University of Cape Town

MSc. Biostatistics

Project Proposal

Multivariate Analysis Project

Author:
Sian Wood

Student Number:
WDXSIA001

March 12, 2025

Contents

1 Introduction **1**

2 Data Description **1**

3 Analysis Approach **2**

A Appendix **6**

1 Introduction

As humans continue to influence the natural environment, conservation efforts become more and more essential to the preservation of species diversity. In order for conservationists to make educated decisions regarding sites for reintroduction programs or sites at which to implement conservation laws, a thorough understanding of the environments in which indigenous species thrive is required (Peterson, 2001).

The fundamental niche of a species refers to the physical conditions under which it can survive. Due to competition with other species, a species might only be found in a subset of its fundamental niche, known as its realised niche. (Peterson, 2001). This concept means that an understanding of the ecological niche of a species must include an understanding of the ecological niches of species with which it interacts.

This project's research question is "What are the environmental factors which most strongly define the realised niches of three indigenous and two non-indigenous fish species along the Kars River in the Western Cape of South Africa?" Despite non-indigenous species not being the target of conservation efforts, an understanding of their ecological niches is necessary to understand the competition they provide to the indigenous species. Understanding the realised niches of these fish species means identifying aspects of the environment which are positively or negatively correlated with species abundance.

It is hypothesised that lower indigenous species abundance will be predicted in regions with higher predicted non-indigenous species abundance. It is also hypothesised that all fish species will be predicted to be more abundant at sites on wider, deeper and more still sections of the Kars River.

2 Data Description

This dataset originates from a study conducted by Keir Lynch in 2018. The aim of this study was to identify predictors of the presence and abundance of the Heuningnes redfin (Morch and Griffiths, 2025). No further information is available regarding the collection methods, and no literature has yet been published which uses this dataset. It was only made publicly available in January of 2025 (Morch and Griffiths, 2025).

Each observations consists of 47 measurements taken at one of 56 sites along the Kars River. The dimensions and locations of each site are not available in the dataset. Two variables are binary, indicating whether any indigenous or non-indigenous species were detected at each site. These will not be used as response or explanatory variables. The remaining variables are numeric. Six are count data for the number

Table 1: The number of sites along the Kars River at which each species was detected. There were 56 sites in total.

Species	Number of Sites
Heuningnes Redfin	16
Cape Kurper	27
Cape Galaxias	8
Spotted Bass	4
Bluegill Sunfish	3
Common Carp	0

of fish of given species that were detected at each site. Variables relating to water quality, geographic features, substrates, debris, macrophyte presence, canopy and river depth and width are available. The variables relating to substrates, debris, macrophyte presence, canopy and river depth and width are values given as the proportion of the site with a given characteristic. The range of each explanatory variable can be seen in Table A.3.

The three indigenous fish species recorded in this dataset are the Heuningnes redfin, the Cape Kurper and the Cape Galaxias. The three non-indigenous species are the Spotted Bass, the Bluegill Sunfish and the Common Carp. The number of sites at which each of these species were detected is recorded in Table 1. With no recorded sightings of the Common Carp, the presence and abundance of this species will have to be removed as a response variable in the multivariate analysis.

As the response variables are abundance data, the majority of these values are zero, indicating that no fish of a given species were observed. When species of interest were observed at a site, the number of individuals varies dramatically between 1 and 225. Summary statistics for the response variables can be seen in Table 2, and the associated histograms can be found in Figure A.1. All response variables with more than a few observations appear to follow a Poisson distribution.

Of the 56 sites at which data were recorded, indigenous species were detected at 40 sites, while non-indigenous species were only detected at six. The very low number of sites at which non-indigenous species were detected will have an impact on the statistical power of any model created using these data, and consideration should be given to this throughout the statistical analysis.

3 Analysis Approach

A multivariate approach to this analysis is advantageous as it provides greater statistical power than analysing each species individually (Wang et al., 2012).

Table 2: Summary statistics for non-zero response variable values. The Common Carp is excluded as no observations of this species were made.

Variable	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Heuningnes Redfin	1	3	7.50	38.19	34.75	225
Cape Kurper	1	2	5	63	18.50	138
Cape Galaxias	2	3.75	5	11.38	10	46
Spotted Bass	1	1	2	2.50	3.50	5
Bluegill Sunfish	2	21.50	41	52.67	78	115

The Heuningnes redfin dataset has no missing data, so no imputation or removal of sites due to missingness will be required. Very little information on this dataset is available, so the degree of processing on these data is unknown. I will therefore operate under the assumption that no observations were removed due to missingness prior to it being made publicly available.

Before any further exploration of this dataset can be conducted, it will be divided into a test and a training set, with 25% of the observations being held in the test set. An exploratory data analysis (EDA) will then be conducted on the training set, which will begin with further exploration of individual explanatory variables, with the goal of determining whether transformations are necessary. Correlations between variables will be investigated, and outliers identified.

Linear Principal Component Analysis (PCA) will then be applied to the explanatory variables in the training dataset. Due to the high dimensionality of our data (39 explanatory variables), it is unlikely that reduction to a more manageable number of variables would be possible using linear PCA, without losing a large amount of the variation in the data. The results of linear PCA will, however, be viewed and analysed to determine whether any significant relationships can be identified, and to assist in choosing a direction for the continuation of the analysis.

Polynomial and kernel PCA as well as other non-linear dimension reduction techniques such as autoencoding and t-SNE will need to be experimented with before determining the optimal low-dimension representation of our dataset. Clustering may also be experimented with, but would require the manufacturing of categorical

variables from species presence and absence data. Observations could be coloured based on the presence/abundance of each fish species in turn, or categorised into sites at which only indigenous species, only non-indigenous species, both indigenous and non-indigenous species or no species were detected.

Once our EDA has been conducted, a predictive model will be built using the `mvabund` package in R (Wang et al., 2022). The EDA will help to inform the type of regression used (e.g. negative binomial, Poisson, Gaussian etc.), as well as providing information regarding any possible violations of model assumptions. For example, the assumption of independence between response variables may be violated due to the competition that is hypothesised between indigenous and non-indigenous species. The `manyglm` function from the `mvabund` package provides functionality to relax this assumption (Wang et al., 2022), which should be explored.

Redundancy Analysis is a second option which can be explored for building a multivariate regression model. Using this method, we could select the most significant explanatory variables using a stepwise selection method (Valette et al., 2023), which would likely help to reduce the dimensionality of the data.

The results of these two techniques can be compared, and conclusions drawn regarding the hypotheses.

References

Morch, C. and Griffiths, C. (2025). Between a Bass and a Hard Place: The Fragmented Distribution of an Endangered Redfin in the Heuningnes River System of the Cape Fold Ecoregion.

Peterson, A. T. (2001). Predicting species' geographic distributions based on ecological niche modeling. *The Condor*, 103(3):599–605.

Valette, T., Leitwein, M., Lascaux, J., Desmarais, E., Berrebi, P., and Guinand, B. (2023). Redundancy analysis, genome-wide association studies and the pigmentation of brown trout (salmo trutta l.). *Journal of fish biology*, 102(1):96–118.

Wang, Y., Naumann, U., Eddelbuettel, D., Wilshire, J., and Warton, D. (2022). *mvabund: Statistical Methods for Analysing Multivariate Abundance Data*. R package version 4.2.1.

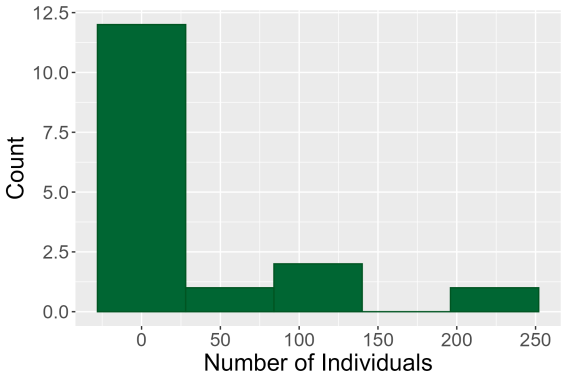
Wang, Y., Naumann, U., Wright, S. T., and Warton, D. I. (2012). mvabund– an r package for model-based analysis of multivariate abundance data. *Methods in ecology and evolution*, 3(3):471–474.

A Appendix

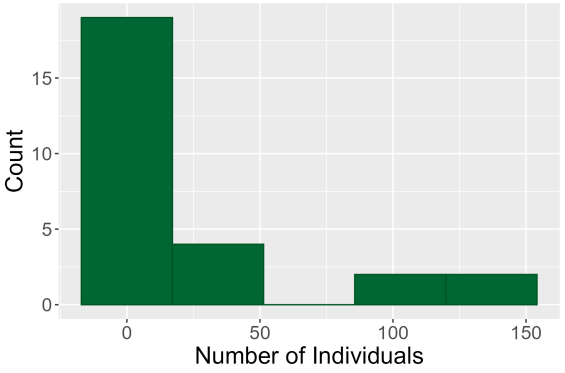
Table A.3: Ranges of the explanatory variables of the Heuningnes redfin dataset.

	Variable	Minimum	Maximum
Water Quality	pH	4.20	10.25
	EC	0.22	20.50
	DO	43.87	126.40
	Temp	11.70	27.87
	Ammonia	0.00	0.94
	Phosphorous	0.00	2.60
	Nitrite	0.00	21.00
	Nitrate	0.00	4.30
	Total_Iron	0.00	3.16
	Phosphonate	0.00	2.89
	TDS	5.00	751.00
	Inorg_Nitrogen	0.10	91.81
Geographic	Elevation	58.00	306.00
	Flow	0.00	0.90
	Slope	0.03	36.86
Substrate	Silt-Sand Substrate	0.00	1.00
	Gravel Substrate	0.00	0.73
	Cobble Substrate	0.00	0.03
	Boulder Substrate	0.00	0.60
	Bedrock Substrate	0.00	0.87
Debris	Woody Debris	0.00	0.70
	No Woody Debris	0.30	1.00
	Undercut Bank	0.00	0.60
	No Undercut Bank	0.40	6.00
Macrophyte	No Macrophytes	0.00	1.00
	Scarce Macrophytes	0.00	1.00
	Moderate Macrophytes	0.00	0.43
	Abundant Macrophytes	0.00	0.47
Canopy	Open Canopy	0.27	1.00
	Partial Canopy	0.00	0.57
	Closed Canopy	0.00	0.23
River Depth	Shallow Water Depth (0 - 50)	0.00	1.00

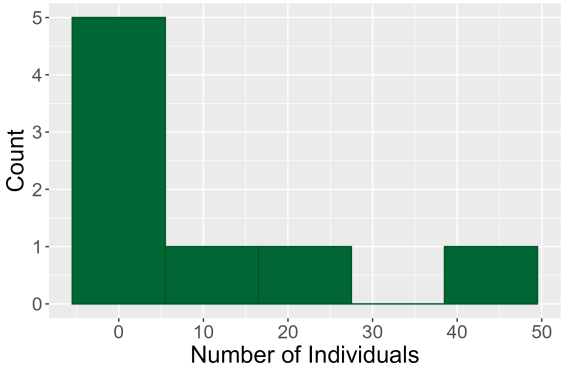
	Moderate Water Depth (51-100)	0.00	0.93
	Deep Water Depth (100 - 180)	0.00	1.00
	Very Deep Water Depth (<180)	0.00	0.90
River Width	Narrow River Width (0 - 3)	0.00	1.00
	Moderate River Width (3 - 6)	0.00	1.00
	Wide River Width (6 - 10)	0.00	1.00
	Very Wide Width (<10)	0.00	1.00



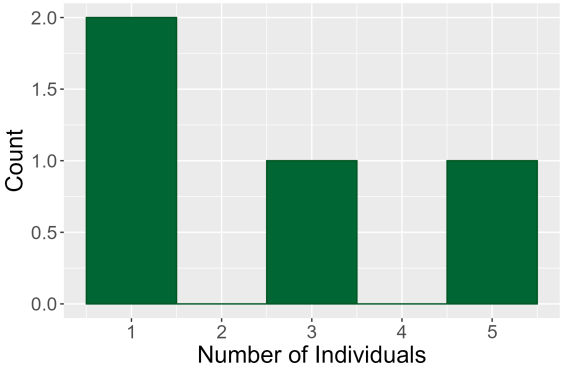
(a) Heuningnes Redfin



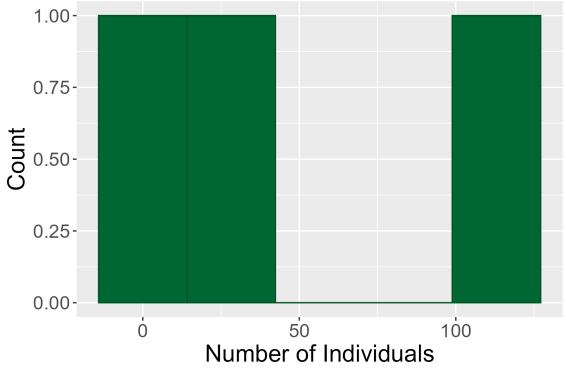
(b) Cape Kurper



(c) Cape Galaxias



(d) Spotted Bass



(e) Bluegill Sunfish

Figure A.1: Distribution of non-zero counts of various fish species at sites on the Kars River (2018).