OPTIMOS PRIME: An R package for autoecological (optima and tolerance range) data calculation

SATHICQ, María Belén¹; NICOLOSI GELIS María Mercedes¹; COCHERO, Joaquín¹

¹Instituto de Limnología “Dr. Raúl A. Ringuelet,” Facultad de Ciencias Naturales y Museo, Universidad Nacional de La Plata, La Plata, Argentina. CONICET, Consejo Nacional de Investigaciones Científicas y Tecnológicas. Boulevard 120 y 62 N° 1437 (1900). Buenos Aires, La Plata, Argentina

KEYWORDS: autoecology, ecological optima, environmental range, paleolimnology, tolerance ranges

RUNNING TITLE: An R package for autoecological data calculation

ACADEMY SECTION: BIOLOGICAL SCIENCES

CORRESPONDENCE AUTHOR: Dr. María Belén Sathicq. Email: mbelen @ilpla.edu.ar
ABSTRACT

1. Calculation of autoecological data, such as optima and tolerance ranges to environmental variables, can be useful to establish the distribution and abundance of the species. These calculations, although mathematically not complex, can be prone to error when using a large database.

2. We present an R package (“optimos.prime”) that uses species’ abundance data and environmental data to calculate the optimum value and tolerance range of each species to each environmental factor, by weighted average. Additionally, the package can create caterpillar plots to show the results.

3. Using sample data from a phytoplankton database, we exemplify the use of the R package and its functions. A stand-alone version for Windows is also provided, and source code and documents are freely available on GitHub to encourage collaborative work.

INTRODUCTION

When using organisms as indicators of environmental conditions, the key aspect to consider is the combination of environmental variables that are optimal for its existence, development, growth and reproduction (Verbitsky and Verbitskaya 2007). These aspects depend mostly on environmental factors, and the reproductive success of the species is related to an “ecological optimum” (Battarbee et al. 2010; ter Braak and Smilauer 1998). However, in natural ecosystems, the ecological optimum includes not only a single point value, but also the oscillations of the environmental variable around this value, and for each environmental variable there are lower and upper limits, below and above which the taxon is less likely to survive. These limits constitute the tolerance range of the taxon (Huggett 2004; Smith and Smith 2009; Cristóbal et al. 2014).

In this regard, the knowledge of autoecological data can be very helpful in the interpretation of the distribution, and presence/absence of different taxa in different environments (Clarke et al. 2006; Licursi et al. 2010). The most common approach used for its calculation is the weighted averaging,
using the number of occurrences to adjust the tolerance assigned to each taxon to estimate optima and tolerance range in cases where taxa have unequal occurrences (Birks et al. 1990).

When employed, for instance, in water research and limnology, these methods have facilitated the reconstruction of past lake conditions (Oksanen et al. 1988; Birks et al. 1990; Bradshaw et al. 2002; Miettinen et al. 2005; Holden et al. 2008), the accurate estimation of contemporary limnological conditions in lentic and lotic environments (e.g: Potapova and Charles 2003; Meador et al. 2008; Yang et al. 2008; Licursi et al. 2010; Sathicq 2017), and the establishment of environmental quality standards for discerning the trends, causes, and consequences of watershed-land use (Cooper 1995; Kattel 2012; Arva et al. 2017). Quantitative autecological characteristics derived from small-scale regional datasets are useful for regional monitoring programmes, but since they are dependent on the restricted range and distribution of the environmental parameters in the dataset, they may not be appropriate for areas with different water chemistry characteristics. Reliable autecological data can be obtained only from datasets with large numbers of observations representing the full range of environmental conditions (Potapova and Charles 2003), and their calculation can be prone to errors, particularly when there is a significant number of species.

OPTIMOS PRIME is an open-source R script to facilitate the calculation of the ecological optima and tolerance ranges for each species to a given set of environmental variables. It is based on the weighted averaging procedure mentioned above. In this article we present its basic functions, providing an example of its use with data from phytoplankton samples that is available through the GitHub repository of the project. OPTIMOS PRIME is developed in R since it is increasing popularity as an analytical tool in the ecological sciences (e.g: Dixon 2003; Colchero et al. 2012; Metcalf et al. 2012; Revell 2012), although a stand-alone version for Windows is also available in the GitHub repository (https://github.com/limnolab/Optimos-Prime). Users can freely download a stable version of the package from the CRAN website (https://cran.r-project.org/web/packages/optimos.prime). Also, OPTIMOS PRIME development is being managed via GitHub to encourage collaborative development.

DATA REQUIREMENTS AND INPUT
The calculation of species’ optima and tolerance range to environmental variables requires two input data frames (in the R terminology). The first data frame (“Environmental data”) contains all the values for all environmental parameters measured in the samples or sampling sites. The second data frame (“Species data”) contains the relative abundance of each species in each of the samples or sampling sites. Sample data frames are included in the package (/data subfolder).

**Environmental data:** This data frame contains the environmental parameters measured (rows) by the sampled sites (columns). The first row has to contain the names of the sites, and the first column has to contain the names of the environmental parameters, or the code that the user will use as the default labels for the plots. A sample data frame is included in the package. Missing data should be left blank in the data frame.

**Species data:** This data frame contains the relative abundance of each species (rows) in each sampled site (columns). The first row, as in the environmental data frame, has to contain the names of the sites, and the first column has to contain the names of the species, or the code that the user will use as the default labels for the plots.

These data frames can be already loaded in R or imported into R from a comma separated value file (CSV) created in a spreadsheet software (e.g. Microsoft Excel). If the data frames are not specified when the function is run, the user will be prompted to select a CSV file.

**CALCULATION OVERVIEW**

There are three functions in the OPTIMOS PRIME package (Table 1). The function `op_calculate()` will use the environmental and species’ data frames to calculate the optima and tolerance range for each species to each environmental parameter. During the calculation, the environmental data needs to be transformed into log10; the `op_calculate` function assumes that the data is raw (not transformed), but if it is already transformed the parameter “islog10” (FALSE by default) of the function has to be set to TRUE.

The species’ data needs to be the relative abundance of each species in each sample. If this is not the case, and the data frame being used is the raw density of the species, then the “isRelAb”
parameter needs to be changed to FALSE, and the function will convert it automatically to relative abundance values.

The weighted average estimates of the species optima ($u_k$) is calculated as:

$$u_k = \frac{\sum_{i=1}^{n} y_{ik} x_i}{\sum_{i=1}^{n} y_{ik}}$$

Where $y_{ik}$ is the relative abundance of species $k$ in sample $i$; $x_i$ is the log10 value of the environmental parameter in sample $i$; and $n$ is the total number of samples in the dataset.

Tolerance, or weighted standard deviation ($t_k$), is calculated as:

$$t_k = \sqrt{\frac{\sum_{i=1}^{n} y_{ik} (x_i - u_k)^2}{\sum_{i=1}^{n} y_{ik}}}$$

The op_lists() function allows the user to extract from the environmental and species data frames a list of the names of the sampling sites (Site list), of the names of the environmental variables (Environmental list), of the names of the species’ (Species list) or a single list including all the previous three lists. This can be used to check if the original data frames were constructed according to the specifications above, or to use the lists for further tests in R.

| Function   | Description                                      | Parameters                      |
|------------|--------------------------------------------------|---------------------------------|
| op_calculate | Uses the environmental data and species’ data to calculate the ecological optima and tolerance ranges. | environtmental_df: The data frame with your environmental data. species_df: The data frame with your species densities. |
| **op_lists** | Generates three lists with names from your data frames: a species list, an environmental factors list and a sample (or sampling sites) list |
| **isRelAb** | Specifies if the species’ data is the relative abundance of each species in each sample. Default=TRUE. |
| **Islog10** | Specifies if the environmental data has already been transformed to log10. Default=FALSE. |
| **op_plot** | Generates caterpillar plots for a specified environmental variable showing optima and tolerance ranges as calculated with the op_calculate() function |
| **environmental_df** | The data frame with your environmental data. |
| **species_df** | The data frame with your species densities. |
| **listOnly** | Specifies which list/s to return. If = 1, it returns only the lists of Sites. If = 2, it returns only the list of Species. If = 3, it returns only the list of Environmental parameters. If = 0 (default), then returns all three lists combined (Sites, Species, Environmental). |
| **optimaDF** | The data frame resulting from the op_calculate() method |
| **label** | (optional) The label for the X axis |
PLOTTING RESULTS

The op_plot() function will provide with caterpillar plots for all species showing their optima and tolerance ranges to a certain chosen variable. This function uses the result data frame obtained from the op_calculate() function as input, provides the user with a choice to select which environmental variable to plot and draws an interactive caterpillar plot (Figure 1). This function depends on three other libraries that need to be installed along with the OPTIMOS PRIME package: ggplot2, tidyverse, and plotly. These three libraries should be automatically installed when Optimos Prime is loaded into R; otherwise they can be obtained from the CRAN repository.

![Fig 1. Caterpillar plot resulting from the op_plot() function with the example data, using “electrical conductivity” as an environmental factor.](image)

EXAMPLE APPLICATION

The following example uses two CSV files that are openly available in the GitHub repository of the project (https://github.com/limnolab/Optimos-Prime/tree/master/sample_data), although both data frames are also included in the package (/data subfolder). The file “environmental_data.csv” contains data from 11 environmental parameters (such as temperature, electrical conductivity, dissolved oxygen, etc.) in 50 samples. The file “species_data.csv” contains the relative abundance...
of 57 species of algal plankton in the same 50 samples. The data corresponds to a subset of the phytoplankton data used in Sathicq (2017).

For this example, we will assume that the user wants to load the data directly from the CSV file, and not from an already existing data frame in R. The following code will load the library and calculate the autoecological data using the input data frames, placing the results into a new data frame called “results_op”. All figures shown for this example are displayed if the user is using RStudio, although similar results will be obtained if using the R core.

```r
> install.packages("optimos.prime")
> results_op <- op_calculate()

"Select CSV matrices"
[1] "Select ENVIRONMENTAL matrix"
[2] "Select SPECIES matrix"
"Calculating…"
"Optimum values and tolerance range calculated and placed in the final data frame"
"Use View() to view data frame with results"
```

Notice that no parameters were passed to the op_calculate() function, which resulted in the user being prompted to choose the CSV files from a dialog box in steps [1] and [2]. Also, the parameter islog10 was not set; its default value is FALSE, which assumes that the data is raw and was not previously transformed to log10.

The results of the analysis were placed in the data frame “results_op”, which can be viewed using View (results_op), as shown in Figure 2. Each row is a species, and each column consists of the ecological optima for each environmental parameter and its range of tolerance (columns HIGH and LOW for each parameter). If some species was absent from all sampling sites where the environmental variables were measured, the results show “NaN”, indicating missing data.
Fig. 2 Sample of the resulting data frame from the optima and tolerance ranges calculation.

The caterpillar plots for (for example) the environmental parameter “Conductivity” can be now obtained by using the `op_plot()` function, specifying the recently obtained data frame.

```r
> op_plot(results_op)
```

1 Conductivity (uS/cm)
2 pH
3 Temperature (°C)
4 Dissolved Oxygen (mg/L)
5 BOD5 (mg/L)
6 COD (mg/L)
7 N-NO3 (mg/L)
8 N-NO2 (mg/L)
9 Turbidity (NTU)
10 N-NH3 (mg/L)
11 P-PO4 (mg/L)

> [Question] What variable number do you want to plot? 1

"Plotting…"

"Plot complete!"

After the function is run, a list of the environmental variables is shown in the terminal, and the user is prompted to type the number of the variable that is to be plotted (1 in this case is Conductivity). The interactive plot resulting is shown in Figure 1; green points mark each optimum
and red points mark both high and low ends of the tolerance range for each species to that variable.

**WINDOWS STAND-ALONE VERSION**

A stand-alone version for Windows is also available from the GitHub repository. The compressed file (.zip) includes an installer, and the software leads the user into loading the data frames in order. To begin the process, the user has to establish how many species, samples and environmental parameters is going to use; in the following tabs, the data frames can be pasted directly from the clipboard in its place (Figure 3). The output of the Windows version is similar to the results data frame from the R package described above, although it does not include a plotting function to create the caterpillar plots. The output can be exported back to CSV format, and the software is available both in English and Spanish.

![Fig 3. Main form of the OPTIMOS PRIME stand-alone version for Windows.](image)

**CONCLUSIONS**

We introduced the open-source R package OPTIMOS PRIME, which is designed to facilitate the calculation of the ecological optima and tolerance ranges for each species to a given set of environmental variables. The OPTIMOS PRIME package contains function for calculate optima and
tolerance ranges, produce caterpillar plots, and generate lists with names from the data frames. In the GitHub repository of the package we provide the source code, along with two CSV files (“environmental_data.csv”, “species_data.csv”) with sample data, which users can download to test the package and build their data frames. A stand-alone Windows version is also available from the GitHub repository for non-R users. Future versions of OPTIMOS PRIME will include other algorithms for optima and tolerance range calculations, besides weighted averaging. Source code is publicly available through GitHub to encourage collaborative work.

AUTHORS’ CONTRIBUTIONS

All authors contributed to software planning and manuscript writing. M.B.S. and M.M.N.G. provided sample data and software testing. J.C. led software development and coding.

DATA ACCESSIBILITY

The OPTIMOS PRIME package can be downloaded from the Comprehensive R Archival Network (https://cran.r-project.org/web/packages/optimos.prime) or GitHub (https://github.com/limnolab/Optimos-Prime). A stand-alone Windows version can be downloaded from the repository as well. OPTIMOS PRIME is open source under GNU Public License.
REFERENCES

ÁRVA D., TÓTH M., MOZSÁR A., AND SPECZIÁR A. 2017. The roles of environment, site position, and seasonality in taxonomic and functional organization of chironomid assemblages in a heterogeneous wetland, Kis-Balaton (Hungary). Hydrobiologia, 787(1), 353-373.

BATTARBE R. W., CHARLES D. F., BIGLER C., CUMMING B. F., AND RENBERG I. 2010. Diatoms as indicators of surface-water acidity. The Diatoms: Applications for the Environmental and Earth Sciences, (Ed.2), 98-121. Retrieved from https://www.cabdirect.org/cabdirect/abstract/20103313511

BIRKS H. J. B., LINE J. M., JUGGINS S., STEVENSON A. C., AND TER BRAAK C. J. F. 1990. Diatoms and pH reconstruction. Philos Trans R Soc B 327:263-278

BRADSHAW E. G., ANDERSON N. J., JENSEN J. P., AND JEPPESEN E. 2002. Phosphorus dynamics in Danish lakes and the implications for diatom ecology and palaeoecology. Freshw Biol 47:1963-1975

CLARKE K. R., SOMERFIELD P. J., AND CHAPMAN M. G. 2006. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. Journal of Experimental Marine Biology and Ecology, 330(1), 55-80.

COLCHERO F., JONES O. R., AND REBKE M. 2012. BaSTA: an R package for Bayesian estimation of age-specific survival from incomplete mark-recapture/recovery data with covariates. Methods in Ecology and Evolution, 3(3), 466-470.

COOPER S. R. 1995. Chesapeake Bay watershed historical land use: impact on water quality and diatom communities. Ecol Appl 5:703-723

CRISTOBAL E., AYUSO S. V., JUSTEL A., AND TORO M. 2014. Robust optima and tolerance ranges of biological indicators: a new method to identify sentinels of global warming. Ecological research, 29(1), 55-68.

DIXON, P. 2003. VEGAN, a package of R functions for community ecology. Journal of Vegetation Science, 14(6), 927-930. doi:10.1111/j.1654-1103.2003.tb02228.x

HOLDEN P. B., MACKAY A. W., AND SIMPSON G. L. 2008. A Bayesian palaeoenvironmental transfer function model for acidified lakes. J Paleolimnol 39:551-566
HUGGETT R. J. 2004 Fundamentals of biogeography. Routledge, London

KATTEL G. R. 2012. Can we improve management practice of floodplain lakes using cladoceran zooplankton? River Research and Applications, 28(8), 1113-1120.

LICURSI M., GÓMEZ N., AND DONADELLI J. 2010. Ecological optima and tolerances of coastal benthic diatoms in the freshwater-mixohaline zone of the Río de la Plata estuary, 418, 105-117. doi:10.3354/meps08865

MEADOR M. R., CARLISLE D. M., AND COLES J. F. 2008. Use of tolerance values to diagnose water-quality stressors to aquatic biota in New England streams. Ecol Indic 8:718-728

METCALF C. J. E., MCMAHON S. M., SALGUERO-GOMEZ R. AND JONGEJANS E. 2012 IPMpack: an R package for integral projection models. Methods in Ecology and Evolution, 4, 195-200.

MIETTINEN J. O., SIMOLA H., GRONLUND E., LAHTINEN J., AND NIINIOJA R. 2005 Limnological effects of growth and cessation of agricultural land use in Ladoga Karelia: sedimentary pollen and diatom analyses. J Paleolimnol 34:229-243

OKSANEN J., LÄÄRÄ E., HUTTUNEN P., AND MERILÄINEN J. 1988. Estimation of pH optima and tolerances of diatoms in lake sediments by the methods of weighted averaging, least squares and maximum likelihood, and their use for the prediction of lake acidity. Journal of Paleolimnology, 1(1), 39-49. doi:10.1007/BF00202192

POTAPOVA M., AND CHARLES D. F. 2003. Distribution of benthic diatoms in U.S. rivers in relation to conductivity and ionic composition. Freshwater Biology, 48(8), 1311-1328. doi:10.1046/j.1365-2427.2003.01080.x

REVELL L. J. 2012. phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution, 3(2), 217-223. doi:10.1111/j.2041-210X.2011.00169.X@10.1111/(ISSN)2041-210X.TOPMETHODS

SATHICQ M. B. 2017. Empleo de descriptores fitoplanctónicos como biomonitores en la evaluación de la calidad del agua en la costa del río de la Plata (Franja Costera Sur). Retrieved from http://sedici.unlp.edu.ar/handle/10915/58915

SATHICQ M. B., GÓMEZ N., BAUER D. E., AND DONADELLI J. 2017. Use of phytoplankton assemblages to assess the quality of coastal waters of a transitional ecosystem: Río de la Plata estuary. Continental Shelf Research, 150, 10-17. doi:10.1016/j.csr.2016.08.009

SMITH T. M., AND SMITH R. L. 2009 Elements of ecology. Cummings, San Francisco
Ter Braak C. J. F., and Smilauer P. 1998. CANOCO Reference Manual and User’s Guide to Canoco for Windows: Software for Canonical Community Ordination (Version 4). Retrieved from http://agris.fao.org/agris-search/search.do?recordID=NL2012032100

VERBITSKY V. V. B., AND VERBITSKAYA T. T. I. 2007. Ecological optimum of ectothermic organisms: static-dynamical approach. *Nature Precedings*. doi:10.1038/NPRE.2007.1324.1

YANG X., ANDERSON N. J., DONG X., AND SHEN J. 2008. Surface sediment diatom assemblages and epilimnetic total phosphorus in large, shallow lakes of the Yangtze floodplain: their relationships and implications for assessing long-term eutrophication. *Freshw Biol* 53:1273-1290