Development of new version MASHv4.01 for homogenization of standard deviation

Tamás Szentimrey

Varimax Limited Partnership

Budapest

This year (2023) we have finished the new version MASHv4.01 that is able to homogenize the monthly and daily series also in the standard deviation i.e. the second order moment.

The software MASH has been developed for many years as an interactive automatic, artificial intelligence (AI) system that simulates the human intelligence and mimics the human analysis on the basis of advanced mathematics.

To illustrate that MASH is indeed an AI, some quotation from the Proceedings of the 4th Seminar for Homogenization (2004)

“**Programmed Statistical Procedure (Software: MASHv2.03)**

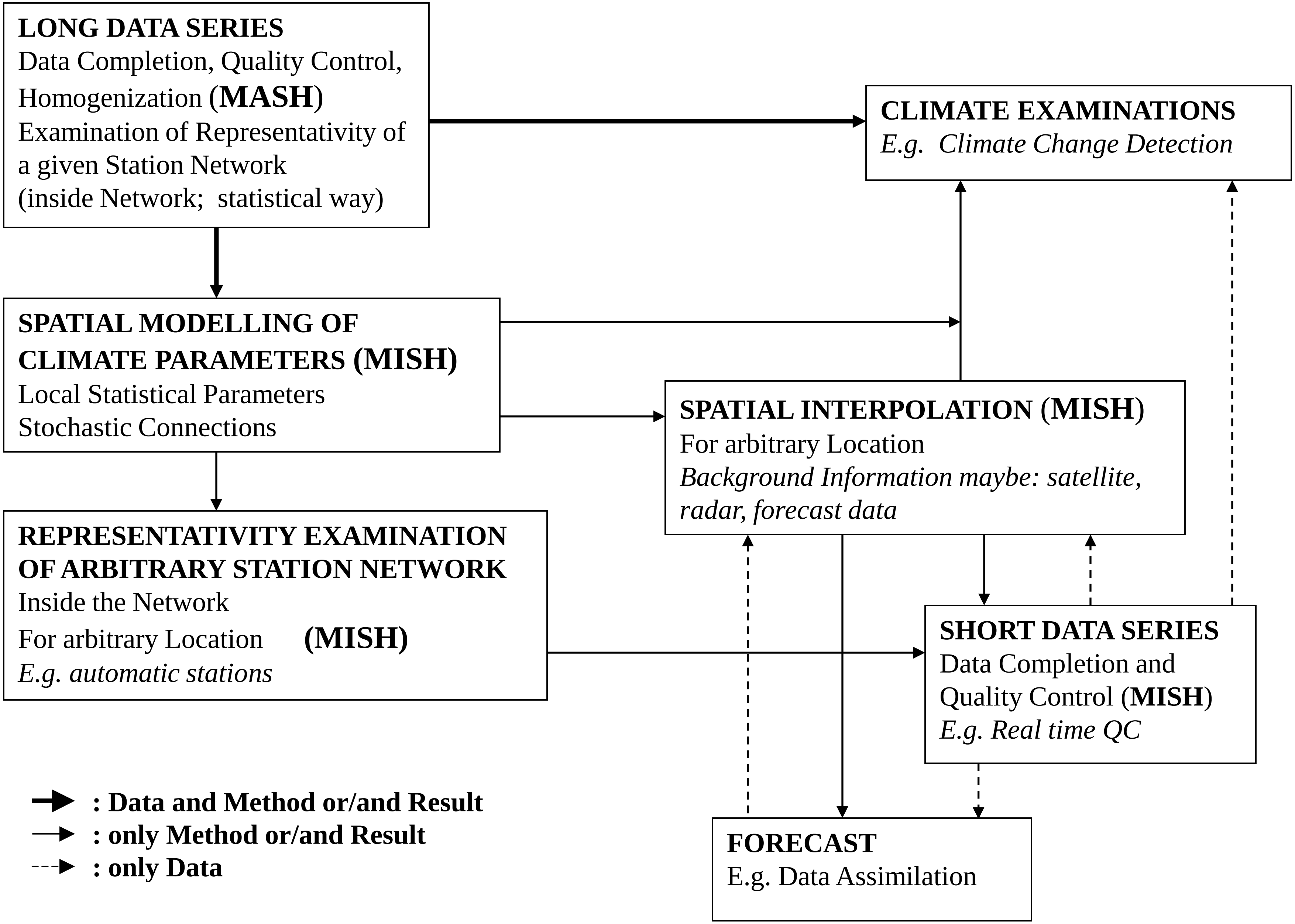
EXAMPLE Let us assume that there is a difficult stochastic problem.

In case of having relatively few statistical information:
– an intelligent human is possibly able to solve the problem, but it is time-consuming;
– the solution of the problem cannot be programmed.

In case of increasing the amount of statistical information:
– one is unable to discuss and evaluate all the information,
– but then the solution of the problem can be programmed. **(CHESS!!)**

AIM, REQUIREMENT
– Development of mathematical methodology in order to increase the amount of statistical information.
– Development of algorithms for optimal using of both the statistical and the ‘metadata’ information.”

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   **(MISH)**
E.g. automatic stations

SHORT DATA SERIES
Data Completion and
Quality Control (**MISH**)
E.g. Real time QC

FORECAST
E.g. Data Assimilation

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

Theoretical Background 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

$1 \le T < n$: break 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):

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

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

Mathematical Formulation of Homogenization

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

Homogenization of $Y_2(t) \ \left(t = 1,2,..,T\right)$**:**

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

Transfer function: $F_{1,t}^{-1}\left(F_{2,t}\left(Y_2(t)\right)\right)$

Remark

The basis of the Quantile Matching methods can be integrated into the general theory. But, good heuristics with poor mathematics.

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

Theorem 1

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:

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

Remarks:

i, A simple linear function and there is no pseudo problem of “tail distribution” for extremes, which problem is the Galton’s phenomenon: “regression towards the mean”.

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 in practice

(Monthly series for a given month in a small region)

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

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

Relative Multiplicative Model (e.g. precipitation)

$$X_j(t) = \mu(t) \cdot E_j \cdot IH_j(t) \cdot \exp\left(\varepsilon_j(t)\right)$$

μ : unknown climate change signal; E : spatial trend;
IH : inhomogeneity signal; ε : normal noise

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

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

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 (*Caussinus*, *Mestre*)
ACMANT (Adapted Caussinus-Mestre; *Domonkos*)

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

Methodology for adjustment of monthly series

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

Possibilities, principles

a, In general: **Point Estimation**

a1, Least-Squares (joint) estimation (ANOVA?):
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?

Theorem 2

Daily data: $Y(t)\ \left(t=1,..,30\right)$, **monthly mean:** $\bar{Y}=\frac{1}{30}\sum_{t=1}^{30}Y(t)$

Monthly variable for examination of standard deviation (*D*): $S=\sqrt{\frac{1}{29}\sum_{t=2}^{30}\left(Y(t)-Y(t-1)\right)^2}$

Daily data with inhomogeneity in mean (*E*) and standard deviation (*D*):

$$\mathrm{E}\left(Y_{ih}(t)\right)=\mathrm{E}\left(Y(t)\right)+\beta\ ,\quad \mathrm{D}\left(Y_{ih}(t)\right)=\alpha\cdot\mathrm{D}\left(Y(t)\right)\qquad \left(t=1,..,30\right)$$

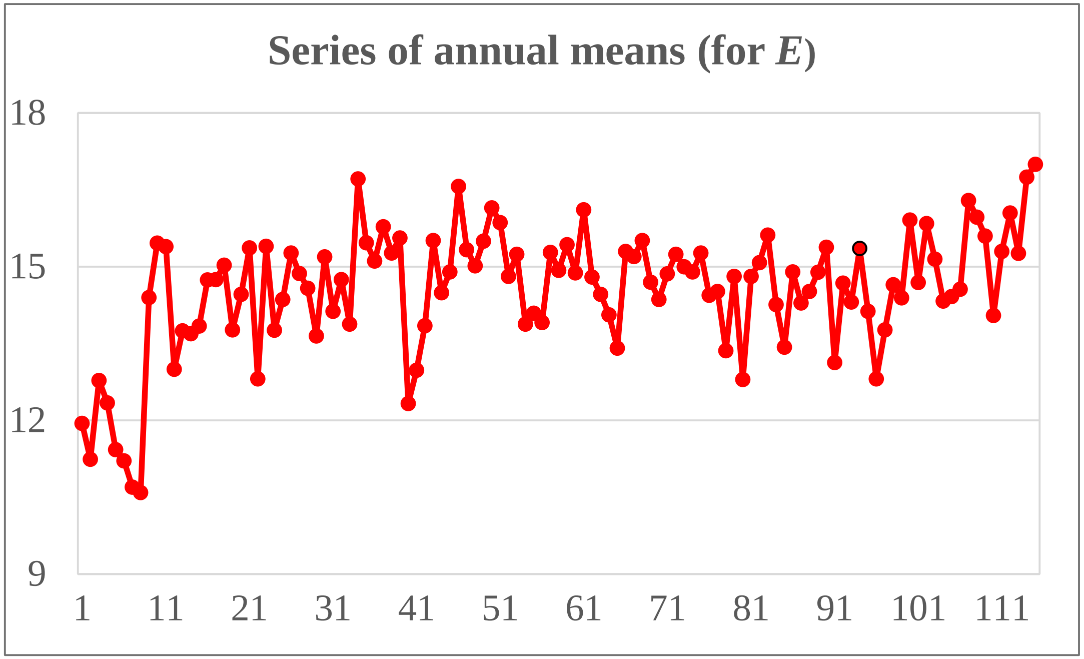
The appropriate monthly variables: $\bar{Y}_{ih}=\frac{1}{30}\sum_{t=1}^{30}Y_{ih}(t)$, $S_{ih}=\sqrt{\frac{1}{29}\sum_{t=2}^{30}\left(Y_{ih}(t)-Y_{ih}(t-1)\right)^2}$

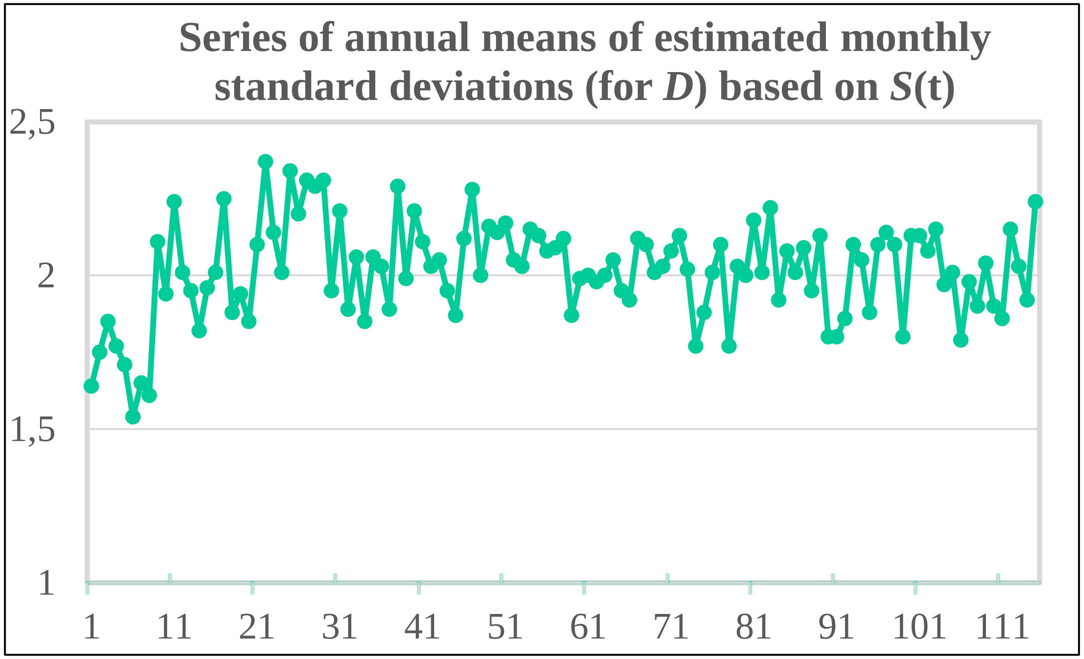
i, Then the monthly mean is also inhomogeneous in mean (*E*) and st. deviation (*D*):

$$\mathrm{E}\left(\bar{Y}_{ih}\right)=\mathrm{E}\left(\bar{Y}\right)+\beta \quad \text{and} \quad \mathrm{D}\left(\bar{Y}_{ih}\right)=\alpha\cdot\mathrm{D}\left(\bar{Y}\right)$$

ii, Moreover variable S_{ih} **can be used to estimate the inhomogeneity of st. deviation (*D*):**

$$\mathrm{E}\left(S_{ih}\right)=\alpha\cdot\mathrm{E}\left(S\right)$$





EXAMPLE (it is a real problem)

Maximum temperature series of Miskolc (in Hungary) 1901-2015

Inhomogeneity in 1901-1908, measured Réaumur: 1 ºC=0.8 Re

$$\mathrm{E}\left(Y_{ih}(t)\right) = 0.8 \cdot \mathrm{E}\left(Y(t)\right)$$

$$\mathrm{D}\left(Y_{ih}(t)\right) = 0.8 \cdot \mathrm{D}\left(Y(t)\right)$$

An alternative procedure developed in MASHv4.01

1. Homogenization of monthly series $S(t)$, $\overline{Y}(t)$.

Homogenization of series $S(t)$ by multiplicative model.

- Break points detection, estimation of inhomogeneity of st. deviation (*D*).

Adjustment of standard deviation (*D*) of series $\overline{Y}(t)$.

Homogenization of adjusted series $\overline{Y}(t)$ by additive model.

- Break points detection, estimation of inhomogeneity of mean (*E*).

Adjustment of mean (*E*) of series $\overline{Y}(t)$.

Assumption: homogeneity of higher order (>2) moments.

This assumption is always right in case of normal distribution **(*Theorem 1*)**!

2. Homogenization of daily series

Homogenization of mean (*E*) and standard deviation (*D*) on the basis of the monthly results. The used monthly information are the break points and the monthly adjustments of the mean (*E*) and standard deviation (*D*).

Adjustment of $Y_2(t)$ in mean (*E*) and st. deviation (*D*), the transfer formula:

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

<u>1. Adjustment of $Y_2(t)$ in standard deviation (*D*)</u>

Theoretical formula: $Y_{1,2hD}(t) = \frac{D_1(t)}{D_2(t)}\left(Y_2(t) - E_2(t)\right) + E_2(t) \qquad \left(t = 1,2,..,T\right)$

but $E_2(t)$ $\left(t = 1,2,..,T\right)$ are unknown.

Therefore, the applied formula: $Y_{1,2hD}(t) = \frac{D_1(t)}{D_2(t)}\left(Y_2(t) - \bar{E}_2\right) + \bar{E}_2$

where $\bar{E}_2$ is the mean value of $E_2(t)$ $\left(t = 1,2,..,T\right)$. $\bar{E}_2$ can be estimated by mean $\bar{Y}_2$.

Inhomogeneity of st. deviation $IH_D(t) = \frac{D_2(t)}{D_1(t)}$ **can be estimated by homogenizing the monthly standard deviation series** $S(t)$ **using multiplicative model.**

Adjustment of $Y_2(t)$: $Y_{1,2hD}(t) = \frac{Y_2(t)}{IH_D(t)} - IH_{D,E}(t)$, where $IH_{D,E}(t) = \left(\frac{D_1(t)}{D_2(t)} - 1\right)\bar{E}_2$

2. Adjustment of $Y_{1,2hD}(t)$ in mean (E)

$$Y_{1,2h}(t) = Y_{1,2hD}(t) - IH_{E,D}(t)$$

Inhomogeneity of $Y_{1,2hD}(t)$ in mean

$$IH_{E,D}(t) = \mathrm{E}\left(Y_{1,2hD}(t)\right) - E_1(t) = \left(\frac{D_1(t)}{D_2(t)} - 1\right)\left(E_2(t) - \bar{E}_2\right) + E_2(t) - E_1(t)$$

can be estimated by homogenizing the monthly mean series $Y_{1,2hD}(t)$ using additive model.

3. Summary of the adjustment of montly mean $Y_2(t)$ and the daily series

The adjustment of $Y_2(t)$ in mean (E) and standard deviation (D) can be written in the following linear function form:

$$Y_{1,2h}(t) = \frac{Y_2(t)}{IH_D(t)} - IH_E(t)\text{, where } IH_E(t) = IH_{D,E}(t) + IH_{E,D}(t)$$

For homogenization of daily data series in mean (*E*) and standard deviation (*D*) we also use this linear function form (*Theorem 2*).

Software MASHv4.01 (Multiple Analysis of Series for Homogenization) (2023, *T. Szentimrey*)

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 in mean (*E*).

Quasi lognormal distribution (e.g. precipitation)

Monthly mean or sum series are homogenized by multiplicative model.

Software MASHv4.01 (Multiple Analysis of Series for Homogenization)
(**2023**, *T. Szentimrey*)

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. Homogenization in mean (*E*) and st. deviation (*D*).
- 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.

Incorrect Sentences about MASH from a Book of Elsevier

P. Domonkos, R. Tóth and L. Nyitrai, 2022: "Climate observations: data quality control and time series homogenization"

- **"MASHv3 is better than MASHv4."** Sorry, but it is a misleading **BULLSHIT!**

Publication of Book: in 2022   Publication of MASHv4: in 2023 , 2022 < 2023

The authors couldn't know MASHv4! And MASHv3 is part of MASHv4.

- **"Novelties in MASHv4:** …the proposed algorithm easily detects false breaks of the standard deviation around the break points for the means. It is because the **empirical standard deviation** is elevated for periods including shifts in the means."

Sorry, but it is a false **Fake News!** We don't use the **empirical standard deviation!**

- **A funny personal note about me as the creator of MASH in the book:**

"The creator often chose unique mathematical solutions differing both from the traditional tools of climate data homogenization and from those suggested by other statisticians."   **Yes, because I am a mathematician!**

Conclusion: The Credibility of the Content of this Book is doubtful for me!

15 Hungarian July Mean Temperature Series 1901-2015

Estimated Inhomogeneities for St. Deviation (D)(%)

Series	IHD	Series	IHD	Series	IHD
8	8.05	9	7.98	4	6.73
12	4.88	7	4.08	11	3.59
6	3.33	2	2.43	15	2.22
5	2.16	13	2.02	10	1.70
1	1.57	14	1.34	3	0.54

AVERAGE: 3.51

$$D_{ih}(t) = D(t)\cdot IHD(t) \quad \left(t = 1,.,n\right),$$

$$IHD = \frac{100}{n}\sum_{t=1}^{n}\left|IHD(t)-1\right|$$

Estimated Inhomogeneities for Mean (E) (°C)

Series	IHE	Series	IHE	Series	IHE
3	0.80	8	0.55	15	0.53
7	0.52	12	0.48	10	0.48
14	0.31	6	0.31	5	0.29
11	0.24	1	0.23	4	0.14
9	0.13	2	0.09	13	0.08

AVERAGE: 0.35

$$E_{ih}(t) = E(t) + IHE(t) \quad \left(t = 1,.,n\right),$$

$$IHE = \frac{1}{n}\sum_{t=1}^{n}\left|IHE(t)\right|$$

There is no royal road!

(Archimedes)

Thank you for your attention!

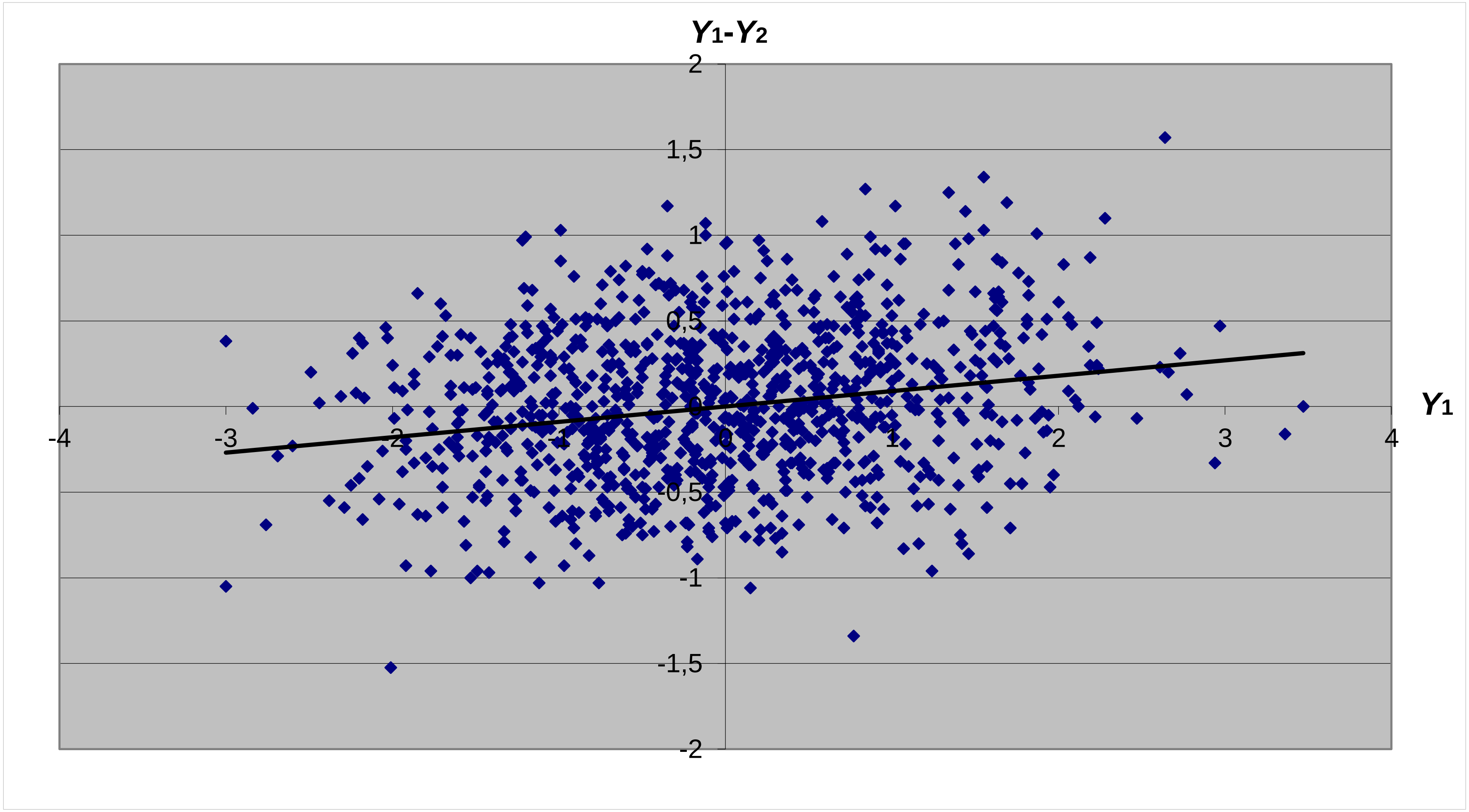
Parallel measurements, “tail distribution” problem or rather a natural phenomenon?

(GALTON: regression towards the mean)   (Szentimrey (2011), 7th Seminar for Homogenization)

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

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

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



Szentimrey (2008): “Series Comparison”, 6th Seminar for Homogenization

Maximum likelihood estimation for $\mu(t)$, $\mathbf{E}$, $\mathbf{T}(t)$, $\mathbf{v}$, if K is given:

(Generalized (!)-least-squares estimation, $\mathbf{C} \neq \sigma^2 \mathbf{I}$)

$$\min_{\mu(t),\mathbf{E},\mathbf{T}(t),\mathbf{v}} \left(\sum_{t=1}^{n} \left(\mathbf{X}(t) - \left(\mu(t)\mathbf{1} + \mathbf{E} + \mathbf{T}(t)\mathbf{v}\right)\right)^{\mathrm{T}} \mathbf{C}^{-1} \left(\mathbf{X}(t) - \left(\mu(t)\mathbf{1} + \mathbf{E} + \mathbf{T}(t)\mathbf{v}\right)\right) \right) =$$

$$= \min_{\mathbf{T}(t),\mathbf{v}} \left(\sum_{t=1}^{n} \left(\mathbf{Z}_c(t) - (\mathbf{I} - \mathbf{\Lambda})\mathbf{T}_c(t)\mathbf{v}\right)^{\mathrm{T}} \mathbf{C}_{\mathbf{Z}}^{-1} \left(\mathbf{Z}_c(t) - (\mathbf{I} - \mathbf{\Lambda})\mathbf{T}_c(t)\mathbf{v}\right) \right)$$

where $\mathbf{Z}_c(t) = \mathbf{Z}(t) - \overline{\mathbf{Z}}$, $\mathbf{T}_c(t) = \mathbf{T}(t) - \overline{\mathbf{T}}$,

and $\mathbf{C}_{\mathbf{Z}}^{-1}$ is a generalized inverse of the covariance matrix of $\mathbf{Z}(t)$.

<u>Consequently</u>: The solution(s) for $\mathbf{T}(t)$ and $\mathbf{v}$ are functions of the optimal difference series. The Maximum Likelihood method examines implicitly also the optimal difference series $\mathbf{Z}(t)\left(t = 1,\ldots,n\right)$!