Niche Overlap:

A unified definition and analysis

for data of different types

Shirley Pledger[1]

Shane W. Geange[2]

15 June 2009

[1]School of Mathematics, Statistics and Operations Research,
[2] School of Biological Sciences,
Victoria University of Wellington,
P.O. Box 600, Wellington 6140, New Zealand.

Email: shirley.pledger@vuw.ac.nz

School of Mathematics, Statistics and Operations Research
Research Report 2009-05
ISSN 1174-2011

SUMMARY

The partitioning of available resources within ecological communities has been suggested as contributing towards the determination of community structure (Pianka 1974). As such, quantifying the degree of species overlap in the utilisation of resources such as food, space, and shelter has become a valuable approach in studies of both community structure and species coexistence. Traditionally, overlap in resource use has been quantified as the degree of niche overlap between species, where niche overlap is simply the joint use of a resource (or resources) by two or more species (Hutchinson 1957, Colwell & Futuyma 1971).

Niche overlap between species may be viewed as the volume in multidimensional hyperspace in which two or more species maintain a viable population(s) in the presence of one another (Mouillot *et al.* 2005). Depending upon the organisms in question, a number of discrete or continuous biotic or abiotic variables may be considered as hypervolume axes. These may include (i) an environmental condition (e.g. altitude, temperature, salinity, soil type), (ii) resource use (e.g. prey type, refuge type, host identity), (iii) a physical trait which indicates the type of resource used (e.g. for fish, gut length may be indicative of diet), or (iv) a measure of resource usage (e.g. habitat selection or food preference). Furthermore, a single study may incorporate a number of different types of variables, including categorical, continuous count or binary data, as well as electivity scores (on a 0,1 range); however, incorporating axes' described by different data types into a single measure of niche space is potentially problematic because statistically, different data types cannot be dealt with in the same way.

Here, we propose general methods for combining different data types within a unified multivariate analysis of niche overlap. Using appropriate transformations and density estimation techniques, each data type gives rise to a standard measure of niche overlap ranging from 0 (no overlap) to 1 (complete overlap). The use of a standard measure of niche overlap ensures that the geometric interpretation of the overlapping density functions or probability is the same for each data type. Once estimated probability distributions are available for each data type, the overlap statistic between two species is simply the overlapping area between the distributions for each species. It is then possible to blend measurements derived from different types of data into a single multivariate analysis of niche overlap by averaging over multiple axes. We then use null models to differentiate between species occupying similar and different niches.

We outline the measurement of niche overlap in Section 2. Section 3 details the different types of data that can be incorporated into our unified analysis of niche overlap, while Section 4 provides methods for graphically representing overlap between species. Finally, Section 5 outlines the construction of null models to statistically test for significant niche separation between pairs of species at the same site, or over different sites for the same species.

Contents

1 Introduction

Niche overlap between taxa is a measure used in the description and analysis of biological community structure. The taxa examined may be species, subspecies, functional groups, or some other classification. In this report we will refer to species, on the understanding that this is interchangeable with other groupings.

A niche may be visualized as the volume in multidimensional hyperspace within which a species can maintain a viable population (Hutchison 1957, Mouillot *et al.* 2005). Each axis of the hypervolume may be one of the following types.

1. An environmental condition (e.g. altitude, temperature, salinity, soil type).
2. Resource use (e.g. prey type, refuge type, host identity).
3. A physical trait indicating the type of resource used (e.g. for fish, gut length may be indicative of diet).
4. A measure of resource use (e.g. habitat selection or food preference). This may be modelled as an electivity score (e.g. Manly's alpha, Manly 1974, Chesson 1978, 1983), which measures resource use relative to resource availability. For example, the proportion of time spent in each habitat type relative to the availability of each habitat type, or the proportion of each food type found in stomach contents analysis relative to the availability of each food type. Electivity scores are usually represented on a scale of $0 - 1$.

A single study may incorporate one or more axes of hypervolume space. Some axes' are measured on each individual (e.g. physical traits), while others are site-specific (e.g. the environmental conditions above). A measure of resource usage is based on individual usage, but becomes site-specific when the electivity is calculated.

Traditionally, niche overlap was calculated using categorical data (e.g. presence or absence of a particular predator, or level of availability of a certain type of food, low, medium or high) (Ludwig and Reynolds 1988, Chapter 10). Because continuous data is seldom normally distributed, continuous variables were converted to categorical data (e.g. maximum daily temperature to low, medium and high categories); however, there is a loss of information when continuous data is replaced with ordered categories.

Mouillot *et al.* (2005) provided a breakthrough, describing an approach based on kernel distribution estimators that models niche overlap from continuous data independently of the underlying distribution of the data. Their approach extended niche overlap studies to allow the construction of broad, multivariate indices of niche overlap based on probability distributions of continuous measurements constructed using density estimation, that were therefore, comparable across different axes.

Here, we extend the work of Mouillot *et al.* (2005) by proposing general statistical methods for combining different kinds of measures within a unified multivariate analysis. The traditional binary and categorical measures may be combined with the continuous-variable quantitative functional traits of Mouillot *et al.* (2005), and other measures such as ratios, count data, and electivity scores on a scale of 0-1, using transformations and density estimates appropriate for each particular data type.

2 Niche Overlap Measurement

Several measures of niche overlap are discussed in Ludwig and Reynolds (1988). We first show how multivariate niche overlap is defined if niche overlap on individual axes is available, and then discuss the individual axes.

Assume there are T measures (resources, environmental conditions, quantitative functional traits, etc.), each of which is one axis within a broader multidimensional niche overlap estimate. Suppose that each axis provides a measure of niche overlap NO between species i and j, and that these measures are comparable over the different axes. Then niche overlap (NO_{ijt}) between species i and j is calculated for each axis t separately, and the overall niche overlap for these two species is defined as the mean niche overlap over all the measures:

$$NO_{ij} = \frac{\sum_{t=1}^{T} NO_{ijt}}{T} \tag{1}$$

with associated variance

$$s^2_{NO_{ij}} = \frac{\sum_{t=1}^{T} (NO_{ijt} - NO_{ij})^2}{T}$$

as in Mouillot *et al.* (2005) equations (6) and (7). If desired, highly correlated measures may be downweighted by using a weighted average and variance, with measure t having weight w_t such that $\sum_{t=1}^{T} w_t = 1$ (Mouillot *et al.* 2005, equations (8) and (9)).

The extension from one to multiple dimensions (above) is valid provided the individual axes have comparable measures of NO. In (Mouillot *et al.* 2005), all axes had continuous measurements, and empirical probability distributions were constructed using density estimation. This made all axes comparable, ensuring that the average measure above was balanced and appropriate.

However, niche overlap axes may arise from various kinds of data. There is a loss of information involved in either (i) making all the measures categorical (e.g. replacing continuous data with categories, or abundance with presence/absence or low, medium and high categories), or (ii) discarding categorical or discrete measures in order to use only continuous variables.

We propose direct modelling of each data type, extending the work of Mouillot *et al.* (2005) to provide comparable NO measures, whether from categorical, discrete or continuous data. Each data type gives rise to a standard niche overlap measure, making it possible to blend the disparate measures into a single multivariate analysis. For each data type, the index of niche overlap is constructed to range from 0 (no overlap) to 1 (complete overlap), ensuring that the geometric interpretation of overlapping density functions or probability functions is the same in each case.

For each combination of axis t ($t = 1, 2, \ldots T$) and species i, the data provide an **estimated probability distribution**. Discrete or categorical data give an estimated probability function, represented by a bar chart. Continuous data give an estimated probability density function (using density estimation, as in Mouillot *et al.* 2005), represented by a probability density curve from 0 to 1, with area under the curve equal to one. There are also combined data types, with a mixture of discrete and continuous probabilities, represented geometrically by a combination of bars and a curve.

Once estimated probability distributions are available, the overlap statistic between two species is a simple overlapped area between two distributions. Since this method is used for each data type, averaging over the different axes gives a valid multivariate overlap measure.

3 Types of Data

We now consider the restriction to a single axis, t. Different possible types of data are listed and discussed below. For each data type, a probability distribution may be estimated for each species, and the overlap of species i and j is defined as the overlap of the probability distributions. This generic measure of niche overlap along axis t will lie between 0 and 1 (inclusive).

Data may be collected at a single site, or over multiple sites. At a single site, individual-based data are used to provide niche overlap measures. Over multiple sites, site-based data (e.g. environmental variables) may also be incorporated.

Sections 3.1 to 3.8 discuss data types which may be

- individually based, in which case niche overlap may be calculated
  - for a single site (if there is only one),
  - separately for each site (if there are multiple sites), or
  - as a single niche overlap measure obtained from pooling data over all sites,

- site based, in which case a single measure of niche overlap is calculated using multiple sites, with just one observation per site.

Section 3.9 discusses data which come from individuals, but must have a site-specific adjustment made before being comparable between sites.

NOTE: For environmental variables (e.g. frost), we assume the numbers of each species at each site reflects the density of those species.

3.1 Binary Data

Examples are (i) the presence/absence of a certain prey type in the diet (a resource use axis for animals), with indicator 1 for presence and 0 for absence, (ii) an indicator of whether the locality has frosts (an environmental condition axis for plants).

We assume a Bernoulli (binary) distribution for the indicator, with Species i having probability p_i for "success" (value 1) and $q_i = 1 - p_i$ for "failure" (value 0). The proportion of individuals of species i with indicator 1 is the estimate of p_i. Similarly p_j for species j is estimated by the proportion of individuals of species j with indicator value 1.

The niche overlap between species i and j (on axis t) is defined as

$$NO_{ijt} = \min(p_i, p_j) + \min(q_i, q_j).$$

Example

Suppose $\hat{p}_i = 0.85$ and $\hat{p}_j = 0.30$. The niche overlap calculation is in this table:

	Failure	Success	Total
Species i:	0.15	0.85	1
Species j:	0.70	0.30	1
Minimum:	0.15	0.30	0.45

This niche overlap of 0.45 is represented graphically in Figure 1. Species i and j are represented respectively by red and blue hatched probability bar charts. If the bars have width one, the red and blue hatched areas are each one. The niche overlap is the cross-hatched area.

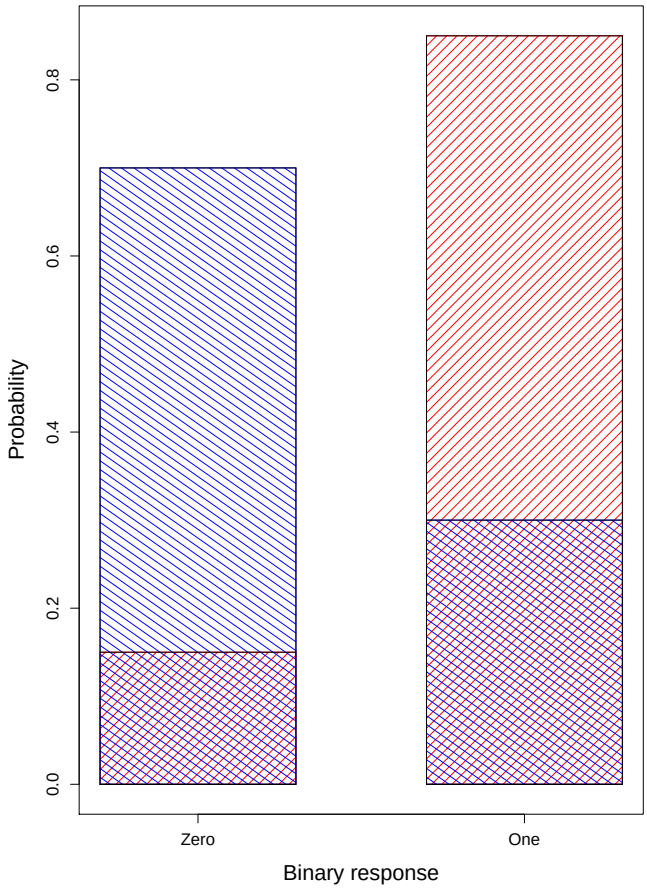

Figure 1: Niche overlap from binary data. Species i (red hatching) has 15% zeros and 85% ones, and Species j (blue hatching) has 70% zeros and 30% ones. With bar widths equal to one, the red and blue hatchings each have total area one, and the niche overlap is the cross-hatched area common to both species, $NO_{ij} = 0.45$.

This measure of niche overlap from binary data ranges from zero (no overlap) to 1 (completely identical distributions).

Note that a binary variable which is intrinsic to the species will give data with all zeros or all ones, and $p_i = 0$ or 1. For example, if birds of species i have bills long enough to reach nectar from flowers of plant species t, $p_i = 1$ and $q_i = 0$. If bird species j also has a long enough bill, $p_j = 1$ and $q_j = 0$, giving an exact match of probability distributions and hence $NO_{ijt} = 1$. If, however, bird species j has a short bill and cannot reach the nectar, $p_j = 0$, $q_j = 1$ and $NO_{ijt} = 0$.

3.2 Categorical Data

An environmental example of axis t could be the habitat type in which animal species i spends its time.

Suppose there are K categories (e.g. of habitat), all assumed to be equally available to species i. The proportional use of category k by species i is written p_{ik}, assuming $\sum_{k=1}^{K} p_{ik} = 1$. In the habitat example, p_{ik} might be estimated by the proportion of observed individuals of species i associating with habitat k. Similarly species j has proportions p_{jk}.

In an extension from binary data (two categories) to K categories, the niche overlap between species i and j (on axis t) is defined as

$$NO_{ijt} = \sum_{k=1}^{K} \min\left(p_{ik}\ p_{jk}\right).$$

Example

An example of data and calculations with $K = 4$ follows:

Category:	A	B	C	D	Total
Species i	0.10	0.35	0.40	0.15	1
Species j	0.45	0.25	0.20	0.10	1
Minimum	0.10	0.25	0.20	0.10	0.65

This example, with niche overlap 0.65, is illustrated in Figure 2. Species i and j are represented by bar charts with red and blue hatching respectively. If each bar has width 1, red and blue each have total area 1. The niche overlap is cross-hatched, with area 0.65.

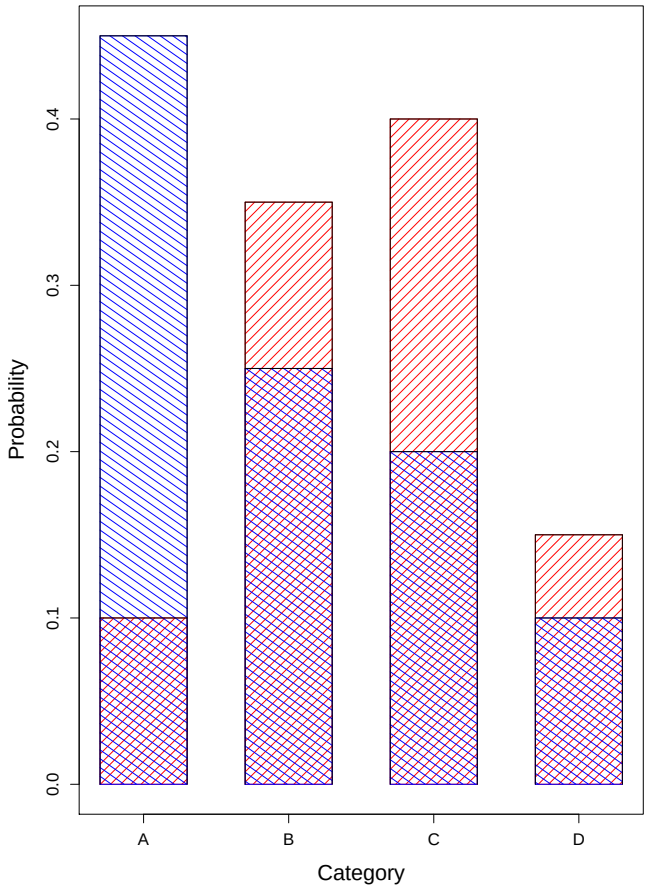

Figure 2: Niche overlap from categorical data. Species i has red hatching, and Species j has blue. If the widths of the bars are one, each species has hatched area equal to one, and the niche overlap is the cross-hatched area common to both species. In this figure, $NO = 0.65$.

3.3 Continuous Data

Examples of continuous measures include many quantitative functional traits (e.g. fish eye diameter), and environmental covariates (e.g. soil pH, average minimum overnight temperature).

Mouillot *et al.* (2005) employed kernel-based density estimation to convert a finite data set into a continuous probability density of flexible shape, thus avoiding two problems, (i) the loss of information involved in replacing continuous measurements with discrete categories (e.g. low, neutral and high pH), and (ii) the unwarranted assumption of normality (or some other particular shape of distribution) if a single continuous distribution is fitted to the data. Density estimation by the kernel method (Silverman, 1986) gives a smooth, flexible, nonparametric curve for a probability density function over the data points. Niche overlap between species i and j is defined as the overlap of their density functions. Logistically, it is necessary to use the same set of x-axis values for both species, in order to calculate the overlap curve on a point-by-point basis.

Example

As an example, we take the following data:

Species i: {2,2,3,5,6,6,7,8,9,10}
Species j: {4,4,5,6,7,7,8,9,12,13,14,15}

After density estimation for each species using R (R Development Core Team 2007, function `density`), the fitted density curves and overlap are shown in Figure 3. The area under each probability density curve is one, and the overlap in this example is 0.72.

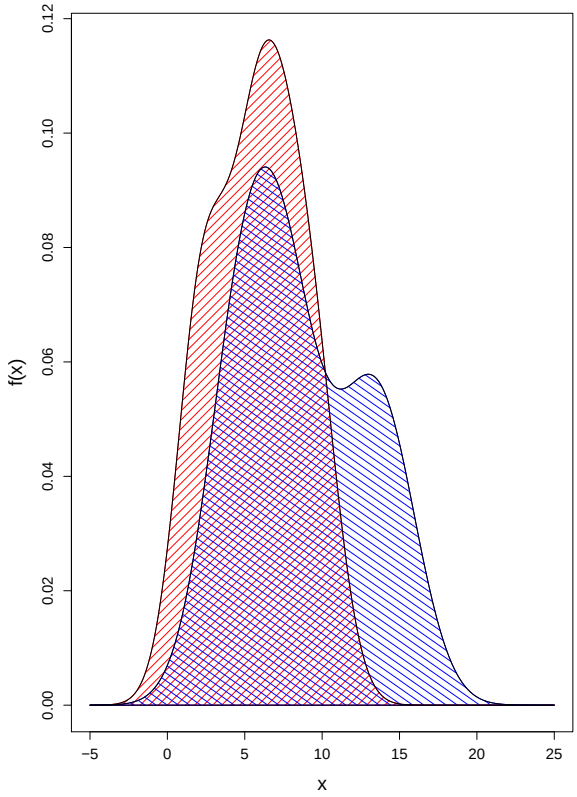

Figure 3: Niche overlap from continuous data. This plot of the fitted probability density functions has Species i hatched in red and Species j in blue, each with hatched area equal to one. The niche overlap of 0.72 is the cross-hatched area.

A potential problem arising from the use of density curves is the extension to $x < 0$ when the data were positive measurements. An overflow of this type is seen in Figure 3, where both curves have some probability assigned to $x < 0$. This results from modelling the observed data with a mixture of normal distributions, which do not observe a restriction to positive values.

The method above is appropriate if the data can go below zero (e.g. minimum overnight temperature in ^{0}C); however, if the data are measurements, which must be positive, a slightly different method is required. Silverman (1986, section 2.10) states that with positive data, it is preferable to obtain density estimates with $f(x) = 0$ for negative x, and suggests estimating the density of log(x) instead. This is considered in the next section.

3.4 Measurement Data

If the data are positive (e.g. measurements), we estimate a mixture of normal densities for the log data.

Figure 4 shows the data from the previous section with densities estimated on the log scale.

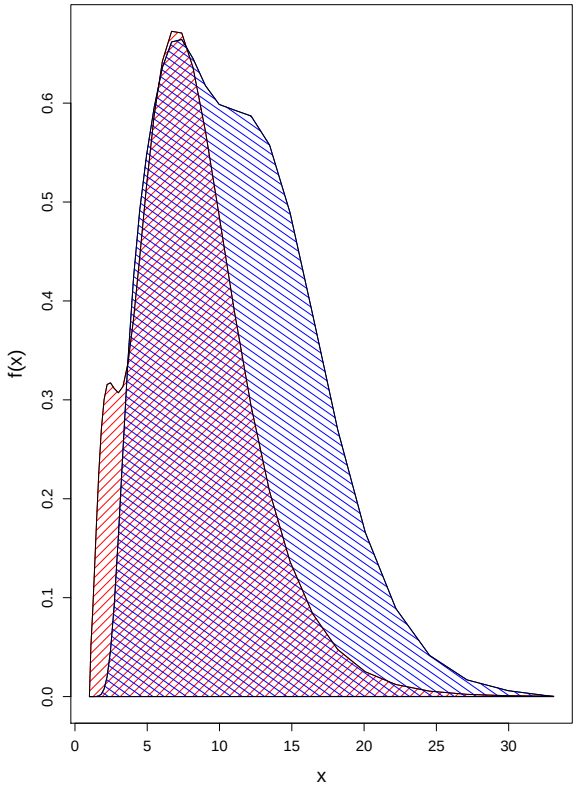

Figure 4: Niche overlap from log-transformed data. This figure uses the same data as in Figure 3, but with a log transformation. For ease of comparison with Figure 3, , the x axis has been back-transformed to the original scale. Each probability density function has area one, with species i hatched in red and Species j in blue. The niche overlap of 0.77 is the cross-hatched area.

Although Figures 3 and 4 show different densities, the niche overlap statistic appears stable (0.72 and 0.77 respectively), with the log-transformations making little difference; however, the differences may be more pronounced in other cases. We therefore, on statistical grounds, recommend using log-transformations for data restricted by a lower bound of zero.

3.5 Ratio Data

Examples of ratio data include quantitative functional traits, e.g. ratio of length:depth of the caudal fin which characterizes the swimming performance of reef fish.

This is a continuous positive measurement, and so (as with the previous measurement data type), density estimation is appropriate, preferably on log-transformed data values to give a better fit to the data.

3.6 Proportion Data

Some ratios (for example proportion data) may be bounded below by zero and above by one (e.g. the ratio of tail length to total body length in lizards, a surrogate for fat storage). Since this ratio is bounded between 0 and 1, an estimated density curve on the raw data may overflow either of the (0,1) bounds. To prevent this overflow, density on the logit-transformed data is recommended. If X is the proportion (e.g. (tail length)/(total length)), we use

$$Y = \log\left(\frac{X}{1-X}\right)$$

for the density estimation.

3.7 Percentage Data

Percentage data (e.g. tail length is 40% of total length) is similar to proportion data in that it is bounded above and below. The appropriate transformation is

$$Y = \log\left(\frac{X}{100-X}\right)$$

Example

As an example, we take the following percentage data:

Species i: {0,4,15,18,27,34,39,55,68,73}
Species j: {11,16,32,45,50,63,74,77,80,86,98,100}

Both raw and logit-transformed data (log $\frac{X}{100-X}$) had density curves fitted, with curves and overlaps shown in Figure 5 and 6. The area under each probability density curve is one, and the overlaps are 0.677 and 0.684 respectively. Again, the transformation has not greatly changed the niche overlap statistic with this data set. Taking logits has removed the concern about fitting positive probabilities for some unrealistic data values less than 0% or greater than 100%.

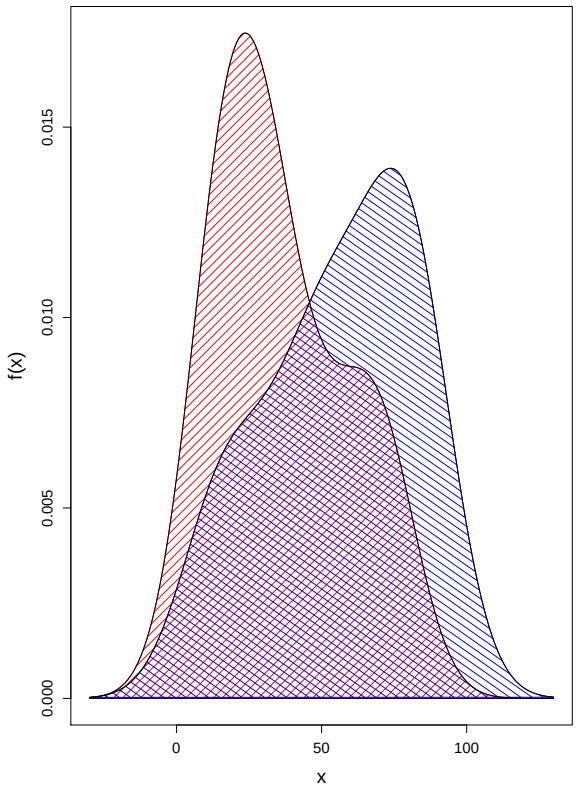

Figure 5: Niche overlap from untransformed percentage data. This plot of the fitted probability density functions has Species i hatched in red and Species j in blue, each with hatched area equal to one. The niche overlap of 0.677 is the cross-hatched area. The use of untransformed data has resulted in parts of the fitted distributions going outside the bounds of [0,100].

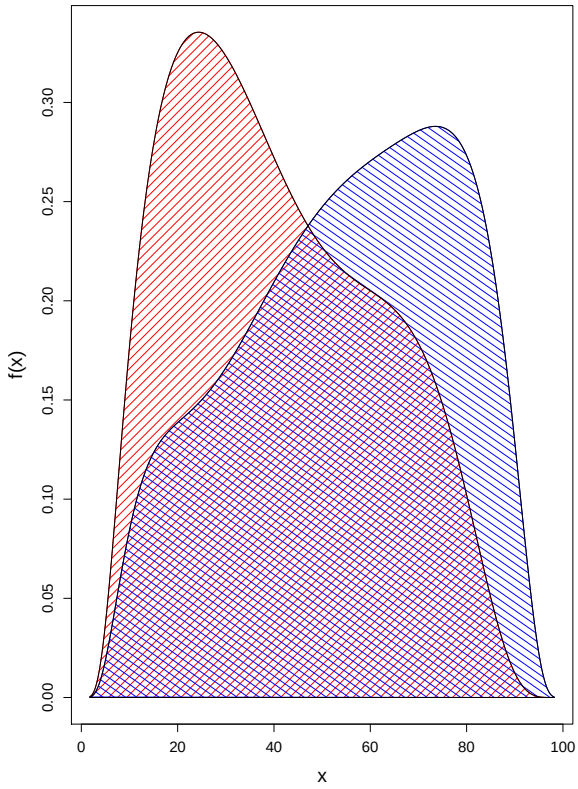

Figure 6: Niche overlap from percentage data. A logit transformation was used for the analysis, with back-transformation for presenting the plot. The fitted probability density function for Species i is hatched in red, and Species j in blue, each with hatched area equal to one. The niche overlap of 0.684 is the cross-hatched area. The logit transformation has ensured the fitted distributions are within the [0,100] bounds.

3.8 Count Data

Examples of count data include the number of leaves on the terminal shoot of a plant (a plant functional trait, Stubbs and Wilson 2004), or the number of prey eaten by an individual in one hour of observation.

Instead of treating count data as continuous, direct modelling provides more information and avoids having density estimates $f(x)$ outside the natural range of X (i.e. fractional values, when the data were integers). For species i there will be records from several individuals. Assuming a single Poisson distribution with parameter λ for all the observations imposes a restrictive shape which may be inappropriate for the observed data, especially in the case of heterogeneity and overdispersion, where the data show a variance above that of a single Poisson distribution. For example, larger individuals of species i may have a higher average rate of prey consumption than smaller individuals.

A more flexible shape is provided by taking a finite mixture of Poisson distributions, which is analogous to the mixture of normal distributions used in kernel density estimation for continuous data. Using **non-parametric maximum likelihood estimation** (NPMLE, Lindsay 1983), a mixture of finitely many Poisson distributions is fitted to the count data for individuals of this species, with the AIC criterion (Akaike, 1973) providing an objective choice of the number of components needed to provide a good fit while keeping the number of parameters down. This choice of the number of components is similar to the bandwidth selection in the continuous case density estimation (Mouillot *et al.* 2005).

Example

Suppose individuals of Species i and j have counts for axis t, as follows.

Species i: {0,0,0,0,0,1,1,1,1,2,2,2,3,4,6,6,7,8,12}
Species j: {0,2,4,6,9,9,10,15,19,22,24}

Maximum likelihood fitting of Poisson mixtures gave the following information, where relative AIC = AIC - minimum AIC for each species.

Species	No. components	Relative AIC
i	1	22.32
	2	0.00
	3	4.00
	4	8.04
j	1	30.18
	2	2.96
	3	0.00
	4	3.08

We select the two-component mixture for Species i and three components for Species j, as indicated by the lowest AIC. If the fitted distributions for Species i and j have probabilities p_{ix}, p_{jx} respectively, for values $x = 0, 1, 2, \ldots$, the niche overlap for axis t is calculated as

$$NO_{ijt} = \sum_{x=0}^{\infty} \min\{p_{ix},\ p_{jx}\}.$$

In practice, the sum is taken from 0 to $x_{\max}$, where $x_{\max}$ is set high enough for the probability to be almost zero for both species.

Figure 7 shows the fitted Poisson mixtures, with Species i and j hatched in red and blue respectively. The niche overlap of 0.44 is cross-hatched.

The first fitted mixture of Poisson distributions had 67% with $\lambda = 1.0$ and 33% with $\lambda = 6.90$, while the second species had the fitted mixture with three Poisson distributions, with Poisson parameters 1.1, 7.7 and 19.9, in proportions 18%, 46% and 36% respectively. Essentially this second mixture has grouped the data into {0,2}, {4,6,9,9,10} and {15,19,22,24}.

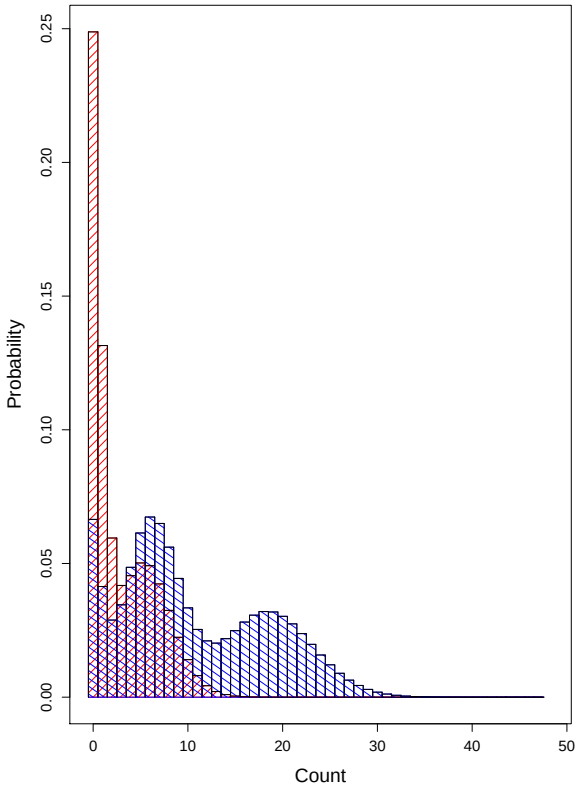

Figure 7: Niche overlap from count data. The fitted Poisson mixture distribution for Species i is hatched in red, with Species j in blue, each having hatched area equal to one. The niche overlap of 0.44 is the cross-hatched area.

A variation on this Poisson mixture model is possible, if the nature of the data dictates that a zero cannot occur - for example, the number of different prey species in the stomach contents, or the number of leaves on the terminal shoot, may have a minimum of one. In this case, the Poisson mixture model is easily adapted to be made conditional on readings of zero not occurring.

3.9 Electivity Data

Electivity scores are used where selection of a resource category is related to the availability of that resource. The resource could be habitat, substrate or prey, with different categories representing the different resource options available.

Manly (1974) proposed an electivity score (Manly's Alpha), which ranges between zero and one, with zero indicating that a resource type is never used, and one indicating that a resource type is exclusively used. For each species at one site, two vectors are needed. If there are R resource types (different habitats, different prey types), there is a **usage vector** f of length R, with component f_r being the usage of resource type r by the chosen species. The **availability vector** g has component g_r being the availability of resource type r at this site. The definition of Manly's alpha is

$$\alpha_r = \frac{\frac{f_r}{g_r}}{\sum_{r=1}^{R} \frac{f_r}{g_r}}.$$

The denominator in the α formula is used to scale the values to be on the interval [0,1].

We distinguish two cases: (i) calculating niche overlap at one site; and (ii) calculating niche overlap over multiple sites. At a single site, each species has a single vector of electivity, one component per resource type, with values adding to one. This may be treated as categorical data, with niche overlap calculated as in Section 3.2. The result will be a single niche overlap of (say) Species i and j indicating whether these species select that resource (habitat, diet) in a similar or different way.

Alternatively, with a large number of sites (e.g. 10 or more) each species has a matrix of electivities, with (say) rows being sites and columns being resource types. This gives rise to a vector of electivity (Manly's alpha) scores for each resource type (one for each site). Each resource type therefore contributes to a dimension to overall niche overlap. For each pair of species and each resource type (e.g. Species i and j selecting a forest habitat), the electivity scores (alpha values) over different sites may be used to calculate a niche overlap between the two species.

Electivity scores are similar to fraction data (between 0 and 1), but they are not strictly continuous data because there may be clusters of observations at zero or at one. If axis t is the usage of a certain resource, for example fish species i selecting a substrate of sand, if all fish of species i are over sand the electivity (Manly's alpha) is one, and if no fish of species i are over sand, the electivity is zero. However, between zero and one, with partial usage of the sand substrate, the electivity measure is essentially continuous.

This implies there is a composite or mixed statistical distribution, neither fully continuous nor fully discrete.

Such a distribution is not easy to represent as a density. If a continuous curve is used for the probability density function on $0 < x < 1$, there should be infinite spikes at zero and one. Similarly, the bar charts used in discrete distributions cannot accommodate continuous density functions. We note that representation as the cumulative distribution function is possible, $y = F(x) = \text{Prob}(X \leq x)$. With discrete distributions, this rises from 0 to 1 in a step function, while a continuous distribution gives a continuously rising curve. A distribution with a mixture of discrete and continuous probability simply includes both steps and sections of continuous curve. However, it is not possible to display the niche overlap on the cumulative probability graph.

Instead, we illustrate such distributions in a **triptych graph**, which has left and right panels to display the probabilities of zero and one, and a central panel for the continuous probability. The density estimation for the central panel uses logit-transformed data, to avoid an overflow outside (0,1). By giving all panels width one, the niche overlap is again represented by a cross-hatched area.

Example

Suppose we have the following Manly alpha values (electivity scores) for selecting a certain habitat, e.g. forest, over 20 sites. At each site, a number of individuals of Species i or j were seen. The alpha estimate is the proportion of individuals at that site choosing that particular substrate, after allowing for the availability of that habitat type at that site.

Species i:	0	0.15	0.30	0.35	0.45	0.50	0.53	0.56	0.60	0.62
	0.60	0.70	0.73	0.75	0.79	0.81	0.86	0.92	1	1
Species j:	0	0	0	0.05	0.10	0.15	0.21	0.23	0.25	0.35
	0.47	0.53	0.58	0.60	0.64	0.68	0.73	0.74	0.79	1

For Species i the estimated probability of a zero is 1/20 (the proportion of zeros), and similarly the probability of a one is estimated by 2/20. The remaining 17 observations have density estimation as for continuous data, using logit-transformed data to avoid overflow beyond 0 or 1, but the probabilities from the density estimation (on logit data) are downscaled by a factor of 17/20 to reduce the area under the density curve to 17/20. A similar procedure is used for Species j.

Figure 8 shows the results of this probability distribution fitting, with each distribution being a mixture of two discrete and one continuous distribution. As in the earlier graphs, Species i is hatched in red, and Species 2 in blue. Both blue and red hatchings have area one, and the cross-hatched overlap is the niche overlap on this axis (i.e. for this resource). In this example, $NO = 0.70$.

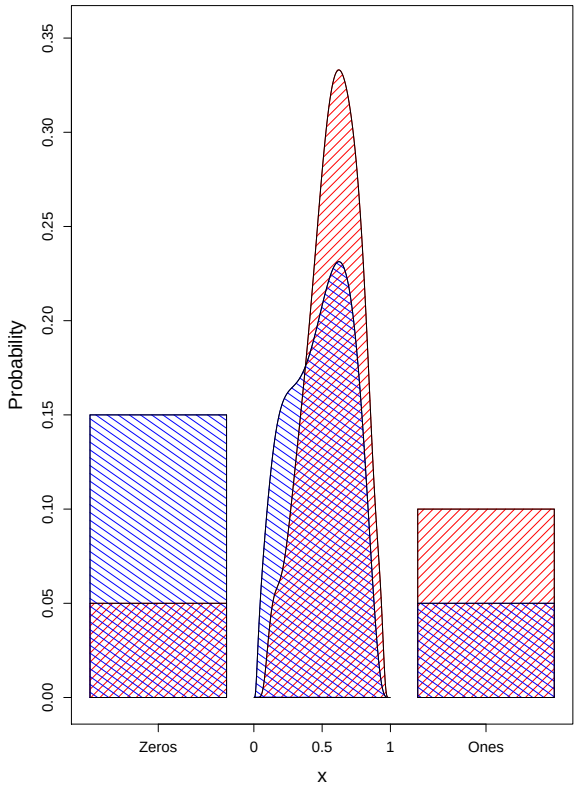

Figure 8: A triptych graph illustrating niche overlap from electivity data. The left panel shows the estimated probability of zero, and the right shows the estimated probability of one. The central panel is a probability density estimate, using a logit transformation on the data between 0 and 1. The red-hatched and blue-hatched areas (Species i and j respectively) each total to 1, and the niche overlap is the total cross-hatched area, 0.70.

4 Multivariate Graphical Representation

Using the overall niche overlap from averaging over all axes (equation 1), it is clear that some pairs of species overlap more than others.

Using an appropriate distance measure between pairs of species, multidimensional scaling can be used to graphically represent the amount of overlap between pairs of species.

4.1 Distance Measures

The niche overlap measures NO_{ij}, ranging from zero to one, may be seen as measures of association between pairs of species. Hence $1 - NO_{ij}$ may be seen as a distance between species i and j.

This produces a (symmetric) distance matrix D which is $n \times n$, where n is the number of species in the study. An exact graphical representation of these distances would require $n-1$ dimensions (e.g. representing three species as a triangle in two dimensions). However, an approximate representation may be possible in fewer dimensions, and can be achieved using multidimensional scaling.

4.2 Multidimensional Scaling

Since all pairs of species now have a distance measurement between them, the distance matrix can be used in multidimensional scaling (MDS), which reduces the dimension of the data below $n - 1$.

The reduction of the distance matrix to two dimensions, if justified by the stress levels in the scaling, provides a 2-D graphical representation of species associations, illustrating which pairs of species have similar niches, and those which have dissimilar niches.

5 Statistical Testing

When studying niche overlap, a major objective is to detect whether two species occupy the same or different niches. A useful extension is to test whether a group of species occupy the same niche, or at least one is different.

5.1 Sampling Variation

The comparison of niches between two species must be done statistically, to rule out "detecting" as different two niches which really only differ from sampling variation. On each axis, it is expected that even if the niches are identical, there will be some differences in the data purely by chance. The question is whether the two species may be described by the same probability distribution, or whether there is evidence of some difference.

On any one axis t, species i has (say) n_i readings, $\{x_{i1}, x_{i2}, \ldots x_{in_i}\}$, which are realisations of the random variable X_{it}. Similarly, species j has n_j readings, $\{x_{j1}, x_{j2}, \ldots x_{jn_j}\}$ from the random variable X_{jt}. The observed readings are used to estimate the parameters and hence the probability structure of each distribution.

However, if the random variables X_{it} and X_{jt} are the same, sampling variation will almost certainly cause the sampled data values to differ, giving $NO_{ijt} < 1$.

The question of interest is whether NO_{ijt} is sufficiently far below one to give evidence of different niches (different random variables).

The same argument applies to the combined NO measure, averaged over all the axes. Even with identical niches, sampling variation will probably ensure the observed NO_{ij} is below one.

The usual assumptions for statistical analysis are unlikely to be met. For example, count data may be overdispersed, with higher variance than would be expected for a binomial distribution, and electivity scores have unknown distributions.

For these reasons, we recommend analysis using null models and their associated permutation tests.

5.2 Null Models

We recommend the construction of null models (Gotelli and Graves 1996, Gotelli 2000) to compare niche overlap between species. Null models use randomisation or permutation tests, which do not rely on distributional assumptions (Manly 2007).

Instead, the null distribution (the distribution of the test statistic under the null hypothesis) is generated by calculating pseudo-values of the test statistic which would arise if H_0 is true. The position of the value of the data-based test statistic (NO) in relation to the pseudo-values provides the p-value for the test. With sufficient permutations, the p-value may be generated to a chosen number of decimal places.

We may compare niche overlaps for two or more species at the same site, or for the same species over two or more sites.

5.2.1 Comparing two species at the same site

To compare Species i and j, with a p-value to 3 decimal places, the procedure is as follows:

1. Select out the subset of the data pertaining to the two species of interest.
2. Using these data, for each axis t calculate and store the vector of NO_{ijt} values (for $t = 1, 2, \ldots T$). Also store the average value, NO_{ij}.
3. Do the following 999 times:
   (a) Select the data set for these two species only.
   (b) Randomly permute the species labels in the data set for species i and j. This permutation is done at the level of the individual animal or plant.
   (c) For each axis $t = 1, 2, \ldots T$, calculate and store a pseudo-value for NO_{ijt}, using the pseudo-data with the permuted labels.
   (d) Also store the average pseudo-value NO_{ij}.
4. For each axis t, sort the 999 pseudo-values of NO_{ijt}. A histogram shows the generated null distribution of NO_{ijt}.
5. Find where the data-based value of NO_{ij} lies on this distribution. If the niches differ, this value should be lower than might be found by chance, i.e. at the low end of the null distribution.
6. The number of pseudo-values below the data-based value is divided by 1000 to give the p value of the test of
   H_0: Species i and j have the same niche, versus
   H_A: the niches differ.
   The p-value is the proportion of the null distribution in the lower tail.

If the procedure above is run again, the p-value will probably differ, because the random permutations will be different this time; however, if over several runs the p-values look fairly stable, the procedure is working as desired. If very different p-values are obtained, the procedure needs longer runs (e.g. 9999) for stability.

5.2.2 Comparing two species on one axis

If an overall niche difference is found between i and j, details of which axes contribute to the difference may be obtained. For each axis in turn, compare the previously-stored data-based NO_{ijt} with the 999 pseudo-values of NO_{ijt}, and find the p-value for axis t. However, doing T tests brings the problem of multiple comparisons, so an adjustment to the p-values is necessary to protect against false positives ("detecting" a difference which is not really there). We suggest a sequential Bonferroni adjustment (see, e.g. Quinn and Keough 2002).

5.2.3 Comparing multiple species at the same site

An overall test (analogous to oneway ANOVA) is also possible. For a set of three or more species, to test

H_0: All species occupy the same niche, versus

H_A: At least one species occupies a different niche,

we use the same procedure as above, for the comparison of two species. The data set is restricted to the set of species being tested, and the procedure outlined above is followed. The permutation of species labels, again at the level of the individual, is now over more than two species. For the test statistic now, we may use the average niche overlap, averaged over all the distinct pairs of species in the set. If the set of species in the test is A, the test statistic is now

$$\overline{NO} = \text{mean}_{ij \in A,\ i<j}\ NO_{ij}.$$

The rationale for this test is that if all the species in question occupy the same niche, the actual labelling of each species is irrelevant. Hence the null distribution of $\overline{NO}$ is generated by permuting the labels to calculate pseudo-values of $\overline{NO}$. The proportion of pseudo-values less than the data-based value is the p-value for the test.

5.2.4 Comparing one species over two sites

To compare niches at two sites for a single species, any resource selection variables must take account of different availabilities at the two sites, i and j. A resource which is available at only one site cannot be included in the data set, as its unavailability at the other site means we have missing information on whether or not the individuals would have used that resource, if available. Any resource selection variable should have its electivities (alpha values) spread over only those resources available at both sites.

Under a null hypothesis that the chosen species occupies the same niches at both sites, there will be a permutation of site labels. However, since the availability of resources is intrinsically attached to each site, when site labels are permuted, each availability vector must be kept with its correct site. An equivalent approach would be to reallocate the individuals to the sites (with the correct number on each site), and in this case any measurements intrinsic to the individuals (e.g. length, sex) should be carried along with the individual to its new site.

The algorithm follows:

1. Select out the subset of the data, with a restriction to the chosen species and the two sites under consideration.
2. Imposed a further restriction: for each resource selection variable, remove the data for any resource which is not available at both sites. For example, in a study of lizards, if habitats forest, grass, rock and sand are available at Site i, but sand is missing from Site j, remove any data from lizards on sand at Site i. The comparisons can only use the completely available resources, forest, grass and rock.
3. Using these data, for each axis t calculate and store the vector of NO_{ijt} values (for $t = 1, 2, \ldots T$). Also store the average value, NO_{ij}.
4. Do the following 999 times:
   (a) Select the data set for these two species only.
   (b) Randomly permute the site labels in the data set for sites i and j. This permutation is done at the level of the individual animal or plant.
   (c) Use the same permutation to keep all availability vectors of resources matched with their correct sites.
   (d) For each axis $t = 1, 2, \ldots T$, calculate and store a pseudo-value for NO_{ijt}, using the pseudo-data with the permuted site labels and availabilities.
   (e) Also store the average pseudo-value NO_{ij}.
5. For each axis t, sort the 999 pseudo-values of NO_{ijt}. A histogram shows the generated null distribution of NO_{ijt}.
6. Find where the data-based value of NO_{ij} lies on this distribution. If the niches differ, this value should be lower than might be found by chance, i.e. at the low end of the null distribution.
7. The number of pseudo-values below the data-based value is divided by 1000 to give the p-value of the test of
   H_0: this species occupies the same niche at sites i and j, versus
   H_A: the niches differ between the sites.
   The p-value is the proportion of the null distribution in the lower tail.

5.2.5 Comparing one species at two sites using one axis

If an overall niche difference is found between sites i and j, details of which axes contribute to the difference may be obtained. For each axis in turn, compare the previously-stored data-based NO_{ijt} with the 999 pseudo-values of NO_{ijt}, and find the p value for axis t. However, doing T tests brings the problem of multiple comparisons, so an adjustment to the p values is necessary to protect against false positives ("detecting" a difference which is not really there). We suggest a sequential Bonferroni adjustment (see, e.g. Quinn and Keough 2002).

5.2.6 Comparing one species at multiple sites

An overall test (analogous to oneway ANOVA) is also possible. For a set of three or more species, to test

H_0: This species occupies the same niche at all sites, versus

H_A: It occupies a different niche at at least one site,

we use the same procedure as above, for the comparison of two sites. The data set is restricted to the set of species being tested, with a further restriction to resources which are available at all the sites, and the procedure outlined above is followed. The permutation of site labels, again at the level of the individual, is now over more than two sites. For the test statistic now, we may use the average niche overlap, averaged over all the distinct pairs of sites in the set. If the set of sites in the test is A, the test statistic is now

$$\overline{NO} = \text{mean}_{ij \in A,\ i<j}\ NO_{ij}.$$

The rationale for this test is that if the chosen species occupies the same niche over all sites, the actual labelling of each site is irrelevant. Hence the null distribution of $\overline{NO}$ is generated by permuting the labels to calculate pseudo-values of $\overline{NO}$. The proportion of pseudo-values less than the data-based value is the p-value for the test.

6 Examples

Worked examples and R code are available from the first author. There are two examples using artificial data, one with a range of variables at one site, and the other with fewer variables but over several sites. Between them, the examples illustrate most of the techniques suggested in this report.

7 Discussion

This report has a combination of existing and new methods, blended into a single unified approach to measuring and testing niche overlap. A full discussion will be in a research article (Geange et al., in preparation).

References

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. em Second International Symposium on Information Theory, Eds. Petrov, B. N. and Csaki, F. Academiai Kiado, pages 267-281.

Chesson, J. (1978). Measuring preference in selective predation. *Ecology* **59:** 211–215.

Chesson, J. (1983). The estimation and analysis of preference and its relationship to foraging models. *Ecology* **64:** 1297–1304.

Colwell, R.K. and Futuyma, D.J. (1971). On the measurement of niche breadth and overlap. *Ecology* **52:** 567-576

Gotelli, N.J. (2000). Null model analysis of species co-occurrence patterns. *Ecology* **81:** 2606–2621.

Gotelli, N.J. and Graves, G.R. (1996). *Null Models in Ecology.* Washington D.C.: Smithsonian Institute Press.

Hutchinson G. (1957). Concluding Remarks. *Cold Spring Harbor Symposia on Quantitative Biology* **22:** 415-427.

Lindsay, B.G. (1983). The geometry of mixture likelihoods: A general theory. *The Annals of Statistics* **11:** 86-94.

Ludwig, J.A. and Reynolds, J.F. (1988). *Statistical Ecology: A Primer on Methods and Computing.* New York: John Wiley and Sons.

Manly, B.F.J. (1974). A model for certain types of selection experiments. *Biometrics* **30:** 281–294.

Manly, B.F.J. (2007). *Randomization, Bootstrap and Monte Carlo Methods in Biology.* London: Chapman and Hall. 3ed. 455pp.

Mouillot, D., Stubbs, W., Faure, M., Dumay, O., Tomasini, J.A., Wilson, J. B. and Chi, T. D. (2005). Niche overlap estimates based on quantitative functional traits: a new family of non-parametric indices. *Oecologia* **145:** 345–353.

Pianka ER (1972). r and K selection or b and d selection. *The American Naturalist* **106:** 581-588

Quinn, G.P. and Keough, M.J. (2002). *Experimental design and data analysis for biologists.* Cambridge, UK ; New York : Cambridge University Press.

R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL: http://www.R-project.org.

Schmid, F. and Schmidt, A. (2006). Nonparametric estimation of the coefficient of overlapping – theory and empirical application. *Computational Statistics and Data Analysis* **50:** 1583–1596.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

Stubbs, W.J. and Wilson, J.B. (2004). Evidence for limiting similarity in a sand dune community. *Journal of Ecology* **92:** 557–567.