Journal of Statistical Software

September 2007, Volume 22, Issue 3. http://www.jstatsoft.org/

Maximum Likelihood Method for Predicting Environmental Conditions from Assemblage Composition: The R Package bio.infer

Lester L. Yuan
US Environmental Protection Agency

Abstract

This paper provides a brief introduction to the R package **bio.infer**, a set of scripts that facilitates the use of maximum likelihood (ML) methods for predicting environmental conditions from assemblage composition. Environmental conditions can often be inferred from only biological data, and these inferences are useful when other sources of data are unavailable. ML prediction methods are statistically rigorous and applicable to a broader set of problems than more commonly used weighted averaging techniques. However, ML methods require a substantially greater investment of time to program algorithms and to perform computations. This package is designed to reduce the effort required to apply ML prediction methods.

Keywords: maximum likelihood, paleolimnology, weighted averaging, environmental prediction, R.

1. Introduction

Different organisms require different environmental conditions to persist. These differences are thought to have arisen over evolutionary timescales as organisms have specialized to optimally exploit certain habitat conditions and resources. One example of different environmental requirements can be observed in the differences between the mayfly, *Ameletus*, and the midge, *Tanytarsus*. The first insect has evolved to persist in cold, mountain streams, whereas the second insect has evolved to persist in warmer, lowland streams.

Observations of biota can often be used to predict environmental conditions at a site because of their differences in environmental requirements. For example, if one were to observe *Ameletus* in a stream, one would infer that the temperature of the stream is likely to be relatively

low. Conversely, an observation of *Tanytarsus* would suggest that the stream temperature is relatively high. Such predictions can be particularly useful in cases in which biological observations are the only data that are available. For example, skeletal remains of diatoms can be recovered from lake sediments, and the environmental conditions within the lake inferred from these remains (e.g., Birks, Line, Juggins, Stevenson, and ter Braak 1990; ter Braak and Juggins 1993). Remains collected from deeper (and older) sediments can then provide predictions of historical environmental conditions in the lake. A second application in which biologically-based environmental predictions may be useful is in the management of stream water quality. Water quality in streams and rivers is increasingly monitored by collecting and evaluating benthic invertebrates. By predicting environmental conditions from invertebrate observations, we may be able to extract useful information regarding the types of environmental stress that are present in different streams (e.g., Hämäläinen and Hüttünen 1996).

One of the most popular approaches for predicting environmental conditions from biological observations is weighted averaging. In weighted averaging, the environmental preferences of different taxa are quantified by computing the average environmental conditions of sites in which a given taxon is found, weighted by the abundance of the taxon at each site. This calculation yields a single, optimum value, which is an estimate of the environmental condition that is most preferred by the taxon. Predictions of the environmental conditions at a new site are then calculated by averaging the optima of all taxa that are observed at the site. Weighted averaging has been shown to be robust to noisy data and to provide relatively accurate predictions of environmental conditions (ter Braak and van Dam 1989).

For certain applications weighted averaging may not be the most appropriate technique to apply. Two issues, in particular, can limit the applicability of weighted average inferences. First, weighted average inferences are known to reduce the true range of environmental observations (i.e., they "shrink" the length of the environmental gradient). This shrinkage can be attributed to the fact that each environmental prediction is based on two distinct averaging operations (Birks *et al.* 1990). Different approaches for post hoc corrections of the inferences exist; however, the corrections themselves can introduce bias into the inferences (Yuan 2005). Thus, it can be very difficult to obtain an accurate prediction of true environmental conditions using weighted averaging. A second issue arises when weighted averaging is applied to environmental gradients that are correlated with one another. Because weighted average predictions can only be calculated with respect to a single environmental variable, changes in taxon occurrences that are a consequence of a variable that is correlated with the one of interest are interpreted only as an effect of the main variable. This limitation can reduce the accuracy of weighted average inferences when they are applied to covarying environmental gradients (Yuan 2007b). Furthermore, the limitation to single variable models virtually eliminates the possibility of using weighted average inferences to distinguish between the effects of different covarying environmental variables.

Predictions of environmental conditions that are based on a maximum likelihood (ML) formulation address many of these issues. In a ML-based approach, one first uses regression to model the observations of a taxon as a function of the values of one or more environmental variables. This reliance upon regression equations permits one to explore single and multivariate models as well as different curve shapes for representing the relationship between taxon observations and environmental conditions. Once taxon–environment relationships are estimated, inferences are computed by employing the same likelihood formulation as used

for fitting the original regression equations. Thus, ML inferences are internally consistent and preserve the original scaling of the modeled environmental variables. Also, the entire taxon–environment relationship is used to calculate the inference, rather than the single optimum point used in weighted averaging. However, ML methods have been used relatively infrequently (e.g., ter Braak and van Dam 1989; Oksanen, Läärä, Huttunun, and Meriläinen 1990) likely in part because computing ML inferences is considerably more demanding. Computational requirements are higher than for weighted averaging, and programming the algorithms to solve the ML inference problem are more involved.

Here, I introduce a set of R scripts (R Development Core Team 2007) for computing ML predictions of environmental conditions. I first describe the example data sets that will be used to illustrate the scripts (Section 2). Then, I discuss different scripts provided in the package, in the order in which one is likely to use them to compute ML predictions of environmental conditions (Section 3). Final concluding remarks are presented in Section 4. Hopefully, the availability of these scripts will facilitate the broader use of ML-based inference approaches.

2. The example data sets

The package is demonstrated with two distinct data sets. Regional-scale calibration data were collected by the US Environmental Protection Agency Environmental Monitoring and Assessment Program (EMAP) at randomly selected wadeable stream reaches across 12 states in the western United States in the summers of 2000 to 2002 (Stoddard, Peck, Olsen, Larsen, van Sickle, Hawkins, Hughes, Whittier, Lomnicky, Herlihy, Kaufmann, Peterson, Ringold, Paulsen, and Blair 2005). Independent data used for validation were collected at a smaller geographic scale in western Oregon in 1999 to 2000 (OR).

Models in the package are illustrated using stream temperature as the environmental gradient of interest. Stream temperatures change naturally along elevational gradients and can also change as a result of human activities. Instantaneous stream temperature was measured at all sites in EMAP and at selected OR sites at the time of sampling.

Benthic macroinvertebrates were sampled in both OR and EMAP data sets following the same protocol. At each stream site eight samples were collected at randomized locations in riffles with a modified D-frame kicknet (500 μm mesh) by disturbing a 0.09 m^2 area for 30 seconds. Samples were then combined and preserved with 95% ethanol. In the laboratory samples were spread on a gridded pan and organisms picked from randomly selected grid squares until at least 500 organisms were collected. Each organism was then identified to the lowest possible taxonomic level.

A total of 1089 samples in EMAP and 271 samples in OR are used in the examples shown here.

3. The package

Two steps are required for predicting environmental conditions from biological observations. First, relationships are estimated between the occurrence of different taxa and the environmental variables of interest (i.e., the "regression" step). This step is usually accomplished empirically using a calibration data set in which paired biological and environmental obser-

vations are available. Second, taxon–environment relationships defined in the first step are combined with observations of biota in an independent data set, and environmental conditions are predicted at new sites (i.e., the "prediction" step). In this second data set, environmental measurements are not required.

With real biological data, the regression and prediction steps become more involved because we must account for differences in the format with which taxon names are recorded and differences in the taxonomic resolution at which individual organisms are identified (e.g., species-level vs. genus-level identifications). These types of differences almost always accompany data sets of biological observations. To address this issues, I have included in **bio.infer** a tool that standardizes the processing of these data by matching biological observations to a standard data set that describes the full taxonomic hierarchy of each taxon.

Once a full taxonomic hierarchy is established, the computation of taxon–environment relationships requires that logistic regression models are fit for each distinct taxon. The script `taxon.env` facilitates this process and formats the resulting models for each taxon such that they can be easily used in the subsequent steps of the inference calculation.

After taxon–environment relationships are computed using a calibration data set, we typically use these relationships to predict environmental conditions based on a second, independent set of biological observations. Full taxonomic information for this second set of observations must again be standardized. Then, operational taxonomic units (OTUs) must be designated for this data. OTUs are designated to guarantee not only that each organism listed in the test data is associated with a taxon–environment relationship, but also that particular individuals in a sample are not counted more than once at different taxonomic levels.

Once, OTUs have been designated, a ML formulation can be used to predict environmental conditions in the test data using taxon–environment relationships from the calibration data and observations of the presence or absence of different OTUs in the test data.

3.1. Standardizing taxonomy

Biological data sets vary tremendously in the degree to which a complete taxonomic hierarchy is reported for different observed taxa. Some data sets report only the taxon name, while others report an incomplete taxonomic hierarchy for each observation. For example, family and genus are often reported, while tribe or subfamily are omitted. Furthermore, among datasets for which a full hierarchy is included, one can often find differences in the higher order taxonomy that is reported for a given taxon. Before different data sets are used for computing biological predictions (e.g., the calibration and test data sets), differences in taxonomy must be reconciled. This task is often time-consuming.

My solution to this problem is to standardize the taxonomy of different biological data sets to a single reference set of taxa, provided by the Integrated Taxonomic Information System (ITIS, `http://www.itis.gov/`). These data provide a consistent taxonomic hierarchy from phylum to genus. (Species-level data were not included to keep the database to a manageable size.) The ITIS database for the Kingdom Animalia was retrieved on November 16, 2006, and is provided in **bio.infer** as the data frame `itis.ttable`. Other Kingdoms will be added as needed.

The script `get.taxonomic` parses each taxon name in a biological data set into distinct character strings, and then queries the ITIS database to determine whether each character string is a valid taxon name. Several distinct types of taxon names are handled automatically

by the script. The first, and simplest case consists of a single character string that matches an ITIS entry. In this case, the full ITIS taxonomic hierarchy can be directly merged with the taxon name. A second common case occurs when an individual organism can only be identified as two possible taxa. In this case, two character strings extracted from a single taxon name (e.g., *Bezzia/Palpomyia*) are both found in the ITIS data. The script merges these types of records with the ITIS record for the *next coarser* taxonomic level. For example, if a taxon name consists of two genera, this record is merged with the tribe that contains the two genera (if it exists). In the case mentioned above, the two biting midges *Bezzia* and *Palpomyia* often cannot be distinguished from one another, and the compound identification is coarsened to the tribe *Palpomyiini* and matched with the ITIS data. A third type of taxon name is a species-level identification (e.g., the mayfly *Baetis bicaudatus*). Because the ITIS database only includes taxon names to the genus level, only the genus portion of the name will successfully match with an ITIS entry. The second character string extracted from this taxon name will not be found. In this case, the taxon is matched with the correct ITIS entry for genus, but the character string that identifies the species is saved in a separate field.

In the following example, biological observations from EMAP data are matched with the ITIS taxonomy table.

```
R> library("bio.infer")
R> options(width = 60)
R> data("itis.ttable")
R> data("bcnt.emapw")
R> bcnt.tax <- get.taxonomic(bcnt.emapw, itis.ttable,
+     outputFile = "sum.tax.table.txt")

Correct misspellings or synonyms:
Corrections entered.
             TAXANAME     CORRECTION
1               ACARI        ACARINA
2        ALBERTATHYAS
3            CHYRANDA      CHYRANDRA
4            NEOPORUS
5             PANISUS
6            PARTUNIA
7  PSEUDOTORRENTICOLA
8         RADOTANYPUS MACROPELOPIINI
9     SANFILIPPODYTES
10      SPERCHONTIDAE   SPERCHONIDAE
11     STYGOTHROMBIUM

Summary of taxa without matches:
            TAXANAME NUMBER OF OCCURRENCES
1       ALBERTATHYAS                     9
2           NEOPORUS                     7
3            PANISUS                     2
4           PARTUNIA                     1
5 PSEUDOTORRENTICOLA                     2
```

```
6    SANFILIPPODYTES                    18
7     STYGOTHROMBIUM                    11

The following taxa match with multiple ITIS records.
Select appropriate taxon from list.
      TAXON     PHYLUM      CLASS          ORDER
384 MENETUS ARTHROPODA    INSECTA        DIPTERA
385 MENETUS   MOLLUSCA GASTROPODA BASOMMATOPHORA
         FAMILY
384  TACHINIDAE
385 PLANORBIDAE

Check final taxa name assignments in sum.tax.table.txt
```

Usually, in any biological data set, the script fails to match a small number of taxon names with entries in ITIS. These taxon names may be misspelled or may represent taxa that are not officially recognized and included in ITIS. These remaining unmatched taxa must be corrected by hand by the user to any extent possible. Corrections are entered using the R utility `edit`. In the example shown above, the genus *Chyrandra* is misspelled and must be corrected, and the genus *Radotanypus* is not an officially recognized taxon, and its identification must be downgraded to tribe. Other corrections are entered for some of the taxa, but taxa for which no obvious correction is available can be left uncorrected. After corrections are entered, the script provides a summary of the number of occurrences of the remaining unmatched taxa, and provides the user an opportunity to further correct taxa names.

Taxon names are not guaranteed to be unique in the ITIS database. So, in the final step of this script, the biological data are examined to determine whether any taxon was matched with more than one record from the ITIS database. These taxa are displayed with possible matching taxonomic hierarchies, and the user is prompted to select the correct match. In the present example, I selected the Mollusca entry for the taxon, Menetus, as this is the taxon most likely to be observed in a stream ecosystem.

The final output of this script is the original biological observational data with each taxon now linked to a standardized taxonomic hierarchy (e.g., Phylum, Class, Order, Family, Genus, and Species). A summary of the resulting taxonomic hierarchy is also provided in the tab-delimited text file specified by the user (`outputFile`). The full taxonomic hierarchy for all taxa is too large to show here, but for illustration, a portion of the hierarchy for the family Ephemeridae is shown below.

```
            ORDER      FAMILY     GENUS           SPECIES
734 EPHEMEROPTERA EPHEMERIDAE
735 EPHEMEROPTERA EPHEMERIDAE  EPHEMERA
736 EPHEMEROPTERA EPHEMERIDAE HEXAGENIA
737 EPHEMEROPTERA EPHEMERIDAE HEXAGENIA HEXAGENIA.LIMBATA
             TAXANAME
734       EPHEMERIDAE
735      EPHEMERA SP.
736     HEXAGENIA SP.
737 HEXAGENIA LIMBATA
```

The field `TAXANAME` is the original taxon name provided in the biological data set, and the remaining information has been extracted from appropriately matched records in the ITIS database.

If necessary, species names in this file can be further edited by hand and reloaded using the script `load.revised.species`.

3.2. Estimating taxon–environment relationships

Once taxonomic information in the calibration data set is standardized, the script `taxon.env` computes taxon–environment relationships at different taxonomic levels. This script provides a convenient tool for fitting generalized linear models (`glm`) for every possible taxon in the biological data and for formatting the resulting models appropriately for subsequent calculation of inferences using ML.

Logistic regression models can be expressed mathematically for two explanatory variables, x_i and z_i, observed in sample i, as follows:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = g_{ij} = b_{0j} + b_{1j}x_i + b_{2j}x_i^2 + b_{3j}z_i + b_{4j}z_i^2 \tag{1}$$

where g_{ij} is the logit transform of the probability of occurrence p_{ij} of taxon j in sample i and $b_{0j} \ldots b_{4j}$ are the regression coefficients. The presence of the quadratic terms allows for unimodal responses, a common form for taxon–environment relationships (ter Braak and Looman 1986). Higher order terms are not expected and therefore not allowed in the model.

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = g_{ij} = b_{0j} + b_{1j}x_i + b_{2j}x_i^2 + b_{3j}z_i + b_{4j}z_i^2 + b_{5j}x_iz_i \tag{2}$$

Specifying pairwise interaction terms in Equation (2) can occasionally improve the fit of certain models, so the scripts provided in this package allow for this possibility.

The script, `taxon.env`, first summarizes biological observations at different taxonomic levels. Then, at each level, it identifies taxa that were present at and absent from at least M samples, where the value of M is set at $10\times$ the number of degrees of freedom of the specified model. This criterion has been cited as the minimum requirement to prevent overfitting regression models (Harrell 2001). Models are fit for each of the selected taxa.

We can compute taxon–environment relationships with respect to the water temperature in the stream (`STRMTEMP`) using the EMAP data as follows:

```
R> data("envdata.emapw")
R> coef <- taxon.env(form = ~STRMTEMP + STRMTEMP^2,
+     bcnt = bcnt.tax, envdata = envdata.emapw,
+     bcnt.siteid = "ID.NEW", bcnt.abndid = "ABUND",
+     env.siteid = "ID.NEW", tlevs = "SPECIES",
+     dumpdata = TRUE)

Model formula:  resp ~STRMTEMP+I(STRMTEMP^2)
Minimum number of occurrences:  30
```

```
Warning: fitted probabilities numerically 0 or 1 occurred
    for DESPAXIA.AUGUSTA
Warning: fitted probabilities numerically 0 or 1 occurred
    for DRUNELLA.COLORADENSIS/FLAVILINEA
Warning: fitted probabilities numerically 0 or 1 occurred
    for MOSELIA.INFUSCATA
Number of taxa modeled:
  TAXON.LEVEL NUM.MODS
1     SPECIES       64
```

The function call to `taxon.env` is similar to calling statements for `glm`. The user specifies the model equation using an identical format as used with `glm` except that the response variable is not identified. The names of data frames containing biological and environmental data are also specified, as well as the names of fields required for merging the data frames and running the models. For this first example, two other options are selected. First, the parameter `tlevs` is specified as `SPECIES`, forcing the script to only estimate taxon–environment relationships at the `SPECIES` level. The default value for this parameter (`all`) forces the script to estimate relationship at all possible taxonomic levels. Second, the logical parameter `dumpdata` is specified as `TRUE`, which forces the script to include the raw data used to fit the models in the output. These raw data can be useful if one wishes to further examine the model fit to the data. The default value for this parameter is `FALSE` to minimize the size of the output file.

The main output of this script are the regression coefficients for each of the models stored in a single matrix (`csave`). Other items include a vector of taxon names that are represented in the coefficient matrix (`tnames`), the range of observations in the environmental data (`xlims`), the formula that specifies the regression model (`form`), the names of the explanatory variables (`xvar`), a measure of the predictive accuracy of the model (`roc`), and the raw data used to fit the model (`raw.data`).

```
R> names(coef)

[1] "tnames"   "csave"    "xvar"     "xlims"    "form"
[6] "roc"      "raw.data"
```

Plots of estimated mean taxon–environment relationships can be examined as follows by using the script `view.te`.

```
R> view.te(coef, plotform = "windows")
```

This script allows the user to interactively select the taxa that are plotted. Plots can be directed to either the console window or a pdf file. If the coefficient file, `coef`, was generated with the option `dumpdata = TRUE`, then observed data are included with the plots of the mean relationships. Example plots of the fitted relationships are shown for two species stored in `coef` in Figure 1. The stream temperature at which these two species are most likely to be observed is very similar (the maximum point on the curve). However, *Zapada columbiana* exhibits very little tolerance for higher stream temperatures and is not observed in temperatures greater than 17.5°. In contrast, *Zapada cinctipes* is occasionally observed in temperatures exceeding 25°.

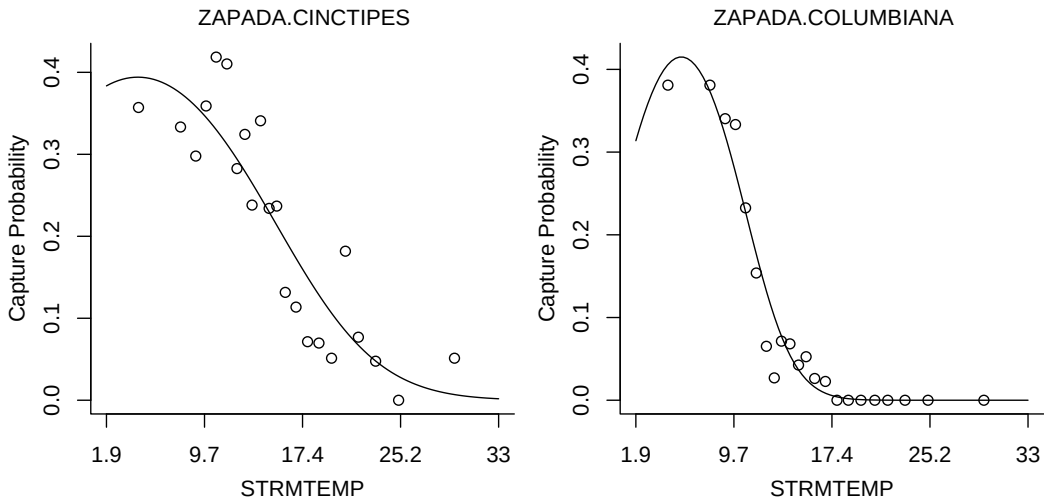

Figure 1: Examples of estimated taxon–environment relationships for two stonefly species. Open circles represent mean capture probability estimated from approximately 40 samples with stream temperature centered around plotted location.

The script `taxon.env` also automatically computes the area under the receiver operator characteristic curve (ROC) for each of the models. This statistic provides an indication of the classification strength of the model. Values of the area under ROC near 1 indicate that the model very accurately predicts sites where the taxon is present and sites where the taxon is absent. Conversely, values of area under ROC near 0.5 indicate that the model poorly predicts taxon occurrences.

For the species–temperature models developed for this example, area under ROC ranges from 0.60 to 0.87, with a mean value of 0.73 (Figure 2), so overall, we can conclude that stream temperature is a fairly strong predictor for the occurrence of different taxa.

Note that even after restricting taxa to those that occur in a minimum number of samples, occurrences of some selected taxa may be too tightly clustered in a small portion of the modeled gradient. For these taxa, the modeled mean probability of occurrence will approach values of zero or one that exceed the machine accuracy of the computer, and a warning message will be displayed. The model fit for these taxa can then be inspected more closely. In the present example, models for three species trigger the warning message, but further inspection of these models indicates that the model fit is appropriate.

To compute inferences, we need to estimate taxon–environment relationships for as many taxa as possible, so we rerun `taxon.env` with option `tlev` set to `all` and `dumpdata = FALSE`.

```
R> coef <- taxon.env(form = ~STRMTEMP + STRMTEMP^2,
+     bcnt = bcnt.tax, envdata = envdata.emapw,
+     bcnt.siteid = "ID.NEW", bcnt.abndid = "ABUND",
+     env.siteid = "ID.NEW", tlevs = "all", dumpdata = FALSE)

Model formula:  resp ~STRMTEMP+I(STRMTEMP^2)
Minimum number of occurrences:  30
```

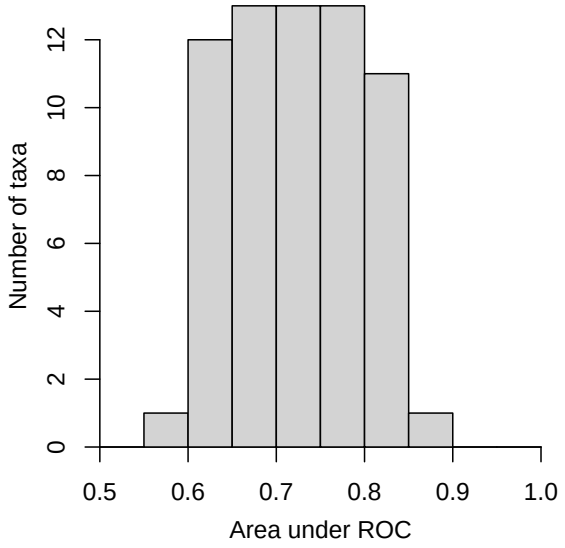

Figure 2: Histogram of area under ROC values for species–temperature relationships in EMAP.

```
Warning: fitted probabilities numerically 0 or 1 occurred
    for PLEUROCERIDAE
Warning: fitted probabilities numerically 0 or 1 occurred
    for DIPLOPERLINI
Warning: fitted probabilities numerically 0 or 1 occurred
    for DESPAXIA
Warning: fitted probabilities numerically 0 or 1 occurred
    for JUGA
Warning: fitted probabilities numerically 0 or 1 occurred
    for MOSELIA
Warning: fitted probabilities numerically 0 or 1 occurred
    for DESPAXIA.AUGUSTA
Warning: fitted probabilities numerically 0 or 1 occurred
    for DRUNELLA.COLORADENSIS/FLAVILINEA
Warning: fitted probabilities numerically 0 or 1 occurred
    for MOSELIA.INFUSCATA
Number of taxa modeled:
   TAXON.LEVEL NUM.MODS
1       PHYLUM        4
2    SUBPHYLUM        2
3        CLASS        7
4     SUBCLASS        6
5   INFRACLASS        1
6   SUPERORDER        3
7        ORDER       18
```

```
8     SUBORDER       16
9   INFRAORDER       13
10 SUPERFAMILY       19
11      FAMILY       83
12   SUBFAMILY       44
13       TRIBE       34
14       GENUS      205
15     SPECIES       64
```

3.3. Designating operational taxonomic units

The primary motivation for designating operational taxonomic units (OTUs) is to identify a consistent set of taxon names to use across the entire biological data set that maximizes the amount of information that can be extracted from the observations. This process also guarantees that individual observations of taxa are not counted more than once in the inference computation and that the absence of a given taxon from a site is meaningful. The concept of operational taxonomic units (OTUs) is not unique to this application and has been used previously in predictive models of assemblage composition (e.g., Ostermiller and Hawkins 2004).

OTUs are designated using a two-step process. First, each taxon in the data set is compared with the list of taxa for which taxon–environment relationships are available. If the taxon of interest is not found on the list, it is downgraded to the next coarser taxonomic level until a corresponding taxon–environment relationship is found. This first step ensures that every taxon observed in the data set is associated with a taxon–environment relationship, if an appropriate match exists.

Second, OTUs are specified to ensure that no individual organism in a sample is counted as more than one taxon in the inference computation. This process is best illustrated with an example. In Table 1 data from OR for the mayfly genus *Epeorus* are shown. Four species of *Epeorus* (*E. albertae*, *E. deceptivus*, *E. grandis*, and *E. longimanus*) were observed in the data set. Individuals were also observed that could not be identified to the species-level and were only identified as *Epeorus*. Of the four species, three had taxon–environment relationships defined previously from the EMAP data (stored in `coef`). *E. albertae* had no taxon–environment relationship, so it effectively could only be considered as a genus-level identification. So, if we summarize the number of samples in which different observations occur, we find that in 97 samples $(59 + 38)$ only genus-level identifications were available, whereas in 140 samples $(7 + 59 + 74)$ species-level observations were available. Since more samples have species-level data, we retain the observations of *E. deceptivus*, *E. grandis*, and *E. longimanus*, and observations of *Epeorus* and *E. albertae* are not used in the inference computation. This process is repeated for all taxa at all taxonomic levels.

The criterion that is used to decide whether to retain coarser versus finer taxonomic identifications simply compares the number of samples at which taxa at each level of identification was observed. Then, the taxonomic level at which more samples are observed is retained. This criterion is arbitrary and based on the notion that more refined taxonomy provides more specific inference information. Note, though, that the effects of choosing coarser versus finer taxonomy differ. In the example shown above, if we chose to use genus-level identification, we would not be able to use species-level taxon–environment relationships. However, in those

Taxon name	Number of occurrences	Taxon–environment available?
Epeorus	59	Y
E. albertae	38	N
E. deceptivus	7	Y
E. grandis	59	Y
E. longimanus	74	Y

Table 1: Number of occurrences of *Epeorus* as genus and as species.

samples where *Epeorus* individuals were identified to the species level, we could still use genus-level taxon–environment relationships to account for those individuals when computing the inference. Conversely, when we select species-level identifications, genus-level identifications of *Epeorus* do not contribute at all to any of the inferred values. For certain taxonomic groups, the relationship between a given genus and the environmental variable of interest may be very strong, and it would make sense to retain genus-level rather than species-level information, even if more samples had species-level data. The present version of `get.otu` provides only a very simple approach for designating consistent taxa that does not take into account the strength of different taxon–environment relationships. More sophisticated decision criteria may be implemented in future versions of this script.

Commands for preparing the OR biological data set for computing inferences using taxon–environment relationships stored in `coef` are as follows:

```
R> data("bcnt.OR")
R> bcnt.tax.OR <- get.taxonomic(bcnt.OR, itis.ttable)

Correct misspellings or synonyms:
Corrections entered.
      TAXANAME     CORRECTION
1 ALBERTATHYAS
2  HYDRACARINA        ACARINA
3 PALAEGAPETUS  PALAEAGAPETUS
4  RADOTANYPUS MACROPELOPIINI

Summary of taxa without matches:
      TAXANAME NUMBER OF OCCURRENCES
1 ALBERTATHYAS                     2

Check final taxa name assignments in sum.tax.table.txt

R> bcnt.otu.OR <- get.otu(bcnt.tax.OR, coef)

Review the changes in species names:
          Original name                      Revised name
1 DRUNELLA.COLORADENSIS DRUNELLA.COLORADENSIS/FLAVILINEA
2    EUBRIANAX.EDWARDSI              EUBRIANAX.EDWARDSII
```

```
3   RHYACOPHILA.PELLISA        RHYACOPHILA.PELLISA/VALUMA
4    RHYACOPHILA.VALUMA        RHYACOPHILA.PELLISA/VALUMA

Final OTU/taxa table stored in sum.otu.txt
```

In the example shown above, we first standardize the taxonomy of the Oregon data with `get.taxonomic`. Then, the script `get.otu` is used to designate OTUs for Oregon data, with respect to taxon–environment relationships estimated from EMAP (saved in `coef`).

Operationally, `get.otu` first attempts to match species names in the OR data to species names with taxon–environment relationships saved in `coef`. Because species names are not included in the ITIS data, their format does not adhere to consistent standards. Therefore, slightly more flexibility is required in performing this match. For example, in contrast to higher level taxonomy, compound names are allowed at the species level (e.g., *Rhyacophila pellisa/valuma*). Then, when OR species names are matched to these taxon–environment relationships, both *Rhyacophila pellisa* and *Rhyacophila valuma* are matched to the same compound species from EMAP. Also, slight misspellings of species names are allowed (e.g., *Eubrianax edwardsii*). Once species names are reconciled, the script systematically matches observations of different taxa at all taxonomic levels to existing taxon–environment relationships specified in `coef`. As described above, taxa that cannot be matched are downgraded to coarser identifications until a match is found. Then, the decision criterion described above for selecting the "most informative" taxonomic level is applied to each taxon, starting at the coarsest taxonomic level and proceeding through to the genus level.

The output of `get.otu` is the original biological observation data frame with an `OTU` field appended for each taxon. Also, the script produces a summary table that provides the full taxonomic hierarchy for each taxon in the data set, the number of occurrences of that taxon, the associated taxon name for which a taxon–environment relationship is available (`otufin1`), and the final assigned OTU (`otufin2`). This file is provided as tab-delimited text so it can easily viewed and edited. The user can change OTU designations manually in this file, and these changes can be reloaded using the script `load.revised.otu`.

3.4. Maximum likelihood inference

Maximum likelihood inferences are computed by expressing a binomial likelihood function as a function of the set of possible taxa that could occur at a site.

$$L_i = \prod_{j=1}^{N} p_j^{Y_{ij}} (1 - p_j)^{1-Y_{ij}} \tag{3}$$

where p_j is the probability of occurrence of taxon, j, modeled in Equation (1), and the product is computed over all N OTUs designated for the data set. Note that we assume now that we have a functional representation of p_j for all possible values of the explanatory variables and so the subscript i is no longer included in p_j. The variable Y_{ij} is equal to 1 when taxon, j, is present at site i, and zero when the taxon is absent. Here, i indexes different sites in the second data set (i.e., the OR data in this example). As noted earlier, each p_j has been modeled as a function of one or more environmental variables. The values of these environmental variables at which the likelihood is maximized gives the most likely environmental conditions for the sample, given the observed biological assemblage.

Maximizing the likelihood function is facilitated by first taking the log of the likelihood:

$$\log L_i = \sum_{j=1}^{N} Y_{ij} g_j + \log\left(\frac{1}{1+\exp g_j}\right) \tag{4}$$

where g_j is the logit transform of the probability of occurrence defined in Equation (1). Identifying the maximum point of the log-likelihood is a constrained optimization problem, where the box constraints are set by the limits of the observed environmental variables in the calibration data set. In other words, I restrict inferences to within the range of conditions observed in the calibration data set. The optimization problem is solved using the script, `optim`, provided in R, with the iterative solution method, `L-BFGS-B`, selected. The efficiency of the iterative solution is greatly enhanced when the gradient of the function being optimized is provided. The x-component of this gradient can be written as follows.

$$\frac{\partial(\log L_i)}{\partial x} = \sum_{j=1}^{N} Y_{ij} \frac{\partial g_j}{\partial x} - \frac{\exp g_j}{1+\exp g_j} \frac{\partial g_j}{\partial x} \tag{5}$$

Similar equations can be written for other components of the gradient.

Computing biological inferences for the OR data proceeds as follows:

```
R> ss <- makess(bcnt.otu.OR)
R> inferences <- mlsolve(ss, coef, site.sel = "99046CSR",
+     bruteforce = TRUE)

R> print(inferences)

       SVN STRMTEMP Inconsistent
1 99046CSR  15.6759        FALSE
```

In **bio.infer** two scripts are run to compute inferences. First, biological observations are reformatted as a sample-OTU matrix (`makess`), in which each OTU corresponds to a single column and each distinct sample corresponds to a single row of the matrix. This format is convenient for evaluating likelihood values at each site. Then, the script `mlsolve` computes maximum likelihood inferences for different samples based on the sample-OTU matrix and the set of regression coefficients that describe taxon–environment relationships (`coef`). The script first specifies functions that evaluate the log-likelihood, as defined in Equation (4), and the gradient of the log-likelihood, as defined in Equation (5), given values of the environmental variables, a matrix of regression coefficients (`coef`), and observations of the presence and absence of different OTU in a sample (one row of the sample-OTU matrix). Then, the script calls `optim` for each sample. Because the optimization problem is solved iteratively, a possibility exists that local, rather than global optima will be found. To guard against this possibility, solutions are computed for each sample using several different initial guesses. Cases in which different optimum points are found that have similar log-likelihood values are flagged as `Inconsistent` in the output data frame.

For illustrative purposes, the example shown above is computed for a single site (selected with `site.sel`). Also, the solution routine is forced to compute log-likelihoods for a set of

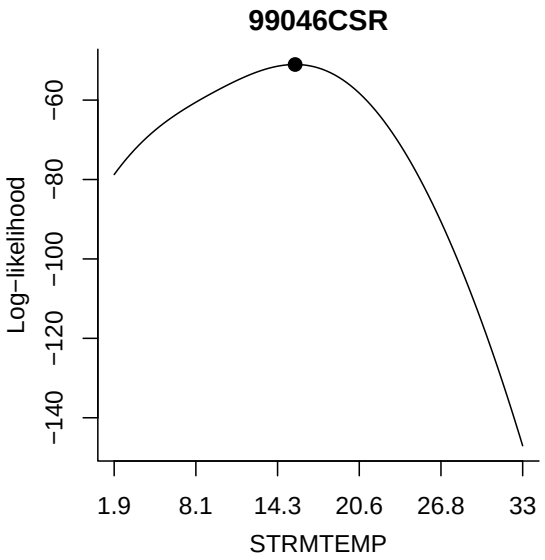

Figure 3: Example of log-likelihood curve at a single site, as a function of stream temperature. Solid circle shows the point of maximum log-likelihood.

values spanning the entire range of possible stream temperatures (`bruteforce = TRUE`). Using `bruteforce` allows one to plot the entire log-likelihood curve (Figure 3); however, such a computation can be very time-consuming. The default option of `bruteforce = FALSE` only evaluates the likelihood function at selected points required to iteratively identify the location of the maximum likelihood, and therefore runs much more quickly. The advantage of the `bruteforce = TRUE` option is that it allows one to graphically assess the accuracy of the iterative solution. In the present case, the inferred stream temperature of 15.6° calculated iteratively does appear to correspond to the point at which the log-likelihood value is maximized.

We now set `bruteforce = FALSE` to compute predictions at all sites in the OR data set.

```
R> inferences <- mlsolve(ss, coef, site.sel = "all",
+     bruteforce = FALSE)
```

In the OR data used as an example here, direct measurements of stream temperature are available, so we can assess the accuracy of the inferred temperatures. These measurements are plotted against observations in Figure 4 (the dashed line in the figure shows a 1:1 correspondence). The root mean square error for the predictions in this case was a relatively low 2.04.

3.5. Additional tools

Several additional scripts are included to help increase the practical utility of this package. Two scripts that allow the user to manually correct species names (`load.revised.species`) and OTU designations `load.revised.otu` have already been mentioned in previous sections.

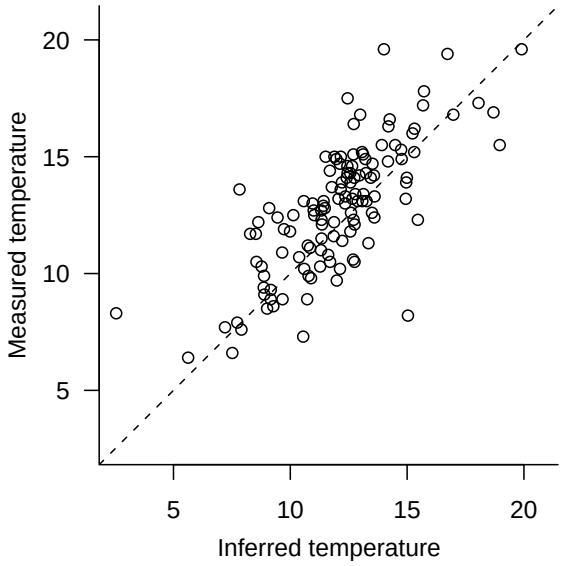

Figure 4: Inferred versus measured temperature in Oregon. Dashed line shows the 1:1 relationship.

Both of these scripts require that the user first edit and save tab-delimited taxa table files using other software programs (e.g., spreadsheets or text editors). These two scripts then incorporate the changes into the biological observation data frames within R. In general, manual corrections should not be required for the output of `get.taxonomic` and `get.otu`. However, in the case of `get.taxonomic` it is possible that an unusual taxonomic naming convention leads to errors in the species names provided by the script. In the case of `get.otu` it is possible that the user will know of certain taxa in which a change in the OTU designations will yield more accurate inferences.

The script `view.te` allows the users to view plots of taxon–environment relationships and has also been discussed briefly in a previous section. This script uses the regression coefficients in a `coef` file and plots contour plots (in the case of two variable models) or line plots of taxon–environment relationships for different taxa. No option is provided at this time for plots of relationships based on greater than two environmental variables. Plots are returned by default to the file `taxon.env.pdf`.

3.6. Pre-computed taxon–environment relationships

A data frame of pre-computed taxon–environment regression coefficients is also included in the package that addresses two of the main flaws associated with the simple regression models provided in `taxon.env`. First, the regression models computed with `taxon.env` assume that environmental measurements are observed with no error and that these measurements are directly relevant to the persistence of different taxa. Both of these assumption are not likely to be true. For example, the instantaneous temperature measurements used in the examples shown above are unlikely to be directly related to the probability of occurrence of different taxa. Instead, they provide a relatively inaccurate estimate for the average temperature in

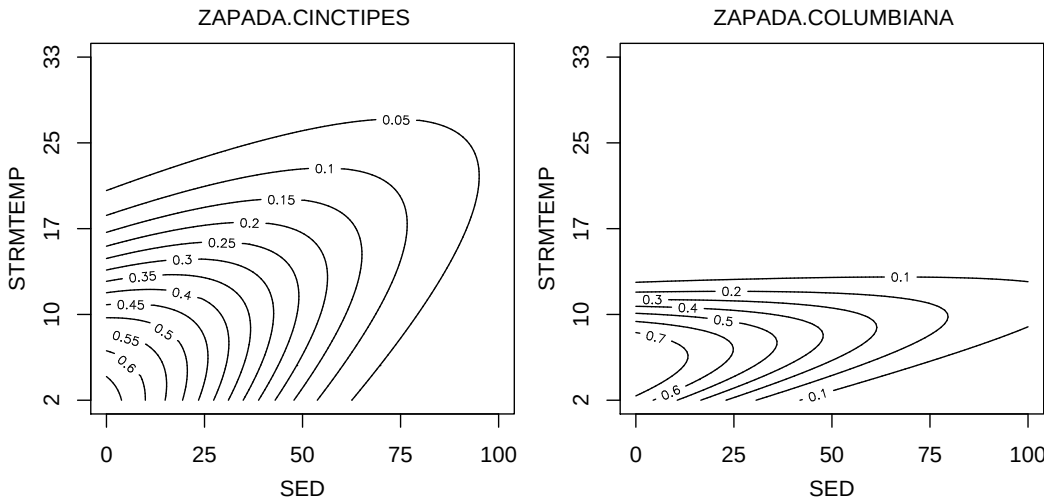

Figure 5: Examples of pre-computed taxon–environment relationships for two stonefly species with respect to percent sand/fines in substrate (SED) and stream temperature (STRMTEMP). Compare to relationships for only stream temperature shown in Figure 1

the stream, which in turn, is more closely linked to the occurrence of different invertebrates. Regression model results can be adjusted for these "measurement errors", and the predictions based on these adjusted results have been shown to be more accurate than predictions using simple regressions (Yuan 2007a).

A second refinement one can introduce to estimates of taxon–environment relationships is to explicitly incorporate sampling designs in the models. The EMAP data used in this paper was collected with a stratified random sampling design, and each sample represented a known portion of stream network (Stoddard *et al.* 2005). By incorporating sample weights in the regression models, we ensure that the estimated taxon–environment relationships are representative of the sampled region.

Taxon–environment relationships for the western United States that take into account both sampling design and measurement error are included in the **bio.infer** as `coef.west.wt`. These coefficients model the occurrences of different taxa as a function of stream temperature and the percentage of fine sediment in the stream substrate. Consequently, using `view.te` to view the taxon–environment relationships produces contour plots (Figure 5). The temperature preferences of each species remain generally the same as observed in the temperature-only model (Figure 1). However, we can now observe that both species also prefer relatively low levels of sand and fines in the substrate. Furthermore, for both species, a small interactive effect can be seen, in which the optimal temperature increases as the amount of sand and fines increase.

These pre-computed coefficients can be used in place of the coefficients calculated from `taxon.env`, using the same sequence of steps to infer environmental conditions from biological assemblage composition.

```
R> bcnt.otu.OR <- get.otu(bcnt.tax.OR, coef.west.wt)
```

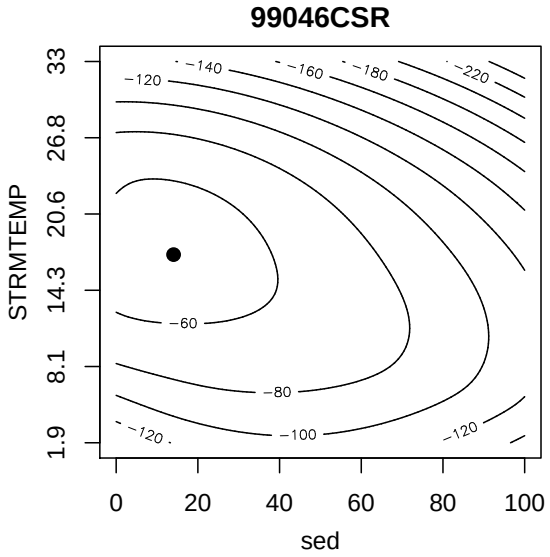

Figure 6: Example of log-likelihood surface at a single site, as a function of stream temperature and percent sand/fines in the substrate. Solid circle shows the point of maximum log-likelihood.

```
Review the changes in species names:
          Original name                       Revised name
1 DRUNELLA.COLORADENSIS DRUNELLA.COLORADENSIS/FLAVILINEA
2    EUBRIANAX.EDWARDSI              EUBRIANAX.EDWARDSII
3   RHYACOPHILA.BRUNNEA        RHYACOPHILA.BRUNNEA/VEMNA

Final OTU/taxa table stored in sum.otu.txt

R> ss <- makess(bcnt.otu.OR)

R> inference <- mlsolve(ss, coef.west.wt, site.sel = "99046CSR",
+     bruteforce = TRUE)

R> print(inference)

       SVN      sed STRMTEMP Inconsistent
1 99046CSR 14.05123 17.24963        FALSE
```

Because the taxon–environment relationships are multivariate, calculations of inferred temperatures and sediment require that the maximum point of a log-likelihood surface is identified (Figure 6). Using pre-computed taxon–environment relationships slightly changes the inferred stream temperature at the same site as shown in Figure 3. In general, these pre-computed taxon–environment relationships provide more accurate predictions of stream temperature

and substrate composition for western streams. Pre-computed taxon–environment relationships for other geographic locations and other environmental variables will soon be available at http://www.epa.gov/caddis/.

4. Concluding remarks

I have described in this paper a set of scripts that facilitates the computation of maximum likelihood predictions of environmental conditions. Hopefully, this package will provide a useful tool for paleolimnologists and other ecologists interested in estimating environmental conditions from biological observations. Other tools provided in this package for matching biological observations to standardized taxonomy and for designating operational taxonomic units may also be useful to anyone who analyzes large biological and environmental data sets.

Acknowledgments

The author acknowledges the field data collection efforts of the US EPA EMAP Surface Waters Program and the Oregon Department of Environmental Quality. This paper represents the views of the author and does not represent those of the US Environmental Protection Agency. Mention of trade names does not constitute endorsement.

References

Birks HJB, Line JM, Juggins S, Stevenson AC, ter Braak CJF (1990). "Diatoms and pH Reconstruction." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **327**, 263–278.

Hämäläinen H, Hüttünen P (1996). "Inferring the Minimum pH of Streams from Macroinvertebrates Using Weighted Average Regression and Calibration." *Freshwater Biology*, **36**, 697–709.

Harrell FE (2001). *Regression Modeling Strategies*. Springer-Verlag, New York, NY.

Oksanen J, Läärä E, Huttunun P, Meriläinen J (1990). "Maximum Likelihood Prediction of Lake Acidity based on Sedimented Diatoms." *Journal of Vegetation Science*, **1**, 49–56.

Ostermiller JD, Hawkins CP (2004). "Effects of Sampling Error on Bioassessments of Stream Ecosystems: Applications to RIVPACS-type Models." *Journal of the North American Benthological Society*, **23**(2), 363–382.

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/.

Stoddard JL, Peck DV, Olsen AR, Larsen DP, van Sickle J, Hawkins CP, Hughes RM, Whittier TR, Lomnicky G, Herlihy AT, Kaufmann PR, Peterson SA, Ringold PL, Paulsen SG, Blair R (2005). "Western Streams and Rivers Statistical Summary." *Technical Report EPA 620/R-05/006*, US Environmental Protection Agency, Washington, DC.

ter Braak CJF, Juggins S (1993). "Weighted Averaging Partial Least Squares Regression (WA-PLS): An Improved Method for Reconstructing Environmental Variables from Species Assemblages." *Hydrobiologia*, **269/270**, 485–502.

ter Braak CJF, Looman CWN (1986). "Weighted Averaging, Logistic Regression and the Gaussian Response Model." *Vegetatio*, **65**, 3–11.

ter Braak CJF, van Dam H (1989). "Inferring pH from Diatoms: A Comparison of Old and New Calibration Methods." *Hydrobiologia*, **178**, 209–223.

Yuan LL (2005). "Sources of Bias in Weighted Average Inferences of Environmental Conditions." *Journal of Paleolimnology*, **34**, 245–255.

Yuan LL (2007a). "Effects of Measurement Error on Inferences of Environmental Conditions." *Journal of the North American Benthological Society*, **26**, 152–163.

Yuan LL (2007b). "Using Biological Assemblage Composition to Infer the Values of Covarying Environmental Factors." *Freshwater Biology*, **52**, 1159–1175.

Affiliation:

Lester L. Yuan
Office of Research and Development
US Environmental Protection Agency
Washington DC 20460, United States of America
E-mail: yuan.lester@epa.gov

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 22, Issue 3 *Submitted:* 2007-01-10
September 2007 *Accepted:* 2007-06-22