Package 'outfillingR'

December 3, 2024

Title Infill Missing Historical Climatic Data

Version 0.0.1

Description Infills missing or incomplete rainfall data based on specified parameters, statistical distributions, and historical calibration data.

License LGPL (>= 3)

Encoding UTF-8

Roxygen list(markdown = TRUE)

RoxygenNote 7.3.2

Suggests knitr,
testthat (>= 3.0.0)

Imports dplyr,
lubridate,
magrittr,
stats,
utils

Depends R (>= 2.10)

LazyData true

Config/testthat/edition 3

Contents

`calculate_and_plot_conditional_probabilities` *Calculate and Plot Conditional Probabilities*

Description

This function calculates conditional probabilities of rainfall based on ranges of rainfall estimates (RFE) and generates statistical summaries for each bin, including probability of rainy and dry days. Results are filtered to exclude bins with a low number of observations.

Usage

```
calculate_and_plot_conditional_probabilities(
  filtered_df,
  rfe_bin_edges,
  count_filter = 10
)
```

Arguments

`filtered_df`	A data frame containing rainfall data with columns "rainfall" and "rfe".
`rfe_bin_edges`	A numeric vector defining the edges of RFE bins.
`count_filter`	A numeric value specifying the minimum count required for a bin to be included in the results (default is 10).

Value

A list containing:

- `filtered_probabilities_df`: A data frame with conditional probabilities, mean and std of rainfall, and rainy/dry day probabilities by RFE bin.
- `p0`: Probability of rainfall > 0 when RFE is zero.
- `p0_rainyday`: Probability of rainfall > 0 for rainy days when RFE is zero.
- `p0_dryday`: Probability of rainfall > 0 for dry days when RFE is zero.

`calculate_b0_b1` *Calculate Logistic Model Coefficients (b0, b1)*

Description

Estimates the coefficients b0 and b1 for a logistic model based on mean rainfall estimate (RFE) and observed conditional probabilities of rain. Returns 0 for both coefficients if the total count of rainy days is below the specified threshold or if there are insufficient data points.

Usage

```
calculate_b0_b1(
  mean_rfe,
  conditional_prob_rain,
  bin_counts,
  min_rainy_days_threshold
)
```

Arguments

`mean_rfe`	A numeric vector representing mean RFE values for each bin.
`conditional_prob_rain`	A numeric vector representing conditional probabilities of rain for each bin.
`bin_counts`	A numeric vector indicating the count of observations in each RFE bin.
`min_rainy_days_threshold`	A numeric threshold for the minimum required total count of rainy days across bins.

Value

A named numeric vector containing `b0` and `b1` logistic model coefficients, or `NA` values if model fitting fails.

`calibrate_kappa_theta` *Calibrate Kappa and Theta via Log-Log Linear Regression*

Description

This function estimates the parameters kappa (k) and theta for rainfall estimation by performing a linear regression on the log-transformed mean and variance of rainfall. Kappa (k) is derived from the intercept, and theta from the slope of the regression line.

Usage

```
calibrate_kappa_theta(filtered_probabilities_df)
```

Arguments

`filtered_probabilities_df`	A data frame containing columns `Rainfall_Mean` (mean rainfall) and `Rainfall_Std` (standard deviation of rainfall).

Value

A named list with `k` (kappa) and `theta` values, or `NA` for both if there are fewer than 4 data points.

`compute_monthly_parameters` *Compute Monthly Parameters*

Description

Calculates monthly parameters based on rainfall data from a CSV file.

Usage

```
compute_monthly_parameters(
  data,
  custom_bins = c(1, 3, 5, 10, 15),
  count_filter = 10,
  min_rainy_days_threshold = 10
)
```

Arguments

`data`	Either a path to a CSV file containing historical rainfall data or a data frame.
`custom_bins`	A numeric vector specifying custom bins for RFE.
`count_filter`	A numeric threshold defining the minimum number of values in each bin to include in the calculations
`min_rainy_days_threshold`	A numeric threshold for minimum rainy days.

Value

A data frame with the computed monthly parameters.

`do_infilling` *Perform Rainfall Data Infilling*

Description

This function infills missing or incomplete rainfall data based on specified parameters, statistical distributions, and historical calibration data. It uses monthly parameters, supports a Markov chain approach, and allows for user-defined dry seasons and custom bins for rainfall estimates.

Usage

```
do_infilling(
  data,
  station,
  date,
  rainfall,
  rfe,
  metadata = NULL,
  metadata_station = NULL,
```

```
  lon,
  lat,
  station_to_exclude,
  rainfall_estimate_column,
  custom_bins = c(1, 3, 5, 10, 15, 20),
  count_filter = 10,
  min_rainy_days_threshold = 50,
  target_months = 5:9,
  distribution_flag = c("gamma", "lognormal"),
  markovflag = TRUE,
  b0 = -0.2,
  b1 = 0.05,
  a0 = 0.1,
  a1 = 1,
  b0_dryday = -0.2,
  b0_rainyday = -0.2,
  b1_dryday = 0.05,
  b1_rainyday = 0.05,
  kappa = 2,
  theta = 2,
  p0 = 0.001,
  p0_rainyday = 0.001,
  p0_dryday = 0.001
)
```

Arguments

`data`	Either a path to a CSV file containing historical rainfall data or a data frame.
`station`	A string specifying the column name in `data` that contains the station names.
`date`	A string specifying the column name in `data` that contains the date values.
`rainfall`	A string specifying the column name in `data` that contains the original rainfall values.
`rfe`	A string specifying the column name in `data` that contains rainfall estimates (RFE) used as predictors for rainfall generation.
`metadata`	An optional data frame containing additional metadata to merge with the historical data. Default is `NULL`.
`metadata_station`	A string specifying the column name in the metadata data frame that corresponds to `station`. If `NULL`, it defaults to the value of `station`.
`lon`	A string specifying the column corresponding to longitude values. If `metadata` is `NULL`, this column is from `data`, otherwise this is from `metadata`
`lat`	A string specifying the column corresponding to latitude values. If `metadata` is `NULL`, this column is from `data`, otherwise this is from `metadata`
`station_to_exclude`	A string specifying the station to exclude from calibration data.
`rainfall_estimate_column`	A string specifying the column name in `data` that contains rainfall estimates used for calibration.
`custom_bins`	A numeric vector specifying custom bins for RFE values. Default is `c(1, 3, 5, 10, 15, 20)`.

`count_filter` A numeric threshold specifying the minimum number of values required in each bin to include in calculations. Default is `10`.

`min_rainy_days_threshold` A numeric threshold for the minimum number of rainy days required for inclusion in calculations. Default is `50`.

`target_months` A numeric vector specifying the months (as integers) to apply dry season parameters. Default is `5:9` (May to September).

`distribution_flag` A character string specifying the statistical distribution to use for rainfall generation. Options are `"gamma"` (default) or `"lognormal"`.

`markovflag` A logical value indicating whether to use a Markov chain approach for rainfall occurrence. Default is `TRUE`.

`b0` A numeric value for the intercept in the dry season model. Default is `-0.2`.

`b1` A numeric value for the slope in the dry season model. Default is `0.05`.

`a0` A numeric value for the intercept in the rainfall distribution model. Default is `0.1`.

`a1` A numeric value for the slope in the rainfall distribution model. Default is `1`.

`b0_dryday` A numeric value for the intercept for dry days. Default is `-0.2`.

`b0_rainyday` A numeric value for the intercept for rainy days. Default is `-0.2`.

`b1_dryday` A numeric value for the slope for dry days. Default is `0.05`.

`b1_rainyday` A numeric value for the slope for rainy days. Default is `0.05`.

`kappa` A numeric value specifying the scaling factor for the rainfall distribution variance. Default is `2`.

`theta` A numeric value specifying the exponent for the rainfall distribution variance. Default is `2`.

`p0` A numeric value for the baseline probability of precipitation. Default is `0.001`.

`p0_rainyday` A numeric value for the probability of precipitation following a rainy day. Default is `0.001`.

`p0_dryday` A numeric value for the probability of precipitation following a dry day. Default is `0.001`.

Value

A data frame containing the infilled rainfall data, including columns for station name, date, RFE, original rainfall, and generated rainfall.

See Also

`weather_generator`, `compute_monthly_parameters`

Examples

```
## Not run:
# Example with the Zambia data set
data("zambia_data")

infill_data <- do_infilling(data = zambia_data,
                            station = "station",
                            date = "date",
```

```
                              rainfall = "rainfall",
                              rfe = "rfe",
                              lon = "lon",
                              lat = "lat",
                              station_to_exclude = "PETAUKE MET",
                              rainfall_estimate_column = "chirps")

## End(Not run)
```

`extract_rows_by_date_range_across_years` *Extract Rows by Date Range Across Multiple Years*

Description

This function filters rows in a data frame based on a specified date range, ignoring the year component and treating the date range as recurring annually. Useful for selecting seasonal data.

Usage

```
extract_rows_by_date_range_across_years(df, start_date, end_date)
```

Arguments

`df`	A data frame with a column named `date`, which should be of class `Date` or convertible to `Date`.
`start_date`	A string in "MM-DD" format specifying the start of the date range.
`end_date`	A string in "MM-DD" format specifying the end of the date range.

Value

A filtered data frame containing rows where the `date` falls within the specified date range, across all years.

`fill_nearest` *Fill Nearest Non-NA Values*

Description

This function takes a vector and fills NA values with the nearest non-NA value. It uses a data frame to facilitate filling both upwards and downwards to ensure the closest available value is used to fill each NA.

Usage

```
fill_nearest(x)
```

Arguments

`x`	A numeric vector that may contain NA values.

Value

A numeric vector with NA values filled by the nearest non-NA values. The original order of the vector is preserved.

`logistic_function` *Logistic Function*

Description

Computes the logistic function for a given RFE value and parameters. This function can be used to estimate conditional probabilities using a logistic model.

Usage

```
logistic_function(rfe, b0, b1)
```

Arguments

`rfe`	A numeric value or vector representing the rainfall estimate (RFE).
`b0`	A numeric value representing the intercept of the logistic model.
`b1`	A numeric value representing the slope of the logistic model.

Value

A numeric value or vector giving the probability output from the logistic function.

`select_calibration_data` *Select Calibration Data*

Description

This function reads a CSV file containing rainfall data, selects the relevant column for rainfall estimates, and excludes data from a specified station for calibration. It outputs a CSV file with the filtered calibration data.

Usage

```
select_calibration_data(
  data,
  station,
  rainfall_estimate_column,
  station_to_exclude,
  save = FALSE
)
```

Arguments

data	Either a path to a CSV file containing historical rainfall data or a data frame.
station	A string specifying the column name in data that contains the station names.
rainfall_estimate_column	A string representing the column name with rainfall estimates to be used in calibration.
station_to_exclude	A string representing the name of the station to exclude from the calibration data.
save	A logical column indicating whether to save the resulting data frame or not. Default FALSE.

Value

A data frame with the calibration data (without rows from the excluded station).

---

weather_generator *Generate Synthetic Rainfall Data*

Description

This function generates synthetic rainfall data using historical rainfall data and monthly parameters. It applies a specified statistical distribution (gamma or lognormal) to model rainfall amounts, optionally using a Markov chain to simulate the probability of precipitation based on past conditions.

Usage

```
weather_generator(
  data,
  station,
  date,
  rfe,
  rainfall,
  metadata = NULL,
  metadata_station = NULL,
  lon,
  lat,
  monthly_params_df,
  distribution_flag = "gamma",
  markovflag = TRUE
)
```

Arguments

data	Either a path to a CSV file containing historical rainfall data or a data frame.
station	A string specifying the column name in data that contains the station names.
date	A string specifying the column name in data that contains the date values.
rfe	A string specifying the column name in data that contains rainfall estimates (RFE) used as predictors for rainfall generation.

`rainfall` A string specifying the column name in `data` that contains the original rainfall values.

`metadata` An optional data frame containing additional metadata to merge with the historical data. Default is `NULL`.

`metadata_station` A string specifying the column name in the metadata data frame that corresponds to `station`. If `NULL`, it defaults to the value of `station`.

`lon` A string specifying the column corresponding to longitude values. If `metadata` is `NULL`, this column is from `data`, otherwise this is from `metadata`

`lat` A string specifying the column corresponding to latitude values. If `metadata` is `NULL`, this column is from `data`, otherwise this is from `metadata`

`monthly_params_df` A data frame containing monthly parameters for rainfall generation. This must include columns for `Month`, `b0`, `b1`, `b0_rainyday`, `b1_rainyday`, `b0_dryday`, `b1_dryday`, `a0`, `a1`, `kappa`, `theta`, `p0`, `p0_rainyday`, and `p0_dryday`.

`distribution_flag` A string specifying the statistical distribution to use for generating rainfall amounts. Options are `"gamma"` (default) or `"lognormal"`.

`markovflag` A logical value indicating whether to use a Markov chain approach for rainfall occurrence. If `TRUE` (default), the probability of rainfall depends on previous conditions.

Value

A data frame containing the generated synthetic rainfall data with the following columns: - `Station_name`: The station name. - `date`: The date. - `lon`: Longitude of the station. - `lat`: Latitude of the station. - `rfe`: Rainfall estimates (RFE) used in the generation process. - `original_rainfall`: The original rainfall value from `data`. - `generated_rainfall`: The generated synthetic rainfall value.

`zambia_data` *Zambia Meteorological Data*

Description

This dataset contains weather and climate-related variables recorded at Meteorological Stations in Zambia. The data spans multiple dates, stations, and provides key meteorological and model-derived information.

Usage

```
zambia_data
```

Format

A data frame with several rows and 15 variables:

station The name of the meteorological station. There are five stations in this data: PETAUKE MET, MSEKERA AGROMET, MFUWE MET, LUNDAZI MET and CHIPATA MET.

date Date of observation from 1st January 1983 until 31st December 2022. Format: YYYY-MM-DD.

lon Longitude of the station's location, in degrees.

lat Latitude of the station's location, in degrees.

rainfall Measured rainfall in millimeters.

rfe Rainfall estimates.

chirps Rainfall estimate from CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data), in millimeters.

imerg_cal Calibrated rainfall estimates from IMERG (Integrated Multi-satellite Retrievals for GPM), in millimeters.

imerg_uncal Uncalibrated rainfall estimates from IMERG, in millimeters.

era5 Rainfall data derived from ERA5 (ECMWF Reanalysis 5th Generation), in millimeters.

uwnd_925 U-component of wind (zonal wind) at 925 hPa, in meters per second.

uwnd_600 U-component of wind (zonal wind) at 600 hPa, in meters per second.

vwnd_925 V-component of wind (meridional wind) at 925 hPa, in meters per second.

vwnd_600 V-component of wind (meridional wind) at 600 hPa, in meters per second.

tamsat Rainfall estimates from TAMSAT (Tropical Applications of Meteorology using SATellite data), in millimeters.

Source

Data collected from five Meteorological Stations in Zambia and derived from multiple meteorological sources.

Index