A blind test of correction algorithms for inhomogeneities in daily time series

Introduction

As part of the COST Action E0601 HOME (Advances in homogenisation methods of climate series: an integrated approach, 2007-2011), within working group WG5 (Methods for homogenisation of daily data) a dataset was generated that serves as a validation tool for correction of daily inhomogeneities. The dataset contains daily air temperature data with a length of 100 years and was generated based on the temperature series from the Czech Republic.

The validation dataset contains both small networks as well as pairs. Different types of inhomogeneities have been inserted into the series. Parametric breaks in the first three moments were introduced and the influence of relocation was simulated by exchanging the distribution of two nearby stations. The surrogate daily temperature data reproduce: distribution, distribution of difference series (cross-correlations) and power spectra (auto-correlation function). They do not reproduce: cross-spectra (cross-correlation function), outliers, missing data nor local trends. For the further analysis, only data with known breaks were applied (but with unknown structure of inhomogeneities).

Data and methods

Benchmark or validation?

- Benchmark
  - blind validation (no feedback)
  - error metrics known in advance
  - error metrics computed by third party
  - realistic problem
  - maturity field

- Validation
  - does not have to be blind (correct errors)
  - introduce metrics as needed during analysis
  - test limits with unrealistic problems
  - new developing field

* do not get to use a realistic case
* do not know the error metrics that well

Structure of the validation dataset

- Surrogate daily temperature data
- Based on data from the Czech Republic
- 1000a of data
- Breaks known & breaks unknown
  - used only data with date of break known
  - five breaks per 100a
- Three data sections
  - network (12)
  - pairs (42)
  - pairs-dedicated (21)
  - three set with 7 networks with the same break dates

Homogeneous data

- Surrogate daily temperature data
- Original data was made 100a by mirroring
- Surrogates reproduce
  - distribution
  - distribution of difference time series
  - Cross-correlations:
    - pairs: moderate 0.70 to high 0.98
    - network: 0.92 to 0.98
  - Three types of breaks
    - switching of stations
    - simulate relocation (only for pairs)
  - change in mean
  - change in higher moments

Inhomogeneities

- Five breaks per 100a
  - date random
  - minimum HSP four years
  - station and its best correlated pair
  - no annual cycle
  - No local trends (no outliers, no missing data)
  - Cross-correlations:
    - pairs: moderate 0.70 to high 0.98
    - network: 0.92 to 0.98
  - Three types of breaks
    - switching of stations
    - simulate relocation (only for pairs)
    - outliers
    - missing data

Validation metrics

- CRMSE or trends in mean are not on the focus
  - at least daily CRMSE
  - Distribution of HSP
  - Kolmogorov-Smirnov test
  - Cramér-von Mises test
  - Percentile regression

Potential problems

- Only Central European temperature available
- Lower correlations than usual – problems for some of the methods
- Some methods contributed only with certain networks

Current Conclusions

- Absolute methods are "dangerous": it is better to apply no methods than to use absolute ones since they produce datasets with errors
- As expected, the simpler methods, correcting only the mean, are the best ones at reducing CRMSE
- Evaluation of individual breaks (for various inhomogeneities) still to come
- We need to better understand the nature of inhomogeneities in daily data, i.e. what are the factors behind differences between measurements at two neighbouring sites. For this purpose a new COST Action dealing with comparative measurements could represent an important opportunity to improve the current understanding (http://tinyurl.com/paralleldata)

Some of the Results

Contributions

- Stepanek P (1,2), Venema V. (3), Toreti A. (7), Nemec J. (5), Zahradnicke P (1,2)
- Contributed to the COST Action E0601 HOME (Advances in homogenisation methods of climate series: an integrated approach, 2007-2011), within working group WG5 (Methods for homogenisation of daily data)
- Submitted results for Benchmark or Validation?

Network dataset

- Mean RMSE of the first six networks (100a) / temp (°C)

<table>
<thead>
<tr>
<th>Network</th>
<th>Mean RMSE (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homogeneous</td>
<td>0.12</td>
</tr>
<tr>
<td>Inhomogeneous</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Pair dataset

- Mean RMSE of the first six pairs (100a) / temp (°C)

<table>
<thead>
<tr>
<th>Pair</th>
<th>Mean RMSE (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homogeneous</td>
<td>0.13</td>
</tr>
<tr>
<td>Inhomogeneous</td>
<td>0.19</td>
</tr>
</tbody>
</table>

References

- For more information on the COST Action on homogenisation see: http://www.homogenisation.org/
- Victor Venema's homogenization page: http://www2.meteo.uni-bonn.de/venema/themes/homogenisation/

Contacts:

petr.stepanek@chmi.cz, victor.venema@uni-bonn.de, jguijarrop@aemet.es