Package 'specleanr'

November 25, 2025

Type Package

Title Detecting Environmental Outliers in Data Analysis Pipelines

Version 1.0.0

Description A framework used to detect and handle outliers during data analysis workflows. Outlier detection is a statistical concept with applications in data analysis workflows, highlighting records that are suspiciously high or low. Outlier detection in distribution models was initiated by Chapman (1991) (available at <https://www.researchgate.net/publication/332537800_Quality_control_and_validation_of_point-sourced_environmental_resource_data>), who developed the reverse jackknifing method. The concept was further developed and incorporated into different R packages, including 'flexsdm' (Velazco et al., 2022, <doi:10.1111/2041-210X.13874>) and 'biogeo' (Robertson et al., 2016 <doi:10.1111/ecog.02118>). We compiled various outlier detection methods obtained from the literature, including those elaborated in Dastjerdy et al. (2023) <doi:10.3390/geotechnics3020022> and Liu et al. (2008) <doi:10.1109/ICDM.2008.17>. In this package, we introduced the ensembling aspect, where multiple outlier detection methods are used to flag the record as either an absolute outlier. The concept can also be applied in general data analysis, as well as during the development of species distribution models.

License GPL (>= 3)

Encoding UTF-8

LazyData true

URL https://anthonybasooma.github.io/specleanr/

BugReports https://github.com/AnthonyBasooma/specleanr/issues

RoxygenNote 7.3.2

Suggests dplyr, knitr, rmarkdown, testthat (>= 3.0.0), ggplot2, ggpmisc, tibble, rinat, rvertnet, rgbif, curl, rfishbase (>= 5.0.1), sf, terra, tidytext, scatterplot3d

Config/testthat/edition 3

VignetteBuilder knitr

Imports cluster, dbscan, e1071, isotree, methods, utils, robust, robustbase, usdm, mgcv

Depends R (>= 4.1.0)

NeedsCompilation no

Author Anthony Basooma [aut, cre] (ORCID:
<https://orcid.org/0000-0002-8994-9989>),
Thomas Hein [ctb, fnd, ths] (ORCID:
<https://orcid.org/0000-0002-7767-4607>),
Astrid Schmidt-Kloiber [ctb, fnd, dtc] (ORCID:
<https://orcid.org/0000-0001-8839-5913>),
Merret Buurman [ctb],
Sami Domisch [ctb],
Martin Tschikof [ctb],
Florian Borgwardt [ctb, fnd] (ORCID:
<https://orcid.org/0000-0002-8974-7834>)

Maintainer Anthony Basooma <anthony.basooma@boku.ac.at>

Repository CRAN

Date/Publication 2025-11-25 20:20:02 UTC

Contents

`abdata` *Alburnoides bipunctatus species data from GBIF and iNaturalist*

Description

A `tibble` Data from GBIF (https://www.gbif.org/) and iNaturalist (https://www.inaturalist.org/)

Usage

```
data(abdata)
```

Format

A tibble 2130 rows and 3 columns.

Details

The species data was collated from the Global Biodiversity Information Facility and iNaturalist

Examples

```
data("abdata")
abdata
```

`adjustboxplots` *Adjust the boxplots bounding fences using medcouple to flag suspicious outliers.*

Description

Adjust the boxplots bounding fences using medcouple to flag suspicious outliers.

Usage

```
adjustboxplots(
  data,
  var,
  output = "outlier",
  a = -4,
  b = 3,
  coef = 1.5,
  pc = FALSE,
  pcvar = NULL,
  boot = FALSE
)
```

Arguments

data	dataframe. Dataframe to check for outliers.
var	string. Environmental predictor considered in flagging suspicious outliers.
output	string Either clean: for dataframe with no suspicious outliers or outlier: to return dataframe with only outliers.
a	numeric. Constant for adjusted boxplots.

b	numeric. Constant for adjusted boxplots.
coef	numeric. Constant for adjusted boxplots.
pc	Whether principal component analysis will be computed. Default FALSE
pcvar	Principal component analysis to e used for outlier detection after PCA. Default PC1
boot	Whether bootstrapping will be computed. Default FALSE

Value

dataframe. Dataframe with or with no outliers.

References

Hubert M, Vandervieren E. 2008. An adjusted boxplot for skewed distributions. Computational Statistics and Data Analysis 52:5186-5201.

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                      lat = 'decimalLatitude', lon= 'decimalLongitude',
                       colsp = "scientificName",
                       bbox = db,
                       minpts = 10)

adout <- adjustboxplots(data = refdata[["Thymallus thymallus"]], var = 'bio6', output='outlier')
```

---

bestmethod *Identifies the best method for outlier detection for a single species.*

Description

Identifies the best method for outlier detection for a single species.

Usage

```
bestmethod(
  x,
  sp = NULL,
  threshold = NULL,
  autothreshold = FALSE,
  warn = FALSE,
  verbose = FALSE
)
```

Arguments

x	List of dataframes for each methods used to identify outliers in `multdetect` function.
sp	species name or index if multiple species are considered during outlier detection.
threshold	Maximum value to denote an absolute outlier. The threshold ranges from `0` which indicates a point has not been flagged by any outlier detection method as an `outlier` or `1`, when means the record is an absolute or true outlier sicen it has been identified by all methods. At both extremes, at low threshold values, many records are classified, which may be due to individual method weakness or strength and data distribution. Also, at higher threshold values, the true outliers are retained Fo example, if 10 methods are considered and 9 methods flags a record as an outlier, If a cut off 1 is used, then that particular record is retained. Therefore the `default` cutoff is 0.6 but `autothreshold` can be used to select the appropriate threshold.
autothreshold	Identifies the threshold with mean number of absolute outliers.The search is limited within 0.51 to 1 since thresholds less than are deemed inappropriate for identifying absolute outliers. The autothreshold is used when `threshold` is set to `NULL`.
warn	If `TRUE`, warning on whether absolute outliers obtained at a low threshold is indicated. Default `TRUE`.
verbose	if `TRUE` then messages and warnings will be produced. Default `FALSE`.

Value

best method for identifying outliers.

Examples

```
data("efidata")
data("jdsdata")

matchdata <- match_datasets(datasets = list(jds = jdsdata, efi=efidata),
                            lats = 'lat',
                            lons = 'lon',
                            species = c('speciesname','scientificName'),
```

```
                            date = c('Date', 'sampling_date'),
                            country = c('JDS4_site_ID'))


danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

worldclim <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

rdata <- pred_extract(data = matchdata,
                      raster= worldclim ,
                      lat = 'decimalLatitude',
                      lon= 'decimalLongitude',
                      colsp = 'species',
                      bbox = db,
                      minpts = 10,
                      list=TRUE,
                     merge=FALSE)


out_df <- multidetect(data = rdata, multiple = TRUE,
                      var = 'bio6',
                      output = 'outlier',
                      exclude = c('x','y'),
                      methods = c('zscore', 'adjbox','iqr', 'semiqr','hampel', 'kmeans',
                                  'logboxplot', 'lof','iforest', 'mahal', 'seqfences'))

bmout <- bestmethod(x = out_df, sp= 1, threshold = 0.2)
```

`boots` *To implement bootstrapping procedures. Sampling with replacement.*

Description

To implement bootstrapping procedures. Sampling with replacement.

Usage

```
boots(data, boots, seed, pca)
```

Arguments

`data`	Environmental data
`boots`	Number of bootstraps
`seed`	Random seed to ensure reproduciblity
`pca`	Whether bootstrapping is conducted on data after principal component analysis.

broad_classify *Outlier detection method broad classification.*

Description

Outlier detection method broad classification.

Usage

```
broad_classify(category)
```

Arguments

category	The different outlier categories including mult, uni and ref

Value

vector method broad categories

Examples

```
x <- broad_classify(category = "mult")
```

check.exclude *indicate excluded columns.*

Description

indicate excluded columns.

Usage

```
check.exclude(x, exclude, quiet = TRUE)
```

Arguments

x	dataframe with columns to where the columns are supposed to be removed.
exclude	string or vector column names to be checked if it is in the data.
quiet	TRUE if implementation messages to be shown. Default FALSE.

Value

columns that are not in the dataframe.

checks *Post checks for PCA and bootstrapping*

Description

Post checks for PCA and bootstrapping

Usage

```
checks(y, nboots, th, var)
```

Arguments

y	list of PCA and bootstrapped output.
nboots	Number of bootstrapping
th	threshold for identifying absolute outlier from bootstrapped samples.
var	variable of interest.

check_names *Check species names for inconsistencies*

Description

Check species names for inconsistencies

Usage

```
check_names(
  data,
  colsp = NULL,
  verbose = FALSE,
  pct = 90,
  merge = FALSE,
  sn = FALSE,
  ecosystem = FALSE,
  rm_duplicates = FALSE
)
```

Arguments

data	dataframe. Data frame with species names to checked from FishBase.
colsp	string. A column in the data with the species column names.
verbose	logical. To indicate the merges during checking of names. The default is FALSE not to show whether the species are in Fish base or not found.
pct	numeric. The percentage similarity used to assign a relatively similar name from FishBase if the exact match is absent. Default 90 The higher the values, the higher percentage similarity are considered to replace a species name that is cheeked from Fishbase.
merge	logical. Default is **FALSE**, not to merge the cleaned species column on to the data frame but rather only two columns returned.
sn	logical. Whether to consider synonyms. Default FALSE so accepted names will be considered from FishBase database.
ecosystem	logical Returns whether the species is freshwater, marine , or brackish or a combination for for estuarine loving species.
rm_duplicates	logical. If TRUE, removes all duplicate species names especially when a dataframe is is the output from the function.

Details

The function produces a data set with species names corresponding with Fishase. If synonym is provided in the data set, the function will by defualt return the accepted name. However, if the synoymn is desired, then set the sn parameter to **TRUE**. The function also check for spellings of species names and returns a name that is closer to the one in FishBase with a particular degree of similarity set with pct parameter. pct of 1 indicates the name must 100 The user can iterate with different pct and decide if the return name is right or wrong. This function is not necessary if the species names are clean and also for other taxa.

Value

Data frame or names of corrected or cleaned species names.

See Also

match_datasets for standardizing and binding datasets.

Examples

```
## Not run:

data(jdsdata)

data(efidata)

#step 1. match and bind datasets if more than one datasets

matchdata <- match_datasets(datasets = list(jds = jdsdata, efi = efidata),
```

```
                          lats = 'lat',
                          lons = 'lon',
                          species = c('speciesname','scientificName'),
                          country=c('JDS4_site_ID'),
                          date=c('Date', 'sampling_date'))

#clean species names to produce one dataset.

datafull <- check_names(data= matchdata, colsp='species', pct = 90, merge = TRUE)

data2col <- check_names(data = matchdata, colsp='species', pct = 90) #two columns generated

cleansp_name <- check_names(data= 'slamo trutta', pct=90) #wrong names vs FB suggestion

clean_sp_epithet <- check_names(data = 'Salmo trutta fario') #Salmo trutta will be returned

speciesepithet2 <- check_names(data = 'Salmo trutta lacustris', pct=90)


## End(Not run)
```

---

check_packages *Check for packages to install and respond to use*

Description

Check for packages to install and respond to use

Usage

```
check_packages(pkgs)
```

Arguments

pkgs	list of packages to install

Value

error message for packages to install

`classify_data` *Extract final clean data using either absolute or best method generated outliers.*

Description

Extract final clean data using either absolute or best method generated outliers.

Usage

```
classify_data(
  refdata,
  outliers,
  var_col = NULL,
  threshold = 0.1,
  warn = FALSE,
  verbose = TRUE,
  classify = "med",
  EIF = FALSE
)
```

Arguments

`refdata`	`dataframe`. The reference data for the species used in outlier detection.
`outliers`	`string`. Output from the outlier detection process.
`var_col`	`string`. A parameter to be used if the `data` is a data frame and the user must indicate the column with species names.
`threshold`	`numeric`. Value to consider whether the outlier is an absolute outlier or not.
`warn`	`logical`. If **FALSE**, warning on whether absolute outliers obtained at a low threshold is indicated. Default **TRUE**.
`verbose`	`logical`. Produces messages or not. Default **FALSE**.
`classify`	`string`. Categorize data base on the correlation coefficient manner based on `Akoglu 2018`. For more information check in the details section.
`EIF`	`logical` To calculate the empirical influence function for each value.

Details

Outlier cluster weights were based on statistical classification of coefficients mostly for correlation based on `Akoglu 2018`. They are classified based on three naming standards, namely Dancey & Reidy (Physchology), Quinni piac University (Politics) and Chan YH medicine. All classifications have been used in the function and each affects the data clusters. The default is Chan YH (medicine).

Value

Either a `list` or `dataframe` of cleaned records for multiple species.

References

Akoglu, H. 2018. User's guide to correlation coefficients. - Turk J Emerg Med 18: 91–93.

See Also

search_threshold

Examples

```
data(jdsdata)
data(efidata)
matchdata <- match_datasets(datasets = list(jds = jdsdata, efi = efidata),
                            lats = 'lat',
                            lons = 'lon',
                            species = c('speciesname','scientificName'),
                            country= c('JDS4_site_ID'),
                            date=c('sampling_date', 'Date'))


danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)


worldclim <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

rdata <- pred_extract(data = matchdata,
                      raster= worldclim ,
                      lat = 'decimalLatitude',
                      lon= 'decimalLongitude',
                      colsp = 'species',
                      bbox = db,
                      minpts = 10,
                      list=TRUE,
                      merge=FALSE)


out_df <- multidetect(data = rdata, multiple = TRUE,
                      var = 'bio6',
                      output = 'outlier',
                      exclude = c('x','y'),
                      methods = c('zscore', 'adjbox','iqr', 'semiqr','hampel'))

#extracting use the absolute method for one species

extractabs <- classify_data(refdata = rdata, outliers = out_df)
```

cosine *Cosine similarity index based on (Gautam & Kulkarni 2014; Joy & Renumol 2020)*

Description

Cosine similarity index based on (Gautam & Kulkarni 2014; Joy & Renumol 2020)

Usage

```
cosine(x, sp = NULL, threshold = NULL, warn = FALSE, autothreshold = FALSE)
```

Arguments

x	`datacleaner` class for each methods used to identify outliers in `multidetect` function.
sp	`string`. Species name or index if multiple species are considered during outlier detection.
threshold	`numeric`. Maximum value to denote an absolute outlier. The threshold ranges from `0`, which indicates a point has not been flagged by any outlier detection method as an `outlier`, to `1`, which means the record is an absolute or true outlier since all methods have identified it. At both extremes, many records are classified at low threshold values, which may be due to individual method weakness or strength and data distribution. Also, at higher threshold values, the true outliers are retained. For example, if ten methods are considered and 9 methods flag a record as an outlier, If a cutoff of 1 is used, then that particular record is retained. Therefore, the `default` cutoff is `0.6`, but `autothreshold` can be used to select the appropriate threshold.
warn	`logical`. If **TRUE**, warning on whether absolute outliers obtained at a low threshold is indicated. Default **TRUE**.
autothreshold	`vector`. Identifies the threshold with mean number of absolute outliers.The search is limited within 0.51 to 1 since thresholds less than are deemed inappropriate for identifying absolute outliers. The autothreshold is used when `threshold` is set to `NULL`.

Value

best method for identifying outliers.

Examples

```
data(efidata)

danube <- system.file('extdata/danube.shp.zip', package='specleanr')
```

```
db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package = "specleanr"))

extdf <- pred_extract(data = efidata, raster = wcd,
                    lat = 'decimalLatitude', lon = 'decimalLongitude',
                     colsp = "scientificName",
                     list = TRUE,verbose = FALSE,
                     minpts = 6,merge = FALSE)#basin removed

 #outlier detection
outliersdf <- multidetect(data = extdf, output='outlier', var = 'bio6',
                          exclude = c('x','y'), multiple = TRUE,
                          methods = c('mixediqr', "iqr", "mahal", "iqr", "logboxplot"))

consineout <- cosine(x = outliersdf, sp= 1, threshold = 0.2)#
```

`datacleaner-class` *Outlier detection class for multiple methods*

Description

Outlier detection class for multiple methods

Slots

`result` List of data sets with outliers detected.

`mode` Either ´TRUE´ for multiple species and FALSE for one species.

`varused` The variable used for outlier detection, useful for univariate outlier detection methods.

`out` Either outliers or clean dataset outputted.

`methodsused` The methods used in outlier detection.

`dfname` the dataframe name to aid tracking it during clean data extraction.

`excluded` whether some columns were excluded during outlier detection. useful for multivariate methods where coordinates are removed from the data.

`pc` parameters for principal component analysis.

`bootstrap` parameters for bootstrapping for small data sets.

`nboots` the number of bootstraps during bootstrapping.

`pcvariable` variable to be considered during PCA.

`pcretained` the number data columns retained. the default is 3.

`maxrecords` the maximum number of records used for bootstrapping.

`distboxplot` *Distribution boxplot*

Description

Distribution boxplot

Usage

```
distboxplot(
  data,
  var,
  output,
  p1 = 0.025,
  p2 = 0.975,
  boot = FALSE,
  pc = FALSE,
  pcvar = NULL
)
```

Arguments

`data`	Dataframe or vector where to check outliers.
`var`	Variable to be used for outlier detection if **data** is not a vector file.
`output`	Either **clean**: for clean data output without outliers; **outliers**: for outlier data frame or vectors.
`p1, p2`	Different pvalues for outlier detection `Schwertman et al. 2004`).
`boot`	Whether bootstrapping will be computed. Default `FALSE`
`pc`	Whether principal component analysis will be computed. Default `FALSE`
`pcvar`	Principal component analysis to e used for outlier detection after PCA. Default `PC1`

Value

Either clean or outliers.

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))
```

```
refdata <- pred_extract(data = efidata, raster= wcd ,
                          lat = 'decimalLatitude', lon= 'decimalLongitude',
                          colsp = "scientificName",
                          bbox = db,
                          minpts = 10)

 bxout <- distboxplot(data = refdata[["Thymallus thymallus"]], var = 'bio6', output='outlier')
```

---

ecological_ranges *Check for environmental outliers using species optimal ranges.*

Description

Check for environmental outliers using species optimal ranges.

Usage

```
ecological_ranges(
  data,
  var,
  output = "outlier",
  species = NULL,
 optimumSettings = list(optdf = NULL, optspcol = NULL, mincol = NULL, maxcol = NULL,
    ecoparam = NULL, direction = NULL),
  minval = NULL,
  maxval = NULL,
  lat = NULL,
  lon = NULL,
  ecoparam = NULL,
  direction = NULL,
  pct = 80,
  checkfishbase = FALSE,
  mode = NULL,
  warn = TRUE
)
```

Arguments

data	Dataframe with environmental predictors for a species or multiple species.
var	Environmental parameter considered in flagging suspicious outliers.
output	output Either clean: for dataframe with no suspicious outliers or outlier: to retrun dataframe with only outliers.
species	The species should be indicated if the minimum minval and maximum values maxval are provided.

optimumSettings
: A list of optimal parameters are provided mostly when multiple species are examined.
  - optdf: is the dataframe with species optimal values (min, max, or ecoparam). This dataset can be generated from literature for different species.
  - optspcol: Is the column with species names in the optdf dataset.
  - mincol: Is the column name in the optdf with minimum values.
  - maxcol: Is the column name in the optdf with maximum values.
  - ecoparam: If in the optdf the minimum and maximum values are not found, then the the column with ecoparam should be provided.
  - direction: If ecoparam is provided in the optdf, then column for direction should be provided.

minval, maxval
: Minimum and maximum values (ranges) for a particular that are used to flag out values outside the ranges.

lat, lon
: If the checkfishbase and mode are set, then the columns for latitude longitude should be provided.

ecoparam
: This parameter is used only when the lower bound (minimum) and upper bound maximum or ranges are absent. For example, if only minimum value is present for a particular species, then ecoparam is set and the direction is provided whether lower, greater, equal, less/equal or greater/equal the ecoparam value provided.

direction
: This indicates if the provided ecological threshold ecoparam or ranges is greater than greater, less than less, equal equal, less or equal le or greater or equal ge. Íf the minimum and maximum values are known, then the ecoparam and direction should not be used.

pct
: The percentage similarity of the species name provided by the user and the one in FishBase. Only fish species names are checked with Fishbase but other taxa can be checked using taxize package.

checkfishbase
: Either TRUE to check for both temperatures temp and latitudinal or geographical ranges geo. If the checkfishbase is set to TRUE then the mode parameter must be set to either geo or temp. This function applies for only fish species.

mode
: Either geo or temp for latitudinal ranges or temperature ranges respectively. See thermal_ranges or geo_ranges on how to obtain the data.

warn
: Either TRUE to return warning messages or FALSE for no warning messages. the defualt is FALSE:

Value

Dataframe with or with no outliers.

Examples

```
## Not run:

data("efidata")
data("jdsdata")
```

```
datafinal <- match_datasets(datasets = list(jds = jdsdata, efi=efidata),
                            lats = 'lat',
                            lons = 'lon',
                            species = c('speciesname','scientificName'),
                            date = c('Date', 'sampling_date'),
                            country = c('JDS4_site_ID'))

efidata <- check_names(data = datafinal, colsp='species', pct=90, merge=TRUE)

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                         lat = 'decimalLatitude', lon= 'decimalLongitude',
                        colsp = "scientificName",
                       bbox = db,
                        minpts = 10)

saldata <- refdata[["Thymallus thymallus"]]
#1. checking the annual maean temperature (bio1) are within the ranges in FishBase
salmotherange <- thermal_ranges(x = "Salmo trutta")

sdatatemp <- ecological_ranges(data = saldata, var = 'bio1', species = "Salmo trutta",
                               checkfishbase = TRUE, mode = 'temp', output = 'outlier')
#zero record no outliers
#====
#2. geographical ranges: latitude longitude
#geo ranges in fishbase
salgeorange <- geo_ranges(data = "Salmo trutta")
sdatageo <- ecological_ranges(data = saldata, lat = 'y', lon = 'x', output = 'outlier',
                              species = "Salmo trutta",
                              checkfishbase = TRUE, mode = 'geo')
#3. GENERAL LITERATURE RANGES
#======
#1. when the min and and max are provided
#multiple FALSE SHOULD BE SET
#3.1: If only the minimum value is present: assuming minimum temperature is 6, varible: bio1
#direction less than 6.0 is outlier and greater is not
sdata <- ecological_ranges(data = saldata, ecoparam = 6.0, var = 'bio1',
                           direction = 'greater' )
#3.2
sdata2 <- ecological_ranges(data = saldata, var = 'bio1', minval = 2,
                            maxval = 24, species = "Salmo trutta" )

#4. Multiple TRUE
#the optimal parameters should be provided in a dataframe format with min max, or ecoparam
#4.1 optimal dataset

optdata <- data.frame(species= c("Salmo trutta", "Abramis brama"),
                      mintemp = c(6, 1.6),maxtemp = c(20, 21),
```

```
                      meantemp = c(8.5, 10.4), #ecoparam
                      direction = c('greater', 'greater'))

#parameter used is annual mean temperature (WORLDCLIM)
#provide the column with species names in the environment dataset
#set optimal list parameter
#
# #optimal parameters
sdata3 <- ecological_ranges(data = saldata, species = 'Salmo trutta',
                            var = 'bio1', output = "outlier",
                            optimumSettings = list(optdf = optdata,maxcol = "maxtemp",
                                                   mincol ="mintemp",optspcol = "species"))
#
#
#only one ecological parameter (ecoparam is provided) and direction
sdata4 <- ecological_ranges(data = saldata, species = 'Salmo trutta', var = 'bio1',
                            output = "outlier",
                            optimumSettings = list(optdf = optdata,
                                                   ecoparam = "meantemp",
                                                   optspcol = "species",
                                                   direction= "direction"))

## End(Not run)
```

---

efidata *EFIPLUS data used to develop ecological sensitivity parameters for riverine species in European streams and rivers.*

Description

A `tibble`

Usage

```
data(efidata)
```

Format

A `tibble` 99 rows and 23 columns.

Details

BQEs sensitivity to global/climate change in European rivers: implications for reference conditions and pressure-impact-recovery chains (Logez et al. 2012). An extract has been made for usage in this package but for more information write to ihg@boku.ac.at

References

Logez M, Belliard J, Melcher A, Kremser H, Pletterbauer F, Schmutz S, Gorges G, Delaigue O, Pont D. 2012. Deliverable D5.1-3: BQEs sensitivity to global/climate change in European rivers: implications for reference conditions and pressure-impact-recovery chains.

Examples

```
data("efidata")
efidata
```

`eif` *Computes the empirical influence function for each values in the dataset*

Description

Computes the empirical influence function for each values in the dataset

Usage

```
eif(x, var)
```

Arguments

`x`	Outlier checked data
`var`	variable of interest

`extentvalues` *To check for a bounding box*

Description

To check for a bounding box

Usage

```
extentvalues(x, par = NULL)
```

Arguments

`x`	raster, shapefile or list of bounding box values.
`par`	indicate the database being queried to handing the issues of bounding box settings.

Value

extent values from raster, shapefile and bounding box

extractMethods *List of outlier detection methods implemented in this package.*

Description

List of outlier detection methods implemented in this package.

Usage

```
extractMethods()
```

Value

List of methods

Examples

```
extractMethods()
```

extractoutliers *Extract outliers for a one species*

Description

Extract outliers for a one species

Usage

```
extractoutliers(x, sp = NULL)
```

Arguments

x	`list`. Outlier outputs for both single and multiple species.
sp	`string`. Species name or index in the list from datacleaner output. NULL for a single species

Value

`data frame` Outliers for each method

Examples

```
data(efidata)

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package = "specleanr"))


extdf <- pred_extract(data = efidata, raster = wcd,
                      lat = 'decimalLatitude', lon = 'decimalLongitude',
                      colsp = 'scientificName',
                      list = TRUE,verbose = FALSE,
                    minpts = 6,merge = FALSE)

#outlier detection

outliersdf <- multidetect(data = extdf, output='outlier', var = 'bio6',
                          exclude = c('x','y'), multiple = TRUE,
                          methods = c('mixediqr', "iqr", "iqr", "logboxplot"),
                          silence_true_errors = FALSE,
                         verbose = FALSE, sdm = TRUE)

extoutlier <- extractoutliers(x=outliersdf, sp = 3)
```

`extract_clean_data` *Extract final clean data using either absolute or best method generated outliers.*

Description

Extract final clean data using either absolute or best method generated outliers.

Usage

```
extract_clean_data(
  refdata,
  outliers,
  mode = "abs",
  var_col = NULL,
  threshold = NULL,
  warn = FALSE,
  verbose = FALSE,
```

```
  autothreshold = FALSE,
  pabs = 0.1,
  loess = FALSE,
  outlier_to_NA = FALSE,
  cutoff = 0.6
)
```

Arguments

refdata	dataframe. The reference data for the species used in outlier detection.
outliers	string. Output from the outlier detection process.
mode	character. Either abs to use absolute outliers to filter data or best to outliers from best method.
var_col	string. A parameter to be used if the data is a data frame and the user must indicate the column wih species names.
threshold	numeric. Value to consider whether the outlier is an absolute outlier or not.
warn	logical. If **FALSE**, warning on whether absolute outliers obtained at a low threshold is indicated. Default **TRUE**.
verbose	logical. Produces messages or not. Default **FALSE**.
autothreshold	vector. Identifies the threshold with mean number of absolute outliers.The search is limited within 0.51 to 1 since thresholds less than are deemed inappropriate for identifying absolute outliers. The autothreshold is used when threshold is set to NULL.
pabs	numeric. Percentage of outliers allowed to be extracted from the data. If best is used to extract outliers and the pabs is exceeded, the absolute outliers are removed instead. This because some records in the best methods are repeated and they will likely to remove true values as outliers.
loess	logical. Set to TRUE to use loess threshold optimization to extract clean data.
outlier_to_NA	logical If TRUE a clean dataset will have outliers replaced with NAs. This parameter is experimented to ouput dataframe when multiple variables of concerns are considered during outlier detection. ###param multiple TRUE for multiple species and FALSE for single species considered during outlier detection.
cutoff	numeric. Ranging from 0.5 to 0.8 indicating the cutoff to initiate the LOESS model to optimize the identification of absolute outliers.

Value

Either a list or dataframe of cleaned records for multiple species.

See Also

search_threshold

Examples

```
data(jdsdata)
data(efidata)
matchdata <- match_datasets(datasets = list(jds = jdsdata, efi = efidata),
                            lats = 'lat',
                            lons = 'lon',
                            species = c('speciesname','scientificName'),
                            country= c('JDS4_site_ID'),
                            date=c('sampling_date', 'Date'))


danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)


worldclim <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

rdata <- pred_extract(data = matchdata,
                      raster= worldclim ,
                      lat = 'decimalLatitude',
                      lon= 'decimalLongitude',
                      colsp = 'species',
                      bbox = db,
                      minpts = 10,
                      list=TRUE,
                      merge=FALSE)


out_df <- multidetect(data = rdata, multiple = TRUE,
                      var = 'bio6',
                      output = 'outlier',
                      exclude = c('x','y'),
                      methods = c('zscore', 'adjbox','iqr', 'semiqr','hampel'))

#extracting use the absolute method for one species

extractabs <- extract_clean_data(refdata = rdata, outliers = out_df,
                                 mode = 'abs', threshold = 0.6,
                                 autothreshold = FALSE)

bestmout_bm <- extract_clean_data(refdata = rdata, outliers = out_df,
                                  mode = 'best', threshold = 0.6,
                                 autothreshold = FALSE)
```

---

geo_ranges *Checks for geographic ranges from FishBase*

Description

Checks for geographic ranges from FishBase

Usage

```
geo_ranges(
  data,
  colsp = NULL,
  verbose = FALSE,
  pct = 90,
  sn = FALSE,
  warn = FALSE,
  synonym = fishbase(tables = "synonym"),
  ranges = fishbase(tables = "ranges")
)
```

Arguments

data	Dataframe or vector to retrieve ranges from FishBase.
colsp	Column with species names from the data set.
verbose	TRUE and messages will show. Default FALSE:
pct	The percentage similarity of species names during standardization from FishBase.
sn	TRUE and synonyms will be generated and not accepted ones. Default is FALSE, where species accepted names will be produced.
warn	FALSE, not to generate warnings and TRUE for warnings. Default is FALSE:
synonym	A standard database for species synonym names from FishBase. See FishBase for more information.
ranges	A standard database for ecological ranges from FishBase. See FishBase for more information.

Value

Dataframe with geographical corrected ranges for species from FishBase.

Examples

```
## Not run:

gr <- geo_ranges(data= "Lates niloticus")


## End(Not run)
```

getdata *Download species records from online database.*

Description

Download species records from online database.

Usage

```
getdata(
  data,
  colsp = NULL,
  extent = NULL,
  db = c("gbif", "vertnet", "inat"),
  gbiflim = 50000,
  vertlim = 1000,
  inatlim = 3000,
  verbose = FALSE,
  warn = FALSE,
  pct = 80,
  sn = FALSE,
  ...
)
```

Arguments

data	dataframe, list, vector, string. data to retrieve records from online databases (GBIF, VertNET, and iNaturalist).
colsp	string. A variable of species names. Provided if data is a data frame, so not required for lists and vector.
extent	vector or sf. Bounding box to limit the download of records within a particular area. Otherwise all records from the GBIF will be downloaded. These can be provided in two forms, either a shapefile (sf) class accepted or provide a list of named xmin, ymin, xmax, and ymax in this particular order.
db	vector. The different databases allowed including 'gbif', 'vertnet', and 'inat'.
gbiflim	integer. Limits on the records from the Global Biodiversity Information Platform
vertlim	integer. Limits on the records from VertNET.
inatlim	integer. Limits on the records from iNaturalist database.
verbose	logical. **TRUE** if detailed messages should be indicated and **FALSE** if download messages are not needed. Default **TRUE**.
warn	logical. To indicate if warning messages should be shown. Default FALSE.

pct	numeric. The percentage similarity used to assign a relatively similar name from FishBase if the exact match is absent. Default 90 The higher the values, the higher percentage similarity are considered to replace a species name that is cheeked from Fishbase.
sn	logical. Whether to consider synonyms. Default FALSE so accepted names will be considered from FishBase database.
...	More function for species data download can be used. See rgbif::occ_data for more information, rinat::get_inat_obs, and rvertnet::searchbyterm.

Details

Note always check the validity of the species name with standard database FishBase or World Register of Marine Species. If the records are more than 50000 in GBIF, and extent can be provide to limit the download.

Value

Lists of species records from online databases

Examples

```
## Not run:

gbdata <- getdata(data = 'Gymnocephalus baloni', gbiflim = 100, inatlim = 100, vertlim = 100)

#Get for two species
sp_records <- getdata(data=c('Gymnocephalus baloni', 'Hucho hucho'),
                            gbiflim = 100,
                            inatlim = 100,
                            vertlim = 100)
#for only two databases
sp_records_2db <- getdata(data=c('Gymnocephalus baloni', 'Hucho hucho'),
                           db= c('gbif','inat'),
                            gbiflim = 100,
                            inatlim = 100,
                            vertlim = 100)


## End(Not run)
```

---

getdiff *get dataframe from the large dataframe.*

Description

get dataframe from the large dataframe.

Usage

```
getdiff(x, y, full = FALSE)
```

Arguments

x	Small dataset
y	Large dataset for intersection
full	Whether the whole column names are checked or not. Default FALSE where only the first column is considered. if FALSE; then the returned columns may be few or more if the considered column has less or more similar rows across the two data sets.

Value

Data to extracted from large dataset.

Examples

```
x = data.frame(id=c(1,2,3,4,5),  name=c('a','b','c', 'd','e'))

y=data.frame(id=c(1,2,3,4,7,6,5), tens=c(10,29,37,46,58, 34, 44),
                 name=c('a','b','c','d','e', 'f','g'))
```

---

ggenvironmentalspace *Title Plotting to show the quality controlled data in environmental space.*

Description

Title Plotting to show the quality controlled data in environmental space.

Usage

```
ggenvironmentalspace(
  qcdata,
  xvar = NULL,
  yvar = NULL,
  zvar = NULL,
  labelvar = NULL,
  type = "2D",
  xlab = NULL,
  ylab = NULL,
  zlab = NULL,
  ncol = 2,
  nrow = 2,
  scalecolor = "viridis",
```

```
  colorvalues = "auto",
  legend_position = "right",
  legend_inside = NULL,
  pointsize = 1,
  themebackground = "bw",
  fontsize = 13,
  legtitle = "blank",
  ggxangle = 1,
  xhjust = 0.5,
  xvjust = 1,
  main = NULL,
  pch = "auto",
  lpos3d = "left",
  cexsym = NULL
)
```

Arguments

qcdata	dataframe Data output from quality controlled function extract_clean_data and classify_data.
xvar	string The variable to be on the x-axis.
yvar	string The variable to be on the y-axis.
zvar	string The variable to be on the z-axis only if the 3D plot type is selected..
labelvar	string Column name in the quality controlled data that has the labels. This applies is the 3D plot is selected.
type	string Its 1D, 2D for a two dimensional ggplot2 graph or 3D for a 3-dimensional graph for multivariate data.
xlab, ylab, zlab	string x-axis, y-axis, and z-axis label.
ncol, nrow	integer If number of groups are greater than 1, then number of rows and columns can be set. Check ggplot2 facet parameters on how the columns are set.
scalecolor	string The scale color themes supported are grey, manual, viridis. If manual is selected, then the colorvalues should be provided for the different colors for each data label.
colorvalues	If manual is selected, then the colorvalues should be provided for the different colors for each data label. If 3D is selected and colorvalues is not auto, then colors should determined.
legend_position	string Its either bottom, top or inside. If the inside is selected then the vector with graph coordinates should be provided to avoid the legend overlap with the graph contents.
legend_inside	vector If the inside for legend position is selected then the vector with graph coordinates should be provided to avoid the legend overlap with the graph contents.
pointsize	decimal The size of the points.

themebackground
: string Either classic, bw or gray to set the plot theme. This is based on ggplot2.

fontsize
: integer Indicates the sizes of fonts for the whole graph.

legtitle
: string Either blank or TRUE to set the legend title for the 2D plot.

ggxangle
: integer Indicates the angle of the x-axis text. The dafualt is 45 but depend on the data.

xhjust
: numeric Indicates the distance of the x-axis text from the x-axis line in a vertical direction.

xvjust
: numeric Indicates the distance of the x-axis text from the x-axis line in a horizontal direction.

main
: string Plot title

pch
: string Either auto: the point characters will be automatically set or different pch are set.

lpos3d
: string Indicates the legend position for the 3D graph. bottom, left, and right are accepted.

cexsym
: numeric The size of pch in the 3D plot.

Value

If "2D" or "1D" is the selected type, then a ggplot2 graph will be the output and a "3D" type will return a scatterplot3D plot.

---

ggoutlieraccum *Identify if enough methods are selected for the outlier detection.*

Description

Identify if enough methods are selected for the outlier detection.

Usage

```
ggoutlieraccum(
  x,
  boots = 5,
  select = NULL,
  ncol = 3,
  linecolor = "blue",
  seed = 1134,
  sci = FALSE,
  xlab = "Number of methods",
  ylab = "Number of outliers",
  scales = "free"
)
```

Arguments

x	datacleaner. The output from the outlier detection in multidetect function.
boots	interger. The number of bootstraps to sample the outliers obtained during outlier detection process. Start from a lower number such as 10 and increase serially to get a smoother curve. High bootstrap may lead to crashing the Generalized Additive Model used to fit the bootstraps and cumulative number of outliers.
select	vector. If more than 10 groups are considered, then the at least should be seclected to hvae meaningful visualization.
ncol	integer. Number of columns if the groups are greater 4, to allow effective vizualisation.
linecolor	string A parameter to indicate the color of the lines. The default is 'purple'.
seed	integer To fix the random sampling during bootstrapping.
sci	logical. If sci is TRUE, then the species names will be italised otherwise normal names will displayed. Default FALSE
xlab, ylab	string. inherited from ggplot2 to changes x and y axis texts.
scales	string Define if the x oy y axis will be shared or free. check ggplot2 for details.

Value

ggplot2 output with cumulative number of outliers and number of methods used.

---

ggoutliers *Visualize the outliers identified by each method*

Description

Visualize the outliers identified by each method

Usage

```
ggoutliers(x, select = NULL, color = "purple", desc = TRUE, ncol = 2, nrow = 2)
```

Arguments

x	. the datacleaner object
select	vect. Enter selected groups to be displayed especially if they are greater than 10. For example if the species are more than 10, the plot will be done in batches.
color	string. Color of the bars. Default is grey.
desc	logical To either arrange the bars in ascending or descending order.
ncol, nrow	integer If number of groups are greater than 1, then number of rows and columns can be set. Check ggplot2 facet parameters on how the columns are set.

Value

ggplot object indicating outlier detection methods and number of outlier flagged.

`hamming` *Identify best outlier detection method using Hamming distance.*

Description

Identify best outlier detection method using Hamming distance.

Usage

```
hamming(x, sp = NULL, threshold = NULL, warn = FALSE, autothreshold = FALSE)
```

Arguments

`x`	`datacleaner` class for each methods used to identify outliers in `multidetect` function.
`sp`	`string`. Species name or index if multiple species are considered during outlier detection.
`threshold`	`numeric`. Maximum value to denote an absolute outlier. The threshold ranges from `0`, which indicates a point has not been flagged by any outlier detection method as an `outlier`, to `1`, which means the record is an absolute or true outlier since all methods have identified it. At both extremes, many records are classified at low threshold values, which may be due to individual method weakness or strength and data distribution. Also, at higher threshold values, the true outliers are retained. For example, if ten methods are considered and 9 methods flag a record as an outlier, If a cutoff of 1 is used, then that particular record is retained. Therefore, the `default` cutoff is `0.6`, but `autothreshold` can be used to select the appropriate threshold.
`warn`	`logical`. If **TRUE**, warning on whether absolute outliers obtained at a low threshold is indicated. Default **TRUE**.
`autothreshold`	`vector`. Identifies the threshold with mean number of absolute outliers.The search is limited within 0.51 to 1 since thresholds less than are deemed inappropriate for identifying absolute outliers. The autothreshold is used when `threshold` is set to `NULL`.

Value

best method based on hamming distance

Examples

```
data(efidata)

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package = "specleanr"))

extdf <- pred_extract(data = efidata, raster = wcd,
                    lat = 'decimalLatitude', lon = 'decimalLongitude',
                     colsp = "scientificName",
                     list = TRUE,verbose = FALSE,
                     minpts = 6,merge = FALSE)#basin removed

 #outlier detection

outliersdf <- multidetect(data = extdf, output='outlier', var = 'bio6',
                          exclude = c('x','y'), multiple = TRUE,
                          methods = c('mixediqr', "iqr", "mahal", "iqr", "logboxplot"))

hamout <- hamming(x = outliersdf, sp= 1, threshold = 0.2)#
```

`hampel` *Flag suspicious outliers based on the Hampel filter method..*

Description

Flag suspicious outliers based on the Hampel filter method..

Usage

```
hampel(data, var, output, x = 3, pc = FALSE, pcvar = NULL, boot = FALSE)
```

Arguments

`data`	Data frame to check for outliers
`var`	Environmental parameter considered in flagging suspicious outliers
`output`	Either clean: for dataframe with no suspicious outliers or outlier: to retrun dataframe with only outliers
`x`	A constant to create a fence or boundary to detect outliers.
`pc`	Whether principal component analysis will be computed. Default `FALSE`
`pcvar`	Principal component analysis to e used for outlier detection after PCA. Default `PC1`
`boot`	Whether bootstrapping will be computed. Default `FALSE`

Details

The Hampel filter method is a robust decision-based filter that considers the median and MAD. Outliers lies beyond

$$[x - *lmbda * MAD; x + lmbda * MAD]$$

and lmbda of 3 was considered (Pearson et al. 2016).

Value

Data frame with or with no outliers.

References

Pearson Ronald, Neuvo Y, Astola J, Gabbouj M. 2016. The Class of Generalized Hampel Filters. 2546-2550 2015 23rd European Signal Processing Conference (EUSIPCO).

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')
db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                          lat = 'decimalLatitude',
                          lon= 'decimalLongitude',
                          colsp = "scientificName",
                          bbox = db,
                          minpts = 10)

 hampout <- hampel(data = refdata[["Thymallus thymallus"]], var = 'bio6', output='outlier')
```

---

handle_true_errors *Catch errors during methods implementation.*

Description

Catch errors during methods implementation.

Usage

```
handle_true_errors(
  func,
  fname = NULL,
  spname = NULL,
  verbose = FALSE,
  warn = FALSE,
  silence_true_errors = TRUE
)
```

Arguments

func	Outlier detection function
fname	function name for messaging or warning identification.
spname	species name being handled
verbose	whether to return messages or not. Default FALSE.
warn	whether to return warning or not. Default TRUE.
silence_true_errors	show execution errors and therefore for multiple species the code will break if one of the methods fails to execute.

Value

Handle errors

---

interquartile *Computes interquartile range to flag environmental outliers*

Description

Computes interquartile range to flag environmental outliers

Usage

```
interquartile(
  data,
  var,
  output,
  x = 1.5,
  pc = FALSE,
  pcvar = NULL,
  boot = FALSE
)
```

Arguments

data	Dataframe to check for outliers
var	Variable considered in flagging suspicious outliers
output	Either clean: for dataframe with no suspicious outliers or outlier: to retrun dataframe with only outliers.
x	A constant to create a fence or boundary to detect outliers.
pc	Whether principal component analysis will be computed. Default FALSE
pcvar	Principal component analysis to e used for outlier detection after PCA. Default PC1
boot	Whether bootstrapping will be computed. Default FALSE

Details

Interquartile range (IQR) uses quantiles that are resistant to outliers compared to mean and standard deviation (Seo 2006). Records were considered as mild outliers if they fell outside the lower and upper bounding fences [Q1 (lower quantile) -1.5*IQR (Interquartile range); Q3 (upper quantile) +1.5*IQR] respectively (Rousseeuw & Hubert 2011). Extreme outliers were also considered if they fell outside \[Q1-3*IQR, Q3+3*IQR\] (García-Roselló et al. 2014). However, using the interquartile range assumes uniform lower and upper bounding fences, which is not robust to highly skewed data (Hubert & Vandervieren 2008).

Value

Dataframe with or with no outliers.

References

Rousseeuw PJ, Hubert M. 2011. Robust statistics for outlier detection. Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery 1:73-79.

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd , lat = 'decimalLatitude',
                          lon= 'decimalLongitude',
                          colsp = "scientificName",
                          bbox = db,
                          minpts = 10)

 iqrout <- interquartile(data = refdata[["Thymallus thymallus"]], var = 'bio6', output='outlier')
```

`isoforest` *Identify outliers using isolation forest model.*

Description

Identify outliers using isolation forest model.

Usage

```
isoforest(
  data,
  size,
  cutoff = 0.5,
  output,
  exclude = NULL,
  pc = FALSE,
  boot = FALSE,
  pcvar = NULL,
  var
)
```

Arguments

`data`	Dataframe of environmental variables extracted from where the species was recorded present or absent.
`size`	Proportion of data to be used in training isolation forest n´model. It ranges form 0.1 (fewer data selected) to 1 to all data used in training isolation model.
`cutoff`	Cut to select where the record was an outlier or not.
`output`	Either clean: for a data set with no outliers or outlier: to output a dataframe with outliers. Default is 0.5.
`exclude`	Exclude variables that should not be considered in the fitting the one class model, for example x and y columns or latitude/longitude or any column that the user doesn't want to consider.
`pc`	Whether principal component analysis will be computed. Default `FALSE`
`boot`	Whether bootstrapping will be computed. Default `FALSE`
`pcvar`	Principal component analysis to e used for outlier detection after PCA. Default `PC1`
`var`	The variable of concern, which is vital for univariate outlier detection methods

Value

Dataframe with or with no outliers.

References

1. Liu FeiT, Ting KaiM, Zhou Z-H. 2008. Isolation Forest. Pages 413–422 In 2008 Eighth IEEE International Conference on Data Mining. Available from https://ieeexplore.ieee.org/abstract/document/4781136 (accessed November 18, 2023).

Examples

```
data("efidata")
danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                        lat = 'decimalLatitude',
                        lon= 'decimalLongitude',
                        colsp = "scientificName",
                       bbox = db,
                        minpts = 10)

iosd <- isoforest(data = refdata[["Thymallus thymallus"]], size = 0.7,  output='outlier',
                  exclude = c("x", "y"))
```

---

jaccard *Identifies the best outlier detection method using Jaccard coefficient.*

Description

Identifies the best outlier detection method using Jaccard coefficient.

Usage

```
jaccard(x, sp = NULL, threshold = NULL, warn = FALSE, autothreshold = FALSE)
```

Arguments

x	`datacleaner` class for each methods used to identify outliers in `multidetect` function.
sp	`string`. Species name or index if multiple species are considered during outlier detection.

threshold	numeric. Maximum value to denote an absolute outlier. The threshold ranges from 0, which indicates a point has not been flagged by any outlier detection method as an outlier, to 1, which means the record is an absolute or true outlier since all methods have identified it. At both extremes, many records are classified at low threshold values, which may be due to individual method weakness or strength and data distribution. Also, at higher threshold values, the true outliers are retained. For example, if ten methods are considered and 9 methods flag a record as an outlier, If a cutoff of 1 is used, then that particular record is retained. Therefore, the default cutoff is 0.6, but autothreshold can be used to select the appropriate threshold.
warn	logical. If **TRUE**, warning on whether absolute outliers obtained at a low threshold is indicated. Default **TRUE**.
autothreshold	vector. Identifies the threshold with mean number of absolute outliers.The search is limited within 0.51 to 1 since thresholds less than are deemed inappropriate for identifying absolute outliers. The autothreshold is used when threshold is set to NULL.

Value

string best method for identifying outliers.

Examples

```
data(efidata)

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package = "specleanr"))

extdf <- pred_extract(data = efidata, raster = wcd,
                    lat = 'decimalLatitude', lon = 'decimalLongitude',
                     colsp = "scientificName",
                     list = TRUE,verbose = FALSE,
                     minpts = 6,merge = FALSE)#basin removed

 #outlier detection

outliersdf <- multidetect(data = extdf, output='outlier', var = 'bio6',
                          exclude = c('x','y'), multiple = TRUE,
                          methods = c('mixediqr', "iqr", "mahal", "iqr", "logboxplot"))

jaccardout <- jaccard(x = outliersdf, sp= 1, threshold = 0.2)#
```

jdsdata *Joint Danube Survey Data*

Description

A `tibble` Data on a five year periodic data collection within the Danube River Basin. For more information, please visit https://www.danubesurvey.org/jds4/about

Usage

```
data(jdsdata)
```

Format

A `tibble` 98 rows and 24 columns.

Details

Species ecological parameters such as ecological ranges both native and alien

References

https://www.danubesurvey.org/jds4/about

Examples

```
data("jdsdata")
jdsdata
```

jknife *Identifies outliers using Reverse Jackknifing method based on Chapman et al., (2005).*

Description

Identifies outliers using Reverse Jackknifing method based on Chapman et al., (2005).

Usage

```
jknife(
  data,
  var,
  output = "outlier",
  mode = "soft",
  pc = FALSE,
  pcvar = NULL,
  boot = FALSE
)
```

Arguments

data	Dataframe to check for outliers
var	Variable considered in flagging suspicious outliers.
output	Either clean: for data frame with no suspicious outliers or outlier: to return data frame with only outliers
mode	Either robust, if a robust mode is used which uses median instead of mean and median absolute deviation from median or mad instead of standard deviation.
pc	Whether principal component analysis will be computed. Default FALSE
pcvar	Principal component analysis to e used for outlier detection after PCA. Default PC1
boot	Whether bootstrapping will be computed. Default FALSE

Details

Reverse jackknifing was specifically developed to detect error climate profiles (Chapman 1991, 1999). The method has been applied in detecting outliers in environmental data (García-Roselló et al. 2014; Robertson et al. 2016) and incorporated in the DIVAS-GIS software (Hijmans et al. 2001).

Value

Data frame with or with no outliers.

References

1. Chapman AD. 1991. Quality control and validation of environmental resource data in Data Quality and Standards. Pages 1-23. Canberra. Available from https://www.researchgate.net/publication/332537824.
2. Chapman AD. 1999. Quality Control and Validation of Point-Sourced Environmental Resource Data. eds. . Chelsea,. Pages 409-418 in Lowell K, Jaton A, editors. Spatial accuracy assessment: Land information uncertainty in natural resources, 1st edition. MI: Ann Arbor Press., Chelsea.

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                        lat = 'decimalLatitude',
                        lon= 'decimalLongitude',
                        colsp = "scientificName",
                        bbox = db,
                        minpts = 10)

jkout <- jknife(data = refdata[["Thymallus thymallus"]], var = 'bio6', output='outlier')
```

---

kdat *Sequential fences constants*

Description

A `tibble` data with k constants for sequential fences method.

Usage

```
data(kdat)
```

Format

A `tibble` 101 rows and 2 columns.

Details

k constants fro flagging outliers with several chnages in the fences.

References

Schwertman NC, de Silva R. 2007. Identifying outliers with sequential fences. Computational Statistics and Data Analysis 51:3800–3810.

Examples

```
data("kdat")
kdat
```

`logboxplot` *Log boxplot based for outlier detection.*

Description

Log boxplot based for outlier detection.

Usage

```
logboxplot(data, var, output, x = 1.5, pc = FALSE, pcvar = NULL, boot = FALSE)
```

Arguments

`data`	Dataframe or vector where to check outliers.
`var`	Variable to be used for outlier detection if **data** is not in a vector format.
`output`	Either **clean**: for clean data output without outliers; **outliers**: for outlier data frame or vectors.
`x`	The constant for creating lower and upper fences. Extreme is 3, but default is 1.5.
`pc`	Whether principal component analysis will be computed. Default `FALSE`
`pcvar`	Principal component analysis to e used for outlier detection after PCA. Default `PC1`
`boot`	Whether bootstrapping will be computed. Default `FALSE`

Details

The loxplot for outlier detection **Barbato et al. (2011)** modifies the the interquartile range method to detect outlier but considering the sample sizes while indicating the fences (lower and upper fences).

$$lowerfence = [Q1 - 1.5 * IQR[1 + 0.1 * log(n/10)]$$

$$upperfence = [Q3 + 1.5 * IQR[1 + 0.1 * log(n/10)]$$

Where; Q1 is the lower quantile and Q3 is the upper quantile. The method consider the sample size in setting the fences, to address the weakness of the interquartile range method *(Tukey, 1977)*. However. similar to IQR method for flagging outlier, log boxplot modification is affected by data skewness and which can be address using distboxplot, `seqfences`, `mixediqr` and `semiIQR`.

Value

Dataframe with our without outliers depending on the output.

clean Data without outliers.

outlier Data with outliers.

References

Barbato G, Barini EM, Genta G, Levi R. 2011. Features and performance of some outlier detection methods. Journal of Applied Statistics 38:2133-2149

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                        lat = 'decimalLatitude', lon= 'decimalLongitude',
                          colsp = "scientificName",
                          bbox = db,
                          minpts = 10)

 logout <- logboxplot(data = refdata[["Thymallus thymallus"]], var = 'bio6', output='outlier')
```

`mahal` *Flags outliers based on Mahalanobis distance matrix for all records.*

Description

Flags outliers based on Mahalanobis distance matrix for all records.

Usage

```
mahal(
  data,
  exclude = NULL,
  output = "outlier",
  mode = "soft",
  pdf = 0.95,
  tol = 1e-20,
  pc = FALSE,
  boot = FALSE,
  var,
  pcvar = NULL
)
```

Arguments

data	dataframe. Dataframe to check for outliers or extract the clean data.
exclude	vector or string Variables that should not be considered in the executing the Mahalanobis distance matrix. These can be coordinates such as latitude/longitude or any column that the user doesn't want to consider.
output	string Either clean for a data set with no outliers or outlier to output a data frame with outliers.
mode	string Either robust, if a robust mode is used which uses auto estimator to instead of mean. Default mode is soft.
pdf	numeric chisqure probability distribution value used for flagging outliers (Leys et al. 2018). Default is 0.95.
tol	numeric tolernce value when the inverse calculation are too small. Default 1e-20.
pc	Whether principal component analysis will be computed. Default FALSE
boot	Whether bootstrapping will be computed. Default FALSE
var	The variable of concern, which is vital for univariate outlier detection methods
pcvar	Principal component analysis to e used for outlier detection after PCA. Default PC1

Value

Either clean or outliers dataset

References

Leys C, Klein O, Dominicy Y, Ley C. 2018. Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. Journal of Experimental Social Psychology 74:150-156.

Examples

```
data("efidata")



danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                        lat = 'decimalLatitude',
                        lon= 'decimalLongitude',
                        colsp = "scientificName",
                       bbox = db,
```

```
                        minpts = 10)

#outliers
outliers <- mahal(data = refdata[["Thymallus thymallus"]], exclude = c("x", "y"),
                        output='outlier')
```

`match.argc` *Customized match function*

Description

Customized match function

Usage

```
match.argc(x, choices, quiet = TRUE)
```

Arguments

`x`	The category with words to match
`choices`	The different options or choices in a particular category that are allowed.
`quiet`	Default `FALSE` not to return messages.

Value

choices

`match_datasets` *Data harmonizing for offline data based on Darwin Core terms .*

Description

Data harmonizing for offline data based on Darwin Core terms .

Usage

```
match_datasets(
  datasets,
  country = NULL,
  lats = NULL,
  lons = NULL,
  species = NULL,
  date = NULL,
  verbose = FALSE
)
```

Arguments

datasets	List of offline or online data to be merge. Each offline data set should be given a specific name for identification in the match data set.
country	Indicate the **country** column names as they appear in the data sets to be merged.
lats	Match the column names for latitude for each data set to be matched. The default latitude name is **decimalLatitude**. So, indicate the latitude name as it is referenced in all data sets to be matched.
lons	Match the column names for latitude for each data set to be match. The default longitude name is **decimalLongitude**. So, indicate the longitude name as it is referenced in all data sets to be match.
species	Indicate the species columns as they appear in the data sets to be matched. The default is **species**, so if the data set doesn't have species as the column name for scientific species names names, indicate the column name here.
date	Indicate the **date** column names as they appear in the data sets to be matched.
verbose	Messages during data matching. Default **FALSE**

Details

If a data set being matched has standard columns, namely decimalLatitude, decimalLatutide, and species, then they are not indicated while matching. Otherwise all column names with varying names for the 5 parameters should be indicated.

Value

Harmonized data set with standardized column names foe species names, latitude, longitude, country and dates.

References

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, et al. (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715. https://doi.org/10.1371/journal.pone.0029715.

Examples

```
data(jdsdata)

data(efidata)

matchdfs <- match_datasets(datasets = list(jds = jdsdata, efi = efidata),
                           lats = 'lat',
                           lons = 'lon',
                           species = c('speciesname','scientificName'),
                           country=c('JDS4_site_ID'),
                           date=c('Date', 'sampling_date'))
```

`medianrule` *Median rule method*

Description

Median rule method

Usage

```
medianrule(data, var, output, x = 2.3, pc = FALSE, pcvar = NULL, boot = FALSE)
```

Arguments

`data`	Dataframe or vector where to check outliers.
`var`	Variable to be used for outlier detection if **data** is not a vector file.
`output`	Either **clean**: for clean data output without outliers; **outliers**: for outlier data frame or vectors.
`x`	A constant for flagging outliers.
`pc`	Whether principal component analysis will be computed. Default `FALSE`
`pcvar`	Principal component analysis to e used for outlier detection after PCA. Default `PC1`
`boot`	Whether bootstrapping will be computed. Default `FALSE`

Value

Either clean or outliers.

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                          lat = 'decimalLatitude',
                          lon= 'decimalLongitude',
                          colsp = "scientificName",
                          bbox = db,
                          minpts = 10)

 medout <- medianrule(data = refdata[["Thymallus thymallus"]], var = 'bio6', output='outlier')
```

mixediqr *Mixed Interquartile range and semiInterquartile range* Walker et al., 2018

Description

Mixed Interquartile range and semiInterquartile range Walker et al., 2018

Usage

```
mixediqr(data, var, output, x = 3, pc = FALSE, pcvar = NULL, boot = FALSE)
```

Arguments

data	Dataframe or vector where to check outliers.
var	Variable to be used for outlier detection if **data** is not a vector file.
output	Either **clean**: for clean data output without outliers; **outliers**: for outlier data frame or vectors.
x	A constant for flagging outliers Walker et al., 2018).
pc	Whether principal component analysis will be computed. Default FALSE
pcvar	Principal component analysis to e used for outlier detection after PCA. Default PC1
boot	Whether bootstrapping will be computed. Default FALSE

Value

Either clean our outliers

References

Walker ML, Dovoedo YH, Chakraborti S, Hilton CW. 2018. An Improved Boxplot for Univariate Data. American Statistician 72:348-353. American Statistical Association.

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                            lat = 'decimalLatitude', lon= 'decimalLongitude',
```

```
                            colsp = "scientificName",
                            bbox = db,
                            minpts = 10)

logout <- mixediqr(data = refdata[["Thymallus thymallus"]], var = 'bio6', output='outlier')
```

---

mth *mth datasets with constant at each confidence interval levels.*

Description

A `tibble` The data consist the

Usage

```
data(mth)
```

Format

A `tibble` 7 rows and 9 columns.

Details

The data is extracted from (Schwertman & de Silva 2007).

References

Schwertman NC, de Silva R. 2007. Identifying outliers with sequential fences. Computational Statistics and Data Analysis 51:3800–3810.

Examples

```
data("mth")
mth
```

`multiabsolute` *Identifies absolute outliers for multiple species.*

Description

Identifies absolute outliers for multiple species.

Usage

```
multiabsolute(
  x,
  threshold = NULL,
  props = FALSE,
  warn = FALSE,
  autothreshold = FALSE
)
```

Arguments

`x`	`datacleaner` class for each methods used to identify outliers in `multidetect` function.
`threshold`	`numeric`. Maximum value to denote an absolute outlier. The threshold ranges from `0`, which indicates a point has not been flagged by any outlier detection method as an `outlier`, to `1`, which means the record is an absolute or true outlier since all methods have identified it. At both extremes, many records are classified at low threshold values, which may be due to individual method weakness or strength and data distribution. Also, at higher threshold values, the true outliers are retained. For example, if ten methods are considered and 9 methods flag a record as an outlier, If a cutoff of 1 is used, then that particular record is retained. Therefore, the `default` cutoff is `0.6`, but `autothreshold` can be used to select the appropriate threshold.
`props`	`dataframe`. To output the proportional absoluteness for each outlier.
`warn`	`logical`. If **TRUE**, warning on whether absolute outliers obtained at a low threshold is indicated. Default **TRUE**.
`autothreshold`	`vector`. Identifies the threshold with mean number of absolute outliers.The search is limited within 0.51 to 1 since thresholds less than are deemed inappropriate for identifying absolute outliers. The autothreshold is used when `threshold` is set to `NULL`.

Value

vector or absolute outliers, best outlier detection method or data frame of absolute outliers and their proportions

See Also

`ocindex`

Examples

```
data(efidata)

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package = "specleanr"))

extdf <- pred_extract(data = efidata, raster = wcd,
                      lat = 'decimalLatitude', lon = 'decimalLongitude',
                     colsp = "scientificName",
                     list = TRUE,verbose = FALSE,
                     minpts = 6,merge = FALSE)#basin removed

 #outlier detection
outliersdf <- multidetect(data = extdf, output='outlier', var = 'bio6',
                          exclude = c('x','y'), multiple = TRUE,
                          methods = c('mixediqr', "iqr", "mahal", "iqr", "logboxplot"))

totabs_counts <- multiabsolute(x = outliersdf, threshold = 0.2)
```

`multibestmethod` *Identify best method for outlier removal for multiple species using majority votes.*

Description

Identify best method for outlier removal for multiple species using majority votes.

Usage

```
multibestmethod(
  x,
  threshold = NULL,
  warn = FALSE,
  verbose = FALSE,
  autothreshold = FALSE
)
```

Arguments

`x`	Output from the outlier detection.

threshold	value to consider whether the outlier is an absolute outlier or not.
warn	If **TRUE**, warning on whether absolute outliers obtained at a low threshold is indicated. Default **FALSE**.
verbose	Produce messages on the process or not. Default **FALSE**.
autothreshold	Identifies the threshold with mean number of absolute outliers.The search is limited within 0.51 to 1 since thresholds less than are deemed inappropriate for identifying absolute outliers. The autothreshold is used when threshold is set to NULL.

Value

best method for outlier detection for each species

Examples

```
danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package = "specleanr"))

preddata <- pred_extract(data = efidata, raster = wcd,
                      lat = 'decimalLatitude', lon = 'decimalLongitude',
                     colsp = 'scientificName',
                     list = TRUE,verbose = FALSE,
                     minpts = 6,merge = FALSE)#'basin removed

 #outlier detection

outliersdf <- multidetect(data = preddata, multiple = TRUE,
                      var = 'bio6',
                      output = 'outlier',
                      exclude = c('x','y'),
                      methods = c('zscore', 'adjbox','iqr', 'semiqr','hampel', 'kmeans',
                                  'logboxplot', 'lof','iforest', 'mahal', 'seqfences'))

multbm <- multibestmethod(x = outliersdf, threshold = 0.2)#
```

---

multidetect *Ensemble multiple outlier detection methods.*

Description

The function allows to ensemble multiple outlier detection methods to ably compare the outliers flagged by each method.

Usage

```
multidetect(
  data,
  var,
  select = NULL,
  output = "outlier",
  exclude = NULL,
  multiple,
  var_col = NULL,
 optpar = list(optdf = NULL, ecoparam = NULL, optspcol = NULL, direction = NULL, maxcol
  = NULL, mincol = NULL, maxval = NULL, minval = NULL, checkfishbase = FALSE, mode =
    NULL, lat = NULL, lon = NULL, pct = 80, warn = FALSE),
  kmpar = list(k = 6, method = "silhouette", mode = "soft"),
  ifpar = list(cutoff = 0.5, size = 0.7),
  mahalpar = list(mode = "soft"),
  jkpar = list(mode = "soft"),
  zpar = list(type = "mild", mode = "soft"),
  gloshpar = list(k = 3, metric = "manhattan", mode = "soft"),
  knnpar = list(metric = "manhattan", mode = "soft"),
  lofpar = list(metric = "manhattan", mode = "soft", minPts = 10),
  methods,
 bootSettings = list(run = FALSE, nb = 5, maxrecords = 30, seed = 1135, th = 0.6),
  pc = list(exec = FALSE, npc = 2, q = TRUE, pcvar = "PC1"),
  verbose = FALSE,
  spname = NULL,
  warn = FALSE,
  missingness = 0.1,
  silence_true_errors = TRUE,
  sdm = TRUE,
  na.inform = FALSE
)
```

Arguments

data	dataframe or list. Data sets for multiple or single species after of extraction of environment predictors.
var	character. A variable to check for outliers especially the one with directly affects species distribution such as maximum temperature of the coldest month for bioclimatic variables (IUCN Standards and Petitions Committee, 2022)) or stream power index for hydromorphological parameters (Logez et al., 2012). This parameter is necessary for the univariate outlier detection methods such as Z-score.
select	vector The columns that will be used in outlier detection. Make sure only numeric columns are accepted.
output	character. Either clean: for a data set with no outliers, or outlier: to output a dataframe with outliers. Default outlier.

exclude — vector. Exclude variables that should not be considered in the fitting the one class model, for example x and y columns or latitude/longitude or any column that the user doesn't want to consider.

multiple — logical. If the multiple species are considered, then multiple must be set to TRUE and FALSE for single species.

var_col — string. A column with species names if dataset for species is a dataframe not a list. See pred_extract for extracting environmental data.

optpar — list. Parameters for species optimal ranges like temperatures ranges. For details check ecological_ranges.

kmpar — list. Parameters for k-means clustering like method and number of clusters for tuning. For details, check xkmeans.

ifpar — list. Isolation forest parameter settings. Parameters of the isolation model that are required include the **cutoff** to be used for denoting outliers. It ranges from 0 to 1 but Default 0.5. Also, the **size** of data partitioning for training should be determined. For more details check (Liu et al. 2008)

mahalpar — list. Parameters for Malahanobis distance which includes varying the mode of output mahal.

jkpar — list. Parameters for reverse jackknifing mainly the mode used. For details jknife.

zpar — list. Parameters for z-score such as mode and x parameter. For details zscore

gloshpar — list. Parameters for global local outlier score from hierarchies such as distance metric used. For details xglosh.

knnpar — list. Parameters for varying the distance matrix such as Euclidean or Manhattan distance. For details xknn

lofpar — list. Parameters for local outlier factor such as the distance matrix and mode of method implementation such as robust and soft mode. For details xlof.

methods — vector. Outlier detection methods considered. Use extractMethods to get outlier detection methods implemented in this package.

bootSettings — list. A list of parameters to implement bootstrapping mostly for records below 30. For details checks boots.

pc — list. A list of parameters to implement principal component analysis for dimesnion reduction. For details checks pca.

verbose — logical. whether to return messages or not. Default FALSE.

spname — string. species name being handled.

warn — logical. Whether to return warning or not. Default TRUE.

missingness — numeric. Allowed missing values in a column to allow a user decide whether to remove the individual columns or rows from the data sets. Default 0.1. Therefore, if a column has more than 10% missing values, then it will be removed from the dataset rather than the rows.

silence_true_errors — logical. Show execution errors and therefore for multiple species the code will break if one of the methods fails to execute.

sdm	logical If the user sets TRUE, strict data checks will be done including removing all non-numeric columns from the datasets before identification of outliers. If set to FALSE non numeric columns will be left in the data but the variable of concern will checked if its numeric. Also, only univariate methods are allowed. Check broad_classify for the broad categories of the methods allowed.
na.inform	logical Inform on the NAs removed in executing general datasets. Default FALSE.

Details

This function computes different outlier detection methods including univariate, multivariate and species ecological ranges to enables seamless comparison and similarities in the outliers detected by each method. This can be done for multiple species or a single species in a dataframe or lists or dataframes and thereafter the outliers can be extracted using the extract_clean_data function.

Value

A list of outliers or clean dataset of datacleaner class. The different attributes are associated with the datacleaner class from multidetect function.

- result: dataframe. list of dataframes with the outliers flagged by each method.
- mode: logical. Indicating whether it was multiple TRUE or FALSE.
- varused: character. Indicating the variable used for the univariate outlier detection methods.
- out: character. Whether outliers where indicated by the user or no outlier data.
- methodsused: vector. The different methods used the outlier detection process.
- dfname: character. The dataset name for the species records.
- exclude: vector. The columns which were excluded during outlier detection, if any.

References

1. IUCN Standards and Petitions Committee. (2022). THE IUCN RED LIST OF THREATENED SPECIESTM Guidelines for Using the IUCN Red List Categories and Criteria Prepared by the Standards and Petitions Committee of the IUCN Species Survival Commission. https://www.iucnredlist.org/documents/RedListGuidelines.pdf.
2. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In 2008 eighth ieee international conference on data mining (pp. 413-422). IEEE.

Examples

```
#' #====
#1. Mult detect for general data analysis using iris data
#===
# the outliers are introduced for testing purposes
irisdata1 <- iris

#introduce outlier data and NAs
```

```
rowsOutNA1 <- data.frame(x= c(344, NA,NA, NA),
                         x2 = c(34, 45, 544, NA),
                         x3= c(584, 5, 554, NA),
                         x4 = c(575, 4554,474, NA),
                         x5 =c('setosa', 'setosa', 'setosa', "setosa"))

colnames(rowsOutNA1) <- colnames(irisdata1)

dfinal <- rbind(irisdata1, rowsOutNA1)

#===========

setosadf <- dfinal[dfinal$Species%in%"setosa",c("Sepal.Width", 'Species')]

setosa_outlier_detection <- multidetect(data = setosadf,
                                        var = 'Sepal.Width',
                                        multiple = FALSE, #'one species
                                        methods = c("adjbox", "iqr", "hampel","jknife",
                                                    "seqfences", "mixediqr",
                                                    "distboxplot", "semiqr",
                                                   "zscore", "logboxplot", "medianrule"),
                                        silence_true_errors = FALSE,
                                        missingness = 0.1,
                                        sdm = FALSE,
                                        na.inform = TRUE)
#======
#2.all species
#=====
multspp_outlier_detection <- multidetect(data = dfinal,
                                         var = 'Sepal.Width',
                                        multiple = TRUE, #'for multiple species or groups
                                         var_col = "Species",
                                         methods = c("adjbox", "iqr", "hampel","jknife",
                                                     "seqfences", "mixediqr",
                                                     "distboxplot", "semiqr",
                                                    "zscore", "logboxplot", "medianrule"),
                                         silence_true_errors = FALSE,
                                         missingness = 0.1,
                                         sdm = FALSE,
                                         na.inform = TRUE)

ggoutliers(multspp_outlier_detection)


#======
#3. Multidetect for environmental data
#======
#'Species data
data("abdata")

#area of interest
danube <- system.file('extdata/danube.shp.zip', package='specleanr')
```

```
db <- sf::st_read(danube, quiet=TRUE)

worldclim <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

abpred <- pred_extract(data = abdata,
                       raster= worldclim ,
                       lat = 'decimalLatitude',
                       lon= 'decimalLongitude',
                       colsp = 'species',
                       bbox = db,
                       minpts = 10,
                       list=TRUE,
                       merge=FALSE)


about_df <- multidetect(data = abpred, multiple = FALSE,
                       var = 'bio6',
                       output = 'outlier',
                       exclude = c('x','y'),
                       methods = c('zscore', 'adjbox','iqr', 'semiqr','hampel', 'kmeans',
                                   'logboxplot', 'lof','iforest', 'mahal', 'seqfences'))

ggoutliers(about_df)


#==========
#4. For mulitple species in species distribution models
#======
data("efidata")
data("jdsdata")

matchdata <- match_datasets(datasets = list(jds = jdsdata, efi=efidata),
                            lats = 'lat',
                            lons = 'lon',
                            species = c('speciesname','scientificName'),
                            date = c('Date', 'sampling_date'),
                            country = c('JDS4_site_ID'))
#extract data
rdata <- pred_extract(data = matchdata,
                       raster= worldclim ,
                       lat = 'decimalLatitude',
                       lon= 'decimalLongitude',
                       colsp = 'species',
                       bbox = db,
                       minpts = 10,
                       list=TRUE,
                       merge=FALSE)

#optimal ranges in the multidetect: made up
multspout_df <- multidetect(data = rdata, multiple = TRUE,
                        var = 'bio6',
                        output = 'outlier',
                        exclude = c('x','y'),
```

```
                      methods = c('zscore', 'adjbox','iqr', 'semiqr','hampel', 'kmeans',
                                  'logboxplot', 'lof','iforest', 'mahal', 'seqfences'))

ggoutliers(multspout_df, "Anguilla anguilla")

#====================================
#5. use optimal ranges as a method
#create species ranges
#==================================
#max temperature of "Thymallus thymallus" is made up to make it appear in outliers

optdata <- data.frame(species= c("Phoxinus phoxinus", "Thymallus thymallus"),
                      mintemp = c(6, 1.6),maxtemp = c(20, 8.6),
                      meantemp = c(8.69, 8.4), #'ecoparam
                      direction = c('greater', 'greater'))

ttdata <- rdata["Thymallus thymallus"]

#even if one species, please indicate multiple to TRUE, since its picked from pred_extract function

thymallus_out_ranges <- multidetect(data = ttdata, multiple = TRUE,
                      var = 'bio1',
                      output = 'outlier',
                      exclude = c('x','y'),
                      methods = c('zscore', 'adjbox','iqr', 'semiqr','hampel', 'kmeans',
                         'logboxplot', 'lof','iforest', 'mahal', 'seqfences', 'optimal'),
                      optpar = list(optdf=optdata, optspcol = 'species',
                                    mincol = "mintemp", maxcol = "maxtemp"))

ggoutliers(thymallus_out_ranges)
```

---

ocindex *Identifies absolute outliers and their proportions for a single species.*

Description

Identifies absolute outliers and their proportions for a single species.

Usage

```
ocindex(
  x,
  sp = NULL,
  threshold = NULL,
  absolute = FALSE,
  props = FALSE,
  warn = FALSE,
  autothreshold = FALSE
)
```

Arguments

x	datacleaner class for each methods used to identify outliers in multidetect function.
sp	string. Species name or index if multiple species are considered during outlier detection.
threshold	numeric. Maximum value to denote an absolute outlier. The threshold ranges from 0, which indicates a point has not been flagged by any outlier detection method as an outlier, to 1, which means the record is an absolute or true outlier since all methods have identified it. At both extremes, many records are classified at low threshold values, which may be due to individual method weakness or strength and data distribution. Also, at higher threshold values, the true outliers are retained. For example, if ten methods are considered and 9 methods flag a record as an outlier, If a cutoff of 1 is used, then that particular record is retained. Therefore, the default cutoff is 0.6, but autothreshold can be used to select the appropriate threshold.
absolute	logical. To output absolute outliers for a species.
props	dataframe. To output the proportional absoluteness for each outlier.
warn	logical. If **TRUE**, warning on whether absolute outliers obtained at a low threshold is indicated. Default **TRUE**.
autothreshold	vector. Identifies the threshold with mean number of absolute outliers.The search is limited within 0.51 to 1 since thresholds less than are deemed inappropriate for identifying absolute outliers. The autothreshold is used when threshold is set to NULL.

Value

vector or dataframe of absolute outliers, best outlier detection method or data frame of absolute outliers and their proportions

Examples

```
data(efidata)

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package = "specleanr"))


extdf <- pred_extract(data = efidata, raster = wcd,
                      lat = 'decimalLatitude', lon = 'decimalLongitude',
                     colsp = "scientificName",
                     list = TRUE,verbose = FALSE,
                     minpts = 6,merge = FALSE)#basin removed
```

```
 #outlier detection

outliersdf <- multidetect(data = extdf, output='outlier', var = 'bio6',
                          exclude = c('x','y'), multiple = TRUE,
                          methods = c('mixediqr', "iqr", "mahal", "iqr", "logboxplot"))

ociss <- ocindex(x = outliersdf, sp= 1, threshold = 0.2, absolute = TRUE)#
#No outliers detected in more than two methods
```

---

onesvm *Identify outliers using One Class Support Vector Machines*

Description

Identify outliers using One Class Support Vector Machines

Usage

```
onesvm(
  data,
  kernel = "radial",
  tune = FALSE,
  exclude = NULL,
  output,
  tpar = list(gamma = 1^(-1:1), epislon = seq(0, 1, 0.1), cost = 2^2:4, nu = seq(0.05, 1,
    0.1)),
  boot = FALSE,
  pc = FALSE,
  var,
  pcvar = NULL
)
```

Arguments

data	Dataframe of environmental variables extracted from where the species was recorded present or absent.
kernel	Either radial, linear
tune	To performed a tuned version of one-class svm. High computation requirements needed.
exclude	Exclude variables that should not be considered in the fitting the one class model, for example x and y columns or latitude/longitude or any column that the user doesnot want to consider.
output	Either clean: for a dataset with no outliers or outlier: to output a dataframe with outliers.

tpar	A list of parameters to be varied during tunning from the normal model.
boot	Whether bootstrapping will be computed. Default FALSE
pc	Whether principal component analysis will be computed. Default FALSE
var	The variable of concern, which is vital for univariate outlier detection methods
pcvar	Principal component analysis to e used for outlier detection after PCA. Default PC1

Value

Dataframe with or with no outliers.

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                        lat = 'decimalLatitude',
                        lon= 'decimalLongitude',
                        colsp = "scientificName",
                       bbox = db,
                        minpts = 10)

nedata <- onesvm(data = refdata[["Thymallus thymallus"]], exclude = c("x", "y"), output='outlier')
```

---

optimal_threshold *Optimize threshold for clean data extraction.*

Description

Optimize threshold for clean data extraction.

Usage

```
optimal_threshold(
  refdata,
  outliers,
  var_col = NULL,
  warn = FALSE,
  verbose = FALSE,
```

```
  plotsetting = list(plot = FALSE, group = NULL),
  cutoff = 0.6
)
```

Arguments

refdata	dataframe. Species data frame from precleaned analysis.
outliers	datacleaner. Datacleaner output with outliers flagged in multidetect function.
var_col	string. A column with species names if dataset for species is a dataframe not a list. See pred_extract for extracting environmental data.
warn	logical. If **TRUE**, warning on whether absolute outliers obtained at a low threshold is indicated. Default **TRUE**.
verbose	logical. If true, then messages about the outlier flagging will be displayed.
plotsetting	list. to show plot of loess fitted function with local and global maxima (optimal threshold and clean data). The list had two parameters. 1) plot to indicate the plot and group to provide the plot title.
cutoff	numeric. Ranging from 0.5 to 0.8 indicating the cutoff to initiate the LOESS model to optimize the identification of absolute outliers.

Value

Either a list or dataframe of cleaned records for multiple species.

---

overlap *Identifies best outlier detection method using Overlap coefficient.*

Description

Identifies best outlier detection method using Overlap coefficient.

Usage

```
overlap(x, sp = NULL, threshold = NULL, warn = FALSE, autothreshold = FALSE)
```

Arguments

x	datacleaner class for each methods used to identify outliers in multidetect function.
sp	string. Species name or index if multiple species are considered during outlier detection.

threshold	numeric. Maximum value to denote an absolute outlier. The threshold ranges from 0, which indicates a point has not been flagged by any outlier detection method as an outlier, to 1, which means the record is an absolute or true outlier since all methods have identified it. At both extremes, many records are classified at low threshold values, which may be due to individual method weakness or strength and data distribution. Also, at higher threshold values, the true outliers are retained. For example, if ten methods are considered and 9 methods flag a record as an outlier, If a cutoff of 1 is used, then that particular record is retained. Therefore, the default cutoff is 0.6, but autothreshold can be used to select the appropriate threshold.
warn	logical. If **TRUE**, warning on whether absolute outliers obtained at a low threshold is indicated. Default **TRUE**.
autothreshold	vector. Identifies the threshold with mean number of absolute outliers.The search is limited within 0.51 to 1 since thresholds less than are deemed inappropriate for identifying absolute outliers. The autothreshold is used when threshold is set to NULL.

Value

best method for identifying outliers.

Examples

```
data(efidata)

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package = "specleanr"))

extdf <- pred_extract(data = efidata, raster = wcd,
                    lat = 'decimalLatitude', lon = 'decimalLongitude',
                     colsp = "scientificName",
                     list = TRUE,verbose = FALSE,
                     minpts = 6,merge = FALSE)#basin removed

 #outlier detection

outliersdf <- multidetect(data = extdf, output='outlier', var = 'bio6',
                          exclude = c('x','y'), multiple = TRUE,
                          methods = c('mixediqr', "iqr", "mahal", "iqr", "logboxplot"))

overlapout <- overlap(x = outliersdf, sp= 1, threshold = 0.2)#
```

pca *Implement principal component analysis for dimension reduction*

Description

Implement principal component analysis for dimension reduction

Usage

```
pca(data, npc, q)
```

Arguments

data	Environmental dataframe
npc	Number of principal components to be retained. Default is 2
q	To show the cumulative total variance explained by the npc selected.

pcboot *To package both principal component analysis and bootstrapping.*

Description

To package both principal component analysis and bootstrapping.

Usage

```
pcboot(pb, var, pc, boot, pcvar)
```

Arguments

pb	the principal component or bootstrapped data
var	The variable of concern, which is vital for univariate outlier detection methods
pc	Whether principal component analysis will be computed. Default FALSE
boot	Whether bootstrapping will be computed. Default FALSE
pcvar	Principal component analysis to e used for outlier detection after PCA. Default PC1

pred_extract *Preliminary data cleaning including removing duplicates, records outside a particular basin, and NAs.*

Description

Preliminary data cleaning including removing duplicates, records outside a particular basin, and NAs.

Usage

```
pred_extract(
  data,
  raster,
  lat = NULL,
  lon = NULL,
  bbox = NULL,
  colsp,
  minpts = 10,
  mp = TRUE,
  rm_duplicates = TRUE,
  na.rm = TRUE,
  na.inform = FALSE,
  list = TRUE,
  merge = FALSE,
  verbose = FALSE,
  warn = FALSE,
  coords = FALSE
)
```

Arguments

data	dataframe. Data frame with multiple species or only one species for checking records with no coordinates, duplicates, and check for records that fall on land, sea, country or city centroids, and geographical outliers(Zzika et al., 2022).
raster	raster. Environmental layers from different providers such as WORLDCLIM (), Hydrogaphy90m (), CHELSA, Copernicus ().
lat, lon	coordinates. variable for latitude and longitude column names.
bbox	sf or vector. Object of class 'shapefile' If only a particular basin is considered. Bounding box vector points can also be provided in the form "c(xmin, ymin, xmax, ymax)". xmin is the minimum longitude, ymin is the minimum latitude, xmax is the maximum longitude and xmax is the minimum latitude.
colsp	string. variable already in the data that determine the groups to considered when extracting data.
minpts	numeric. Minimum number of records for the species after removing duplicates and those within a particular basin.

mp	logical. If TRUE, then number of minimum records minpts should be provided to allow dropping groups with less records. This is significant if species distribution are going to be fitted.
rm_duplicates	logical TRUE if the duplicates will removed based species coordinates and names. Default TRUE.
na.rm	logical If TRUE, the missing values will be discarded after data extracted. DEFAULT TRUE.
na.inform	logical If TRUE, the missing values will be discarded after data extracted and message will be returned. DEFAULT FALSE.
list	logical. If TRUE the a list of multiple species data frames will be generated and FALSE for a dataframe of species data sets. Default TRUE
merge	logical. To add the other columns in the species data after data extraction. Default **TRUE**.
verbose	logical. if TRUE message and warnings will be produced. Default TRUE.
warn	logical. indicating to whether to show implementation warning or not. Default FALSE.
coords	logical. If TRUE, the original coordinates are also returned attached on the extracted dataset. Default FALSE.

Value

dataframe or list of precleaned data sets for single or multiple species.

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

danubebasin <- sf::st_read(danube, quiet=TRUE)

#Get environmental data

worldclim <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

referencedata <- pred_extract(data = efidata,
                          raster= worldclim ,
                          lat ="decimalLatitude",
                          lon = 'decimalLongitude',
                          colsp = 'scientificName',
                          bbox = danubebasin,
                          list= TRUE, #list will be generated for all species
                          minpts = 7, merge=TRUE)
```

search_threshold *Determine the threshold using Locally estimated or weighted Scatterplot Smoothing.*

Description

Determine the threshold using Locally estimated or weighted Scatterplot Smoothing.

Usage

```
search_threshold(
  data,
  outliers,
  sp = NULL,
  plotsetting = list(plot = FALSE, group = NULL),
  var_col = NULL,
  warn = FALSE,
  verbose = FALSE,
  cutoff,
  tloss = seq(0.1, 1, 0.1)
)
```

Arguments

data	Dataframe. The reference dataframe were absolute outliers will be removed.
outliers	datacleaner. Datacleaner output with outliers flagged in multidetect function.
sp	string. Species name or index if multiple species are considered during outlier detection.
plotsetting	list. to show plot of loess fitted function with local and global maxima (optimal threshold and clean data). The list had two parameters. 1) plot to indicate the plot and group to provide the plot title.
var_col	string. A column with species names if dataset for species is a dataframe not a list. See pred_extract for extracting environmental data.
warn	logical. If **TRUE**, warning on whether absolute outliers obtained at a low threshold is indicated. Default **TRUE**.
verbose	logical. If true, then messages about the outlier flagging will be displayed.
cutoff	numeric. Ranging from 0.5 to 0.8 indicating the cutoff to initiate the LOESS model to optimize the identification of absolute outliers.
tloss	seqences Indicates the sequence for tuning the the span parameter of the LOESS model.

Value

Returns numeric of most suitable threshold at globalmaxima or localmaxima of the loess smoothing.

semiIQR *Computes semi-interquantile range to flag suspicious outliers*

Description

Computes semi-interquantile range to flag suspicious outliers

Usage

```
semiIQR(data, var, output, x = 3, pc = FALSE, pcvar = NULL, boot = FALSE)
```

Arguments

data	Dataframe to check for outliers
var	Environmental parameter considered in flagging suspicious outliers
output	Either clean: for dataframe with no suspicious outliers or outlier: to retrun dataframe with only outliers
x	A constant to create a fence or boundary to detect outliers.
pc	Whether principal component analysis will be computed. Default FALSE
pcvar	Principal component analysis to e used for outlier detection after PCA. Default PC1
boot	Whether bootstrapping will be computed. Default FALSE

Details

SemiInterquantile Ranges introduced adjusts for whiskers on either side to flag suspicious outliers [Q1 – 3(Q2 (median) - Q1); Q3 + 3(Q3 - Q2)] ((Kimber 1990)). However, SIQR introduced the same constant values for bounding fences for the lower and upper quartiles (Rousseeuw & Hubert 2011), which leads to outlier swamping and masking.

Value

Dataframe with or with no outliers.

References

Kimber AC. 1990. Exploratory Data Analysis for Possibly Censored Data From Skewed Distributions. Page Source: Journal of the Royal Statistical Society. Series C (Applied Statistics).

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')
```

```
db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                          lat = 'decimalLatitude', lon= 'decimalLongitude',
                          colsp = "scientificName",
                          bbox = db,
                          minpts = 10)

 semiout <- semiIQR(data = refdata[["Thymallus thymallus"]], var = 'bio6', output='outlier')
```

`seqfences` *Sequential fences method*

Description

Sequential fences method

Usage

```
seqfences(
  data,
  var,
  output,
  gamma = 0.95,
  mode = "eo",
  pc = FALSE,
  pcvar = NULL,
  boot = FALSE
)
```

Arguments

`data`	Dataframe or vector where to check outliers.
`var`	Variable to be used for outlier detection if **data** is not a vector file.
`output`	Either **clean**: for clean data output without outliers; **outliers**: for outlier data frame or vectors.
`gamma`	`numeric`. the p-values used to classify a record as an outlier. The lower the p-value, the extremeness is the outlier `Schwertman & de Silva 2007`.
`mode`	`string`. Indicates the extremeness of the outlier.
`pc`	Whether principal component analysis will be computed. Default `FALSE`
`pcvar`	Principal component analysis to e used for outlier detection after PCA. Default `PC1`
`boot`	Whether bootstrapping will be computed. Default `FALSE`

Details

Sequential fences is a modification of the TUKEY boxplot, where the data is divided into groups each with its own fences Schwertman & de Silva 2007. The groups can range from 1, which flags mild outliers to 6 for extreme outliers ()

Value

Dataframe or vector with or without outliers

References

1. Schwertman NC, de Silva R. 2007. Identifying outliers with sequential fences. Computational Statistics and Data Analysis 51:3800-3810.
2. Schwertman NC, Owens MA, Adnan R. 2004. A simple more general boxplot method for identifying outliers. Computational Statistics and Data Analysis 47:165-174.
3. Dastjerdy B, Saeidi A, Heidarzadeh S. 2023. Review of Applicable Outlier Detection Methods to Treat Geomechanical Data. Geotechnics 3:375-396. MDPI AG.

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                           lat = 'decimalLatitude', lon= 'decimalLongitude',
                         colsp = "scientificName",
                         bbox = db,
                         minpts = 10)

 sqout <- seqfences(data = refdata[["Thymallus thymallus"]], var = 'bio6', output='outlier')
```

---

show,datacleaner-method

set method for displaying output details after outlier detection.

Description

set method for displaying output details after outlier detection.

Usage

```
## S4 method for signature 'datacleaner'
show(object)
```

Arguments

object	The data model for outlier detection.

Value

prints the datacleaner class for this package.

---

smc *Identify best outlier detection method using simple matching coefficient.*

Description

Identify best outlier detection method using simple matching coefficient.

Usage

```
smc(x, sp = NULL, threshold = NULL, warn = FALSE, autothreshold = FALSE)
```

Arguments

x	datacleaner class for each methods used to identify outliers in multidetect function.
sp	string. Species name or index if multiple species are considered during outlier detection.
threshold	numeric. Maximum value to denote an absolute outlier. The threshold ranges from 0, which indicates a point has not been flagged by any outlier detection method as an outlier, to 1, which means the record is an absolute or true outlier since all methods have identified it. At both extremes, many records are classified at low threshold values, which may be due to individual method weakness or strength and data distribution. Also, at higher threshold values, the true outliers are retained. For example, if ten methods are considered and 9 methods flag a record as an outlier, If a cutoff of 1 is used, then that particular record is retained. Therefore, the default cutoff is 0.6, but autothreshold can be used to select the appropriate threshold.
warn	logical. If **TRUE**, warning on whether absolute outliers obtained at a low threshold is indicated. Default **TRUE**.
autothreshold	vector. Identifies the threshold with mean number of absolute outliers.The search is limited within 0.51 to 1 since thresholds less than are deemed inappropriate for identifying absolute outliers. The autothreshold is used when threshold is set to NULL.

Value

best method for identifying outliers based on simple matching coefficient.

Examples

```
data(efidata)

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package = "specleanr"))

extdf <- pred_extract(data = efidata, raster = wcd,
                      lat = 'decimalLatitude', lon = 'decimalLongitude',
                     colsp = "scientificName",
                     list = TRUE,verbose = FALSE,
                     minpts = 6,merge = FALSE)#basin removed

#outlier detection

outliersdf <- multidetect(data = extdf, output='outlier', var = 'bio6',
                          exclude = c('x','y'), multiple = TRUE,
                          methods = c('mixediqr', "iqr", "mahal", "iqr", "logboxplot"))

smcout <- smc(x = outliersdf, sp= 1, threshold = 0.2)#
```

`sorensen` *Identifies best outlier detection method suing Sorensen Similarity Index.*

Description

Identifies best outlier detection method suing Sorensen Similarity Index.

Usage

```
sorensen(x, sp = NULL, threshold = NULL, warn = FALSE, autothreshold = FALSE)
```

Arguments

`x`	`datacleaner` class for each methods used to identify outliers in `multidetect` function.
`sp`	`string`. Species name or index if multiple species are considered during outlier detection.

threshold	numeric. Maximum value to denote an absolute outlier. The threshold ranges from 0, which indicates a point has not been flagged by any outlier detection method as an outlier, to 1, which means the record is an absolute or true outlier since all methods have identified it. At both extremes, many records are classified at low threshold values, which may be due to individual method weakness or strength and data distribution. Also, at higher threshold values, the true outliers are retained. For example, if ten methods are considered and 9 methods flag a record as an outlier, If a cutoff of 1 is used, then that particular record is retained. Therefore, the default cutoff is 0.6, but autothreshold can be used to select the appropriate threshold.
warn	logical. If **TRUE**, warning on whether absolute outliers obtained at a low threshold is indicated. Default **TRUE**.
autothreshold	vector. Identifies the threshold with mean number of absolute outliers.The search is limited within 0.51 to 1 since thresholds less than are deemed inappropriate for identifying absolute outliers. The autothreshold is used when threshold is set to NULL.

Value

best method for identifying outliers.

Examples

```
data(efidata)

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package = "specleanr"))

extdf <- pred_extract(data = efidata, raster = wcd,
                      lat = 'decimalLatitude', lon = 'decimalLongitude',
                     colsp = "scientificName",
                     list = TRUE,verbose = FALSE,
                     minpts = 6,merge = FALSE)#basin removed

 #outlier detection

outliersdf <- multidetect(data = extdf, output='outlier', var = 'bio6',
                          exclude = c('x','y'), multiple = TRUE,
                          methods = c('mixediqr', "iqr", "mahal", "iqr", "logboxplot"))

sordata <- sorensen(x = outliersdf, sp= 1, threshold = 0.2)#
```

thermal_ranges *Collates minimum, maximum, and preferable temperatures from FishBase.*

Description

Collates minimum, maximum, and preferable temperatures from FishBase.

Usage

```
thermal_ranges(
  x,
  colsp = NULL,
  verbose = FALSE,
  pct = 90,
  sn = FALSE,
  synonym = fishbase(tables = "synonym"),
  ranges = fishbase(tables = "ranges")
)
```

Arguments

x	dataframe or string. species names or a dataframe of species to aid in retrieving temperature ranges from FishBase.
colsp	string. if x is a data frame, then the column species is required. Otherwise for list of species or vector, the colsp is NULL.
verbose	logical To return implementation messages. Default FALSE.
pct	numeric. Provide the perecentage similarity of the species name provided and the one in FishBase. The lower the pct value, the higher the chances of getting a wrong species in the standard databases (FishBase). The plausible pct value should be greater than **0.9**.
sn	logical. Either to output synonym or only accepted names. This parameter reduces duplication of species synonyms and old name etc. For more information see FishBase.
synonym	fishbasedataframe. A standard database for species synonym names from FishBase. See FishBase for more information.
ranges	fishbasedataframe. A standard database for ecological ranges from FishBase. See FishBase for more information.

Value

Data table for minimum, maximum and preferable species temperatures from FishBase.

Examples

```
## Not run:

x <- thermal_ranges(x = "Salmo trutta")

## End(Not run)
```

`ttdata` *Thymallus thymallus species data from GBIF and iNaturalist*

Description

A `tibble` Data from GBIF (https://www.gbif.org/) and iNaturalist (https://www.inaturalist.org/)

Usage

```
data(ttdata)
```

Format

A `tibble` 100 rows and 8 columns.

Details

The species data was collated from the Global Biodiversity Information Facility and iNaturalist

Examples

```
data("ttdata")
ttdata
```

`xglosh` *Global-Local Outlier Score from Hierarchies*

Description

Global-Local Outlier Score from Hierarchies

Usage

```
xglosh(
  data,
  k,
  output,
  exclude = NULL,
  metric = "manhattan",
  mode = "soft",
  pc = FALSE,
  boot = FALSE,
  var,
  pcvar = NULL
)
```

Arguments

data	Data frame of species records with environmental data.
k	The size of the neighborhood (Hahsler et al 2022).
output	Either clean: for data frame with no suspicious outliers or outlier: to return dataframe with only outliers.
exclude	Exclude variables that should not be considered in the fitting the one class model, for example x and y columns or latitude/longitude or any column that the user doesn't want to consider.
metric	The different metric distances to compute the distances among the environmental predictors. See dist function and how te different distances are applied. The different measures are allowed including "euclidean", "maximum", "manhattan", "canberra", "binary".
mode	This includes soft when the outliers are removed using mean to compute the z-scores or robust when median absolute deviation.
pc	Whether principal component analysis will be computed. Default FALSE
boot	Whether bootstrapping will be computed. Default FALSE
var	The variable of concern, which is vital for univariate outlier detection methods
pcvar	Principal component analysis to e used for outlier detection after PCA. Default PC1

Value

Dataframe with or with no outliers.

References

1. Campello, Ricardo JGB, Davoud Moulavi, Arthur Zimek, and Joerg Sander. Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Transactions on Knowledge Discovery from Data (TKDD) 10, no. 1 (2015). doi:10.1145/2733381
2. Hahsler M, Piekenbrock M (2022). dbscan: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Related Algorithms. R package version 1.1-11, <https://CRAN.R-project.org/package=dbscan>

Examples

```
data("efidata")
danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                        lat = 'decimalLatitude',
                        lon= 'decimalLongitude',
                        colsp = "scientificName",
                        bbox = db,
                        minpts = 10)

gloshout <- xglosh(data = refdata[["Thymallus thymallus"]], exclude = c("x", "y"),
                output='outlier', metric ='manhattan', k = 3,
                 mode = "soft")
```

---

xkmeans *Flags outliers using kmeans clustering method*

Description

Flags outliers using kmeans clustering method

Usage

```
xkmeans(
  data,
  k,
  exclude = NULL,
  output,
  mode = "soft",
  method = "silhouette",
  seed = 1135,
  verbose = FALSE,
  pc = FALSE,
  boot = FALSE,
  var,
  pcvar = NULL
)
```

Arguments

data	Dataframe to check for outliers
k	The number of clusters to be used for optimization. It should be greater than 1. For many species k should be be greater 10 to ably cater for each species search for optimal k using the different optimization methods in kmethod
exclude	Exclude variables that should not be considered in the fitting the one class model, for example x and y columns or latitude/longitude or any column that the user doesn't want to consider.
output	Either clean: for a data set with no outliers or outlier: to output a data frame with outliers.
mode	Either robust, if a robust mode is used which uses median instead of mean and median absolute deviation from median.
method	The method to be used for the kmeans clustering. Default is `silhouette`. `Elbow method` can be used but user input is required, and therefore multiple outlier detection method is not possible.
seed	An integer to fix the maintain the iterations by during the kmeans method optimisation.
verbose	To indicate messages and the default is FALSE.
pc	Whether principal component analysis will be computed. Default `FALSE`
boot	Whether bootstrapping will be computed. Default `FALSE`
var	The variable of concern, which is vital for univariate outlier detection methods
pcvar	Principal component analysis to e used for outlier detection after PCA. Default `PC1`

Value

Dataframe with or with no outliers.

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                        lat = 'decimalLatitude',
                        lon= 'decimalLongitude',
                        colsp = "scientificName",
                        bbox = db,
                        minpts = 10)
```

```
kmeansout <- xkmeans(data = refdata[["Thymallus thymallus"]],
                     output='outlier', exclude = c('x', 'y'), mode = 'soft', k=3)
```

---

xknn *k-nearest neighbors for outlier detection*

Description

k-nearest neighbors for outlier detection

Usage

```
xknn(
  data,
  output,
  exclude = NULL,
  metric = "manhattan",
  mode = "soft",
  pc = FALSE,
  boot = FALSE,
  var,
  pcvar = NULL
)
```

Arguments

data	Data frame of species records with environmental data.
output	Either clean: for data frame with no suspicious outliers or outlier: to return dataframe with only outliers.
exclude	Exclude variables that should not be considered in the fitting the one class model, for example x and y columns or latitude/longitude or any column that the user doesn't want to consider.
metric	The different metric distances to compute the distances among the environmental predictors. See `dist` function and how te different distances are applied. The different measures are allowed including `"euclidean"`, `"maximum"`, `"manhattan"`, `"canberra"`, `"binary"`.
mode	This includes `soft` when the outliers are removed using mean to compute the z-scores or `robust` when median absolute deviation.
pc	Whether principal component analysis will be computed. Default `FALSE`
boot	Whether bootstrapping will be computed. Default `FALSE`
var	The variable of concern, which is vital for univariate outlier detection methods
pcvar	Principal component analysis to e used for outlier detection after PCA. Default `PC1`

Value

Dataframe with or with no outliers.

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                        lat = 'decimalLatitude',
                        lon= 'decimalLongitude',
                        colsp = "scientificName",
                        bbox = db,
                        minpts = 10)

lofout <- xknn(data = refdata[["Thymallus thymallus"]], exclude = c("x", "y"),
                output='outlier', metric ='manhattan',
                 mode = "soft")
```

---

xlof *Flags suspicious using the local outlier factor or Density-Based Spatial Clustering of Applications with Noise.*

Description

Flags suspicious using the local outlier factor or Density-Based Spatial Clustering of Applications with Noise.

Usage

```
xlof(
  data,
  output,
  minPts,
  exclude = NULL,
  metric = "manhattan",
  mode = "soft",
  pc = FALSE,
  boot = FALSE,
  var,
  pcvar = NULL
)
```

Arguments

data	Data frame of species records with environmental data
output	Either clean: for data frame with no suspicious outliers or outlier: to return dataframe with only outliers.
minPts	Minimum neighbors around the records.
exclude	Exclude variables that should not be considered in the fitting the one class model, for example x and y columns or latitude/longitude or any column that the user doesn't want to consider.
metric	Distance-based measure to examine the distance between variables. Default manhattan.
mode	Either soft if mean is used or robust if mad is used. Default soft.
pc	Whether principal component analysis will be computed. Default FALSE
boot	Whether bootstrapping will be computed. Default FALSE
var	The variable of concern, which is vital for univariate outlier detection methods
pcvar	Principal component analysis to e used for outlier detection after PCA. Default PC1

Value

Dataframe with or with no outliers.

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                        lat = 'decimalLatitude',
                        lon= 'decimalLongitude',
                        colsp = "scientificName",
                       bbox = db,
                        minpts = 10)

lofout <- xlof(data = refdata[["Thymallus thymallus"]], exclude = c("x", "y"),
               output='outlier', metric ='manhattan',
               minPts = 10, mode = "soft")
```

zscore *Computes z-scores to flag environmental outliers.*

Description

Computes z-scores to flag environmental outliers.

Usage

```
zscore(
  data,
  var,
  output = "outlier",
  type = "mild",
  mode = "soft",
  pc = FALSE,
  pcvar = NULL,
  boot = FALSE
)
```

Arguments

data	Dataframe or vector to check for outliers.
var	Variable considered in flagging suspicious outliers.
output	Either **clean**: for data frame with no suspicious outliers or outlier: to return dataframe with only outliers.
type	Either **mild** if zscore cut off is 2.5 or **extreme** if zscore is >3.
mode	Either robust, if a **robust** mode is used which uses median instead of mean and median absolute deviation from median.
pc	Whether principal component analysis will be computed. Default FALSE
pcvar	Principal component analysis to e used for outlier detection after PCA. Default PC1
boot	Whether bootstrapping will be computed. Default FALSE

Details

The method uses mean as an estimator of location and standard deviation for scale (Rousseeuw & Hubert 2011), which both have zero breakdown point, and their influence function is unbounded (robustness of an estimator to outliers) (Seo 2006; Rousseeuw & Hubert 2011). Because both parameters are not robust to outliers, it leads to outlier masking and swamping (Rousseeuw & Hubert 2011). Records are flagged as outliers if their Z-score exceeds 2.5 (Rousseeuw & Hubert 2011).

Value

Data frame with or with no outliers.

Examples

```
data("efidata")

danube <- system.file('extdata/danube.shp.zip', package='specleanr')

db <- sf::st_read(danube, quiet=TRUE)

wcd <- terra::rast(system.file('extdata/worldclim.tiff', package='specleanr'))

refdata <- pred_extract(data = efidata, raster= wcd ,
                          lat = 'decimalLatitude', lon= 'decimalLongitude',
                          colsp = "scientificName",
                          bbox = db,
                          minpts = 10)

 zout <- zscore(data = refdata[["Thymallus thymallus"]], var = 'bio6', output='outlier')
```

Index