ACMANT homogenization software Manual, version 5.0 subversion 10

(monthly homogenization, automatic or interactive)

2021

Content

1. Introduction

ACMANT (Applied Caussinus – Mestre Algorithm for homogenizing Networks of climatic Time series) is known about its outstanding accuracy among tested automatic homogenization methods (Domonkos et al., 2021). This method is suitable for the homogenization of several surface climate variables, namely temperature, precipitation amount, relative humidity, wind speed, sunshine duration, radiation (of any kind) and atmospheric pressure. This subversion, ACMANTv5_S10, is applicable for the homogenization of monthly time series, either in automatic or interactive mode. ACMANTv5 offers metadata use together with the statistical procedure. Further principal properties of ACMANTv5 are:

(i) ACMANTv5 includes an automatic networking procedure, therefore the spatial similarity in the climate of the source area of the input data is not required. An exception will be mentioned in Sect. 2.1. Input datasets may have any size between 4 and 5000 time series.

(ii) ACMANTv5 removes automatically the physically impossible values from the input, and users may define thresholds for climatic outliers. Beyond the control of climatic outliers, ACMANTv5 offers the filtering of spatial outlier values and short-term spatial outlier periods.

(iii) Interrupted homogenization of datasets can be continued without repeating the homogenization of the series whose homogenization results have once been produced.

(iv) Input time series may cover varied time periods, and strict rules must be followed in their preparation.

(v) Metadata can be used either in interactive or automatic modes. Regarding the automatic mode, the list of metadata dates must be prepared in the prescribed form. If metadata list is not provided and automatic homogenization is opted, the program will run without metadata use.

(vi) The major part of the theoretical structure of the previous ACMANT version (Domonkos, 2020) has been kept. However, a novelty of ACMANTv5 is the combined time series comparison of time series (Domonkos, 2021) in the first cycle of the homogenization procedure. This provides improved accuracy when spatially systematic biases affect data homogeneity (Domonkos, 2021).

(vii) The software offers default solutions regarding several details of the homogenization, but users may select personalized options. The list of optionally alterable characteristics are as follows:

- Minimum of spatial correlations (default = 0.4);
- For input datasets of larger than 40 time series, or those including time series with lower than minimum threshold spatial correlations, the software divides the input dataset to the set of the default networks. Each network includes one candidate series and a limited number of (max. 98) and sufficiently correlating neighbour series. The content of the networks is editable manually;
- Thresholds for climatic outliers (default = thresholds of physically possible range)
- Time series subjected to homogenization (default = all);
- Time series subjected to interactive homogenization (in interactive mode, default = all);
- Gap filling;
- Items of homogenization output;

Any input dataset for ACMANTv5 should be free of strikingly large, climatically impossible outlier values, values generated by spatial interpolation or by a previous homogenization procedure in which spatial comparisons were applied. Input datasets should preferably contain all the existing observed climatic data of the regions under study, as the efficiency and accuracy of homogenization increase with the spatial density of data.

If you have questions, comments or any unexpected experience in using ACMANTv5 or ACMANTv4.3, please do not hesitate to contact with the method developer:

Peter Domonkos
dpeterfree@gmail.com

References:

Domonkos, P**.** 2020: ACMANTv4: Scientific content and operation of the software. 71pp. https://github.com/dpeterfree/ACMANT.

Domonkos, P. 2021: Combination of using pairwise comparisons and composite reference series: a new approach in the homogenization of climatic time series with ACMANT. In: Atmosphere special issue: Application of Homogenization Methods for Climate Records (ed. Domonkos, P.). https://doi.org/10.3390/atmos12091134.

Domonkos, P., Guijarro, J.A., Venema, V., Brunet, M., Sigró, J. 2021: Efficiency of time series homogenization: method comparison with 12 monthly temperature test datasets. J. Climate, 34, 2877-2891. https://doi.org/10.1175/JCLI-D-20-0611.1.

2. Content of software package ACMANTv5_S10

2.1. Essence of scientific content

This software homogenizes the first moment of climatic variables listed in Sect. 1, provides the probable positions of sudden shifts in the means (hereafter breaks) and short-term large size inhomogeneities (outlier values and outlier periods), assesses the size of breaks and outliers, provides the homogenized time series and fills data gaps with interpolated values (optional). For temperature, relative humidity, wind speed, sunshine duration, radiation and atmospheric pressure, the homogenization is based on the concept that inhomogeneities result in linear biases. Although the linearity is not perfect, the use of linear model gives good results in the homogenization of annual and multiannual means and linear trends. For precipitation homogenization, the bias is supposed to be multiplicative.

Seasonally varying biases may be treated a) with the inclusion of a sinusoid model of seasonal changes (S-model), b) with individually estimated monthly correction terms (for irregular seasonality, I-model), or c) with the exclusion of the seasonal variation of correction terms (flat, F-model). Users must choose between these three options according to the advice below:

In the middle and high latitudes not only the mean climate, but also the inhomogeneities often have quasi-harmonic shape with modes around the solstices. In these regions, the selection of S-model is recommended for the homogenization of mean temperature, maximum temperature, relative humidity, sunshine duration, radiation and sea level pressure, while the I-model is recommended for the homogenization of minimum temperature and wind speed. In the homogenization of precipitation there is no free options of seasonality model. There, the F-model is applied when rain is the dominant precipitation form in all the year round, while distinct corrections of inhomogeneities for the rainy season and snowy season are applied where the dominant precipitation form is snow in some part of the year. The starting and ending months of the snowy season are defined by the user.

For tropical belt regions and regions of subtropical monsoon climate, the application of I-model is recommended for all climate variables except for precipitation.

When the source area of the input dataset covers regions for which varied inhomogeneity seasonality models are favourable, the division of the initial dataset is recommended, since in ACMANTv5 varied seasonality models cannot be applied within one homogenization procedure.

In the homogenization of sunshine or radiation the F-model cannot be chosen.

2.2. Structure of directories and files

The software is put into a directory named “ACMANTv5.0”. In its main directory one can find the main executive file “ACMANT5run.BAT” together with 4 subdirectories (hereafter they are referred to as directories) and the file of this guide. The BAT file of the main directory opens an executive file of directory “Works”, which manages the homogenization procedure using further 10 BAT files and 10 executive files.

Directory “Input” is for the deposition of input dataset, while the homogenization products will appear in directory “Output”. Directory “Partial results” is for keeping some partial results of the homogenization procedure there. Users use partial results when they opted for interactive homogenization.

Within directory Works one can find several files and one subdirectory “Auxiliary”. Note, however, that during the homogenization procedure the temporary subdirectory “Subnetws” is created when the procedure divides the input dataset to networks. This subdirectory is deleted at the end of a homogenization procedure. Users may do some editions in subdirectory Subnetws, but they should not touch any other part of directory Works.

3. Preparation of input data

Before starting a homogenization procedure with ACMANTv5, you are expected to prepare your data in the required format and put them into directory Input. It is strongly recommended to remove any other files from “Input” (e.g. the input data of an earlier homogenization) before this step. In Sects. 3.1 – 3.2, the rules for the preparation of input climatic datasets are presented. Note that ACMANT can be applied if the amount and temporal compactness of spatially fairly comparable data reaches certain (rather low) thresholds. The thresholds are shown in Sect. 3.3. Sect. 3.4 presents the rules for preparing metadata lists. Users may change default parameterization with showing the requested parameters in request files. The rules for editing such files are shown in Sects. 3.5 – 3.8.

Note again that all kinds of input information must be put into directory Input before starting a homogenization procedure.

3.1. Rules for preparing monthly climate data files

i) Number of time series per dataset is between 4 and 5000. As a general rule, a time series or its section will be homogenized if it has at least 3 neighbour series with temporally coincidental periods of observed data and sufficient spatial correlations (default threshold = 0.4) with the candidate series. A neighbour series or its section is usable only when that can be homogenized based on the same conditions as which are prescribed for the candidate series.

ii) Length of time series (may include data gaps): between 10 years and 200 years.

iii) Time series are separated into distinct files according to observing stations. In other words: 1 file is for 1 station time series only.

iv) Each file must contain a headline with the name of the station or other station identifier. There is no restriction for the length of the headline, but only its first 40 characters will be usable for identifying the station.

v) Default units of climatic elements: °C for temperature, mm for precipitation, % for relative humidity, hour for sunshine duration, m/s for wind speed and hPa for atmospheric pressure. You are allowed to use other units (except for precipitation) with the following precautions: a) It is indispensable to use the same unit throughout a given homogenization process. b) The wrong use of units might cause overshooting of physical thresholds inbuilt in programs or overshooting of user defined climatic thresholds even when the input data is correct. c) Climatic values below -998 are considered missing data and those higher than 9999 cannot be treated by the software.

vi) Each line (after the headline) contains a pre-determined number of values (i.e. date identifier(s) and 12 monthly values. The values in the same line are separated with one or more space characters. TAB also can be applied for separation (optional). The use of other characters (comma, semicolon, etc.) is not allowed. Date identifiers are shown with whole numbers, while the data of climatic variables are usually presented with 1 decimal preciseness, but note that other forms are also accepted. Decimals are separated by dot from the other digits. E.g. value – 4°C can be presented in any form of – 4, – 4.0 or – 4.00.

vii) Missing values are represented by –999.9.

viii) Trace precipitation must be coded with 0.

ix) Gap filling with missing value code

Input time series must be temporally continuous from the first calendar month of the year of the first observed data until the last calendar month of the year of the last observed data, although some exceptions are allowed (Sect. 3.2).

Best Site Station

1951 -999.9 -999.9 4.5 7.3 15.6 17.6 -999.9 -999.9 -999.9 -999.9 -999.9 -999.9
1952 -1.5 0.0 7.1 9.8 10.9 15.6 19.8 20.3 13.8 -999.9 8.1 0.4
1953 -999.9 -999.9 -999.9 -999.9 -999.9 -999.9 -999.9 -999.9 -999.9 -999.9 -999.9 -999.9
1954 -999.9 -999.9 -999.9 8.6 14.2 16.8 18.8 18.5 16.4 9.5 3.6 -0.2
…….
…….
…….

In Best Site Station the temperature observation started in March 1951, but there are several gaps in the first few years. Please remember that it is not allowed to jump over the year 1953, in spite of that there was no observation at all in that year. It is also important to note that the time series starts with the January of the first observed data (missing data codes for January and February of 1951).

x) Name of files: "ANAMEJJJJ.txt".

- "ANAME" is the name of the dataset. This substring is always 5-character wide and may contain any alphanumeric characters.

- "JJJJ" is the serial number of time series, it always consists of 4 digits.

- ".txt": these 4 characters are fixed, they are always ".txt"

Example: If the name of the dataset is "Lucky" and it consists of 275 time series, then the names of the time series must be:

Lucky0001.txt
Lucky0002.txt
Lucky0003.txt
……..
……..
Lucky0274.txt
Lucky0275.txt

As it was shown by the example of "Best site station", each file includes a headline, and thereafter each line is for one year section of the data series. In each line of the data series first the calendar year is shown, and then the 12 monthly climate data of the year.

3.2. Exceptions when input time series are allowed to be fragmented

Before starting the homogenization, the user must define the target period of the homogenization (Sect. 4.2). The input data out of the target period are allowed to be fragmented if they comply with the following conditions.

i) Before the beginning of the target period: If data for at least one date is included from a given year, then all the monthly data of that year must be represented either with observed data or with missing data code. However, years without any observed data can be jumped out.

ii) After the target period: If the input time series includes the last year of the target period, then the reading of that file will be terminated with the data of that year, and thus anything might be present in the later part of the file. However, if the series of continuous data terminates before the last year of the target period, the file must be ended together with the end of the series of continuous data.

Note that if you use any of these alleviations, this might cause problems if you would like to repeat the homogenization of the same time series with varied target periods.

3.3. Conditions for treatment and homogenization by ACMANTv5

i) Requirements for the treated period of a time series

Sometimes time series include large missing data fields at the beginning and/or at the end of their period. During the homogenization procedure, the input time series are split into three sections: a treated section (referred to as treated period), and two not treated sections, one in the beginning part and another one in the ending part of the time series. Not treated sections remain out of most steps of the homogenization procedure, yet inhomogeneity adjustments and data completion by spatial interpolation may be applied in them. When the number of missing values is low, the whole period of the time series is treated period. The ratio of missing values within the treated period is limited according to the following rules:

a) In the first *k* year and last *k* year of the treated period the expected ratio of available observed data is higher than 25% for any *k* between 1 and 20, or between 1 and the length of the treated period if the latter is the shorter. For instance, if the first year is 1953 and the last year is 2002, the minimum number of observed monthly data for periods 1953-1953, 1953-1954, 1953-1967 and 2000-2002 are 4, 7, 46 and 10, respectively. Also, at least two neighbour series with sufficient ratio of observed data are needed.

b) The minimum length of a treated period is 10 years.

c) The minimum number of observed monthly data is 9 for each month of the year.

d) The minimum number of the total of observed monthly data is 114.

e) In case of too few or too fragmented data, the above conditions might remain unsatisfied throughout the time series. In this case, the time series will fully be excluded from the homogenization procedure: for these series neither inhomogeneity adjustments nor data completion are applied. Note that the homogenization procedure does not stop for the exclusion of one or more time series.

ii) Requirements for homogenizing series or their sections

a) If a station series (or its section) does not have at least 3 neighbour series with sufficient spatial correlation, that series (or its section) will not be homogenized, but otherwise the program will run normally. If none of the series can be homogenized, the program will stop with the adequate control message. Note that when no section of a candidate series can be homogenized, neither inhomogeneity adjustments nor data completion are applied to that series.

b) When there is a section of candidate series without 3 neighbour series, that section will not be homogenized. Sometimes both before and after that section the number of comparable time series is higher, thus two distinct sections could be homogenized. However, in such cases only one section, i.e. the late section will be homogenized, even if a longer section could have been homogenized in the early part of the candidate series. If more than one separate homogenizable sections are found by ACMANT, then this information will appear in the output file “Control message.txt”.

The section of time series which can be homogenized according to the presented conditions is referred to as the homogenized period of the time series. Often the whole period of the input time series is homogenized period. Not treated sections cannot be part of the homogenized period. The use of data falling within the treated period but out of the homogenized period is strongly limited in the homogenization procedure.

3.4. Metadata list

Users may provide a list of metadata dates before starting a homogenization procedure. When such a list is provided, the homogenization procedure will use the metadata dates in combination with the break dates detected by statistical methods.

In the evaluation of pairwise detection results, the value of a piece of metadata equals to 1 statistically detected break of high certainty (score = 1.0). The minimum threshold score for validating a break position at that phase of the homogenization procedure is 2.1 (see the explication of pairwise detection result scores in Domonkos, 2021). Later in the homogenization procedure, metadata dates are used also to refine the dates of statistically detected breaks.

All the metadata of a given input dataset must be shown in one common file. Metadata must be shown with daily preciseness. Even when you do not know the exact date, the estimated date must be shown with daily preciseness.

The name of metadata list file: “ANAMEmeta.txt”, where

- “ANAME” is the dataset name (the same as for the climatic data)
- “meta.txt” is constant

The size of this file depends on the number of known metadata dates, and there is no need to order the pieces of metadata either according to time or station serial numbers.

In each line of the file, 4 whole numbers must be shown, they are from left to right: a) serial number of station, b) day, c) month, d) year. For instance, LUCKYmeta.txt has 20 time series, but only 4 metadata:

14 31 11 1970
14 12 4 1970
5 31 12 1988
9 7 11 1963

The example shows that more than one metadata may occur for the same year and same station, but note that too high metadata frequency is generally neither recommended nor reasoned. The maximum number of pieces of metadata for a station is 40, while for an entire input dataset 40,000. Warning: Any erroneous date will stop the program with the relevant error message. However, when metadata file is not provided, it does not cause problem, and the program will run without metadata use.

3.5. Thresholds for climatic outliers

ACMANTv5 can be run with the inbuilt physical thresholds, or user defined climatic thresholds can be applied.

i) Inbuilt thresholds
temperature: -98 and 60
precipitation: 0 and 6000
relative humidity and wind speed: 0 and 100
sunshine and radiation: 0 and 744
atmospheric pressure: 0 and 1099

ii) User defined thresholds
User defined thresholds overwrite the inbuilt threshold values. While the inbuilt thresholds are season-independent, user defined thresholds consist of monthly low threshold and high threshold values for each calendar month. These thresholds are written to the request file “Thresholds.txt”. The file must include 12 lines, with the serial number of calendar month and the monthly low threshold and high threshold values in each line. An example of Thresholds.txt for the homogenization of temperature maximums in a Mediterranean region:

1 -10 35
2 -10 35
3 -10 40
4 0 50
5 5 50
6 10 55
7 10 55
8 10 55
9 10 50

10 5 45
11 -5 40
12 -10 35

Note that lower (higher) threshold values than -997.9 (9999) cannot be defined, and in precipitation homogenization negative values are not allowed.

When the use of inbuilt physical thresholds of the software is preferred, the input package should not contain Thresholds.txt.

3.6. Spatial correlation

Two kinds of spatial correlations are used in ACMANT, one is for homogenization, and another is for gap filling. The ways of their calculation differ (Domonkos, 2020). For gap filling, the minimum threshold is 0.4 and users cannot modify it. For homogenization, the spatial correlations are calculated from de deseasonalised values of monthly increment series (see more in Domonkos, 2020). In precipitation homogenization the spatial correlations are calculated from the re-scaled data by a semi-logarithmic transformation (Domonkos, 2020). The default minimum threshold is 0.4, but users can alter this either for the entire dataset, or for some selected networks. If you would like to use another minimum threshold for the entire dataset, then you must show it in the relevant request file.

The filename is constant: “Rth.txt”. The requested spatial correlation threshold must be written into its first line without adding any other thing. The program accept any value between 0.1 and 0.99. When Rth.txt is not shown in directory Input, or a valid correlation threshold cannot be found in that, the program will use the default correlation threshold.

In interactive mode, users may define individual correlation thresholds for selected networks, this will be shown in Sect 4.6.

3.7. Target series for homogenization

For relatively small datasets, ACMANT usually homogenizes all the series together without grouping them into networks.

When networks are constructed by the program, still all the series of a network are homogenized together, but in networks the results are saved only for the candidate series of the network. It is because a principle of the network construction in ACMANT is that a network is optimal for one selected candidate series, and therefore the number of networks equals to the number of the time series of the input dataset (see also Domonkos, 2020).

In homogenizing datasets of larger than 40 time series (multi-network datasets), users may save time with defining which time series are expected to be homogenized when not

all the series are needed to be homogenized. The serial numbers of the requested time series must be written into the request file “ANAMEtarget-homg.txt”. In this filename

- “ANAME” is the dataset name (the same as for the climatic data)
- “target-homg.txt” is constant

Each line of the file must contain one serial number. You may write the relevant serial numbers in any order. If the same serial number is shown more than once, this does not cause error. However, error will occur if a) an invalid serial number is shown, then the program stops with error message; b) the file includes more lines with the presentation of valid serial numbers than the number of time series (*N*) in the dataset (it is possible by repeated presentations of some serial numbers), then the program considers only the first *N* serial numbers.

Example: dataset WHITE includes 1622 time series, but you need the homogenization results only for series 7, 78, 922, 926, 1473 and 1496. Then WHITEtarget-homg.txt file must be created with the following content:

7
78
922
926
1473
1496

If WHITEtarget-homg.txt is not shown in directory Input, ACMANT will homogenize all the 1622 networks. If you write 1622 into the first line of WHITEtarget-homg.txt and nothing else, then ACMANT will homogenize only the last network.

3.8. Target series for interactive homogenization

To perform interactive homogenization, you must select the interactive option (see Sect. 4.2). Then the program will give access to define specific correlation thresholds for networks and interact with the homogenization procedure in the homogenization of any candidate series. In the homogenization of a multi-network dataset, the program suspends its running twice in the interactive homogenization of a candidate series, and continues the running when the user gives permission. If the user does not define target series for interactive homogenization, all the candidate series will be homogenized in interactive mode, which can be tiring when a dataset consists of so many time series as for instance WHITE. You may input a request showing the serial numbers of time series for interactive homogenization. The filename for this request is “ANAMEtarget-edit.txt”. It is a similar file with similar edition rules to the request file presented in Sect. 3.7, but here the constant part of the name is “target-edit.txt” instead of “target-homg.txt”. Each line of the file will contain one serial number.

It is important to note that in ANAMEtarget-edit.txt users may write zero into the first line (and only into the first line). This code means that the user prefers to define specific

correlation thresholds for (some) networks, but he/she does not want to interact with the homogenization procedure in other ways. Once you have written zero into the first line, you should not write any other thing into the file (the program stops reading the file once it has read the zero).

The homogenization request of ANAMEtarget-homg.txt cannot be overwritten by the edition permission request of ANAMEtarget-edit.txt. Coming back to the example of Sect. 3.7, if you want the interactive homogenization of series 78, 916, 922, and 926 of WHITE, you will write into WHITEtarget-edit.txt

78
916
922
926

However, if you do not include 916 in WHITEtarget-homg.txt, the program will exclude that series. With the shown parameterization, ACMANTv5 will provide the interactive homogenization of series 78, 922 and 926, the automatic homogenization of series 7, 1473 and 1496, while it does not generate output results for the other series of the dataset.

4. Running ACMANTv5_S10

4.1. General conditions

The software package with its directory structure must be placed in your computer. Saving a copy is advised. The preferred environment is Windows, since the compiled FORTRAN programs of ACMANT are directly executable under Windows. Under Linux, de program can be run by “wine” application.

The computational time demand widely varies: for datasets of up to 40 time series 1-network homogenization is performed, which may last from a few seconds to a few minutes. In multi-network homogenization the typical time demand is from a few hours to a few days, but for very large datasets (>1000 series) the time demand may be several weeks. In multi-network homogenization the time demand increases linearly with the number of time series and exponentially with the length of the target period. The network size has strong impact on the time demand, which must be taken into consideration in network editions.

Even if you use automatic mode, you must introduce some parameters manually at the beginning of the homogenization procedure, these are described in Sects. 4.2 - 4.3. Thereafter the program can do everything without user assistance, and thus the content of Sects. 4.4 – 4.7 may be less interesting for the users of automatic mode.

4.2. Initiation I. Starting a new homogenization

Once the input dataset and supplementary input files have been prepared, they are placed to directory “Input”. Please check that directories “Input” and “Output” do not include data of earlier homogenizations.

The homogenization starts with clicking on “ACMANT5run.BAT”. The program will ask for typing some parameters on the screen. When the answer includes (or may include) letters, please introduce the answer from the left end of the line, without space characters.

i) Climatic element: Two characters. It is “TT” for temperature, “RR” for precipitation amount, “HH” for relative humidity, “FF” for wind speed, “SS” for sunshine duration and radiation, and “PP” for atmospheric pressure.

ii) Name of the dataset: 5 characters, the same as which is included in the file names of the input time series.

iii) Number of stations: The number of time series in the input dataset must be given here.

iv) – v) First calendar year of time series and number of years in time series: These two parameters determine the target period of homogenization. If all the input time series cover the same time period, it is straightforward to define that period as target period, but the period of observations might vary widely in large datasets. You are allowed to choose either shorter or longer target period than which is covered with input data, there are two requirements only: a) the length of target period must fall between 10 and 200 years; b) All dates of the metadata list (if metadata list is provided) must fall within the target period.

vi) Seasonality of inhomogeneities: It can be sinusoid (“S”), irregular (“I”) or flat (F). See explanation and advice in Sect. 2.1. --- This question does not appear in precipitation homogenization. Furthermore, option “F” is not offered in sunshine duration / radiation (SS) homogenization.

vii) – viii) First and last snowy months: These questions appear only in the homogenization of precipitation (RR). Months are coded here with their serial numbers within a calendar year. In case of no snowy season, the answer is “0” to the first snowy month, and in this case the program will not ask for the last snowy month. Note that if the snowy season would have a non-zero, but shorter than 3-month duration according to the user introduced parameters, the snowy season length is extended by ACMANTv5 with 2 months, by adding the adjacent calendar months to the pre-defined snowy season.

ix) Would you like outlier filtering?: The answer is 1 character, “Y” if yes and “N” if no. The recommended response here is yes.

x) Would you prefer default output package?: The answer can be “Y” or “N”, and if “Y” is chosen, the program starts immediately the homogenization procedure in automatic mode. The default output package comprises the

- homogenized time series in the same format as that of the input time series;

- list of neighbour series and their spatial correlations with the candidate series;
- list of detected breaks and outliers.

A more detailed description of the default output package is presented in Sect. 5.1.

If you prefer interactive homogenization, you may not choose the default output package, and either if you prefer other output items (see Sect. 5.2) than those of the default package. When the default output package has not been accepted, ACMANTv5 puts further questions:

xi) Gap filling. “0”= never, “1”= within the homogenized period, “2”= completion of series for the target period, “3”= all kinds of output items of the previous three options.

xii) Table of confidence indicators: “Y” if yes, “N” if no. --- Indicator values show for individual monthly values if they are homogenized observed data, interpolated data, etc. The indicators are whole numbers between 0 and 9. Lower values (except 0) mean higher reliability, see detailed description in Sect. 5.3.

xiii) List of detected breaks and outliers. “Y” if yes, “N” if no. This item is a part of the default output package, but if you opt here for “N”, the item will not be output.

xiv) List of neighbour series and their spatial correlations with the candidate series. “Y” if yes, “N” if no. This item is a part of the default output package, but if you opt here for “N”, the item will not be output.

xv) Edit option for partial results? “N”= No, “E”= Edit option required, “S”= Save partial results without edit option. If you prefer interactive homogenization, you must choose “E”. If you do not respond with “E”, the program will run in automatic mode without considering “ANAMEtarget-edit.txt”.

4.3. Initiation II. Continuation of a suspended homogenization

You may suspend the homogenization of multi-network datasets. When you want to continue the homogenization, click again on ACMANT5run.BAT, then the program will ask for the climatic element. To that question, please respond with “CC” instead of the code of the climatic element. From this response ACMANTv5 understands that you want to continue a suspended homogenization, and will ask you how many time series have already been homogenized. You can check that from directory Output and introduce the correct response. Then ACMANTv5 will run without putting further initiation questions.

4.4. Networks

In multi-network homogenization, ACMANTv5 creates the subdirectory “Subnetws” within directory Works. Subnetws is divided more to sub-sub directories, as for each candidate series of the dataset a network is constructed, and each of these networks has

an own directory (network directory, hereafter). Network directory names show the serial number of the candidate series of the network. Inside a network directory, the candidate series and its neighbour series are present in the same form as in the input dataset, except for the filenames. In datasets of dense and highly correlated time series, the typical network size is 30-40 time series, although maybe somewhat larger when the observation period widely varies between stations, or for many data gaps. Each network directory includes a file "networksize.txt" presenting the number of time series in network, it is needed for the ACMANT procedure. Any network directory may include a file named "subrth.txt" presenting a specific correlation threshold for the homogenization of the candidate series. This file can be provided by the user, or when ACMANTv5 does not find this file, it creates it as an empty file. The other files of network directories are for the time series belonging to them. The rules of automatic network construction with ACMANT were presented by Domonkos (2020).

In network directories, the filenames of time series hold serial numbers specific for the network structure and independent from the serial numbers shown in the input dataset. Serial number 1 always belongs to the candidate series of the network. The other serial numbers show the order of selection of neighbour series during the network construction. The order between series is neutral for the homogenization procedure (except for serial number 1), but the program must find filenames with every serial number from 1 up to the number indicated in networksize.txt.

In multi-network homogenization with ACMANTv5, the homogenizations of individual candidate series are independent from each-other, in fact a new homogenization procedure starts for the homogenization of each candidate series. The number of homogenized candidate series can be regulated by ANAMEtarget-homg.txt (Sect. 3.7).

4.5. Partial results

Partial results are tables showing some details of the homogenization for all series of the actually examined network. When partial results are requested at the initiation of the homogenization procedure, they are provided for each homogenized candidate series and all its neighbour series. First, ACMANTv5 puts some partial results into the directory Partial results, and moves them to directory Output at the end of the homogenization of a given candidate series. Four kinds of partial results are generated after the first homogenization cycle of the ACMANT procedure, and a fifth kind result table just before the end of the homogenization of a given candidate series. They are:

i) List of detected breaks by the first homogenization cycle
ii) List of detected outliers in the first homogenization cycle
iii) Table of pairwise comparison scores
iv) Anomalies of annual values relative to the long-term mean values, calculated from outlier filtered but non-homogenized values
v) Anomalies of annual values relative to the long-term mean values, calculated from the final homogenized values

In bivariate homogenization i), iv) and v) have two subtypes, i.e. result tables are shown for both variables. Bivariate homogenization is performed for a) S-model inhomogeneities, the two variables are the annual mean and summer – winter difference (Domonkos, 2020), b) precipitation homogenization with alternating rainy and snowy seasons, the two variables are the precipitation total of rainy season and precipitation total of snowy season (Domonkos, 2020).

Note that in precipitation homogenization the logarithmic transformed values are used for all calculations of the partial results.

i) List of detected breaks by the first homogenization cycle. --- It is the only editable file type of the partial results, the other files are only for providing information to the user. The filename has the form of "PBRtypeANAMEJJJJ.res".

"PBR" – constant
"type" – "Year" for annual mean, "S-Wd" for summer - winter difference, "Rain" ("Snow") for precipitation total of rainy (snowy) season
"ANAME" – dataset name
"JJJJ" – serial number of the candidate series
".res" – constant

For a network of *N** series, this file contains *N** blocks. For each block:

- line 1: within network serial number of time series, homogenized period, station identifier
- line 2: number of detected breaks (*k*)
- lines after 2 until 2+*k*: serial number of break, year of the break (the end of the shown year must be considered), break size.

ii) List of detected outlier values or outlier periods in the first homogenization cycle. --- The filename has the form "PoutlieANAMEJJJ.res".

"Poutlie" – constant
"ANAME" – dataset name
"JJJJ" – serial number of the candidate series
".res" – constant

In ACMANTv5 the possible occurrences of monthly outlier values or outlier periods for 1-28 month sections of time series are examined, anything is the time resolution of the input data. In precipitation homogenization only 1-month outlier values are examined, and only when user opted for outlier filtering yes. In the homogenization of other climatic elements, when user opts for outlier filtering no, possible occurrences of 1-4 month outlier periods are not examined, but those of 5-28 month periods yes.

For a network of *N** series, this file contains *N** blocks. For each block:

- line 1: within network serial number of time series, number of detected outliers (*k'*) station identifier
- lines after 1 until 1+*k'*: serial number of outlier period, calendar year and month for the last month of the outlier period, length of outlier period in months.
- line *k'*+2: blank

iii) Table of pairwise comparison scores. --- The filename has the form "PpairwsANAMEJJJ.res".

"Ppairws" – constant
"ANAME" – dataset name
"JJJJ" – serial number of the candidate series
".res" – constant

This is the summary table of pairwise detection results without any combination of the pieces of the results and without considering metadata. The rows are for the years of the homogenized period of the candidate series, while the columns are for the time series of the examined network. The table shows that how many breaks were detected for each time series and each year. In the ACMANTv5 pairwise comparisons such sums can be fractions (see Domonkos, 2021).

iv) Anomalies of annual values relative to the long-term mean values, calculated from outlier filtered but non-homogenized values. --- The filename has the form "rsrtypeANAMEJJJ.csv"

"rsr" – constant
"type" – "Year" for annual mean, "S-Wd" for summer - winter difference, "Rain" ("Snow") for precipitation total of rainy (snowy) season
"ANAME" – dataset name
"JJJJ" – serial number of the candidate series
".csv" – constant

This table can be opened in Excel, and its purpose is to give users an easy access to make graphical comparisons between time series. In the data table the rows are for the years of the homogenized period of the candidate series, while the columns are for the time series of the examined network. The data table can be converted into figures within seconds, it is easy even for whom Excel is not a routine access. Figure 1 shows an example. For simplicity, a small network (6 time series) and a short period (1975-2000) are shown. For transforming a data table to figure, first select the data area (A – G; 1 – 27 in the upper panel of Fig. 1) with mouse, then choose Insert / Graphics / Lines from the menu. The result is shown in the bottom panel of Fig. 1. For larger networks and longer periods you may need the selection of one or more small sets of time series / subperiods. These all are very easy in Excel with moving rows and columns to the required order.

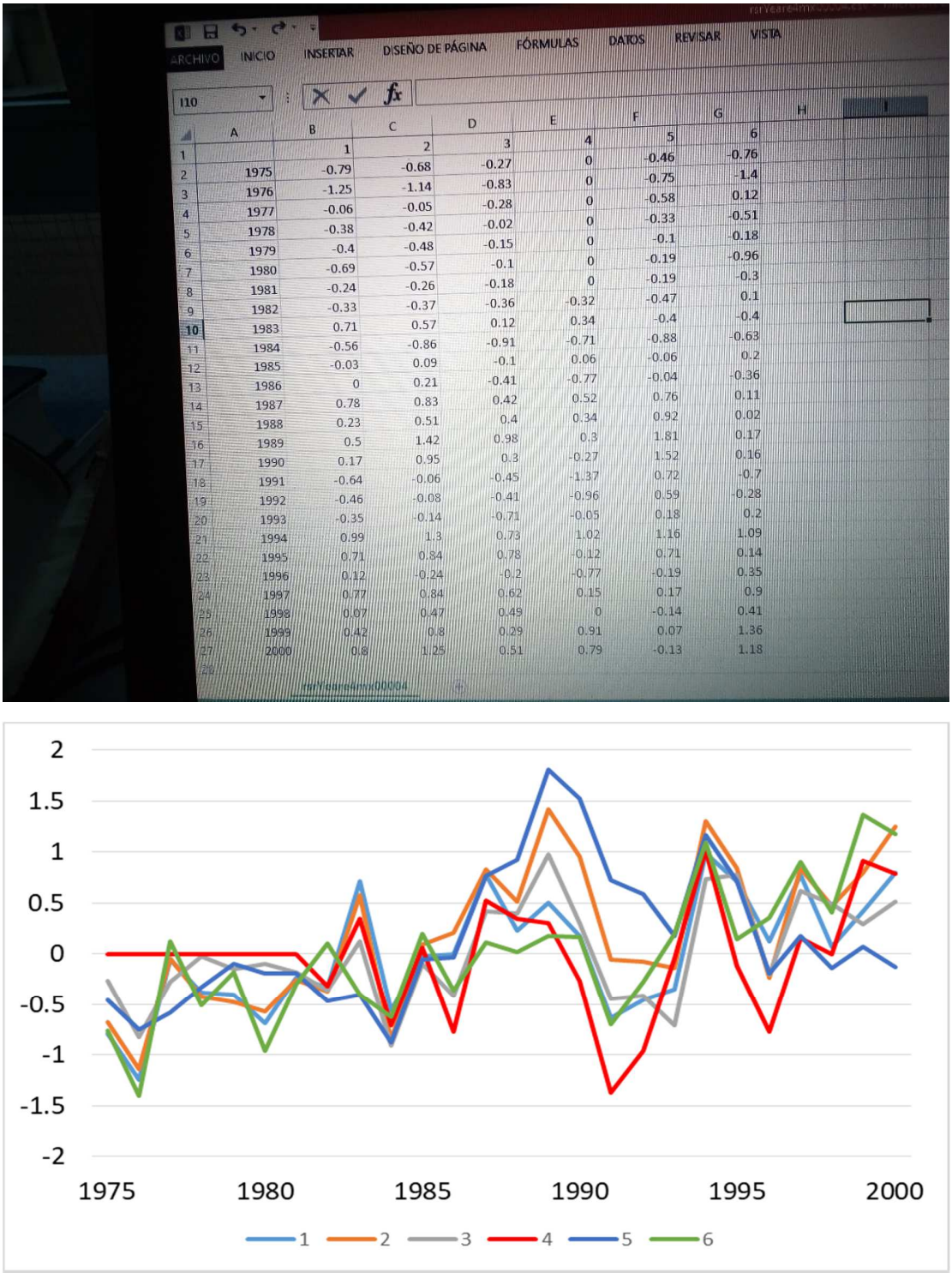

Figure 1. Upper panel: A *.csv data table for the 1975-2000 period of 6 time series. Bottom panel: the figure of the data table by selecting Insert/Graphics/Lines

Note that when a section of a neighbour series cannot be homogenized, series of zero values will appear in the data table and the transformed figure (series 4, 1975-1981 in Fig. 1).

v) Anomalies of annual values relative to the long-term mean values, calculated from the final homogenized values. --- This file is similar to that of item iv) both in its name and content, with two differences a) In the filename the first three letters are “hsr” (constant); b) It shows homogenized time series.

In multi-network homogenization, when a candidate series cannot be homogenized for reasons discussed in Sect. 3.3, partial results are generally not generated, and the final results clarify the fact that the candidate series was not homogenized. In the homogenization of 1-network datasets, requested partial results are not presented only when none of the series can be homogenized. In precipitation homogenization, outlier list is not provided if the option of no outlier filtering was selected.

4.6. Possibilities of user interaction

During the homogenization procedure, users may edit the networks, may change the correlation thresholds for selected networks and may change the break detection results of the first homogenization cycle.

i) Adjusting network contents

- Users may add or remove time series in networks, except for the candidate series must remain the same. The list of the serial numbers of the time series must remain continuous. For instance, if a network of monthly time series contains 20 time series (ANAME01.txt, ANAME02.txt,…ANAME20.txt), and you remove the ones with serial numbers 12 and 15, then you must rename some files. The simplest solution is if ANAME19.txt is renamed to ANAME12.txt and ANAME20.txt is renamed to ANAME15.txt. Remember that changing the number of time series in a network, you must correct the content of file “networksize.txt”.

- Users may introduce specific correlation thresholds for networks with creating “subrth.txt” files. The correlation threshold must be shown in the first line of the file, then it must be put to the selected network.

When you would like to reduce the effect of weakly correlating neighbour series on the homogenization of a candidate series, you may opt to remove some useless neighbour series or elevate the correlation threshold. Note, however, that these two do not have the same effect: when the correlation threshold is elevated, it influences the homogenization of the neighbour series in a not fully presumable way, and this might affect the homogenization results of the candidate series. When some neighbour series are removed, it is clear that they will be absent in the homogenization of all the other series of the network. Remember that in ACMANT, all time series of a network are homogenized together to achieve high quality results for the candidate series.

When after the automatic networking a candidate series cannot be homogenized for its low correlations with the closest station series, you may add more neighbour series to its network manually. However, relatively weakly correlating neighbour series will be effective in the homogenization of the candidate series only if the correlation threshold of the network is lowered accordingly.

You must keep in mind that any increase of network sizes lengthens the computation time. The permitted maximum network size is 99 series, but the computation time for a network doubles with adding 5-7 time series to it, therefore it is recommended to enlarge automatically created networks only when the inclusion of additional neighbour series is clearly reasoned.

ii) Adjusting break lists

With editing PBRtypeANAMEJJJJ.res, users may alter the number of detected breaks or the year of the break position(s) for any time series. Changing the shown break size does not have effect on the homogenization procedure. When you change the number of detected number of break for a time series, remember correcting the content of line 2 (number of detected breaks) of the relevant block of the file.

The changes in PBRtypeANAMEJJJJ.res impact the homogenization results of the first homogenization cycle, but their impact on the final homogenization results is uncertain. However, it is normal that a statistical homogenization method gives only limited possibilities to subjective interventions.

4.7. Interactive homogenization with ACMANTv5

- After the automatic network construction ACMANTv5 asks if the user accepts the networks. The user can check the networks in directory Works/Subnetws. The networks can be edited, and when the user is satisfied with the networks, he/she types “C” (continue).

- After the first homogenization cycle of homogenizing each network, ACMANTv5 indicates that a break list is editable. The user can check the content of the partial results in directory Partial results, and may edit the file PBRtypeANAMEJJJJ.res. When the user is ready, he/she types “C”.

- In bivariate homogenization (Sect. 4.5), ACMANTv5 gives two separate indications for the edition options of the two break files. --- When de user is ready with the check and possible edition of the second break list, he/she types “C”.

- When the homogenization of a network has been finished, ACMANTv5 asks if the user accepts the homogenization results of the candidate series. For the decision of accepting or not accepting the homogenization results, user can examine all the partial results and the requested other output items, everything in the directory Output at this phase. If the user accepts the results, he/she types “A” (accept), and the homogenization continues

with the subsequent network. If the user is unsatisfied with the results, he/she may edit the network content, may prepare to different break list editions after the first homogenization cycle, and when he/she is ready with these, types “R” (repeat). Then ACMANTv5 will repeat the homogenization of the given network. This repetition option makes easy to treat possible visible problems of the homogenization results immediately, even for large datasets.

5. Results of the ACMANTv5_S10 homogenization procedure

The output may include various versions of homogenized series, a summery of the detected breaks and outliers, files with monthly reliability indicators related to the data treatment during the homogenization procedure, the list of neighbour series and their spatial correlations with the candidate series, one or more control messages, and some partial results. All these depend on the user’s choices provided at the beginning of the homogenization procedure (Sect. 4.2). One easy option is to choose the default output package, its content is detailed in Sect. 5.1. However, users may choose other output items than which belong to the default output package. The optional output items are described in Sect. 5.2. Note that the partial results were described in Sect. 4.5.

5.1. Default output package

i) Homogenized monthly series with gap filling within the homogenized period (for possibly existing missing data of the raw series). Their names have the form of “ANAMEJJJJt.txt”, where

“ANAME” – name of the dataset
“JJJJ” – serial number of the time series
“t.txt’ – constant

This file contains the homogenized time series, but note that the requested homogenized data tables are all generated, even when no section of a time series has been homogenized. So that the number of files of this kind is the same as the number of time series in the input dataset. The data format is exactly the same as the input data format.

ii) Summary of the homogenization procedure with the list of the detected breaks and outliers and some other characteristics. The name of this file has the form of “ANAME_breaks.txt”.

“ANAME” – name of the dataset
“_breaks.txt” – constant

The organisation of the data in this file is as follows: The list comprises *N* sub-lists, where *N* stands for the number of stations. A sub-list generally contains the following 3 parts: a) headline, b) list of breaks, c) list of outliers. The headline contains (from the left to the right) the serial number of the time series, the starting and ending years of the treated period (Sect. 3.3), the starting and ending years of the homogenized period, the number of detected breaks, the number of detected outliers (only in outlier filtering "yes" mode), and the station identifier. In part (b), the properties of the detected breaks are presented. They are the serial number of break, the year and month of break position and the break size for the annual mean of the tested variable. In bivariate homogenization (Sect. 4.5) two break sizes are shown, one for each variable. In part (c), the properties of the detected outlier periods are presented, they are the serial number of the outlier period, the year and month of the last month of the outlier period, the duration of outlier period (in months) and the magnitude of deviation. Some notes:
- If no break is detected in the time series, there is no b) part in the sub-list.
- If no outlier is detected in the time series, or outlier filtering "No" mode is applied, there is no c) part in the sub-list.
- The most frequent length of outlier periods is 1 month (i.e. one single outlier value), but lengths may vary between 1 and 4 months. In the partial results the maximum length of outlier periods is 28 months (Sect. 4.5). However, in the final homogenization cycle outlier periods of longer than 4 months are treated as a pair of breaks, and their properties are shown in the break list (part b).
- If homogenization has not been performed, character "0" is written into the places of the starting and ending years of the homogenized period, and "–1" into the place of the number of breaks.
- If a time series does not have treated period, "0" are shown in the places of the starting and ending years of the treated period.

iii) List of neighbour series for each candidate series, and the spatial correlations between candidate series and their neighbour series. The name of the file has the form of "ANAME_rlist.txt".

"ANAME" – name of the dataset
"_rlist.txt" – constant

In multi-network homogenization this file shows the network structures and the correlations between the candidate series and neighbour series. In 1-network homogenization (for small datasets), the candidate series – neighbour series relations are also shown for every time series, since the roles of being candidate series or neighbour series change during the homogenization procedure. Note that when a candidate series does not have homogenized section, no neighbour series or spatial correlation is shown for that. Furthermore, when a neighbour series cannot be homogenized, that series is not shown in any neighbour series list, even if it was selected for some candidate series during the automatic network construction. The form of data is as follows: for each candidate series with non-zero homogenized period has a headline, followed by the list of neighbour series ordered according to decreasing spatial correlations. The headline includes (from the left to the right) the serial number of the candidate series, the starting and ending years of its homogenized period, and its station identifier. Then, in each line

after the headline, the serial number of correlation value, the correlation value itself and the station identifier of the neighbour series are shown.

iv) Control message. In running ACMANT, “Control message.txt” file is generated. If no problem has been detected during the run of the program, “….homogenization has been completed” will be the message. However, other messages are possible. In multi-network homogenization, at least one message is provided for each network homogenization.

5.2. Optional output items

Optional output items (excluding here the partial results, which are presented in Sect. 4.5) are data tables of homogenized time series or tables of reliability indicators. All of them have the same format as “ANAMEJJJJt.txt”. The letter before “.txt” shows the kind of the output item, the other characters of the filenames are the same for a given time series.

v) ANAMEJJJJs.txt – homogenized monthly time series without gap filling for missing data.

vi) ANAMEJJJJh.txt – homogenized monthly time series completed with interpolated data to the target period when data gaps occur in the input data.

vii) ANAMEJJJJj.txt – monthly data table of reliability indicators. The data format is the same as for climate data tables, except for whole numbers (0…9) of reliability indications stand on the places of climate data of climate data tables. The meanings of reliability indicators are presented in Sect. 5.3.

5.3. Meaning of reliability indicators

Although the file ANAME_breaks.txt shows the homogenized period for each time series, reliability indicators might show more about the reliability of the homogenized data for two reasons:

- When less than three time series have data for a specific year, that year is excluded from the homogenization for all the stations. Such years might occur within the homogenized period.

- ACMANT automatically performs gap filling with spatial interpolation, but when the number of neighbour series or the spatial correlations are low, the confidence of interpolated values is limited.

Generally the lower values mean higher reliability, except the 0 code.

0 – sporadic observed value, out of the treated period of the time series, or observed value in a section of the candidate series where the number of neighbour series with observed values is 0 or 1.

1 – observed value within the homogenized period
2 – observed value out of the homogenized period
3 – Interpolated value from at least 4 sufficiently correlating neighbour series.
4 – Interpolated value from 3 highly correlating neighbour series, or from more than 3 but only moderately correlating neighbour series.
5 – Interpolated value from at least 3 fairly correlating neighbour series.
6 – Interpolated value from 2 fairly correlating neighbour series, or from more but poorly correlating neighbour series.
7 – Interpolated value by the use of 1 only neighbour series.
8 – Missing value is substituted with the climatic normal value due to the complete lack of observed values in neighbour series. Frequent occurrence of this code number may be present when the target period is longer than the periods of observed data, and full completion of time series is performed. When no data is available for spatial interpolation, the climatic normal values are included (repeatedly) in the artificially lengthened series. Note, however, that this code sometimes occur within the homogenized period due to synchronous data gaps (e.g. for political events).
9 – missing data for which interpolation has not been performed. This code may occur out of the homogenized period, when data tables of time series completed to the target period are not requested.

Note: The coding between 3 and 7 is based on the number of neighbour series, the spatial correlations between series, the frequency of comparable data pairs around the date of the interpolation, and if the data are within the homogenized period yes or not. Due to this complexity, not all details of the coding rules are presented.

5.4. Documentation of user interventions

In the present ACMANT version not all the user interventions remain documented.

- The applied correlation thresholds and network editions are not documented, but the truly used network structures (neighbour series and their spatial correlations with the candidate series) are documented in ANAME_rlist.txt, except when the user's output requests exclude this file.

- Editions of break lists are documented in PBRtypeJJJJ.res: Once its edit option has been closed, the data format for each of its blocks is as follows:
    - line 1: within network serial number of time series, homogenized period, station identifier
    - line 2: number of ACMANTv5 detected breaks (k)
    - lines after 2 up to 2+k: serial number of break, year of the break, break size.
    - line 3+k: number of breaks after user editions (k*)
    - lines after 3+k up to 3+k+k*: serial number and year of the break after user editions

- When repeated network homogenizations occur, the following details are documented by ACMANTv5 for each homogenization experiment (except when the user's requests exclude the relevant output items):
    - Network structure: in file ANAME_rlist.txt
    - Final lists of treated periods, homogenized periods, detected breaks and outliers: in file ANAME_breaks.txt.
    - Control message

6. Possible technical problems with using ACMANTv5_S10

The program might fail to accomplish the homogenization for various reasons. The use of a managing program (i.e. the program ACMANT5run manages the running of other "inner" programs of the software) has an inconvenience: when an inner program stops during the homogenization of a network, the running of the managing program still continues, and the procedure turns on the homogenization of the next network. As a consequence, users might fail to realise in time that a problem occurred during the homogenization of a dataset.

The following symptoms indicate that the program failed during the homogenization of a candidate series: (i) In the list of control messages ("Control message.txt") a candidate series is missing; (ii) Some of the output items or all the output items of a given candidate series are missing; (iii) The directory "Output" includes error message from the program (the homogenization may stop with or without giving error message). Note that the homogenization of some candidate series might be denied for the insufficient spatial-temporal coherence of the data, but this is not a software error, and in this case control message and some other output items are still generated.

When a failure of homogenization is realised, first the input data should be checked, as errors in the input data preparation may cause stop of the homogenization. In such cases, likely an error message from the program will be present in directory Works. A relatively frequent data preparation error is that the same time series occurs repeatedly, under different station names. In this case, the homogenization will stop with the error message of "Indefinite equation system".

The programs of ACMANTv5 have been tested, but in spite of this, maybe that some software errors will appear only during its practical use. Thus if you do not find irregularity in your input dataset, maybe that a software error causes the failure of homogenization, and then please contact me and report the problem.