Package: kwb.misa (via r-universe)

September 18, 2024

Title Functions to be Used in Project MiSa

Version 0.1.0

Description Assessment of oxygen course in rivers. Assessment is aimed at reducing critical situations and fish deaths.

License MIT + file LICENSE

URL https://github.com/KWB-R/kwb.misa

BugReports https://github.com/KWB-R/kwb.misa/issues

Depends R (>= 3.5.0)

Suggests covr, knitr, rmarkdown, testthat (>= 3.0.0)

VignetteBuilder knitr

Imports kwb.utils (>= 0.9.0.9000)

Remotes KWB-R/kwb.utils@dev

Config/testthat/edition 3

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.1.2

Repository https://kwb-r.r-universe.dev

RemoteUrl https://github.com/KWB-R/kwb.misa

RemoteRef HEAD

RemoteSha abadd7201657246b21b744c2af81976fd66b2f3f

Contents

`adjust_time` *Equal time intervals*

Description

Transforms the time vector, so that the every timestep is shifted to equal equally distanced points of time

Usage

```
adjust_time(time_vector, time_interval)
```

Arguments

`time_vector`	A POSIXct vector
`time_interval`	Temporal resolution in seconds

Value

A POSIX vector with equally distanced points of time. The new timesteps start on the hour and

aggregate_measurements

Average data per predefined timesteps

Description

Creates a data frame with fitted time column and corresponding average data values.

Usage

```
aggregate_measurements(time_vector, data_vector, time_interval = 60 * 15)
```

Arguments

time_vector	A POSIXct vector
data_vector	A numeric vector, with data corresponding to the time_vector
time_interval	Temporal resolution in seconds

Value

Data frame with POSIX time column "t" and data column "d

continuousTimeIntervals

Force data into predefined time intervals

Description

The measurements are fitted into timesteps defined be the first point of time and a temporal resolution

Usage

```
continuousTimeIntervals(
  time_vector,
  data_vector,
  res = 15,
  first_pointOfTime = min(time_vector, na.rm = T),
  last_pointOfTime = max(time_vector, na.rm = T)
)
```

Arguments

time_vector	A POSIXct vector
data_vector	A numeric vector, with data corresponding to the time_vector
res	Temporal resolution in minutes
first_pointOfTime	Starting point (POSIXct) of the newly defined time series. By default the minimum of the time_vector
last_pointOfTime	End point (POSIXct) of the newly defined time series. By default the maximum of the time_vector

Details

In a first step a vactor is generated with continuous timesteps, starting at first_pointOfTime by a defined time interval. Subsequently, the measured data is forced into timesteps with a similar time interval. Here, the measurements are assigned to the timestep that is closest to the actual time of measurements. If more than one measurement are assigned to one timestep, the average is used. If there is no measurement, NA is used.

Value

Dataframe with POSIX column "t" and data column "d"

---

count_def_events *Count Events of deficits*

Description

Counts the Number of intervals where x number of data points in a row are below a predifined threshold value. Events are separated by a specified number of data points above that threshod value. Furthermore, the exceedance of a value can also seperation critirion.

Usage

```
count_def_events(
  data_vector,
  starting_data_points,
  threshold,
  separating_data_points,
  use_recovery_value = FALSE,
  recovery_value = NULL,
  return_event_positions = FALSE
)
```

Arguments

data_vector	Numeric vector (with data in the same unit as the tjreshold)
starting_data_points	Minimal number of data points to define the beginning of an deficiency event
threshold	Numeric in the same unit as the data vector
separating_data_points	Minimal number of data points to seperate two events
use_recovery_value	If TRUE a recovery, two events are only separated if a revocvery value is exceeded between two deficits
recovery_value	Numeric in the same unit as the data vector. Only used if use_recovery_value = TRUE.
return_event_positions	Instead the number of events, the events starting and endpositions are returned, correspoding to the data vector

Value

Either a number of events or a data frame with event start and end position

Examples

```
data_vector <- sin(x = seq(0,50,0.5)) * 1:101/20

a <- count_def_events(
data_vector = data_vector,
starting_data_points = 2,
threshold = 0,
separating_data_points = 4,
use_recovery_value = FALSE,
recovery_value = 7,
return_event_positions = TRUE)

plot(data_vector, pch = 20, type = "b")
rect(xleft = a$tBeg, xright = a$tEnd, ybottom = -10, ytop = 10,
 col = "red", density = 4)

recovery_value <- 3
a <- count_def_events(
data_vector = data_vector,
starting_data_points = 2,
threshold = 0,
separating_data_points = 4,
use_recovery_value = TRUE,
recovery_value = recovery_value,
return_event_positions = TRUE)

plot(data_vector, pch = 20, type = "b")
rect(xleft = a$tBeg[a$start],
 xright = a$tEnd[a$end],
```

```
ybottom = -10, ytop = 10,
col = "red", density = 4)
abline(h = recovery_value, col = "blue")
```

`count_def_hours` *Count hours of deficits*

Description

Count hours of deficits

Usage

```
count_def_hours(data_vector, threshold, res)
```

Arguments

`data_vector`	Numeric vector (with data in the same unit as the threshold)
`threshold`	Numeric in the same unit as the data vector
`res`	Temporal resolution of data in minutes

Value

A single Value (hours of deficits)

`finding_o2Column` *finding_o2Column*

Description

This function looks for the oxygen data column in a data frame by column name

Usage

```
finding_o2Column(dataFrame, tryO2 = c("o2", "oxygen", "ox", "sauerstoff"))
```

Arguments

`dataFrame`	The data frame where the column is searched
`tryO2`	A vector with patterns possible patterns of columnnames with oxygen data (not case sensitive)

Value

A vector with column numbers of oxygen columns

`finding_timestampColumns` *finding_timestampColumns*

Description

This function looks for the timestamp column in a data frame by typical timestamp symbols

Usage

```
finding_timestampColumns(dataFrame)
```

Arguments

`dataFrame`	The data frame where the column is searched

Value

A vector with column numbers of timestamp columns

`interpolate_multipleNA` *Linear interpolation for one or more missing values*

Description

All sections of NA values that are smaller or equal as a defined maximal number of NA's are interpolated

Usage

```
interpolate_multipleNA(data_vector, max_na)
```

Arguments

`data_vector`	Numeric vector of measurements (including NA values)
`max_na`	the maximal number of NA values in a row to be interpolated

Value

A list containing the data vector with interpolated NA value as well as an information about the amount of NA's interpolated

misa_filter_data *Filter MiSa Dataframe*

Description

Filters the loaded data frame by sites and time

Usage

```
misa_filter_data(
  dataFrame,
  sites = "",
  tBeg = min(dataFrame$posixDateTime, na.rm = TRUE),
  tEnd = max(dataFrame$posixDateTime, na.rm = TRUE)
)
```

Arguments

dataFrame	Data frame loaded by a MiSa function (see details)
sites	Names of considered sites, written in the site column of a MiSa Dataframe
tBeg	POSIX-Value with a start time of the observeration interval
tEnd	POSIX-Value with an end time of the observeration interval

Details

The name of the site column must be "site", the name of the timestamp column must be "posixDate-Time". The best way is to load the oxygen data with one of the following functions: read_misa_oneSite(),read_misa_mult or read_misa_files().

Value

A filtered data frame with the same columns as the input data frame

misa_prepare_data *Prepare MiSa Data for MiSa Assessment*

Description

Timestamps are adapted, oxygen data is interpolated, it is filtered for summmer months

Usage

```
misa_prepare_data(df_MiSa, res = 15, max_na_interpolation = 60/res)
```

Arguments

df_MiSa	Data frame loaded with one of the MiSa Load functions
res	Temporal resolution in minutes
max_na_interpolation	Maximal numbers of NA values in a row to be interpolated. The default is one hour without measurements. Number of NA depneds on the temporal resolution (60 / res)

Value

List with data frames per site, that is ready for MiSa Assessmen. Additional information is printed about the number of interpolated NA values. If there are many NA values that are not interpolated it is probably due to the fact of no measurements during winter.

---

negative_deviation *Relative Negative Deviation from a reference*

Description

The cumulative sum of all negative deviations.

Usage

```
negative_deviation(data_vector, reference_vector)
```

Arguments

data_vector	Numeric data vector
reference_vector	Corresponding data of the reference

Details

First the similarity the data vector and the reference vector is calculated. Only complete pairs (no NA values) are used. For each data pair the quotient between data and reference is calcutaled. If data > reference the value is set to 1. All quotients are cumulated (-> absolute similarity). This can be maximum the number of data pairs. When deviding by the number of data pairs, the relative similarity is obtained. One minus the relative similarity is the negative deviation.

Value

Numeric value between 0 and 1

`read_misa_files` *Read MiSa Files*

Description

This function combines the functions read_misa_oneSite and read_misa_multipleSites and is strictly bound the the misa folder structure

Usage

```
read_misa_files(input_path)
```

Arguments

`input_path`	This is the directory where the two folders "files_per_site" and "sites_per_file" are located

Details

All csv files from both folders will be loaded. They must contain a timestamp column and an oxygen column. The timestamp column is identified automatically, by looking for a column where the entries contain ":" and one of the date separating symbols ".", "/" or "-".. The oxygen column is found by its colname. It should conatin "O2", "o2", "Oxygen", "oxygen", "ox", "Ox", "Sauerstoff" or "sauerstoff".

For "file_per_site" files: all letters in the filename before the first "_" are used for the sitename. For "sites_per_file" files: all column names will be used for site names. Thus, all columns except the timestamp must be oxygan concentrations at different sites

Value

A Data frame with 3 columen: Timestamp, Oxygen data and Site name

`read_misa_multipleSites` *read_misa_multipleSites*

Description

This function reads csv tables with one timestamp column and several oxygen data columns. Where the colnames refer to the sites of measurements

Usage

```
read_misa_multipleSites(path, file)
```

Arguments

path	The path to the file
file	Filename (including ".csv" Ending)

Value

A Data frame with 3 columen: Timestamp, Oxygen data and Site name

---

read_misa_oneSite *read_misa_oneSite*

Description

This function reads csv tables with one timestamp column and one oxygen data column

Usage

```
read_misa_oneSite(path, file, siteID)
```

Arguments

path	The path to the file
file	Filename (including ".csv" Ending)
siteID	a character vector specifying the site name

Value

A Data frame with 3 columen: Timestamp, Oxygen data and Site name

---

same_inarow *Repeating values in a row within a Vector*

Description

Describes the Values of a vector the times they are repeated and the start and end position of those values

Usage

```
same_inarow(v)
```

Arguments

v	A character, factor or numeric vector

Value

A data frame with four columns: Value (-> listed value of the input vector), Repeats (times it is repeated in a row), starts_at (start position), ends_at (end position).

SummerMonths *Breaking down POSIX into months and year and filter by month*

Description

Adds month and year column to data frame and filters by month. The addition of the year column is important for the following MiSa Assessment

Usage

```
SummerMonths(df, time_column = "t", months = 5:9)
```

Arguments

df	data frame with a POSIX time column
time_column	Name or number of the time column
months	the number of the months that should be kept. The default is 5:9 which are the important months for oxygen deficits in Berlin caused by CSOs

Value

The filteterd input data frame with months and year column

---

yearly_crit_Events *MiSa Assessment: Yearly Numbers of deficits*

Description

Counting the events below threshold values on a yearly basis

Usage

```
yearly_crit_Events(
  dataFrame,
  res = 15,
  seperating_hours = 5 * 24,
  deficiency_hours = 0.25,
  thresholds = 1.5,
  max_missing = 25,
  use_recovery_value = FALSE,
  recovery_value = NULL
)
```

Arguments

dataFrame	MiSa Dataframe: with columne "d": oxygen data, "year": year
res	Temporal resolution of oxygen data in minutes
seperating_hours	TODO: describe (also: should be "separating_hours" with "a", not "e")
deficiency_hours	TODO: describe
thresholds	Oxygen threshold values used for the assessment in mg/L
max_missing	The maximal allowed percent of missing oxygen data. If NA Values exceed this number, hours below thresholds are set to NA
use_recovery_value	TODO: describe
recovery_value	TODO: describe

Value

Data frane with rows per year and columns per threshold as well es for missing data

---

yearly_deficiency_time

MiSa Assessment: Yearly hours of deficits

Description

Counting the hours on a yearly basis below threshold values

Usage

```
yearly_deficiency_time(
  dataFrame,
  res = 15,
  thresholds = c(0.5, 1, 1.5, 2, 5),
  max_missing = 25
)
```

Arguments

dataFrame	MiSa Dataframe: with columne "d": data, "year": year
res	Temporal resolution of oxygen data in minutes
thresholds	Oxygen threshold values used for the assessment in mg/L
max_missing	The maximal allowed percent of missing oxygen data. If NA Values exceed this number, hours below thresholds are set to NA

Value

Data frane with rows per year and columns per threshold as well es for missing data

`yearly_negative_deviation` *Negative deviation from a reference site*

Description

Functions cumulates the negative deviation (lower O2-Concentrations) compared to a reference site without (significant) urban pollution

Usage

```
yearly_negative_deviation(dataFrame, oxygen_ref, max_missing = 25)
```

Arguments

`dataFrame`	MiSa Dataframe: with columne "d": oxygen data, "year": year
`oxygen_ref`	The course of oxygen of the reference
`max_missing`	The maximal allowed percent of missing oxygen data. If NA Values exceed this number, hours below thresholds are set to NA

Value

Data frane with negative deviation per year

Index