ESM 567 Hw#3

Bernard Romey

Due: Monday, February 16th, 2015

## Introduction

Redundancy analysis is one of the canonical analyses, which allows assessing the relationships between two matrices. In this homework, we analyze a dataset collected by the US EPA. The EPA researchers measured water quality in randomly selected streams and then characterized associated watershed conditions in the Pacific Northwest. The main objective of this homework is to understand the patterns in stream water quality and its relationships with the watershed conditions. Specifically, you are asked to do the followings:

1. Run PCA on water quality variables (Table 1) to identify and characterize major patterns in terms of water quality.
2. Explain the patterns illustrated by PCA by correlating main PC axes to the associated watershed variables.
3. Run RDA on water quality and watershed variables (Table 1) to assess the relationships between water quality in streams and watershed characteristics.

Table 1. Stream water quality (19 variables) and watershed conditions (14 variables) collected by the US Environmental Protection Agency (EPA) from sampling sites() in Washington, Oregon, and Idaho.

## Methods

**Principal Componenet Analysis (PCA) using R statistical software**

*Evaluation:*

* Center the data with Z-score, and transform if distribution is non-normal, or if different measurement units between variables. PCA requires robust multinormal variable distributions. Make sure the number of variables is less than the number of observations/sample size. Only use quantitative data.
* Calculate variance/covariance matrix (S)
* Eigen-analysis () where ,and
* Calculate site scores (, coordinates of each site in a PCA plot)
* Plot site scores and eigenvectors (Biplot), and interpret results
* Check number of principal components to evaluate (broken-stick model)
* Check for amount of distortion when converting from multidimensional space to two dimensions with a Shepard diagram

*Interpretation:*

* PCA biplot
* Variance/Covariance matrix or correlation matrix
* Number of PCs interpreted (broken stick)
* % of variance explained by each interpreted PC (eigenvalues)
* Eigenvectors (loadings)

**Redundancy Analysis (RNA) using R statistical software**

*Evaluation:*

* Center or standardize both response (Y) and explanatory data (X) matrices.
* Regress each response variable on all explanatory variables in X.
* Calculate fitted values for the response variables from the multiple regression.
* Perform PCA on the matrix of the fitted values
* Use eigenvectors from the PCA to calculate site scores in the space defined by X.

*Interpretation:*

RDA full model

* Constrained inertia (how well two matrices are related)
* Is relationship significant (ANOVA)
* Can the model be reduced

RDA reduced model

* Constrained inertia
* Is model relationship significant
* How many axis are significant
* How many axis to interpret
* What each RDA axis represents (scores)
* Triplot interpretation

## Results

**PCA results**

The goal of PCA is to reduce the complexity of a dataset with n observation and p variables (with dependencies) to a small number of synthetic variables (orthogonal=independent) that represent most of the information in the original dataset. To express the covariance in many variables into a smaller number of composite synthetic variables (PCs), PCA seeks the strongest linear correlation structure among variables.

For this analysis, most of the water quality variable distributions were skewed to the right, therefore a log transformation was used to normalize the distributions so that they were reasonably unskewed. Once normality was checked, a variance (primary axis)/covariance matrix was calculated that was standardized (Correlation Matrix) to Z-scores from the original observations using matrix algebra (Poole 2011) . The reason for converting the original data to Z-scores is because not all units of measure for the different water quality variables were homogeneous (of the same scale). A PCA analysis was then performed on all eighteen log transformed water quality variables.

The eigenvalues from the canonical matrix for component one and two were 0.4637 and 0.1470, respectively. The eigenvalues for the broken stick model indicated that component one and two had higher eigenvalues than what would be observed randomly, so they were retained for analysis (Figure 1). Component one represents 46.4% (eigenvalue 1) of the variability along the primary axis, while component two represents 14.7% (eigenvalues 2) of the variation along the secondary component axis. Thus, both component one and two comprise 61.1% of the variability observed in the data.

The Euclidean distance among objects in multidimensional and reduced space for original descriptors are compared in **figure 3**. The Shepard diagram shows the distortion that occurs when converting from multidimensional space to two dimensions. Projection in reduced space accounts for a high fraction of variance, and the relative positions of the majority of objects are similar.

The eigenvector matrix is a relative representation of the correlations between variables and components. If the correlation matrix is used in the PCA analysis, the principal component loadings are directly proportional to correlations between corresponding variables and principal components (McGarigal 2012). Although, these proportional correlations are not actual correlations between original variables and principal components and is the reason why interpretation of the principal component should be primarily based on the actual correlations in the correlation matrix.

The loadings for each variable on component one and two are shown in **table 3**. Conductivity, acid neutralizing capacity , and calculated bicarbonate contribute the most variance to component one at -0.31, respectively. Other variables have a lower contribution, but not by that much. For example, turbidity and amonium are two of the second to lowest variance values and are -0.13, respectively. The concentration of nitrate in the water had the least variance contribution to component one at -0.06. All eigenvectors are negative for component one. These findings for principal component one corroborate with the high correlations for the same variables in the correlation matrix (Table 2).

Total nitrogen concentration had the largest variance contribution (loading) to component two at -0.35. Turbidity and ammonium contributed the the second largest amount of variance to component two with a variance of -0.33, respectively. Dissolved silica concentration was the least variance contributor to total variance on component two with a loading of -0.06.

Water quality variable vectors with a small cosine angle to the component axis have a larger variance contribution, dependent on the length of the vector. With the exception of water color and SO4 concentration, the majority of the variance contributing to principal component one are very similar as observed in the biplot with the length of the vectors being of comparable length (Figure 2). The water quality variable vectors below dissolved organic matter (DOC) have less variance contribution to component one due to their angle associated with the component one axis.

**RDA results**

As with PCA above, the first step in the RDA analysis was to make sure that the variables for water quality and watershed were appropriate for the analysis. Both the explanatory (watershed) and response (water quality) variables were log transformed and scaled/centered to Z-scores for the purpose of normalizing the distributions.

*Full RDA model*

An overall test of significance showed that the canonical relationship between the water quality and watershed matrices was significant ($p=0.001) after 999 permutations (*Table 4*). The proportion of variance explained by all the explanatory variables (Watershed) was 0.4659.

*Reduced RDA model*

I then used a step selection with the AIC approach plus variance inflation factor (VIF) to see if the RDA model could be reduced to fewer variables. The VIF revealed that the watershed variable straight-line distance to ocean (KM\_SEA) had a VIF of 9.27 suggesting multi-collinearity among watershed variables. After running the selection procedure using the Aikaike Information Criterion (AIC) approach; a search method that compares models sequentially, the recomended reduced model was to remove the straight-line distance to ocean and watershed aspect variables leaving the remaining 12 watershed variables in the reduced model. This reduces the redundancy among watershed variables and helps reduce multi-collinearity (predictor vaiabls correlated among themselves).

Checking the reduced model VIF showed that all reduced model variables had a VIF < 5 indicating multi-colinearity had been reduced, and the AIC value starting at 790.11 was reduced to 648.50. The proportion of variance explained by all the explanatory variables in the reduced RDA was 0.4609. An overall test of signifiacance (*Table 5*) showed that the relationship between the reduced watershed matrix and the water quality matrix was significant ( after 1000 permutations). Six of the fourteen possible RDA axes were found to be significant (*Table 6*) with . Only the first two of the RDA axis were used for interpetaion due to their eigenvalues being 6.18 and 1.46, respectively.

## Discussion

**PCA**

Scatterplots of each component scores for each pair of retained components (in this case 1 and 2) are presented in the biplot (*Figure 2*). Distances among sites in the biplot are approximate to the Euclidean distances among sites in the original multidimensional space. This means that scores in close proximity are ecologically similar with respect to the defined environmental gradients in both principal components (McGarigal 2012). Scores on the left side of component one and close the the water quality vectors, are associated with the negative side of the water quality gradient more than scores on the positive side of the gradient. Projecting a site onto a water quality variable vector in that group of vectors at right angle approximates the position of the site along the variable.

Nonlinearities will show up as an arched pattern in the distribution of scores. Since these site scores do not show a defined pattern with a curvilinear shape, than the linear assumption is most likely not violated.

Since all of the water quality variables seem to be somewhat negatively associated with component one, it would seem that component one is a water quality gradient (*Figure 2*). For component two, the gradient seems to be primarily negatively associated with turbidity, total nitrogen concentration, and amonium (*Table 2*). Elevated turbidity could be associated with disturbances in the watershed from land use practices. Water quality gradient for component one might also be associated with land use practices.

A large proportion of the correlations between variables in the correlation matrix are high (r > 0.5), indicating there is probably redundancy, and making it easier to reduce the dimensionality of the data set (Table 2).

*Hypothesis*

After evaluating the PCA analysis information, it looks like there are pattern in water quality that may be associated with other variables that are not included in the PCA. When the principal component one scores are compared with watershed variables (*Figure 4*), principal component one has the highest correlation with precipitation (), mean watershed slope (), and a negative correlation with riparian disturbance index ().

Some of the questions that arise from PCA are:

* How much variability of river’s water quality can be explained by watershed characteristics?
* Which watershed characteristics are most important with relation to water quality?
* If watershed characteristics don’t explain a large amount of variance, then what else may be important?

In addition, some hypothesis that can be tested with the RDA analysis are:

: There is no significant relationship between the water quality and watershed matrices : There are no significant RDA axis

**RDA**

This analysis produced 12 canonical axes () and 19 aditional unconstrained axes for the residuals. When the constrained variance is expressed as a proportion, it is equivalent to the in multiple regression. Since the accumulation of explanatory variables inflates the explained variance due to random correlations, it is best to use the adjusted instead. The cumulative contribution to the variance of the water quality data (unconstrained) is the proportion of the total variance explained by the RDA (). Twelve of the watershed variables(constrained) explained approximately 0.46() of the variance in water quality. The remaining unconstrained variance left in the residuals not explained by the watershed variables was 53.9%. The residual PC1 (3.28) has an eigenvalue greater than RDA2 (1.45), indicating that the first residual axis of the data has more variance than some of the constrained axis that can be explained by the watershed variables.

The constrained variance would equate to approximately 21.39% of the variance in PC1 from the principal component analysis of water quality variables. In other words, the proportion of variance attributed to watershed variables in the PC1 is approximately 0.21.

In the final reduced RDA, only the first two of the RDA axis were used for interpetaion due to their eigenvalues being 6.18 and 1.46, respectively (*Table 4*). Their proportion of contribution to the variance was 0.4, with RDA1 and RDA2 contributing 0.33 and 0.08 variance, respectively (*Figure 5*).

The primary RDA axis is most likely a watershed disturbance scale, whereas the secondary RDA axis is mainly an elevation scale that is negatively correlated with population density.

In the reduced RDA space water quality variables calcium, conductivity, Magnesium, sodium, total phosphorus conentratino, dissolved silica concentration, and total suspended solids are all primarlily correlated with % of reangeland and road density (*Figure 5*). This again indicates that RDA1 axis is most likely a watershed disturbance gradient. The secondary population density/elevation gradient is correlated with nitrate. Percent of agriculture in watershed is primarily correlated with turbidity, Ammonium, and total nitrogen concentration. In general, all the water quality variables are between percent of total range in the watershed and population denstiy indicating that the water quality is primarily influanced by anthropocentric factors. The SO4 concentration and water color have short projections, indicating they are most likely present over most proportions of the watershed, or related to other ecological paramaters not included in this analysis.

Further analysis behond the scope of this assignment could include varaince paritioning with partial RDA. Some of the question that could be addressed with variance partitioning would include:

* Can water quality be grouped into big categories (natural vs. anthropogenic)?
* Which category of water quality is more important than others (varpart) regarding watershed variables?
* What are water qualtiy variables interactive effects on watershed variables?

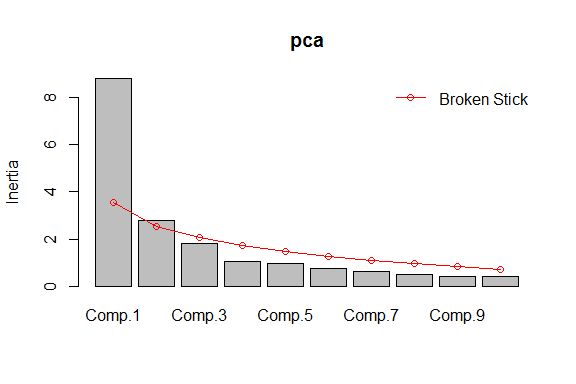
## PCA Figures & Tables

Table 2. Correlation matrix for all 19 water quality variables.

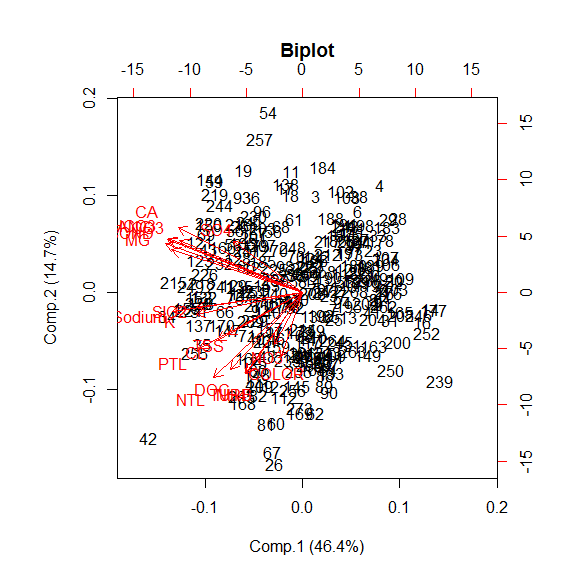
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | CA | COND | K | MG | Sodium | ANC | CL | CO3 | COLOR | DOC | HCO3 | NH4 | NO3 | NTL | PTL | SIO2 | SO4 | TSS | TURB |
| CA | 1.00 | 0.96 | 0.43 | 0.85 | 0.65 | 0.95 | 0.36 | 0.83 | -0.05 | 0.22 | 0.95 | 0.10 | 0.08 | 0.30 | 0.37 | 0.46 | 0.60 | 0.30 | 0.12 |
| COND | 0.96 | 1.00 | 0.55 | 0.93 | 0.77 | 0.98 | 0.47 | 0.87 | 0.02 | 0.30 | 0.98 | 0.20 | 0.10 | 0.39 | 0.48 | 0.57 | 0.58 | 0.36 | 0.19 |
| K | 0.43 | 0.55 | 1.00 | 0.59 | 0.65 | 0.56 | 0.27 | 0.55 | 0.15 | 0.38 | 0.55 | 0.30 | 0.08 | 0.45 | 0.65 | 0.76 | 0.20 | 0.37 | 0.29 |
| MG | 0.85 | 0.93 | 0.59 | 1.00 | 0.68 | 0.93 | 0.42 | 0.84 | 0.09 | 0.31 | 0.93 | 0.20 | 0.06 | 0.35 | 0.49 | 0.60 | 0.44 | 0.33 | 0.17 |
| Sodium | 0.65 | 0.77 | 0.65 | 0.68 | 1.00 | 0.74 | 0.64 | 0.67 | 0.15 | 0.48 | 0.74 | 0.38 | 0.14 | 0.57 | 0.68 | 0.74 | 0.35 | 0.41 | 0.35 |
| ANC | 0.95 | 0.98 | 0.56 | 0.93 | 0.74 | 1.00 | 0.38 | 0.88 | 0.01 | 0.29 | 1.00 | 0.17 | 0.04 | 0.34 | 0.49 | 0.61 | 0.48 | 0.35 | 0.18 |
| CL | 0.36 | 0.47 | 0.27 | 0.42 | 0.64 | 0.38 | 1.00 | 0.32 | 0.16 | 0.31 | 0.38 | 0.49 | 0.44 | 0.58 | 0.39 | 0.23 | 0.40 | 0.26 | 0.34 |
| CO3 | 0.83 | 0.87 | 0.55 | 0.84 | 0.67 | 0.88 | 0.32 | 1.00 | 0.01 | 0.24 | 0.88 | 0.15 | 0.00 | 0.33 | 0.42 | 0.51 | 0.45 | 0.32 | 0.13 |
| COLOR | -0.05 | 0.02 | 0.15 | 0.09 | 0.15 | 0.01 | 0.16 | 0.01 | 1.00 | 0.45 | 0.01 | 0.06 | 0.05 | 0.24 | 0.16 | 0.14 | -0.12 | 0.10 | 0.28 |
| DOC | 0.22 | 0.30 | 0.38 | 0.31 | 0.48 | 0.29 | 0.31 | 0.24 | 0.45 | 1.00 | 0.28 | 0.43 | 0.04 | 0.59 | 0.45 | 0.35 | 0.02 | 0.28 | 0.41 |
| HCO3 | 0.95 | 0.98 | 0.55 | 0.93 | 0.74 | 1.00 | 0.38 | 0.88 | 0.01 | 0.28 | 1.00 | 0.16 | 0.05 | 0.34 | 0.49 | 0.61 | 0.48 | 0.35 | 0.17 |
| NH4 | 0.10 | 0.20 | 0.30 | 0.20 | 0.38 | 0.17 | 0.49 | 0.15 | 0.06 | 0.43 | 0.16 | 1.00 | 0.31 | 0.58 | 0.36 | 0.19 | 0.04 | 0.23 | 0.31 |
| NO3 | 0.08 | 0.10 | 0.08 | 0.06 | 0.14 | 0.04 | 0.44 | 0.00 | 0.05 | 0.04 | 0.05 | 0.31 | 1.00 | 0.62 | 0.11 | -0.05 | 0.26 | 0.09 | 0.06 |
| NTL | 0.30 | 0.39 | 0.45 | 0.35 | 0.57 | 0.34 | 0.58 | 0.33 | 0.24 | 0.59 | 0.34 | 0.58 | 0.62 | 1.00 | 0.53 | 0.34 | 0.21 | 0.42 | 0.37 |
| PTL | 0.37 | 0.48 | 0.65 | 0.49 | 0.68 | 0.49 | 0.39 | 0.42 | 0.16 | 0.45 | 0.49 | 0.36 | 0.11 | 0.53 | 1.00 | 0.75 | 0.05 | 0.50 | 0.53 |
| SIO2 | 0.46 | 0.57 | 0.76 | 0.60 | 0.74 | 0.61 | 0.23 | 0.51 | 0.14 | 0.35 | 0.61 | 0.19 | -0.05 | 0.34 | 0.75 | 1.00 | 0.00 | 0.36 | 0.30 |
| SO4 | 0.60 | 0.58 | 0.20 | 0.44 | 0.35 | 0.48 | 0.40 | 0.45 | -0.12 | 0.02 | 0.48 | 0.04 | 0.26 | 0.21 | 0.05 | 0.00 | 1.00 | 0.17 | 0.01 |
| TSS | 0.30 | 0.36 | 0.37 | 0.33 | 0.41 | 0.35 | 0.26 | 0.32 | 0.10 | 0.28 | 0.35 | 0.23 | 0.09 | 0.42 | 0.50 | 0.36 | 0.17 | 1.00 | 0.51 |
| TURB | 0.12 | 0.19 | 0.29 | 0.17 | 0.35 | 0.18 | 0.34 | 0.13 | 0.28 | 0.41 | 0.17 | 0.31 | 0.06 | 0.37 | 0.53 | 0.30 | 0.01 | 0.51 | 1.00 |

Table 3. Component loadings for PC1 & PC2.

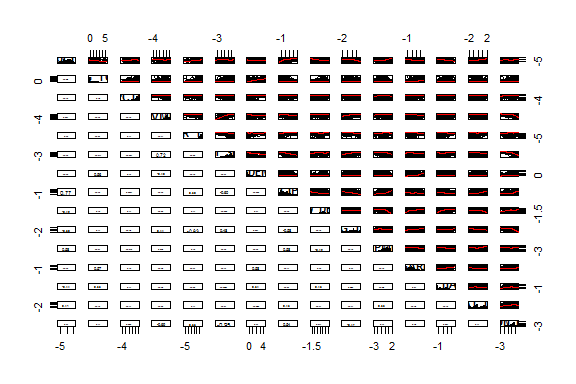
|  |  |  |
| --- | --- | --- |
|  | Comp.1 | Comp.2 |
| CA | -0.28 | 0.26 |
| COND | -0.31 | 0.20 |
| K | -0.24 | -0.09 |
| MG | -0.30 | 0.17 |
| Sodium | -0.30 | -0.08 |
| ANC | -0.31 | 0.21 |
| CL | -0.20 | -0.19 |
| CO3 | -0.28 | 0.21 |
| COLOR | -0.05 | -0.26 |
| DOC | -0.16 | -0.31 |
| HCO3 | -0.31 | 0.22 |
| NH4 | -0.13 | -0.33 |
| NO3 | -0.06 | -0.21 |
| NTL | -0.20 | -0.35 |
| PTL | -0.24 | -0.23 |
| SIO2 | -0.24 | -0.06 |
| SO4 | -0.16 | 0.21 |
| TSS | -0.17 | -0.17 |
| TURB | -0.13 | -0.33 |



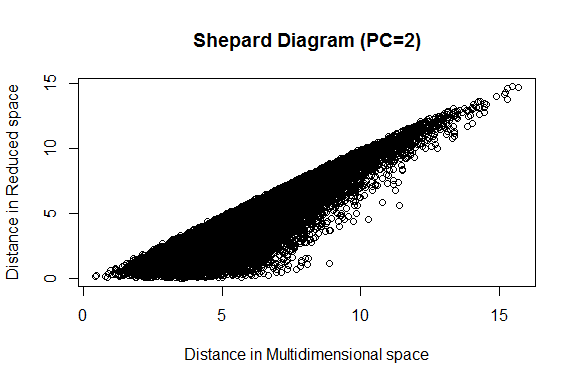
Scree plot showing the inertia/variance (eigenvalues) for first 10 principal component.



Biplot of principal component 1 & 2 with site scores and water quality variable vectos.

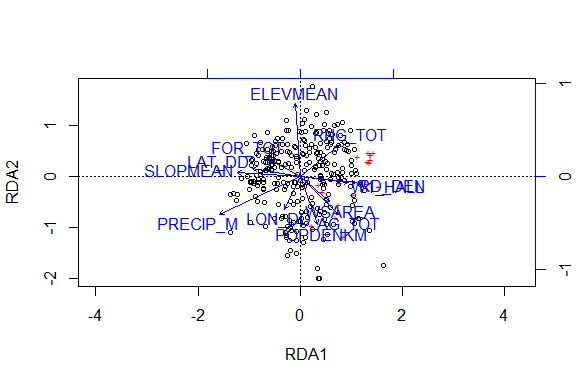


Correlatin matrix of PC1 compared with watershed variables



Shepard plot showing the distortion that occurs when converting from multidimentional space to two dimentions

## RDA Figures & Tables



Triplot of watershed (blue) and water quality variables (red), with site scores (black)

Table 4. ANOVA test if the *full RDA model* relationship between the two matrices significant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Df | Variance | F | Pr(>F) |
| Model | 14 | 8.851795 | 15.76285 | 0.001 |
| Residual | 253 | 10.148205 | NA | NA |

Table 5. ANOVA test if the *reduced RDA model* relationship between the two matrices significant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Df | Variance | F | Pr(>F) |
| Model | 12 | 8.757811 | 18.17028 | 0.001 |
| Residual | 255 | 10.242189 | NA | NA |

Table 6. ANOVA test for how many RDA axes are significant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Df | Variance | F | Pr(>F) |
| RDA1 | 1 | 6.1669342 | 153.5382871 | 0.001 |
| RDA2 | 1 | 1.4467598 | 36.0200087 | 0.001 |
| RDA3 | 1 | 0.4945867 | 12.3137357 | 0.001 |
| RDA4 | 1 | 0.2127226 | 5.2961601 | 0.001 |
| RDA5 | 1 | 0.2068231 | 5.1492802 | 0.002 |
| RDA6 | 1 | 0.1011844 | 2.5191902 | 0.025 |
| RDA7 | 1 | 0.0447568 | 1.1143108 | 0.322 |
| RDA8 | 1 | 0.0363384 | 0.9047168 | 0.486 |
| RDA9 | 1 | 0.0231711 | 0.5768923 | 0.782 |
| RDA10 | 1 | 0.0108554 | 0.2702663 | 0.987 |
| RDA11 | 1 | 0.0103370 | 0.2573601 | 0.987 |
| RDA12 | 1 | 0.0033410 | 0.0831814 | 1.000 |
| Residual | 255 | 10.2421894 | NA | NA |

# REFERENCES

McGarigal. 2012. *Multivariate Statistics for Wildlife and Ecology Research*. 1st ed. 233 Spring Street, New Yourk, NY 10012, USA: Springer Science+Business Media, Inc.

Poole. 2011. *Linear Algebra: A Modern Introduction*. 3rd Edition. Canada: Brooks/Cole.