Theoretical Problems of Homogenization and Spatial Interpolation

Tamás Szentimrey

Varimax Limited Partnership

Budapest

I also warmly welcome all the participants.

The Seminar series began 30 years ago in 1996!

Sándor Szalai and I were the founding fathers. Sándor's idea was us to organize such a seminar for homogenization and he had the leading part in the organization activity. Sándor unfortunately passed away in 2022.

It is already the 12th Homogenization Seminar since 1996 and it is the 7th Interpolation Conference since 2004 in Budapest. The specialty of both series was focusing on the mathematical methodology in meteorology!

Many thanks to the local organizers who continue this series.

Outline

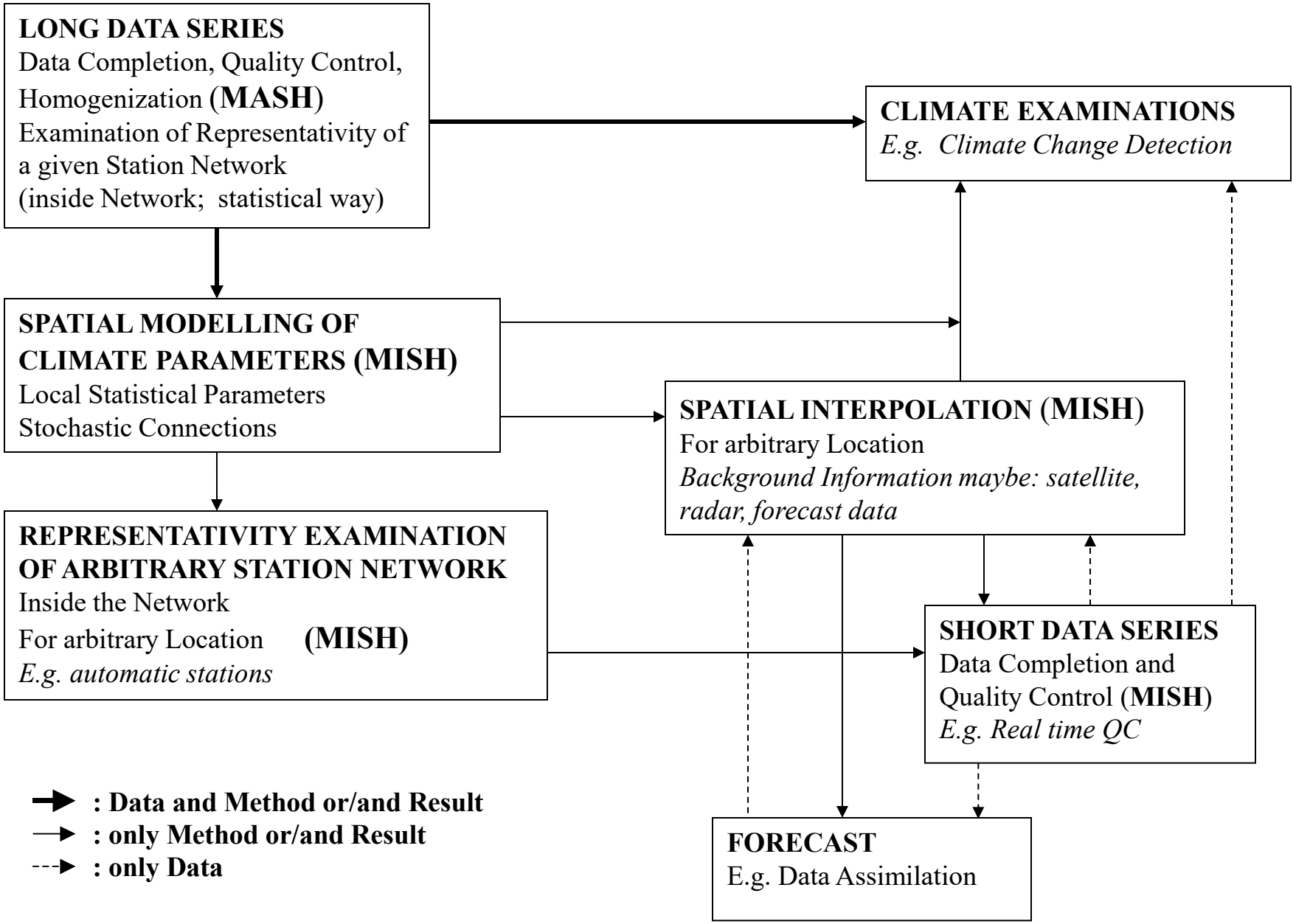
- Meteorological Topics, Methods
- Tools: Meteorology, Mathematics, Software
- Relation of Tools
- Theoretical considerations: methodological role of Climate
- Theoretical formulation of Homogenization
- Theoretical formulation of Spatial Interpolation
- Statistical modelling of present Climate
- Mathematics of Data Assimilation
- Software MISH-MASH

Meteorological Topics, Methods

To ensure high-quality meteorological data that are representative in both space and time, the following procedures and methods are required:

- Homogenization of climate data series, including quality control and missing data completion
- Spatial interpolation, gridding, real time quality control, interpolation with background information
- Data assimilation

Strong Connection between Topics and Systems



Tools: Meteorology, Mathematics, Software

The necessary conditions and tools for developing such procedures and methods are, in principle, meteorological knowledge, a proper meteorological formulation of the problem, and - based on these - advanced mathematical methodology and corresponding software (e.g. AI) development.

In practice, the methods and software used generally either are specialized meteorological methods and software (e.g. homogenzation) or originate from general-purpose statistical procedures and tools (e.g. GIS interpolation).

The main problem is that general-purpose, mathematically correct statistical procedures do not take into account specific meteorological aspects, while specialized meteorological methods typically do not fully meet strict mathematical requirements.

What is the Mathematics of Homogenization of Climate Data Series?

There are several methods and software in meteorology but

- there is no exact mathematical theory of homogenization!

- unreasonable dominance of the practice over the theory.

Miracle waiting: Artificial Intelligence (AI) does not require mathematics.

But: No solution without advanced mathematics!

"The real question is not whether machines think but whether men do."
(a provocative question from famous psychologist B. F. Skinner)

Some positive steps:

COST Action ES0601: Advances in Homogenization Methods of Climate Series: an integrated approach (HOME) (2011)

WMO Guidelines on Homogenization (2020)

What is the Mathematics of Spatial Interpolation in Meteorology?

– Nowadays the geostatistical interpolation methods built in **GIS** are applied in meteorology, e.g. the various kriging methods.

– The geostatistical methods with correct mathematics are based on spatial data only and cannot efficiently use the spatiotemporal data like the meteorological data series.

– While the meteorological data series make possible to obtain the necessary climate information i.e. to model the climate statistical parameters for spatial interpolation.

Theoretical considerations

Consequently, the theoretical problems of the mentioned meteorological procedures primarily arise in the mathematical methodology, and are therefore mathematical in nature.

What are these problems? In these procedures, the spatiotemporal climate probability distribution plays a key role. Therefore, effective methods can only be achieved if we have a quantitative characterization of the climate - which is not surprising!

However, the quantitative description of climate and the development of effective methods based on this knowledge require advanced mathematics.

Theoretical considerations

- The climate can be formulated as the probability distribution of the meteorological events or variables.

- The purpose of the statistical climatology should be to estimate or model the climate probability distribution or equivalently the climate statistical parameters.

- Furthermore the meteorological data series make possible to estimate or model the climate statistical parameters in accordance with the establishments of statistical climatology principles.

"Without a quantitative formulation of meteorological questions, we are not able to answer even the simplest qualitative questions!" (John von Neumann)

Mathematical Formulation of Homogenization

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 probability distribution functions:

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

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

Their precise estimation is impossible!

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

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

Theorem.

Let us assume normal distribution,

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

$E_1(t), E_2(t)$: expected values (means) $D_1(t), D_2(t)$: standard deviations

Then the transfer function of homogenization:

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

Remark

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

Homogenization in Practice

The necessary climate information can be derived directly from the climate data series themselves by statistical estimation.

Main Topics of Homogenization

Relation of daily and monthly data series homogenization

Statistical spatiotemporal modelling of climate data series

Methodology for comparison of climate data series

Break point (changepoint) detection for climate data series

Methodology for adjustment of climate data series

Quality control and missing data completion

Theoretical Formulation of Spatial interpolation

(normal distribution, e.g. temperature)

Daily or monthly data for a date (**s** any location vector)

Predictand: $Z(\mathbf{s}_0)$ Predictors: $Z(\mathbf{s}_i)\ (i = 1,...,M)$

Linear Interpolation Formula

$$\hat{Z}(\mathbf{s}_0) = \lambda_0 + \sum_{i=1}^{M} \lambda_i \cdot Z(\mathbf{s}_i)\ , \quad \text{where } \sum_{i=1}^{M} \lambda_i = 1\ .$$

Optimal Interpolation Parameters: $\lambda_i\ (i = 0,...,M)$ minimize RMSE

The Optimal Interpolation Parameters are known functions of the theoretical climate statistical parameters!

Optimal constant term: $\lambda_0 = \sum_{i=1}^{M} \lambda_i \left(E(\mathbf{s}_0) - E(\mathbf{s}_i) \right)$

where $E(\mathbf{s}_i)$ $(i = 0,...,M)$ are the expected values (means).

Vector of optimal weighting factors: $\boldsymbol{\lambda} = [\lambda_1,...,\lambda_M]^{\mathrm{T}}$

$$\boldsymbol{\lambda} = \mathbf{C}^{-1}\left(\mathbf{c} + \frac{\left(1 - \mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{c}\right)}{\mathbf{1}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{1}}\mathbf{1} \right) \qquad \text{(covariance form)}$$

and covariances $\mathbf{c}$, $\mathbf{C}$ can be written as functions of standard deviations $D(\mathbf{s}_i)$ $(i = 0,...,M)$ and correlations $\mathbf{r}$, $\mathbf{R}$.

Possibility for Spatial Interpolation in Practice, Climate modelling

Since we have not data for the predictand locations in general therefore we are not able to give estimates for the necessary climate statistical parameters. However, the required climate information can be modelled based on the available homogenized data series.
This modelled climate information can be used also for the data assimilation.

Thus, not only the future climate must be modelled, but the present climate as well.

Special advanced MATHEMATICS is needed!

Modelled monthly, daily spatiotemporal statistical parameters in MISH
(for a half minutes grid)

i. Spatial expected values (means) $E(\mathbf{s})$
ii. Spatial standard deviations $D(\mathbf{s})$
iii. Spatial correlations $r(\mathbf{s}_1, \mathbf{s}_2)$
iv. Temporal first-order autocorrelations $\rho(\mathbf{s})$

Consequently the first two spatiotemporal moments can be modelled for daily and monthly data by MISH! The normal distribution is uniquely determined by these moments.
The Optimum Interpolation Parameters $\lambda_i \ (i = 0,...,M)$ can be calculated from the above modelled parameters.

Mathematics of Data Assimilation (normality is assumed)

Theorem 1: The correct mathematical result of the Bayesian estimation for the atmospheric state $\boldsymbol{x}$ is obtained by minimizing the following cost function:

$$J(\boldsymbol{x}) = \left(\boldsymbol{x} - E(\boldsymbol{x}|\boldsymbol{x}_b)\right)^T \boldsymbol{B}^{-1}\left(\boldsymbol{x} - E(\boldsymbol{x}|\boldsymbol{x}_b)\right) + \left(\boldsymbol{y}_0 - E(\boldsymbol{y}_0|\boldsymbol{x})\right)^T \boldsymbol{R}^{-1}\left(\boldsymbol{y}_0 - E(\boldsymbol{y}_0|\boldsymbol{x})\right)$$

$\boldsymbol{x}$: the atmospheric state (predictand)

$\boldsymbol{x}_b$: the background, i.e. short range forecasts

$E(\boldsymbol{x}|\boldsymbol{x}_b)$: conditional expectation of $\boldsymbol{x}$, given $\boldsymbol{x}_b$

$\boldsymbol{y_0}$: the observations of the atmospheric state

$E(\boldsymbol{y}_0|\boldsymbol{x})$: conditional expectation of $\boldsymbol{y}_0$, given $\boldsymbol{x}$

$\boldsymbol{B}$, $\boldsymbol{R}$: covariance matrices

In essence: Interpolation with background information + Quality control

Problem in Forecasting: $J(\boldsymbol{x})$ is the basic formula for analysis field, but with an arbitrary assumption and incorrect practice: $E(\boldsymbol{x}|\boldsymbol{x}_b) = \boldsymbol{x}_b$

The conditional expected value $E(\boldsymbol{x}|\boldsymbol{x}_b) = \boldsymbol{x}_b$ is an incorrect a priori assumption, which has nothing to do with Bayesian theory. Why should it be true? For example, in the case of a completely bad forecast, i.e. if $\boldsymbol{x}_b$ is independent of $\boldsymbol{x}$, then $E(\boldsymbol{x}|\boldsymbol{x}_b) = E(\boldsymbol{x})$, where $E(\boldsymbol{x})$ is the expected value of $\boldsymbol{x}$, which is a climate statistical parameter! Consequently, modelling of climate is very important also for data assimilation!

The reason of this incorrect practice may be that the mathematical concept of conditional expectation is not known in the field of forecasting.

(Formalized by A. Kolmogorov (1933) using Radon–Nikodym theorem)

We published the mathematical derivation of the Theorem 1 with proof in:

Szentimrey, T. (2016): Analysis of the data assimilation methods from the mathematical point of view. In: *Mathematical Problems in Meteorological Modelling*, Springer International Publishing, 193–205

We wanted to publish the meteorological interpretation of the theorem with title:

Izsák, B., Szentimrey, T. (2025): Data assimilation - critical review. Does climate play a role in forecasting?

We sent the manuscript to prestigious meteorological Journals:

Monthly Weather Review (MWR) (American Meteorological Society)

Journal of Meteorological Research (JMR) (Chinese Meteorological Society

In both cases the manuscript was rejected without peer review process!

We received meaningless, dilettant excuses, without understanding the mathematical parts - particularly the conditional expected value. (Publication may be a scandal?)

It would be advisable for meteorological journals to involve qualified mathematicians as reviewers as well!

After this affair, I think the fundamental theoretical problem is:

The relationship between meteorology and mathematics!

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.

Software MISHv2.01 (Meteorological Interpolation based on Surface Homogenized Data Basis)
(*Szentimrey and Bihari*; under development, last shared version MISHv1.03)

I. Modelling system for climate statistical parameters in space (AI)
(expected values, standard deviations, spatiotemporal correlations)

- Based on long homogenized data series and model variables.
- Modelling procedure must be executed only once before the interpolation applications.

II. Spatial interpolation system (AI)

- Additive (e.g. temperature) or multiplicative (e.g. precipitation) model and interpolation formula can be used depending on the climate elements.
- Daily, monthly, annual values and many years' means can be interpolated.
- The expected interpolation error RMSE is modelled too.
- Real time Quality Control for daily and monthly data (additive model).
- Capability for application of background information such as satellite, radar forecast data. (with QC: data assimilation)
- Capability for gridding of data series.

There is no royal road!

(Archimedes)

Thank you for your attention!

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."

(Kasparov, Deep Blue, 1997)