Package 'naturaList'

October 13, 2022

Type Package

Title Classify Occurrences by Confidence Levels in the Species ID

Version 0.5.0

Description Classify occurrence records based on confidence levels of species identification. In addition, implement tools to filter occurrences inside grid cells and to manually check for possibles errors with an interactive shiny application.

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 7.1.2

Imports shiny, shinyWidgets, dplyr, stringr, sp, raster, shinydashboard, leaflet, leaflet.extras, tidytext, magrittr, vegan, fasterize, sf, htmltools, methods, rlang

Suggests knitr, rmarkdown, testthat, tm, rgeos, rgdal, rnaturalearth, lwgeom, shinyLP, maptools

VignetteBuilder knitr

Depends R (>= 2.10)

URL https://github.com/avrodrigues/naturaList

BugReports https://github.com/avrodrigues/naturaList/issues

NeedsCompilation no

Author Arthur Vinicius Rodrigues [aut, cre] (<https://orcid.org/0000-0003-2656-558X>), Gabriel Nakamura [aut] (<https://orcid.org/0000-0002-5144-5312>), Leandro Duarte [aut] (<https://orcid.org/0000-0003-1771-0407>)

Maintainer Arthur Vinicius Rodrigues <rodrigues.arthur.v@gmail.com>

Repository CRAN

Date/Publication 2022-04-20 13:30:02 UTC

R topics documented:

`A.setosa` *Occurrence records of Alsophila setosa downloaded from Global Biodiversity Information Facility (GBIF).*

Description

A GBIF raw dataset containing 508 occurrence records for the tree fern Alsophila setosa.

Usage

```
A.setosa
```

Format

A data frame with 508 rows and 45 variables

Source

GBIF.org (08 July 2019) GBIF Occurrence Download doi: 10.15468/dl.6jesg0

BR *Brazil boundary*

Description

A spatial polygon with the Brazil boundaries

Usage

```
BR
```

Format

A 'SpatialPolygonsDataFrame' with 1 feature

classify_occ *Classify occurrence records in levels of confidence in species identification*

Description

Classifies occurrence records in levels of confidence in species identification

Usage

```
classify_occ(
  occ,
  spec = NULL,
  na.rm.coords = TRUE,
  crit.levels = c("det_by_spec", "not_spec_name", "image", "sci_collection",
    "field_obs", "no_criteria_met"),
  ignore.det.names = NULL,
  spec.ambiguity = "not.spec",
  institution.code = "institutionCode",
  collection.code = "collectionCode",
  catalog.number = "catalogNumber",
  year = "year",
  date.identified = "dateIdentified",
  species = "species",
  identified.by = "identifiedBy",
  decimal.latitude = "decimalLatitude",
  decimal.longitude = "decimalLongitude",
  basis.of.record = "basisOfRecord",
  media.type = "mediaType",
  occurrence.id = "occurrenceID",
```

```
  institution.source,
  year.event,
  scientific.name,
  determined.by,
  latitude,
  longitude,
  basis.of.rec,
  occ.id
)
```

Arguments

occ	data frame with occurrence records information.
spec	data frame with specialists' names. See details.
na.rm.coords	logical. If TRUE, remove occurrences with NA in decimal.latitude or decimal.longitude
crit.levels	character. Vector with levels of confidence in decreasing order. The criteria allowed are det_by_spec, not_spec_name, image, sci_collection, field_obs, no_criteria_met. See details.
ignore.det.names	character vector indicating strings in identified.by that should be ignored as a taxonomist. See details.
spec.ambiguity	character. Indicates how to deal with ambiguity in specialists names. not.spec solve ambiguity by classifying the identification as done by a non-specialist;is.spec assumes the identification was done by a specialist; manual.check enables the user to manually check all ambiguous names. Default is not.spec.
institution.code	column name of occ with the name (or acronym) in use by the institution having custody of the object(s) or information referred to in the record.
collection.code	column name of occ with The name, acronym, code, or initials identifying the collection or data set from which the record was derived.
catalog.number	column name of occ with an identifier (preferably unique) for the record within the data set or collection.
year	Column name of occ the four-digit year in which the Event occurred, according to the Common Era Calendar.
date.identified	Column name of occ with the date on which the subject was determined as representing the Taxon.
species	column name of occ with the species names.
identified.by	column name of occ with the name of who determined the species.
decimal.latitude	column name of occ latitude in decimal degrees.
decimal.longitude	column name of occ longitude in decimal degrees.
basis.of.record	column name with the specific nature of the data record. See details.

media.type	column name of occ with the media type of recording. See details.
occurrence.id	column name of occ with link or code for the occurrence record. See in Darwin Core Format
institution.source	deprecated, use institution.code instead.
year.event	deprecated, use year instead.
scientific.name	deprecated, use species instead.
determined.by	deprecated, use identified.by instead
latitude	deprecated, use decimal.latitude instead
longitude	deprecated, use decimal.longitude instead
basis.of.rec	deprecated, use basis.of.record instead.
occ.id	deprecated, use occurrence.id instead

Details

spec data frame must have columns separating LastName, Name and Abbrev. See create_spec_df function for a easy way to produce this data frame.

When ignore.det.name = NULL (default), the function ignores strings with "RRC ID Flag", "NA", "", "-" and "_". When a character vector is provided, the function adds the default strings to the provided character vector and ignore all these strings as being a name of a taxonomist.

The function classifies the occurrence records in six levels of confidence in species identification. The six levels are:

- det_by_spec - when the identification was made by a specialists which is present in the list of specialists provided in the spec argument;
- not_spec_name - when the identification was made by a name who is not a specialist name provide in spec;
- image - the occurrence have not name of a identifier, but present an image associated;
- sci_collection - the occurrence have not name of a identifier, but preserved in a scientific collection;
- field_obs - the occurrence have not name of a identifier, but it was identified in field observation;
- no_criteria_met - no other criteria was met.

The (decreasing) order of the levels in the character vector determines the classification level order.

basis.of.record is a character vector with one of the following types of record: PRESERVED_SPECIMEN, PreservedSpecimen, HUMAN_OBSERVATION or HumanObservation, as in GBIF data 'basisOfRecord'.

media.type uses the same pattern as GBIF mediaType column, indicating the existence of an associated image with stillImage.

Value

The occ data frame plus the classification of each record in a new column, named naturaList_levels.

Author(s)

Arthur V. Rodrigues

See Also

speciaLists

Examples

```
data("A.setosa")
data("speciaLists")
occ.class <- classify_occ(A.setosa, speciaLists)
```

---

clean_eval *Evaluate the cleaning of occurrences records*

Description

This function compare the area occupied by a species before and after pass through the cleaning procedure according to the chosen level of filter. The comparison can be made by measuring area in the geographical and in the environmental space

Usage

```
clean_eval(
  occ.cl,
  geo.space,
  env.space = NULL,
  level.filter = c("1_det_by_spec"),
  r,
  species = "species",
  decimal.longitude = "decimalLongitude",
  decimal.latitude = "decimalLatitude",
  scientific.name,
  longitude,
  latitude
)
```

Arguments

occ.cl	data frame with occurrence records information already classified by classify_occ function.
geo.space	a SpatialPolygons* or sf object defining the geographical space

env.space	a SpatialPolygons* or sf object defining the environmental space. Use the define_env_space for create this object. By default env.space = NULL, hence do not evaluate the cleaning in the environmental space.
level.filter	a character vector including the levels in 'naturaList_levels' column which filter the occurrence data set.
r	a raster with 2 layers representing the environmental variables. If env.space = NULL, it could be a single layer raster, from which the cell size and extent are extracted to produce the composition matrix.
species	column name of occ.cl with the species names.
decimal.longitude	column name of occ.cl longitude in decimal degrees.
decimal.latitude	column name of occ.cl latitude in decimal degrees.
scientific.name	deprecated, use species instead.
longitude	deprecated, use decimal.longitude instead
latitude	deprecated, use decimal.latitude instead

Value

a list in which:

area data frame remaining area after cleaning proportional to the area before cleaning. The values vary from 0 to 1. Column named r.geo.area is the remaining area for all species in the geographic space and the r.env.area in the environmental space.

comp data frame with composition of species in sites (cells from raster layers) before cleaning (comp$comp$BC) and after cleaning (comp$comp$AC). The number of rows is equal the number of cells in r, and number of columns is equal to the number of species in the occ.cl.

rich data frame with a single column with the richness of each site

site.coords data frame with site's coordinates. It facilitates to built raster layers from results using rasterFromXYZ

See Also

define_env_space

Examples

```
## Not run:

library(sp)
library(raster)


data("speciaLists") # list of specialists
data("cyathea.br") # occurrence dataset
```

```
# classify
occ.cl <- classify_occ(cyathea.br, speciaLists)

# delimit the geographic space
# land area
data("BR")


# Transform occurrence data in SpatialPointsDataFrame
spdf.occ.cl <- sp::SpatialPoints(occ.cl[, c("decimalLongitude", "decimalLatitude")])


# load climate data
data("r.temp.prec") # mean temperature and annual precipitation
df.temp.prec <- raster::as.data.frame(r.temp.prec)

### Define the environmental space for analysis
# this function will create a boundary of available environmental space,
# analogous to the continent boundary in the geographical space
env.space <- define_env_space(df.temp.prec, buffer.size = 0.05)

# filter by year to be consistent with the environmental data
occ.class.1970 <-  occ.cl %>%
  dplyr::filter(year >= 1970)

### run the evaluation
cl.eval <- clean_eval(occ.class.1970,
                      env.space = env.space,
                      geo.space = BR,
                      r = r.temp.prec)

#area results
head(cl.eval$area)


### richness maps
## it makes sense if there are more than one species
rich.before.clean <- raster::rasterFromXYZ(cbind(cl.eval$site.coords,
                                                 cl.eval$rich$rich.BC))
rich.after.clean <- raster::rasterFromXYZ(cbind(cl.eval$site.coords,
                                                cl.eval$rich$rich.AC))

raster::plot(rich.before.clean)
raster::plot(rich.after.clean)

### species area map
comp.bc <- as.data.frame(cl.eval$comp$comp.BC)
comp.ac <- as.data.frame(cl.eval$comp$comp.AC)

c.villosa.bc <- raster::rasterFromXYZ(cbind(cl.eval$site.coords,
                                            comp.bc$`Cyathea villosa`))
c.villosa.ac <- raster::rasterFromXYZ(cbind(cl.eval$site.coords,
```

```
                                        comp.ac$`Cyathea villosa`))

raster::plot(c.villosa.bc)
raster::plot(c.villosa.ac)

## End(Not run)
```

`create_spec_df` *Create specialist data frame from character vector*

Description

Creates a specialist data frame ready for use in `classify_occ` from a character vector containing the specialists names

Usage

```
create_spec_df(spec.char)
```

Arguments

`spec.char`	a character vector with specialist names

Value

a data frame. Columns split the names, surname and abbreviation for the names. If the full name contain any special character, such as accent marks, two lines for that name will be provided, with and without the special characters. See examples.

Examples

```
# Example using Latin accent marks
data(spec_names_ex)

spec_names_ex
create_spec_df(spec_names_ex)
```

cyathea.br *Occurrence records of Cyathea species in Brazil downloaded from Global Biodiversity Information Facility (GBIF).*

Description

A filtered GBIF dataset containing 3851 occurrence records for the fern species from the genus Cyathea in Brazil. We filtered the data after download from GBIF to ensure all occurrences records are from Brazil.

Usage

```
cyathea.br
```

Format

A data frame with 3851 rows and 50 variables

Source

GBIF.org (07 March 2021) GBIF Occurrence Download doi: 10.15468/dl.qrhynv

define_env_space *Define environmental space for species occurrence*

Description

Based on two continuous environmental variables, it defines a bi-dimensional environmental space.

Usage

```
define_env_space(env, buffer.size, plot = TRUE)
```

Arguments

env	matrix or data frame with two columns containing two environmental variables. The variables must be numeric, even for data frames.
buffer.size	numeric value indicating a buffer size around each point which will delimit the environmental geographical border for the occurrence point. See details.
plot	logical. whether to plot the polygon. Default is TRUE.

Details

The environmental variables are standardized by range, which turns the range of each environmental variable from 0 to 1. Then, it is delimited a buffer of size equal to `buffer.size` around each point in this space and a polygon is draw to link these buffers. The function returns the polygon needed to link all points, and the area of the polygon indicates the environmental space based in the variables used.

Value

An object of sfc_POLYGON class

Examples

```
## Not run:
library("raster")

# load climate data
data("r.temp.prec")
env.data <- raster::as.data.frame(r.temp.prec)

define_env_space(env.data, 0.05)

## End(Not run)
```

`get_det_names` *Get the names in the 'identified.by' column*

Description

This function facilitates the search for non-taxonomist strings in the 'identified.by' column of occurrence records data set

Usage

```
get_det_names(
  occ,
  identified.by = "identifiedBy",
  freq = FALSE,
  decreasing = TRUE,
  determined.by
)
```

Arguments

occ	data frame with occurrence records information.
identified.by	column name of occ with the name of who determined the species.
freq	logical. If TRUE output contain the number of times each string is repeated in the identified.by column. Default = FALSE
decreasing	logical. sort strings in decreasing order of frequency. Default = TRUE.
determined.by	deprecated, use identified.by instead.

Value

character vector containing the strings in identified.by column of occ. If freq = TRUE it return a data frame with two columns: 'strings' and 'frequency'.

Examples

```
data("A.setosa")
get_det_names(A.setosa, freq = TRUE)
```

---

grid_filter *Filter the occurrence with most confidence in species identification inside grid cells*

Description

In each grid cell it selects the occurrence with the highest confidence level in species identification made by classify_occ function.

Usage

```
grid_filter(
  occ.cl,
  grid.resolution = c(0.5, 0.5),
  r = NULL,
  institution.code = "institutionCode",
  collection.code = "collectionCode",
  catalog.number = "catalogNumber",
  year = "year",
  date.identified = "dateIdentified",
  species = "species",
  identified.by = "identifiedBy",
  decimal.latitude = "decimalLatitude",
  decimal.longitude = "decimalLongitude",
  basis.of.record = "basisOfRecord",
  media.type = "mediaType",
  occurrence.id = "occurrenceID",
```

```
  institution.source,
  year.event,
  scientific.name,
  determined.by,
  latitude,
  longitude,
  basis.of.rec,
  occ.id
)
```

Arguments

`occ.cl`	data frame with occurrence records information already classified by `classify_occ` function.
`grid.resolution`	numeric vector with width and height of grid cell in decimal degrees.
`r`	raster from which the grid cell resolution is derived.
`institution.code`	column name of `occ.cl` with the name (or acronym) in use by the institution having custody of the object(s) or information referred to in the record.
`collection.code`	column name of `occ.cl` with The name, acronym, code, or initials identifying the collection or data set from which the record was derived.
`catalog.number`	column name of `occ.cl` with an identifier (preferably unique) for the record within the data set or collection.
`year`	Column name of `occ.cl` the four-digit year in which the Event occurred, according to the Common Era Calendar.
`date.identified`	Column name of `occ.cl` with the date on which the subject was determined as representing the Taxon.
`species`	column name of `occ` with the species names.
`identified.by`	column name of `occ.cl` with the name of who determined the species.
`decimal.latitude`	column name of `occ.cl` latitude in decimal degrees.
`decimal.longitude`	column name of `occ.cl` longitude in decimal degrees.
`basis.of.record`	column name with the specific nature of the data record. See details.
`media.type`	column name of `occ.cl` with the media type of recording. See details.
`occurrence.id`	column name of `occ` with link or code for the occurrence record. See in Darwin Core Format
`institution.source`	deprecated, use `institution.code` instead.
`year.event`	deprecated, use `year` instead.

scientific.name	deprecated, use species instead.
determined.by	deprecated, use identified.by instead
latitude	deprecated, use decimal.latitude instead
longitude	deprecated, use decimal.longitude instead
basis.of.rec	deprecated, use basis.of.record instead.
occ.id	deprecated, use occurrence.id instead

Value

Data frame with the same columns of occ.cl.

Author(s)

Arthur V. Rodrigues

See Also

classify_occ

Examples

```
## Not run:

data("A.setosa")
data("speciaLists")

occ.class <- classify_occ(A.setosa, speciaLists)
occ.grid <- grid_filter(occ.class)


## End(Not run)
```

---

map_module *Check the occurrence records in a interactive map module*

Description

Allows to delete occurrence records and to select occurrence points by classification levels or by drawing spatial polygons.

Usage

```
map_module(
  occ.cl,
  action = "clean",
  institution.code = "institutionCode",
  collection.code = "collectionCode",
  catalog.number = "catalogNumber",
  year = "year",
  date.identified = "dateIdentified",
  species = "species",
  identified.by = "identifiedBy",
  decimal.latitude = "decimalLatitude",
  decimal.longitude = "decimalLongitude",
  basis.of.record = "basisOfRecord",
  media.type = "mediaType",
  occurrence.id = "occurrenceID",
  institution.source,
  year.event,
  scientific.name,
  determined.by,
  latitude,
  longitude,
  basis.of.rec,
  occ.id
)
```

Arguments

`occ.cl`	Data frame with occurrence records information already classified by `classify_occ` function.
`action`	a string with ‘"clean"‘ or ‘"flag"‘ which defines the action of ‘map_module‘ function with the occurrence dataset. Default is ‘"clean"‘. If the string is ‘"clean"‘ the dataset returned only the occurrences records selected by the user. If the string is ‘"flag"‘, a column named ‘map_module_flag‘ is added in the output dataset, with tags ‘selected‘ and ‘deleted‘, following the choices of the user in the application.
`institution.code`	column name of `occ` with the name (or acronym) in use by the institution having custody of the object(s) or information referred to in the record.
`collection.code`	column name of `occ` with The name, acronym, code, or initials identifying the collection or data set from which the record was derived.
`catalog.number`	column name of `occ` with an identifier (preferably unique) for the record within the data set or collection.
`year`	Column name of `occ` the four-digit year in which the Event occurred, according to the Common Era Calendar.

`date.identified`
: Column name of `occ` with the date on which the subject was determined as representing the Taxon.

`species`
: column name of `occ` with the species names.

`identified.by`
: column name of `occ` with the name of who determined the species.

`decimal.latitude`
: column name of `occ` latitude in decimal degrees.

`decimal.longitude`
: column name of `occ` longitude in decimal degrees.

`basis.of.record`
: column name with the specific nature of the data record. See details.

`media.type`
: column name of `occ` with the media type of recording. See details.

`occurrence.id`
: column name of `occ` with link or code for the occurrence record. See in Darwin Core Format

`institution.source`
: deprecated, use `institution.code` instead.

`year.event`
: deprecated, use `year` instead.

`scientific.name`
: deprecated, use `species` instead.

`determined.by`
: deprecated, use `identified.by` instead

`latitude`
: deprecated, use `decimal.latitude` instead

`longitude`
: deprecated, use `decimal.longitude` instead

`basis.of.rec`
: deprecated, use `basis.of.record` instead.

`occ.id`
: deprecated, use `occurrence.id` instead

Value

Data frame with the same columns of `occ.cl`.

Author(s)

Arthur V. Rodrigues

See Also

`classify_occ`

Examples

```
## Not run:
data("A.setosa")
data("speciaLists")

occ.class <- classify_occ(A.setosa, speciaLists)
occ.selected <- map_module(occ.class)
occ.selected
```

```
## End(Not run)
```

`r.temp.prec` *Raster of temperature and precipitation*

Description

Raster of Annual Mean Temperature (bio1) and Total Annual Precipitation (bio2). Layers were downloaded from worldclim database and cropped to the extent of `cyathea_br` with a buffer of 100 km.

Usage

```
r.temp.prec
```

Format

A raster with two layers

`speciaLists` *Specialists of ferns and lycophytes of Brazil*

Description

A dataset containing the specialists of ferns and lycophytes of Brazil formatted to be used by `naturaList` package. This data serves as a format example for `spec` argument in `classify_occ`.

Usage

```
speciaLists
```

Format

A data frame with 27 rows and 8 columns:

LastName Last name of the specialist.

Name1 Columns with the names of specialist. Could be repeated as long as needed. In this data Name* was repeated three times.

Name2 Columns with the names of specialist.

Name3 Columns with the names of specialist.

Name4 Columns with the names of specialist.

Abbrev1 Columns with the abbreviation (one character) of the names of specialists. Could be repeated as long as needed. In this data Abbrev* was repeated three times.

Abbrev2 Columns with the abbreviation (one character) of the names of specialists.

Abbrev3 Columns with the abbreviation (one character) of the names of specialists.

Source

The specialists names was derived from the authors of paper: doi: 10.1590/21757860201566410

`spec_names_ex` *Example of specialist names with accent marks*

Description

Example of specialist names with accent marks

Usage

```
spec_names_ex
```

Format

character

Index