SPECIES RESPONSE CURVES

Steven M. Holland

Department of Geology, University of Georgia, Athens, GA 30602-2501

June 2014

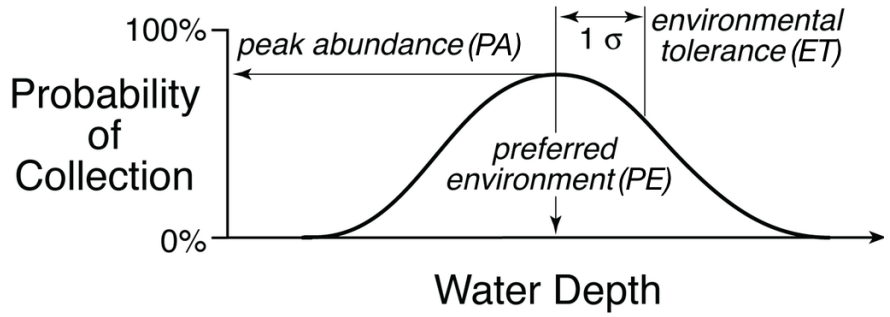
Introduction

Species live on environmental gradients, and we often would like to describe their occurrence along those gradients. The distribution of a species along a gradient is known as a species response curve. In modern environments, many species have unimodal response curves, and this curve is often symmetrical. Such a response curve can easily be modeled as a normal distribution, defined with three parameters.

The **optimum** measures where a species is most likely to be found, that is, the peak of a distribution. For any distribution, the optimum is equivalent to the mode of the distribution. For a symmetrical distribution, the optimum is also equivalent to the mean of the distribution. In our work, we have often called the optimum the **preferred environment (PE)** or, where we are dealing specifically with a water depth gradient, the **preferred depth (PD)**.

The **tolerance** measures the ability of a species to live in non-optimal environments, that is, it describes the spread or width of a distribution. For a symmetrical distribution, this is equivalent to the standard deviation of the distribution. We have often called tolerance by the terms **environmental tolerance (ET)** or **depth tolerance (DT)**.

The **maximum** measures how abundant a species is at its optimum, that is, the height of the response curve at the optimum. The maximum can be described in terms of abundance, but it is more commonly described as a probability, such as the probability of encountering the species at the optimum. We have often used the term **peak abundance (PA)** for the maximum.



Although many species response curves can be described by a normal distribution with these three parameters, some species have different distributions. The most common alternative is an asymmetrical distribution, in which tail on one side of the optimum is much longer than the other. Asymmetrical distributions often arise when a species' optimum lies close to a fixed boundary, such as for a marine species whose optimum is in very shallow water, or a terrestrial species whose optimum is near sea level. Bimodal or polymodal distributions also occur, but they are much rarer than symmetrical and asymmetrical unimodal distributions. We will focus only on symmetrical unimodal response curves.

Computation

Weighted averaging and logistic regression are the two main approaches to estimating optimum, tolerance, and maximum. Both methods start with a standard species-abundance matrix, with species in columns, and samples in rows. Both methods also need the position along an environmental gradient for each sample. The gradient may be **direct**, such as water depth, salinity, or temperature, or the gradient may be **indirect**, such as the sample scores from an ordination method, such as detrended correspondence analysis or non-metric multidimensional scaling.

Weighted averaging

The weighted averaging method gets its name from the calculation of the optimum: it is the abundance-weighted average of the gradient position of every sample bearing the species. By weighting this average by abundance, a sample containing many individuals of a species (and therefore, likely close to the optimum) counts proportionally more than a sample containing only a single occurrence of the species (which is likely far from the optimum).

In ordination methods such as detrended correspondence analysis and non-metric multidimensional scaling, species scores are calculated as abundance-weighted averages. If response curves are being calculated from these ordination methods, no additional calculations need to be made to obtain the optima - they are simply the species scores.

Tolerance is calculated by calculating the standard deviation of the gradient positions of all samples that contain the taxon. For ordination-based methods, tolerance is the standard deviation of the sample scores of samples bearing the taxon. If a species occurs over a limited set of gradient positions, this standard deviation will be small, and if a species occurs over a broad range of gradient positions, the standard deviation will reflect a large tolerance.

Maximum is estimated by first calculating the percent occurrence of a species in all samples that lie within one tolerance (i.e., standard deviation) of the optimum. This is multiplied by a a constant (approximately 1.169) that reflects the ratio of the peak height of a normal distribution to the average height of that distribution over the interval of the mean plus or minus one standard deviation.

Logistic regression

Logistic regression is used when the dependent variable (in this case, the probability of occurrence) is constrained to lie between 0 and 1. A Gaussian function can be fitted to the occurrence data using this function:

$$logit(p) = b_0 + b_1 x + b_2 x^2$$

The right-hand side is the equation for a parabola, where x is the position along a gradient, and b_0, b_1, and b_2 are shape coefficients. The left-hand side, known as the transfer function, converts this parabola to a Gaussian distribution. The function logit(p) is the log-odds of the probability of collection, defined as

$$logit(p) = log\left(\frac{p}{1-p}\right)$$

where p is the probability of collection of a taxon.

From this logistic regression and its three parameters, b_0, b_1, and b_2, estimates of the optimum (O), tolerance (T), and maximum (M) can be calculated:

$$O = \frac{-b_1}{2b_2}$$

$$T = \frac{1}{\sqrt{-2b_2}}$$

$$M = e^{b_0 + b_1 O + b_2 O^2}$$

Species response curves in R

To perform these methods, you will need an abundance matrix and a vector of gradient positions.

The abundance matrix will be the same as you would use in an ordination. Each column should contain the abundance of a single species, with each row reflecting a single sample. Species names should be in the first row of the data set, and sample labels should be in the first column.

The vector of gradient positions should have the same length as the number of samples, and these gradient positions must be in the same order as the samples in the abundance matrix. If you use an ordination method to establish the gradients, the vector will usually be in the correct order.

For this tutorial, we will assume that you will want to use an ordination to establish the gradient. Although this will be illustrated here with detrended correspondence analysis (DCA), you

can follow a similar approach with non-metric multidimensional scaling or any other ordination.

First, read in the data. This tutorial uses the KentuckyCounts.txt data, which consists of Late Ordovician marine invertebrates from the Frankfort, Kentucky area (Holland and Patzkowsky 2004).

```
counts <- read.table(file='KentuckyCounts.txt', header=TRUE,
   row.names=1, sep=',')
```

Next, load the vegan library, which contains the useful data transformation function **decostand()** and the DCA function **decorana()**.

```
library(vegan)
```

Perform two data transformations to correct for differences in sample size and species abundance. First, standardize by row totals, which converts all the abundance counts to percentages, preventing sample size from dominating the ordination. Second, standardize by the maximum in each column, which expresses every value in a column as a fraction of the maximum value in that column, in effect giving all species the same weight in the ordination.

```
counts.t1 <- decostand(counts, method='total')
counts.t2 <- decostand(counts.t1, method='max')
```

Perform a detrended correspondence analysis on this doubly transformed data.

```
counts.dca <- decorana(counts.t2)
```

It is helpful to extract the sample scores and species scores from the decorana output. We will focus here only on axis 1, the most important source of variation in the data.

```
sampleScores1  <- scores(counts.dca, display="sites",
   choices=1)

speciesScores1 <- scores(counts.dca, display="species",
   choices=1)
```

At this point, you have everything you need: the abundance matrix (**counts**), a vector of gradient positions of samples (**sampleScores1**), and a vector of gradient positions of species (**speciesScores1**).

Weighted averaging

To illustrate the approach of weighted averaging, we will calculate the three parameters for one species, the common orthid brachiopod *Hebertella*, which is in column 13 of the abundance matrix and therefore row 13 of the species scores. Getting the optimum (preferred environment, PE) of a species is trivial: it is just the species score.

```
Hebertella <- 13
```

```
PE <- speciesScores1[Hebertella]
```

Next is the value of tolerance (ET). Get the sample scores all samples containing *Hebertella* (i.e., those with an abundance greater than zero), then calculate their standard deviation.

```
scoresWithHebertella <- sampleScores1[counts[ ,Hebertella]>0]

ET <- sd(scoresWithHebertella)
```

Last, calculate the maximum (PA). Find all of the samples that lie on the gradient within one tolerance of the optimum. Count the number of these samples. Find the abundances of *Hebertella* within those samples, and use this to count the number of those samples that bear *Hebertella*. Finally, calculate the percentage of samples lying near the optimum that contain *Hebertella*, and correct this percentage to give the maximum (that is, the percentage at the optimum itself).

```
samplesNearOptimum <- (abs(sampleScores1 - PE) <= ET)

total <- length(samplesNearOptimum[samplesNearOptimum==TRUE])

abundancesNearOptimum <- counts[samplesNearOptimum,
   Hebertella]

present <- length(abundancesNearOptimum[
   abundancesNearOptimum>0])

PA <- present/total * 100 * 1.168739
```

For convenience, all of this can be wrapped in a function that takes the abundance matrix, the sample score vector, and the species score vector as arguments. The function returns a data frame with columns for optimum (PE), tolerance (ET), and maximum (PA).

```
responseCurveWA <- function(abundance, sampleScores,
   speciesScores) {

   numSpecies <- length(speciesScores)

   PE <- speciesScores
   ET <- vector(length=numSpecies)
   PA <- vector(length=numSpecies)

   for (species in 1:numSpecies)
   {
      # get sample scores of all samples containing the taxon
      scoresWithSpecies <-
         sampleScores[abundance[ ,species]>0]

      # calculate ET
      ET[species] <- sd(scoresWithSpecies)
```

```
        # find all samples within 1 ET of PE
        samplesNearOptimum <-
           (abs(sampleScores - PE[species]) <= ET[species])

        # count the number of these samples
        total <- length(samplesNearOptimum
           [samplesNearOptimum==TRUE])

        # find abundances of taxon within those samples
        abundancesNearOptimum <-
           abundance[samplesNearOptimum, species]

        # extract those samples that actually contain the taxon
        present <- length(abundancesNearOptimum
           [abundancesNearOptimum>0])

        # calculate PA, with correction factor from average PA
        # to peak PA. Note that this correction factor will not
        # be accurate for large PA
        PA[species] <- present/total * 100 * 1.168739
     }

     speciesResponse <- data.frame(PE, ET, PA)
     rownames(speciesResponse) <- colnames(abundance)
     speciesResponse
}
```

Call the function as follows, and assign the results so that they can be viewed.

```
paramsWA <- responseCurveWA(counts, sampleScores1,
   speciesScores1)
```

Logistic regression

Logistic regression is based simply on whether a species occurs at any of the gradient positions. For this method, you will need a vector of gradient positions (**sampleScores1**), and you will need a vector of whether the species occurs at that gradient position. To illustrate the method, we will use the common brachiopod *Hebertella* in the Frankfort data (column 13) and convert its abundances to presence/absence. This is done by finding every abundance that is greater than zero and replacing it with a 1, such that all values are now 0 (absent) or 1 (present).

```
Hebertella <- 13
HebertellaPA <- counts[, Hebertella]
HebertellaPA[HebertellaPA > 0] <- 1
```

For clarity, we will call the sample score data x, as it is the independent variable, and we will call the presence/absence data y, as it is the dependent variable.

```
x <- sampleScores1
y <- HebertellaPA
```

The logistic regression is performed with **glm()**. The results need to be assigned to an object.

```
reg <- glm(y ~ x + I(x^2), family=binomial)
```

The **glm()** command will produce warnings if a regression cannot be fit to the data, which happens in some cases.

For clarity, we relabel the coefficients from the regression object to match the logistic equation given above.

```
b0 <- reg$coefficients[1]
b1 <- reg$coefficients[2]
b2 <- reg$coefficients[3]
```

The optimum (**O**), tolerance (**T**), and maximum (**M**) are calculated as shown. Note that M is a probability (0 to 1), rather than as a percentage, unlike in the weighted averaging method, which returns the maximum as a percentage.

```
O <- (-b1)/(2*b2)
T <- 1 / sqrt(-2*b2)
M <- 1 / (1 + exp(b1^2 / (4 * b2) - b0))
```

For convenience, this can be wrapped in a function to calculate the optimum, tolerance, and maximum for all species in a dataset. The first parameter is the matrix of species abundances, and the second is a vector of sample scores. In cases where values cannot be computed, two lines near the end of the function replace the missing values with **NaN** (not a number). If this replacement is necessary, you will see a warning message after issuing the command.

```
responseCurveLR <- function(abundance, sampleScores) {
   numTaxa <- ncol(abundance)

   O <- vector(mode="numeric", length=numTaxa)
   T <- vector(mode="numeric", length=numTaxa)
   M <- vector(mode="numeric", length=numTaxa)

   for (t in 1:numTaxa) {
         x <- sampleScores
         y <- abundance[ ,t]
         y[y>0] <- 1

         reg <- glm(y ~ x + I(x^2), family=binomial)

         b0 <- reg$coefficients[1]
         b1 <- reg$coefficients[2]
         b2 <- reg$coefficients[3]
```

```
            O[t] <- (-b1)/(2*b2)
            T[t] <- 1 / sqrt(-2*b2)
            M[t] <- 1 / (1 + exp(b1^2 / (4 * b2) - b0))
        }

        params <- data.frame(O, T, M)
        params$O[is.na(params$T)] <- NaN
        params$M[is.na(params$T)] <- NaN

        rownames(params) <- colnames(abundance)
        params
    }
```

When calling this function, assign the results to an object so that you can use them. Each row corresponds to a species, in the same order as the columns of your abundance matrix. The columns correspond to optimum (O), tolerance (T), and maximum (M).

```
paramsLR <- responseCurveLR(counts, sampleScores1)
```

References

Coudun, C., and J. C. Gegout, 2006. The derivation of species response curves with Gaussian logistic regression is sensitive to sampling intensity and curve characteristics. Ecological Modeling 199:164–175.

Holland, S.M., 1995. The stratigraphic distribution of fossils. Paleobiology 21: 92–109.

Holland, S. M., and M.E. Patzkowsky, 2004. Ecosystem structure and stability: Middle Upper Ordovician of central Kentucky, USA. Palaios 19:316–331.

Holland, S. M., and A. Zaffos, 2011. Niche conservatism along an onshore-offshore gradient. Paleobiology 37: 270–286.

Jongman, R.H.G., C.J.F. ter Braak, and O.F.R. Van Tongeren, 1995. Data Analysis in Community and Landscape Ecology. Cambridge University Press: Cambridge, 299 p.