Package ‘climatol’

August 29, 2016

Version 3.0

Date 2016-08-27

Title Climate Tools (Series Homogenization and Derived Products)

Author Jose A. Guijarro <jguijarrop@aemet.es>

Maintainer Jose A. Guijarro <jguijarrop@aemet.es>

Description Functions to homogenize climatological series and to produce climatological summaries and grids from the homogenized results, plus functions to draw wind-roses and Walter&Lieth diagrams.

Encoding latin1

License GPL (>= 2)

Depends R (>= 3.0.2)

Imports grDevices, graphics, stats, utils

Suggests fields, gstat, mapdata, maps, ncdf4, raster, RODBC, sp

URL http://www.climatol.eu/

NeedsCompilation no

Repository CRAN

Date/Publication 2016-08-29 17:44:24

R topics documented:

`dahgrid` *Interpolation of normalized homogeneous data on a predefined grid*

Description

Homogenized data generated by `homogen` are normalized and interpolated on a grid provided by the user at every time step, and saved in a NetCDF file.

Usage

```
dahgrid(varcli, anyi, anyf, anyip=anyi, anyfp=anyf, grid, mh=FALSE, std=NA,
ini=NA, obsonly=TRUE)
```

Arguments

`varcli`	Acronym of the name of the studied climatic variable, as in the data file name.
`anyi`	Initial year of the homogenized data.
`anyf`	Final year of the homogenized data.
`anyip`	First year of the desired reference period. (The reference period defaults to the whole period of the data).
`anyfp`	Last year of the desired reference period.
`grid`	Grid on which interpolations must be performed.
`mh`	If `TRUE`, read monthly data from file '`*-mh_*.dat`' (generated after a daily homogenization).
`std`	Type of normalization: **1:** deviations from the mean, **2:** rates to the mean (only for means greater than 1), **3:** standardization (subtract the mean and divide by the standard deviation of the sample).
`ini`	Initial date (first of January of `anyi` by default. This parameter is included in the '`*.rda`' file, but this file is not read when `mh=TRUE`.
`obsonly`	Do not interpolate estimated missing data. (`TRUE` by default).

Details

Homogenized data are read from the binary file 'VAR_ANYI-ANYF.rda' generated by homogen. Only series reconstructed from their longest homogeneous sub-period are retained, and they are normalized by their means (and standard deviations, if std=3), computed for the selected reference period (or for the whole period of the data, by default).

Unless obsonly is set to FALSE, data that were missing in the observed series are deleted to avoid interpolation of already interpolated data.

Finally, the normalized homogeneous data are interpolated on the predefined grid for every time step using an inverse distance weight method, and the resulting grids are stored in a NetCDF file named 'VAR_ANYIP-ANYFP.nc', including grids of the reference means (and standard deviations, if applied).

The user must provide the grid as an object of class SpatialPixel, as in this example defining a grid from 40N,3E to 43N,7E with a resolution of 0.1 degrees:

```
grid <- expand.grid(x=seq(3,7,.1),y=seq(40,43,.1))
library(sp)
coordinates(grid) <- ~ x+y
```

The resolution of this grid need not be too high, but adjusted to the spatial density of the available series.

The user may be more interested in obtaining grids of absolute values, rather than normalized. This can be achieved simply by undoing the normalization on the grids with the help of the provided grids of reference means and standard deviations. However, the resulting grids will only be the product of a geometrical interpolation, and will not reflect the influence of orography and other physiographic effects on the studied climatic variable. Therefore, it is more advisable to derive better reference grids of means (and standard deviations, if needed) by applying a geostatistical model to the reference means (provided in the file 'VAR_ANYIP-ANYFP_means.csv' with their corresponding coordinates).

This better quality climatic maps will probably have a higher resolution than that of the grids of the NetCDF file provided by this function. In that case, these normalized grids must be interpolated to the grid of the geostatistically derived climatic maps before undoing the normalization to obtain the final maps of absolute values at all or selected time-steps of the studied period.

Value

This function does not return any value, since the results are stored in files.

See Also

homogen

Examples

```
## Not run:
# This function reads data from files. In order to run the example, you should
# run the examples of the homogen function first.  Once the results from
# homogen have been generated, you need to define the grid to use. E.g., for
# the monthly means of the homogenized Ttest series:
#
```

```
grd=expand.grid(x=seq(-109,-107.7,.02),y=seq(44,45,.02))
library(sp)
coordinates(grd) <- ~ x+y
#
# Now we are ready to compute the grided homogenized monthly series:
#
dahgrid('Ttest', 1981, 2000, grid=grd, mh=TRUE)
#
# The resulting NetCDF file contains normalized values. You can obtain a new
# file with temperatures in degrees Celsius with externals tools, such as CDO:
#
cdo add -mul Ttest-mh_1981-2000.nc Ttest-mh_1981-2000_s.nc \
  Ttest-mh_1981-2000_m.nc Ttest-mu_1981-2000.nc
#
# But the new grids in Ttest-mu_1981-2000.nc are based on geometric
# interpolations only, and therefore it is better to build new files
# Ttest-mh_1981-2000_m.nc and Ttest-mh_1981-2000_s.nc through geostatistical
# methods.

## End(Not run)
```

---

dahstat *Statistical summaries of the homogenized data*

Description

Listing of means, standard deviations, quantiles or trends for a specified period, from data generated by homogen.

Usage

```
dahstat(varcli, anyi, anyf, anyip=anyi, anyfp=anyf, stat="me", ndc=1, vala=2,
cod=NULL, mnpd=0, mxsh=0, prob=.5, last=FALSE, long=FALSE, mh=FALSE,
pernys=100, ini=NA, estcol=4, sep=',', dec='.', eol='\n')
```

Arguments

varcli	Acronym of the name of the studied climatic variable, as in the data file name.
anyi	Initial year of the homogenized period.
anyf	Final year of the homogenized period.
anyip	First year of the period to analyze. (Defaults to anyi).
anyfp	Last year of the period to analyze. (Defaults to anyf).
stat	Statistical parameter to compute for the selected period: **"me":** Means (default), **"mdn"** Medians, **"max"** Maxima,

"min" Minima,
"std" Standard deviations,
"q" Quantiles (see the prob parameter),
"tnd" Trends,
"series" Do not compute any statistics; only output all homogenized series in individual *.csv files.

ndc Number of decimal places to be saved in the output file (1 by default).

vala Annual values to compute from the sub-annual data:
0: None,
1: Sums,
2: Means (default),
3: Maxima,
4: Minima.

cod Optional vector of codes of the stations to be processed.

mnpd Minimum percentage of original data. (0 = no limit).

mxsh Maximum SNHT. (0 = no limit).

prob Probability for the computation of quantiles (0.5 by default, i.e., medians).

last Logical value to compute statistics only for stations working at the end of the period of study. (FALSE by default).

long Logical value to compute statistics only for series built from the longest homogeneous sub-period. (FALSE by default).

mh If TRUE, read monthly data computed from daily adjusted series. (FALSE by default).

pernys Number of years on which to compute trends. (Defaults to 100).

ini First date of the series, with format "YYYY-MM-DD". (Only used if mh=TRUE. If not set, the first day of the study period will be assumed).

estcol Columns of the homogenized stations file to be included in the output file. (Defaults to 4, the column of station codes).

sep String to use for separating the output data. (' ').

dec Character to use as decimal point in the output data. ('.').

eol Line termination style. ('\n').

Details

Homogenized data are read from the file 'VAR_ANYI-ANYF.rda' saved by homogen, while this function saves the computed data for the specified period in 'VAR_ANYIP-ANYFP.STAT', where STAT is substituted by the stat requested statistic. An exception is when stat="q", since then the extension of the output file will be qPP, where PP stands for the specified prob probability (in percent). The output period ANYIP-ANYFP must of course be comprised within the period of the input data, ANYI-ANYF.

Parameters mnpd, mxsh and last act as filters to produce results only for series that have those minimum percentages of original data and maximum SNHT values, or to select only those stations working at the end of the period studied. No selection is performed by default, listing the desired statistic for all the reconstructed series (from every homogeneous sub-period).

stat='tnd' computes trends by OLS linear regression on time, listing them in a CSV file '*_tnd.csv' and their p-values in '*_pval.csv'

If stat='series' is chosen, two text files in CSV format will be produced for every station, one with the data and another with their flags: 0 for original, 1 for infilled and 2 for corrected data.

Value

This function does not return any value, since outputs are saved to files.

See Also

homogen, dahgrid.

Examples

```
## Not run:
# After having run the example of the homogen function, you can do:
dahstat("Ptest", 1951, 2010, stat='tnd')
# and check the new files Ptest_1951-2010_tnd.csv and Ptest_1951-2010_pval.csv

## End(Not run)
```

---

datcli *Monthly climatic data of a single station*

Description

Monthly climatic data, for a single station, of average precipitation, average daily maximum and minimum temperatures, and daily minimum temperatures.

Usage

```
data(datcli)
```

Format

A matrix of 4 lines (variables) and 12 columns (months) of data

Source

Made up data for the example of diagwl

See Also

diagwl

Examples

```
data(datcli)
print(datcli)
```

`db2dat` *Get daily or monthly data from a database and build input files *.dat and *.est*

Description

This function facilitates the creation of the input files needed by this package by retrieving the data from a database through an RODBC connection.

Usage

```
db2dat(varcli, anyi, anyf, minny=5, daily=TRUE, ch, dformat='%Y-%m-%d',
vtable, vcode, vdate, vval, stable, scode, sname, sx, sy, sz)
```

Arguments

`varcli`	Acronym of the name of the studied climatic variable, as it will appear in all data file names.
`anyi`	Initial year of the data to be included in the file.
`anyf`	Final year of the data to be included in the file.
`minny`	Minimum number of years with data for a series to be included in the file.
`daily`	Logical flag indicating whether the data are daily (the default) or monthly (set `daily=FALSE` in this case).
`ch`	Already open ODBC connection to the climatic database.
`dformat`	Date format in the database.
`vtable`	Name of the table containing our climatic variable.
`vcode`	Name of the variable containing station codes in the `vtable` table.
`vdate`	Name of the variable containing dates in the `vtable` table.
`vval`	Name of the climatic variable in the `vtable` table.
`stable`	Name of the table containing station information (metadata).
`scode`	Name of the variable containing station codes in the table `stable`.
`sname`	Name of the variable containing station names in the `stable` table.
`sx`	Name of the variable containing longitudes (degrees with decimals!) in the `stable` table.
`sy`	Name of the variable containing latitudes (degrees with decimals!) in the `stable` table.
`sz`	Name of the variable containing elevations (meters) in the `stable` table.

Details

This function creates the two input files needed by the homogenization functions of this package, 'VAR_YEAR-YEAR.dat' (holding the data) and 'VAR_YEAR-YEAR.est' (holding station coordinates, codes and names).

The table in the accessed database must contain either daily or monthly data (set daily=FALSE in this case). Otherwise the number of data per series will not be match the expected value and the function will fail.

Moreover, every data item must be in a different record in the database, as in this example table of monthly data (different variables for the same time step are O.K.):

```
Station   Date      T.max T.min Rel.Hum Precip  Wind.speed
S032    1991-01-01  12.1  -2.1    59    128.2     5.4
S032    1991-02-01  13.2  -2.5    62     78.4     6.2
...
```

But if the table in the database arranges all monthly values of one year (or all daily values of one month) in a single record, then this function will not retreive the data correctly.

Value

This function does not return any value.

See Also

homogen, homogsplit

Examples

```
## Not run:
# First we must access our climatic database through RODBC, wich requires to
# have this package installed. System programs that allow ODBC connections to
# databases must also be installed and properly configured.
#
# For this example we will assume that our database is named "climate" and we
# access it with user "USER" and password "PASS".  Then we open the connection
# with:
#
library(RODBC)
ch <- odbcConnect("climate",uid="USER",pwd="PASS")
#
# Now we want to use this function to gather all monthly relative humidity
# averages for the period 1961-2015, requiring a minimum of 10 years of data
# (not necessarily consecutive). We must use the corresponding names of tables
# and headers existing the the database, and putting the parameters in the
# required order we avoid the need to name them:
#
db2dat('HRel',1961,2015,10,FALSE,ch,'
'Station','Date','Value','stations','Station','Name','Longitude',
'Latitude','Elevation')
#
odbcClose(ch) #close the connection if you do not need it anymore
```

```
#
# Our data would now be ready to be homogenized with the homogen function:
#
homogen('HRel',1961,2015,vmin=0,vmax=100)
#

## End(Not run)
```

---

dd2m *Compute monthly data from daily series*

Description

Daily series are aggregated into total, mean, maximum, or minimum monthly values, and saved to files for further processing.

Usage

```
dd2m(varcli, anyi, anyf, anyip=anyi, anyfp=anyf, ndec=1, suf=NA, valm=2,
namax=10, na.strings="NA", homog=FALSE, ini=NA)
```

Arguments

varcli	Acronym of the name of the studied climatic variable, as in the data file name.
anyi	Initial year of the data present in the file.
anyf	Final year of the data present in the file.
anyip	First year of the output period. (Defaults to anyi).
anyfp	Last year of the output period. (Defaults to anyf).
ndec	Number of decimal places to be saved in the output file.
suf	Optional suffix to be added (by a hyphen) to the acronym of the variable to form the name of the input data.
valm	Monthly value to compute: **1:** Sum, **2:** Mean, **3:** Maximum, **4:** Minimum.
namax	Maximum number of missing data in any month to compute its monthly value. (10 by default)
na.strings	Missing data code in the original daily data.
homog	If TRUE, monthly aggregation will be performed on daily adjusted from a previous monthly homogenization.
ini	Initial date of the daily data, with format "YYYY-MM-DD". Its default NA value assume it to be the first day of the period of study.

Details

This function can be applied to either raw or homogenized daily data.

Raw data are read from files 'VAR_YEAR-YEAR.dat' and 'VAR_YEAR-YEAR.est', and monthly data will be saved to files 'VAR-m_YEAR-YEAR.dat' and 'VAR-m_YEAR-YEAR.est'.

Homogenized data are read from 'VAR_YEAR-YEAR.rda'. In this case monthly data will be saved to files named 'VAR-mh_YEAR-YEAR.dat' and 'VAR-mh_YEAR-YEAR.est'.

Value

This function does not return any value.

See Also

homogen, dahstat, dahgrid

Examples

```
## Not run:
#First we must generate the input files from example data:
data(Ttest)
write(dat,'Ttest_1981-2000.dat')
write.table(est.c,'Ttest_1981-2000.est',row.names=FALSE,col.names=FALSE)
rm(dat,est.c) #remove loaded data from memory space
#Now we can apply this function to these monthly precipitation data:
#
dd2m('Ttest',1981,2000)

## End(Not run)
```

---

diagwl *Walter \& Lieth climatic diagram*

Description

Plot of a Walter \& Lieth climatic diagram of a station.

Usage

```
diagwl(dat, est="", alt=NA, per="", margen=c(4, 4, 5, 4), mlab="",
pcol="#005ac8", tcol="#e81800", pfcol="#79e6e8", sfcol="#09a0d1", shem=FALSE,
p3line=FALSE, ...)
```

Arguments

`dat`	Monthly climatic data for which the diagram will be plotted.
`est`	Name of the climatological station
`alt`	Altitude of the climatological station
`per`	Period on which the averages have been computed
`margen`	Margins vector for the plot (to be passed to `par`).
`mlab`	Month labels for the X axis: **"en":** Month initials in English. **"es":** Month initials in Spanish. **Other:** Numeric labels (1-12).
`pcol`	Color pen for precipitation.
`tcol`	Color pen for temperature.
`pfcol`	Fill color for probable frosts.
`sfcol`	Fill color for sure frosts.
`shem`	Set to `TRUE` for southern hemisphere stations.
`p3line`	Set to `TRUE` to draw a supplementary precipitation line referenced to three times the temperature (as suggested by Bogdan Rosca).
`...`	Other graphic parameters

Details

Climatic data must be passed as a 4x12 matrix of monthly (January to December) data, in the following order:

Row 1: Mean precipitation

Row 2: Mean maximum daily temperature

Row 3: Mean minimum daily temperature

Row 4: Absolute monthly minimum temperature

This last row is only used to determine the probable frost months (when absolute monthly minimums are equal or lower than 0°C).

For stations located in the southern hemisphere it is useful to set `shem=TRUE`, in order to keep the summer period in the central zone of the graphic (the diagram will begin the plot with the July data).

As described by Walter and Lieth, when monthly precipitation is greater than 100 mm, the scale is increased from 2mm/ºC to 20mm/ºC to avoid too high diagrams in very wet locations. This change is indicated by a black horizontal line, and the graph over it is filled in solid blue.

When the precipitation graph lies under the temperature graph ($P < 2T$) we have an arid period (filled in dotted red vertical lines). Otherwise the period is considered wet (filled in blue lines), unless `p3line=TRUE`, that draws a precipitation black line with a scale $P = 3T$; in this case the period in which $3T > P > 2T$ is considered semi-arid.

Daily maximum average temperature of the hottest month and daily minimum average temperature of the coldest month are frequently used in vegetation studies, and are labeled in black at the left margin of the diagram.

Value

No value is returned by this function.

References

WALTER H & LIETH H (1960): Klimadiagramm Weltatlas. G. Fischer, Jena.

See Also

plot, par

Examples

```
data(datcli)
diagwl(datcli,est="Example station",alt=100,per="1961-90",mlab="en")
```

---

homogen *Automatic homogenization of climatological series*

Description

Automatic homogenization of climatological series, including missing data filling and detection and correction of outliers and shifts in the mean of the series.

Usage

```
homogen(varcli, anyi, anyf, suf=NA, nm=NA,  nref=c(10,10,4), std=3, swa=NA,
ndec=1, dz.max=5, dz.min=-dz.max, wd=c(0,0,100), snht1=25, snht2=snht1,
tol=.02, mxdif=NA, force=FALSE, wz=.001, trf=0, mndat=NA, gp=3, ini=NA,
na.strings="NA", maxite=50, vmin=NA, vmax=NA, nclust=100,
clustmethod='ward.D2', grdcol=grey(.5), mapcol=grey(.65), hires=TRUE,
expl=FALSE, metad=FALSE, sufbrk='m', verb=TRUE)
```

Arguments

varcli	Acronym of the name of the studied climatic variable, as in the input data files.
anyi	Initial year of the data present in the file.
anyf	Final year of the data present in the file.
suf	Optional suffix appended with a '-' to the name of the variable in the input files.
nm	Number of data per year in each station. (Defaults to NA, and then it will be computed from the total number of data).
nref	Maximum number of references for data estimation. (Defaults to 10 in the detection stages, and to 4 in the final series adjustments).
std	Type of normalization: **1:** deviations from the mean,

2: rates to the mean (only for means greater than 1),

3: standardization (subtract the mean and divide by the sample standard deviation).

`swa` Size of the step forward to be applied to the staggered window application of SNHT. If not set (the default), 365 terms (one year) will be used for daily data, and 60 otherwise.

`ndec` Number of decimal digits to which the homogenized data must be rounded.

`dz.max` Threshold of outlier tolerance, in standard deviations. (5 by default).

`dz.min` Lower threshold of outlier tolerance may be different from the higher one. (By default, they are the same, with opposite signs).

`wd` Distance (in km) at which reference data will weigh half of that of another located at zero distance. (Defaults to `c(0,0,100)`, meaning that no weighting will be applied in the first two stages, and 100 km in the third).

`snht1` Threshold value for the stepped SNHT window test applied in stage 1. (25 by default).

`snht2` Threshold value for the SNHT test when applied to the complete series in stage 2 (same value as snht1 by default).

`tol` Tolerance factor to split several series at a time. The default is 0.02, meaning that a 2% will be allowed for every reference data. (E.g.: if the maximum SNHT test value in a series is 30 and 10 references were used to compute the anomalies, the series will be split if the maximum test of the reference series is lower than 30*(1+0.02*10)=36. Set `tol=0` to disable further splits when any reference series has already been split at the same iteration).

`mxdif` Maximum difference of any data item in consecutive iterations. If not set, defaults to half of the data precision (defined by the number of decimals).

`force` Force break even when only one reference is available. (`FALSE` by default).

`wz` Scale parameter of the vertical coordinate `Z`. 0.001 by default, which gives the vertical coordinate (in m) the same weight as the horizontal coordinates (internally managed in km).

`trf` By default, data are not transformed (`trf=0`), but if the data frequency distribution is very biased, the user can choose to apply a log(x+1) transformation (`trf=1`) or any root of index `trf>1` (2 for square root, 3 for cubic root, etc. Fractional numbers are allowed).

`mndat` Minimum number of data for a split fragment to become a new series. It defaults to half of the `swa` value for daily data, or to `nm` otherwise, with a minimum of 5 terms.

`gp` Graphic parameter:

0: no graphic output,

1: only descriptive graphics of the input data,

2: as with 1, plus diagnostic graphics of anomalies,

3: as with 2, plus graphics of running annual means and applied corrections,

4: as with 3, but running annual totals (instead of means) will be plotted in the last set of graphics. (Better when working with precipitation data).

ini	Initial date, with format 'YYYY-MM-DD'. If not set, it will be assumed that the series begin the first of January of the initial year anyi.
na.strings	Character string to be treated as a missing value. (It can be a vector of strings, if more than one is needed). Defaults to 'NA', the standard R missing data code.
maxite	Maximum number of iterations when computing the means of the series. (50 by default).
vmin	Minimum possible value (lower limit) of the studied variable. Unset by default, but note that vmin=0 will be applied if std is set to 2.
vmax	Maximum possible value (upper limit) of the studied variable. (E.g., for relative humidity or relative sunshine hours it is advisable to set vmax=100).
nclust	Maximum number of stations for the cluster analysis. (If much greater than 100, the default value, the process may be too long and the graphic too dense).
clustmethod	Clustering method. Defaults to 'ward.D2', but the user can choose any hierarchical method available in the hclust function.
grdcol	Color of the graphic background grids. (Middle gray by default).
mapcol	Color of the background map (gray by default).
hires	By default, the background map will be drawn in high resolution. Set this parameter to FALSE if you are studying a big geographical area (>1000 km).
expl	Set this to TRUE to perform an exploratory analysis.
metad	Set this to TRUE if a metadata file is provided (see the details).
sufbrk	Suffix to add to varcli to form the name of the provided metadata file. 'm' by default, to read the breaks detected at the monthly scale. (Only relevant when metad=TRUE).
verb	Verbosity. Set to FALSE to avoid messages being output to the console. (They will be in the output log file anyway).

Details

Input data must be provided in two text files, one with the data (with extension dat) and another with the station coordinates (with extension est). Both have as base name, 'VAR_YEAR-YEAR', composed by the acronym of the climatological variable, and the initial and final years of the data, as set in the first three parameters of the call, varcli, anyi and anyf.

Data are stored in a free blank separated format (any number of data items per line is allowed), in chronological order, station by station (all data from station 1 go first, then all data from station 2, and so on). As dates are not stored in this file, all data must be present in the file, using a code for any missing data in the records (NA by default, but any other code can be used, provided that they are specified in the parameter na.strings).

The stations file, with extension est, is also a blank separated text file where each line identifies a single station, with structure X Y Z CODE NAME: Coordinates X and Y may be in geographical degrees (longitude and latitude, in decimal form), or in m if you are using a UTM projection; elevation Z must be supplied in m; and the identification CODE and the full NAME of the station must be quoted if they contains blanks).

The transformation of the input data may be very useful to normalize highly biased L-shape distributed variables (as is often the case with precipitation, wind speed, ...), but use it with caution.

Alternatively, you can use the rate normalization (`std=2`) on the raw data if the variable has a natural zero lower limit. (This alternative has yielded better results than transformation in some applications, provided that no homogeneous sub-period has means lower than 1. If this is the case, a workaround may be to multiply all data values by a constant prior to their homogenization).

The default values of `dz.max`, `snht1` and `snht2` can be appropriate for monthly values of temperature, but not so much for precipitation or for daily series. Therefore it is advisable to adjust them empirically with the help of the histograms in the graphic output of a first exploratory application, using `expl=TRUE` in the call to the function.

This graphic output includes: a) a summary of the data availability and frequency distribution; b) a correlogram of the first differences of the series; c) a dendrogram based on these correlations and a map with the station locations; d) graphics of anomalies showing the detected breaks, the minimum distance to a reference data and the number of references used; e) a histogram of maximum SNHT values found in overlapping window analysis; d) and e) are repeated for the analysis on the whole series; f) histograms of number of splits per station and per year; g) graphics of final anomalies of the series h) graphics of the reconstructed series and applied corrections; i) a histogram of the normalized anomalies of all data (may help to set rejection thresholds for the outliers); final histograms of SNHT values; and j) a plot of quality/singularity of the stations (a bad score may be due to a bad quality of the series, but also to a singular siting with a peculiar micro-climate).

Note that every time that a significant shift in the mean of the series is detected, it will be split into two (potentially) homogeneous sub-periods, and hence the final number of homogenized series will be increased, as complete homogeneous series will be reconstructed from them. When several homogeneous series have been yielded for the same location, the user can choose to use that reconstructed from the last sub-period (for climate monitoring), from the period with the highest percentage of original data or the highest final SNTH (the more robust reconstruction) or all of them (e.g., for climatic mapping, when no a priori knowledge can indicate which of the sub-periods will be more representative at the studied spatial scale). Aditional columns `pod`, `ios`, `ope` and `snht` in the stations data frame can help the user to choose the most appropriate series for each location when using the post-processing functions `dahstat` and `dahgrid` (see below).

Value

This function does not return any value, its results being saved to files with the same base name as the input files, and extensions:

***.txt:** A text file that logs all the processing output,

***_out.csv:** List of corrected outliers,

***_brk.csv:** List of corrected breaks,

***.pdf:** PDF file with a collection of diagnostic graphics,

***.rda:** Homogenization results in R binary format, used by `dahstat` and `dahgrid` post-processing functions, but can be loaded by the user for further with the function `load`). This file contains the following objects:

- **dat** matrix or array of original data,
- **dah** matrix or array of homogenized data,
- **est.c** data frame with columns:
  - **X** X coordinate,
  - **Y** Y coordinate,

Z elevation,
Code code of the station,
Name name of the station,
pod percentage of original data,
ios index of original station,
ope operating at the end of the period? (0=no, 1=yes),
snht relative SNTH of the homogenized series

nd number of time steps in every series,
ne number of series after the homogenization,
nei initial number of series,
nm number of "months" in a year (0=daily data),
ndec number of decimals in the data,
std type of standardization used (as explained in the details),
ini initial date of the period under study

Author(s)

Jose A. Guijarro

References

Guijarro JA (2014): Quality Control and Homogenization of Climatological Series. In Eslamian S (Ed.), Handbook of Engineering Hydrology, Vol. 1: Fundamentals and Applications. Francis and Taylor, CRC Group, USA, ISBN 9781466552357, 636 pp.

Guijarro JA (2015): Homogenization of Spanish mean wind speed monthly series. In 8th Seminar for Homogenization and Quality Control in Climatological Databases and Third Conference on Spatial Interpolation Techniques in Climatology and Meteorology, Budapest, 12-16 May 2014 (Lakatos et al., Eds.), WMO Climate Data and Monitoring WCDMP-No. 84, pp. 98-106.

Azorin-Molina C, Guijarro JA, McVicar TR, Vicente-Serrano SM, Chen D, Jerez S, Espírito-Santo F (2016): Trends of daily peak wind gusts in Spain and Portugal, 1961-2014. J. Geophys. Res. Atmos., 121, doi:10.1002/2015JD024485, 20 pp.

See Also

dahstat, dahgrid, outrename, dd2m

Examples

```
## Not run:
#First we must generate the input files from example data:
data(Ptest)
write(dat,'Ptest_1951-2010.dat',ncolumns=12)
write.table(est.c,'Ptest_1951-2010.est',row.names=FALSE,col.names=FALSE)
rm(dat,est.c) #remove loaded data from memory space
#Now we can apply this function to these monthly precipitation data:
homogen('Ptest',1951,2010,std=2,dz.max=7,snht1=15,gp=4)
#
#Now with daily mean temperatures. Preparation of the input files:
```

```
data(Ttest)
write(dat,'Ttest_1981-2000.dat')
write.table(est.c,'Ttest_1981-2000.est',row.names=FALSE,col.names=FALSE)
rm(dat,est.c)
#Daily series are quite noisy. Let's aggregate them into monthly values:
dd2m('Ttest',1981,2000)
#Homogenization of the monthly averages:
homogen('Ttest-m',1981,2000)
# If you have a history of the stations, you can edit the file
# Ttest_1981-2000_brk.csv to adjust the dates of the breaks to known relevant
# changes.
# Now we can obtain the homogenized daily series adjusting them to the breaks
# found at the monthly scale:
homogen('Ttest',1981,2000,dz.max=7,metad=TRUE)

## End(Not run)
```

---

homogsplit *Apply homogen() on overlapping rectangular areas.*

Description

If the number of series is too big to be homogenized all at a time (normally several thousands, depending on the computer resources), this function can homogenize them by splitting the geographical domain in overlapping rectangular areas.

Usage

```
homogsplit(varcli, anyi, anyf, xc=NULL, yc=NULL, xo=.5, yo=.38, maponly=FALSE,
suf=NA, nm=NA, nref=c(10,10,4), swa=NA, std=3, ndec=1, dz.max=5,
dz.min=-dz.max, wd=c(0,0,100), snht1=25, snht2=snht1, tol=.02, mxdif=NA,
force=FALSE, wz=.001, trf=0, mndat=NA, gp=3, ini=NA, na.strings="NA",
maxite=50, vmin=NA, vmax=NA, nclust=100, clustmethod='ward.D2',
grdcol=grey(.5), mapcol=grey(.65), hires=TRUE, expl=FALSE, metad=FALSE,
sufbrk='m', verb=TRUE)
```

Arguments

varcli	Acronym of the name of the studied climatic variable, as in the data file name.
anyi	Initial year of the data present in the file.
anyf	Final year of the data present in the file.
xc	Vector of X axis coordinates setting the domain splitting meridians.
yc	Vector of Y axis coordinates setting the domain splitting parallels.
xo	Overlapping width in the East-West direction.
yo	Overlapping width in the North-South direction.
maponly	Do not homogenize. Only draw a map with stations locations and domain partitioning.

suf — Optional suffix appended with a '-' to the name of the variable in the input files.

nm — Number of data per year in each station. (Defaults to NA, and then it will be computed from the total number of data).

nref — Maximum number of references for data estimation. (Defaults to 10 in the detection stages, and to 4 in the final series adjustments).

swa — Size of the step forward to be applied to the staggered window application of SNHT. If not set (the default), 365 terms (one year) will be used for daily data, and 60 otherwise.

std — Type of normalization:

- **1:** deviations from the mean,
- **2:** rates to the mean (only for means greater than 1),
- **3:** standardization (subtract the mean and divide by the sample standard deviation).

ndec — Number of decimal digits to which the homogenized data must be rounded.

dz.max — Threshold of outlier tolerance, in standard deviations. (5 by default).

dz.min — Lower threshold of outlier tolerance may be different from the higher one. (By default, they are the same, with opposite signs).

wd — Distance (in km) at which reference data will weigh half of that of another located at zero distance. (Defaults to `c(0,0,100)`, meaning that no weighting will be applied in the first two stages, and 100 km in the third).

snht1 — Threshold value for the stepped SNHT window test applied in stage 1. (25 by default).

snht2 — Threshold value for the SNHT test when applied to the complete series in stage 2 (same value as snht1 by default).

tol — Tolerance factor to split several series at a time. The default is 0.02, meaning that a 2% will be allowed for every reference data. (E.g.: if the maximum SNHT test value in a series is 30 and 10 references were used to compute the anomalies, the series will be split if the maximum test of the reference series is lower than 30*(1+0.02*10)=36. Set `tol=0` to disable further splits when any reference series has already been split at the same iteration).

mxdif — Maximum difference of any data item in consecutive iterations. If not set, defaults to half of the data precision (defined by the number of decimals).

force — Force break even when only one reference is available. (`FALSE` by default).

wz — Scale parameter of the vertical coordinate `Z`. 0.001 by default, which gives the vertical coordinate (in m) the same weight as the horizontal coordinates (internally managed in km).

trf — By default, data are not transformed (`trf=0`), but if the data frequency distribution is very biased, the user can choose to apply a log(x+1) transformation (`trf=1`) or any root of index `trf>1` (2 for square root, 3 for cubic root, etc. Fractional numbers are allowed).

mndat — Minimum number of data for a split fragment to become a new series. It defaults to half of the `swa` value for daily data, or to `nm` otherwise, with a minimum of 5 terms.

`gp` Graphic parameter:

- **0:** no graphic output,
- **1:** only descriptive graphics of the input data,
- **2:** as with 1, plus diagnostic graphics of anomalies,
- **3:** as with 2, plus graphics of running annual means and applied corrections,
- **4:** as with 3, but running annual totals (instead of means) will be plotted in the last set of graphics. (Better when working with precipitation data).

`ini` Initial date, with format `'YYYY-MM-DD'`. If not set, it will be assumed that the series begin the first of January of the initial year `anyi`.

`na.strings` Character string to be treated as a missing value. (It can be a vector of strings, if more than one is needed). Defaults to 'NA', the standard missing data code in R.

`maxite` Maximum number of iterations when computing the means of the series. (50 by default).

`vmin` Minimum possible value (lower limit) of the studied variable. Unset by default, but note that `vmin=0` will be applied if `std` is set to 2.

`vmax` Maximum possible value (upper limit) of the studied variable. (E.g., for relative humidity or relative sunshine hours it is advisable to set `vmax=100`).

`nclust` Maximum number of stations for the cluster analysis. (If much greater than 100, the default value, the process may be too long and the graphic too dense).

`clustmethod` Clustering method. Defaults to `'ward.D2'`, but the user can choose any hierarchical method available in the `hclust` function.

`grdcol` Color of the graphic background grids. (Middle gray by default).

`mapcol` Color of the background map (gray by default).

`hires` By default, the background map will be drawn in high resolution. Set this parameter to `FALSE` if you are studying a big geographical area (>1000 km).

`expl` Set this to `TRUE` to perform an exploratory analysis.

`metad` Set this to `TRUE` if a metadata file is provided (see the details).

`sufbrk` Suffix to add to `varcli` to form the name of the provided metadata file. `'m'` by default, to read the breaks detected at the monthly scale. (Only relevant when `metad=TRUE`).

`verb` Verbosity. Set to `FALSE` to avoid messages being output to the console. (They will be in the output log file anyway).

Details

If you have not decided the splitting meridians and parallels, do not set them, and the function will provide a map to help in selecting the areas.

If you set the `xc` and `yc` splitting borders, setting `maponly=TRUE` will also produce a map with the stations, plus the required overlapping areas, without doing any homogenization. In this way you can review the limits of the areas to choose new ones if you are not happy with the current partitioning.

All parameters except xc, yc, xo, yo and maponly are the same as in the homogen function, and will be passed to it to perform the homogenization.

If a rectangular area include less than 10 stations, these will be added to the next area. Warning! If it is the last area, they will not be processed, and the homogenization results will have inconsistent number of stations. In this case the user should try a new set of cutting limits.

One graphic output file will be produce for every area containing stations, but the rest of the output will be merged into single files.

Value

This function does not return any value.

See Also

homogen, dahstat, dahgrid

Examples

```
## Not run:
# First we must generate the input files from example data:
#
data(Ptest)
write(dat,'Ptest_1951-2010.dat',ncolumns=12)
write.table(est.c,'Ptest_1951-2010.est',row.names=FALSE,col.names=FALSE)
rm(dat,est.c) #remove loaded data from memory space
#
# Now we can apply this function to these daily temperature data:
#
homogsplit('Ptest',1951,2010,3.0,39.6,.1,.1,std=2,dz.max=7,snht1=15,gp=4)
#
# Note that this is just a trivial example; this function is intended to
# homogenize a network of thousands of long series which might overload
# computer resources.

## End(Not run)
```

---

outrename *Rename homogen's output files*

Description

This function inserts a suffix to the output file names of homogen, to prevent them from being rewritten by any further function run.

Usage

```
outrename(varcli, anyi, anyf, suffix, restore=FALSE)
```

Arguments

`varcli`	Acronym of the name of the studied climatic variable, as in the data file name.
`anyi`	Initial year of the study period
`anyf`	Final year of the study period
`suffix`	Suffix to be inserted (or removed) in the output file names.
`restore`	Set this parameter to `TRUE` to remove the `suffix` previously inserted by this function.

Details

The variable (or file base) name is separated from the appended suffix by a hyphen. The purpose of this function is to allow a new application of `homogen` to the same data with different parameters without overwriting the previous results.

Value

This function does not return any value.

See Also

`homogen`

Examples

```
## Not run:
#After having run the example of the homogen function, you can do:
outrename("Ptest", 1951, 2010, "old")
#
#The previous output files will be renamed to Ptest-old_1951-2010*

## End(Not run)
```

`Ptest` *Monthly precipitation data*

Description

Example of a network of 20 monthly precipitation series for 1951-2010.

Usage

```
data(Ptest)
```

Format

This data set contains:

dat Monthly precipitation data arranged in a 720x20 matrix

est.c Data frame with station's coordinates (X,Y,Z), codes and names

Source

Subset of synthetic precipitation data from the Spanish research project MULTITEST (2015-2017)

See Also

`homogen`, `homogsplit`

Examples

```
data(Ptest)
print(est.c)
str(dat)
rm(est.c,dat)
```

---

rosavent *Wind-rose plot*

Description

Plot of a wind-rose. Wind frequencies may be supplied by speed intervals, and can be absolute or relative.

Usage

```
rosavent(frec, fnum=4, fint=5, flab=2, ang=3*pi/16,
col=rainbow(10,0.5,0.92,start=0.33,end=0.2), margen=c(0,0,4,0), key=TRUE,
uni="m/s", ...)
```

Arguments

`frec`	Data frame containing the wind frequencies.
`fnum`	Number of reference circles to plot.
`fint`	Frequency steps (in %) between reference circles.
`flab`	Parameter indicating which circles must be labelled: **1:** Label outer circle only, **2:** Label all circles, **Other value:** Do not label any circle.
`ang`	Angle along which circles will be labelled.

`col`	Colors to fill the frequency polygons.
`margen`	Margins vector for the plot (to be passed to `par`).
`key`	Set to FALSE if you do not want a legend of the wind-rose, that will otherwise be plotted if frequencies are supplied by speed intervals.
`uni`	Speed units for the legend header.
`...`	Other graphic parameters.

Details

The number of direction classes is indifferent: As many as elements of the frequency vector (or columns of the matrix) will be considered, but the first element (column) must correspond to North. Frequencies will be converted to percentages, no matter how they were supplied.

Value

No value is returned by this function.

See Also

`plot`, `par`

Examples

```
data(windfr)
rosavent(windfr, 4, 4, ang=-3*pi/16, main="Annual windrose")
```

`Ttest` *Daily temperature data*

Description

Example of a network of 12 daily temperature series for 1981-2000.

Usage

```
data(Ttest)
```

Format

This data set contains:

dat Daily mean temperature data arranged in a 7305x12 matrix

est.c Data frame with station's coordinates (X,Y,Z), codes and names (invented)

Source

Subset of synthetic temperature data from Killick, R. E. (2016): Benchmarking the Performance of Homogenisation Algorithms on Daily Temperature Data, PhD Thesis, University of Exeter.

See Also

homogen, dd2m, homogsplit

Examples

```
data(Ttest)
print(est.c)
str(dat)
rm(est.c,dat)
```

---

windfr *Wind frequency table*

Description

Example of wind frequency contingency table.

Usage

```
data(windfr)
```

Format

A table of 4 lines (speed intervals) and 16 columns (wind directions) of data, plus headers.

Source

Bogus data for the example of rosavent

See Also

rosavent

Examples

```
data(windfr)
print(windfr)
```

Index