Data inhomogeneity creates artificial long-range dependence in historical temperature data

H.W. Russ1, O. Mestre 2 and V. Venema 3

1 Physics Institute, University of Potsdam and Potsdam Institute for Climate Impact Research, Potsdam, Germany,
2 Météo-France, Toulouse, France,
3 Institute of Meteorology, University of Bonn, Germany

Introduction

We are investigating the effect of inhomogeneities in data on the estimation of long-range dependence (LRD). Examples of inhomogeneities in empirical temperature data are sudden jumps caused by a relocation of the measurement station, or a new type of shelter. With homogenisation algorithms utilising the differences between multiple nearby stations, it is possible to detect such jumps and correct for them. We analyse a set of temperature time series before and after homogenisation with respect to LRD and find that in most cases the LRD parameter is reduced for the homogenised series. In order to test whether the homogenisation artificially reduces LRD, we create sets of simulated data from a LRD process, introduce jumps and subject them afterwards to the homogenisation procedure. This test provides evidence that the homogenisation algorithm does not remove natural LRD dependencies, it compensates only for the bias in the estimate due to data inhomogeneities.

Methods

Homogenisation Algorithm

The inhomogeneities are detected by a statistical penalised log-likelihood procedure which is used to detect an unknown number of breaks and outliers [1]. The algorithm initially detects jumps in pairs of stations. These jumps are then attributed to a specific station by comparing multiple pairs. To this list of statistically detected inhomogeneities, the known inhomogeneities are added. We then correct each series by using moving neighbourhoods. The size of the correction is different for every month.

Long-Range Dependence

Let \( X \) be a stationary stochastic process with ACF \( \rho(\tau) \). \( X \) is long-range dependent (LRD) if \( \rho(\tau) = o(\tau^{-d}) \), e.g., for a slowly polynomially decaying ACF \( \rho(\tau) = \tau^{-d} \). If the ACF is summable, the process is short-range dependent (SRD). For \( 0 < d < 0.5 \) the process is LRD.

Model parameters can be estimated using exact maximum-likelihood or, as we did here, the Whittle-approximation [2]. Model orders \( p \) and \( q \) can be determined using model selection criteria as the AIC or the likelihood-ratio test [3]. The Hurst exponent estimator is given by \( \hat{H} \) for ARMA processes.

Detrended Fluctuation Analysis (DFA)

DFA recently has become popular as an estimator of the LR parameter or Hurst exponent \( H \). It is based on a straight-line fit to \( \log(F(s)) \) vs. \( \log(s) \) for large scales \( s \), where \( F(s) \) is the fluctuation function for scale \( s \), e.g., [5]. The estimator \( \hat{H}_{\text{DFA}} \) has no known limiting distribution and is thus a heuristic estimator without means to derive confidence intervals. Therefore statistical inference is not possible.

Case Study

We compare the Hurst exponent estimates \( \hat{H}_{\text{ARIMA}} \) and \( \hat{H}_{\text{DFA}} \) from raw temperature series to the corresponding estimates of their homogenised counterpart. In the case study, we have monthly temperature data from 24 stations in France.

Results for LRD Data

<table>
<thead>
<tr>
<th>Farima Modelling</th>
<th>DFA Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{H}_{\text{INH}} )</td>
<td>( \hat{H}_{\text{INH}} )</td>
</tr>
<tr>
<td>( \hat{H}_{\text{ORIG}} )</td>
<td>( \hat{H}_{\text{ORIG}} )</td>
</tr>
</tbody>
</table>

Discussion

Causes of sudden jumps in the data can be a relocation of the measurement station, a new type of shelter, measurement times or a replacement of the measurement instrument. These problems are not anectodical. During the construction of the HISTALP precipitation dataset [1] “on average one break could be detected every 23rd year in a series of 136 years in length.” 192 precipitation series were studied, and none of them could be considered free of inhomogeneities. Also in the French dataset we study none of the original series could be considered reliable, being affected by at least three or four artificial shifts.

References