**An integrative analysis of the environmental structuring of fish communities in Doubs River system.**

Bevan Vollenhoven (3840572)

University of the Western Cape

Robert Sobukwe Rd, Bellville, Cape Town, 7535



**Bevan Vollenhoven (3840572)**

**Quantitative Ecology (2021)**

**An integrative analysis of the environmental structuring of fish communities in Doubs River system.**

**Table of contents**

1. Introduction
2. Aims and Objectives
3. Methods and Materials
   1. Species data and Environmental variables
   2. Ordination plots
   3. Cluster analysis
4. Results
   1. The principal coordinate analysis
   2. Distance-Based Redundancy Analysis (db-RDA)
   3. Cluster Analysis
5. Discussion
6. Conclusion

References

**Abstract**

Running water ecosystems are vulnerable yet understanding on the drivers of species richness and structure remains challenging for ecologist. A study done by Verneaux et al. (2003) found an effective method in interpreting monitoring running water species data. We have tried to use laternative multivariate analysis methods to test their efficiency in showing the same or similar results. It was established that a PCoA, cluster analysis, and a distance-based RDA were able to effectively show the community and species structure as a function of the underlying environmental conditions.

1. **Introduction**

Running water ecosystems contain high levels of biodiversity (Vörösmarty et al. 2010), yet they are one of the most vulnerable habitats due to their extensive exposure to anthropogenic disruptions (Xiong et al. 2016). Having a method effectively identify communities and species structure within river systems will allow for a more efficient methods of conservation and rehabilitation of disturbed river systems. Multivariate analysis of ecological data allows us to such a thing. There are several methods that can used depending on the type of data and what exactly you want to analyse.

Understanding drivers of species richness and variation can be challenging for theoretical ecologists and conservationists (Brown et al., 2019). The current study was based on a study done by Verneaux et al. (2003) regarding the species data monitoring in running waters. In their study, they sort to produce a dataset that can be used to in comparative analyses. They done this through the production of a general species continuum, typological species traits used to establish a general distribution pattern of species, and a theoretical watercourse comprising of reference sampling sites from different rivers. This study will be used as a principle study in order to compare other methods of analysing ecological data.

One of these includes clustering. Clustering aims to recognize discontinuous subsets present in the environment which can be discreet within taxonomic data (Borcard et al., 2011). Clustering data deals with partitioning collections of descriptors, without testing for an hypothesis (Borcard et al., 2011). While the use of cluster analyses can assist in revealing hidden data (at the discretion of the ecologist) (Borcard et al., 2011), ordination methods may reveal the presence of clusters more effective than standard clustering procedures (Anderson, 1971).

Ordination methods represent data along a reduced quantity of orthogonal (uncorrelated) axes representing the main trends of the data in descending order (Borcard et al., 2011). The trends produced can be visually interpreted or in collaboration with clustering or regressions (Borcard et al., 2011). The Principal coordinate analysis (PCoA), unlike the PCA used in the principal study, uses both Euclidean distance and chi-square distance to determine the distance amoung objects (Borcard et al., 2011). Ordinations can be constrained (direct gradient analysis) or unconstraint (indirect gradient analysis). PCoA is based on an association matrix, allowing it to directly depict relations among objects or variables (Borcard et al., 2011).

1. **Aim and objectives**

Since there are various methods used for multivariate data analysis, our aim is to validate alternative statistical methods for analyzing running water ecosystems and to establish if they are as affective as the methods used in the principal study. The main objective of the study is to observe how efficient the alternative methods and if they are appropriate in analyzing river ecosystem data.

1. **Methods and Materials** 
   1. **Species data and Environmental variables**

Two sets of data were used from Verneaux et al. (2003). The first set of data comprised of species abundance data. This dataset contained abundances data of both insect and fish species present in each sample site along the Doubs River system. There were, in total, 27 species recorded across 30 sites. The original species abundance matrix was weighted against five classes of relative density adapted to the highest values in each sample of each genus for all the species. Sites that possessed a biotic index great or equal to 18/20 was deemed as not or slight disturbed between 1970 to 1976, allowing them to be selected as the reference sites. All species present in the matrix were present in the sample of the reference sites. Chosen sites possessed the highest relative species richness along the upstream-downstream gradient (Verneaux et al. 2003).

Environmental variables were obtained by efforts made to collect data to analyse the potential deteriorating water quality at the time the study was conducted. There was a decline in species abundance in the reference sites, which led to the decrease in sites having a biotic index equal to or more than 18/20. Data collected on environmental period were restricted to the period from 1970 to 1976. Data analyses were conducted using R studio ().

* 1. **Ordination plots**

Ordinations can be used to represent data along a reduced number of uncorrelated (orthogonal) axes. They are constructed in a manner as to allow the representation of the main trends or gradients within the data, in a decreasing order of ‘importance’. Data that contain few gradients or trends will have their most useful information regarding their variation present within the first few axes (Borcard et al., 2011).

The data matrices were analysed by Principal Coordinate analysis (PCoA) (Gower, 1966). A PCoA can produce a Euclidean representation of measured relationships between groups through a chosen distance or similarity measure (Borcard et al., 2011). The species data of the Doubs River data was converted to a dissimilarity matrix and PCoA was produced using the capscale() function from the vegan package (Oksanen et al., 2017). Through this, the relationship that species composition had on site similarities could be analysed. An indirect analysis was applied on the effect of the underlying environmental conditions in order to understand its influence on community structuring.

Constrained ordinations can be referred to as “direct gradient analysis”. These ordinations summarise and extract variation in a group of variables that are influenced by constraints such as environmental properties present. A distance-based Redundancy Analysis (db-RDA) is seen as an extension of a PCA (which was done in the principal study). This analysis models species data as a function of the underlying environmental data. RDA is a type of multiple regression analysis, as it uses both response variables and explanatory variables. The species data were loaded, and the 1st column and 8th row were removed. This was done in order to remove rows that held no species data and remove column that held no important data. The function vegdist(), belonging to package vegan, was used to produce a Bray-Curtis dissimilarity matrix. A db-RDA was done in order to analyse species and site relations as a function of environmental variables. Collinear environmental variables were removed as to using variance inflation factors (VIF) (Dormann et al., 2013) and a subset of non-collinear variables was retained. VIF values that were greater than 10 were removed. The significance of the model was tested using ANOVA and a p-value less than 0.05. Significance was tested for both the db-RDA axis and the environmental variables in order to determine the significant values that will be used in the ordination plot.

* 1. **Cluster Analysis**

Trends from ordinations can be used in conjunction with clustering and regressions. Three cluster analyses were conducted in the Doubs River system species and environmental data. Single Linkage Agglomerative Clustering agglomerates objects on the basis of their shortest pairwise distances (or greatest similarities). In order for two groups to be connected on a similar dissimilarity level, only one of the objects in the two groups need to be linked in order to aggregate the groups at the same level of (dis)similarity. This common hierarchical method allows gradients to be interpreted clearly but causes difficulty in the interpretations of partitions (Borcard et al., 2011). The function hclust() from the cluster package (Maechler et al., 2017). Two different dendrograms were produced. One for a single linkage agglomerative clustering and one for a complete linkage agglomerative clustering. Complete linkage clustering only forms groups with objects that has a distance corresponding to the furthest object in the group.

A k-means cluster was also produced using the fviz\_cluster() function from the package cluster. Prior to this, as k-means requires you to input the number of clusters, three methods of finding the required of number of clusters were used: Gap statistic, silhouette method, and the elbow method. These cluster analysis, along with the previous ordinations, were used to analyse and interpret the integrative view of the environmental structuring of the fish community.

1. **Results** 
   1. **The principal coordinate analysis**

It was found that the first three ordinates (MDS1 to MDS3) explained 77.98% of the variation in the eigenvalues. Sites ordinated closer to one another in the PCoA biplot are considered more similar than sites ordinated further away (Borcard et al., 2011). In PCoA biplot scaling 1 (Fig 1A) for MDS1 and MDS2, relationships between sites were emphasized. Sites ordinated on the left side of the plot are more similar due to their similar species composition. The main species contributing to their similarity is Satr, Phph, and Babl. Ththm Teso, and Cogo, also contributed to the similarity between these sites, however they are not as "powerful" of an influence on the species composition. The sites ordinated to the right are more similar to their shared species composition, which is made up of the remaining species ordinated in the same direction. Notably, sites 22-24 are more similar to one another than the other sites ordinated to the right. In figure 1 B, the PCoA biplot scaling 2 for MDS1 and MDS2, the relationship between species, the effect it has on site dissimilarity was observed. Satr, Phph Babl, Ththm Teso, and Cogo are all positively correlated with one another opposed to all the remaining species. As a result, species ordinated in the left quadrants on the biplot have more abundances of these species. All the sites to the right of the ordination have low abundances and are negatively correlated with these species. The rest of the species, are more negatively correlated with these species, signifying that they have lower abundances in the sites ordinated to the right. The sites ordinated to the right are positively correlated with the remainder of the species. Sites 22 - 23 however, are not as strongly correlated with majority of the species plotted to the right. These sites are most strongly positively correlated to species Alal.

Chart, scatter chart

Description automatically generated

Figure 1: The ordinations obtained from the PCoA. A) shows the ordinations produced through scaling 1 (site focused) for MDS1 and MDS2, B) shows the ordinations produced through scaling 2 (species correlation focused) for MDS1 and MDS2, C) shows the ordinations of scaling 1 for MDS1 and MDS3, and D) shows the ordination of scaling 2 for MDS1 and MDS3.

A subset of data was made in order to conduct an indirect analysis on the underlying environmental effect on species distribution and site dissimilarity (Fig 2). In the top left section of the plot, we can see that dissolved oxygen and altitude have a strong positive correlation and the sites found here are described by these conditions. The right side of the plots has strong positive correlations to flo, distance from source, nitrogen concentration, and hardness. The bottom right side of the plot has a strong positive correlation with biological oxygen demand, ammonium concentration and phosphorus concentration.

In the four subsets of data Satr, Alal, Satr, and Teso abundances were analyzed individually. In the subset data, the circles represent the level of abundance (larger circle = higher abundance). The green curves on the plots show areas with similar abundances (i.e., all points along the curve have similar abundances). Sites 22-24 have a large abundance of Alal. Alal is most abundant in both sites 22-24 and the in some of the remaining sites ordinated to the right. The environmental conditions these sites are influenced by is high biological oxygen demand, ammonium concentration, nitrogen concentration, mean minimum discharge (flo) and phosphate concentration levels. Satr, on the other hand, is most abundant in sites found the top-left quadrant. These sites are exposed to high levels of dissolved oxygen, altitude, and water pH (however pH is does not have a prominent significance on the abundance of Satr in those sites). In these plots, we can see a gradient resulting in a change from an oligotrophic environment to a eutrophic environment.

* 1. **Distance-Based Redundancy** **Analysis** **(db-RDA)**

Altitude, slope, flo, nitrogen concentration, dissolved oxygen and biological oxygen demand were the environmental variables retained through the assessment of VIF. In figure 3, sites ordinated to the left are characterized by having high slope, altitude, and dissolved oxygen. While sites ordinated to the right are characterized by having high flo, biological oxygen demand, and nitrogen concentrations. We are able to see the gradient from an oligotrophic to a eutrophic environment as we move further down the river where flo increases and slope decreases.

**Diagram

Description automatically generated**

Figure 2: The subset of the abundances of four species (Satr, Scer, Teso and Alal) in order to indirectly analyze underlying environmental conditions.

![Chart, scatter chart

Description automatically generated]()

Figure 3: The db-RDA plot showing the ordination of sites in relationship to significantly relevant environmental variables.

* 1. **Cluster Analyses**

The "height" of the vertical lines represents the distance between clusters. In figure 4A, we can view the presence of a gradient from site 1 to the last site. The clustering of sites is similar to the ordination of sites in previous ordination plots. Sites 1 -15 from one cluster (with exception to discontinuities). The sites become progressively more different until they are all groups together with being equally different to the second cluster (sites 16 -25). Sites 22-25 form a cluster on their own. This agrees with previous ordinations as these three sites were all significant differently ordinated from the other sites. Figure 4B, we are able to see more clearly how clusters are partitioned. We can the the partitions in the 4 identified clusters and the variation within them. The same cluster present in the single linkage cluster is present in the complete linkage cluster, except the only obvious discontinuity that we have is site 1.

![Diagram

Description automatically generated]()

Figure 4: The dendrograms of A) single linkage cluster showing gradients and B) complete linkage cluster used for establishing partitions.

The elbow method and the silhouette method revealed that 4 clusters were appropriate for the k-means clustering method. The gap statistic method revealed that only one cluster would be appropriate but that makes interpretations challenging, hence 4 clusters were selected for k-means clustering. In figure 5, the same sites that were we allocated as clusters in the single and complete linkage cluster, formed clusters within the k-means plot. The sites are characterized into four clusters. Sites plotted close to one another are most similar. Not only are cluster represented, but discontinuities are also shown in relation to cluster.

**Chart, bubble chart

Description automatically generated**

Figure 5: The k-means clusters (k = 4) produced for the dissimilarity species matrix.

1. **Discussion**

In this paper, the species and environmental data of the Doubs River species system continuum done by Verneaux et al. (2003). In this study, they were able to conclude that the species organization of a theoretical running water system is as a continuum of species where species replacement occurs along the upstream-downstream gradient. Verneaux et al. (2003) also projected the density classes of the upstream-downstream gradient onto a horizontal axis. They found that species with low occurrences and poorly differentiated patterns of distribution along the gradient can be justifiably removed. These species have little contributions towards the axis Produced their PCA. Species that were also plotted furthest from the structural curve can be more related to specific habitat requirements only found in specific sites. This agrees with some of the ordinations that were conducted in the present paper. In the PCoA (Fig1) we see that there the upstream-downstream gradient indicating a change from an oligotrophic to eutrophic conditions. As a result, sites containing specific species that are accustomed to specific environmental conditions are ordinated accordingly. We see in figure how some species plotted furthest away from the origin (0, 0) (Satr and Alal) (Fig 2) appeared on either end of the gradient, further justifying the results obtained in the principal study. Species plotted close to the origin are either at their optimum or mid-optimum ecological conditions or are present along the gradients (Borcard et al., 2011). In figure 2, we see that Teso, species plotted close to the origin, is most abundant along a wider range of underlying environmental conditions compared to species plotter further from the origin of the plot.

In figure 2, we can also see how species can be organized into groups based on underlying environmental conditions. Species that are more adapted to certain sets of environmental conditions will be more abundant in sites sporting the favored conditions, resulting in the sites being ordinated accordingly. This relationship is further supported in figure 3, where a direct analysis of the species abundance distribution as a function of underlying environmental conditions shows similar ordinations to the PCoA.

The principal study also concluded that species with narrower typological amplitudes are found more at both ends of the gradient. This may be due to them being more correlated with certain environmental variables. This behaviour can be seen with species Alal (Fig 2), which is strongly associated with sites 22-24 (Fig 1). This species, along with other species found at these sites, is related to a slower flowing water with eutrophic nature. Thus, causing the sites to form a cluster (Figure 5).

The upstream-downstream gradient described by previous ordinations can also be seen in figure 4a. The relations between sites and groups of sites most similar can easily be seen within the dendrogram. Specifically defining how many clusters present within the dendrogram is not an easy matter but the gradient change is represented by the continuum of sites becoming increasingly different in their respective groupings. Figure 4B make identifying partitions easier. Complete linkages clustering produces small separate groups that are spherical in multivariate space and are grouped at large distances (Borcard et al., 2011). This allows for an interesting way of identifying discontinuities within the data. The possible groupings of species proposed by the dendrogram were justified by the clustering of the sites (Fig 5). Since we’ve established site are influenced by underlying environmental variables, the groupings of similar sites into clusters signify the arrangement of species into identifiable groups.

1. **Conclusion**

The arrangement of species into identifiable groups is possible with the applications done in the principal paper. The conclusion they reached had be justified by the results achieved in the present paper. The use of ordination and clustering are viable methods of identifying discrete groups of data within running water biological monitoring. The use of both species, site, and environmental data are effective tools and can justify the patterns observe within models.

**References**

Anderson, A.J.B., 1971. Ordination methods in ecology. *The Journal of Ecology*, pp.713-726.

Maechler, M., Rousseuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., Roudier, P., Gonzalez, J., Kozlowski, K., Schubert, E., Murphy, K. (2017). *cluster: "Finding Groups in Data": Cluster Analysis Extended.* R package version 2.1.2. Available online at: https://CRAN.R-project.org/package=cluster

Borcard, D., Gillet, F. and Legendre, P., 2011. Numerical Ecology with R. New York, NY: Springer Science+Business Media, LLC, pp.1-140.

Brown, J.J., Mennicken, S., Massante, J.C., Dijoux, S., Telea, A., Benedek, A.M., Götzenberger, L., Májeková, M., Lepš, J., Šmilauer, P. and Hrček, J., 2019. A novel method to predict dark diversity using unconstrained ordination analysis. *Journal of Vegetation Science*, *30*(4), pp.610-619.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36, 027–046. doi: 10.1111/j.1600-0587.2012.07348.x

Gower, J. C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53, 325--338.

Oksanen, J., Guillaume Blanchet, F., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2017). *vegan: Community Ecology Package.* R package version 2.5-7. Available online at: https://CRAN.R-project.org/package=vegan

Xiong W, Li J, Chen Y, et al (2016) Determinants of community structure of zooplankton in heavily polluted river ecosystems. Sci Rep 6:22043. doi: 10.1038/srep22043

Verneaux, J., Schmitt, A., Verneaux, V. and Prouteau, C., 2003. Benthic insects and fish of the Doubs River system: typological traits and the development of a species continuum in a theoretically extrapolated watercourse. Hydrobiologia, 490(1), pp.63-74.

Vörösmarty CJ, McIntyre PB, Gessner MO, et al (2010) Global threats to human water security and river biodiversity. Nature 467:555–561. doi: 10.1038/nature09549