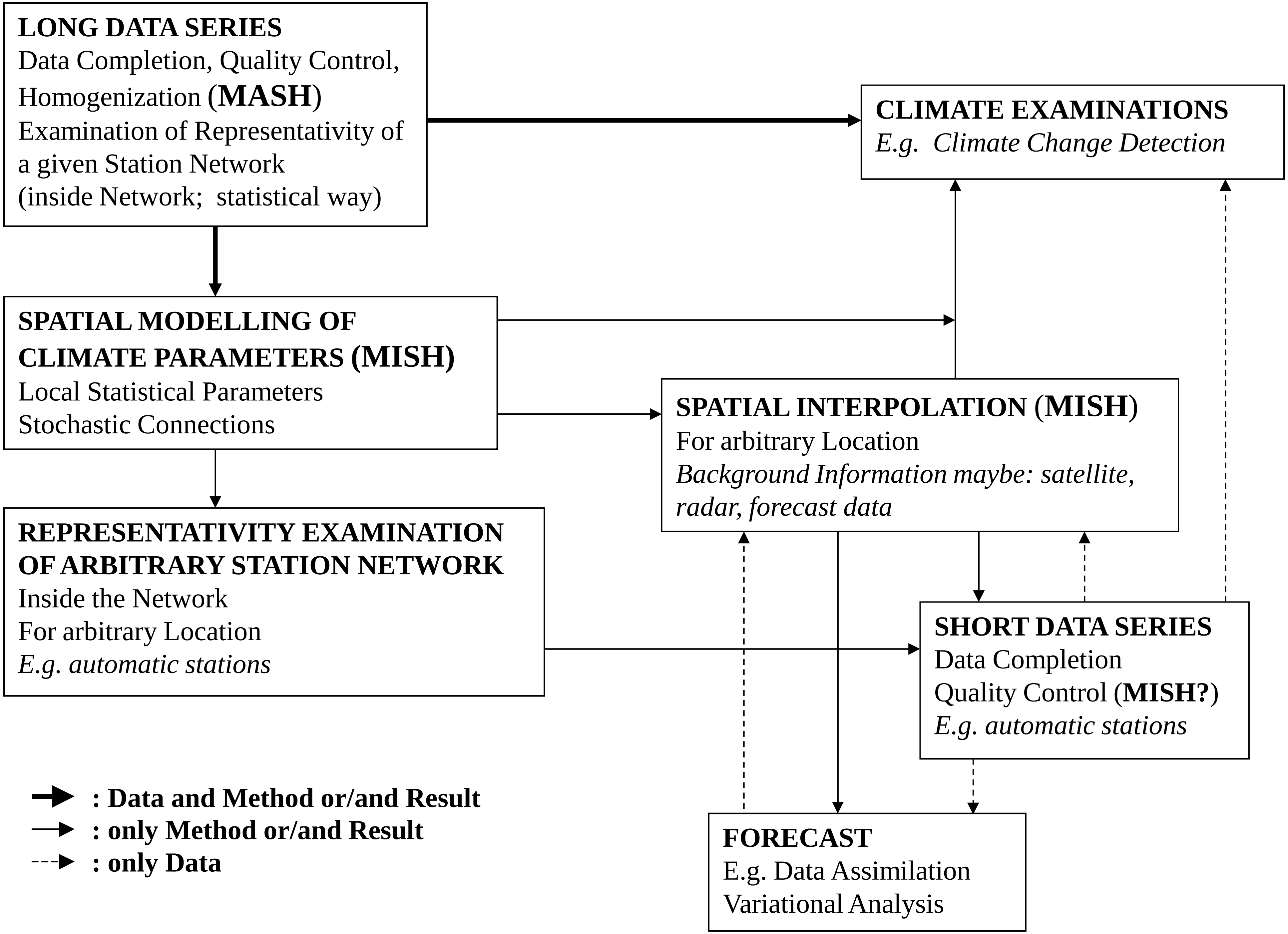
Mathematical questions of homogenization and quality control

Tamás Szentimrey, Mónika Lakatos, Zita Bihari

Hungarian Meteorological Service

Possible Connection of Topics and Systems



LONG DATA SERIES
Data Completion, 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 (MISH)
Local Statistical Parameters
Stochastic Connections

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

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

SHORT DATA SERIES
Data Completion
Quality Control (**MISH?**)
E.g. automatic stations

FORECAST
E.g. Data Assimilation
Variational Analysis

➔ **: Data and Method or/and Result**
→ **: only Method or/and Result**
⇢ **: only Data**

Schema of Meteorological Examinations

1. **Meteorology:** Qualitative formulation of the problem.

2. **Mathematics:** Quantitative formulation of the problem.

3. **Software:** Based on Mathematics.

4. **Meteorology:** Application of Software.

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

Mathematics of Homogenization?

There are several methods and software 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.

The poor mathematics is the main obstacle of the progress.

No solution without advanced mathematics!

Mathematical formulation of homogenization

Distribution problem, not regression!

Let us assume we have daily or monthly data series.

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

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

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

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

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

Probability distribution functions

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

$$y \in \big(-\infty, \infty\big) \ , \ t = 1,2,..,n$$

Climate change

Functions $F_{1,t}(y)$, $F_{2,t}(y)$ $\big(t = 1,2,..,n\big)$ change in time!

Definition of homogeneity

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

is homogeneous if $F_{2,t}(y) = F_{1,t}(y) \quad \big(t = 1,2,..,T\big)$.

Inhomogeneity: otherwise

Homogenization

Adjustment, correction of values $Y_2(t)\ \left(t = 1,2,..,T\right)$

in order to have the corrected values $Y_{1,2h}(t)\ \left(t = 1,2,..,T\right)$

with the same distribution as the elements

of series $Y_1(t)\ \left(t = 1,2,..,T\right)$ have, i.e.:

$$\mathrm{P}\left(Y_{1,2h}(t) < y\right) = F_{1,t}(y), \qquad y \in \left(-\infty, \infty\right),\ t = 1,2,..,T$$

Theorem (for homogenization)

i, Existence:

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

$$\mathrm{P}\left(Y_{1,2h}(t) < y\right) = F_{1,t}(y) \qquad \left(t = 1,2,..,T\right).$$

ii, Unicity:

If $h(s)$ is a strictly monotonous increasing function

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

Problem of calculation

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

Estimation of functions $F_{1,t}(y)$, $F_{2,t}(y)$ $\left(t = 1,2,..,T\right)$?

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

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

The problem is insolvable in general case!

Special but basic case: Normal Distribution

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: $\quad F_{1,t}(y) = \Phi\left(\dfrac{y - E_1(t)}{D_1(t)}\right) \quad , \quad F_{2,t}(y) = \Phi\left(\dfrac{y - E_2(t)}{D_2(t)}\right)$

where $\Phi(s)$ is the standard normal distribution function.

Consequently, the formula 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 on formula:

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

i, A simple linear formula, there is no "tail distribution" problem!

ii, Problem of estimation of $E_1(t)\,, D_1(t)\,, E_2(t)\,, D_2(t)\ \left(t = 1,2,..,T\right)$

- change in time (climate change)
- no sample for $E_1(t)\,, D_1(t)\ \left(t = 1,2,..,T\right)$

Assumptions (simplification)

$$D_2(t) = D_1(t)\ , \quad E_2(t) - E_1(t) = E \quad \left(t = 1,2,..,T\right)$$

$$\Rightarrow \qquad Y_{1,2h}(t) = Y_2(t) - E \qquad \left(t = 1,2,..,T\right)\ ,$$

That is homogenization in mean applied in practice mostly.

An observed phenomenon at extremes

The differences of parallel observations are larger in case of extremes.

Inhomogeneity? Different tail distribution?

In our opinion this observed phenomenon has a simple and logical reason.

The reason is that the extremes may be expected at different moments in case of parallel observations. It is a natural phenomenon.

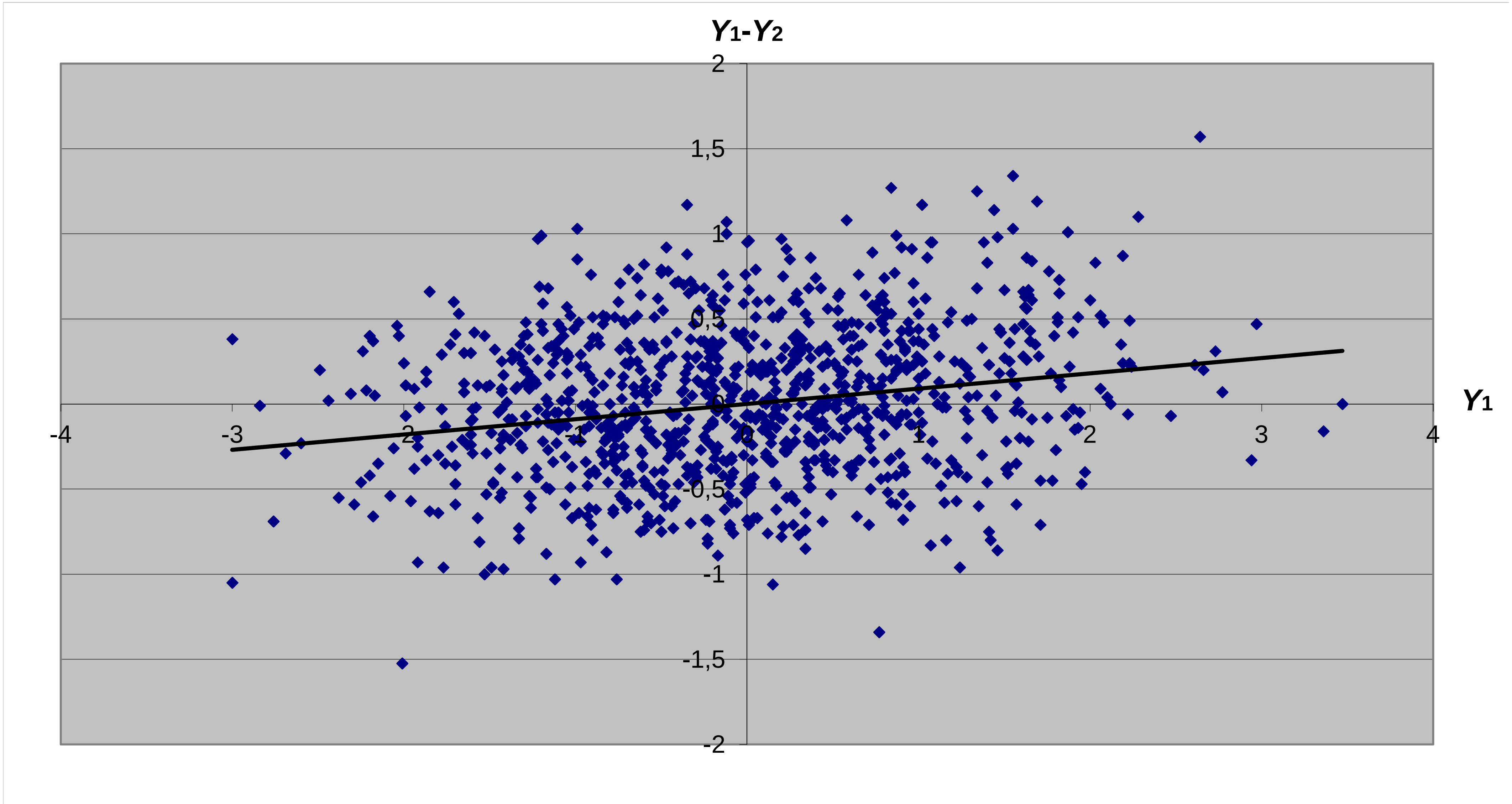
Or with other words, there maybe systematic biases in rank order!

An example is presented.

Example by Monte-Carlo method for natural dependence of $Y_1 - Y_2$ on Y_1

Generated series: $Y_1(t) \in N(0,1)$, $Y_2(t) \in N(0,1)$, $\mathrm{corr}\big(Y_1(t), Y_2(t)\big) = \rho = 0.9 \quad (t = 1,..,1000)$

Difference series: $Y_1(t) - Y_2(t)$, $\quad \mathrm{E}\big(Y_1(t) - Y_2(t) \,|\, Y_1(t)\big) = (1-\rho) \cdot Y_1(t) = 0.1 \cdot Y_1(t)$



Conditional homogenization also can be defined (variable correction?)

Let $X(t)$ $\left(t=1,2,..,n\right)$ be a homogeneous reference series.

Conditional homogenization of $Y_2(t)$ on $X(t)$,

$$Y_{1,2h}(t, X(t)) = F^{-1}_{1,t,x}\left(F_{2,t,x}\left(Y_2(t)\right)\right) \quad \Leftrightarrow \quad X(t)=x \qquad (t=1,2,..,T)$$

where $F_{1,t,x}(y)$, $F_{2,t,x}(y)$ are the conditional distribution functions of $Y_1(t)$, $Y_2(t)$, given $X(t)=x$, that is

$$F_{1,t,x}(y) = \mathrm{P}\left(Y_1(t) < y \,\middle|\, X(t)=x\right),\ F_{2,t,x}(y) = \mathrm{P}\left(Y_2(t) < y \,\middle|\, X(t)=x\right)$$
$$y \in \left(-\infty,\infty\right), \quad t=1,2,..,T$$

Then as a consequence of Bayes theorem:

$$\mathrm{P}\left(Y_{1,2h}(t, X(t)) < y\right) = F_{1,t}(y) \qquad y \in \left(-\infty,\infty\right),\ t=1,2,..,T$$

Theorem

If the joint distribution of $Y_1(t)$, $Y_2(t)$, $X(t)$ $\left(t = 1,2,..,T\right)$ is normal, $Y_1(t) \in N\left(E_1(t), D_1(t)\right)$, $Y_2(t) \in N\left(E_2(t), D_2(t)\right)$ and $\mathrm{corr}\left(Y_2(t), X(t)\right) = \mathrm{corr}\left(Y_1(t), X(t)\right)$ $\left(t = 1,2,..,T\right)$, then the linear formula is obtained again:

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

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

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? Not. Larger st. deviation at daily values implies larger st. deviation at monthly values (can be proved).

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

Overview on homogenization of monthly data in mean

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:

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

μ : unknown climate change signal; E : spatial expected value;
IH : inhomogeneity signal in mean; ε : normal noise

<u>Type of</u> $\mu(t)$: No assumption about the shape of this signal

<u>Type of inhomogeneity</u> $IH(t)$ <u>in general:</u> ’step-like function’

with unknown break points T and shifts $IH(T) - IH(T+1)$.

<u>Noise</u> $\boldsymbol{\varepsilon}(t) = \left[\varepsilon_1(t),\ldots\ldots,\varepsilon_N(t)\right]^{\mathrm{T}} \in N(\mathbf{0},\mathbf{C})$ $\left(t = 1,\ldots,n\right)$ are independent

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

Methodology for comparison of series

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

All the examined series $X_j(t)\ \left(j=1,\ldots,N\right)$:
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 $\mathbf{C}$ may have a key role in
methodology of comparison of series.

Optimal difference series can be applied for Detection
and Correction procedures.

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 Aproach (model selection, segmentation), penalized likelihood methods

Example: HOMER (Caussinus&Mestre), ACMANT (Domonkos)

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

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): HOMER, ACMANT

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

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 (validition statistics?)

Remark (my opinion)**:**
Theoretical evaluation of methods is also necessary!

Our Software

http://www.met.hu/en/omsz/rendezvenyek/homogenizationand_interpolation/software/

MASHv3.03

Multiple Analysis of Series for Homogenization; *Szentimrey, T.*

MISHv1.03

Meteorological Interpolation based on Surface Homogenized Data Basis;

Szentimrey, T.and Bihari, Z.

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.

There is no royal road!

Thank you for your attention!