GOBIERNO DE ESPAÑA

MINISTERIO
DE AGRICULTURA Y PESCA,
ALIMENTACIÓN Y MEDIO AMBIENTE

AEMet
Agencia Estatal de Meteorología

9th Seminar for homogenization and quality control in climatological databases and 4th conference on spatial interpolation techniques in climatology and meteorology (Budapest, 3-7 April 2017)

Homogenization of climatological series with Climatol 3.0

José A. Guijarro <jguijarrop@aemet.es>

State Meteorological Agency (AEMET), Balearic Islands Office, Spain

Introduction

The need to homogenize observational series before its use to assess climate variability emerged long time ago. Efforts were initially focused on annual, seasonal and monthly series, and the successful COST Action ES0601 allowed the exchange of ideas between homogenization specialists and the improvement of the their methodologies. But now the stress is put on the homogenization of daily series, since the study of the variability of indices and extreme values depends on them. The new version 3.0 of the R *climatol* package provides functions to facilitate the homogenization of climatological variables at any temporal scale, as long as the data in the series may be considered synchronous (which is doubtful in the case of sub-daily data).

Climatol homogenization methodology

The homogenization procedure relies on a simple method for estimating data at one point by means of other synchronous data from nearby stations, using a form of orthogonal regression known as Reduced Major Axis (RMA; Leduc, 1987). Orthogonal regression is adjusted by minimizing the perpendicular distance of the scatter points to the regression line, instead of minimizing the vertical distance to that line as in Ordinary Least Squares regression (OLS). Although there is an analytic expression to adjust an Orthogonal regression, a simpler and very close approximation is the RMA expression

$$\hat{y}_i = x_i$$

in which both dependent and independent variables have been previously standardized by removing their mean and dividing by their standard deviation. The climatol package allows the use of one or more reference data, and hence x_i may be an (optionally weighted) mean of several nearby data. (We speak of nearby data and not nearby series because data availability in the surroundings vary along time when series are incomplete).

This procedure allows great flexibility to use nearby data to fill in missing data in a problem series, since no common period of observation is needed between the two, and the closest reference data can be chosen in every time step adapting to the different availability of data in the other series. Its main drawback is the estimation of the right means (and standard deviations in the default normalization) of the series when they have missing data, which is normally the case. This problem is solved here by computing initial values with the available data in every series, estimating the missing data, recomputing their means (and standard deviations), and repeating the process until the maximum difference between the last and the previous means lies below a prescribed threshold.

Apart from the full standardization, climatol allows the “normalization” of the data by means of either only subtracting the mean or dividing by the mean. This latter is more appropriate for variables with a zero lower limit and a biased frequency distribution, as is the case with precipitation or wind, but no series should have mean values lower than 1. (Units may be scaled to avoid this possibility).

Note that the OLS regression with standardized variables can be expressed as $\hat{y}_i = r \cdot x_i$, where r is the correlation coefficient. Therefore, the lower the correlations between the series, the more will be reduced the variance of the estimated data, while RMA regression is free from this effect, which is important for the analysis of extreme values in the homogenized series.

This method is applied to estimate all series in the studied database, and the estimated data can be used to fill in any missing data and to obtain series of anomalies (observed - estimated data) on which to detect outliers or shifts in the mean through the Standard Normal Homogeneity Test (SNHT; Alexandersson, 1986). When the series have missing data, their mean and standard deviation is computed with their available data first, and recomputed after their missing data are filled in, repeating the procedure until a preset degree of convergence is reached. On the other hand, when the SNHT statistic of the series are greater than a prescribed threshold, the series is split at the point of maximum SNHT giving birth to a new series that is incorporated into the data pool. This procedure is done iteratively, splitting only the series with the higher SNHT values at every iteration, until no series is found inhomogeneous. (The SNHT threshold is decided subjectively, since the optimum values depend on the time scale and the variable analyzed. Histograms of SNHT values are provided in an output PDF file to help choosing the thresholds).

Homogenization of daily data

The main problem in the homogenization of daily data is due to the high variability of these series, that lowers the power of detection of shifts in their mean along time. That is why the detection of the inhomogeneities is preferably done on the monthly aggregates of the series, with less inherent variability. In fact the recommended procedure for homogenizing daily temperatures has been to homogenized at the monthly scale and to adjust the daily series with interpolated monthly corrections (Vincent *et al.*, 2002; Brunet *et al.*, 2006), but it did not yield good results when applied to daily peak wind gusts series from Portugal and Spain (Azorín-Molina *et al.*, 2016) and Australia (paper in preparation), nor in experiences with daily precipitation (paper in preparation), that can be attributed to their strongly biased probability distribution frequencies.

But the Climatol package does not need to interpolate monthly corrections, since it just splits the daily series at the break-points detected in the monthly homogenization, and then reconstruct all the series from their homogeneous sub-periods in a final stage by estimating all their missing data.

Therefore the final procedure recommended to homogenize daily series with the Climatol 3.0 package consists in the following steps:

1. Prepare the input files in the format for Climatol.
2. First exploratory analysis of the daily data for quality control.
3. Aggregate the daily data into monthly series.
4. Homogenize the monthly series.
5. Adjust the daily series using the monthly detected break-points.

This methodology will be illustrated in the next sub-sections with the example data accompanying the package.

1. Preparation of the input files for Climatol

Input data must be prepared as indicated in the package manual. If they are stored in a data-base accessible through the ODBC protocol, the provided function db2dat() can be used to generate the input files. Only for the purpose of running the following examples, these files can be generated in the working directory by means of these commands (anything after # is a comment):

```
library(climatol) # load the functions of the package
data(Ttest) #load the example data into R memory space
write(dat, 'Ttest_1981-2000.dat') #save the data file
write.table(est.c, 'Ttest_1981-2000.est', row.names=FALSE, col.names=FALSE) #save the stations file
rm(dat, est.c) #remove the loaded data from memory space
```

2. First exploratory analysis of the daily data for quality control

```
homogen('Ttest', 1981, 2000, expl=TRUE)
```

The user should inspect the output graphics in Ttest_1981-2000.pdf to verify that data does not show weird features. The histogram of anomalies (near the end of the file) may help to choose appropriate thresholds for outlier rejection in the final homogenization step.

3. Aggregate the daily data into monthly series

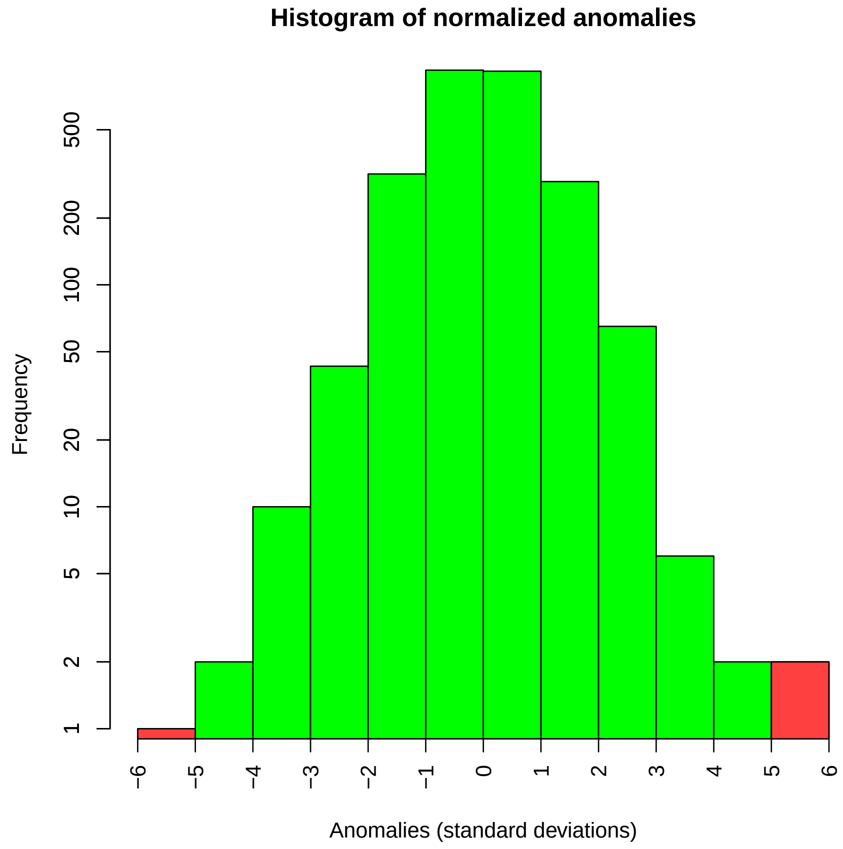
```
dd2m('Ttest', 1981, 2000) #daily data to monthly function
```

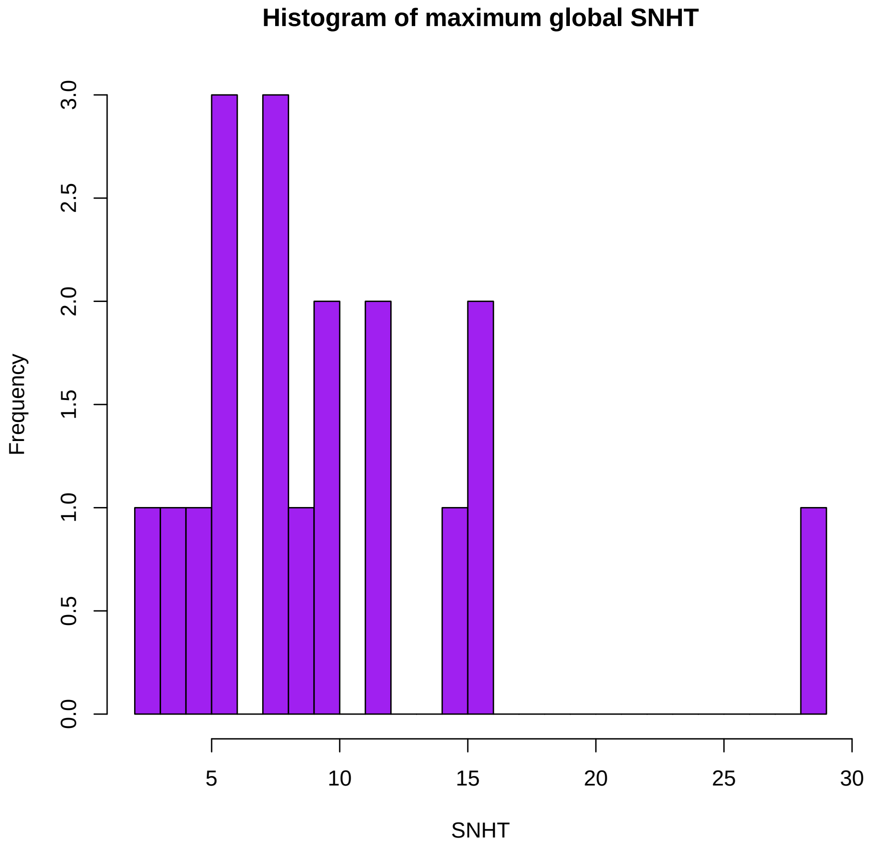
Monthly input files are saved as Ttest-m_1981-2000.dat and Ttest-m_1981-2000.est

4. Homogenize the monthly series

```
homogen('Ttest-m', 1981, 2000) #homogenization of the monthly aggregated data
```

The user should look at the output graphics in Ttest-m_1981-2000.pdf to check if the default threshold values for outlier rejection (dz.max) and SNHT (snht1 and snht2) have been suitable, or re-run the above command specifying better thresholds.





If metadata are available, edit the file Ttest-m_1981-2000_brk.csv containing the break-point list to refine the dates of the shifts in the mean. (Note that not all changes in the history of the station must necessarily produce inhomogeneities).

5. Adjust the daily series using the monthly detected break-points

```
homogen('Ttest', 1981, 2000, metad=TRUE)
```

After every run of the homogen() function, several output files will be generated: a PDF file with lots of diagnostic graphics, lists of breaks and outliers in CSV format, and an R binary file containing raw and homogenized series.

Obtaining products from homogenized data

The user can load the results of the homogenization into the R memory space for any further processing by issuing the command:

```
load('Ttest_1981-2000.rda')
```

But a couple of post-processing functions are provided in the package to help in obtaining common products from the homogenized series, either directly from the daily series, or from their monthly aggregates, which can be generated by:

```
dd2m('Ttest', 1981, 2000, homog=TRUE)
```

Examples for getting some statistical products:

```
dahstat('Ttest', 1981, 2000) #means of the daily series
dahstat('Ttest', 1981, 2000, mh=TRUE) #means of their monthly aggregates
dahstat('Ttest', 1981, 2000, mh=TRUE, stat='tnd') #monthly OLS trends and p-values
dahstat('Ttest', 1981, 2000, stat='q', prob=.2) #first quintile of daily values
```

Another function is provided to obtain homogenized grids, but you must define your grid limits and resolution first, as in:

```
grd=expand.grid(x=seq(-109,-107.7,.02), y=seq(44,45,.02)) #grid specification
library(sp) #load needed package for the following command:
coordinates(grd) <- ~ x+y #convert the grid into a spatial object
```

Now grids can be generated (in NetCDF format) with:

```
dahgrid('Ttest', 1981, 2000, grid=grd) #grids with daily time steps
dahgrid('Ttest', 1981, 2000, grid=grd, mh=TRUE) #grids with monthly time steps
```

These grids are built in normalized, dimensionless values. You can obtain a new file with temperatures in degrees Celsius with externals tools, such as the 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 will be 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 before using them to undo the normalization of the grids.

Conclusion

The Climatol package, that can be freely installed through the means provided in any running R environment, is a convenient tool to homogenize monthly and daily series without much effort, and is adapted to use series with a very high amount of missing data. Although most parameters of their functions are set with default values, the user should tune them to the variable under study and its time resolution to optimize the results. (https://cran.r-project.org/web/packages/climatol/climatol.pdf holds the documentation in PDF format).

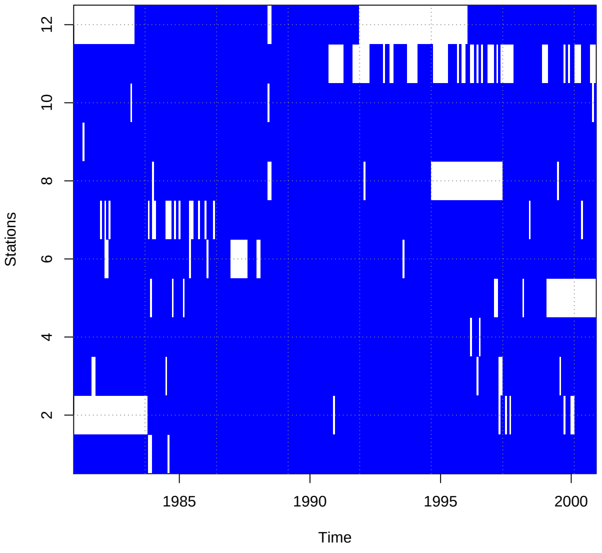
References

Azorín-Molina C, Guijarro JA, McVicar TR, Vicente-Serrano SM, Chen D, Jerez S, Espirito-Santo F (2016): Trends of daily peak wind gusts in Spain and Portugal, 1961-2014. *Journal of Geophysical Research Atmospheres*, DOI: 10.1002/2015JD024485

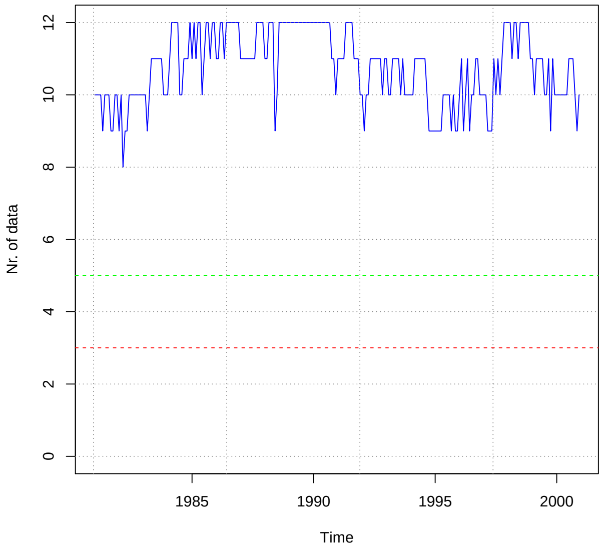
Brunet M, Saladié O, Jones P, Sigró J, Aguilar E, Moberg A, Lister D, Walther A, Lopez D, Almarza C (2006): The development of a new dataset of Spanish daily adjusted temperature series (SDATS) (1850-2003). *Int. J. Climatol.*, 26:1777-1802.

Vincent LA, Zhang X, Bonsal BR, Hogg WD (2002): Homogenization of daily temperatures over Canada. *Journal of Climate*, 15:1322-1334.

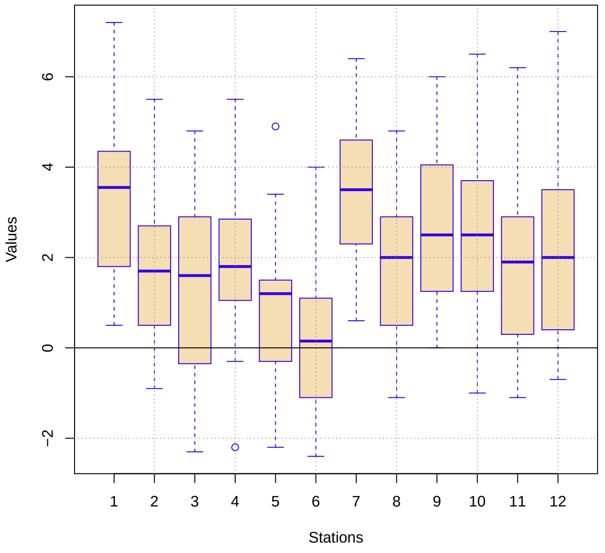
Annex: More examples of the graphic output



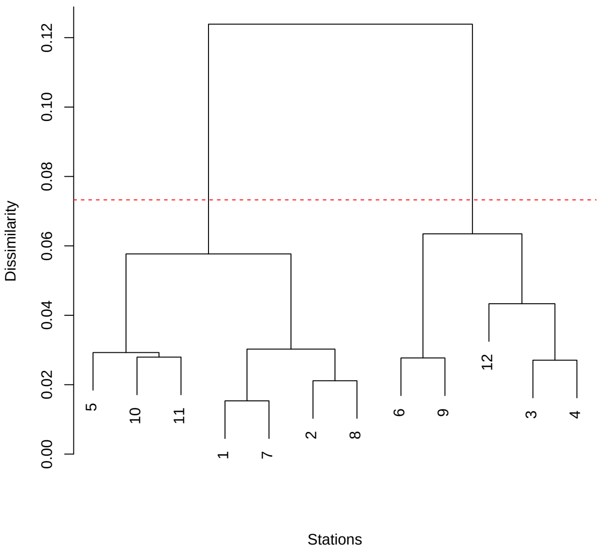
Ttest−m data availability



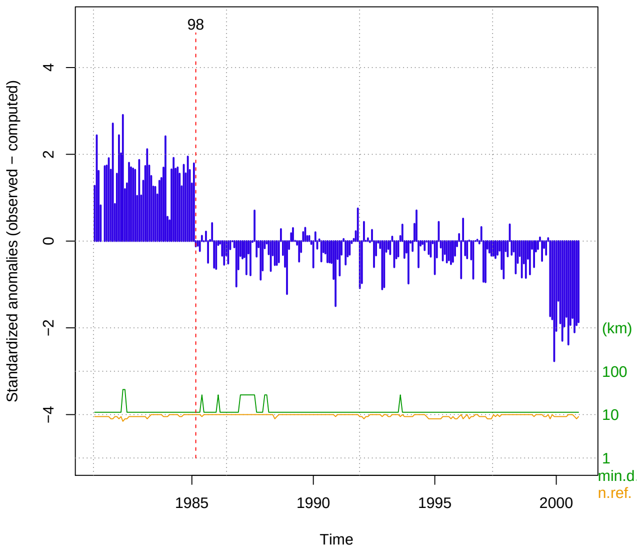
Nr. of Ttest−m data in all stations



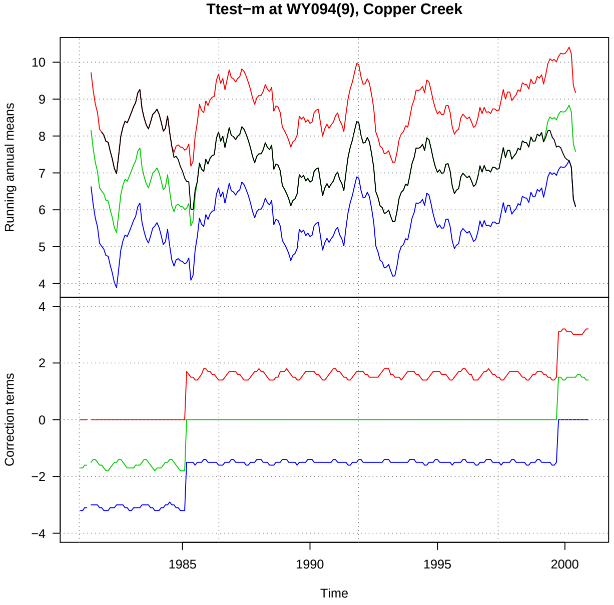
Data values of Ttest−m (Mar)



Dendrogram of station clusters



Ttest−m at WY094(9), Copper Creek



Ttest−m at WY094(9), Copper Creek