General mathematical formulation of homogenization of climate data series

Tamás Szentimrey

Hungarian Meteorological Service
Content

Mathematics of homogenization of climate data series?

Critical remarks

Mathematical formulation of homogenization

- Definition of inhomogeneity of series
- Unconditional homogenization, theorems
- Relation with Quantile Matching methods
- Conditional homogenization, theorems
- Mathematical questions to be solved
- Special but basic case: Normal Distribution

What is in the practice?

- A popular procedure
- Relation of parallel observations and extremes
- An alternative procedure for mean and st. deviation
- Homogenization of standard deviation in MASH!
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.

No solution without advanced mathematics!
Theoretical questions

Mathematical formulation of homogenization of climate data series?

What are the mathematical problems to be solved? And what is the solution?

Evaluation of the mathematical base of the methods applied in practice?

The following contradiction is an often phenomenon at the methods:

good heuristic ideas  versus  poor mathematics

The ratio of the problems to be solved: 90% mathematics, 10% meteorology

But at the methods applied in practice: 10% mathematics, 90% meteorology
Mathematical formulation of homogenization

Distribution problem, not regression!

Let us assume we have daily or monthly data series.

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

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

\( 1 \leq T < n \): change-point

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

After \( T \): series \( Y_1(t) \ (t = T + 1, \ldots, n) \) can be used
Theoretical (!) cumulative distribution functions (CDF)

\[ F_{1,t}(y) = P(Y_1(t) < y) \quad , \quad F_{2,t}(y) = P(Y_2(t) < y) \]

\[ y \in (-\infty, \infty) \quad , \quad t = 1,2,..,n \]

Natural change (annual cycle, climate change)

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

Definition of inhomogeneity

The merged series

\[ Y_2(t) \quad (t = 1,2,..,T) \quad , \quad Y_1(t) \quad (t = T+1,..,n) \]

is inhomogeneous,

if identity \( F_{2,t}(y) \equiv F_{1,t}(y) \quad (t = 1,2,..,T) \) is not true.
Homogenization

Adjustment, correction of values $Y_2(t)\ (t = 1,2,\ldots,T)$

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

with the same distribution as the elements of series $Y_1(t)\ (t = 1,2,\ldots,T)$ have, i.e.:

$$P(Y_{1,2h}(t) < y) = P(Y_1(t) < y) = F_{1,t}(y) \quad t = 1,2,\ldots,T$$
Theorem for (unconditional) homogenization

i, Existence:
If \( Y_{1,2h}(t) = F_{1,t}^{-1}\left(F_{2,t}(Y_2(t))\right) \) then
\[
P\left(Y_{1,2h}(t) < y \right) = F_{1,t}(y) \quad (t = 1,2,..,T).
\]

ii, Unicity:
If \( h(s) \) is a strictly monotonous increasing function and \( P\left(h(Y_2(t)) < y \right) = F_{1,t}(y) \), then \( h(s) = F_{1,t}^{-1}\left(F_{2,t}(s)\right) \).
Quantile Matching (QM) methods

The basis of these methods can be integrated into the general theory since the transfer function,

\[ Y_{1,2h}(t) = F_{1,t}^{-1} \left( F_{2,t} \left( Y_2(t) \right) \right) \]

is equivalent with,

\[ Y_{1,2h}(t) = Y_2(t) + \left( F_{1,t}^{-1}(p) - F_{2,t}^{-1}(p) \right) \]

where \( F_{1,t}^{-1}(p) \), \( F_{2,t}^{-1}(p) \) are the quantile functions and \( F_{2,t}(Y_2(t)) = p \).

However the QM methods developed in practice mainly for daily data are very weak empiric methods. It is not real mathematics! (good heuristics with poor mathematics; brave people!)
Conditional homogenization based on given events

Let \( B = \{ B_j : j = 1, 2, \ldots, M \} \) be a complete system of events:

\[
B_i \cap B_j = \emptyset, \quad \sum_{j=1}^{M} P(B_j) = 1 \quad (\text{e.g. macrosynoptic weather situations})
\]

Conditional homogenization of \( Y_2(t) \) on given events \( B \),

\[
Y_{1,2h}(t, B) = F_{1,t,B_j}^{-1} \left( F_{2,t,B_j} \left( Y_2(t) \right) \right) \quad \Leftrightarrow \quad B_j \text{ occurs at } t \quad (t = 1, 2, \ldots, T)
\]

where \( F_{1,t,B_j}(y), F_{2,t,B_j}(y) \) are the conditional distribution functions of \( Y_1(t), Y_2(t) \), given \( B_j \), that is

\[
F_{1,t,B_j}(y) = P(Y_1(t) < y \mid B_j), \quad F_{2,t,B_j}(y) = P(Y_2(t) < y \mid B_j)
\]

\( y \in (-\infty, \infty), \quad t = 1, 2, \ldots, T \)

Then as a consequence of Bayes and total probability theorems:

\[
P(Y_{1,2h}(t, B) < y) = F_{1,t}(y) \quad y \in (-\infty, \infty), \quad t = 1, 2, \ldots, T
\]
Mathematical questions to be solved

The simpler case: unconditional homogenization

The merged series are given: \( Y_2(t) \ (t = 1, 2, \ldots, T), \ Y_1(t) \ (t = T + 1, \ldots, n) \)

The transfer function is: \( Y_{1,2h}(t) = F_{1,t}^{-1}\left(F_{2,t}\left(Y_2(t)\right)\right) \) \( (t = 1, 2, \ldots, T) \)

Problems:

Estimation, detection of change point(s) \( T \)?

Estimation of distribution functions \( F_{1,t}(y), F_{2,t}(y) \) \( (t = 1, 2, \ldots, T) \)?

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

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

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,

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

\[ E_1(t), E_2(t) : \text{means} \quad D_1(t), D_2(t) : \text{standard deviations} \]

Then the transfer function of homogenization:

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

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!
What is in the Practice?

A popular procedure

1. Homogenization of monthly series:
   Break points detection, correction of mean
   Assumption: homogeneity of higher order moments (e.g. st. deviation)

2. Homogenization of daily series:
   Trial to homogenize also the higher order moments
   (Quantile Matching (HOM, RHtests), Spline (SPLIDHOM))
   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 st. deviation!

- Why are not used the monthly correction factors for daily homogenization?
What is the reason of this “popular procedure”?

An observed phenomenon at extremes

The differences of parallel observations are larger in case of extremes. What is this? Inhomogeneity in the tails of the distributions?
No, 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: $Y_1(t) \in N(0,1)$, $Y_2(t) \in N(0,1)$, $\text{corr}(Y_1(t), Y_2(t)) = \rho = 0.9$ \hspace{1em} (t = 1,..,1000)

Difference series: $Y_1(t) - Y_2(t)$, \hspace{1em} $E(Y_1(t) - Y_2(t) \mid Y_1(t)) = (1 - \rho) \cdot Y_1(t) = 0.1 \cdot Y_1(t)$
An alternative procedure

1. Homogenization of monthly series:
   Break points detection, correction of the mean and standard deviation.
   The correction is based on the normal distribution transfer function.
   Assumption: homogeneity of higher order (>2) moments.
   This assumption is always right in case of normal distribution!

2. Homogenization of daily series:
   Homogenization of mean and standard deviation on the basis of the monthly results. The used monthly information are the break points and the monthly corrections of the mean and standard deviation.
Developing of MASH for homogenization of Standard Deviation
(Multiple Analysis of Series for Homogenization; Szentimrey, T.)

Order of steps
1. Homogenization of Standard Deviation of data series
2. Homogenization of Mean of data series

Principles of Methodology

Multiple break points detection procedures for Mean and St. Deviation.
Procedures based on Test of Hypothesis.
Confidence Intervals are also given for the break points.
(make possible automatic use of metadata).

Estimation of the correction factors for Mean and St. Deviation.
Estimation is based on Confidence Intervals.
Test Statistics for St. Deviation Before Homogenization

Critical value (significance level 0.05): 26.8

<table>
<thead>
<tr>
<th>Series</th>
<th>TSB</th>
<th>Series</th>
<th>TSB</th>
<th>Series</th>
<th>TSB</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>141.13</td>
<td>11</td>
<td>50.79</td>
<td>14</td>
<td>33.13</td>
</tr>
<tr>
<td>6</td>
<td>28.19</td>
<td>12</td>
<td>26.07</td>
<td>13</td>
<td>24.93</td>
</tr>
<tr>
<td>5</td>
<td>22.69</td>
<td>2</td>
<td>22.11</td>
<td>4</td>
<td>21.54</td>
</tr>
<tr>
<td>1</td>
<td>20.87</td>
<td>3</td>
<td>19.44</td>
<td>10</td>
<td>17.46</td>
</tr>
<tr>
<td>9</td>
<td>15.31</td>
<td>8</td>
<td>11.74</td>
<td>15</td>
<td>10.26</td>
</tr>
</tbody>
</table>

AVERAGE: 31.04

Test Statistics for Mean Before Homogenization

Critical value (significance level 0.05): 21.76

<table>
<thead>
<tr>
<th>Series</th>
<th>TSB</th>
<th>Series</th>
<th>TSB</th>
<th>Series</th>
<th>TSB</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>1262.47</td>
<td>5</td>
<td>926.31</td>
<td>10</td>
<td>831.81</td>
</tr>
<tr>
<td>7</td>
<td>637.72</td>
<td>6</td>
<td>558.66</td>
<td>3</td>
<td>506.91</td>
</tr>
<tr>
<td>15</td>
<td>500.07</td>
<td>8</td>
<td>463.98</td>
<td>11</td>
<td>320.14</td>
</tr>
<tr>
<td>13</td>
<td>288.51</td>
<td>14</td>
<td>249.40</td>
<td>9</td>
<td>197.54</td>
</tr>
<tr>
<td>1</td>
<td>166.10</td>
<td>4</td>
<td>134.41</td>
<td>2</td>
<td>88.90</td>
</tr>
</tbody>
</table>

AVERAGE: 475.53
17 Annual mean Maximum Temperature Series 1950-2007
(Network real 000005 of COST Benchmark)

Test Statistics for St. Deviation Before Homogenization

<table>
<thead>
<tr>
<th>Series</th>
<th>TSB</th>
<th>Series</th>
<th>TSB</th>
<th>Series</th>
<th>TSB</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>41.27</td>
<td>5</td>
<td>21.28</td>
<td>7</td>
<td>20.34</td>
</tr>
<tr>
<td>12</td>
<td>13.23</td>
<td>13</td>
<td>12.69</td>
<td>4</td>
<td>12.01</td>
</tr>
<tr>
<td>1</td>
<td>11.06</td>
<td>14</td>
<td>9.89</td>
<td>11</td>
<td>9.17</td>
</tr>
<tr>
<td>8</td>
<td>9.15</td>
<td>17</td>
<td>7.53</td>
<td>2</td>
<td>6.78</td>
</tr>
<tr>
<td>15</td>
<td>6.62</td>
<td>9</td>
<td>6.24</td>
<td>16</td>
<td>5.29</td>
</tr>
<tr>
<td>10</td>
<td>5.07</td>
<td>3</td>
<td>4.14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AVERAGE: 11.87

Test Statistics for Mean Before Homogenization

<table>
<thead>
<tr>
<th>Series</th>
<th>TSB</th>
<th>Series</th>
<th>TSB</th>
<th>Series</th>
<th>TSB</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>351.61</td>
<td>14</td>
<td>245.58</td>
<td>12</td>
<td>227.40</td>
</tr>
<tr>
<td>16</td>
<td>189.04</td>
<td>2</td>
<td>183.80</td>
<td>3</td>
<td>168.03</td>
</tr>
<tr>
<td>8</td>
<td>138.70</td>
<td>5</td>
<td>130.57</td>
<td>6</td>
<td>108.53</td>
</tr>
<tr>
<td>1</td>
<td>96.95</td>
<td>17</td>
<td>80.80</td>
<td>15</td>
<td>56.89</td>
</tr>
<tr>
<td>7</td>
<td>47.16</td>
<td>10</td>
<td>37.68</td>
<td>11</td>
<td>24.90</td>
</tr>
<tr>
<td>13</td>
<td>21.09</td>
<td>9</td>
<td>18.47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

AVERAGE: 125.13
15 Hungarian Annual Mean Temperature Series 1901-2014

Test Statistics for St. Deviation After Homogenization

Critical value (significance level 0.05):  26.8

<table>
<thead>
<tr>
<th>Series</th>
<th>TSA</th>
<th>Series</th>
<th>TSA</th>
<th>Series</th>
<th>TSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>28.19</td>
<td>12</td>
<td>27.66</td>
<td>14</td>
<td>26.34</td>
</tr>
<tr>
<td>5</td>
<td>24.13</td>
<td>2</td>
<td>22.11</td>
<td>4</td>
<td>21.54</td>
</tr>
<tr>
<td>1</td>
<td>20.87</td>
<td>3</td>
<td>20.54</td>
<td>10</td>
<td>18.92</td>
</tr>
<tr>
<td>6</td>
<td>18.18</td>
<td>11</td>
<td>15.61</td>
<td>13</td>
<td>14.96</td>
</tr>
<tr>
<td>9</td>
<td>14.85</td>
<td>15</td>
<td>12.53</td>
<td>8</td>
<td>11.74</td>
</tr>
</tbody>
</table>

AVERAGE: 19.88

Test Statistics for Mean After Homogenization

Critical value (significance level 0.05):  21.76

<table>
<thead>
<tr>
<th>Series</th>
<th>TSA</th>
<th>Series</th>
<th>TSA</th>
<th>Series</th>
<th>TSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>32.81</td>
<td>13</td>
<td>31.46</td>
<td>9</td>
<td>29.11</td>
</tr>
<tr>
<td>12</td>
<td>27.27</td>
<td>8</td>
<td>25.37</td>
<td>7</td>
<td>25.29</td>
</tr>
<tr>
<td>5</td>
<td>22.93</td>
<td>1</td>
<td>21.91</td>
<td>4</td>
<td>21.85</td>
</tr>
<tr>
<td>6</td>
<td>18.96</td>
<td>2</td>
<td>18.52</td>
<td>15</td>
<td>17.23</td>
</tr>
</tbody>
</table>

AVERAGE: 24.09
There is no royal road!

Thank you for your attention!