XIX International Conference on Water Resources
CMWR 2012
University of Illinois at Urbana-Champaign
June 17-22, 2012

STATISTICAL PRE-PROCESSING AND RECONSTRUCTION METHODS FOR SUBSURFACE HYDRO-METEOROLOGICAL AND CRACK APERTURE TIME SERIES

D. Bailly[*], J-M. Matray[*] and R. Ababou[†]

[*] Institut de Radioprotection et de Sûreté Nucléaire,
Laboratoire de Recherche sur le Stockage géologique des déchets et les transferts dans les Sols
31 Avenue du Général Leclerc, 92260 Fontenay-aux-Roses, France
e-mail: david.bailly-cnrs@irsn.fr, jean-michel.matray@irsn.fr

[†] Institut de Mécanique des Fluides de Toulouse
1 Allée du professeur Camille Soula, 31400 Toulouse, France
e-mail: ababou@imft.fr (*corresponding author*)

Key words: Hydro-meteorological signals, crack aperture signal, statistical pre-processing, reconstruction, moving average, residual, auto-regressive, AR1 process, multimodal, URL.

Summary. This paper focuses on statistical methods for pre-processing, reconstructing and/or synthesizing hydro-meteorological time series and shrinkage crack aperture signals collected in the galleries of an Underground Research Laboratory in clay rock (Tournemire, France).

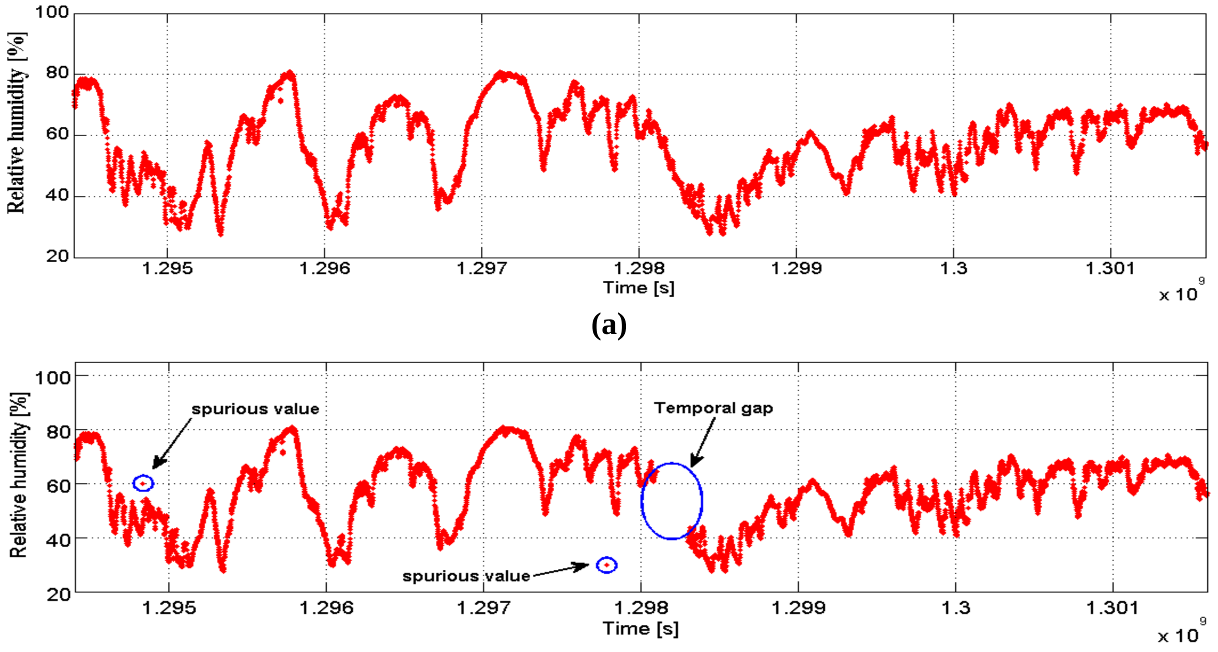
1 INTRODUCTION

We focus here on the effects of hydro-meteorological disturbances on the dynamics of shrinkage crack apertures measured in the walls of subsurface galleries in an URL (Underground Research Laboratory) in Tournemire, France. This URL is dedicated to the study of radioactive waste disposal in a clay rock geological repository. The present work is part of more extensive ongoing study involving statistical analyzes and cross-analyses of pore pressure signals, hydro-meteorological signals (air pressure, relative humidity, temperature), shrinkage crack dynamics, and water contents at various distances from the gallery walls. These analyses are being conducted in order to characterize the clay rock behavior hydraulically and mechanically.

However, the collected time series have data gaps, spurious values, and irregular time steps. In order to obtain the longest possible "clean" signals, the raw data need to be pre-processed. This implies homogenizing the time steps, detecting spurious values, and reconstructing at least some parts of the missing data on a regular time step grid. (Beyond pre-processing, we are also interested in complete synthesis of the signals based on their statistical properties).

In the present work, typically, the longest continuous sequences are on the order of half a year or a bit more (30 weeks), the reference time step is on the order of 15 mn, and the longest data gaps which have been reconstructed were on the order of one week, using a residual Auto-Regressive reconstruction method. In fact, sequences longer than one month have also been reconstructed or synthesized using a spectral method close to (but not identical to) the so-called Wold decomposition (see further below, Spectral Singular Harmonic identification and reconstruction method). The data pre-processing and synthesis algorithms are implemented by the authors in MATLAB (custom made toolbox MultiSAT: Multi Statistical Analysis Tool).

First, let us describe some of the pre-processing procedures. We use for illustration a signal of relative humidity, $H_R(t)$, collected at the Tournemire URL (**Figure 1.a)**. Two additional spurious values and a 58 h long temporal gap are inserted in the original signal for testing (**Figure 1.b)**.



(a)

(b)

Figure 1: **(a)** Signal of relative humidity Hr(t) with time step 1/4h, duration ~ 6 months (07 December 2010 - 22 June 2011); **(b)** modified time series, with 2 artificial spurious data and one artificial data gap (58 hours, 232 time steps).

Secondly, we will use a signal of shrinkage crack apertures collected at the Tournemire URL (**Figure 2**) to illustrate a multi-harmonic spectral method of signal synthesis for partial of complete reconstruction (Spectral Singular Harmonic identification).

2 PRE-PROCESSING STEPS

2.1 Detection of data gaps in the time series

The first step required for data pre-processing is to automatically identify data gaps using a time criterion (**eq. 1)**. If the time span between successive measurements is significantly greater than the proposed "reference time step" Δt_0, then a time gap is created (in some datasets, there will be markers for data gaps, and our algorithm can also use those markers directly). Thus:

$$\frac{t(i+1)-t(i)}{\Delta t_0} > s_{\Delta T} \tag{1}$$

where $s_{\Delta T}$ is the time step criterion used for time gap determination (generally on the order of unity); Δt_0 is the homogeneous time step chosen and used in the final homogenization procedure.

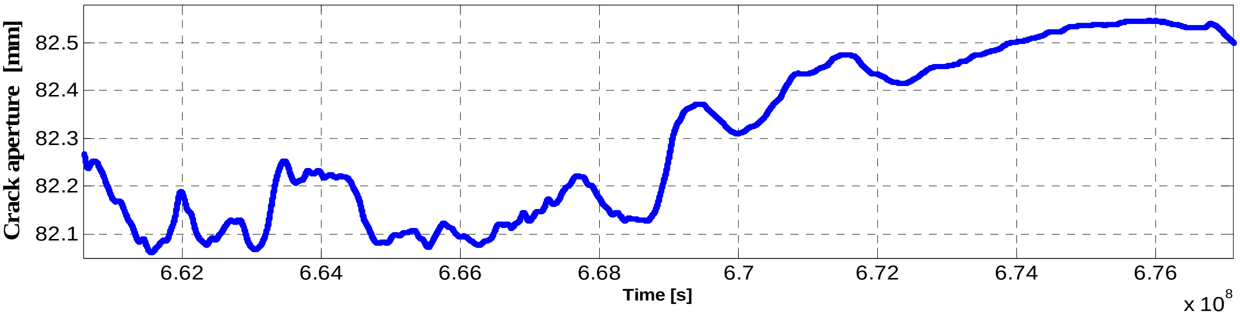


Figure 2: **(a)** Time series of crack aperture used to illustrate signal synthesis (Δt=1/4h; 7Dec2010-22June2011).

2.2 Preliminary reconstruction and detection of spurious values (outliers)

The second step is the detection of spurious values (**Figure 3**). The method is based on the following detection criterion (**eq. 2**).

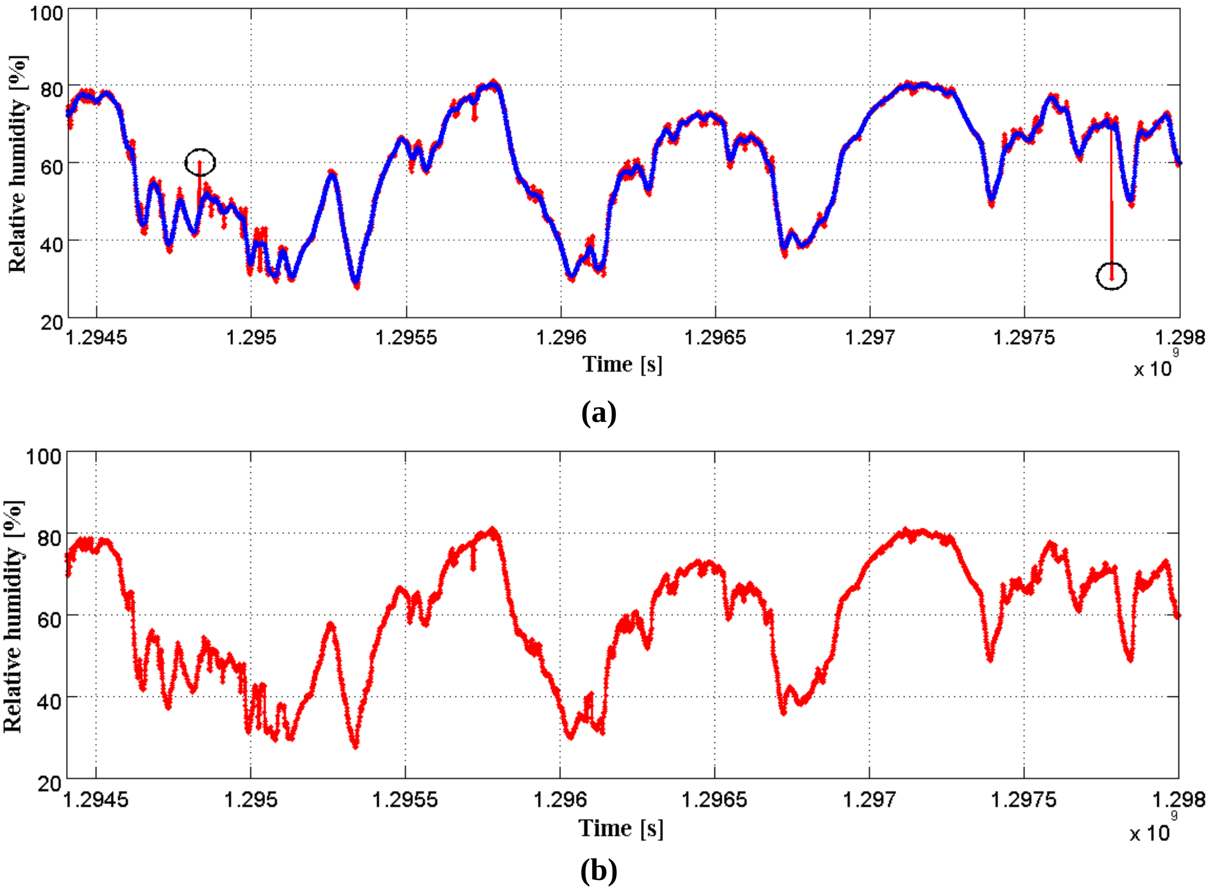
$$\frac{X(i)-\tilde{X}(i)}{2\tilde{\sigma}_{XLONG}} > S_{DETECT} \tag{2}$$

In **eq. 2**, $\tilde{X}(i)$ is the filtered signal obtained by moving average; $\tilde{\sigma}_{XLONG}$ is the standard deviation of the moving average for the longest time series available; S_{DETECT} is a dimensionless detection threshold determined by user. The moving average filtered signal $\tilde{X}(i)$ is defined as:

$$\tilde{X}(i) = \frac{\sum_{i-w}^{i+w} X(i)\,\Delta t(i)}{\sum_{i-w}^{i+w} \Delta t(i)} \quad \text{with } \Delta t(i) = \frac{t_i - t_{i-1}}{2} + \frac{t_{i+1} - t_i}{2} = \frac{t_{i+1} - t_{i-1}}{2} \tag{3}$$

where W is the discrete half-size of the moving average window, and *Δt(i)* is the variable time step centered on time instant *t(i)*.

To pre-compute the statistic $\tilde{\sigma}_{XLONG}$ in **eq. 2**, the user may decide to use the longest contiguous subsequence in the signal. The other option for computing $\tilde{\sigma}_{XLONG}$ is to construct the longest possible time series using a preliminary filling of data gaps. For this sake, a pre-reconstruction step is performed. The maximum length of data gaps to be treated in this pre-reconstruction step is determined by the user. We use linear interpolation for very small gaps (1 or 2 Δt_0), and a moving average for medium size gaps (3 to 10 Δt_0 at most). The method to compute the moving average in the presence of gaps is implemented in two steps: (i) linear interpolation between the left and right side of the gap interval (t_{LEFT}, t_{RIGHT}); (ii) computation of the moving average of the resulting signal locally around the linearly interpolated gap (the algorithm chooses automatically the relevant size of the moving window depending on gap length). **Figure 3** below illustrates the detection of two groups of spurious values.



(a)

(b)

Figure 3: **(a)** Time series of relative humidity $H_R(t)$ (**red curve**) and its moving average (**blue curve**), the latter being used for detection of spurious values (outliers). The normalized detection threshold S_{DETECT} is chosen equal to 0.7, and the half-size of the moving window is 24 Δt_0. **(b)**: Reconstructed sequence after detection of outliers.

2.3 Parametric reconstruction method for longer data gaps

For longer gaps, we have tested a parametric reconstruction method based on a stationary random process model, such as the AR1 model (Auto Regressive 1rst order process). More precisely, we propose to apply the AR1 model to the residual of X(t) obtained after filtering X(t) with a moving average. The steps of the AR1 reconstruction are the following:

1. Identification of the longest subsequence, and homogenization of the time steps of this subsequence using Δt_0 (see **eq. 10** further below).
2. Identification/calibration of parameters of the AR1 process for the residual of the longest subsequence (parameters: lag-1 correlation ρ_1, standard deviation σ_{Xlong}) (see **eqs.4-5**).
3. Preliminary reconstruction of the longest gaps in the original signal (by linear interpolation, using Δt_0).
4. Identification of the left and right values $X_{RES}(t_{LEFT})$ and $X_{RES}(t_{RIGHT})$ for each gap, by computing the residual of the preliminary reconstructed signal (see **eqs.5-6**).
5. Preliminary reconstruction of data gaps by applying a moving average filter (**eq. 3**), with a moving window automatically adjustable to the gap length, and using Δt_0.
6. Finally, the two calibrated AR1 processes (Left-Right "LR", and Right-Left "RL") are averaged arithmetically, and added to the moving average using Δt_0 (see **eq.6**).

$$X_{RES}^{\mathbf{LR}}(0) = X_{RES}(t_{LEFT})\,;\, X_{RES}^{\mathbf{LR}}(i+1) = \rho_1 X_{RES}(i) + \left[\sqrt{(1-\rho_1^2)}\;\sigma_{Xlong}\right]\varepsilon(i+1) \tag{4}$$

$$X_{RES}^{\mathbf{LR}}(N_{GAP}+1) = X_{RES}(t_{RIGHT})\,;\, X_{RES}^{\mathbf{RL}}(i) = \rho_1 X_{RES}(i+1) + \left[\sqrt{(1-\rho_1^2)}\;\sigma_{XLONG}\right]\varepsilon(i) \tag{5}$$

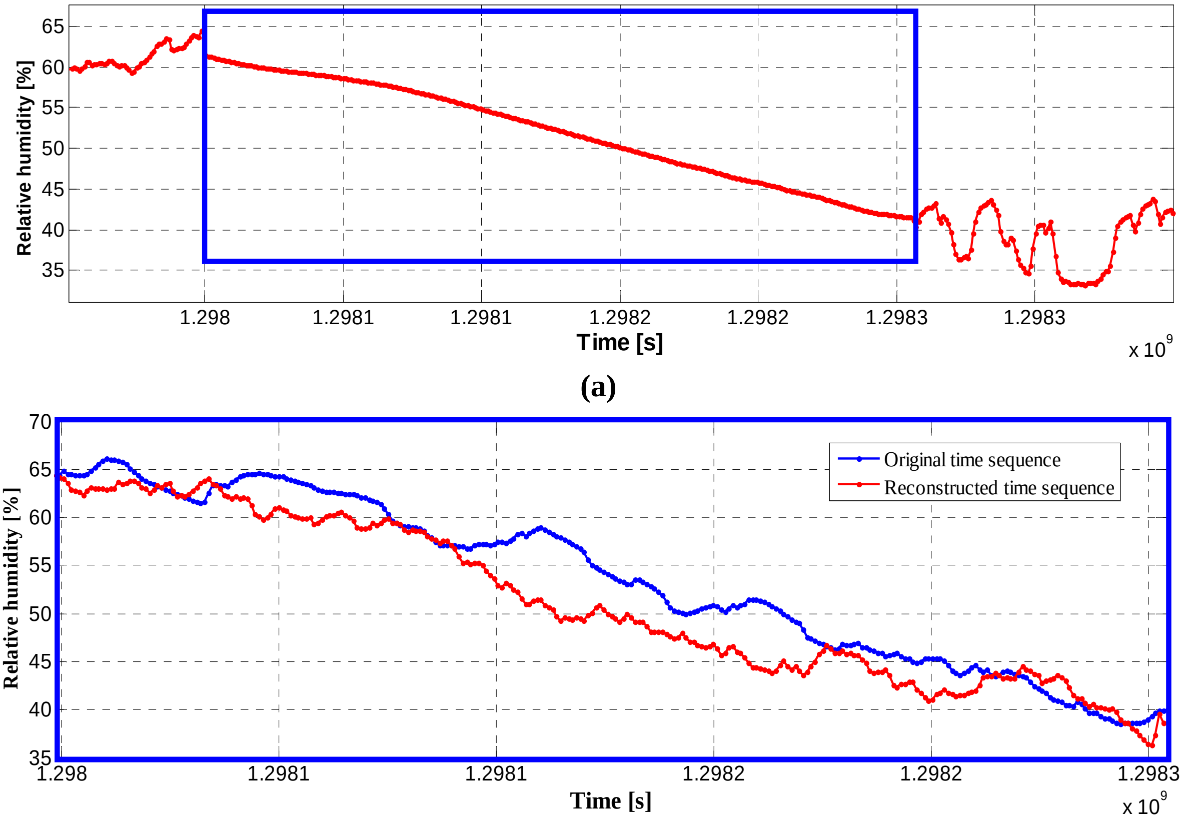
$$X_{RES}(i) = \frac{X_{RES}^{\mathbf{LR}}(i) + X_{RES}^{\mathbf{RL}}(i)}{2} \rightarrow X_{RECONSTRUCTED}(i) = \tilde{X}(i) + X_{RES}(i) \tag{6}$$

In these equations, ρ_1 is the lag-1 auto-correlation and σ_{XLONG} is the standard deviation of the longest subsequence (residual); $\varepsilon(i)$ is a purely random sequence (uncorrelated, N(0,1)).

Residual signal (using a moving average with W = 24 hours)		
	Standard deviation	Auto-correlation ρ_1
Longest subsequence	2.94	0.9860
Original signal	2.45E-1	0.9912
Reconstructed signal	3.55E-1	0.9805
Raw signals		
	Standard deviation	Auto-correlation ρ_1
Longest subsequence	14.35	0.9989
Original signal	8.54	0.9923
Reconstructed signal	8.37	0.9905

Table 1: Statistics of various signals (pre-processing / bidirectional AR1 method): longest subsequence; original subsequence; and reconstructed subsequence with the.

Figure 4 shows the application of this algorithm to the relative humidity signal X(t) of **Figure 1**. In this case, AR1 parameters are identified on the longest subsequence using a moving average with a half-size window equal to 12 hours (same window used also for the residual of the complete signal). The results are satisfactory, albeit not perfect. The details of the reconstructed signal are not the same as the true original signal, but this is to be expected since we use a statistical method (with random numbers). However, some statistics of the true signal have been preserved, as can be seen in **Tab.1**.



(a)

(b)

Figure 4: Reconstruction of X(t) using the AR1 process applied to the moving average residual of X(t). The top graph **(a)** shows the result of step 5 (moving average reconstruction), and the bottom graph **(b)** shows the final result, i.e., the true original signal (**blue curve**) compared to the bidirectional AR1 reconstruction (**red curve**). In both graphs, the **blue** rectangular frame indicates the part of the signal that is being reconstructed.

We have also compared other statistics of the signals (original *vs.* reconstructed), such as their Root Mean Square (RMS) spectrum, estimated by the Wiener-Khinchine method as the Fourier transform of the covariance function - with a Tuckey filter "D" inserted in the transform ([1]):

$$S_{XX}^{RMS}(f_i) = 2\sqrt{S_{XX}(f_i)\ \Delta f} \tag{7}$$

$$S_{XX}(f_i) = 2\Delta t\,\sigma_X^2\left[R_{XX}(0) + 2\sum_{j=1}^{M} D_j . R_{XX}(j).\cos\left(2\pi\, j\Delta t\, f_i\right)\right] \tag{8}$$

$$\tau_j = j\Delta t;\ R_{XX}(\tau_j) \equiv R_{XX}(j) = \frac{1}{\sigma_X^2}\frac{1}{N-1}\sum_{n=1}^{n=N-j}\left(X(t_n)-\overline{X}\right)\left(X(t_{n+j})-\overline{X}\right) \qquad (9)$$

Here, S_{XX}^{RMS} is the Root Mean Square spectrum, S_{XX} is the dimensional spectrum of X(t), R_{XX} is the unbiased estimate of the auto-correlation function (*vs.* discrete time lag "j"), σ_X is the standard deviation of X(t), $\overline{X}$ is the mean of X(t), D_j is the Tukey filter (*vs.* time lag "j"), N the number of data, M the number of discrete time lags, and f_i the discrete dimensional frequencies.

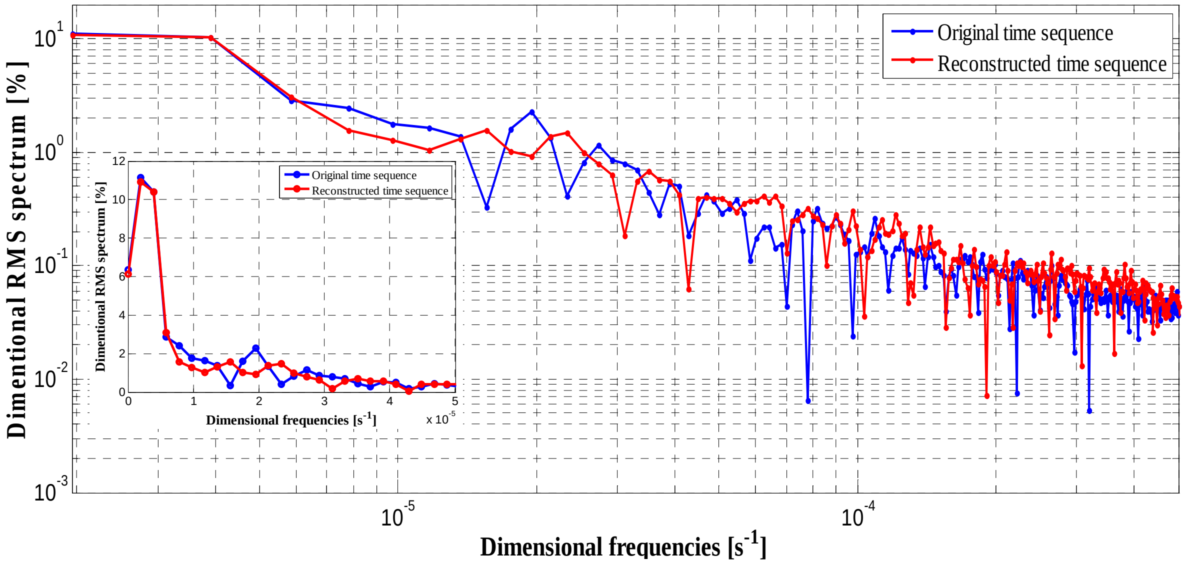


Figure 5: Comparison between the RMS spectrum of the original subsequence (**blue curve**) and of the reconstructed subsequence (**red curve**). The frequency spectra are plotted in log-log scales (log(S) *vs.* log(f)).

Figure 5 compares the spectra of raw *vs.* reconstructed signals. There are some differences. Low frequencies are correctly reproduced, thanks to the moving average pre-reconstruction (step 5). The decreasing structure of the spectrum at high frequencies is also fairly reproduced (except at some isolated frequencies), thanks to the AR1 reconstruction method (step 6). However, the original signal has a few intermediate dominant frequencies, e.g. at f ≈ 2E-5 Hz (close to semi-diurnal), which are not "seen" in the reconstructed signal. The method to be presented in section 3 will address directly the case of isolated dominant harmonics.

2.4 Time step homogenization

In order to implement statistical analyses on the signal(s), such as spectral and cross-spectral analysis, or other decomposition methods like multi-resolution wavelet analyzis ([1]), the time step must be constant. The transformation from a variable time step grid to a constant time step grid is called "time step homogenization". It is performed by a linear interpolation algorithm:

$$X_{HOM}(t_j) = X(t_j) + \rho(t_j)\left\lfloor X(t_j+1) - X(t_j)\right\rfloor \text{ with } \rho(t_j) = \frac{\left(j\,\Delta t_0 + t_{INI}\right) - t(j)}{t_{j+1} - t_j} \qquad (10)$$

where Δt_0 is the chosen reference time step, and t_{INI} is the initial time of the signal. An example of time step homogenization is shown in **Figure 6**.

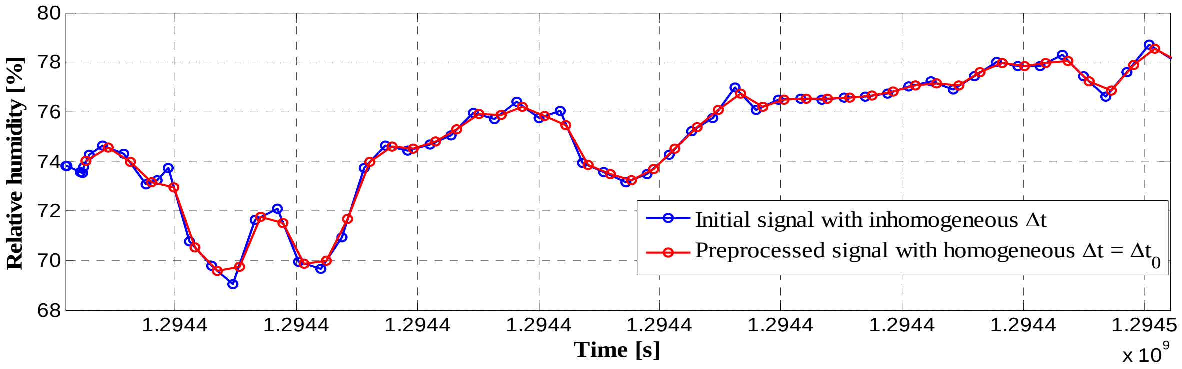


Figure 6: Comparison between the original relative humidity time series with inhomogeneous time steps (**blue curve**) and the final pre-processed relative humidity time series after time step homogenization (**red curve**).

3 FROM RECONSTRUCTION TO SYNTHESIS (MULTI-HARMONIC APPROACH)

Another example of parametric reconstruction or synthesis of a signal is shown below for the case of multi-modal signals, where typically a few discrete (singular) frequencies dominate the spectrum. Our approach is similar to the Wold decomposition ([3]), although we consider here the singular spectrum as deterministic rather than random. Thus, we model the signal as follows:

$$X_{SYNTH}(t) = A_0 + \sum_{i=1}^{H} A_i \cos(\omega_i t + \varphi_i) \text{ where } A_0 = \overline{X} \qquad (11)$$

In **eq.11**, H is the number of singular harmonics, ω_i is the set of singular frequencies, A_i and φ_i are respectively the amplitude and phase associated to harmonic (i), and $\overline{X}$ is the mean of X(t).

For illustration, we analyze a shrinkage crack displacement signal [mm] measured at the Tournemire URL in gallery Ga2003 over a period of 6 months (**Figure 2**). We assume that eleven discrete harmonics (H=11) are sufficient for reproducing the signal accurately. The discrete frequencies (ω_i) and their amplitudes (A_i) are determined from the estimated RMS spectrum. The phases (φ_i) associated with these harmonics are obtained using an optimization procedure. Two kinds of optimization procedures can be applied to identify the phases in **eq.11**:

1. Using a least square optimization procedure on $X(t)\text{-}X_{SYNTH}(t;\underline{\varphi})$ to obtain the best fitted (φ_i)'s from **eq.11**, the other parameters (A_i) being already known.
2. Determining the phases φ_i from the cross-spectrum (phase spectrum) of $(X(t), X_{SYNTH}(t;\underline{0}))$, where X(t) is the original signal, and $X_{SYNTH}(t;\underline{0})$ is the synthetic signal of **eq.11** with zero phases ($\varphi_i = 0$).

The synthetic signal is compared to the original signal in **Figure 7** (crack aperture signal, duration ~6 months, time step Δt=1/4h). As can be seen, the two signals agree fairly well[1]. One advantage of this method is that it is possible to choose predetermined harmonics corresponding to natural hydro-meteorological or geophysical periods (..., 12h, 24h, 6 months, 1 year, ...).

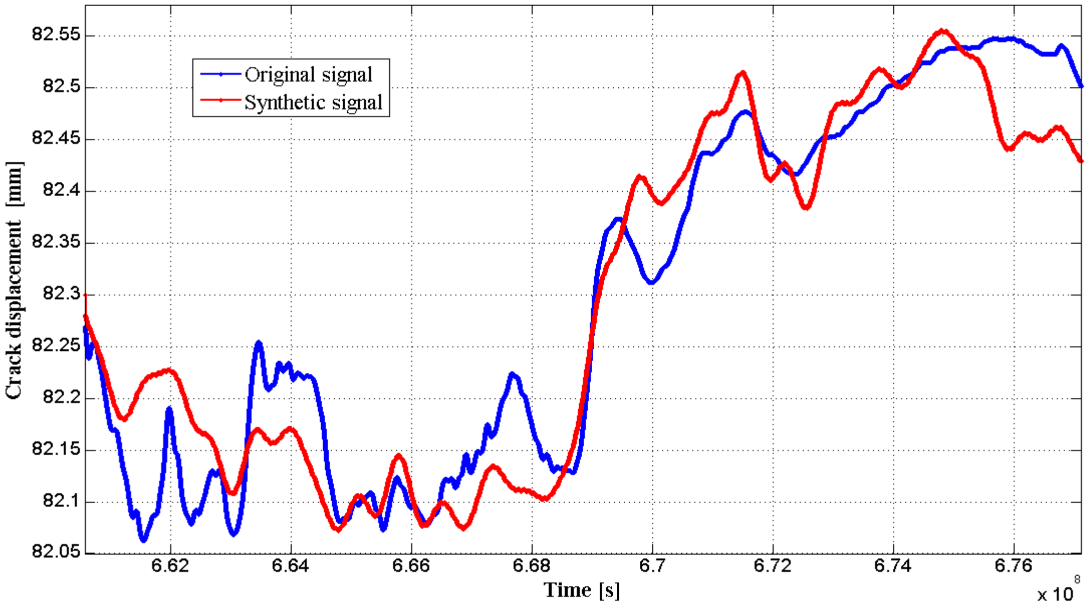


Figure 7: Crack aperture signal (aperture variation or displacement vs. time, in *mm*). Comparison between the original signal (**blue curve**) and the synthetic crack aperture signal (**red curve**).

Finally, **Figure 8** shows the importance of optimizing the phases. It compares the optimized synthetic signal with the zero phase synthetic signal (the former is better than the latter when compared to the true signal).

4 CONCLUSIONS AND OUTLOOK

We are currently working towards a more extensive set of statistical processing and analyses of geophysical, hydrogeological and hydro-meteorological signals, with the aim of characterizing the isolation properties and the hydromechanical behavior of a potential clay rock geologic repository for radioactive waste. Beyond pre-processing (e.g. reconstitution of data gaps), we are also interested in performing a complete joint analyzis and synthesis of the signals based on their statistical properties. Ongoing work focuses on the following techniques:

- Replacing moving average filtering with wavelet multi-resolution analyzis (dyadic decomposition): e.g., wavelet approximation at a chosen dyadic scale serves as a filter.
- Using statistical envelopes to characterize the evolution of signals in the presence of trends

[1] Here, the method used is that of least square optimization of the phases (the other method based on phase cross-spectrum works well too, provided the signals are somewhat filtered, e.g. using wavelet approximations).

(including the effect of evolving excavation fronts / "mine by" tests).

- Developing joint or cross-analyzis and cross-generation of the signals (X(t), Y(t), …) by a spectral Fourier/Wiener-Khinchine synthesis method (similar to Rice theory). This is a non parametric method. It uses the estimated continuous spectra and cross-spectra assuming jointly stationary processes. It can be combined with the parametric multi-harmonic approach ("singular harmonics").

Ongoing work focuses on cross-amplitude, spectral gain, and phase spectrum analyses at various frequencies, e.g., relative humidity *vs.* crack displacement signals, and also, clay water content signals at different distances from the gallery wall.

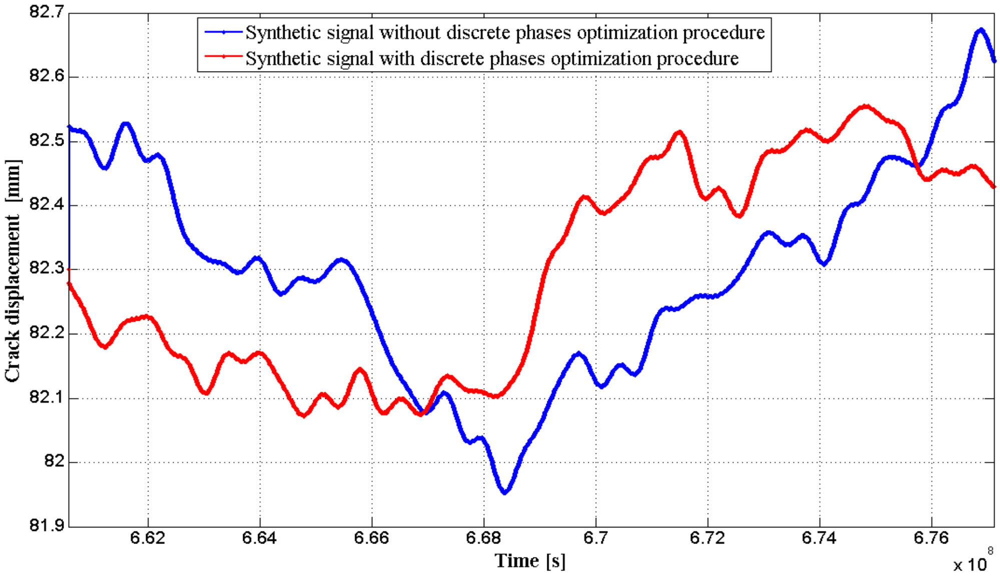


Figure 8: Comparison between two synthetic signals obtained with 11 harmonics (H=11). Non optimized synthetic signal $X_{SYNTH}(t;\underline{0})$ with zero phases (φi = 0): **blue curve**. Optimized synthetic signal (best fitted φi's): **red curve**.

REFERENCES

[1] H. Fatmi, R. Ababou, and J.-M. Matray, “Statistical pre-processing and analyses of hydrogeo-meteorological time series in a geologic clay site (methodology and first results for Mont Terri’s PP experiment)”. *Journal of Physics & Chemistry of the Earth (JPCE),* Special Issue «Clays in Natural & Engineered Barriers for Radioactive Waste Confinement» (CLAY'2007), 33:S14-S23 (2008).

[2] H. Fatmi, R. Ababou, J.-M. Matray, and Ch. Nussbaum, "Statistical analyses of pressure signals, hydrogeologic characterization and evolution of Excavation Damaged Zone (claystone sites of Mont Terri and Tournemire)." *Proceedings MAMERN11*: *4th Internat. Conf. Approx. Methods & Numer. Modelling in Envir. & Natural Resources*, Saidia (Morocco), May 23-26, 2011. B. Amaziane, D. Barrera, H. Mraoui, M.L. Rodriguez & D. Sbibih (eds.), Univ. de Granada, ISBN:078-84-338-5230-4, pp. 325-329 (2011).

[3] M.B. Priestley, "*Spectral analysis and time series (Vol.1: Univariate Series; Vol.2: Multivariate Series, Prediction and Control)"*. Elsevier / Academic Press, 890 pp. (1981).