Unknown-length motif discovery methods in environmental monitoring time series

Lisa Poirier-Herbeck, Elisabeth Lahalle, Nicolas Saurel, Sylvie Marcos

Unknown-length motif discovery methods in environmental monitoring time series

Lisa Poirier--Herbeck
CEA Valduc
Is-sur-Tille, France
CNRS, Laboratoire des signaux et systèmes
Université Paris-Saclay, CentraleSupélec
Gif-sur-Yvette, France
lisa.poirier-herbeck@cea.fr

Elisabeth Lahalle
CNRS, Laboratoire des signaux et systèmes
Université Paris-Saclay, CentraleSupélec
Gif-sur-Yvette, France
elisabeth.lahalle@centralesupelec.fr

Nicolas Saurel
CEA Valduc
Is-sur-Tille, France
nicolas.saurel@cea.fr

Sylvie Marcos
CNRS, Laboratoire des signaux et systèmes
Université Paris-Saclay, CentraleSupélec
Gif-sur-Yvette, France
sylvie.marcos@centralesupelec.fr

***Abstract*—The search for information of interest in massive time series is crucial in many industrial applications. Companies need their data to be analyzed or modeled in real time, which often requires to extract some patterns, also referred as motifs. However, for diverse and ever more signals, human expertise is overwhelmed by time and by huge amount of data. It is the case for environmental monitoring where it is question to detect radiological phenomena from environmental signals. In this paper, we propose an unsupervised and unknown length motif discovery method based on the Matrix Profile with a low computational cost. Its performance is evaluated on a dataset of simulated radiological signals dedicated to environmental monitoring, and compared to a similarity DTW based method and to a classical standard deviation based method. The advantages and drawbacks of each method are highlighted in terms of performance, runtime, accuracy and robustness to different types of noisy signals.**

***Index Terms*—pattern discovery, time series, data mining, motif discovery, unsupervised, variable-length, real time, matrix profile, DTW, Crossmatch**

I. Introduction

Environmental monitoring signals come from the recording of both background noise and patterns, also called motifs, corresponding to some phenomena [1]. Those phenomena correspond to natural and human-made environmental events. For several years, motif discovery has become an important topic in data mining [2]. It allows to efficiently discover phenomena by extracting motifs of interest from the background of the signal and measure their similarities. The main problems of motif discovery algorithms are the difficulty to detect motifs with unknown lengths, the large processing time, discovering ill-known (warped) motifs, and online processing [3].

Among recent methods of the state-of-art [4], approaches based on discretization [5], autoencoder [6], Dynamic Time Warping (DTW) [7] and Matrix Profile (MP) [8] have been studied. Following on from results presented in the comparative study in [4], approaches based on discretization and autoencoder have not been selected here due to their low performance compared to DTW [7] based approach. Moreover, the MP has shown its flexibility to many problems in data mining [8], [9], which makes it interesting for our problem.

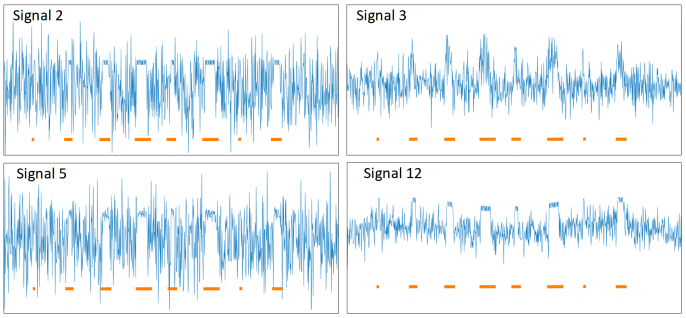


Fig. 1: Time series and motif locations (orange).

In this paper, we propose two unsupervised methods to discover unknown-length motifs with low time complexity for the purpose of being applicable in real time.

The first method we propose is an improvement of the Pan Matrix Profile (PMP), introduced by Madrid et al. [9]. The principle of the Pan Matrix Profile is to compute a Matrix Profile (MP) [8] for all lengths of potential patterns, to discover motifs of different lengths and their starting indices. This leads to very high computational cost. Unlike PMP our method needs only one computation of the MP to directly get the length and the motifs locations of all patterns. Our proposed method needs only one computation of the MP, which greatly reduces the time complexity.

The second method we propose is an adaptation of the Crossmatch algorithm [10] to discover unknown length motifs

with low complexity. The latter computes a similarity matrix equivalent to the DTW distance, for two evolving time series. Noering et al. [4] adapted the similarity matrix of the Crossmatch algorithm and combined it with the motif extraction and clustering method of [11], to discover motifs in one signal. Our modification consists in thresholding the diagonal of the similarity matrix. This new version greatly reduces the time complexity.

Then, we compare the proposed methods based on the MP and on the Crossmatch to a classical approach based on sliding standard deviation thresholding.

As our motivation is to discover motifs online in radiological signals of environmental monitoring, we evaluate our methods on signals constructed with different background noises and motifs simulated by an environmental software, owned by CEA. Discovery methods are evaluated with the ground truth of positions of motifs. Some of those signals are illustrated in Fig. 1.

In Section II, we present our proposed methods. In Section III, we evaluate their performance on synthetic signals which model real environmental signals. Then, we conclude in Section IV.

II. Methodology

A. Proposed method based on Matrix Profile

Matrix Profile (MP) m_i is a sliding Euclidean distance as:

$$m_i = \min_{j>i+L \text{ or } i<i-L} d(x_{i:i+L-1}, x_{j:j+L-1}), \qquad (1)$$

with d the Euclidean distance, $x = \{x_i\}$ the input time series and L the length of the subsequence $x_{i:i+L-1}$. Its minimum values identify diverse repetitive and unique patterns in the signal. Extreme values are zero at positions of a motif if the length of the searched motif corresponds to the subsequence length L (Fig. 2a). However, for noisy signals (Fig. 2b), due to its implementation with the MSTAMP algorithm [8], [12], the MP behaves differently: it creates local maxima at the position of the motif instead of a local minima.

The Pan Matrix Profile (PMP) [9] corresponds to the MP calculated with the MSTAMP algorithm for different L values of subsequence lengths. According to [13], L must be between 4 and a maximum length L_{max}. In Fig. 3b, we represent Matrix Profiles of the signal (Fig. 3a) for some values around the local maxima (light blue line). In Fig. 3c, we represent the PMP image for all values of L such as $4 \leq L \leq L_{max}$ = 40. The abscisses and ordonates correspond to the time index i and the length L respectively. It can be noticed in Fig. 3c, that motifs of 1 point length are represented by a single triangle, whereas longer motifs are represented by two same size triangles that intersect. All triangles have the same size and the coordinates α of the hollow between two intersecting triangles exactly correspond to the starting index i_s and the length L_M of the corresponding motif (see notations in Fig. 3c). This hollow corresponds to the local maxima of the MP when the subsequence length is equal to motifs length see Fig. 2b and 3c). Based on these observations, and to reduce

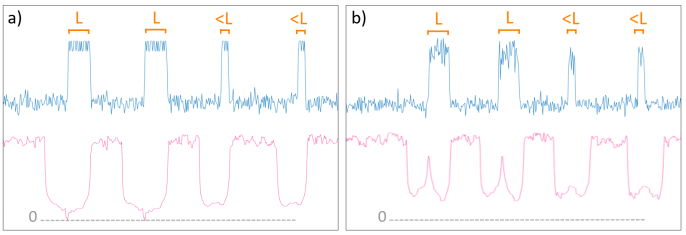


Fig. 2: a) Non-noisy and b) noisy signals (blue) and their Matrix Profiles (pink) obtained with a subsequence length L corresponding to some motif lengths (orange).

the complexity, we propose to compute a unique MP (uMP) for a single value of subsequence length denoted L_{max}, which must be chosen larger than the longest expected pattern. With Thales's theorem, we establish the following formula:

$$\frac{l_1}{l_2} = \frac{\epsilon_1}{\epsilon_2}, \quad with: \quad \epsilon_1 = l_1 \times 2 - (p_2 - p_1), \quad l_2 = L_{max} - 4. \qquad (2)$$

The starting index of the motif is $i_s = p_1 + l_1$ and the ending index is $i_e = i_s + L_M$ with $L_M = L_{max} - \epsilon_2$. To resume, the algorithm consists in obtaining p_1 and p_2 by thresholding the average of MP calculated for L_{max}, to deduce $\alpha(i_s, L_M)$. Moreover, since the MP is sensible to noise as highlighted in Fig. 2, we also test our proposed uMP method with a smoothing applied to the input signal as pre-processing.

B. Proposed method based on Crossmatch

Crossmatch [10], [14] is based on the computation of a similarity matrix v. We here use the same formula than

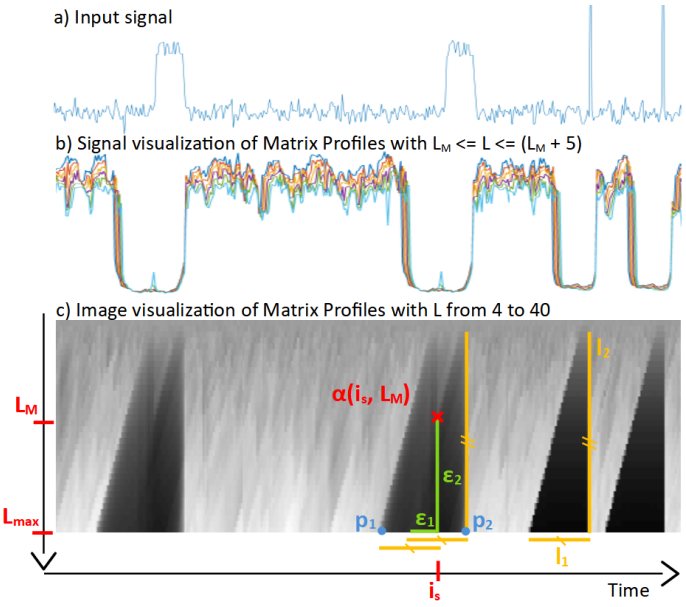


Fig. 3: a) input signal and its b) signal and c) image visualizations of Matrix Profile for different subsequence lengths L

Noering et al. [4] to get the similarity matrix v. Each value of $v(i,j)$ depends on the previous values of the matrix v:

$$v(i,j) = max \begin{cases} \epsilon_{bd} - ||x_i - y_j|| + v(i-1, j-1) \\ \epsilon_{bv} - ||x_i - y_j|| + v(i, j-1) \\ \epsilon_{bh} - ||x_i - y_j|| + v(i-1, j) \\ 0 \end{cases} \quad (3)$$
$$v(i,j) = 0 \ \text{ if } \ |i-j| \leq \text{offset},$$

where x_i and y_j represent the sample points of time series x and y. We compute the similarity matrix v for $x = y$, and we use it to extract motifs. We set $\epsilon_{bd} = 0.03$ and $\epsilon_{bv} = \epsilon_{bh} = -0.01$ as time scaling penalization, i.e. horizontal or vertical steps in v. Find more details about the Crossmatch algorithm in [10]. Noering et al. use an offset region around the diagonal of the similarity matrix (Fig. 4a) to eliminate trivial matches. Focusing on trivial matches of the diagonal, we propose to fix the offset to the minimal value 1 and, we extract a diagonal of empirical width of 10 pixels (Fig. 4b). Consequently, all other elements of our modified similarity matrix are zero. Then, to get the intensity signal of the diagonal, the elements of each column of the matrix are summed. The intensity signal is thresholded using its mean (Fig. 4c) to detect location of motifs and extract them from the background signal. We name our Crossmatch adaptation the dCrossmatch, in reference to the use of the diagonal.

C. Method based on moving standard deviation

With the aim of discovering motifs of interest in time series, we compare our proposed methods to a classical approach based on the moving standard deviation. The input parameter l_w is the window length for the moving standard deviation (STD). Since signal's background noise follows a normal distribution, we apply the following threshold th to the signal: $th = \mu_m + 3\sigma_m$, with σ_m the moving STD and μ_m the signal's moving mean. Signal's moving mean is computed with n times l_w as window length. To obtain a large enough time horizon according to time series' length, an empirical value of $n = 20$ is chosen. The method is illustrated in Fig. 5.

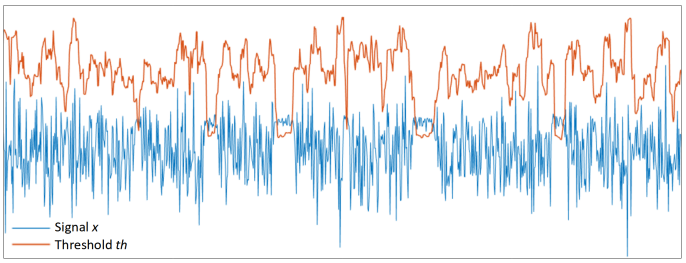


Fig. 5: Moving standard deviation thresholding.

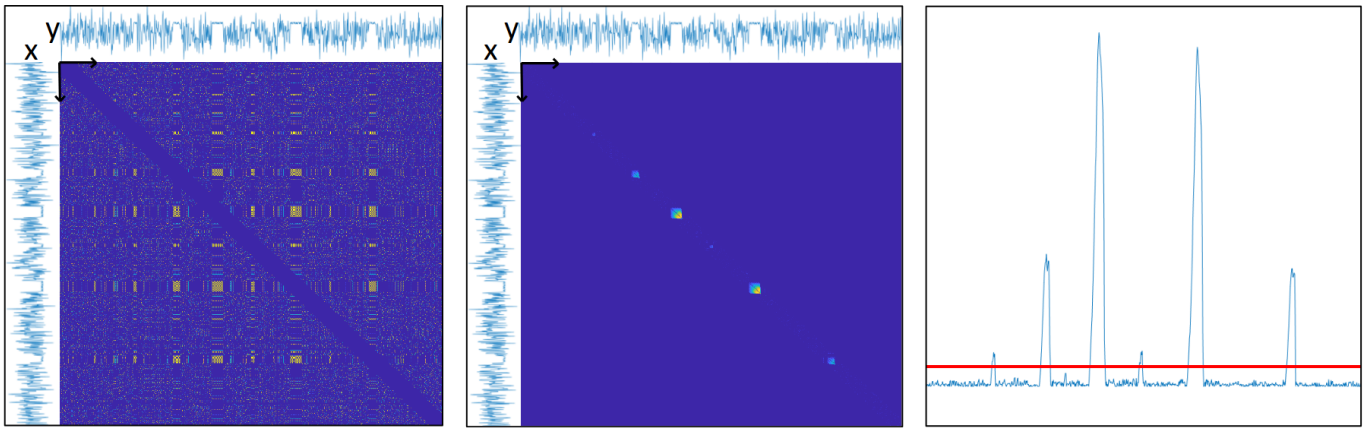
III. Experiments

A. Signals

To evaluate performance of the methods proposed in this paper, we create synthetic time series made of different sizes and composed of normally distributed background noise and motifs of different sizes, obtained by a radiological signal simulation software owned by CEA. Synthetic signals are constructed according to our observations on the real environmental signals: different noise levels for motifs and backgrounds are considered. For some signals, the amplitudes of motifs vary and we make slightly variations on the background. In the end, the dataset contains different cases of signals. For each case, we make 100 signals with different noise realizations. Fig. 1 shows few examples of our synthetic time series. Table I gives signal to background noise ratio (SNR_{BKG}) and signal to motif noise ratio ($\text{SNR}_{\text{motifs}}$), which we formulate as:

$$\text{SNR}_{\text{BKG}} = \frac{P_{\text{motifs}}}{\sigma^2_{\text{BKG}}}, \qquad \text{SNR}_{\text{motifs}} = \frac{P_{\text{motifs}}}{\sigma^2_{\text{motifs}}}, \quad (4)$$

where P_{motifs} is the mean power of motif signals, and σ^2_{BKG} and σ^2_{motifs} are the noise variances of the background and motifs, respectively.



a) Noering et al. adaptation of similarity matrix *v* b) Similarity matrix v, with the extracted diagonal c) Intensity signal of the diagonal and its threshold

Fig. 4: Steps of our Crossmatch adaptation, dCrossmatch, a) is the similarity matrix of Noering et al. [4], b) our proposed similarity matrix with the extracted diagonal, and c) the intensity signal of the diagonal and it threshold.

B. Evaluation metrics

The outputs of the algorithms are the indices of starting and ending points of discovered motifs. Proposed evaluation metrics are the F1-score (F_1), the True Positive Rate (TPR) and the False Discovery Rate (FDR) to evaluate the performance of motif detection. The index position errors (err_i) allows to evaluate the performance of motif locations. We define a found motif as True Positive with a tolerance of 10 points on the ground truth indices, and then we quantify the motif location error with e, L1-norm of index position errors divided by the number of True Positive since we only calculate errors for the True Positive.

$$\begin{aligned} F_1 &= 2 \times \tfrac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, & \text{FDR} &= \tfrac{\text{FP}}{\text{FP}+\text{TP}}, \\ L_1 &= \sum_i |\text{err}_i|, & e &= \frac{L_1}{\text{TP}}, \\ \text{precision} &= \tfrac{\text{TP}}{\text{TP}+\text{FP}}, & \text{recall} = \text{TPR} &= \tfrac{\text{TP}}{\text{TP}+\text{FN}}, \end{aligned} \tag{5}$$

where TP is the number of True Positive, FP the number of False Positive and FN the number of False Negative.

C. Results

To compare the performance of the methods, Fig. 6 exhibits the different metrics. The left sub-figure illustrates the couple of the F1-score and the error e which measures both motif detection and location performance, whereas the right sub-figure shows the couple of FDR and TPR metrics to quantify the motif detection performance. These metrics are calculated on the signals represented by numbers, as illustrated in table I. A good method has its points located in the upper-left corner of both graphs.

Scores and errors illustrated in Fig. 6 exhibit that the moving STD and the proposed uMP method with smoothed input signal are, without contest, the most efficient methods for our dataset. The moving STD gives very good results in both detection and location metrics and for all signals, except for the signal case n°3. Despite this, this method is remarkably stable.

Signal case n°3 is the most difficult one since motifs are very noisy and close to the background (see Table I) then, it provides lower detection precision (F_1) and higher location error (e) than other signals. All methods give limited or poor results for it, except the uMP with smoothed input signal (Fig. 7). This method is more efficient than the moving STD one, except for signal cases n°2, 5, 7 and 11, which contain motifs drowned in the background (low SNR_{BKG}). However, these signals have slightly or non-noisy motifs, which might not correspond to real environmental radiological time series. On the other hand, this method is a bit less accurate than the moving STD one, referring to L1-norm position errors. Besides, the raw uMP method is, in any case, less efficient than the uMP with smoothed input signal.

The dCrossmatch method gives particularly low scores for signal cases n°3, 4, 6 and 10. The common characteristic of these signals is that their motifs and their background are very noisy (see Table I).

Concerning the processing time, for an input signal of length 1000, the uMP, the dCrossmatch and the moving STD approaches have a runtime of about 10^{-1}s, 10^{-2}s and 10^{-3}s respectively (Table II). We use a computer with processor Intel(R) Core(TM) i7-10875H CPU 2.30 GHz, with a RAM memory of 64.0 Go.

TABLE II: TIME COMPUTATION OF EVALUATED METHODS FOR AN INPUT SIGNAL OF LENGTH 1000.

Methods	**uMP**	**dCrossmatch**	**moving STD**
Runtime (s)	10^{-1}	10^{-2}	10^{-3}

D. Discussion

The dCrossmatch has an important dispersion of scores. This is caused by noisy signal cases which provide bad results. Smoothing the noisy signal is not possible since it impacts the similarity measured by the dCrossmatch.

The moving STD gives the best and stable results, indeed, dispersion is very low compared to other methods. It is also the fastest method. Its only drawback is that it struggles dealing with signal cases where motifs are very close to the background and have the same power of noise.

To make up for it, the uMP with smoothed signal overcomes this limitation which makes it also very useful to explore our real environmental signals. In addition, for some signal cases, it gives even slightly better results than the moving STD one.

To benefit of advantages of both the moving STD and the uMP methods for different signal cases, we could use the two methods in parallel. Indeed, since they have different drawbacks depending on signal characteristics, the compensation of both should allow an enough complete motif discovery procedure.

IV. CONCLUSION

In this paper, we proposed methods which allow to tackle the difficult problems of discovering motifs of variable and unknown lengths. These methods have a low runtime, offering the possibility of online processing. We proposed an adaptation of the Matrix Profile (uMP) for which we compared performance with a DTW based Crossmatch method (dCrossmatch) and a classical moving standard deviation thresholding method. The classical thresholding and the uMP methods

TABLE I: SIGNAL TO BACKGROUND NOISE RATIO SNR_{BKG} AND SIGNAL TO MOTIF NOISE RATIO $\text{SNR}_{\text{MOTIFS}}$.

Signals	1	2	3	4	5	6	7	8	9	10	11	12
SNR_{BKG} (dB)	8.2	0.24	7.19	14.37	0.28	8.25	0.24	8.2	8.2	15.18	1.08	8.2
$\text{SNR}_{\text{motifs}}$ (dB)	no noise	no noise	7.19	14.37	20.27	20.28	34.22	34.23	no noise	15.18	21.09	no noise

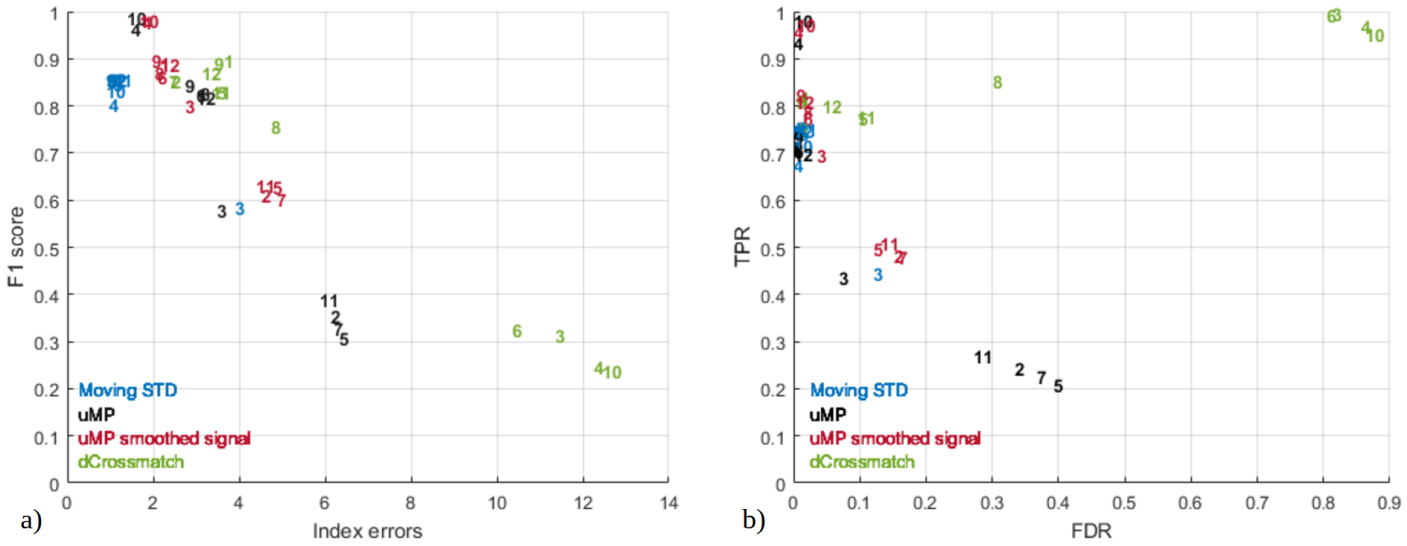


Fig. 6: Visualization of a) the couple of F1-score F_1 and motif location error e, and of b) TPR vs. FDR, for all cases of time series. Each case is represented by a number.

provide good performance for the studied signals. Their limitations could be compensated by combining both methods, since both have fast runtime and are adjustable in real time.

Following this work, we are currently using the selected methods of this paper on real environmental signals to then classify discovered motifs and obtain a better knowledge of radiological environment [15].

ACKNOWLEDGMENTS

We thank Dr. Fabian Noering for his help concerning the implementation of the Crossmatch.

REFERENCES

[1] A. Mueen, "Time series motif discovery: dimensions and applications,' in Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2014, vol. 4, no 2, p. 152-159.

[2] T. C. Fu, "A review on time series data mining," in Engineering Applications of Artificial Intelligence, 2011, vol. 24, no 1, p. 164-181.

[3] S. Torkamani and V. Lohweg, "Survey on time series motif discovery," in Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 7, no. 2, p. e1199, 2017.

[4] F. K.-D. Noering, Y. Schroeder, K. Jonas and F. Klawonn, "Pattern discovery in time series using autoencoder in comparison to nonlearning approaches," in Integrated Computer-Aided Engineering, no. Preprint, pp. 1–20, 2021.

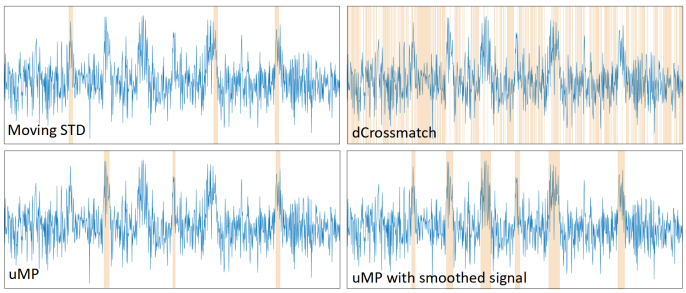


Fig. 7: Outputs of signal n°3: orange highlights represent indices of discovered motifs.

[5] J. Lin, E. Keogh, S. Lonardi and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. 2003. p. 2-11.

[6] K. Bascol, R. Emonet, E. Fromont and J. M. Odobez, "Unsupervised interpretable pattern discovery in time series using autoencoders," in : Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer, Cham, 2016. p. 427-438.

[7] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in KDD workshop, vol. 10, pp. 359–370, Seattle, WA, USA:, 1994.

[8] C.-C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen and E. Keogh, "Matrix profile i: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets," in IEEE 16th international conference on data mining (ICDM), pp. 1317–1322, Ieee, 2016.

[9] F. Madrid, S. Imani, R. Mercer, Z. Zimmerman, N. Shakibay and E. Keogh, "Matrix profile xx: Finding and visualizing time series motifs of all lengths using the matrix profile," in 2019 IEEE International Conference on Big Knowledge (ICBK), pp. 175–182, IEEE, 2019.

[10] M. Toyoda, Y. Sakurai, and Y. Ishikawa, "Pattern discovery in data streams under the time warping distance,"in The VLDB Journal, vol. 22, no. 3, pp. 295–318, 2013.

[11] P. Jancovic, M. Köküer, M. Zakeri and M. Russell, "Unsupervised discovery of acoustic patterns in bird vocalisations employing dtw and clustering," in 21st European Signal Processing Conference (EUSIPCO 2013), pp. 1–5, IEEE, 2013.

[12] https://www.cs.ucr.edu/ eamonn/MatrixProfile.html.

[13] C.-C. M. Yeh, N. Kavantzas and E. Keogh, "Matrix profile vi: Meaningful multidimensional motif discovery," in IEEE international conference on data mining (ICDM), pp. 565–574, IEEE, 2017.

[14] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," in Intelligent Data Analysis, 2007, vol. 11, no 5, p. 561-580.

[15] L. Poirier–Herbeck, E. Lahalle, N. Saurel and S. Marcos, "Classification des séries temporelles de longueurs variables pour la surveillance radiologique de l'environnement" in GRETSI: XXVIIIe Colloque. GRETSI, 2022, in press.