Basic methodological questions of homogenization and the MASH procedure

Tamás Szentimrey

Hungarian Meteorological Service
I. Homogenization of monthly data
   Relative models and methods
   Methodology for comparison of series
   Break point (changepoint) detection
   Methodology for correction of series

II. Homogenization of daily data
   Relation of daily and monthly homogenization

*Remark on absolute methods*: Impossible to separate the climate signal and the inhomogeneity.
OUR SOFTWARES

MASHv3.01:
Multiple Analysis of Series for Homogenization
(Szentimrey, T.)

MISHv1.01:
Meteorological Interpolation based on
Surface Homogenized Data Basis
(Szentimrey, T., Bihari, Z.)
I. HOMOGENIZATION OF MONTHLY DATA

Relative Additive Model (e.g. temperature)

Monthly series for a given month in a small region:

\[ X_j(t) = \mu(t) + E_j + IH_j(t) + \varepsilon_j(t) \quad (j = 1, 2, \ldots, N; t = 1, 2, \ldots, n) \]

\( \mu \): unknown climate change signal;  \( E \): spatial expected value;
\( IH \): inhomogeneity signal;  \( \varepsilon \): normal noise

**Type of \( \mu(t) \):** No assumption about the shape of this signal

**Type of inhomogeneity \( IH(t) \) in general:** 'Step-like function’
with unknown break points \( T \) and shifts \( IH(T - 1) - IH(T) \)

\( \varepsilon(t) = [\varepsilon_1(t), \ldots, \varepsilon_N(t)]^T \in N(\mathbf{0}, \mathbf{C}) \quad (t = 1, \ldots, n) \) are independent

\( \mathbf{C} \): spatial covariance matrix, very important!
Type of errors of homogenization

**Type one error:** detection of false inhomogeneity

**Type two error:** neglecting real inhomogeneity

The significance and the power of the procedures can be defined according to the probabilities of these errors.
METHODOLOGY FOR COMPARISON OF SERIES

Related to the questions: reference series creation, difference series constitution, multiple comparison of series etc.

All the examined series $X_j(t) \ (j = 1, \ldots, N)$: candidate and reference series alike. Reference series are not assumed to be homogeneous!

Aim: to filter out $\mu(t)$ and to increase signal to noise ratio (power)

Optimal difference series can be applied for Detection and Correction procedures (MASH).
Optimal difference series constitution (interpolation between series) (MASH)

To filter out unknown $\mu(t)$ and to increase signal to noise ratio (power):

$$Z_j(t) = X_j(t) - \sum_{i \neq j} \lambda_{ji} X_i(t) = IH_j(t) - \sum_{i \neq j} \lambda_{ji} IH_i(t) + \varepsilon_{Z_j}(t)$$

where $\sum_{i \neq j} \lambda_{ji} = 1$, and $\text{Var}(Z_j(t))$ are minimal ($j = 1,\ldots,N$)

$\sum_{i \neq j} \lambda_{ji} X_i(t)$: created optimal reference series for candidate series $X_j(t)$.

Vector of weighting factors $\lambda_{ji}$:

$$\lambda = C_{\text{ref}}^{-1} \left( c_{\text{c,ref}} + \frac{(1 - 1^T C_{\text{ref}}^{-1} c_{\text{c,ref}})}{1^T C_{\text{ref}}^{-1} 1} 1 \right)$$

$c_{\text{c,ref}}$: cand-ref covariance vector, $C_{\text{ref}}$: ref-ref covariance matrix
MULTIPLE BREAK POINTS (CHANGEPOINT) DETECTION

Examination (more) difference series to detect the break points and to separate for the candidate series.

Difference series: \( Z(t) = IH_Z(t) + \varepsilon_Z(t) \quad (t = 1,\ldots,n) \),

\( IH_Z(t) \): inhomogeneity with \( K \) break points, \( T_1 < T_2 < \ldots < T_K \)

\( \varepsilon_Z(t) \in N\left(E_Z, \sigma^2_Z\right) \ (t = 1,\ldots,n) \) are independent

Possibilities, principles for estimation of \( K; T_1 < T_2 < \ldots < T_K \):

a, **Bayesian approach** (model selection), penalized likelihood methods

b, Multiple break points detection based on **test of hypothesis** (MASH)

Notation of estimates: \( \hat{K}; \hat{T}_1 < \hat{T}_2 < \ldots < \hat{T}_{\hat{K}} \)
Bayesian Approach (model selection), penalized likelihood methods

Decision on estimates $\hat{K}; \hat{T}_1 < \hat{T}_2 < \ldots < \hat{T}_{\hat{K}}$:

$$\max_{0 \leq k \leq n-1} \left[ \hat{L}_k \cdot p_{IH}^k (1 - p_{IH})^{n-k} \right], \text{ where}$$

$\hat{L}_k$ : the maximized likelihood function for $k$ break points

$p_{IH}$ : 'a priori' probability of break at each time $(t = 1, \ldots, n - 1)$

Equivalent: $\min_{0 \leq k \leq n-1} \left[ \ln \frac{RSS_k}{n} + pen_k (p_{IH}) \right], \text{ where}$

$RSS_k$ : the minimized residual sum of squares for $k$ break points

$pen_k (p_{IH}) = -\frac{2k}{n} \ln \frac{p_{IH}}{1 - p_{IH}}$ : penalty terms depend on $p_{IH}$
Example for penalty terms applied in practice

Akaike criterion (AIC): \[ \text{pen}_k(p_{IH}) = \frac{2k}{n} \quad \Leftrightarrow \quad p_{IH} = \frac{e^{-1}}{1 + e^{-1}} \]

Schwarz criterion (BIC): \[ \text{pen}_k(p_{IH}) = \frac{k}{n} \ln n \quad \Leftrightarrow \quad p_{IH} = \frac{\sqrt{n^{-1}}}{1 + \sqrt{n^{-1}}} \]

My problems, questions for example:

- \( p_{IH} \) is ‘a priori’ probability of break instead of ‘a posteriori’!
  
  BUT: \( p_{IH} \) depends on the station network!

- What is the probability of errors for number of break points:
  
  Probability of type one error: \( P\left( \hat{K} > K \right) = ? \) (superfluous break points),
  
  Probability of type two error: \( P\left( \hat{K} < K \right) = ? \) (neglected break points)
Multiple break points detection procedure based on test of hypothesis on a given significance level (MASH)

The result of test procedure

If the detected break points of \( Z(t) \) are \( \hat{K}; \quad \hat{T}_1 < \hat{T}_2 < \ldots < \hat{T}_{\hat{K}} \), then on the given significance level \( p \) (e.g.: \( p=0.1 \)):

i, \( Z(t) \) is not homogeneous above the intervals \( (\hat{T}_{k-1}, \hat{T}_{k+1}] \) because,

\[
P \left( \exists (\hat{T}_{k-1}, \hat{T}_{k+1}] \text{ above that } Z(t) \text{ homogeneous} \right) = p
\]

Consequently the detected break points \( \hat{T}_k \) are not superfluous.
(NO serious type one error)

ii, \( Z(t) \) can be accepted to be homogeneous above the intervals \( (\hat{T}_{k-1}, \hat{T}_k] \).
(NO serious type two error)
METHODOLOGY FOR CORRECTION OF SERIES

Examination of difference series for estimation of shifts at the detected break points: 

\[ IH(\hat{T}_k - 1) - IH(\hat{T}_k) \quad (k = 1, \ldots, n - 1) \]

Possibilities, principles

a, In general: point estimation (maximum likelihood)

b, MASH: estimation is based on confidence intervals (test of hypothesis)

The philosophy of MASH:

Careful break points detection and correction iteration procedure in order to decrease the probability of type one error.

But using optimal series comparison for decreasing the probability of type two error i.e to increase the power.
II. HOMOGENIZATION OF DAILY DATA

Relation of daily and monthly homogenization

Alternative possibilities
– To use the detected monthly inhomogeneities directly for daily data homogenization
– Direct methods for daily data homogenization

Problems
– The direct use of the detected monthly inhomogeneities is probably not sufficient.
– Direct methods for daily data homogenization is probably not enough efficient thinking of the larger variability (less signal to noise ratio).

The Question
How can we use the valuable information of detected monthly inhomogeneities for daily data homogenization?
Software MASH  (Multiple Analysis of Series for Homogenization)

Homogenization of monthly series:
– Relative homogeneity test procedure.
– Step by step iteration procedure: the role of series (candidate, reference) changes step by step in the course of the procedure.
– Additive (e.g. temperature) or multiplicative (e.g. precipitation) model can be used depending on the climate elements.
– Including quality control and missing data complementing.
– Providing the homogeneity of the seasonal and annual series as well.
– Meta data (probable dates of break points) can be used automatically.
– The homogenization results can be verified.

Homogenization of daily series:
– Based on the detected monthly inhomogeneities.
– Including quality control and missing data complementing for daily data.
Possible Connection of Topics and Systems

LONG DATA SERIES
Complementing, Quality Control, Homogenization (MASH)
Examination of Representativity of a given Station Network (inside Network; statistical way)

CLIMATE EXAMINATIONS
E.g. Climate Change Detection

SPATIAL MODELLING OF CLIMATE PARAMETERS
Local Statistical Parameters
Stochastic Parameters

SPATIAL INTERPOLATION (MISH)
For arbitrary Location
Background Information maybe: satellite, radar, forecasting data

REPRESENTATIVITY EXAMINATION OF ARBITRARY STATION NETWORK
Inside the Network
For Optional Location
E.g. automatic stations

SHORT DATA SERIES
Complementing Quality Control
E.g. automatic stations

FORECASTING
E.g. Data Assimilation
Variational Analysis

--- Data and Method or/and Result
---- only Method or/and Result
---- only Data
There is no royal road!

Thank you for your attention!