Detection and correction of breakpoints in longterm German precipitation series

E. Rustemeier, A. Kapala, V. Venema und C. Simmer

Meteorological Institute
University of Bonn

Cost „HOME” and 7th Homogenization Seminar Meeting
26.10.2011
• About 2000 longterm monthly precipitation station
• NOT quality controlled
• Few meta data
Meta data

- Almost not available for old data
  - Lost during the wars
  - Disposed
- Incomplete
  - Relocation
  - Missing values
- Sometimes wrong
  - Observation time
- Only partly digitized
  - Cannot be included in programs
There are many reasons for inhomogeneities
Distribution of daily precipitation stations

Daily data digitized by the German Weather Service and the Meteorological Institute of the University of Bonn.

- 118 precipitation stations
- 100 years of daily data
- Few missing values
- (Quality controlled)
1 Introduction
   • Data base
2 Homogenization
   • Network selection
   • Automatization I
   • Automatization II
   • Monthly correction
3 Summary and outlook
Homogenization course

- Selection of neighbor stations
- Detection of breakpoints
- Correction of breaks

- Calculate log ratios
- Optimization method
- Shortest path algorithm
- Binary coding
- Multiple linear regression
Network selection

For homogenization measuring stations with similar climate conditions are needed. → Network selection

- Monthly indices
- Removal of the 30 years moving trend
- Precipitation < 1 mm → becomes 0 mm
- Principal component analysis S-Mode
- Varimax Rotation
Neighbor stations

- Pairwise comparison of the stations $i, j$

$$y_{i,j} = \ln \left( \frac{\text{Precipitation totals station}}{\text{Precipitation totals neighbor station}} \right)$$
Homogenization course

- Selection of neighbor stations
- Detection of breakpoints
- Correction of breaks

- Calculate log ratios
- Optimization method
- Shortest path algorithm
- Binary coding
- Multiple linear regression
Overview original Prodigé

- Homogenization program Météo France
- Developed by Olivier Mestre (Caussinus and Mestre, 2004)

Detection

- Pairwise comparison of the stations
- \[ y = \ln \left( \frac{\text{precipitation totals station}}{\text{precipitation totals reference station}} \right) \]

Selecting of breaks

- By comparison of pairings

Correction

- By linear regression
Overview original Prodigie

Assumptions

- Between two breaks is the time series homogeneous
- Artificial trends can be corrected in setting breaks
- Normal distribution of the logarithmic ratio of two stations
- Constant variance during the entire time series

Disadvantage

- Only partly automatic
  Prodigie proposes breakpoint positions - the user decides
Detection of breaks
Manual part in original Prodigie.
For each pairing of stations $ij$ there exists:

- a **break vector** with entries for every year consisting of
  - 0 (no break)
  - 1 (break)

  e.g.:

  $$break_{i,j} = [0 0 0 1 0 0 0 1 0 \ldots]$$

- the **correlation** between these two for example

  $$cor_{i,j} = 0.8$$
Automatization
Part 1 - Thresholds

Only neighbor stations with a correlation > 0.5 are chosen

\[
\begin{array}{cccccc}
& 1935 & 1936 & 1937 & 1938 & 1939 & \text{Korr.} \\
\text{break}_{i,1} & = & \ldots & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 & \ldots & 0.64 \\
\text{break}_{i,2} & = & \ldots & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & \ldots & 0.57 \\
\text{break}_{i,3} & = & \ldots & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & \ldots & 0.49 \\
\text{break}_{i,4} & = & \ldots & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & \ldots & 0.56 \\
\text{break}_{i,5} & = & \ldots & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & \ldots & 0.57 \\
\text{break}_{i,6} & = & \ldots & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & \ldots & 0.62 \\
\text{break}_{i,7} & = & \ldots & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & \ldots & 0.55 \\
\text{break}_{i,8} & = & \ldots & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & \ldots & 0.62 \\
\text{break}_{i,9} & = & \ldots & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & \ldots & 0.58 \\
\text{break}_{i,10} & = & \ldots & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & \ldots & 0.59 \\
\text{break}_{i,11} & = & \ldots & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & \ldots & 0.57 \\
\text{break}_{i,12} & = & \ldots & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & \ldots & 0.46 \\
\hline
\text{break}_i & = & \ldots & 0.00 & 0.11 & 0.20 & 0.69 & 0.00 & \ldots & 0.69
\end{array}
\]

\[
\text{break}_{i,t} = \frac{\sum_j \text{break}_{i,j,t} \cdot \text{cor}_j}{\sum_j \text{cor}_j}
\]

Thresholds:
\[
\begin{bmatrix}
0.9 & 0.7 & 0.5 & 0.4 & 0.35 \\
\end{bmatrix}
\]
Comparison Benchmark data set

Standard deviation diff. (anomalies) - Surrogate & Precipitation

- Climatol (170)
- SNHT DWD (170)
- C3SNHT (170)
- RhTestV2 rel (170)
- RhTestV2 abs (170)
- AnClim S Z (170)
- Craddock Brunetti (015)
- MASH Marinova (014)
- MASH SL (170)
- PRODIGE monthly (170)
- PRODIGE trendy (170)
- PRODIGE regular (170)
- PRODIGE Automatic (170)
- Inhom. data

-201816141210-8 -6 -4 -2 0 2 4 6 8 101214161820
First results

Number of breaks

![Map of Germany with breaks and bar chart showing the number of found inhomogeneities per station. The bar chart indicates a peak at 1 and 2 breaks per station, with fewer stations having 0 or 3 breaks.]
First results
Example Zugspitze

- Original data
- — homogenized data

1910 building constructions
1930 cutting middle mountain top
1938 cutting western top
1945+ instrumental change
First results
Example Zugspitze
Pairwise comparison of the stations $i, j$

$$y_{i,j} = \ln \left( \frac{\text{Precipitation totals station}}{\text{Precipitation totals neighbor station}} \right)$$
Combining neighbor stations to one reference station

- The smaller the difference in climate signal between target and reference station the smaller are the breaks that can be found.
- Without breaks criging would be the first choice

- Strong breaks in the nearest station cause a break in the reference series
Pairwise detection in Prodigie

\[ y_{i,j} = \ln(\text{Station}_i) - \ln(\text{Station}_j) \]

Automatically
Neighbor stations are combined to a single reference station

\[ y_i = \text{standardized}(\ln(\text{Station}_i)) - \text{Reference station}_i \]

\[ \text{Reference station}_i = \frac{\sum_{k \neq i} \text{standardized}(\ln(\text{Station}_k))}{\text{Number of neighbor stations}} \]
Monthly correction of breaks

Monthly correction

- Estimation of a monthly correction
- Maintaining of the annual cycle
- Last segment is not considered to be correct
## Validation: Benchmark

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>HSS</th>
<th>norm. CRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>yearly</td>
</tr>
<tr>
<td>Prodige thresh.</td>
<td>0.281</td>
<td>0.599</td>
</tr>
<tr>
<td>Reference station</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all stations</td>
<td>0.318</td>
<td>0.593</td>
</tr>
<tr>
<td>least correlated neighbor station excluded</td>
<td>0.292</td>
<td>0.635</td>
</tr>
</tbody>
</table>

HSS (Break point position): the higher the better
normalized CRMSE (Trend): the smaller the better
1 Introduction
   • Data base
2 Homogenization
   • Network selection
   • Automatization I
   • Automatization II
   • Monthly correction
3 Summary and outlook
Summary

- Automatization of Prodige
  - Using thresholds
  - Construction of a reference station
- Monthly instead of yearly correction
  ⇒ Yearly trends do not change
  ⇒ Monthly trends improve
- Indices can be used to estimate the break point position
Sensitive study - Networks
  - Influence of single stations
  - Less correlated
  - Results for number of breaks and position

Comparison with other algorithms
  - MASH
  - Genetic algorithms

Quality control
Thank you for your attention!