Converts Common Biological (Response) and Environmental (Predictor) Data Formats to a GDM Site-Pair Table

Description

This function takes input biological (response) and environmental, geographic, and other predictor variables and builds a site-pair table required for fitting a Generalized Dissimilarity Model using the gdm function. NOTE: x-y coordinates of sites MUST be present in either the biological or the environmental data.

The input biological data can be in one of the following four formats. Note that the general term "species" is used, but any classification of biological entities (e.g. functional types, haplotypes, etc) can be used as long as an appropriate distance metric is also supplied (see "dist" argument):

1. site-by-species matrix

2. x, y, species list

3. site-by-site biological distance (dissimilarity) matrix

4. an existing site-pair table (see Details)

Predictor data can be provided in three formats:

- a site-by-predictor matrix with a column for each predictor variable and a row for each site

- a raster stack, with one raster for each predictor variable

- one or more site-by-site distance matrices using the "distPreds" argument (see below).

Usage

```
formatsitepair(bioData, bioFormat, dist="bray", abundance=FALSE, siteColumn=NULL, XColumn,
        YColumn, sppColumn=NULL, abundColumn=NULL, sppFilter=0, predData, distPreds=NULL,
        weightType="equal", custWeights=NULL, sampleSites=1)
```

Arguments

bioData	The input biological (response) data table, in one of the four formats defined above (see Details).
bioFormat	An integer code specifying the format of bioData. Acceptable values are 1, 2, 3, or 4 (see Details).
dist	Default = "bray". A character code indicating the metric to quantify pairwise site distances / dissimilarities. Calls the vegdist function from the **vegan** package to calculate dissimilarity and therefore accepts any method available from vegdist.
abundance	Default = FALSE. Indicates whether the biological data are abundance data (TRUE) or presence-absence (0, 1) data (FALSE).
siteColumn	The name of the column in either the biological or environmental data table containing site codes/names. If a site column is provided in both the biological and environmental data, the site column name must be the same in both tables.
XColumn	The name of the column containing x-coordinates of sample sites. X-coordinates can be provided in either the biological or environmental data tables, but MUST be in at least one of them. If an x-coordinate column is provided in both the biological and environmental data, the column name must be identical.
YColumn	The name of the column containing y-coordinates of sample sites. Y-coordinates can be provided in either the biological or environmental data tables, but MUST be in at least one of them. If a y-coordinate column is provided in both the biological and environmental data, the column name must be identical.
sppColumn	Only used if bioFormat = 2 (x, y, species list). The name of the column containing unique name / identifier for each species.
abundColumn	If abundance = TRUE, this parameter identifies the column containing the measure of abundance at each site. Only used if bioFormat = 2 (i.e., x, y, species list), though in the case of abundance data, the format would be: x, y, species, abundance.
sppFilter	Default = 0. To account for limited sampling effort at some sites, sppFilter removes all sites at which the number of recorded species (i.e., observed species richness) is less than the specified value. For example, if sppFilter = 5, all sites with fewer than 5 recorded species will be removed.
predData	The environmental predictor data. Accepts either a site-by-predictor table or a raster stack.
distPreds	An optional list of distance matrices to be used as predictors in combination with predData. For example, a site-by-site dissimilarity matrix for one biological group (e.g., trees) can be used as a predictor for another group (e.g., ferns). Each distance matrix must have as the first column the names of the sites (therefore the matrix will not be square). The name of the column containing the site names should have the same name as that provided for siteColumn argument. Site IDs are required here to ensure correct ordering of sites in the construction of the site-pair table. Note that the formatsitepair function will not accept only distances matrices as predictors (i.e., at least one predictor variable is required). If you wish to fit GDM using only distance matrices, provide one fake predictor (e.g., with all sites have the same value), plus site and coordinate columns if needed. The s1 and s2 columns for this variable can then be removed before fitting the GDM.
weightType	Default = "equal". Defines the weighting for sites. Can be either: (1) "equal" (weights for all sites set = 1), (2) "richness" (each site weighted according to number of species recorded), or (3) "custom" (user defined). If weightType="custom", the user must provide a vector of site weights equal to the number of rows in the full site-pair table (i.e., before species filtering (sppFilter argument) or sub-sampling is taken into account (sampleSites argument)).
custWeights	A two column matrix or data frame of user-defined site weights. The first column should be the site ID and should be named the same as that provided for siteColumn argument. The second column should be numeric weight values and should be named "weights". The weight values represent the importance of each site in model fitting, and the values in the output site-pair table is an average of the two sites in each site-pair. Required when weightType = "custom". Ignored otherwise.
sampleSites	Default = 1. A number between 0-1 indicating the fraction of sites to be used to construct the site-pair table. This argument can be used to reduce the number of sites to overcome possible memory limitations when fitting models with very large numbers of sites.

Details

bioData and bioFormat: The function accepts biological data in the following formats:

bioData = site-by-species matrix; bioFormat = 1: assumes that the response data are provided with a site ID column (specified by siteCol) and, optionally, two columns for the x & y coordinates of the sites. All remaining columns contain the biological data, with a column for each biological entity (most commonly species). In the case that a raster stack is provided for the environmental data (predData), x-y coordinates MUST be provided in bioData to allow extraction of the environmental data at site locations. The x-y coordinates will be intersected with the raster stack and, if the number of unique cells intersected by the points is less than the number of unique site IDs (i.e. multiple sites fall within a single cell), the function will use the raster cell as the site ID and aggregate sites accordingly. Therefore, model fitting will be sensitive to raster cell size. If the environmental data are in tabular format, they should have the same number of sites (i.e., same number of rows) as bioData. The x-y coordinate and site ID columns must have the same names in bioData and predData.

bioData = x, y, species list (optionally a fourth column with abundance can be provided); bioFormat = 2: assumes a table of 3 or 4 columns, the first two being the x & y coordinates of species records, the third (sppCol) being the name / identifier of the species observed at that location, and optionally a fourth column indicating a measure of abundance. If an abundance column is not provided, presence-only data are assumed. In the case that a raster stack is provided for the environmental data (predData), the x-y coordinates will be intersected with the raster stack and, if the number of unique cells intersected by the points is less than the number of unique site IDs (i.e. multiple sites fall within a single cell), the function will use the raster cell as the site ID and aggregate sites accordingly. Therefore, model fitting will be sensitive to raster cell size.

bioData = site-by-site distance (dissimilarity) matrix; bioFormat = 3: is used when a site-by-site distance (dissimilarity) matrix has already been created for the biological response (e.g., Fst for genetic data). The distance matrix must have as the first column the names of the sites (therefore the matrix will not be square). The column of site names should have the same name as the siteColumn argument. Only the lower half (triangle) of the matrix is needed to create the site-pair output table, but this function automatically removes the upper half if present. This is the only bioFormat in which the environmental data CANNOT be provided as a raster object.

bioData = site-pair table; bioFormat = 4: with an already created site-pair table, this option allows the user to add one or more distance matrices (see distPreds above) to the existing site-pair table and/or sub-sample the site-pair table (see sample above). If the site-pair table was not created using the formatsitepair function, the user will need to ensure the order of the sites matches that in other tables being provided to the function.

NOTES: (1) The function assumes that the x-y coordinates and the raster stack (if used) are in the same coordinate system. No checking is performed to confirm this is the case. (2) The function assumes that the association between the provided site and x-y coordinate columns are singular and unique. Therefore, the function will fail should a given site have multiple coordinates associated with it, as well as multiple sites being given the exact same coordinates.

Value

A site-pair formatted table containing the response (biological distance or dissimilarity), predictors, and weights as required for fitting Generalized Dissimilarity Models.

Examples

```
##table data, species and environmental
load(system.file("./data/gdm.RData", package="gdm"))
sppData <- gdmExpData[, c(1,2,13,14)]
envTab <- gdmExpData[, c(2:ncol(gdmExpData))]

##environmental raster data
##commented out to reduce example run time
#rastFile <- system.file("./extdata/stackedVars.grd", package="gdm")
#envRast <- stack(rastFile)

#########table type 1
##site-species table without coordinates
testData1a <- reshape2::dcast(sppData, site~species)
##site-species table with coordinates
coords <- unique(sppData[, 2:ncol(sppData)])
testData1b <- merge(testData1a, coords, by="site")
##site-species, table-table
exFormat1a <- formatsitepair(testData1a, 1, siteColumn="site", XColumn="Long", YColumn="Lat",
        predData=envTab)

##site-species, table-raster
##not run
#exFormat1b <- formatsitepair(testData1b, 1, siteColumn="site", XColumn="Long", YColumn="Lat",
#       predData=envRast)

#########table type 2
##site xy spp list, table-table
exFormat2a <- formatsitepair(sppData, 2, XColumn="Long", YColumn="Lat", sppColumn="species",
        siteColumn="site", predData=envTab)
##site xy spp list, table-raster
##commented out to reduce example run time
#exFormat2b <- formatsitepair(sppData, 2, XColumn="Long", YColumn="Lat", sppColumn="species",
#       siteColumn="site", predData=envRast)

#########table type 3
##dissim matrix model
site <- unique(sppData$site)
gdmDissim <- cbind(site, gdmDissim)
exFormat3 <- formatsitepair(gdmDissim, 3, XColumn="Long", YColumn="Lat", predData=envTab,
        siteColumn="site")

#########table type 4
##adds a predictor matrix to an existing site-pair table, in this case, predData needs to be
##filled, but is not actually used
exFormat4 <- formatsitepair(exFormat2a, 4, predData=envTab, siteColumn="site",
        distPreds=list(as.matrix(gdmDissim)))
```