to achieve the aforementioned objectives, it was necessary to acquire some basic theoretical background about mathematical methods applied to ecological data. In fact, solving a problem was basically a three-step procedure:

1. Research of theoretical background about the problem
2. Research of ways to solve it
3. Research of ways to implement the solution

This procedure allows for maximal rigour (which is necessary for code-writing), while highlighting the mechanics of the numerical methods and their implementation in R. It was therefore the chosen approach for handling the challenges brought about by the objectives of the internship.

2. Protocol

2.1 Numerical Ecology

Since the internship involved a lot of research for each objective, this section only aims at introducing the theoretical *frame* of it. For the sake of clarity, specific methods and concepts will be presented along their practical use/implementation in the context of this work. In other words: after a quick overview of the ecological data and the *JUICE* software, followed by an introduction to ordination, the protocol will track the chronological progress of this project.

2.1.1 *JUICE* and ecological data; an overview

As previously stated[1], *JUICE* allows the user to manage phytosociological data sets, and analyse them. Those datasets are ecological community matrices, illustrated in Table 2.1.

The data contained by each cell is a species abundance value, i.e. density, biomass, cover in percentage, or cover as code (as defined by the Braun-Blanquet scale). Additional data, such as layer, pH, coordinates, altitude etc., can be stored outside of the community matrix, in the so-called headers; they can be used by some classification/ordination methods, or be useful for bioindication (e.g. Ellenberg's Indicator Values).

Table 2.1: Community matrix, where each row i represents a species, and each column j a sampling site/relevé. The cells *(i,j)* contain an abundance value taken by species i for a sampling site j.

Ecological data matrix					
			Descriptors		
Objects	$\mathbf{y}_1$	$\cdots$	$\mathbf{y}_j$	$\cdots$	$\mathbf{y}_p$
$\mathbf{x}_1$	$y_{1,1}$	$\cdots$	$y_{1,j}$	$\cdots$	$y_{1,p}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$\mathbf{x}_i$	$y_{i,1}$	$\cdots$	$y_{i,j}$	$\cdots$	$y_{i,p}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$\mathbf{x}_n$	$y_{n,1}$	$\cdots$	$y_{n,j}$	$\cdots$	$y_{n,p}$

[1] See section 1.1, p.1

2.1.2 Ordinations

Analysing such data set, means studying the simultaneous influence of numerous environmental factors on a large number of species. The methods devised for such a task are derived from multidimensional statistics, and are designed to evaluate complex data that goes beyond unidimensional normal distribution, but they also take into account the covariances among the descriptors of the data, revealing this way its underlying structures. Most of these approaches can be assigned to the following two groups: classification and ordination.

- Classification (or clustering) consists in partitioning the species or the sample units into subsets, using pre-established rules of agglomeration or division.
- Ordination is an operation by which the data is represented in a space that contains fewer dimensions than the original data set. In practice, this means ordering the species or sample units along gradients; these gradients form the axes of multidimensional space, where the data is arranged into. Thus, the coordinates of each point in this newly created space, reveal how much this given piece of data is affected by the gradients. According to the type of ordination, this multidimensional space can then be projected into a two- or three-dimensional space.

Indirect gradient analysis

It is possible to create categories of ordinations according to their premise, as shown on p.6, in table 2.2. First, it is important to distinguish Indirect gradient analysis, which takes only the samples by species matrix as input; information about the environment may be used afterwards, in order to interpret the results.

> "When we perform an indirect analysis, we are essentially asking the species what the most important gradients are."
>
> - Palmer, M. [11]

Direct gradient analysis

On the other hand, Direct gradient analysis uses external environmental data besides the species data. Put simply, Direct gradient analysis comes down to performing a regression, in order to explain species composition through the available environmental variables. This method allows for statistical hypothesis testing, where H_0 would be that species composition is unrelated to measured variables. However, Direct gradient analysis and the ordination techniques that fall into this category were not addressed in detail during this project.

Table 2.2: Categories of common ordination techniques (derived from ter Braak, C. & Prentice, I. *A Theory of Gradient Analysis* [4])

1. Indirect gradient analysis
   - (a) Distance-based approaches
     - - **PO** (Polar ordination)
     - - **PCoA** (Principal Coordinates Analysis, *aka* Metric multidimensional scaling)
     - - **NMDS** (Non-metric multidimensional scaling)
   - (b) Eigenanalysis-based approaches
     - i. Linear Model
       - - **PCA** (Principal Component Analysis)
     - ii. Unimodal model
       - - **CA** (Correspondance Analysis, *aka* Reciprocal Averaging)
       - - **DCA** (Detrended Correspondance Analysis)
2. Direct gradient analysis
   - (a) Linear model
     - - **RDA** (Redundancy Analysis)
   - (b) Unimodal model
     - - **CCA** (Canonical Correspondance Analysis)
     - - **DCCA** (Detrended Canonical Correspondence Analysis)

Distance-based techniques

Distance-based ordinations do not take the whole community matrix as input; they rely on a distance matrix, which is square (and mostly symmetric), and where rows and columns usually represent the samples. This matrix contains values that are representative of the difference between the samples. Those values could be an Euclidean distance[1], or one of many similarity/dissimilarity indexes, which will be addressed in section ..., p. ... One drawback of those methods, is the fact that the species information isn't utilized to perform the ordination. Therefore, distance-based ordinations "hide information".

[1] Straight line distance between two points in a Cartesian coordinate system. In two dimensions, the Euclidean distance is:

$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Eigenanalysis-based techniques

Ordination aims primarily at generating a small number of variables, each one explaining a large portion of the variation. Yet, most ecological data sets contain many variables that are associated (e.g. linearly correlated) to one another.

> "The basic idea underlying several methods of data analysis is to reduce this large number of intercorrelated variables to a smaller number of composite, but linearly independent variables, each explaining a different fraction of the observed variation." [1]

Mathematically, this boils down to "simplifying" a square matrix (a distance/similarity matrix) by finding an equivalent diagonal matrix. Please see Appendix A for an introduction to these ordination methods.

2.2 The *ordijuice* R-package

2.2.1 Internal structure

The *ordijuice* R-package focuses, as its name suggests, on ordination methods. It is designed to perform those ordinations on the data output generated by *JUICE*, in a non-interactive R-session (i.e. the user of *JUICE* doesn't have to give R any input). Indeed, the desired ordination can be selected within *JUICE's* interface, and the software communicates with R in the background through *ordijuice* in order to produce an ordination plot. To do this, *ordijuice* regroups 22 functions. A simplified strucuture of *ordijuice*, with the most essential functions, can be consulted in figure 2.1.

As stated in the beginning, the first objective consisted in updating the code-writing of *ordijuice*. Several new R-versions have been published since the creation of this package in 2006, and with them updates of other packages used in *ordijuice*. Hence, some portions of its code had to be rewritten, so that they would comply with the new syntax of the most recent package- and R-versions. Making a list of all the adjustments seems pointless; nevertheless, the whole procedure of detecting errors, finding their origin(s) and providing a fix, took almost a week to complete. At the end of it, *ordijuice* had been successfully rendered compatible with the current package versions and R-3.4.0.

[1] See Legendre, P. & Legendre, L. *Numerical Ecology*, section 2.9 [3]

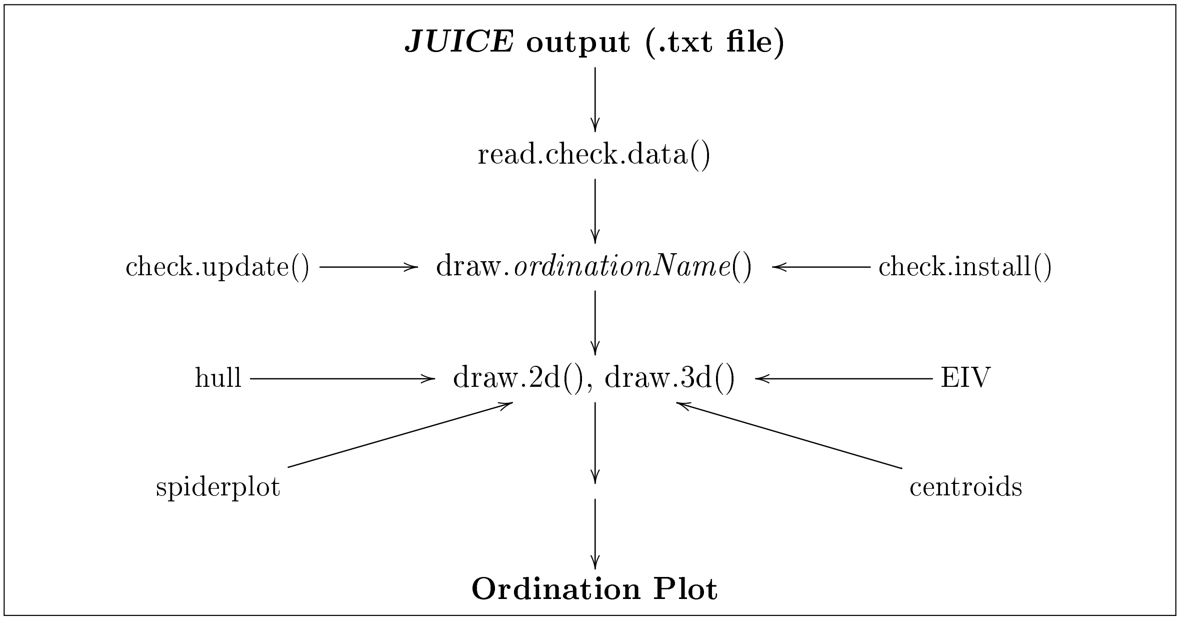


Figure 2.1: Simplified structure and function calls of the *ordijuice* package.

2.2.2 Ordinations in R

PCA - Principal Component Analysis (Basic feature)

PCA is the "simplest" eigenanalysis-based technique; it consists in rotating the original data matrix, projecting it onto a new set of axes, in a way that these (uncorrelated) axes extract a maximum of the data's variance. The graphical representation of this is achieved in *ordijuice* thanks to the command:

```
prcomp(x, ...)
```

where x is the species data exported from *JUICE*. The obtained Principal Components are then plotted thanks to the *draw.2d()* or *draw.3d()* functions.

Often, vegetation data presents unequal proportions of species data; the data sets contain a lot of zeros. This can lead to the so-called *double-zero problem*, whose effect makes differences between samples/sites less meaningful. One way to solve this problem, is the Hellinger transformation, which I have implemented during the internship.[1]

CA - Correspondance Analysis (New feature)

This method, also known as "Reciprocal Averaging", maximizes the correspondence between species and sample scores. Those scores are obtained for each row and column of a community matrix, and calculated in such way that the correlation between species and sample is extracted, and weighed accordingly to species abundance. In R, it can be calculated thanks to the command:

[1] See subsection 2.2.3

```
vegan::cca(x)
```

DCA - Detrended Correspondance Analysis (Basic feature)

Similar to CA, DCA provides a solution for the distortion that can be observed after CA; indeed, this is the result of the so-called *arch-effect*, and the compression of the axis extremes. It is achieved in *ordijuice* with the function:

```
vegan::decorana(x, iweigh=T/F)
```

The argument `iweigh` is a logical operator, indicating whether rare species should be weighed down or not; it is possible to do this in the *JUICE* interface as well.

PCoA - Principal Coordinates Analysis (New feature)

PCoA is an ordination technique that represents a distance between the samples. As stated before, PCoA takes a distance matrix as input; the algorithm then maximizes the linear correlation between these distances, and projects the result into a low-dimensional space[1]. *ordijuice* uses the following command for it:

```
cmdscale(vegan::vegdist(x, method), k)
```

The argument `k` of the `cmdscale` function specifies the dimensions of the space into which the data are to be represented in. Function `vegdist` from the *vegan* package, allows the use of various dissimilarity indices[2], specified thanks to the `method` argument.

NMDS - Non-metric Multidimensional Scaling (Basic feature)

NMDS attempts to rectify the flaws of PCoA (especially the *arch-effect*), by maximizing rack order correlation, instead of linear correlation. In R:

```
vegan::metaMDS(x, distance, k)
```

`distance` allows to specify a dissimilarity index, while `k` represents the number of dimensions.

2.2.3 Data transformation and dissimilarities

According to the type of data in a community matrix and the choice of an appropriate ordination technique, one might want to transform the data, in order to eliminate unwanted side-effects or the infamous *double-zero problem*. Indeed, when using Euclidean distance with large and raw abundance data, absent species from two sites lead to both sites being more similar, rather than different. This can be avoided using various indices/distances/transformations; therefore, it seemed necessary to include those possibilities in *ordijuice*.

[1] If these distances are euclidean, PCoA is equivalent to PCA.

[2] See subsection 2.2.3

Hellinger distance

Legendre, P. & Legendre, L. demonstrated that the Hellinger transformation offers the best compromise between linearity and resolution, thus making it suitable for linear ordination (PCA, RDA) [2]. It is defined as follows:

$$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}} \tag{2.1}$$

y_{i+} represents the sum of the values in column (= sample/site) j.

In *ordijuice*, the Hellinger transformation has been made available for PCA, which can therefore be used as tb-PCA (transformation based PCA). The following code has been added to the *read.check.data()* function, executed prior to the ordination method:

```
if (hellinger.trans) {
loadNamespace("vegan")
spec.data <- vegan::decostand(spec.data, method="hellinger")
}
```

The `hellinger.trans` argument is a logical operator, provided by *JUICE*, and `spec.data` is the species data exported from it.

Dissimilarities

For distance based ordinations, i.e. techniques that take a distance matrix as input (PCoA, NMDS), it is possible to replace the Euclidean distance measure with various indices representing the difference between samples/sites as well. In such cases, the distance matrix has been transformed into a similarity/dissimilarity matrix. The difference between a distance and a (dis)similarity, is the fact that the latter do not satisfy triangle inequality. It was our goal to make the most commonly used dissimilarities in ecology available to the end-user of *JUICE*; the following indices and the associated R code have therefore been added to the *ordijuice* package, in order to be used with PCoA and NMDS:

- Bray-Curtis
- Jaccard
- Manhattan
- Simpson

The first three dissimilarities could be specified using the following function:

```
vegan::vegdist(x, method=distance)
```

with `distance` being a string, such as `"Bray-Curtis"`, `"Jaccard"`, `"Manhattan"`.

However, the Simpson dissimilarity (aka overlap-coefficient) being not available in this function's repertoire, it had to be added manually, according to its definition:

$$overlap(X,Y) = 1 - \frac{|X \cap Y|}{min(|X|,|Y|)} \tag{2.2}$$

In R, it was necessary to create a function that was able to compute the Simpson dissimilarity; this function could then be called by `metaMDS()` (for NMDS) or `vegdist()` to be used on the data:

```
simpson <- function(x) {
            (vegan::designdist(x,
            method="1-((J)/(min(A,B)))",
            terms="minimum"))
            }
```

2.2.4 Documentation and package conformity for CRAN upload

In order to upload a package to the CRAN repositories, some rules must be followed. The first and most essential one, is providing the package with appropriate documentation, so that its user can acquire some deeper knowledge about how the package does what it does. However, *ordijuice* being used in a non-interactive R-session, the documentation would hardly be useful. Nonetheless, documentation is a requirement for CRAN, so Hadley Wickham's package *roxygen2* [16] was used to generate documentation for each and every function of *ordijuice*. The main advantage of *roxygen2* is the fact that it creates the documentation files from specially formatted comments situated at the beginning of a function. It is also fully embedded in R-Studio, which makes it an easy and accessible tool for package creation. Example of documentation for the function *check.install()*:

```
#' Check for the correct installation of necessary packages for ordijuice.
#'
#' \code{check.install} checks if the required packages are installed, and
#' downloads the missing ones. It also checks for a new JUICE-version, as
#' well as the current version of R.
#'
#' @param display.spider Logical, provided by the JUICE software
#' @return A logical, indicating whether the user's R-version is up to date or
#' not. If the R-version is inferior to 3.2.2, ordijuice will ask the user to
#' update to the latest R-version in order to ensure full functionality.
#'
#' @seealso \code{\link{install.libraries}} \code{\link{check.update}}
```

Finally, it was also necessary for the package to pass the R CMD check without throwing any error. Checking a package in R-Studio is an easy task, with the integrated developers tools. Besides the documentation, a package needs to present a proper NAMESPACE and DESCRIPTION file. On one hand, the first one is used to specify the package-dependencies and the functions that are to be exported from the source-files into the active R-session, to become available for use. On the other hand, the latter is used to provide CRAN (and the future users of the package) with the necessary information on what the package does, and how. As of now, the *ordijuice* package still needs some minor improvements (especially function calls from other packages need to be properly written using the "package_name::function_name" syntax); at the end of the internship, the number of warnings produced by the R CMD check procedure has been significantly reduced to the number of 2. I will continue troubleshooting the code after the internship, in order to finalize the whole endeavor.

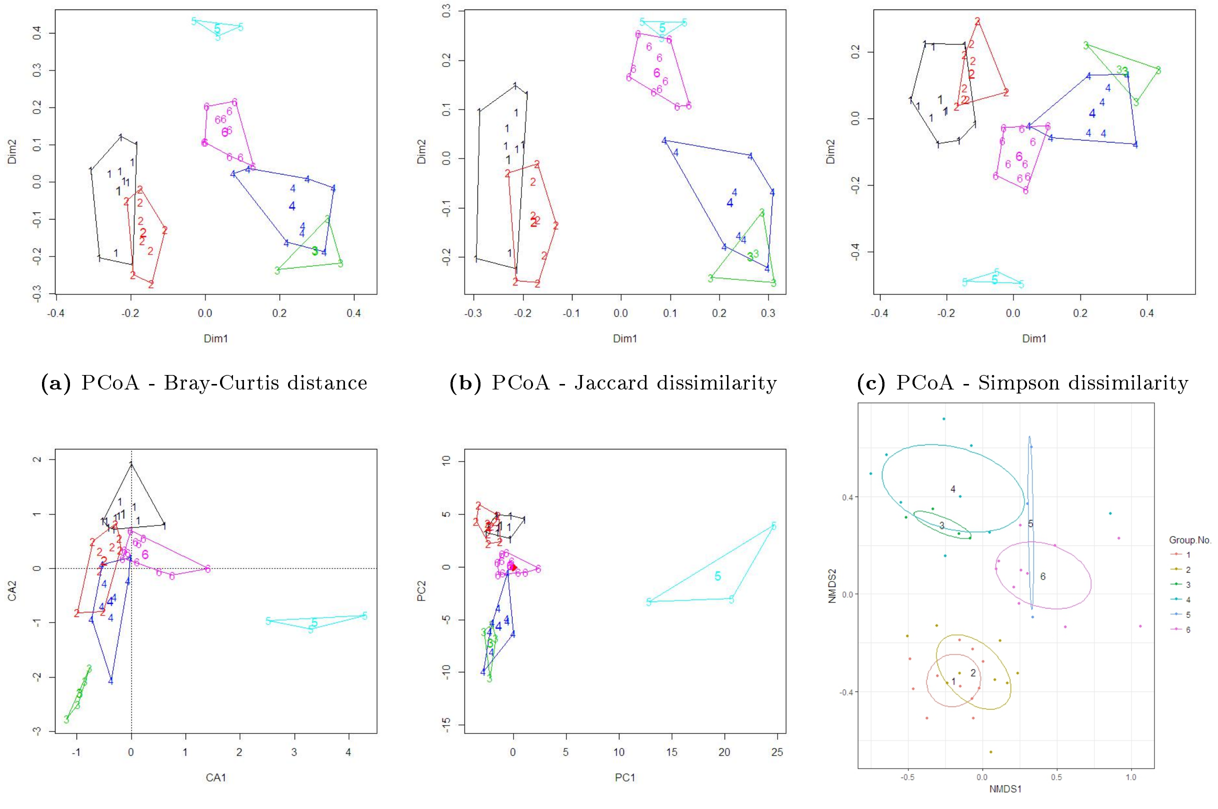
2.2.5 Graphical Improvements

Although not finished, I have been able to produce prototypes of ordination graphics with *ggplot2*; it is still necessary to check how they would fit in in the package as a whole, and adjust it accordingly. However, the results are encouraging, and I will pursue the implementation of such graphics in the future.

3. Results

At the end of the internship, Lubomír Tichý was able to distribute a new version of *JUICE* on its website, containing the following new features: Updated functionality for R-3.4.0, PCoA and CA, as well as a total of 5 new distances and dissimilarities for ordinations.

Furthermore, the documentation and description files for the package have been created and finalized, although they have not been deployed; in the coming weeks, it should be possible to eliminate the remaining warnings that occur when checking the package, and send it to CRAN. I will also be able to embed the new *ggplot2* graphics in the code of *ordijuice*, so that the produced plots may serve a higher purpose than merely previewing the results of the ordination.



(a) PCoA - Bray-Curtis distance **(b)** PCoA - Jaccard dissimilarity **(c)** PCoA - Simpson dissimilarity

(d) CA **(e)** PCA - with Hellinger transformation **(f)** NMDS - using ggplot2

Figure 3.1: Graphical representations of the added features

4. Discussion

It is difficult to confront the hereby obtained results with previously lead work; ordinations are a relatively recent tool, and so are informatics. Moreover, the existing literature was mostly used as a blueprint (or even as a model) to implement various procedures. In other words, this project did not aim at providing new insight into this subject, but more a practical solution to making it more accessible.

However, the *JUICE* software and its associated R-package *ordijuice* can be compared to other programs, performing similar tasks, such as *Canoco*[1]. Indeed, the latter is specialized in multivariate statistical analysis using ordination methods. Compared to it, *JUICE* is non-commercial, and offers also plenty of tools and possibilities to edit the phytosociological data prior to analyzing it.

In short, the project carried out during this internship should not be seen as innovative work, but more as a convergence of ordination techniques and computer science. In particular, it could therefore be described as a natural evolution of the *JUICE* software and its features, and as a means of spreading the use of ordination methods by making them available without expecting much programming-knowledge from the user.

There wasn't any substantial obstacle to tackle; indeed, most of the research provided practicable solutions to the problems at hand. I should however mention that more time than initially expected went into troubleshooting fatal errors caused by compatibility conflicts between package-versions, R-versions and modified R source code.

[1] See the Canoco5 Support Site for further information [14]

5. Conclusion & Personal Appreciation

5.1 Conclusion & Perspectives

At the end of this period spent in Brno, significant progress has been made on *ordijuice*. More ordination techniques have been made available, as well as more options to carry them out. Moreover, progress has been made to provide new graphical representations, and the R-package is almost ready to be sent to CRAN. In this manner, *JUICE* has gained new features and improved user-friendliness. Eventually, with progressing improvements, the software may one day be officially released and distributed.

I will stay in touch with Lubomír Tichý and David Zelený; I have been granted some exclusive knowledge about *ordijuice*, and I would like to use it to complete the previously mentioned advancements, and pursue its general improvement.

5.2 Personal Appreciation

I have thoroughly enjoyed my stay in Brno. I was able to learn a lot about ecological data and how to use it; I have gained substantial insight on ordinations and mathematical transformations, as well as advanced programming in R, and how to create an R-package. On top of that, I have discovered that the research-based nature of the work carried out here suited my preferences completely; I was able to solve problems and implement their solutions independently and at a convenient pace. In other words, not only have I acquired new and useful knowledge in the field of ecology, but I also felt like I was being very efficient, and that I could account for proper progress every day.

5. Appendix

A. Eigenanalysis-based ordination methods

The first step of eigenanalysis aims at obtaining variables that are no longer intercorrelated. Let **A** be a square distance/similarity matrix:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & & & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix} \tag{5.1}$$

Λ is the desired diagonal matrix, such as:

$$\Lambda = \begin{bmatrix} \lambda_{1,1} & 0 & \cdots & 0 \\ 0 & \lambda_{2,2} & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \lambda_{n,n} \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \tag{5.2}$$

Λ is then the *matrix of eigenvalues* (which are also called *characteristic roots*)[2]. It describes the distance/similarity of a new set of variables, which are called *eigenvectors* (or *characteristic vectors*). Since the terms outside of the diagonal are equal to zero, these variables are linearly independent of one another[3].

It is possible to determine the eigenvalues λ_i and the eigenvectors u_i of matrix **A** with the help of the following equation:

$$Au_i = \lambda_i u_i \tag{5.3}$$

Proof and examples of this equation are beyond the scope of this report. Suffice it to know that the square matrix of eigenvectors $U = [u_1 \cdots u_n]$ is the input for Eigenanalysis-based ordination techniques. Indeed, each eigenvector will give birth to an ordination axis, that is associated to its

[2] Matrix Λ is also known as the *canonical form* of matrix A.

[3] Reminder: a set of vectors $(v_i)_{1 \leq i \leq n}$ is composed of linearly independent vectors if, in a vector space E:

$$\forall (a_1, ..., a_n), \quad (a_1 v_1 + \cdots + a_n v_n = 0_E \Rightarrow a_1 = a_2 = \cdots = a_n = 0)$$

eigenvalues. It is important to note that, according to the chosen ordination, those eigenvalues can be manipulated, and be given further mathematical meaning. The ordination axes are finally ranked by their eigenvalues, and the higher the eigenvalue, the more this axis is representative of underlying structures of the data set. In most cases, two or three axes suffice to explain a majority of the values taken by the data.

This is how Eigenanalysis-based ordination can successfully represent relationships within an ecological data set in a low-dimensional space.

Bibliography

[1] Gauch, H. G. *Multivariate Analysis in Community Ecology.* Cambridge University Press, 1982.

[2] Legendre, P., and Gallagher, E. D. Ecologically meaningful transformations for ordination of species data. *Oecologia, vol. 129* (July 2001), 271–280.

[3] Legendre, P., and Legendre, L. *Numerical Ecology.* Elsevier, 2012.

[4] Ter Braak, C. J., and Prentice, C. A theory of gradient analysis. *Advances in Ecological Research, vol. 18* (1988), 271–317.

[5] Tichý, L. Juice, software for vegetation classification. *Journal of Vegetation Science, vol.13, Issue 3* (2002), 451–453.

[6] Wickham, H. *ggplot2 - Elegant Graphics for Data Analysis - Second Edition.* Springer, 2016.

Webography

[7] Clark, C. An introduction to ordination.
http://online.sfsu.edu/efc/classes/biol710/ordination/ordination.htm.

[8] CRAN. The comprehensive r archive network.
https://cran.r-project.org/.

[9] MU. Department of botany and zoology homepage.
http://botzool.sci.muni.cz/en.

[10] MU. Masaryk university homepage.
https://www.muni.cz/en.

[11] Palmer, M. Ordination methods for ecologists.
http://ordination.okstate.edu.

[12] ThePlotlyTeam. Plotly.
https://plot.ly/.

[13] Tichý, L. Juice version 7.0.
http://www.sci.muni.cz/botany/juice/.

[14] Šmilauer, P. Canoco5 support site.
https://http://canoco5.com/.

[15] Wickham, H. ggplot2.
http://ggplot2.org/.

[16] Wickham, H. Introduction to roxygen2.
https://cran.r-project.org/web/packages/roxygen2/vignettes/roxygen2.html.

[17] Zelený, D. Analysis of community ecology data in r.
http://www.davidzeleny.net/anadat-r/doku.php.

Summary

From April to May 2017, I carried out an internship in the department of Botany and Zoology of Masaryk University Brno, Czech Republic. During this period, I was in charge of updating and adding new features to an R-package called *ordijuice*, which links R and *JUICE*, a software for vegetation analysis, created by Lubomír Tichý. The package uses R to perform various ordination techniques on the data from *JUICE*. Its code had to be updated to comply with new R-versions, and new features were to be added. At the end of this seven week period, new ordination techniques were implemented in *ordijuice*, as well as various dissimilarities. Finally, the package had been modified to fit the requirements of CRAN, in order to be uploaded to its servers. Improvements of the graphical outputs have also been initialized.

Keywords

R, Juice, ordinations, ecological data, Czech Republic