Mathematical questions of homogenization and summary of MASH

Tamás Szentimrey

Varimax Limited Partnership

Budapest

Introduction

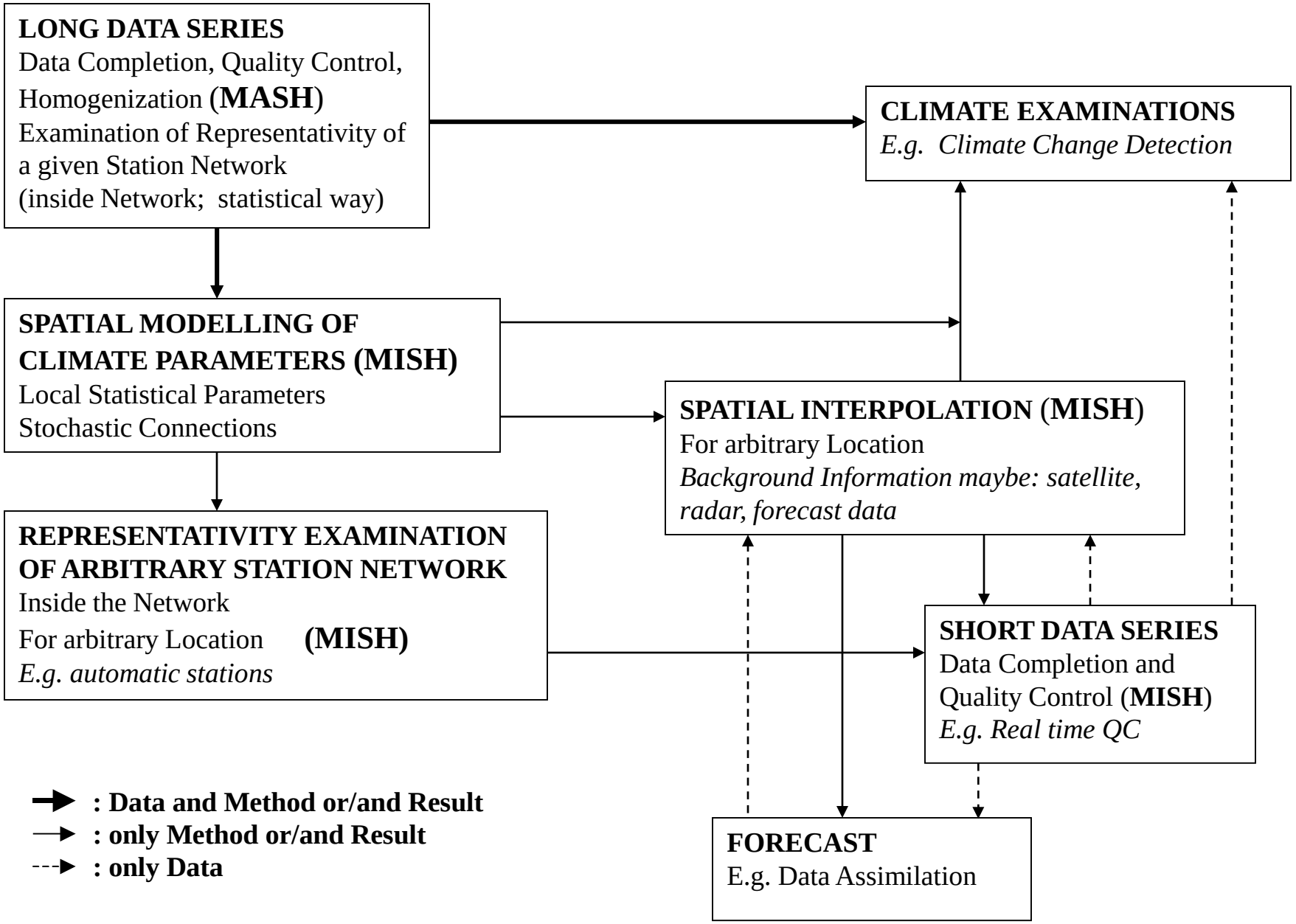
I retired from the Hungarian Meteorological Service.

I continue my activity in VARIMAX Limited Partnership.

Software: the new versions MASHv4.01 and MISHv2.01 are planned to share on the website of VARIMAX next year.

Presentation: summary of my conception on homogenization.

Possible Connection of Topics and Systems



Mathematics of homogenization of climate data series?

There are several methods and software in meteorology but

- there is no exact mathematical theory of homogenization!

Moreover,

- the mathematical formulation is neglected in general,

-“mathematical statements” without proof are in the papers,

- unreasonable dominance of the practice over the theory.

Fake News (miracle waiting): Artificial intelligence (AI),

Big data technology do not require mathematics.

But: No solution without advanced mathematics!

FUTURE? AI=Algorithms+Mathematics (Harari: Homo Deus)

WMO Guidelines on Homogenization

(for monthly data series)

CHAPTER 5. THEORETICAL BACKGROUND OF HOMOGENIZATION……40

5.1 General structure of the additive spatio-temporal models. 40

5.2 Methodology for comparison of series in the case of an additive model. 42

5.3 Methodology for breakpoint (changepoint) detection. 43

5.3.1 Breakpoint detection based on maximum likelihood estimation. 43

5.3.2 Breakpoint detection based on hypothesis testing. 43

5.3.3 Attribution of the detected breakpoints for the candidate series.. . 44

5.4 Methodology for adjustment of series. 44

5.5 Possibilities for evaluation and validation of methods. 44

MATHEMATICAL FORMULATION OF HOMOGENIZATION
(Distribution problem, not regression!)

Let us assume we have daily or monthly data series.

$Y_1(t)\ (t = 1,2,..,n)$: candidate series of the new observing system

$Y_2(t)\ (t = 1,2,..,n)$: candidate series of the old observing system

$1 \le T < n$: change-point

Before T: series $Y_2(t)\ (t = 1,2,..,T)$ can be used

After T: series $Y_1(t)\ (t = T+1,..,n)$ can be used

Theoretical cumulative distribution functions (CDF):

$$F_{1,t}(y) = \mathrm{P}(Y_1(t) < y) \ , \ F_{2,t}(y) = \mathrm{P}(Y_2(t) < y) \ , \quad t = 1,2,..,n$$

Functions $F_{1,t}(y)$, $F_{2,t}(y)$ change in time (e.g. climate change)!

Theoretical formulation of homogenization

Inhomogeneity: $F_{2,t}(y) \neq F_{1,t}(y) \quad (t = 1,2,..,T)$

Homogenization of $Y_2(t) \quad (t = 1,2,..,T)$:

$Y_{1,2h}(t) = F_{1,t}^{-1}\left(F_{2,t}\left(Y_2(t)\right)\right)$, then $\mathrm{P}\left(Y_{1,2h}(t) < y\right) = F_{1,t}(y)$

Transfer function: $F_{1,t}^{-1}\left(F_{2,t}\left(y\right)\right)$, Quantile function: $F_{1,t}^{-1}\left(p\right)$

Remark

The basis of the Quantile Matching methods can be integrated into the general theory. However these methods developed in practice mainly for daily data are very weak empiric methods.

It is not real mathematics! (good heuristics with poor mathematics)

Mathematical questions to be solved

The merged series: $Y_2(t)\ (t=1,2,..,T),\ Y_1(t)\ (t=T+1,..,n)$

The adjustment formula: $Y_{1,2h}(t)=F_{1,t}^{-1}\left(F_{2,t}\left(Y_2(t)\right)\right)\quad (t=1,2,..,T)$

Problem of calculatiom

- Estimation, detection of change point(s) T ?
- Estimation of distribution functions $F_{1,t}(y)$, $F_{2,t}(y)\ (t=1,2,..,T)$?

i, $F_{1,t}(y)$, $F_{2,t}(y)$ change in time (annual cycle, climate change)

ii, No sample for $F_{1,t}(y)\ (t=1,2,..,T)$

The problem is insolvable in general case!
Only relative methods can be used with some assumptions.
In addition some simplifications are necessary.

Special but basic case: Normal Distribution (e.g. temperature)

Theorem.

Let us assume normal distribution,

$$Y_1(t) \in N\left(E_1(t), D_1(t)\right), \quad Y_2(t) \in N\left(E_2(t), D_2(t)\right) \quad \left(t = 1,2,..,n\right)$$

$E_1(t), E_2(t)$: means $\quad D_1(t), D_2(t)$: standard deviations

Then the transfer function of homogenization:

$$Y_{1,2h}(t) = F_{1,t}^{-1}\left(F_{2,t}\left(Y_2(t)\right)\right) = E_1(t) + \frac{D_1(t)}{D_2(t)}\left(Y_2(t) - E_2(t)\right) \quad \left(t = 1,2,..,T\right)$$

Remarks:

i, A simple linear function and there is no “tail distribution” problem!

ii, Only the mean (***E***) and standard deviation (***D***) must be homogenized!

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?

Statistical spatiotemporal modelling of monthly series

Relative Additive Model (normal distribution, e.g. temperature)

Monthly series for a given month in a small region:

$$X_j(t) = \mu(t) + E_j + IH_j(t) + \varepsilon_j(t) \qquad \left(j = 1,2,\ldots,N\,;\; t = 1,2,\ldots,n \right)$$

μ : unknown climate change signal; E : spatial expected value;
IH : inhomogeneity signal in mean; ε : 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 $\boldsymbol{\varepsilon}(t) = [\varepsilon_1(t), \ldots\ldots, \varepsilon_N(t)]^{\mathrm{T}} \in N(\mathbf{0}, \mathbf{C})$ $(t = 1, \ldots, n)$ are independent

$\mathbf{C}$: spatial covariance matrix, very important!

Methodology for comparison of monthly series

All the examined series $X_j(t)\ \left(j = 1,..., N\right)$:
candidate and reference series alike.

Reference series are not assumed to be homogeneous!

Related questions: weighting of reference series,
difference series constitution, multiple comparison of series etc.

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

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

Break point (changepoint) detection for monthly series

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

Multiple break points detection procedures for difference series

(Classical ways in mathematical statistics!)

a, Bayesian Aproach (model selection, segmentation), penalized likelihood methods: HOMER, ACMANT

b, Multiple break points detection based on **Test of Hypothesis,** confidence intervals for the break points, that make possible automatic use of metadata: MASH

Methodology for adjustment of monthly series

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

Possibilities, principles

a, In general: **Point Estimation**

a1, Least-Squares (joint) estimation: HOMER, ACMANT

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

b, Estimation is based on **Confidence Intervals** (Test of Hypothesis): MASH

What is the practice for daily series?

A popular procedure

1. Homogenization of monthly mean series:

   Break points detection, adjustment of mean (*E*)

   Assumption: homogeneity of higher order moments (e.g. st. deviation (*D*))

2. Homogenization of daily series:

   Trial to homogenize also the higher order moments
   (e.g. Quantile Matching, Spline)

   Used monthly information: only the detected break points

Contradiction

- Inhomogeneity of higher moments, **daily: yes** versus **monthly: no** ?

  It is not adequate mathematical model for standard deviation (*D*)! (can be proved)

- Why are not used the monthly adjustment factors for daily homogenization?

An alternative procedure developed in MASH

MASHv4.01 (Multiple Analysis of Series for Homogenization, *T. Szentimrey*)
(https://www.met.hu/en/omsz/rendezvenyek/homogenization_and_interpolation/software/)

The MASH system is based on homogenization of monthly series derived from daily series. The procedures depend on the distribution of climate elements.

Quasi normal distribution (e.g. temperature)

Beside the monthly mean series another type of monthly series are also derived to estimate the inhomogeneity of standard deviation (*D*). These series are homogenized in standard deviation (*D*) by multiplicative model. The monthly mean series adjusted with the estimated inhomogeneity of standard deviation (*D*) are homogenized by additive model for mean (*E*).

Quasi lognormal distribution (e.g. precipitation)

Monthly mean or sum series are homogenized by multiplicative model.

The most important features of MASH

Homogenization of monthly series:

- Relative homogeneity test procedure.
- Step by step automatic iteration (Artificial Intelligence) procedure:
  the role of series (candidate, reference) changes step by step in the course of the procedure.
- Additive or multiplicative model can be used depending on the distribution of 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 (*E*, *D*).
- Including Quality Control and missing data completion for daily data.

Possibilities for evaluation and validation of methods

1. Theoretical, mathematical evaluation

2. Benchmarking

However the benchmark results depend on:

- Methods (quality, manual or automatic?)
- Benchmark dataset (quality, mathematics, adequacy?)
- Testers (skilled or unskilled?)
- Mathematics of evaluation (validition statistics?)

There is no royal road!

Thank you for your attention!