European Geosciences Union General Assembly 2010 | Vienna, Austria, 2-7 May 2010 | Session HS5.5: Stochastics in hydrometeorological processes

Optimal infilling of missing values in hydrometeorological time series

Y. Dialynas, P. Kossieris, K. Kyriakidis, A. Lykou, Y. Markonis, C. Pappas, S.M. Papalexiou and D. Koutsoyiannis
Department of Water Resources and Environmental Engineering, National Technical University of Athens (www.itia.ntua.gr)

1. Abstract

Being a group of undergraduate students in the National Technical University of Athens, attending the course of Stochastic Methods in Water Resources, we study, in cooperation with our tutors, the infilling of missing values of hydrometeorological time series from measurements at neighbouring times. The literature provides a plethora of methods, most of which are reduced to a linear statistical interpolating relationship. Assuming that the underlying hydrometeorological process behaves like either a Markovian or a Hurst-Kolmogorov process we estimate the missing values using two techniques, i.e., (a) a local average (with equal weights) based on the optimal number of measurements referring to a number of forward and backward time steps, and (b) a weighted average using all available data. In each of the cases we determine the unknown quantities (the required number of neighbouring values or the sequence of weights) so as to minimize the estimation mean square error. The results of this investigation are easily applicable for infilling time series in real-world applications.

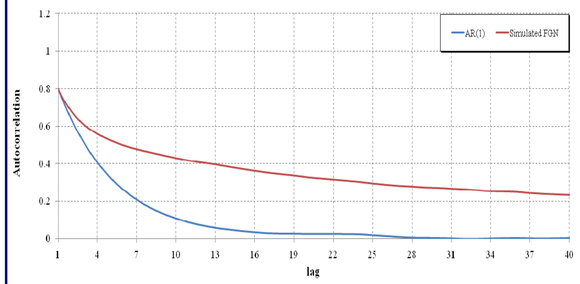
2. Motivation

- Missing values in hydrometeorological time series is a common problem.
- The literature provides a plethora of methods for infilling missing values in time and space (Thiessen, Normal Ratio, BLUE, Organic Correlation, etc.). Almost all methods are based on a weighted average of (non missing) observations (e.g. Koutsoyiannis and Langousis, 2010).
- Here we consider the simplest method where all weights are equal to 1.
- Interpolation in space is typically made by taking the average of nearest points (e.g. within a certain distance), rather than the global average of all points with measurements.
- Paradoxically, infilling in time the global average of all available data is usually preferred.
- We seek to explore if this practice is justified and under which conditions. In other words, we investigate when a global time average is preferable over a local average and when not.
- We investigate two types of time dependence of the underlying process: Markovian (short-range dependence) and Hurst – Kolmogorov (HK; long-range dependence).

3. Long range vs. short range dependence (1)

- Long range dependence, also known as the Hurst phenomenon, is common in hydrological and other geophysical processes. In simple words it expresses the tendency of wet years to cluster into wet periods, or of dry years to cluster into drought periods (e.g. Koutsoyiannis, 2002).
- On the other hand, short range dependence appears in Markovian processes where the future does not depend on the past when the present is known.
- The AR(1) (autoregressive of order 1) model is a very popular representative of Markovian processes, whereas the Fractional Gaussian Noise (FGN) process (e.g. Mandelbrot, 1965), or HK process, is widely used for reproducing long range dependence.
- The major difference in the above models is that while in AR(1) the autocorrelation declines exponentially, in HK it is a power function of lag thus resulting in slow decay of autocorrelation.

4. Long range vs. short range dependence (2)



Autocorrelation function for $\rho_1 = 0.8$ for the AR(1) and FGN processes

For the FGN process (for example when the Hurst coefficient is $H = 0.96$), the autocorrelation for lag 40 (years) is as high as 0.23, whereas in the Markovian process the autocorrelation is practically zero even for much shorter lags.

5. Methodology (1)

- The interpolation problem refers to the estimation of an unknown quantity y from known values x^i ($i = 1, \ldots, n$) (measurements) of the same quantity and the same time period at different points (or at different time periods on the same point). Mathematically the interpolation problem can be simply expressed as a linear equation:

$$\underline{y} = w^1 \underline{x}^1 + \ldots + w^n \underline{x}^n + \underline{e}$$

- where $\underline{x}, \underline{y}, \underline{e}$: random variables
  w^i: weighting factor
  $\underline{e}$: estimation error

- In vector form:

$$\underline{y} = \boldsymbol{w}^T \underline{\boldsymbol{x}} + \underline{e} \quad \leftrightarrow \quad \underline{e} = \underline{y} - \boldsymbol{w}^T \underline{\boldsymbol{x}}$$

where $\boldsymbol{w} := [w^1, \ldots, w^n]^T$
$\underline{\boldsymbol{x}} := [\underline{x}^1, \ldots, \underline{x}^n]^T$

6. Methodology (2)

- Assumptions:
  - the setting is stationary
  - equal weighting factors $w^i = 1$ (for simplicity and direct applicability)
- The optimal infilling produces minimum Mean Square Error (MSE):

$$\mathrm{MSE} := E[\underline{e}^2] = \sigma_e^2 + \mu_e^2$$

where $\mu_e := E[\underline{e}]$
$\sigma_e := (\mathrm{Var}[\underline{e}])^{1/2}$

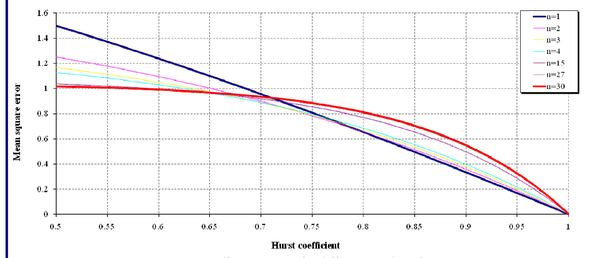
- Under the above assumptions, the MSE can be calculated as:

$$E\left[\underline{e}^2\right] = E\left[\left(\underline{x}_\cdot - \frac{\sum_{i=1}^{-n} \underline{x}_{t+i} + \sum_{i=1}^{n} \underline{x}_{t+i}}{2n}\right)^2\right]$$

$$E\left[\underline{e}^2\right] = \frac{1}{2}\left(\frac{\sigma}{n}\right)^2 \left[(2n+1)\left(n - 2\sum_{i=1}^{n} \rho_i\right) + \sum_{i=1}^{2n} (2n+1-i)\,\rho_i\right]$$

where n: number of neighbouring time steps

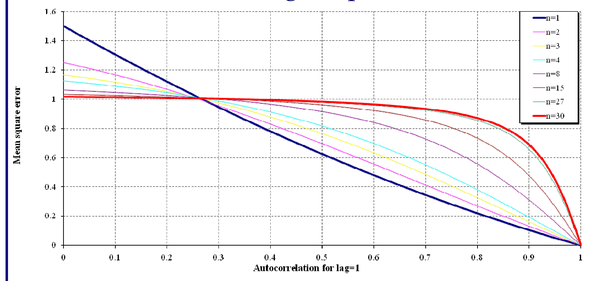
7. Results for long range dependence



MSE – Hurst coefficient curves for different number of time steps

- As Hurst coefficient increases, the optimal number of neighbouring time steps for the interpolation decreases (minimum MSE)
- A global average is preferable for low values of Hurst coefficient, while a small number of neighbouring time steps is required for high values of Hurst coefficient.
- In particular, for $H \geq 0.80$ the optimal n is 1.

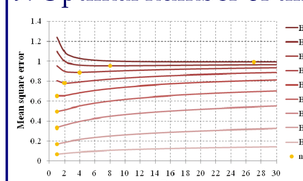
8. Results for short range dependence



MSE – Autocorrelation coefficient curves for different time steps

- In a Markovian process the number of the optimal neighbouring time steps (n) depends on a critical value of autocorrelation for lag-one ($\rho_{cr} \approx 0.24$).
- Thus, for $\rho < \rho_{cr}$ a global average is preferable, while for $\rho > \rho_{cr}$ only one step forward and backward is needed.

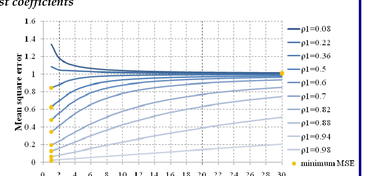
9. Optimal number of time steps



MSE – n curves for different Hurst coefficients

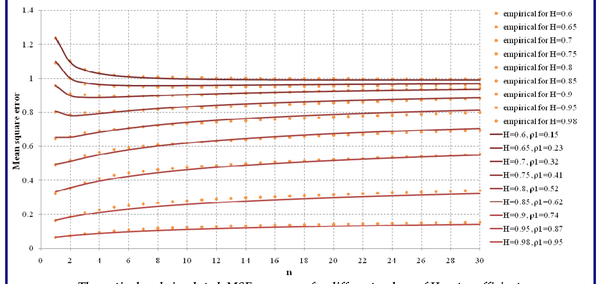
In long range dependence (on the left) for high values of the Hurst coefficient, the optimal number of time steps is limited to one, whereas low values require a global average.

In short range dependence (on the right) the global average is required for the interpolation in processes with low autocorrelation, whereas the number of optimal time steps reduces rapidly to 1 for high values of autocorrelation



MSE – n curves for different autocorrelation coefficients

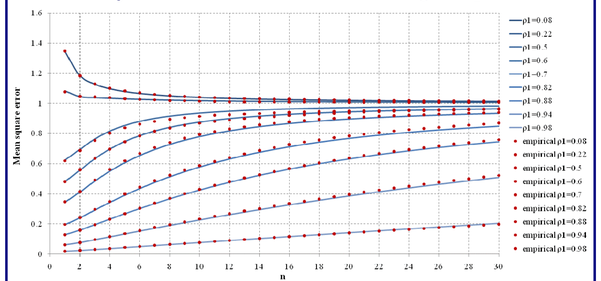
10. Validation by simulation (Long Range Dependence)



Theoretical and simulated MSE - n curves for different values of Hurst coefficient

As clearly shown, the estimation of the MSE from the simulation (FGN model) is nearly identical to the theoretical MSE.

11. Theory vs. Simulation (Short Range Dependence)



Theoretical and simulated MSE - n curves for different values of autocorrelation coefficient

Again, the estimation of the MSE from the simulation (AR(1) model) is nearly identical to the theoretical MSE.

12. Conclusions

- In time series with Hurst – Kolmogorov behaviour, a local average using an optimal number of neighbouring measurements is preferable over a global average on the infilling of a missing value. Depending on the Hurst coefficient, only a few forward and backward time steps are needed for infilling the missing value, which makes the process simpler. Particularly for an HK process with:
  - $H = 0.50\text{-}0.60$, the global average is preferable
  - $H = 0.70$, 4 time steps before and 4 after the interpolation time
  - $H = 0.72$, 3 time steps before and 3 after the interpolation time
  - $H = 0.74$, 2 time steps before and 2 after the interpolation time
  - $H \geq 0.80$, 1 time step before and 1 after the interpolation time
- In time series with Markovian behaviour the optimal number of neighbouring time steps depends on a critical value of autocorrelation. Specifically:
  - $\rho \leq 0.24$, the global average is preferable
  - $\rho \geq 0.24$, 1 time step before and 1 after the interpolation time
- The methodology is appropriate for quick infilling of very few missing values (not for long periods with missing values). The use of local average has an additional advantage, over the global average, that it does not reduce the variance of the time series.

REFERENCES
Koutsoyiannis, D., and A. Langousis, Precipitation, ch. 27 in *Treatise on Water Science*, Elsevier, 2010 (in press).
Koutsoyiannis, D., The Hurst phenomenon and fractional Gaussian noise made easy, *Hydrological Sciences Journal*, 47 (4), 573 – 795, 2002
Koutsoyiannis, D., Lecture notes on *Stochastic Methods in Water Resources*, Edition 3, 100 pages, National Technical University of Athens, 2007
Mandelbrot, B.B., Une classe de processus stochastiques homothetiques a soi : application a la loi climatologique de H. E. Hurst *C. R Acad. Sci. Paris* 260, 1965