Homogenisering av månadsmedeltemperatur 1860–2021

Bilaga 1: Bartguide

Magnus Joelsson

12 oktober 2022

Bart guide

HOMER automated

Basics

Bart is an automated version of HOMER v2.6. The script is a modified version of HOMER-script, written i R. There is a dependency on cghseg.0.0.1 which is not compatible with R version over 3.5 The main script is called bart.x.x.x.R. The current version of Bart is 0.17.12.

Bart runs with a set of input parameters which controls:

- The selection of references
- The order of functions to be run
- The weights for the point system

Table 1: Input parameters

Name in R	**Possible value** (Default value underlined)	**Description**
From the original HOMER-script		
unit.str	"c", "..."	Unit string for figures
comp.option	"a", "r","log"	
dev.str	"pdf", "ps", "svg", "png"	Option for format of graphics files
inter.option	"c", "g", "m", “”	Selects the procedure for selection the references: “c” = correlation, “g” = geographic, “m” = mix-neighbours (bart-specific), “” = all stations
inter.number	0.95, ...	The threshold for the references. If the correlation method is used, the inter.number is the threshold of the correlation, if geographical method is used the inter.number is the distance threshold. If the “mix” method is used the number is used as a correlation measure to divide the time series in groups, see below.
n.min	8, ...	The minimum number of references
season.option	"", "a", ("m")	On what level the detections are taken place seasonal, annual or monthly. Only seasonal (and then only summer, winter) is implemented properly so far
season.corr	"m", ""	On what level the correction is monthly or annual, for comp.option="a", season.corr is forced to "m". Nothing else is implemented so far
trend.str	"y"	Linear trend, for visu function
smooth.str	"y"	Smoothing option, for visu function

col.str	"y"	Polygon fill, for visu function
col.fill	"rb"	Colour option for graphics
col1.str	"red"	Colour option for graphics
col2.str	"blue"	Colour option for graphics
Bart specific		
n.max	0, ...	Maximum number of neighbours, if the maximum number of neighbours is smaller than the minimum number of neighbours, the number of neighbours is unlimited
stand.proc.iter	2, ...	The number of interations for the standard procedure of functions: "d, j, c, a, c". Before the last "c", an "m" is added. The default procedure is thus "d, j, c, a, c, d, j, c, a, m, c" All "d" and "j" functions after the first "c" will be run of corrected data. This procedure corresponds to stand.proc.iter = 2
num.cores	6, ...	Number of cores to for the parallelisation
log.file	"log", ...	Name of the log-file where the process is recorded and input parameters stored
clean.sheet	FALSE, TRUE	If TRUE, the folder is cleared from "ho", "tmp", "fig", "meta"-folders and "000099detected.txt"
r.command	R	The R-command used by the crash safe version mode
crash.safe	FALSE, TRUE	If TRUE Bart runs the joint.detection function in an extern R session, such that if it crashes it is equivalent to no breaks found in the time series by the joint detection
old.auto.homer	FALSE, TRUE	If TRUE all weights are set to mimic the old automatic HOMER mode
ext.brk.info	TRUE, FALSE	If TRUE Bart saves information of all the breaks in the meta-folder as "nnnnnnxx_breakinfo.txt"
brk.conv	FALSE, TRUE	If TRUE, the standard procedure is repeated until no more breaks are reported by any function
validate.meta	c(6), c(1, 2, 3, 4, 5, 6)	A list of meta data type breaks that locks the position of breaks: If a break is confirmed by Bart ±2 years from a meta data break of a type that is included in the list, then the position of the confirmed break is taken from that meta data break and is not subject to the change month function
j.skip	c()	An alternative to crash safe mode is to add known problematic series to the j.skip list. Index number of the station list is used, if j.skip =c('all') then j detection function is skipped altogether
plot.fig.flag	TRUE, FALSE	If TRUE, no figures are plotted, for example good for optimisation runs and other faster slimmer applications
gap.fill.neighbours	FALSE TRUE	An option that modifies the procedure to select references if the mix-neighbour-option (inter.option = “m”) is selected. Instead of making sure that there are at least four references for 95 % of the candidate station’s active period, the method makes sure that at least three references for 95 % of the entire period

		(period including times when the station has missing values)

Weights and similar

restrict.time	1, 0, ...	The restricted time controls time before and after a break no additional breaks are allowed. Only 1 and 0 is tested
w.acmant	3, ...	Weight of the acmant-style detection function points
w.joint	3, ...	Weight of the joint detection function
w.pair	3, ...	Weight of the pairwise detection function, annual
w.season	3, ...	Weight of pairwise
w.meta	3, ...	Weight of meta data breaks
w.amp	0, ...	Amplitude power for the amplitude power term
brk.norm	0, ...	If 0, brk.norm is set to the median of the breaks found by the pairwise detect function. brk.norm is the denominator for the break size if the the amplitude power is used
w.corr	0, ...	Correlation power for the corrrelation power term
amp.thres	0, ...	The threshold for how small break that can be confirmed
adj.pen	3, ...	The penalty a candidate break pays for adding break points for the one year before and after
brk.thres	12, ...	The threshold for when a break is confirmed
note.thres	9, ...	The threshold for when a break is noted
neighbours.normal	0, "median", "mean", "max", "min", "off", ...	A factor for the pairwise detection points if the points are divided by the number of references, if neighbours normal factor is 0 (or "off"), the points will not be divided by the number of references
test.order.index	0, ...	Numbers the stages of the procedure, test.order.index is increased by 1 for each task

corr.cycle	0, ...	Numbers the correction cycle, corr.cycle is increased by 1 each corr-task

Note that all the numeric inputs can be set to be vectors such that their values in updated for every step

There is two ways to run Bart:

- For the command line Bart can be run be giving the input as argumnets. For example
  R CMD BATCH "--args workdir='/path/to/the/working/directory' clean.sheet=T neighbours.normal='min'" bart.0.17.12.R theoutfile.txt &
- If R is run inside a R console or in the command line without arguments the input parameters are fetched from a file called "bartflags.R". This is simply a R-file that is run inside the Bart script.

Files

Input

The basic input is similar to HOMER:

- A stations file called "*nnnnnn*stations.txt" where "*nnnnnn*" is the number of the network. This file is kept in the working directory. Each stations gets a row in the stationsfile. Each row has nine columns seperated by tab:
  filnamn latitude, degree latitude, minute latitude, second longitude,degree latitude, minute latitude, second altitude in m name
  Example:

qctmm00180760d.txt	67	8	24	20	38	60	361.99	Gallivare
qctmm00180960d.txt	67	50	60	20	14	24	502.5	Kiruna

- The data input files are kept in a "qc" folder in the working directory. They are named as qctmm00180760d.txt. From the HOMER guide:

  HOMER file names (see Example 2.1) follow strict conventions to facilitate data input into the software. Data files must have 18 characters, responding to the format *xxvvrsssssssssc*.txt, where:

  - *xx* is data status, one of ra (raw data), qc (quality controlled data) or ho (homogenized data).
  - *vv* is the climate variable, being the most commonly used:
    - tn: monthly mean of daily minimum temperature
    - tx: monthly mean of daily maximum temperature
    - tm: monthly mean of daily mean temperature
    - rr: monthly accumulated precipitation
    - pp: pressure

        - sd: sunshine duration
    - *r*: resolution or averaging period of the data: for HOMER purposes this is always m (monthly). Other valid resolution descriptors in the HOME-format are y (annual), d (daily), the later useful for the daily data homogenization software SPLIDHOM.
    - *ssssssss*: The station number should be specified with 8 digits. If the station has an internationally recognized code such an WMO station number, its utilization is preferable, as this avoids duplicates. In absence of internationally recognized codes, the user can create own internal numbers. If the number is less than 8 digits, it should be completed with leading zeros.
    - *c*: this is a contents indicator. For HOMER purposes, d – for data – should be specified. Another possible options – useful in other applications –is f (quality flags).
    - ".txt": the last four characters are literally ".txt", introduced to identify the plain-text context of the file, for easy processing.
    - The format of the file content is each row represents a year , each row has 13 column seperated by tabs. The year is stated in the first column, the other twelwe column representes the average data for each month. Example:

      1982 -14.231 -6.866 -4.905 -0.886 4.016 7.124 13.303 10.823 4.734 -0.018 -5.954 -9.169
      1983 -10.982 -9.808 -7.352 0.052 6.59 10.069 13.389 9.736 6.238 -0.694 -9.769 -11.689
      1984 -14.882 -7.289 -8.674 -0.027 9.05 11.17 12.116 10.339 4.551 1.087 -999.9 -999.9

- A Bart-specific input is meta data list file. It is kept in the working directory and is named "*nnnnnn*metabrk.txt" where "*nnnnnn*" is the number of the network. Every row is one break. Every row has five column separated by tabs as:

  Station number year month meta break type note
  For example:
  00180760 1996 1 6 Koppling-00180760
  00180760 1959 5 1 Meta-00180750

The meta break type has six + one possible values

Table 2: Meta data break number

Meta data break type number	**Description (suggestions)**
1	Station move
2	Change of observer
3	Change of instrument
4	Change
5	Other
6	Coupling, in the default mode, this meta break type locks a confirmed break in position

7	Real breaks (benchmark data sets), this is only for plotting, not for break detection

Temporary/Help files

Bart-specific

Two important help files that are stored in the working directory

- corrmat.sav: Here the correlations between the stations (unhomogenised) are stored as a matrix in an R object "corr.mat". The index is the same as the linenumber in the stations file
- annlist.sav: Here the annual average (unhomogenised) data is stored in a list of data frames (years and annual average data for each station) in an R object ann.list. The names of the lists are the names of station data files.

There is additional temporary Bart-specific auxillary file stored in the tmp-folder, which means that it is removed with the "clean.sheet" flag turned on:

- listneighbour.sav: Here the vectors of the references are stored in a list as an R. Loading of the file is not implemented yet.
- brknorm.sav: Here the brk.norm R object is saved to be used in the amplitude power term of the break point counting function
- option-file *xxvvrnnnnnn*options.sav has the input parameters saved
- *nnnnnn*noted.txt (where *nnnnnn* is the network code): the noted breaks are stored here. Each line represent a noted break, each row contains three column separated by tabs. The columns are station index, year, and iteration cycle index. Example:
  00180760 1948 1

HOMER

There is a few different types of files in the tmp-folder:

- Temporary detection files: Each time the correction function is run, the detection file is stored. The files are numbered. In Bart also the final version stored, since in Bart discarded breaks are kept in temporary detected files (marked with “-“), but the in main detected file (in the main folder) the discarded breaks are removed.
- .sav-files on the format “detect_*xxvvssssssss*.sav” (annual) and “detect_*xxvvssssssss_MMM*.sav” (*MMM* = DJF [winter], JJA [summer]) where information on breaks from the pairwise detected functions are stored
- temporary data files are also stored here on the same format as in the “qc” and “ho” folder, they are f.x. used in the acmant function
- .sav-files on the format “jdetect_*xxvvssssssss*.sav” (annual) where information on breaks from the joint detected functions are stored

Output

HOMER

- The most important output files are the homogenised time series stored in the ho-folder where time time series are stored in the same format as the qc.
- nnnnnndetected.txt (where nnnnnn is the network code): Here the list of confirmed breaks is stored. Each row represents a break. Each row has six columns separated by tabs. The columns presents the station index, break-type, year, month, meta data verification code, and station name. Example:

00180760	BREAK	1890	12	n	Gallivare
00180760	BREAK	1956	12	n	Gallivare
00180760	BREAK	1964	1	v	Gallivare
00180760	BREAK	1970	12	n	Gallivare
00180760	BREAK	1975	11	n	Gallivare
00180760	BREAK	1996	12	n	Gallivare

  The break-type does only have one option implemented, namely "BREAK". The meta data verification code originally has two options "n" for no and "v" for verified. Verified breaks are not moved by the change month function. In Bart there is two additional code: "m" which is a softer meta data verification such that the change month function will move the break. For the the temporary detected-files, the meta data verification code can also be "-" which means that the break is removed in the correction function.
- meta-files are save in the “meta”-folder on the format “meta*vvssssssss*d.txt”. Here the reference network for each station is stored. A novelty in Bart in that each time the network is recalculated (twice in the default run), the networks are stored. The periods of correction and the amplitude of correction month by month is also stored here.

Bart-specific

The break info files contains information about break points house keeping. Each section represents a break. The first integer represents the test order index. The subsequent sentence informs whether the break count was above the note threshold or the detection threshold, the station and year in question and the total score. The next few rows reports the number of breaks from the different detection functions for the year in question and whether there is a meta data break at the year in question. The respective information for the previous and following year with the additional if the next previous or next following year has a higher total count than the year i question. If for example the year+2 has higher point counts than the candidate year, year+1 will be counted to year+2 and not to the candidate year. Finally the five-year points summarise this. The break info files are stored in the meta-folder. They are named as *ssssssss*_breakinfo.txt where *ssssssss* is the station index.

If the gap-fill function is run, the gapfilled times series which can be either qc or ho (perhaps also ra) is stored in the gapfilled folder. The data format is the same as the qc-files.

Functions

Bart-specific

manage.arguments

This section goes through what the manage.arguents-function works. The manage.arguments function first sets the default settings of all the input parameters. The function then replaces the parameter values of those parameters that are given in the argument. Any input parameter that is not given in arguments are given the default value, see Table 2. If no arguments are given, Bart runs the “bart_flags”-file. Any input parameter that in not listed in the “bart_flags”-file will take the default value.

There is a special flag for the old homer auto mode: “old.auto.homer ← TRUE”. It sets the following parameters:

- restrict.time ← 0
- w.acmant ← 1
- w.joint ← 1
- w.pair ← 0
- w.season ← 0
- w.meta ← 0
- w.amp ← 0
- w.corr ← 0
- brk.norm ← 0
- amp.thres ← 0
- adj.pen ← 999
- brk.thres ← 1
- note.thres ← 999
- neighbours.normal ← 0
- season.option ← “a”
- validate.meta ← c() #meta break types that will not be moved by assess month

Some parameters than are given as characters are translated to numbers.

The working directory is set to the “workdir” parameter. If the workdir is empty, Bart will run in the current directory.

If the neighbours.normal is given any of the following inputs: “median”, “mean”, “max”, “min”, the neighbours.normal is assigned a negative number that serves as a code.

If the two test.order.index- and corr.cycle-parameters are not given, there are given the value 0. Here there could be a better solution to find test.order.index and corr.cycle automatically if a run is broken and should be picked up.

All the numeric parameters are moved into objects that can be vectors, so that the object parameters can be used as numbers.

If a raw.str parameter is not given, Bart will try to select either “qc*vv*” or “ra*vv*”, depending on which exists. If par.str is not given, it is taken from raw.str

If no net.str is given but there is only one *xxxxxx*stations.txt-file, then the net.str will be taken from the stations file

If “clean.sheet” is true, the log file, the detected file, and all content in the “ho”, “tmp”, “fig”, and “meta” folders are removed.

If the convergence option in on (“brk.nov = TRUE”) then the counter “conv.i”, the flag that signals that convergence is reached (“conv.reached”), and action and head orders (decides whether the functions work on corrected or raw data) are set. Otherwise the action and head orders are set according to the given number of standard procedure iterations (stand.proc.iter) or what is given as inputs.

If the num.cores (number of cores) are < 0 then the the value of num.cores are subtracted from the cores available.

Finally, all the inputs are written in the log-file.

make.action.order, make.head.order

These two functions constructs the action order (the order of which the functions should be applied) from the number of iterations given and the head order (decides whether the functions work on corrected or raw data) from the action order.

remove.lonely.stations

This function removes a station and resets all the objects accordingly. If a station does not have any accepted references, it causes trouble and needs to removed.

count.brks

This function counts the breaks for a certain station. The inputs are the

- years of interest (all the years with breaks +/- 1)
- the position of breaks for all functions (incl. meta breaks)
- the amplitude of the pairwise and joint detection breaks
- the correlation of the references and the candidate in the pairwise detection function breaks
- the number of references
- the index of the station

The procedure here is a bit complicated

1. factors for the amplitude of the amplitude of the break and the correlation of the references that signal the break is calculated:

   f_{amp} = (abs(break amplitude) / N_{brk})$^{\text{power of amplitude factor}}$
   f_{corr} = correlation$^{\text{power of correlation factor}}$

   N_{brk} is the normal break amplitude, the default value is the median of the breaks found by the pairwise detection function.

2. vectors of breaks of positive, negative and without amplitude (ACMANT) are calculated separately such that positive and negative breaks do not mix. Pairwise and joint break points are added:

   pairwise break points = $f_{amp,\ pairwise}$ × $f_{corr,\ pairwise}$ × pairwise weight × amplitude flag × max(1, (normal number of neighbours / number of neighbours))

   Note that max(1, (normal number of neighbours / number of neighbours)) means that the factor (normal number of neighbours / number of neighbours) cannot be smaller than 1. This is to prevent candidate series with few references (stations in sparse parts of the network) to gain more weight than the series with normal number of references.

   joint break points = $f_{amp,\ joint}$ × $f_{corr\ joint}$ × joint weight × amplitude flag

   amplitude flag = 1 if break amplitude $\geq$ amplitude threshold, = 0 if break amplitude < amplitude threshold

   total break points = annual pairwise points + seasonal pairwise points + joint breakpoints + acmant

   summer, winter and annual breaks are calculated separately such that in a year the seasonal breaks and the annual breaks can have different signs

3. For each year (and for each seasonal and annual category), it is decided whether there are higher break points for negative or positive breaks. Each category get assign the higher points.

4. If the previous year has lower points than the current year and the current year has higher points than the year two years previous the points from the previous year is added, such that no year can be double counted. An adjacent year penalty is subtracted. Similar calculation is done for the following year. Note that the compared break points are with the same composition of breaks regarding amplitude regardless of what composition is the highest for the adjacent years.

For each of the years of reported meta data breaks, a meta data point is assigned the year with highest scores of the previous, current, and following years.

make.ann.list

This is a modified HOMER function which reads in annual data from a list of file names into the a list in workspace such that the program does not need to read in the same data several times. The output-format is a list of data frames with filenames as indecis. The data frames has a column of times and data.

make.corrmat

This function takes the list of files and calculate a matrix of the correlations.

mix.neighbours

The function is a function to select references. The flag “gap.fill.neighbours” is used here.

All the time series are divided in categories according to their correlation with the candidate series. Within the categories the time series are sorted according to their distance to the candidate station. The categories are as follows:

1. All the time series with higher correlation than the correlation threshold (e.g. 0.95)
2. All the time series with higher correlation than the correlation threshold to the power of 2 (e.g. 0.95 × 0.95 = 0.9025)
3. All the time series with higher correlation than the correlation threshold to the power of 3 (e.g. 0.95 × 0.95 × 0.95 = 0.857375)
4. The following categories follow the same format: Category n: Correlation threshold$^{(n-1)}$ ≥ correlation > Correlation threshold^{n}

The overlap between the candidate and the potential reference must be at least 15 years.

A ranked list of time series are constructed from the categories. The first step is to add time series to a output list of references from the top of the ranked list and down if the time series have data in years where there is less than four reference series with data until 95 % of all years have at least four references with data. If “gap.fill.neighbours” is true, at least three references with data is required for the entire time series where the candidate station either has data or miss data. The second step is to add time series to the output reference list from the top of the ranked list down until i) the minimum number of reference is reached and ii) the correlation is lower than the threshold or the number of references exceed the maximum number.

The idea is to make sure that as much of the time series have references data points. The geographical distance is used since the correlation can depend heavily of the overlap.

make.neighbours

The function makes a list of lists of references for every time series. The function is parallelised. The inter.option can be geographic (“g”), correlation (“c”), all stations (“”), or mixed (default). Both qc-data (unhomogenised) or ho-data (homogenised) can be used. The function operates on the global list.neighbour.list-object.
In other functions the list.neighbour is fetched from the list.neighbour.list instead of calculated as in the original HOMER.

g.part, gap.fill

The gap filling part of the correc-function is isolated in these two functions. The function is repeated until there is no empty data left.

Modified and divided original functions

read.series

The read.series function is modified such that it reads the data from the ann.list object instead from the text files for the annual data. For other data, the series are read as usual from the text files.

distance.serie, correlation.distance

The maximum number of references (n.max) is applied, otherwise the functions are as the original function.

c.d.part

This calculates the correlation between two series. This calculation is isolated in a separate function to easily parallelise it.

extern.joint.detection

This function saves the Y object which is the input to the joint.detection function to a Y.sav-file. The function than runs the j.R script. The j.R script loads the Y.sav-file, run the join detection function and save the out object to a out.sav-file. The out.sav is finally loaded by the extern.joint.detection-function. The reason for this procedure is to make the joint.detection function crash safe (a try/catch statement has not been sufficient). If the crash.safe flag is TRUE, the extern.joint.detection-function replaces the joint.detection-function.

j.part, a.part, j.detect, acmant.detect

The calculations of the j.detect and acmant.detect functions are isolated to enable parallelisation. The part-functions replaced the loops over all candidate time series. The plotting of the break signals can include meta data break and true breaks if the data set is a benchmark data set with known true breaks. In the j.part-function, the count.brks function is applied to get the break points each year. The years with points above the break threshold are added to the detected data frame. The years with points above the note threshold are added to the noted data frame. If the year is already on the noted list (from another iteration cycle), the year is added to the detected data frame. The detected and noted data frames are combined to enable parallelisation. The breaks get one of three meta data break code:

- “n”: The breaks is not validated, the month will be 12 and will be subject to the “change month”-function
- “v”: The break is validated by a meta data break with a type that is included in the “validate.meta”-list (by default only coupling) and will not be subject to “change month”-function. The month is taken from the meta data break.
- “m”: The break is validated by a meta data break with a type that is not included in the “validate.meta”-list (by default all types except coupling) and will be subject to “change month”-function. The month is taken from the meta data break.

- There is also a fourth code that occurs: “-“ means that the break is removed in the correction due to missing data. In HOMER the break is simply removed but is Bart, the break is kept in the temporary detected lists to enable convergence (otherwise a removed break could be added back after the removal and repeal the convergence).

From the j.part and a.part functions a combined detect/note frame is obtained. This is separated into a detected and a noted data frame.

m.part, change.month

The calculations of the change.month function is isolated in the m.part-function to enable parallelisation. The part-functions replaced the loops over all candidate time series.

d.part.1, d.part.2, detect

The calculations of the original detect functions are performed in two separate loops over the reference time series and the calculations are therefore divided into two part-functions. Note that as opposed to j.part, a.part, m.part, and c.part where these replaces loops over all candidate time series, in this case the loop over the candidate time series are still in the parent function “detect”. The part-functions replaced the loops over all the references of each candidate time series in the original detect-function.

In the detect function neighbours.normal (if there is no good neighbours.normal value) is calculated by taking the mean, median, minimum, or maximum (in effect n.min) of the number of references (called neighbours). What statistic means that is used is decided by a code that is stored in the neighbours.normal object as a negative integer.

The brk.norm object is calculated by taking the median of the list of break amplitudes, if there is no good brk.norm value.

A blacklist is constructed where stations that have no good references are stored. These “lonely stations” are removed in the remove.lonely.stations-function.

c.part, correc

The calculations of the correc function is isolated in the c.part-function to enable parallelisation. The part-functions replaced the loops over all candidate time series.

In the c.part there is a bug fix which replaces the n.period.candidate-object with a n.period.candidate.all-object which is common for all candidates.

A tricky thing in the c.part function is to automatically remove breaks in empty levels. An empty level has two breaks and the later break is removed (unless the level is the last in the series, than the last break before the empty level is removed). Rather than removing the break from the temporary lists, the break is marked with “-“ in the meta column.

Main script

There is do.correction- and do.acmant-flags to disable runs of correction and acmant if no change.

The action.str and the head.str (decide whether the function is on corrected or uncorrected data) is taken from the vectors action.order and head.order with the test.order.index. The different parameters are changed if they are given as vectors. You can thus f.x. decrease the break threshold from one step of the homogenisation to the next.

Note that the in built-in pairwise detection in the correction action of HOMER is removed.

The last temporary detect-file is saved as ... detect.final.