Chapter 3

Multivariate Models and Analyses

Erica F. Stuber, Christopher C. Chizinski, Jeffrey J. Lusk, and Joseph J. Fontaine

*The search for association among variables is a basic activity in all the sciences*

—Green (1978:38)

WORKING WITH MULTIVARIATE DATA

The interrelated elements that shape biological systems create unique challenges for ecologists hoping to understand the causes and correlations of system processes. The diversity of species in a community, the multitude of environmental structures in a field, and the assortment of behaviors an individual expresses may be inexplicably related. By sampling broadly there is an opportunity to understand system complexity, but practical problems arise. To understand community diversity, for example, a system with 5 species requires consideration of 10 potential correlations, while 25 species may yield 300 potential correlations (Dillon and Goldstein 1984). Multivariate analysis techniques were developed to reduce the dimensionality of datasets, and explore or interpret how multiple response variables are related and influenced by explanatory variables.

Multivariate data are often represented as a matrix with columns representing the measured variables (e.g., species or environmental variables) and rows of observations (e.g., sites or plots sampled). As with univariate analyses, variables can be binary, quantitative, qualitative, rank-ordered, or a mixture of data types, with the data types guiding the choice of analysis. Because there are many multivariate techniques, a decision tree can help biologists narrow down the possible analyses to implement on their data by considering goals (exploratory versus testing), ordination versus classification, and data structures (Box 3.1).

<Box 3.1 approximately here>

Many multivariate datasets that contain non-categorical data can benefit by using data transformations for either statistical or ecological reasons. Choices made regarding outliers and transformation in data-handling are critical in multivariate analyses and should be thoughtfully considered. Values of transformed variables will reflect the original data, but should be more amenable to particular analyses or ecological interpretation. Data transformations (e.g., log, square root, arcsine) help meet the statistical assumptions of the chosen sampling distribution (e.g., normality, linearity, homogeneity of variances) or make units comparable when the observations are measured on different scales. For example, an analysis of water quality might include temperature data that ranges from 0 - 70 ℃, turbidity from 10 to 10000 NTU units, and microcystin data from 0.35 - 32 µg/L, all of which need to be transformed to represent the same scale. Data transformations and standardizations can also be performed on the ecological distance measures commonly computed in clustering or ordination methods, to reduce the influence of absolute quantities, or equalize the relative importance of ubiquitous or rare species (Legendre and Gallagher 2001; Legendre and Legendre 1998). Some of the most common types of standardizations include relativized (dividing by row or column totals or maximums), z-score (subtract the mean of the row or column and dividing by the standard deviation), Hellinger (for zero-inflated presence-absence data), and chord (weights data by rarity).

ECOLOGICAL DISTANCES

In most multivariate analyses, it is important to represent measurement (i.e. rows) and observations (i.e. columns) in your data set in terms of ecological resemblance (Fig. 3.1).

<Fig. 3.1 approximately here>

**Similarity**, which ranges from 1 (complete similarity) to 0 (no similarity), is the most common measure of resemblance and considers the number of measurements observations have in common, divided by the total number of measurements taken. The complement to similarity is **dissimilarity** (i.e., dissimilarity = 1 – similarity). Alternatively, **distance** between measurements can also quantify ecological resemblance. The most common distance measure is Euclidean distance, which uses the Pythagorean Theorem to measure between two points in multidimensional space (rather than geographic space) and is applicable to data of any scale. Some multivariate techniques are limited to a single distance measure (i.e., principal component analysis, correspondence analysis) based on the type of data collected, but other techniques are more flexible (i.e., principal coordinate analysis, non-metric multidimensional analysis). Each distance measure (Table 1) has its own strengths and weaknesses that must be considered before choosing an analysis technique (Legendre and Legendre 1998; McCune et al. 2002; Greenacre and Primicerio 2013).

<Table 1 approximately here>

EXPLORING RELATIONSHIPS IN MULTIVARIATE DATA

PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is one of the earliest methods developed to reduce the dimensionality of multivariate data (Hotelling 1933) and is widely used in ecology (James and McCulloch 1990). In many cases, PCA is used for exploratory data analysis (Ramette 2007) or developing composite values for more complex methods (Janžekovič and Novak 2012). Principal Component Analysis reduces data into so-called components through linear combinations of variables that maximize the variance explained in the original data (James and McCulloch 1990; Legendre and Legendre 1998; Ellison and Gotelli 2004; Robertson et al. 2001). For example, a PCA of fish morphometric data might help determine which measurements to discontinue to reduce animal handling time (Fig. 3.2). Or, PCA could also be used to identify whether population parameters differ based on conservation or management practices such as no-take areas or species protected areas.

<Fig. 3.2 approximately here>

Using Euclidean distances, each successive composite principal component is orthogonal to the last (reducing correlation) and explains less variation than the previous component (i.e. the first component explains the most variation). In many cases, most of the variation in the original data is explained by the first two dimensions and is displayed as a biplot (Fig. 3.2). The points on the biplot represent observations positioned in ordination space by their principal components such that component 1 is represented on the x-axis and component 2 on the y–axis (Jolicoeur and Mosimann 1960). The interpretation of ordination is dependent on whether a distance biplot (intersample relationships; scaling 1) or a correlation biplot (interspecies correlations; scaling 2) is used, because observation scores are rescaled as a function of the scaling choice (Ramette 2007; Legendre and Legendre 1998).

A Principal Component Analysis can be performed on a correlation matrix, or if the data are collected on the same scale, a variance-covariance matrix generated from the original data. An eigenanalysis of the variance-covariance matrix can calculate variances within each principal component, and the percent of variance explained by the components (Table 2; James and McCulloch 1990). The mathematics behind the eigenvectors and eigenvalues fall outside the scope of this chapter, but more detail is available in Rao (1952), Legendre and Legendre (1998), or most linear algebra textbooks.

<Table 2 approximately here>

When the data are not on the same scale, the correlation matrix must be used (Ramette 2007), but it is necessary to standardize the data, which removes the original variance and is therefore not reflected in the distance between variables. Thus, the interpretation of "variance explained" or the amount accounted for by each component depends on whether a correlation matrix or variance-covariance matrix is used (James and McCulloch 1990). This is because correlation matrices are first standardized, thus, distances between principal component scores are independent from the scales of the original data. Using a correlation matrix or a variance-covariance matrix will lead to different principal components.

Principal Component Analysis has no formal assumptions (James and McCulloch 1990), but there are a few limitations:

* There are no rules determining the number of potential principal components. Principal Component Analysis generates orthogonal statistical relationships from the data which in some cases may have no biological interpretation.
* Principal Component values have little meaning outside of the data being considered (e.g., a principal component value of 0.021 is not relevant in a different data set).
* Principal Component Analysis should be limited to data sampled on relatively short gradients representing linear relationships. Unless data are transformed, PCA cannot account for non-linear relationships, so if data are clustered or non-linear, other techniques such as correspondence analysis or non-metric multidimensional scaling are more appropriate.

PRINCIPAL COORDINATE ANALYSIS

Principal Coordinate Analysis (PCoA), or metric multidimensional scaling, is similar to a Principal Component Analysis but can accept distance measures other than Euclidean. Indeed, a PCoA using Euclidean distances produces the same results as a PCA, but, there are instances where Euclidean distances are not appropriate (e.g., many double zeros in species presence absence, abundance data [high beta diversity], genetic distances). For example, when studying community composition, Euclidean distance represents the distance between two sampling points in multidimensional space, and does not account for species identity. As a result, two sampling points with the same species may be depicted being distinct when ordinated. Bray-Curtis distances, which account for species identity, may be more appropriate. Similar to principal components in PCA, principal coordinates in PCoA are combinations of original variables; however, principal coordinates are determined by the similarity or distance function selected (Legendre and Legendre 1998), making the interpretation of variable contributions distance-function-specific. Thus, the choice of dissimilarity function is critical and may require special attention when negative eigenvalues are produced (Legendre and Legendre 1998). Negative eigenvalues are indicative of convergence trouble, when PCoA cannot accurately represent a dissimilarity matrix, commonly due to the use of a non-metric dissimilarity matrix.

Despite the ability to accommodate a broader array of distance functions, PCoA is used relatively infrequently in ecology. Many of the examples in the published literature focus on community composition or genetics across broad geographic ranges. For example, Montaño-Centellas and Garitano-Zavala (2015) used PCoA with a Gower distance (Gower 1966) to reduce the dimensionality of categorical environmental variables related to landscape type and human disturbance across an elevation gradient in the Andes. The synthesized variables from the first component and elevation were used in further analysis to predict avian species richness in the Andes.

Like Principal Component Analysis, there are few formal assumptions with Principal Coordinate Analysis. However, there are some limitations with this methodology that should be noted:

* The choice of the dissimilarity index is important and can influence the interpretation of the data. It is important to research the strengths and weakness of the distance measure and how it will influence your results.
* As mentioned above, depending on the choice of dissimilarity matrix chosen there can be negative eigenvalues that will need to be remedied. In most applications, this does not influence the first few principal axes.
* Like the PCA, PCoA assumes short linear gradients. Non-linearity can result in an arch effect (see detrended correspondence analysis below for flattening of curvature).

CORRESPONDENCE ANALYSIS

Correspondence analysis (CA) is a multivariate statistical technique originally developed to provide a graphical representation of contingency tables (e.g., standard 2-way frequency tables), but are applicable for analyzing descriptors of any precision (Legendre and Legendre 1998). As such, ecologists most frequently use CA to compare the similarity (correspondence) between sites (samples) and species abundances to represent these similarities in ordination space (Hill 1974, 1973). Correspondence analysis uses chi-squaredistances, which requires observations be positive and measured on the same scale (e.g., species abundances, or percent cover of vegetation types).

Correspondence analysis can use numerous mathematical approaches, but the most common is reciprocal averaging. Reciprocal averaging follows five iterative steps: 1) a random number is assigned for all species (i.e. initial species scores); 2) for each site, a sample score is calculated as the weighted average of the initial species scores and the abundance of species; 3) for each species, a new species score is calculated as the weighted average of all site scores; 4) species and site scores are standardized again to get a mean of zero and a standard deviation of one; and 5) steps 2 through 4 are repeated until there are no further change in the values.

Following a similar process as PCA and PCoA, CA decomposes the variance within the data into uncorrelated components that successively explain less variation. For each component, the overall correspondence between species scores and sample scores is summarized by an eigenvalue, which is equivalent to a correlation between species scores and sample scores (Gauch 1982). The results of CA are represented in a joint plot depicting both the site and species scores (Fig. 3.3), with the origin of the plot indicating the centroid of all scores. Like the biplot for PCA, the choice of scaling influences the interpretation of the points in space. When the species composition changes along an environmental gradient, the sample positions may be displayed as an arch. In CA, this curve may be a mathematical consequence that both maximizes the difference between species and minimizes correlation between axes (ter Braak 1987). To remove this curvature, detrended correspondence analysis (DCA) was developed. Detrended correspondence analysis flattens the distribution of the sites along the first CA axis without changing the ordination on the axis.

With CA, there are few statistical assumptions to consider:

* CA assumes that species have bell shaped or Gaussian response curves to environmental variables such that a species is most abundant in the space that has the optimum environmental variables.
* CA uses the chi-square distance function, so the data must be scaled the same and be non-negative.

<Fig. 3.3 approximately here>

NONMETRIC MULTIDIMENSIONAL SCALING

Nonmetric multidimensional scaling (NMDS) has increasingly found acceptance in ecology, particularly due to its capacity to deal with different types of data, including missing data. Non-metric multidimensional scaling has the best performance characteristics of the unconstrained ordination methods in ecology (Minchin 1987).

While most ordination methods attempt to maintain the distance measures between points in multivariate space (i.e., the distance between points in an ordination represent the distance measures between points), there are situations where the exact distance is immaterial and the rank order of the relationships is sufficient. For example, when data are skewed, or multimodal, and may not neatly fit any statistical distribution. Non-metric analyses, such as NMDS, allow for data without an identifiable distribution, and like the PCoA, NMDS is able to use a variety of distance measures. Nonmetric multidimensional scaling, as opposed to other ordination methods, does not rely on eigenanalysis and does not maximize the variation explained by each axis. In fact, NMDS axes are arbitrary and can be flipped or rotated. We can compare the clustering suggestions of metric CA ordination with NMDS using the same data regarding fish species sampled from multiple waterbodies (Fig. 3.4).

The process to perform NMDS follows several iterative steps. The first step is to standardize the data. One of the most common methods is to standardize by the column maximum and then by the row totals (i.e., Wisconsin double standardization). Unlike the eigenanalysis methods described in PCA, CA, and PCoA, the number of axes are described *a priori* in NMDS. The points are then oriented around the *a priori* specified number of axes at random locations (Oksanen et al. 2007). The distances from the random locations are compared to the distances in the original data using isotonic regression with a stress function (ranges between 0 and 1). The stress (i.e., measure of goodness of fit) indicates how different the ranks are in the initial configuration from the original data (Fig. 4). The points are iteratively moved in the direction of decreasing stress until the lowest stress value is found. The process can then be re-run with new random start locations to ensure the outcome represents the lowest stress value rather than a local minimum.

<Fig. 3.4 approximately here>

The results of NMDS are presented in a fashion like PCA and CA, where species and sites are represented on a scatterplot and the proximity between observations represents the similarity between samples. However, the distances between observations do not correspond to the distances measured in the original data. Given this, the data can be rotated or flipped to help interpret the relationships between points.

Although NMDS makes few statistical assumptions, there are limitations to consider:

* Explaining the relationships between or among the points can require qualitative or subjective interpretations.
* NMDS uses an iterative process so the computation power to run the process can be substantial.

CLUSTER ANALYSIS

Cluster analysis is a broad group of multivariate techniques that seeks to identify homogenous groups, by maximizing between-group variation and minimizing within-group variation. The outcome is a reduction of observations into fewer groups (James and McCulloch 1990; Legendre and Legendre 1998). For example, a cluster analysis can spatially aggregate cases of disease on the landscape relative to non-cases to investigate whether landscape features can predict disease. Cluster analysis is often used in data-mining, or exploratory approaches, as it can identify previously undefined subgroups in data such as identifying species assemblages and biogeographic patterns (Jackson et al. 2010). For example, classifying a community based on similarities in species composition (Fig. 3.5). The technique works best when there are discontinuities in the data such as communities that only occur within discrete ecological boundaries (e.g., an agricultural field vs. a forest), rather than a continuous gradient (e.g., a short-grass prairie that transitions into tall-grass prairie) (Legendre and Legendre 1998). If there is continuous structure within the data, ordination techniques are generally preferred over classification as they assume groups respond independently to gradients, whereas cluster analysis may end up forcing the data into groups when groups do not exist. For example, we might either use classification techniques to separate sampled individuals into one of three different subspecies, or use ordination to represent the degree of similarity between the same individuals. Classification provides binary groupings, whereas ordination places subjects closer or further from each other in multidimensional space.

<Fig. 3.5 approximately here>

Cluster analysis occurs in two general steps. First, a measure of similarity between observations is specified. For example, similarity in species abundance at different field sites (Table 3.3) is calculated in a distance matrix for pairwise combinations of field sites (rows) based on the species composition (columns) of each site. Second, using the distance measures, and after selecting a clustering rule, observations are clustered based on either a hierarchical or partitioning technique (Kauffman and Rousseeuw 1990) (Box 3.2). When a new cluster is formed, distances between clusters and other observations (circular diagrams in middle row, Fig. 3.5) are calculated based on Single linkage, or minimum distance (the distance between clusters is equal to the shortest distance from any member of one cluster to any member of the other); Complete linkage, or maximum method (the distance between clusters is equal to the largest distance from any member of one cluster to any member of the other); or Average linkage (the distance between clusters is equal to the average distance from any member of one cluster to any member of the other).

<Table 3.3 approximately here; Box 3.2 approximately here>

Hierarchical techniques are useful as they can reveal relationships in a nested fashion, similar to how evolutionary relationships are depicted in a phylogenetic tree. Because hierarchical clustering requires a pairwise distance matrix between all observations, it is not very efficient for large sample sizes (e.g., > 500 observations). Partitioning methods are typically applied when there are *a priori* reasons to group data, or when groups must be mutually exclusive (e.g., non-nested). For example, non-hierarchical cluster analyses can classify satellite imagery into land-use categories to investigate species-habitat relationships. Both defined (e.g., k-means) and data driven (e.g., iterative self-organizing) partitioning methods follow four iterative steps: 1) randomly assign cluster centroids, 2) classify observations based on the closest centroid, 3) recalculate the group mean centroid after each observation is added, and 4) repeat steps 1-3 until the within cluster variation is minimized (Jensen 1986; Legendre and Fortin 1989). Unlike hierarchical clustering, partitioning methods do not require a data dissimilarity matrix, and are generally computationally efficient, but partitioning methods are sensitive to outliers and outlier influence should be explored.

Interpretation of cluster analysis output is generally done graphically (e.g., dendrograms, or geographic plots) to identify the number of groups contained in the data, or with summary statistics (e.g., cluster means) that define the profile of each cluster. For example, when classifying satellite images into classes of land use types, we can compare the means of each land use type based on separation in multispectral light reflectance (Fig. 3.6). Separation in means of the clustering variables can highlight differences between groups and determine whether clusters are biologically distinguishable. In the case of classifying land use cover, it is possible to visualize the classifications, and compare how the selection of the number of possible groups effects the classification of habitats on the landscape (Fig. 3.7). Once clusters are defined, the results can be visualized sequentially using biplots to explore how classified groups segregate over the variables measured.

<Fig. 3.6 approximately here; Fig. 3.7 approximately here>

Because cluster analysis is an exploratory or hypothesis-generating tool, it is acceptable to explore data that do not fully meet assumptions for optimal statistical performance; however, there are limitations to consider:

* Partitioning methods are not suited to mixed data types (i.e., continuous and categorical data), and specific distance measures (e.g., Gower’s distance) should be used on mixed data types in hierarchical analysis (Banfield and Raftery 1993).
* Cluster analysis assumes distance measures are based on independent variables that are normally distributed (continuous variables) or follow a multinomial distribution (categorical variables) such that clusters appear spherical, with similar variance.
* Clustering variables are appropriate for group separation; otherwise, cluster solutions are not valid.
* Variables are mean-centered and scaled to standard deviation units (Gelman 2008). Because distance measures are strongly influenced by measurement units and magnitude, the variable with the greatest raw magnitude has an overwhelming impact on distance.
* Visual classifications are subjective, depending on the researchers’ idea or expert opinion defining how much distance constitutes a distinct group (Fig. 3.8).

<Fig. 3.8 approximately here>

TESTING RELATIONSHIPS IN MULTIVARIATE DATA

MULTIVARIATE REGRESSION MODELS

As an extension of univariate models introduced in Chapter 2 (this volume), multivariate regression models (Box 3.3) are particularly useful when the goal is to describe, explain, or decompose relationships between measures of flexible, or covarying attributes such as genes, behavior, physiology, or life history traits (Searle 1961; Pigliucci 2003).

<Box 3.3 approximately here>

Two common multivariate regression models are the multivariate analysis of variance (MANOVA; an extension of the univariate ANOVA) when all predictor variables are categorical, and the multivariate analysis of covariance (MANCOVA; extension of univariate ANCOVA) when predictor variables are categorical and continuous. Contrary to other multivariate techniques, multivariate regression models are well-suited in cases where the goal is not dimension reduction, or description of latent constructs, or where it is not clear which variables should be considered a predictor versus a response (Dingemanse and Dochtermann 2013). For example, in community ecology the abundance of two species might correlate, but not be causally related, if an environmental condition such as precipitation is driving populations of both species independently.

Multivariate regression is a powerful tool for examining the variance within and covariance among multiple interrelated measurements such as bi-directional relationships between multivariate datasets (Pigliucci 2003). Multivariate regression is particularly useful when relationships might be influenced by confounding variables (Fig. 3.9), or when we repeatedly collect multiple flexible response variables from subjects (between- and within-subject correlations may be estimated). For example, an individual may change its foraging strategy, antipredator response, or reproductive investment depending on its energy reserves. When considering highly flexible traits such as these, a simple correlation based on single measurements may not be informational because phenotypic associations are jointly shaped by correlations both between-individuals and within-individuals which can be decomposed by multivariate regression (Brommer 2013; Roff and Fairbairn 2007).

<Fig. 3.9 approximately here>

The results of a multivariate regression include estimated univariate models (one for each dependent variable), and the covariance or correlations between multivariate responses. The univariate results are interpreted like a simple linear regression model with coefficient estimates and associated uncertainty for each predictor on each dependent variable in the multivariate set. For example, as the temperature at a nest box increases by 1°C, roosting birds delay sleep onset at night by 0.84 min (95% credible interval (CI): 0.19, 1.61min), awaken 0.75 min (95%CI 0.39, 1.05) earlier in the morning, and slightly increase the number of awakenings during the night (95%CI 0.01, 0.03): Table 4). Running separate univariate regression models for each dependent variable would result in similar mean coefficients and errors (Table 3.4). In addition to the estimates the tri-variate model provides pairwise correlations between responses after adjusting for fixed effects.

<Table 3.4 approximately here>

Similar to partitioning variance to within- and between-subject components, it is also possible to partition the covariance between multivariate responses within-subjects or between-subjects. For example, the variance in metabolic rate, daily energy expenditure, and exploratory tendencies of individual birds can be partitioned into between-individual (e.g., consistent differences) and within-individual (e.g., plasticity) components. Making it possible to partition the covariance of traits into between-individual (e.g., syndrome), and within-individual (e.g., integration of plasticity) components.

As with univariate regressions, before drawing conclusions it is important to consider model fit, which is typically done through graphical residual analysis. Failure to fit the data can indicate where model assumptions were violated (Fig. 3.10).

Similar to univariate regression models, multivariate regression models assume:

* Observations are independent.
* The error term in our model is normally-distributed with zero mean and constant variance.
* There is no pattern left unexplained by our predictors (Mertler and Reinhart 2016).

Violations of these assumptions can result in biased estimates or overly confident representations of uncertainty.

<Fig. 3.10 approximately here>

STRUCTURAL EQUATION MODELS

Understanding inherently complex ecological systems requires the ability to understand networks of relationships simultaneously at work (Fig. 3.11) (Grace 2008). Structural Equation Modeling (SEM) provides a framework for encoding and testing hypotheses about how systems function as a whole, by analyzing the nature and magnitude of relationship webs between one or more predictor and response variables simultaneously. For example, a wildlife ecologist interested in the direct and indirect effects of avian predators on a community of prey may pose alternative hypotheses about species interactions by developing multiple possible community interaction webs based on previous work or prior knowledge of a system (Wootton 1994). The ability of SEM to evaluate support for interrelated webs of ecological processes is an analytical advance over sequential univariate analyses that can allow ecologists to simultaneously evaluate support for biological hypotheses (Dochtermann and Jenkins 2011), and guide explicit tests of proposed models through prediction and subsequent experimentation.

<Fig. 3.11 approximately here>

Path diagrams, a graphical representation of a hypothesis about sets of relationships, form the foundation of SEM (Fig. 3.11), and correspond to the underlying mathematical regression equations used for analysis (Box 3.4). Graphical path diagrams help clarify how multiple variables are related, considering measured (observed) and latent (unobserved) variables as well as direct and indirect relationships (Grace et al. 2010). Based on the path diagram the inputs to SEM analysis include data (either raw data, or correlation/covariance matrices), and the graphical paths that encode the *measurement*, and *structural models* (Fig. 3.12).

<Fig. 3.12 approximately here; Box 3.4 approximately here>

The output of an SEM analysis is a set of estimated coefficients, or *loadings*, wherever a path is included in the graphical diagram. Variance associated with the measured variables, correlations between latent factors (Box 3.5), and a set of tests (e.g., Akaike information criteria) help to assess how well the proposed diagram fits the observed data (Grace et al. 2010; Mitchell 1992). Standardized path coefficients can be interpreted similarly to estimated regression coefficients (Fig. 3.13).

<Fig. 3.13 approximately here; Box 3.5 approximately here>

When latent variables (i.e. unmeasured constructs) are involved in the analysis, we will estimate factor loadings which are similar to standardized regression coefficients, and represent the relationship between observed variables and their presumed underlying latent construct. Loadings also tell us how well any one measured variable could substitute for multiple variables when describing the latent factor.

Estimates of error associated with path estimates and factor loadings are also given, such that a p-value might be calculated for significance testing, and estimates of residual error variance are provided for observed variables. Double-headed arrows (e.g., Box 3.5, between latent variables) represent correlations between variables of interest (e.g. no directional causality is implied). For example, a behavioral ecologist might want to understand whether spatial food caching preferences are correlated with the amount of flexibility in caching behavior (Dochtermann and Jenkins 2007).

Overall SEM model fit (e.g. how well our hypothesized relationships fit the observed correlation structure in the network) is often assessed through chi-square tests, which evaluate the amount of difference between the expected and observed correlation matrices (Jöreskog and Bollen 1993; Kline 2015) or root mean squared error (RMSE), which represents the amount of variance left unexplained by the proposed model (MacCallum et al. 1996; Steiger 1990).

If the hypothesized model does not adequately fit the observed data, it is possible to perform steps for model re-specification, including identifying previously overlooked relationships, or even formulating new alternative hypotheses. Once a SEM fits the data, the only way to test for causality is to perform experiments. For example, researchers have used SEM to understand and validate the relationships between winter storms and food web dynamics in kelp forests (Byrnes et al. 2011) by using estimates from the fitted SEM as predictions to test field experimentation.

In general, SEM is not as well suited for exploratory analysis because the causal relationships defined in the graphical diagram must be encoded *a priori*. Beyond this limitation, the structure of an SEM also requires assumptions underlying the input data as well as the graphical models are met (Karlin et al. 1983), including:

*Input data and estimation*:

* Data are expected to conform to statistical assumptions of simple linear regression (see Chapter 2: Regression, this volume) as both regression and SEM are typically estimated through maximum-likelihood techniques.
* Observations are independent, or dependence is accounted for with a correlated error term.
* The joint distribution of constructs is multivariate normal, otherwise a transformation or a non-maximum likelihood estimator is necessary.
* The model must be correctly specified.

*Graphical diagrams*:

* Indicators (observed variables) and constructs (latent variables) are correctly identified.
* Direction of the presumed effect or association is correctly specified.

Because regression-based structural equation models have many of the same assumptions as univariate and multivariate regressions, violations of these assumptions will result in similarly biased estimates of predictor effects and over-confident estimates of uncertainty.

LINEAR DISCRIMINATION ANALYSIS

In ecology, groups are often predefined with questions focused on identifying what makes groups different. Linear discrimination analysis (LDA), similar to multinomial logistic regression, attempts to find linear combinations of variables that best separate groups when predicting two or more dependent variables, using continuous independent variables. For example, using truss-based morphometrics, Chizinski et al. (2010) used LDA to describe which morphometric measurements best differentiated the differences in body shape between stunted and non-stunted fish (Fig. 3.14). Features of the head were relatively larger in stunted fish, whereas the mid-body was larger in non-stunted fish.

Similar to logistic regressions, LDA performs best for multivariate normal data when within group variance and covariance are similar (James and McCulloch 1990).

<Fig. 3.14 approximately here>

PERMUTATIONAL MULTIVARIATE ANALYSIS OF VARIANCE

Permutational MANOVA (PERMANOVA; also sometimes referred to as adonis, or AMOVA in multivariate molecular ecology) partitions sums of squares of a multivariate data set equivalent to MANOVA. Permutational MANOVA is an alternative to metric MANOVA and to ordination methods for describing how variation is attributed to different experimental treatments (e.g., LDA) or uncontrolled covariates (Anderson and Walsh 2013). Permutational MANOVA works with any distance function that is appropriate for the data being used. The permutations (generally 1000 or more) of group membership are used to assess the statistical significance. A p-value is calculated as the proportion of the values of the statistic obtained under permutation that are equal to or more extreme than the observed value (Anderson and Braak 2003). Opposite of LDA, PERMANOVA is suited for predicting continuous dependent variables from categorical independent variables. For example, Edgar et al. (2009) used PERMANOVA to test whether two continuous variables, fish density and biomass were significantly different inside, versus outside of marine protected areas.

Because PERMANOVA is a permutation-based test, it does not have distribution-based assumptions (e.g., normality), and has only one formal assumption:

* Observations (rows of the original data matrix) are exchangeable under a true null hypothesis (Anderson 2001).

PERMANOVA is sensitive to heterogeneity of variance within groups, particularly in unbalanced experimental design (see Table 3.5).

MANTEL TEST

The Mantel test compares two independent data sets for similar observations (Mantel 1967) using similarity matrices. Because the Mantel test compares similarity matrices, dependent and independent variables can be a mixture of continuous and categorical. The Mantel statistic is bounded between –1 and 1 and behaves like a correlation coefficient. The significance (p-value) of the Mantel statistic is tested by permutations of the groups as is done in PERMANOVA.

Another application of the Mantel test is a goodness of fit test, to test ecological hypotheses against observed data (Legendre and Legendre 1998). For example, Chizinski et al. (2006) proposed several hypotheses (e.g., equal occurrence, fish invasiveness, proximity, human preference) that would account for patterns of fish presence-absence in urban lakes. They used Mantel tests to compare the observed-presence absence patterns to matrices indicating the nine hypotheses explaining the patterns. Results indicated that the model representing fish invasiveness best fit the observed incidence patterns.

The Mantel test has a few key assumptions:

* The relationships between the two matrices are linear.
* Observations are spatially independent (a widespread issue in ecological and evolutionary data). Spatial autocorrelation in data should be minimized during study design (Guillot and Rousset 2013) or addressed during analysis because it can produce spurious relationships (see Table 6).

ANALYSIS OF SIMILARITIES

Analysis of similarities (ANOSIM) is a non-metric test for differences between two or more groups using any distance measure (Clarke 1993). ANOSIM compares the ranks of distances between groups with the rank distances within groups, producing an *R*-statistic. The *R*-statistic varies from approximately 0 to 1, with 1 indicating large separation between the groups and 0 indicating little separation. The significance (p-value) of the *R*-statistic is tested by permutations of the groups as is done in the Permutational MANOVA.

ANOSIM has one key assumption:

* The ranges of dissimilarities within groups should be similar. Running ANOSIM on groups with very different within group dispersions can lead to spurious results.

INDIRECT GRADIENT ANALYSIS

While ordination techniques like PCA and CA do not account for direct relationships between suites of species, it can be inferred that there are underlying environmental conditions that influence community composition. In PCA or CA, the synthesized observation scores and traditional statistical techniques (e.g., ANOVA, linear regression) can be used to assess the influence of an environmental variable on the synthesized scores. For example, before performing a regression analysis, Bai et al. (2016) first derived PCA scores describing the wing shape of multiple insect species. Subsequently, Bai et al. (2016) investigated how environmental characteristics including temperature, precipitation, and elevation predicted the morphometric PCA scores using linear regression. Similarly, we can use the ranks of the observations on each axis in PCoA and NMDS and the ranks of environmental variables using Spearman’s rank correlation (Legendre and Legendre 1998). For example, the plot-level NDMS scores from an NDMS analysis of a vegetation community from plots located across a large spatial extent could be correlated with temperature, soil characteristics, and precipitation metrics across the study sites. Because this type of analysis uses the synthesized variables created from a multivariate technique and then compares them to environmental variables it is termed an indirect gradient analysis.

DIRECT GRADIENT ANALYSIS

As opposed to an indirect gradient analysis, which analyses the relationship between ordination scores and environmental variables *a posteriori*, a direct gradient analysis uses the environmental variables directly in the ordination, forcing the scores to be maximally related to the environmental gradients. As such, only the variation in species composition explained by the environmental variables is analyzed, not all the variation in the species composition (Legendre and Legendre 1998). Thus, direct gradient analysis requires that the environmental gradient is known and represented by measured variables. For example, it is possible to perform an ordination procedure based on all of the data from a vegetation community rather than only on plant species composition data. Gradient analysis can be performed with any ordination technique, but because ordination computes optimal linear combinations of variables, there is always the assumption of linearity between ordination scores and gradients, and non-linear relationships will likely be undetected.

*Redundancy Analysis*

Redundancy analysis (RDA) is an extension of PCA, but is generally used for statistical testing, rather than exploration, where each canonical axis (similar to principal components) corresponds to a direction in the multivariate dependent variables (e.g., species), measured in Euclidean distance, that is maximally related to linear combinations of the multivariate independent variables (e.g., environmental conditions) (Rao 1952). Multiple linear regression is used to explain the variation between the independent and dependent variables and through iteration, the best ordination of the observations is found. Redundancy analysis preserves the Euclidean distances between the *fitted* dependent variables (Legendre and Fortin 1989). To conduct an RDA requires a data frame corresponding to the abundance of species across sites and a data frame corresponding to environmental conditions at the sites (Fig. 3.15). RDA provides the species variation that can be explained by environmental variation, the cumulative percentage of variance of the species–environment relationship, and the overall statistical significance of the relationships between the species and environmental data frames (Legendre et al. 2011).

<Fig. 3.15 approximately here>

An extension of redundancy analysis is distance-based redundancy analysis (dbRDA) that uses non-Euclidean distance measures (Legendre and Anderson 1999). The general methodology of dbRDA is similar to RDA described above, but it uses PCoA rather than PCA, which allows for the use of other distance measures like Bray-Curtis. For instance, Geffen et al. (2004) used dbRDA to assess the variation of genetic distances in the gray wolf (*Canis lupus*) that is explained by temperature, rainfall, habitat type, water barriers, and climate.

Like PCA, RDA can be represented in a biplot of the species score and site scores, but RDA can also have environmental variables overlaying the scores (i.e., triplots). The environmental variables are represented as arrows for continuous variables or as additional shapes for qualitative or nominal variables (ter Braak 1994). Also similar to the biplot projections in PCA, the choice of scaling (focus on the sites or the species) also influences how the points in space are interpreted (ter Braak 1994).

*Canonical Correspondence Analysis*

Canonical correspondence analysis (CCA) is similar to RDA, except that it is based on the unimodal species-environment relationship in correspondence analysis. The calculations involved in CCA, however, are more complex than RDA and can be approached using three different techniques, details of which are found in ter Braak (1986, 1987) and Legendre and Fortin (1989). Like CA, CCA preserves the chi-squaredistances among observations rather than the Euclidean distances in RDA (Legendre and Fortin 1989). CCA is well adapted to relate environmental data to species abundances and occurrence even when a species is absent at some sites. However, CCA is particularly sensitive to rare species that occur in species-poor samples (Legendre and Legendre 1998). To conduct a CCA, a matrix corresponding to the abundance of species across sites and another matrix of environmental data corresponding to the sites in the species data frame are compared. The analysis provides the species variation explained by environmental variation, the cumulative percentage of variance of the species–environment relationship, and the overall statistical significance of the relationships between the species and environmental data frames.

The results of a CCA can also be represented in the biplot or triplot format, of which the interpretation of the points in multidimensional space are dependent on the type of scaling chosen. The same interpretation of the relationships between sample and species points is found in CA and CCA. Species scores are often interpreted as species’ environmental optima, but this assumes that abundance or probability of occurrence follows a symmetric, unimodal function of position over the environmental gradients considered.

CONCLUSION

Many interesting questions in wildlife science regard complex patterns in ecological communities, and how these patterns are influenced by interactions with multifaceted environments. Multivariate methods provide a formal framework to evaluate datasets with large numbers of often correlated variables, which are common to wildlife biology and ecology. The statistical techniques outlined in this chapter provide methods for summarizing redundancy (e.g., for variable reduction in subsequent studies), or combining variables (e.g., that reflect unmeasured latent variables, or describe composite variables), to allow us to detect and quantify multivariate patterns. The purpose of this chapter was to provide an overview of some of the more popular multivariate statistical techniques in wildlife ecology. If you are interested in more detail on one of the processes discussed or other techniques available to you, you are encouraged to read much more detailed books on the material (McCune et al. 2002; Legendre and Legendre 2012; McGarigal et al. 2000; Borcard et al. 2011). Further, most modern statistical programs (e.g., R, MATLAB, SAS) can conduct most types of multivariate statistics, thus you are free to explore these techniques and processes in much more detail in the statistical environment you are most comfortable in.

Multivariate analyses are often descriptive in nature, acting as important hypothesis-generating mechanisms, but are also used to make inferences when appropriate study design is followed (e.g. probabilistic samples are drawn from a population), and analytical assumptions are met. Although some multivariate techniques often require the investigator to make decisions regarding the level of similarity defining a group, or the interpretation of latent or composite variables, it can also be viewed as a highly flexible mechanism for exploring alternative hypotheses about ecological systems. Regardless of method, multivariate or univariate, inference based on experimental manipulation is the gold-standard for science. Once understood, multivariate methods are powerful tools in wildlife scientists’ repertoire to describe the multivariate nature of wildlife systems and test complex interrelationships.

ACKNOWLEDGEMENTS

We are grateful to Leonard Brennan for inviting us to contribute this chapter, and all the editors for their patience and encouragement during the writing process. We would like to thank Mark Burbach and two anonymous reviewers for their insightful comments on an earlier draft, and Lutz Gruber for help in compiling mathematical equations for Table 1. CJC was supported by Hatch funds through the Agricultural Research Division at the University of Nebraska-Lincoln and from Federal Aid in Wildlife Restoration project W-120-T, administered by the Nebraska Game and Parks Commission. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. The Nebraska Cooperative Fish and Wildlife Research Unit is supported by a cooperative agreement among the U.S. Geological Survey, the Nebraska Game and Parks Commission, the University of Nebraska, the U.S. Fish and Wildlife Service, and the Wildlife Management Institute.

Literature Cited

Anderson, M.J., 2001 A new method for non‐parametric multivariate analysis of variance. *Austral ecology* 26 (1):32-46.

Anderson, M.J., and C.T. Braak, 2003 Permutation tests for multi-factorial analysis of variance. *Journal of statistical computation and simulation* 73 (2):85-113.

Anderson, M.J., and D.C. Walsh, 2013 PERMANOVA, ANOSIM, and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? *Ecological Monographs* 83 (4):557-574.

Bai, Y., J.J. Dong, D.L. Guan, J.Y. Xie, and S.Q. Xu, 2016 Geographic variation in wing size and shape of the grasshopper Trilophidia annulata (Orthoptera: Oedipodidae): morphological trait variations follow an ecogeographical rule. *Scientific reports* 6 (32680).

Banfield, J.D., and A.E. Raftery, 1993 Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49 (3):803-821.

Borcard, D., F. Gillet, and P. Legendre, 2011 Introduction, pp. 1-7 in *Numerical Ecology with R*. Springer.

Brommer, J.E., 2013 On between-individual and residual (co) variances in the study of animal personality: are you willing to take the “individual gambit”? *Behavioral Ecology and Sociobiology* 67 (6):1027-1032.

Byrnes, J.E., D.C. Reed, B.J. Cardinale, K.C. Cavanaugh, S.J. Holbrook *et al.*, 2011 Climate‐driven increases in storm frequency simplify kelp forest food webs. *Global Change Biology* 17 (8):2513-2524.

Chizinski, C., C. Higgins, C. Shavlik, and K.L. Pope, 2006 Multiple hypotheses testing of fish incidence patterns in an urbanized ecosystem. *Aquatic Ecology* 40 (1):97-109.

Chizinski, C., K.L. Pope, G. Wilde, and R. Strauss, 2010 Implications of stunting on morphology of freshwater fishes. *Journal of Fish Biology* 76 (3):564-579.

Clarke, K.R., 1993 Non‐parametric multivariate analyses of changes in community structure. *Australian journal of ecology* 18 (1):117-143.

Deguines, N., J.S. Brashares, and L.R. Prugh, 2017 Precipitation alters interactions in a grassland ecological community. *Journal of Animal Ecology* 86 (2):262-272.

Dillon, W.R., and M. Goldstein, 1984 *Multivariate analysis: Methods and applications*. NJ, USA: Wiley Publishing.

Dingemanse, N.J., and N.A. Dochtermann, 2013 Quantifying individual variation in behaviour: mixed-effect modelling approaches. *Journal of Animal Ecology* 82 (1):39-54.

Dochtermann, N.A., and S.H. Jenkins, 2007 Behavioural syndromes in Merriam's kangaroo rats (Dipodomys merriami): a test of competing hypotheses. *Proceedings of the Royal Society of London B: Biological Sciences* 274 (1623):2343-2349.

Dochtermann, N.A., and S.H. Jenkins, 2011 Developing multiple hypotheses in behavioral ecology. *Behavioral Ecology and Sociobiology* 65 (1):37-45.

Edgar, G.J., and R.D. Stuart-Smith, 2009 Ecological effects of marine protected areas on rocky reef communities—a continental-scale analysis. *Marine Ecology Progress Series* 388:51-62.

Ellison, G.N., and N. Gotelli, 2004 A primer of ecological statistics. *Sinauer, Sunderland, Massachusetts, USA*.

Gauch, H.G., 1982 *Multivariate analysis in community ecology*: Cambridge University Press.

Geffen, E., M.J. Anderson, and R.K. Wayne, 2004 Climate and habitat barriers to dispersal in the highly mobile grey wolf. *Molecular Ecology* 13 (8):2481-2490.

Gelman, A., 2008 Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* 27 (15):2865-2873.

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

Grabmeier, J., and A. Rudolph, 2002 Techniques of cluster algorithms in data mining. *Data Mining and Knowledge Discovery* 6 (4):303-360.

Grace, J.B., 2008 Structural equation modeling for observational studies. *Journal of Wildlife Management* 72 (1):14-22.

Grace, J.B., T.M. Anderson, H. Olff, and S.M. Scheiner, 2010 On the specification of structural equation models for ecological systems. *Ecological Monographs* 80 (1):67-87.

Greenacre, M., and R. Primicerio, 2013 *Multivariate data analysis for ecologists*. Bilbao, Spain: Foundation BBVA.

Guillot, G., and F. Rousset, 2013 Dismantling the Mantel tests. *Methods in Ecology and Evolution* 4 (4):336-344.

Hill, M.O., 1973 Reciprocal averaging: an eigenvector method of ordination. *The Journal of Ecology* 61 (1):237-249.

Hill, M.O., 1974 Correspondence analysis: a neglected multivariate method. *Applied statistics* 23 (3):340-354.

Hotelling, H., 1933 Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* 24 (6):417.

Jackson, D.A., S.C. Walker, and M.S. Poos, 2010 Cluster analysis of fish community data: “new” tools for determining meaningful groupings of sites and species assemblages, pp. 503-527 in *Community ecology of stream fishes: concepts, approaches, and techniques. American Fisheries Society, Symposium*.

James, F.C., and C.E. McCulloch, 1990 Multivariate-Analysis in Ecology and Systematics - Panacea or Pandora Box. *Annual Review of Ecology and Systematics* 21:129-166.

Janžekovič, F., and T. Novak, 2012 PCA – a powerful method for analyze ecological niches, pp. 127-142 in *Principal Component Analysis- multidisciplinary applications*, edited by P. Sanguansat.

Jensen, J.R., 1986 Introductory digital image processing: a remote sensing perspective.

Jolicoeur, P., and J.E. Mosimann, 1960 Size and shape variation in the painted turtle. A principal component analysis. *Growth* 24:339-354.

Jöreskog, K.G., and K.A. Bollen, 1993 Testing structural equation models. *Sage focus editions* 154:294-294.

Karlin, S., E. Cameron, and R. Chakraborty, 1983 Path analysis in genetic epidemiology: a critique. *American journal of human genetics* 35 (4):695.

Kauffman, L., and P. Rousseeuw, 1990 Finding groups in data. *An introduction to cluster analysis: Wiley Interscience*.

Kline, R.B., 2015 *Principles and practice of structural equation modeling*: Guilford publications.

Lambert, T.D., J.R. Malcolm, and B.L. Zimmerman, 2006 Amazonian small mammal abundances in relation to habitat structure and resource abundance. *Journal of Mammalogy* 87 (4):766-776.

Legendre, P., and M.J. Anderson, 1999 Distance‐based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs* 69 (1):1-24.

Legendre, P., and M.J. Fortin, 1989 Spatial pattern and ecological analysis. *Vegetatio* 80 (2):107-138.

Legendre, P., and E.D. Gallagher, 2001 Ecologically meaningful transformations for ordination of species data. *Oecologia* 129 (2):271-280.

Legendre, P., and L. Legendre, 1998 *Numerical ecology*. Amsterdam ; New York: Elsevier.

Legendre, P., and L. Legendre, 2012 *Numerical ecology*. Amsterdam: Elsevier.

Legendre, P., J. Oksanen, and C.J. ter Braak, 2011 Testing the significance of canonical axes in redundancy analysis. *Methods in Ecology and Evolution* 2 (3):269-277.

MacCallum, R.C., M.W. Browne, and H.M. Sugawara, 1996 Power analysis and determination of sample size for covariance structure modeling. *Psychological methods* 1 (2):130.

Mantel, N., 1967 The detection of disease clustering and a generalized regression approach. *Cancer research* 27 (2 Part 1):209-220.

McCune, B., J.B. Grace, and D.L. Urban, 2002 *Analysis of ecological communities*: MjM software design Gleneden Beach.

McGarigal, K., S. Cushman, S.G. Stafford, and S.G. Stafford, 2000 *Multivariate statistics for wildlife and ecology research*: Springer New York.

Mertler, C.A., and R.V. Reinhart, 2016 *Advanced and multivariate statistical methods: Practical application and interpretation*: Routledge.

Minchin, P.R., 1987 An evaluation of the relative robustness of techniques for ecological ordination, pp. 89-107 in *Theory and models in vegetation science*. Springer.

Mitchell, R., 1992 Testing evolutionary and ecological hypotheses using path analysis and structural equation modelling. *Functional Ecology* 6 (2):123-129.

Montaño-Centellas, F.A., and Á. Garitano-Zavala, 2015 Andean bird responses to human disturbances along an elevational gradient. *Acta Oecologica* 65:51-60.

Oksanen, J., R. Kindt, P. Legendre, B. O’Hara, M.H.H. Stevens *et al.*, 2007 The vegan package. *Community ecology package* 10:631-637.

Orlóci, L., 1978 Ordination by resemblance matrices, pp. 239-275 in *Ordination of Plant Communities*. Springer.

Pielou, E.C., 1984 *The interpretation of ecological data: a primer on classification and ordination*: John Wiley & Sons.

Pigliucci, M., 2003 Phenotypic integration: studying the ecology and evolution of complex phenotypes. *Ecology Letters* 6 (3):265-272.

Ramette, A., 2007 Multivariate analyses in microbial ecology. *Fems Microbiology Ecology* 62 (2):142-160.

Rao, C.R., 1952 *Advanced statistical methods in biometric research*. Oxford, England: Wiley.

Robertson, M., N. Caithness, and M. Villet, 2001 A PCA‐based modelling technique for predicting environmental suitability for organisms from presence records. *Diversity and Distributions* 7 (1‐2):15-27.

Roff, D., and D. Fairbairn, 2007 The evolution of trade‐offs: where are we? *Journal of Evolutionary Biology* 20 (2):433-447.

Searle, S., 1961 Phenotypic, genetic and environmental correlations. *Biometrics* 17 (3):474-480.

Steiger, J.H., 1990 Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research* 25 (2):173-180.

Stuber, E.F., C. Baumgartner, N.J. Dingemanse, B. Kempenaers, and J.C. Mueller, 2016 Genetic Correlates of Individual Differences in Sleep Behavior of Free-Living Great Tits (Parus major). *G3: Genes| Genomes| Genetics*:g3. 115.024216.

ter Braak, C.J., 1986 Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67 (5):1167-1179.

ter Braak, C.J., 1987 The analysis of vegetation-environment relationships by canonical correspondence analysis, pp. 69-77 in *Theory and models in vegetation science*. Springer.

ter Braak, C.J., 1994 Canonical community ordination. Part I: Basic theory and linear methods. *Ecoscience* 1 (2):127-140.

Wootton, J.T., 1994 Predicting direct and indirect effects: an integrated approach using experiments and path analysis. *Ecology* 75 (1):151-165.

**Table 1.** Comparison and formulation of distance measures commonly used in multivariate analyses. In each equation we are calculating the distance, in p dimensions, between sample units j and k from a matrix of observations of measured variables on sampling units (e.g., plots or individuals).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Input data (x)** | **Range of Distance value** | **Metric** | **Equations** | **Notes** |
| **Euclidean Distance** | any | non-negative | y1 |  | Pythagorean Theorem applied to *p* dimensions;  used in eigenvector ordination;  literal distances, directly interpretable;  sensitive to large outliers |
| **Manhattan Distance** | any | non-negative | y1 |  | "City-block" distance;  not compatible with LDA, CCA |
| **Chi-square Distance** | ≥ 0 | d ≥ 0 | y1 |  | Euclidean but standardized by the mean;  used for species abundance data in CA and CCA |
| **Chord Distance** | any | 0 ≤ d ≤ √2 | y1 |  | Euclidean distance on a hypersphere of radius = 1;  adjusted ED, by normalizing observations |
| **Percent Dissimilarity** | ≥ 0 | 0 ≤ d ≤ 1\* | n2 |  | Overlap between area under curves;  not compatible with LDA, CCA |
| **Bray-Curtis/Sorensen Distance** | ≥ 0 | 0 ≤ d ≤ 1\* | n3 |  | Used for raw count data in PCA and NMDS;  sensitive to large outliers |
| **Jaccard Distance** | ≥ 0 | 0 ≤ d ≤ 1\* | Y1 |  | Proportion measure in Manhattan space |

**\***: proportional measurements

**1**: Metric measurements satisfy four criteria: 1) minimum distance is 0 and occurs between identical observations; 2) distance will be positive if observations are not identical; 3) distances are symmetric; and 4) satisfies the triangle inequality

**2**: Non-metric measurements can take negative values

**3**: Semi-parametric measurements do not satisfy the triangle inequality

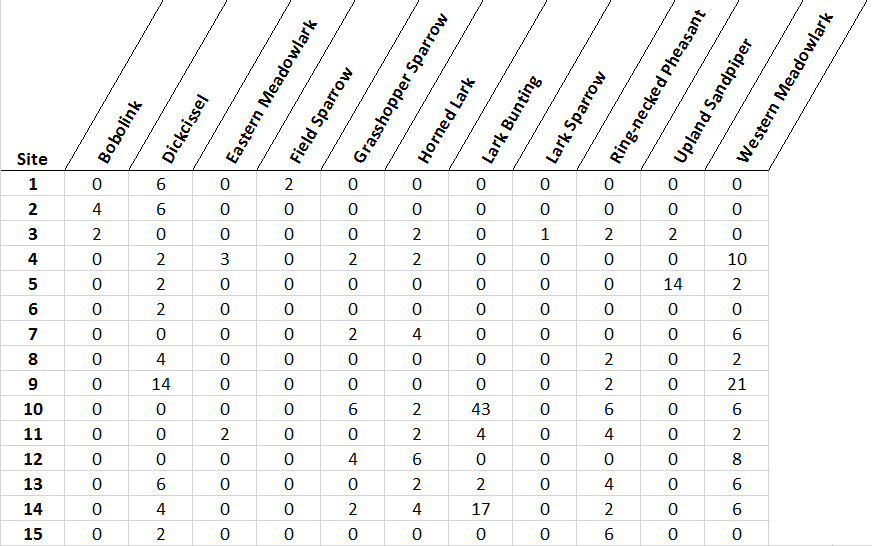
**j+ and k+** denote row sums

**+i** denotes a sum of row sums

**Table 2.** Variable loadings of measured variables on each principal component which represent the coefficients of their linear combination as well as the standard deviation associated with each principal component. For example, the 5 measures of Annulus all load strongly on component 1, but weakly on component 2. While Fish length and weight do not load as strongly on component 1 as they do on component 2. We could assume that component 1 mostly reflects fish’s annuli while component 2 described length, and weight. The proportion of variance explained by each principal component represents how well any single principal component explains the variance in the multivariate dataset, while the cumulative proportion represents total variance explained by sequentially adding each principal component. In this case, the fish biologist could safely reduce the variables (s)he considers in analytical procedures from the 9 initial variables to 6 principal components, without substantial loss of variance explained. Furthermore, as all 5 annulus measurements have very similar loadings, our fish biologist might consider only measuring 1 annulus variable in the future.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Measurement variables** | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** | **PC7** | **PC8** | **PC9** |
| **Waterbody** | -0.20 | 0.35 | 0.62 | -0.34 | -0.56 | 0.04 | -0.12 | 0.06 | 0.04 |
| **Fish Length** | -0.28 | 0.53 | 0.06 | 0.31 | 0.21 | 0.06 | 0.20 | -0.67 | 0.10 |
| **Fish Weight** | -0.29 | 0.51 | -0.09 | 0.36 | 0.18 | -0.02 | -0.18 | 0.66 | -0.13 |
| **Age** | 0.04 | 0.42 | -0.71 | -0.49 | -0.26 | -0.02 | 0.02 | -0.05 | 0.02 |
| **Annulus1** | -0.38 | -0.07 | 0.10 | -0.57 | 0.58 | 0.41 | 0.10 | 0.09 | -0.02 |
| **Annulus2** | -0.41 | -0.14 | -0.03 | -0.17 | 0.16 | -0.72 | -0.46 | -0.18 | 0.00 |
| **Annulus3** | -0.41 | -0.18 | -0.06 | 0.03 | -0.19 | -0.31 | 0.76 | 0.22 | 0.22 |
| **Annulus4** | -0.40 | -0.23 | -0.16 | 0.14 | -0.30 | 0.23 | -0.01 | -0.16 | -0.76 |
| **Annulus5** | -0.39 | -0.25 | -0.23 | 0.20 | -0.25 | 0.41 | -0.35 | -0.02 | 0.59 |
| **Standard Deviation** | 2.28 | 1.39 | 1.04 | 0.60 | 0.53 | 0.28 | 0.20 | 0.15 | 0.09 |
| **Proportion of Variance** | 0.58 | 0.21 | 0.12 | 0.04 | 0.03 | 0.01 | 0.00 | 0.00 | 0.00 |
| **Cumulative Proportion** | 0.58 | 0.79 | 0.91 | 0.95 | 0.98 | 0.99 | 1.00 | 1.00 | 1.00 |

**Table 3.** Example of grassland bird abundances sampled from 15 different field sites. An ecologist might be interested in differences between species (R-mode analysis) by calculating a matrix of distances between pairs of species (columns in community matrix), or differences between sites (Q-mode analysis) by calculating a distance matrix of distances between pairs of sites (rows in community matrix) (Legendre and Legendre 1998).



**Table 4.** Coefficient estimates and 95% credible intervals of the fixed effects of sex and temperature on three sleep variables measured from great tits (*Parus major*) roosting in nest boxes in the wild. Results were generated from a single tri-variate model simultaneously evaluating all three response variables (left), or three separate univariate models (right). Re-analyzed from Stuber et al. (2016).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Intercept (β0)** | **Sex (β1)** | **Evening Temperature (β2)** | **Intercept (β0)** | **Sex (β1)** | **Evening Temperature (β2)** |
| **Sleep Onset** | -13.21 (-16.77, -9.59) | 4.59 (-0.11, 9.66) | 0.84 (0.19, 1.61) | -13.35 (-17.50, -9.76) | 4.58 (-0.52, 9.61) | 0.81 (0.12, 1.56) |
| **Awakening Time** | -15.21 (-16.85, -13.37) | -5.26 (-7.91, -2.93) | -0.76 (-1.05, -0.39) | -15.19 (-16.92, -13.37) | -5.19 (-7.67, -2.96) | -0.78 (-1.11, -0.46) |
| **# Awakenings** | 4.11 (4.06, 4.17) | 0.05 (-0.03, 0.13) | 0.02 (0.01, 0.03) | 4.11 (4.05, 4.17) | 0.05 (-0.03, 0.13) | 0.02 (0.01, 0.03) |

**Table 5.** PERMANOVA results comparing average thoracic ventilation of anuran species between control and atrazine treated plots. The anuran community (10 species) was sampled from 30 control, and two sets of 30 treatment sites. Data from the first set of 30 treatment sites had the same variance as data collected from control sites (e.g. model assumption homogeneity of variance between groups satisfied), whereas data from the second set of 30 treatment sites had much higher variance associated with it (e.g. model assumption homogeneity of variance between groups *not* satisfied). Both sets of treatment sites had the same mean thoracic ventilation, and only differed in the variance around the mean. While we were able to detect a difference in means between the control and first set of treatment sites (left side), we were unable to detect the same difference in mean between the control and second set of treatment sites with unequal variance (right side). Based on simulated data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **df** | **MS** | **Pseudo-F** | **P(perm)** | **df** | **MS** | **Pseudo-F** | **P(perm)** |
| **Treatment** | 1 | 3986 | 2.52 | 0.02 | 1 | 5389 | 0.64 | 0.75 |
| **Residuals** | 58 | 1581 |  |  | 58 | 8358 |  |  |
| **Total** | 59 |  |  |  | 59 |  |  |  |

df= degrees of freedom

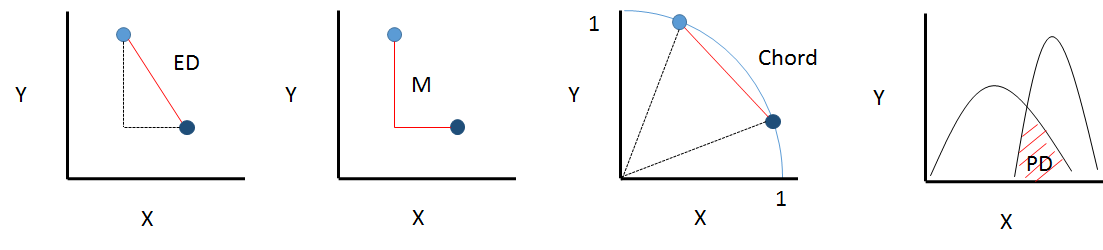
MS = mean sum of squares

Pseudo-F = pseudo F-statistic

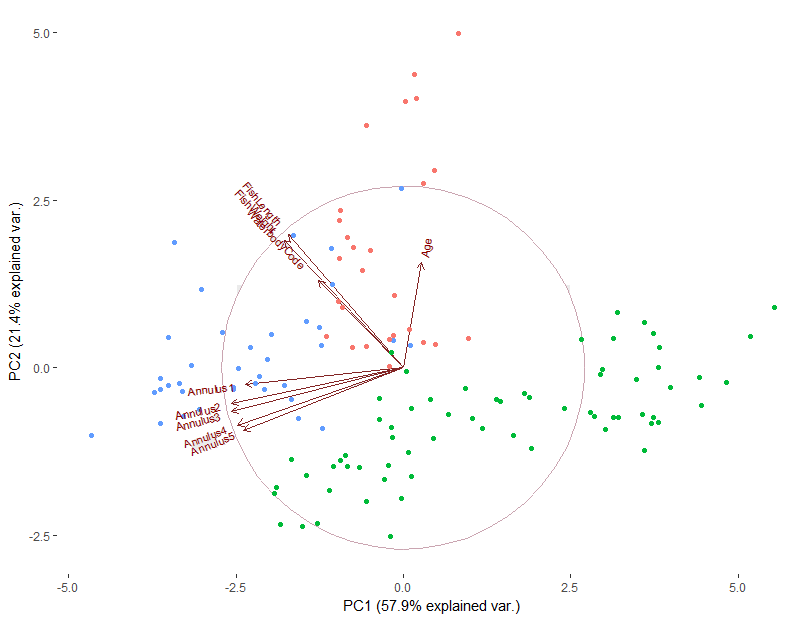
P(perm) = p-value based on 999 permutations

**Table 6.** Mantel test results testing the similarity between two matrices: simulated insect community species abundance from 5 species, and 3 simulated environmental covariates measured from 50 sampling sites. We generated environmental variables under conditions of spatial independence, and spatial autocorrelation by adding an exponential variogram (sill= 1 pixel, range= 5 pixels) to the model of spatial independence. Community abundance data were only weakly dependent on the environmental characteristics generated. Using the mantel test to determine whether two matrices are independent (community abundance matrix, and environmental variables matrix, measured at 50 sampling sites), we are unable to reject the null hypothesis that the matrices are independent when data come from spatially uncorrelated sites. However, when environmental data are collected under conditions of spatial autocorrelation, the mantel test suffers from inflated type-1 error (false positives) when the test’s assumptions of independent samples is violated (Guillot and Rousset 2013). Indeed, data generated with the same underlying characteristics but with spatial autocorrelation in the environmental predictors (a common occurrence in wildlife science data), the mantel test produced a much smaller p-value.

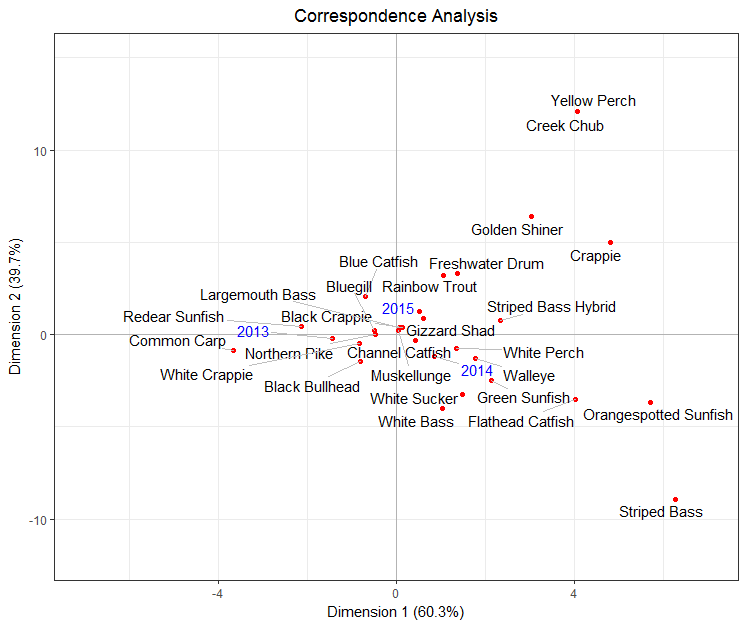
|  |  |  |
| --- | --- | --- |
|  | **r** | **P(perm)** |
| **Spatially Independent** | -0.033 | 0.665 |
| **Spatially Autocorrelated** | 0.089 | 0.086 |



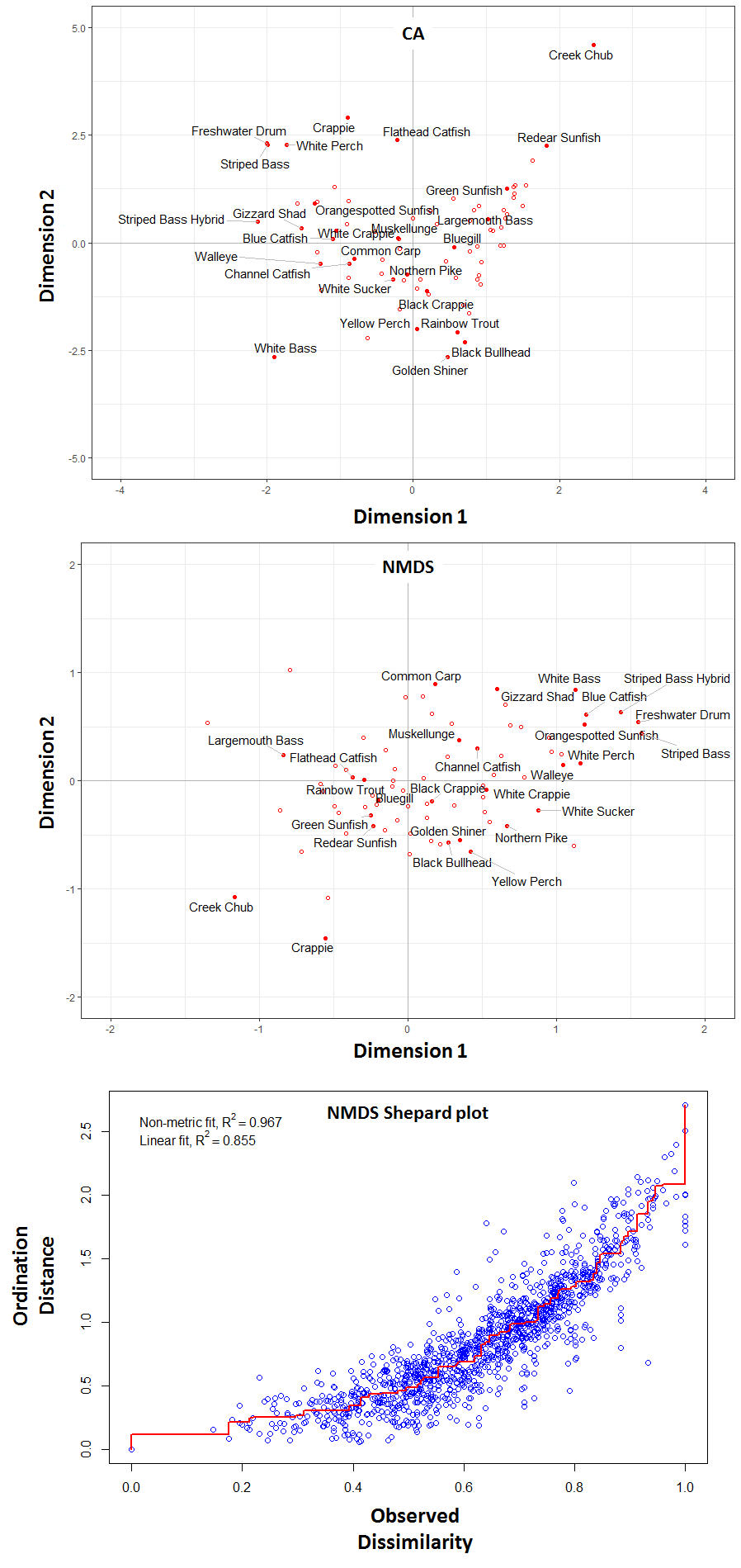
**Fig. 1.** Diagrammatic visualizations of common distance metrics. Distances between two points (e.g. species, or sites) are shown in red. ED= Euclidean distance; M= Manhattan distance; Chord= Chord distance; PD= percent dissimilarity (red hatched overlap), for example between abundance of two species along an environmental gradient.



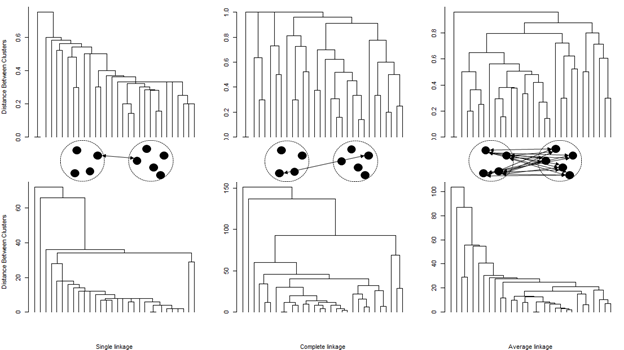
**Fig. 2.** A principal component biplot depicting fish data collected from multiple waterbodies (see Table 2). Points correspond to measures of individual fish colored by species: blue: largemouth bass, pink: channel catfish, green: bluegill. Points that are close together are individuals with similar scores for the measured variables (e.g., length and weight). Vectors represent loadings on the first two principal components and its length approximates the standard deviation in each original variable; all annulus measurements load strongly on PC1 (x-axis), and fish weight, length and age, as well as waterbody load on PC2 (y-axis). Taking the cosines of the angles between two vectors approximates the correlation between original variables. Generally speaking, if a point lies in the top right quadrant, it can be characterized by positive values of both components 1 and 2, while points that lie in bottom right quadrant would be characterized by positive component 1 values and negative component 2 values. The depiction demonstrates that fish species separate in multivariate space, as principal component scores do not overlap much between species. A fisheries biologist might use these results to classify unidentified fish species, or decide to stop collecting both fish length and weight when handling fish as they are highly redundant in this system.



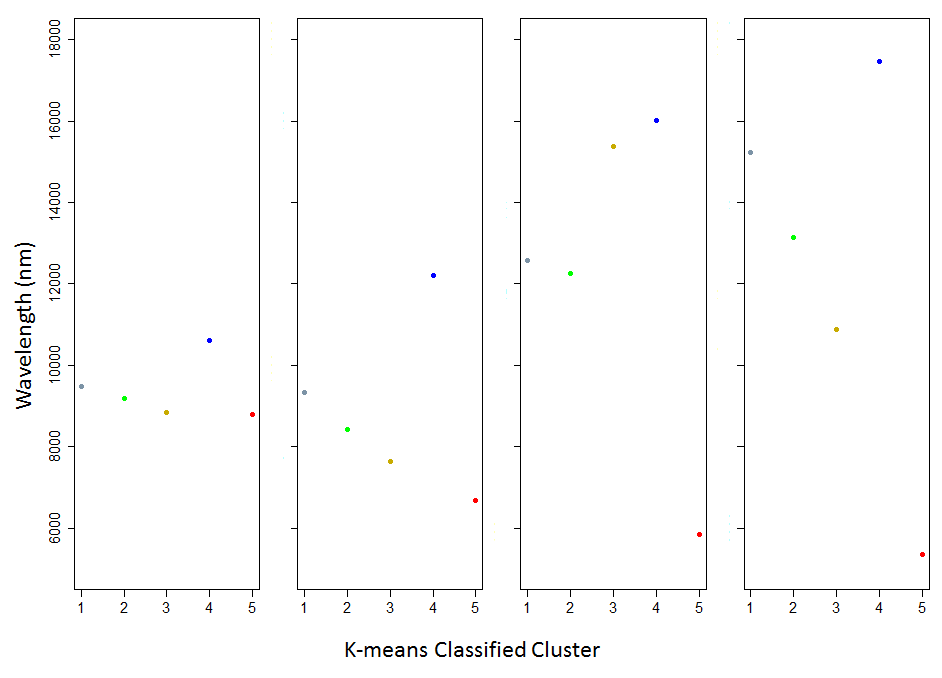
**Fig. 3.** Correspondence analysis of two-way contingency table of time series data of the number of fish caught representing 28 species (rows), from multiple waterbodies, over three field years (columns). Species frequencies (row values) are shown in black, while years (column values) are shown in blue. The distance between species is a measure of similarity; species that are closer together have similar frequency profiles over the three years (e.g., White Bass (*Morone chrysops*) and White Sucker (*Catostomus commersonii*) have similar profiles, while Striped Bass (*Morone saxatilis*) and Crappie (*Pomoxis annularis*) do not). Similarly, the distance between years is a measure of similarity. In this case, the sampling profiles of each field season distinct. The distance between row and column points is not directly comparable, however right angle projections of points onto vectors represents the value of that variable for each observation.



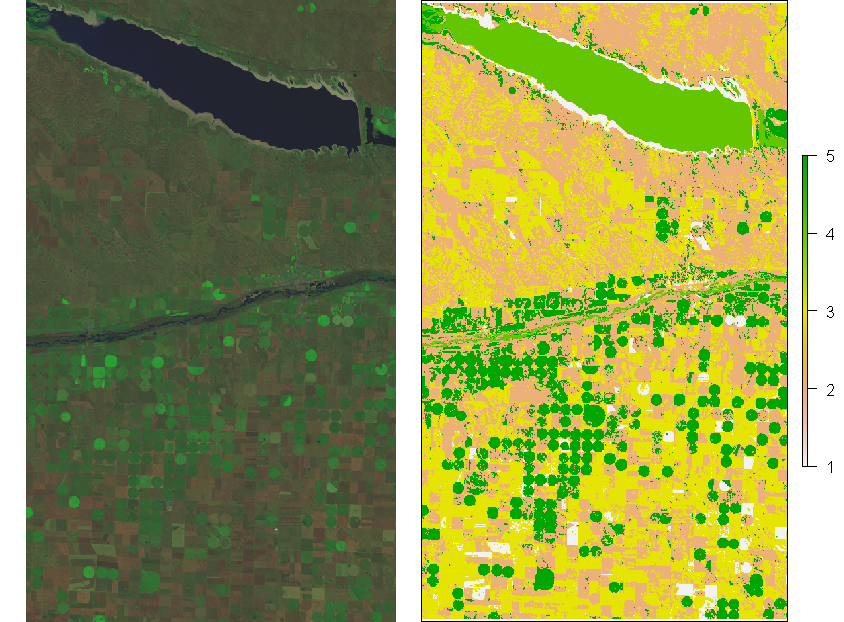
**Fig. 4.** Comparison visualizations of CA (top) and NMDS (middle) ordinations of fish species composition sampled from numerous waterbodies (open circles in ordination plots; unlabeled). In NMDS we specified 2 dimensions and used Bray-Curtis distances in our fish community × waterbody matrix. In CA, we do not specify the number of dimensions, but we show the first two components, which explain the greatest amount of variation in the data. Because the two analyses are based on different methods (e.g., metric vs non-metric, chi-square distance vs Bray-Curtis) we might expect differences in the relationships. For example, the CA clusters species closer together around the origin than the NMDS, and we can see a weak arch effect suggesting that detrended correspondence analysis might be a better fit. The NMDS Shepard plot (bottom) depicts the degree of agreement between the final and original configurations. The low degree of scatter around the regression line (red), and low ‘stress score’ (stress = 0.18) suggests that reducing the dimensionality of the data to 2 adequately represents the original dissimilarities between communities.



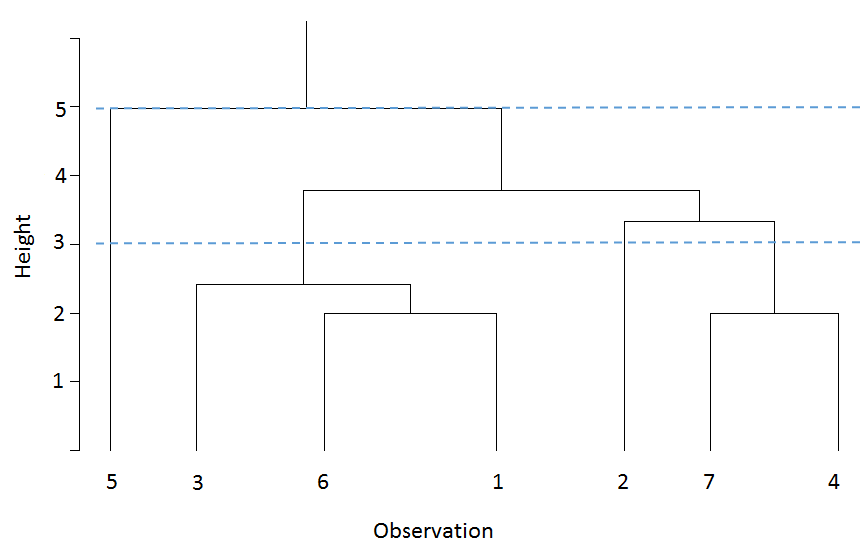
**Fig. 5.** Hierarchical agglomerative clustering of avian community sampled from 30 sites visualized as a dendrogram. Agglomerative clustering merges individual units and clusters that have the highest similarity using a linkage criterion, most commonly single, complete, or average. For example, field sites with the same species in the same abundances would have the highest similarity and be the first to cluster together. Clustering ends when all observations are merged into a single cluster (i.e., the top of the dendrogram). The composition within clusters is sensitive to the choice of clustering method and measures of distance (Orlóci 1978; Pielou 1984). For example, Bray-Curtis (top row), and Manhattan (bottom row) indices of distance produce disparate classifications and have different distance interpretations.

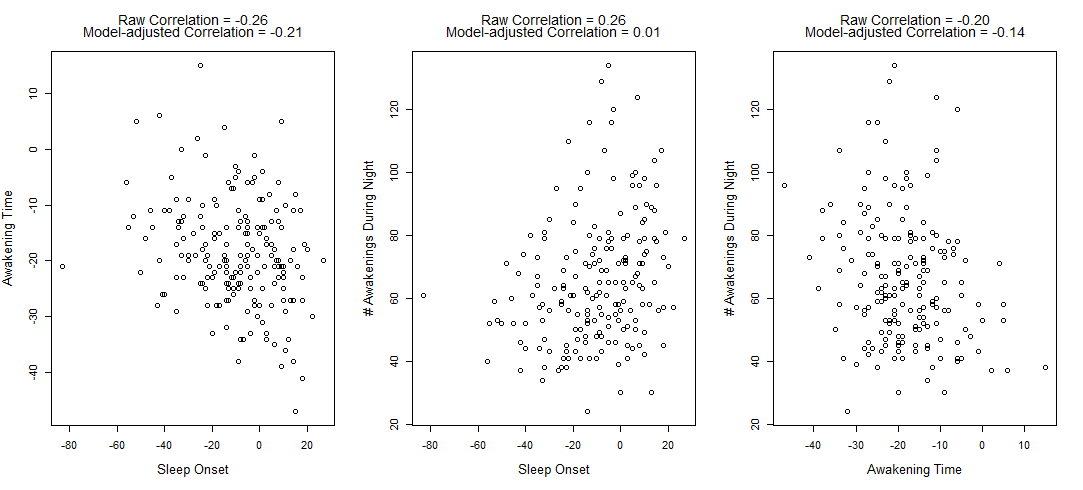


**Fig. 6.** Centroid values of five clusters segmented based on multiple clustering variables using k-means. Each panel represents a spectral band over specific ranges of light wavelengths (e.g. infra-red, ultra-violet). A light reflectance of the earth measured over seven spectral bands was used to classify a satellite image into five distinct habitat types based on separation in the mean wavelengths across multiple bands. For example, observations in cluster 5 always have the lowest average wavelengths in each of the clustering variables.

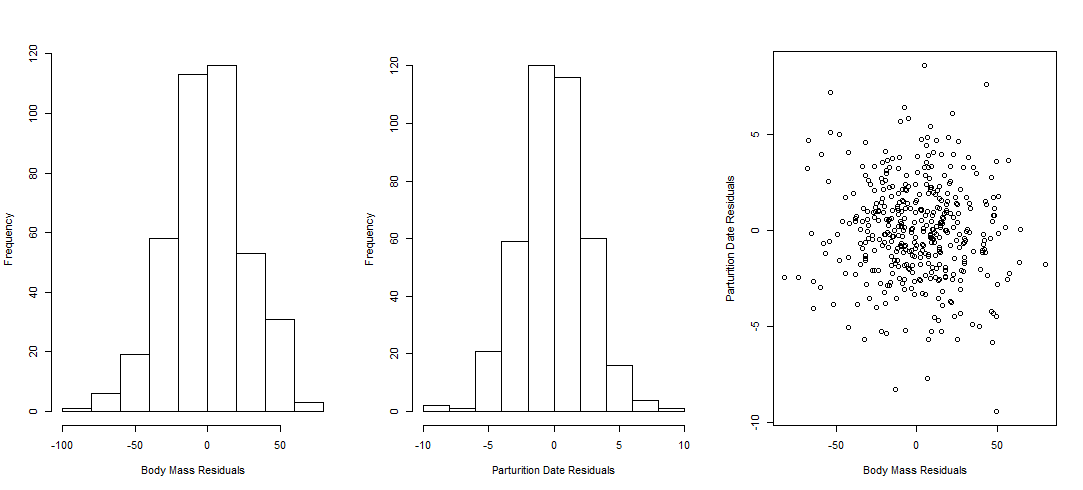


**Fig. 7.** K-means classification of multispectral Landsat 8 satellite imagery at 30 × 30m spatial resolution. The left panel depicts an un-classified satellite image near Lake McConaughy, Nebraska, USA. The right panel reflects the outcome of k-means clustering with an initial cluster value set to 5, representing hypothesized groups: water, corn fields, small grain fields, grassland, and “other”. Typically, following such unsupervised classification, field data are used to “ground-truth” the classification. If data collected from the field indicate that the k-means classification did not perform well, it is necessary to investigate the characteristics of clusters that were mis-classified by changing the number of groups, or including additional clustering variables to see if it is possible to increase accuracy. Once the classification reflects reality, the classification can be used for species distribution modeling, or investigations of resource selection.

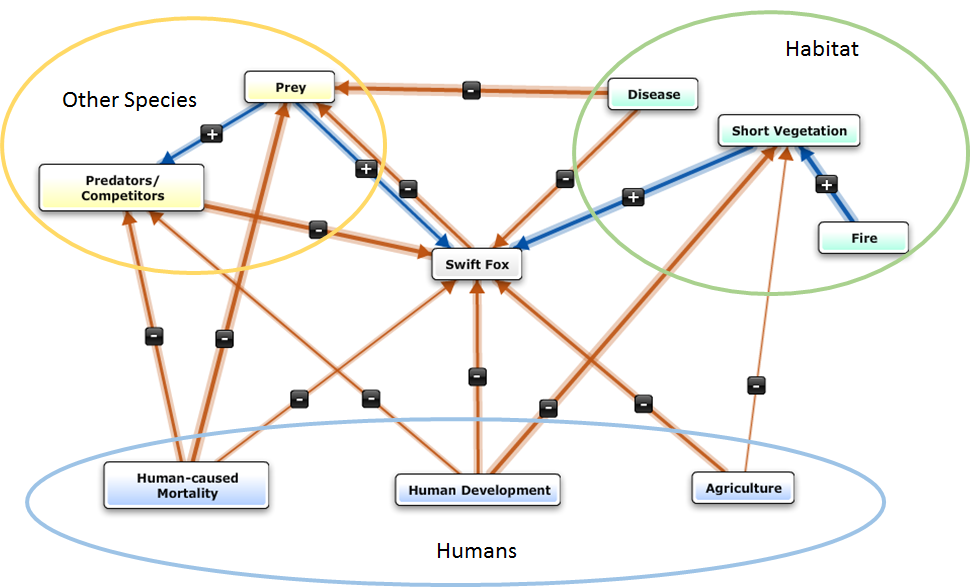
**Fig. 8.** Schematic diagram of a hierarchical cluster analysis dendrogram. The dendrogram is read from bottom to top (in the typical agglomerative analysis), with solid horizontal lines depicting when clusters are formed and becoming increasingly nested toward the top. The position of the black horizontal lines shows the distance at which the observations were clustered (e.g. cluster 7-4 was joined at height = 2). Determining the number of groups represented by the dendrogram is subjective, if the similarity cutoff for defining a group (dashed blue lines) is 3, the dendrogram suggests four clusters in the data, whereas a less restricted cutoff of 5 would suggest two clusters. The cutoff criteria represent a trade-off between the accepted similarity of observations within defined clusters and the number of clusters considered. A lack of long stems in a dendrogram, as well as evidence of ‘chaining’, when single observations are sequentially added, are evidence that clusters are not well-defined by the chosen clustering variables.



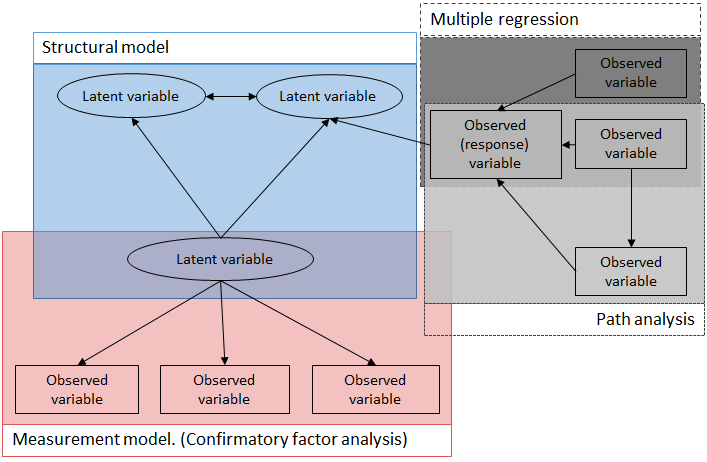
**Fig. 9.** Relationships between three variables related to sleeping behavior: sleep onset time (minutes relative to sunset; normally-distributed), awakening time (minutes relative to sunrise; normally-distributed), and the number of awakening bouts during the course of the night (Poisson distribution). Sleep measures were taken on individual birds roosting in nest boxes in the wild. Although it is possible to calculate raw correlations (Pearson correlation) between pairwise combinations of sleep variables the different sleep variables are affected by various environmental characteristics, including temperature, and intrinsic variables such as sex. A tri-variate model will provide estimates of correlations after correcting for such confounding variables. Re-analyzed from Stuber et al. (2016).



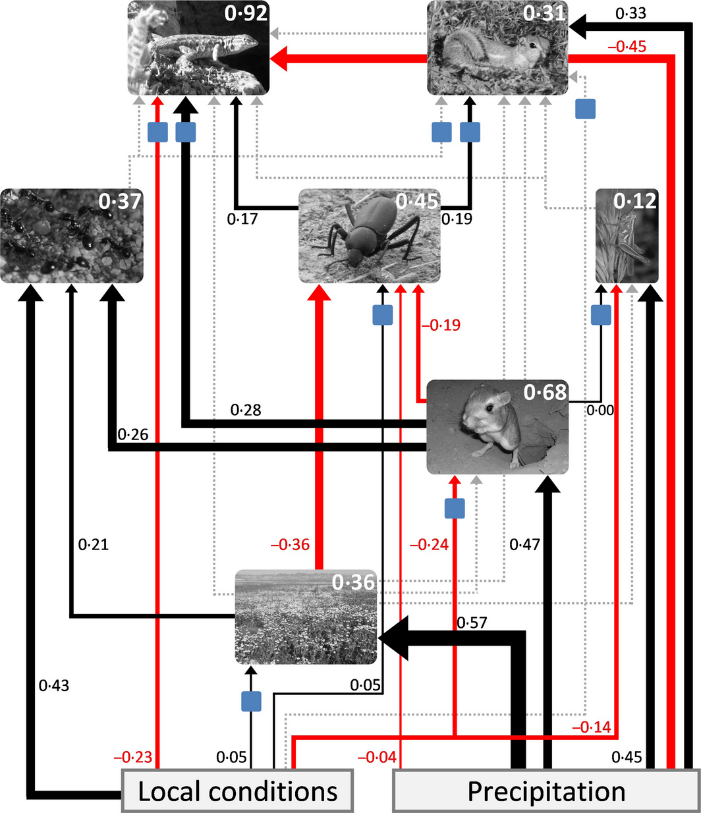
**Fig. 10.** Graphical residual analysis of multivariate regression. A biologist might be interested in the effects of supplemental feeding on size at reproduction and reproductive timing in ungulates. After collecting spring body mass and data of parturition from individual elk that were either provided supplemental feeding over-winter or not, we could model this data using a MANOVA to look for mean differences in body mass and timing of reproduction between the treatment groups (supplemental feeding or control). Once residuals are extracted from our fitted model, we can plot histograms of residuals by each predictor variable, and bivariate scatterplots of residuals for each response variable. Histograms that are not symmetric, and scatterplots that are not elliptical are indicative of data not sampled from a multivariate normal distribution, and we should consider transforming the data Histograms of MANOVA residuals for elk body mass (left panel), and date of parturition (middle panel), are symmetric around zero, and appear multivariate-normal (mean = 0, elliptical, uncorrelated; right panel). Because residual analysis does not show any indication of mis-specification, we can draw conclusions from this model.



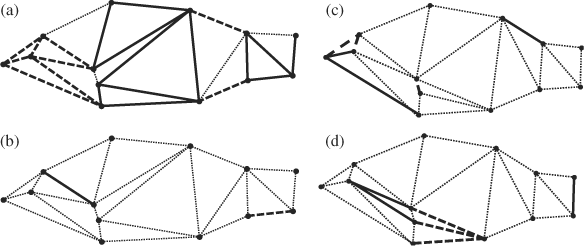
**Fig. 11.** Hypothesized network of relationships showing the factors that affect swift fox populations. Relationship webs, which are often used in community ecology, and wildlife science generally, are particularly amenable to analysis by structural equation modeling, which can simultaneously evaluate all hypothesized relationships, rather than sequentially through univariate modeling. The arrows represent direct relationships between ecological factors (e.g., other species, habitat, humans) and swift fox (the focal species), and relationships among factors that indirectly affect the focal species. The arrow represents the direction of hypothesized causality of each relationship, line thickness represents expected strength of the relationship, and positive/negative values qualifying relationships.



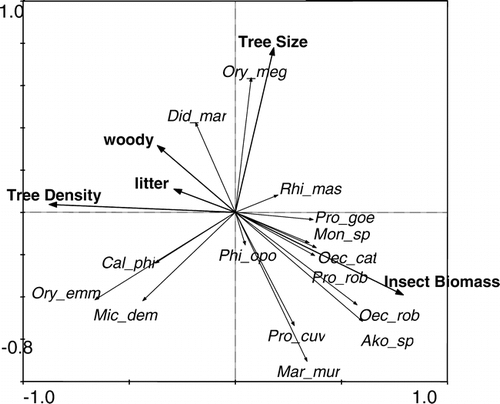
**Fig. 12.** SEM is an umbrella term encompassing many statistical techniques, including confirmatory factor analysis, path analysis, and multiple regression. S*tructural models* represent the hypothesized causal relationships between dependent and independent measured (observed) and latent (unmeasured) variables, and *measurement models* depict relationships between observed variables and the expected latent variable construct.



**Fig. 13.** Estimated structural equation model investigating the direct and indirect effects of precipitation on a grassland community (pictures). Precipitation can affect the animals of the community directly through modification of physiology or behavior, or indirectly through effects on the plant community. Coefficients of estimated effects are interpreted similar to regressions. For example, holding all other values constant, a 1 standard deviation increase in precipitation results in a 0.57 standard deviation increase in plant biomass. SEM model fit is assessed similarly to regression models. SEM output can return R2 values for each variable in the model (e.g., numbers inside picture boxes). For example, local conditions and precipitation explain 68% of the variation in giant kangaroo rat density. Fig. from Deguines et al. (2017).

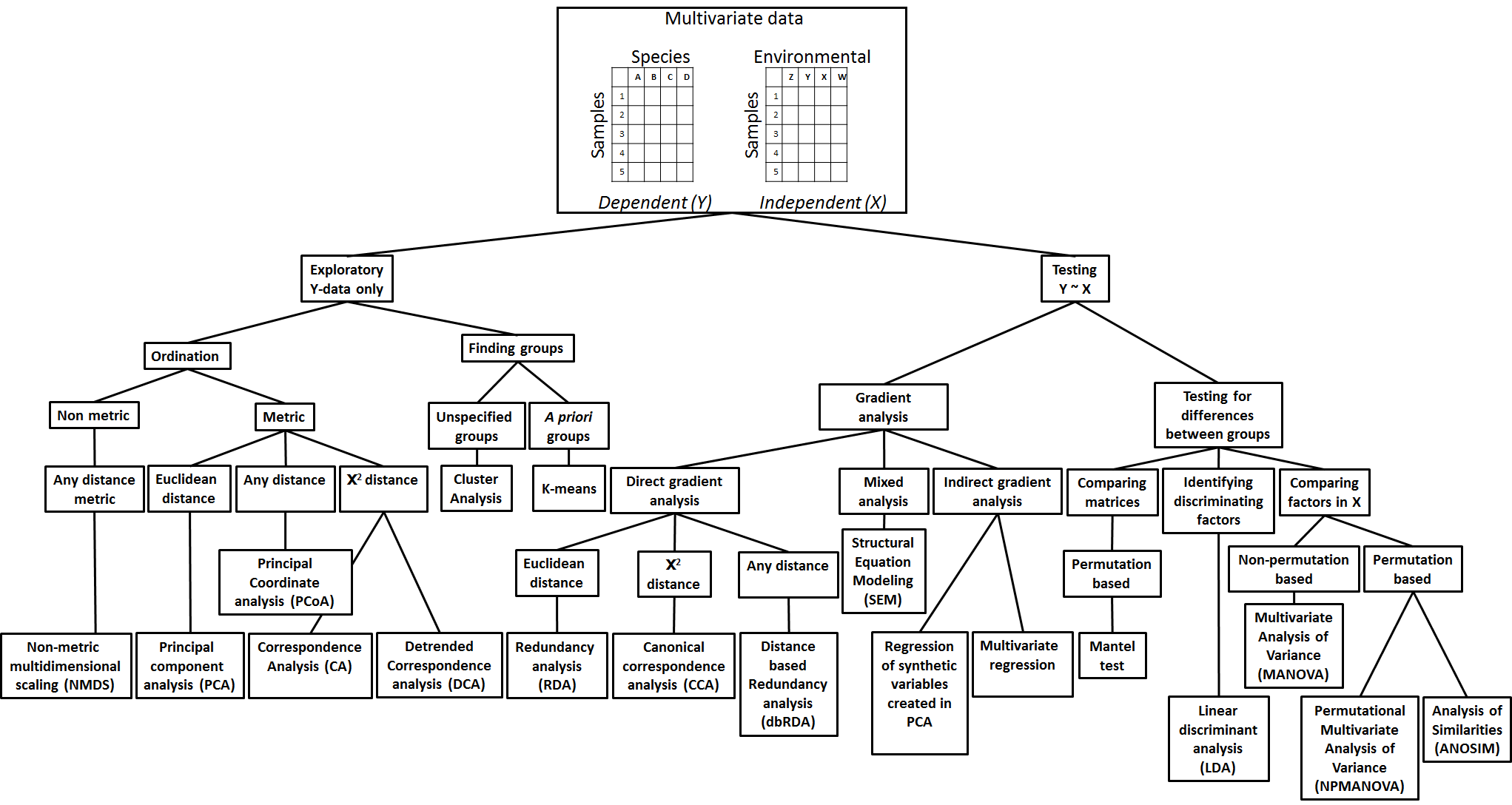


**Fig. 14.** Significant loadings and landmark co-ordinates between (a) stunted and non-stunted, and (b) adult and juvenile *Morone americana*. Solid lines depict morphological distances associated with greater distances in the groups and dashed lines depict morphological distances associated with shorter distances in the groups. Dotted lines represent measures with no association between groups. Fig. adapted from Chizinski et al. (2010).

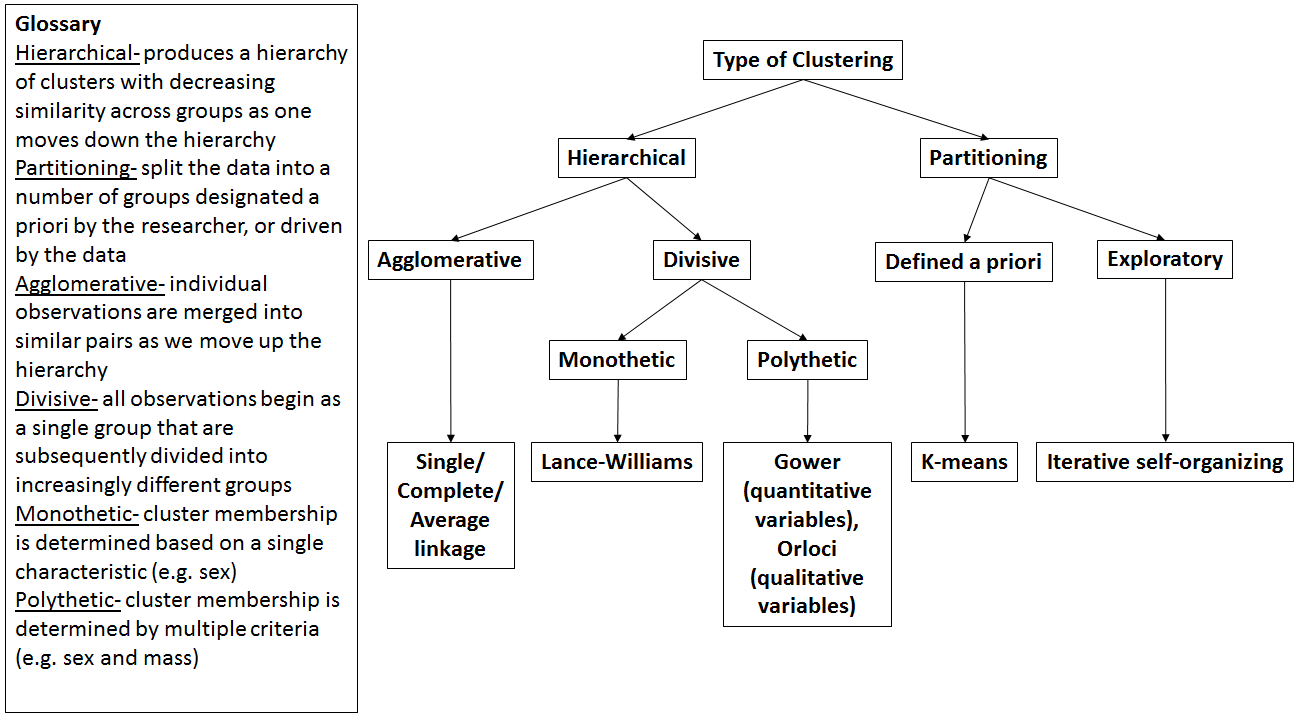


**Fig. 15.** Biologists used redundancy analysis to describe habitat use (tree density, woody understory, litter depth, tree size) by a small mammal community, and the relationships with resources (insect biomass). The angles between response (species) and explanatory vectors (tree size, tree density, woody, litter, and insect biomass) represent their correlations and distance between species represent Euclidean distances in ordination space. Two axes of redundancy analysis explained 47% of variance in species abundance and largely agreed with multiple univariate regression analyses (from Lambert et al. 2006).

**Box 3.1.** Multivariate statistics decision tree



**Box 3.2**. Overview and taxonomy of clustering process (based on Grabmeier and Rudolph 2002).



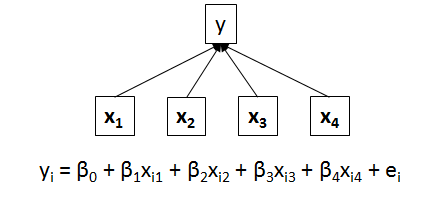
**Box 3.3.** Although the term multivariate regression is often used interchangeably with multivariable regression and multiple regression in the social science and ecological literature, they are distinct statistical concepts. Multivariable, and multiple linear regression refer to cases of a single dependent (response) variable, , but multiple independent (predictor) variables:

but multivariate regression refers to cases with multiple dependent variables, , and :

,

The equations look like two univariate linear models; however, the residual error components, and the random grouping effect in the case of repeated measures ; e.g for repeated-measures on subjects) are estimated differently. Although both components (, ) are still assumed to have a mean of 0 as in univariate models, they are no longer considered independent across dependent variables. Therefore, we can use multivariate regression models to concurrently estimate the covariance between multiple dependent variables (, ) measured simultaneously.

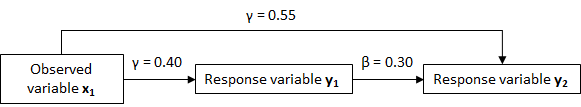
**Box 3.4.** Simple linear regressions can also be drawn as path diagrams:



When variables are connected by a single-headed arrow, the estimated coefficient corresponds with a simple linear regression coefficient. The strength of an indirect relationship is the product of the coefficients along the multi-step path (e.g., x1 on y2 = 0.40 × 0.30 = 0.12):

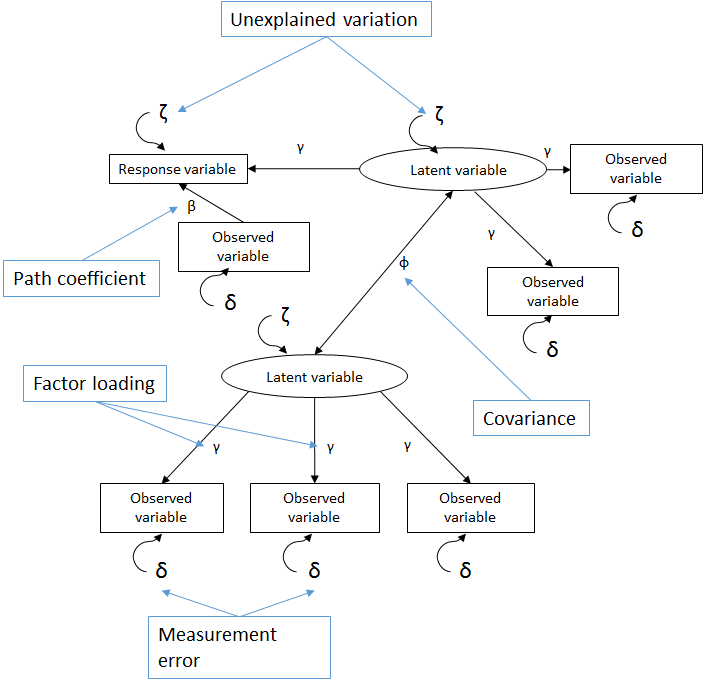


When variables are connected by multiple paths, the estimated coefficients represent partial regression coefficients. The total effect a variable has on another is the sum of its direct (x1 on y2: 0.55) and indirect (x1 on y2 via y1: 0.40 × 0.30 = 0.12) effects (sum = 0.67):



**Box 3.5.** Anatomy of a Structural Equation Model diagram.

An arrow implies causality and is an unelaborated summary of underlying causal mechanisms. Although little mechanistic information is contained, it represents the expected behavior of the system based on our previous knowledge. We learn about possible mechanisms by confronting our graphical models with data.



Single-headed arrows represent a regression while double-headed arrows represent covariation. Fitting a structural equation model will provide us with estimates of path coefficients (β; like regression coefficients) for relationships between observed (dependent) variables and response variables, factor loadings (γ; like standardized regression coefficients) for the relationship between observed variables and an underlying (unmeasured) latent variable, and correlations (φ) between response variables and/or latent variables.