Theoretical aspects of homogenization

Tamás Szentimrey

Hungarian Meteorological Service
Theoretical aspects = Mathematics

Topics:

Pending problems

Contradictions between theory and practice
Schema of Meteorological Examinations

1. Meteorology: Qualitative formulation of the problem.


John von Neumann: Without quantitative formulation of the meteorological questions we are not able to answer the simplest qualitative questions either.
Relation of daily and monthly homogenization

If we have daily series the general way is,
- calculation of monthly series
- homogenization of monthly series (larger signal to noise ratio)
- homogenization of daily series based on monthly inhomogeneities

Question

How can we use the valuable information of estimated monthly inhomogeneities for daily data homogenization?
A popular procedure ("variable correction" methods):

Homogenization of monthly series:
   Break points detection, correction in the first moment (mean)
   Assumption: homogeneity in higher moments (e.g. st. deviation)

Homogenization of daily series:
   Trial to homogenize also in higher moments
   Used monthly information: only the detected break points

My problems
- Inhomogeneity in higher moments: daily: yes versus monthly: no?
  Is it adequate model? Probability theory?
- Why are not used the monthly correction factors for daily homogenization?
Theorem (trivial)

Daily data: \( X(t) \) \((t = 1,2,..,30)\), monthly mean: \( \bar{X} = \frac{1}{30} \sum_{t=1}^{30} X(t) \)

Daily data with inhomogeneity in standard deviation:

\[
X_{ih}(t) = \alpha \cdot (X(t) - \mathbb{E}(X(t))) + \mathbb{E}(X(t)), \quad (t = 1,2,..,30)
\]

\[
\mathbb{E}(X_{ih}(t)) = \mathbb{E}(X(t)), \quad \mathbb{D}(X_{ih}(t)) = \alpha \cdot \mathbb{D}(X(t))
\]

Monthly mean: \( \bar{X}_{ih} = \frac{1}{30} \sum_{t=1}^{30} X_{ih}(t) \)

Then, monthly mean is also inhomogeneous in standard deviation:

\[
\mathbb{E}(\bar{X}_{ih}) = \mathbb{E}(\bar{X}), \quad \mathbb{D}(\bar{X}_{ih}) = \alpha \cdot \mathbb{D}(\bar{X})
\]
Homogenization of monthly data

Statistical spatiotemporal modelling of the series
Relative models and methods
Methodology for comparison of series
Break point (changepoint), outlier detection
Methodology for correction of series
Missing data completion
Usage of metada
Manual versus automatic methods
Relation of monthly, seasonal, annual series
Benchmark for methods
Statistical spatiotemporal modelling of monthly series

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 \( IH(T) – IH(T+1) \).

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

\( C \) : spatial covariance matrix, very important!
Theoretical question
What is with other models?

In Biology and Medicine:
Mixed Linear Model (MLM) is used for segmentation.

Is MLM adequate also for climate series?
The Mixed Linear Model is false for climate series theoretically!
Contradictions could be obtained if MLM were true for climate series.

Climate series: \( X(s, t) \)  
\( s \in D \): space, \( t \): time

There would be a consequence of MLM that,
\[
\text{var}(X(s_1, t) - X(s_2, t)) \leq \text{var}(X(s_1, t) - X(s_3, t)) + \text{var}(X(s_2, t) - X(s_4, t))
\]
if \( s_1 \neq s_3 \) and \( s_2 \neq s_4 \).

But let us assume:
\[
\|s_1 - s_3\| \ll \|s_1 - s_2\| \quad \text{and} \quad \|s_2 - s_4\| \ll \|s_1 - s_2\|
\]

Then just the opposite inequality is expected,
\[
\text{var}(X(s_1, t) - X(s_2, t)) > \text{var}(X(s_1, t) - X(s_3, t)) + \text{var}(X(s_2, t) - X(s_4, t))
\]

Contradiction!!!!!!!!!!
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)

The spatial covariance matrix $C$ may have a key role in methodology of comparison of series.

Optimal difference series can be applied for Detection and Correction procedures (MASH).
Break points (changepoint) detection

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

Key question of the homogenization software:
Automatic procedures for attribution of the break points for the candidate series!!!
Multiple break points detection for a difference series

Possibilities, principles for joint estimation of break points:
(Classical ways in mathematical statistics!)

**a, Bayesian Approach** (model selection, segmentation), penalized likelihood methods

Example: PRODIGE (Caussinus&Mestre)

**b, Multiple break points detection based on Test of Hypothesis**, confidence intervals for the break points

Example: MASH (Szentimrey)
Methodology for correction of series

Examination of difference series for estimation of shifts (correction factors) at the detected break points.

Possibilities, principles

a, In general: **Point Estimation**

a1, Least-Squares estimation (ANOVA): PRODIGE

a2, Maximum Likelihood method, Generalized-Least-Squares estimation (based on spatial covariance matrix $C$)

b, Estimation is based on **Confidence Intervals** (Test of Hypothesis): MASH
Remark:
The confidence intervals given for the break points and shifts make possible automatic use of metadata at MASH!
Automation of methods and software

Manual versus interactive or automatic methods?

In the practice numerous stations must be examined!
Stations per network: more than 100 instead of 10-15

Key questions for the methods and software:
- quality of homogenized data
- quantity of stations (automation!)

Necessary conditions for automation of methods, software:
- automatic attribution of break points for the candidate series
- automatic use of metadata
(mathematics!!)
Benchmark for methods
(to test methods on benchmark dataset)

Benchmark results depend on:
- Methods (quality, manual or automatic?)
- Benchmark dataset (quality, adequacy?)
- Testers (skilled or unskilled?)
- Methodology of evaluation (validation statistics?)

Remark (my opinion):
Theoretical evaluation of methods is also necessary!
Comparison of manual and automatic methods?

Similar to:
Comparison of handmade and factory products

For example:
A manual time consuming method with a skilled tester, versus an interactive method with an unskilled tester. Which method will be the better? (bad question)
Validation Statistics?

A tricky problem of reference period! (unsolved)

Some notations in case of additive model (temperature)

Month: \( k = 1,2, \ldots, 12 \), year: \( t = 1,2, \ldots, n \)

Original homogeneous series (orig): \( X_H^{(k)}(t) \)

Inhomogeneous series (inhom): \( X_{IH}^{(k)}(t) = X_H^{(k)}(t) + IH^{(k)}(t) \)

Estimated inhomogeneity series: \( IH^{(k)}(t) \)

Homogenized series (h0..): \( \hat{X}_H^{(k)}(t) = X_{IH}^{(k)}(t) - IH^{(k)}(t) \)

Reference time: in general it is \( n \), i.e.,

\[
IH^{(k)}(n) = 0 \quad \text{or} \quad \hat{X}_H^{(k)}(n) = X_{IH}^{(k)}(n)
\]
Problem 1

Let us assume \( n-1 \) is the only one detected break, i.e.,

\[ \exists k : I\hat{H}^{(k)}(n-1) \neq I\hat{H}^{(k)}(n) \]

Then the better methods do not change all the values for \( t = 1, \ldots, n-1 \). In this case the reference time is rather \( n-1 \) than \( n \), and the last values are taken as outliers, since we do not want damage the series!

The homogenized series:

\[ \hat{X}^{(k)}(t) = X_{IH}^{(k)}(t) \ (t = 1, \ldots, n-1), \quad \hat{X}^{(k)}(n) = X_{IH}^{(k)}(n) - \hat{O}_k \]
Problem 2

Let us assume that the benchmark series was defined with only one break at \( n-1 \),

\[
IH^{(k)}(t) \equiv C_k \ (t = 1, \ldots, n-1), \ IH^{(k)}(n) = 0
\]

Then the difference series:

\[
Z^{(k)}(t) = \hat{X}^{(k)}_H(t) - X^{(k)}_H(t) \equiv C_k \ (t = 1, \ldots, n-1)
\]

and

\[
Z^{(k)}(n) = \hat{X}^{(k)}_H(n) - X^{(k)}_H(n) = -\hat{O}_k
\]

In this case the validation statistics as \( RMSE(Z) \), \( CRMSE(Z) \) may be very large, opposite to the fact, that the “homogenization” was quite good.
The contradiction is arising from the non trivial problem of definition, interpretation of the reference period
Software MASHv3.03 (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 completion.
- Providing the homogeneity of the seasonal and annual series as well.
- Metadata (probable dates of break points) can be used automatically.
- The homogenization results and the metadata can be verified.

Homogenization of daily series:
- Based on the detected monthly inhomogeneities.
- Including quality control and missing data completion for daily data.
New version MASHv3.03
- COST ES0601 (HOME) format also can be used at I/O
- developments for automation

Can be downloaded:
http://www.met.hu/pages/seminars/seeera/downloads.htm

Some MASH speciality
- neighbourhood comparison of series
- use of spatial covariance for comparison of series
- automatic attribution of break points for candidate series
- automatic use of metadata
There is no royal road!

Thank you for your attention!