CLIMATOL: FREE SOFTWARE FOR ERROR DETECTION AND HOMOGENIZATION OF CLIMATOLOGICAL DATA

José A. GUIJARRO. *Instituto Nacional de Meteorología. CMT en Illes Balears, Palma de Mallorca, Spain (jaguijarro@inm.es)*

6th European Conference on Applied Climatology (Ljubljana, Slovenia, 4-8/Sep-06)

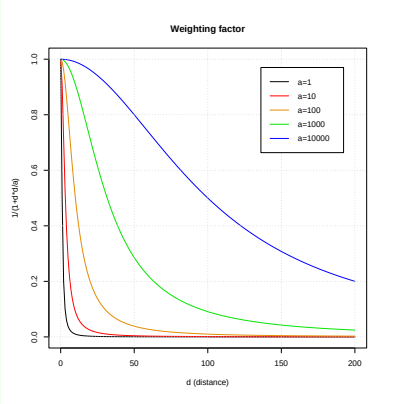
1. INTRODUCTION

Any climatological study using more or less long series of observations has to face the study of their homogeneity, to fill the missing data and detect anomalous behaviours produced by non meteorological causes. A set of functions to address this tasks has been developed in this work, implemented in a module of the statistical package R, with the advantage of been multi-platform and able to run under different operating systems (GNU-Linux, Solaris, Windows, etc), hence enabling its use in a wide range of working environments. Moreover, and not least, its code is open (under the GPL licence) and can be modified by other climatologists to meet their needs and, at their will, make their contributions available to the community. This is facilitated by the wide variety of statistical tests and graphic representations of the R package, and its simple yet powerful programming language to develop new functionalities.

2. METHODOLOGY

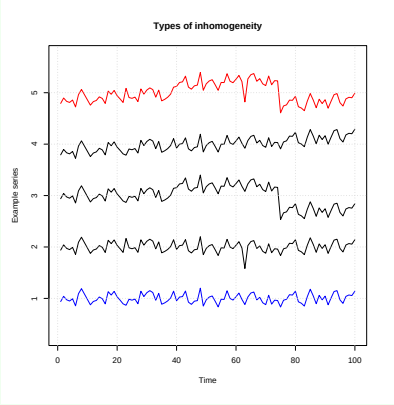
The provided functions are devised for the treatment of monthly data series, though can also be used for either daily or seasonal/annual data. The basic process consist in a comparison of each series with a reference series built from all the other. The commonest criterion to select the reference stations to built the series is the correlation coefficient, but here the inter-station distance is preferred, to allow the use of the nearest stations even it they do not have a common period of observation long enough to enable correlation computations.

But first all the series are standardized by one of the following methods: 1) Deviations from the mean; 2) Proportions of the mean (only for means greater than 1); 3) Full standardization (subtract the mean and divide by the standard deviation). Then the reference series (standardized estimated values for each station) are computed as a weighted averages of the standardized data of all the other stations. The weighting factor is an inverse function of the distance d: $1/(1+d^2/a)$, where parameter a allows the investigator to modulate the relative weight of nearby stations to the more distant ones:



As means (and standard deviations) vary depending of the period of observation of the stations, this process is repeated iteratively until their averages get stable.

Afterwards, each original series is compared with its estimate (reference), studying the differences of their standardized values. If a series is homogeneous (and so are their neighbors), the difference series should behave as a random normally distributed series (a white noise). In real cases what we usually find are point errors, shifts in the mean, trends, or rather a combination of them all:



Types of inhomogeneity in climatological series. 1) Homogeneous series; 2) Point errors; 3) Shifts in the mean; 4) Trends; 5) All together.

3. USE OF *CLIMATOL*

`http://cran.r-project.org/src/contrib/PACKAGES.html` is the page where you can find the package. Once downloaded and installed, the first thing to do is to prepare the data files. Monthly (or whatever) data must be provided in a text file named `VAR_AI-AF.dat`, where `VAR` stands for any acronym of the involved climatological variable, and `AI` and `AF` are the two last digits of the initial and final year of the data. Data are stored station by station, and chronologically within each station block, with no constraints about where to cut lines. Missing values must be specified as `NA` (the usual way in R).

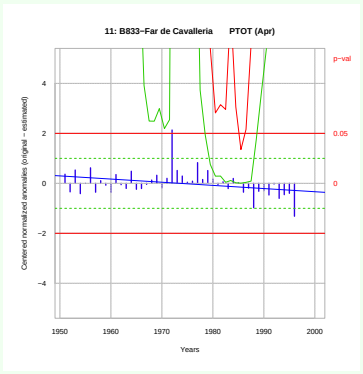
Station coordinates and names on its turn go in a file `VAR_AI-AF.est`, in lines with the structure `'X Y Z ID NAME'` (one per line), where `X` and `Y` are the UTM coordinates in km, `Z` the altitude in m, `ID` an identifier of the station, followed by its `NAME` (between double quotes if it contains any blank space):

```
607.60 412.20   60 B801  "Sant Lluís"
606.00 414.70   50 B802  "Maó Llucmaçanes"
611.70 413.80   16 B803  "Far Port de Maó"
... (etc)
```

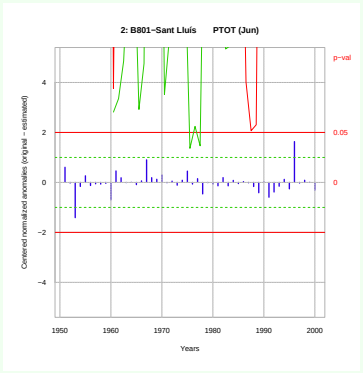
Then, after having started the R session and loaded the package with the command `library(climatol)`, we can proceed to study the homogeneity of monthly precipitation (`PTOT` acronym was chosen) between 1951 and 2000, standardizing with the rate method (`ttip=2`), and looking at the graphical representation of the series (`graf=TRUE`):

```
> depudm("PTOT", 1951, 2000, ttip=2, graf=TRUE)
```

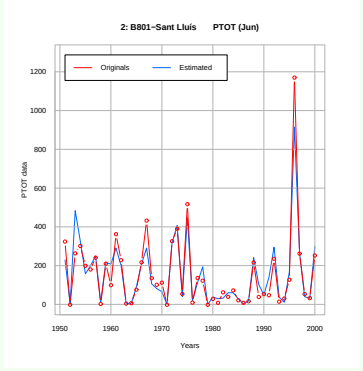
Here are some examples of the graphics obtained: 1) Series with a point error, trend and/or shift in the mean. (Green and red lines show the 20 and 10 terms of running window t-tests):



Some point anomalies, but no (very) significant shifts or trend:



Plot of the original and estimated values of the former series:



Other possible options are to apply a square root transformation to the data if we work with precipitation, or to automatically purge the data substituting all data differing more than 2.5 (or other value) standard deviations from the reference series by their estimated values:

```
> depudm("PTOT", 1951, 2000, auto=TRUE, dz.max=2.5)
```

This will yield a file with the data purged from point errors and missing values filled named `PTOT_51-00.dep`, from which we can obtain the normal averages for the international period 1961-90 listed in a file `PTOT_61-90.med` with the command:

```
> depstat("PTOT",1951,2000,1961,1990)
```

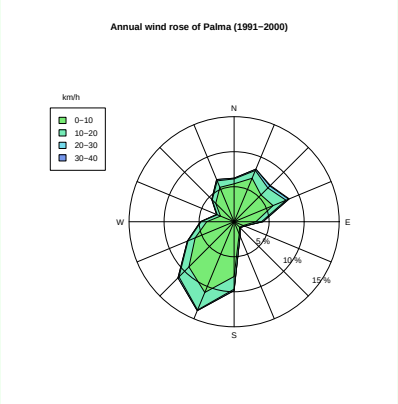
However, climatological data base homogenization must be faced with great doses of patience, working iteratively: looking for and correcting only major inhomogeneities in the first place, and refining the process step by step. Otherwise we will be exposed to taking as inhomogeneities anomalies due to errors in nearby stations. Parameter a must be chosen quite big at the beginning, in order to let the errors get diluted in a fairly great number of series, and small in the final steps, to take into account the peculiarities of the local climatology and avoid a dramatic loss of the series variances.

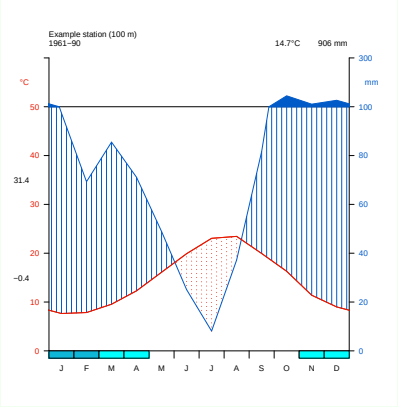
4. FURTHER ANALYSIS

Objects created by *CLIMATOL* (original data, `dat.d`; normalized data, `dat.z`; estimated data, `dat.e`; ...) will remain resident in the memory space during the rest of the R session while not explicitly removed. Therefore, the power and variety of R functions can be readily applied to further analysing our climatological series: correlation analysis, time series analysis (smoothing, lag correlation, ...), etc, or plotting histograms and many other graphical representations of the data.

5. ADDITIONAL FUNCTIONS

Though the main objective of *CLIMATOL* is to study the homogeneity of the climatological series, a couple of additional functions have been added to help in descriptive climatology work, implementing the plot of rose-winds and Walter & Lieth diagrams:





REFERENCES

Guijarro JA (2004): Climatol: Software libre para la depuración y homogeneización de datos climatológicos. In García-Codron *et al.* (Eds.), *El clima, entre el Mar y la Montaña*, Asociación Española de Climatología, A-4:493-502.

Guijarro JA (2006): Homogenization of a dense thermopluviometric monthly database in the Balearic Islands using the free contributed R package "CLIMATOL". Fifth Seminar for Homogenization and Quality Control in Climatological Databases, Budapest, 29-May to 2-June. (In press).