Classification de séries temporelles de longueurs variables pour la surveillance radiologique de l'environnement

Lisa Poirier--Herbeck[1,2], Elisabeth Lahalle[1], Nicolas Saurel[2], Sylvie Marcos[1]
[1]*Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France*
[2]*CEA Valduc, France*
lisa.poirier-herbeck@cea.fr elisabeth.lahalle@centralesupelec.fr
nicolas.saurel@cea.fr sylvie.marcos@centralesupelec.fr

***Résumé*—**De par l'émergence de séries temporelles massives dans de multiples domaines, l'extraction et la classification dans ces séries de signaux d'intérêt sont d'importants sujets de data mining. Dans ce contexte, nous proposons la classification de motifs extraits du signal par seuillage à l'aide d'un écart-type glissant. Nous proposons des méthodes de classification basées sur les mesures de distance *Dynamic Time Warping (DTW)*, *MPdist* et *area-based shape*, et sur les algorithmes de classification *k-Means* et hiérarchique. Leurs performances sont évaluées sur des signaux simulés de surveillance radiologique environnementale. Nos méthodes sont illustrées sur des signaux réels.

***Abstract*—**Due to the emergence of massive time series in multiple domains, the discovery and classification in these series of phenomena of interest are important subjects in data mining. In this context, we propose the classification of patterns extracted from the signal by thresholding using a sliding standard deviation. We propose clustering methods based on *Dynamic Time Warping (DTW)*, *MPdist* and *area-based shape* distance measurements, and on *k-Means* and hierarchical classification algorithms. Their performance is evaluated on simulated environmental radiological monitoring signals. Our methods are illustrated on real signals.

I. Introduction

Extraire et classer les événements enregistrés sous forme de motifs dans des séries temporelles massives est crucial dans de nombreuses applications industrielles. Cependant, l'expertise humaine est souvent dépassée par l'énorme quantité de données. Les algorithmes d'extraction de motifs et de classification pallient cette difficulté. En revanche, ils doivent être étudiés pour s'adapter aux caractéristiques des signaux d'intérêt. Dans le contexte de la surveillance radiologique environnementale, les séries temporelles sont composées de bruits de fond et de signaux d'intérêt de durées et d'intensités variables. Après avoir été extraits pour constituer des motifs, ces signaux d'intérêt doivent être classés selon leurs formes en faisant abstraction de leurs durées, de leurs amplitudes et de leurs niveaux de bruit. Définir la meilleure mesure de distance est une étape essentielle qui sera exploitée ensuite par les algorithmes de classification [1].

Dynamic Time Warping (DTW) [2] est une mesure de distance invariante aux changements de temporalité, permettant de comparer deux séries temporelles de longueurs différentes. Depuis 1994, *DTW* est largement utilisée dans de nombreux domaines scientifiques [3; 4; 5]. La distance *area-based shape* [6] est définie comme étant la distance *DTW* calculée dans un espace linéaire par morceaux. En effet, la distance est mesurée sur la suite des segments linéaires obtenus par estimation L_1 de la tendance des signaux [7]. L'avantage de cette mesure de distance comparée à *DTW* est qu'elle prend en compte la forme globale du signal et est d'autant plus intéressante pour les signaux longs [8]. Gharghabi et al. [9] proposent la mesure de distance *MPdist* suivant le principe que deux séries temporelles sont similaires si elles partagent des sous-séquences similaires. En plus des propriétés que couvrent *DTW* et l'*area-based shape*, *MPdist* est invariante aux déphasages et aux répétitions dans les signaux [9].

Dans cet article, nous évaluons les performances des mesures de distance *DTW*, *area-based shape* et *MPdist* pour une classification à l'aide des algorithmes usuels : *k-Means* et classification hiérarchique pour laquelle les clusters sont regroupés itérativement selon la méthode de Ward (*ward HC*) ou à l'aide d'une distance moyenne pondérée (*weigthed HC*) [1]. De plus, nous adaptons la mesure de distance *MPdist* afin de la rendre non-paramétrique.

Nous présentons en partie II les différentes distances retenues permettant de comparer des motifs de longueurs différentes. Dans la partie III, nous évaluons et discutons leurs influences sur les performances de classification des différents motifs extraits de séries temporelles. Ces performances sont évaluées sur des signaux synthétiques qui modélisent les signaux réels de surveillance radiologique environnementale. Nous concluons en partie IV sur l'apport de ces mesures de distance pour les données réelles.

II. Méthodologie

Dans un premier temps, nous extrayons les signaux d'intérêts, considérés comme motifs, des séries temporelles.

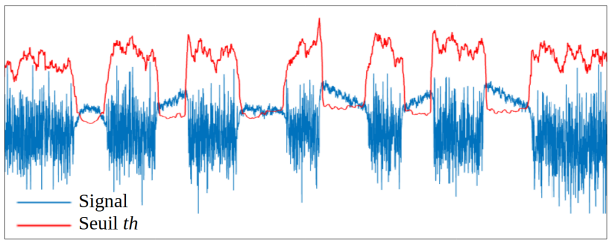

Figure 1 – Seuillage par écart-type glissant.

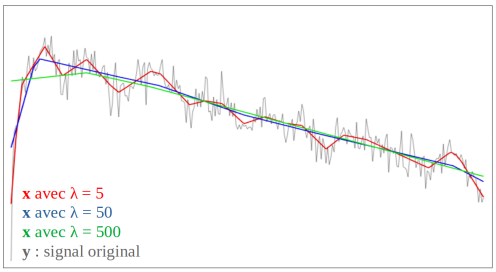

FIGURE 2 – Approximation L_1 (**x**) du signal (**y**).

Ensuite, nous étudions trois mesures de distance qui seront utilisées en partie III avec les algorithmes de classification *k-Means* et *HC* afin de classer les signaux d'intérêt extraits.

A. Extraction de motifs dans les séries temporelles

Une méthode d'extraction de motifs de séries temporelles par seuillage d'écart-type glissant est sélectionnée selon [10] et illustrée en figure 1. Les bruits de fond suivant une loi normale, le seuil retenu est $th = \mu_m + 3\sigma_m$, avec σ_m l'écart-type glissant du signal, obtenu sur une fenêtre de longueur l_w et μ_m la moyenne glissante du signal, obtenue sur une fenêtre de longueur N fois l_w. Afin d'obtenir un large horizon temporel des signaux considérés, la valeur $N = 20$ est choisie.

B. Mesures de distance étudiées

DTW. Une distance cumulative entre les deux séries temporelles d'entrée (les motifs extraits), est calculée. Le dernier élément de celle-ci définit la valeur de *DTW* [2]. Nous choisissons d'appliquer *DTW* telle quelle puis d'employer deux prétraitements différents afin de réduire le bruit dans les signaux : un lissage des signaux avec une moyenne glissante, et une estimation L_1 afin d'obtenir la tendance des signaux, obtenue par minimisation du critère des moindres carrés régularisé par la norme L_1 [7] :

$$\min\ Q(\mathbf{x}) = \frac{1}{2}\sum_{t=1}^{n}(y_t - x_t)^2 + \lambda \sum_{t=2}^{n-1} |x_{t-1} - 2x_t + x_{t+1}|,$$

où la série temporelle $\mathbf{y} = \{y_t\}_{t=1}^n$ est constituée d'une composante de tendance sous-jacente $\mathbf{x} = \{x_t\}_{t=1}^n$ et d'un élément aléatoire $\mathbf{z} = \{z_t\}_{t=1}^n$. Le paramètre $\lambda > 0$ prend en compte la régularité de $\mathbf{x}$ et de l'amplitude du résidu $\mathbf{z}$. L'estimation $\mathbf{x}$ est illustrée en figure 2 pour différentes valeurs du paramètre de régularisation λ.

MPdist. *MPdist* [9] est une mesure de distance basée sur la *Matrix Profile* [11]. Elle est définie par la formule suivante :

$$MPdist = \begin{cases} k^{th} \text{ valeur de } P_{ABBA} \text{ trié}, & |P_{ABBA}| > k \\ \max(P_{ABBA}), & |P_{ABBA}| \leq k \end{cases}$$

où P_{ABBA} est un vecteur contenant la distance euclidienne entre toutes les sous-séquences de longueur L d'une série temporelle A avec leurs sous-séquences les plus proches dans la série temporelle B et vice versa. Dans notre approche de classification, les séries temporelles A et B sont les motifs extraits dans la partie II-A. Nous choisissons de fixer le paramètre L à la taille du plus petit des deux motifs dont on calcule la distance. Ainsi, *MPdist* est rendue non-paramétrique.

Area-based shape. Les signaux d'entrée $X = \{(t_1, v_1), (t_2, v_2), ..., (t_m, v_m)\}$ et $Y = \{(s_1, u_1), (s_2, u_2), ..., (s_n, u_n)\}$, où v_i et u_i sont les valeurs des signaux aux instants t_i et s_i, sont filtrés pour obtenir une approximation L_1 de ces signaux sous forme de segments linéaires [7]. On note deux de ces segments $\bar{X} = \langle (t_L, v_L), (t_R, v_R) \rangle$ et $\bar{Y} = \langle (s_L, u_L), (s_R, u_R) \rangle$, où (t_L, v_L), (t_R, v_R), (s_L, u_L) et (s_R, u_R) sont les coordonnées de début et de fin de ces segments, dont on calcule la distance $d(\bar{X}, \bar{Y})$ définie par [6; 8] :

$$d(\bar{X}, \bar{Y}) = \begin{cases} w_1 \cdot \min(|s_L - t_L|, |s_R - t_R|) + w_2 \cdot |L_{\bar{X}} - L_{\bar{Y}}|, & \text{si } \bar{X} \parallel \bar{Y} \\ w_1 \cdot \min(|s_L - t_L|, |s_R - t_R|) + w_2 \cdot \sqrt{\frac{1}{2}\left|\begin{vmatrix} s & v_L & 1 \\ t_R & v_R & 1 \\ s'_R & u'_R & 1 \end{vmatrix}\right|}, & \text{sinon} \end{cases}$$

où

$$L_{\bar{X}} = \sqrt{(t_R - t_L)^2 + (v_R - v_L)^2},$$

$$L_{\bar{Y}} = \sqrt{(s_R - s_L)^2 + (u_R - u_L)^2},$$

$$w_1 = \frac{|\hat{X}| + |\hat{Y}|}{\check{X} + \check{Y} + |\hat{X}| + |\hat{Y}|}, \qquad w_2 = 1 - w_1,$$

avec $s'_R = t_L - s_L + s_R$, $u'_R = v_L - u_L + u_R$. On note $\hat{X} = \sum_{i=1}^m v_i/m$ et $\check{X} = \sum_{i=1}^m t_i/m$ pour $X = \{(t_1, v_1), (t_2, v_2), ..., (t_m, v_m)\}$. Puis, dans cet espace linéaire par morceaux, la distance entre les séries temporelles approximées par segments $\{\bar{X}_1, \bar{X}_2, ..., \bar{X}_{k1}\}$ et $\{\bar{Y}_1, \bar{Y}_2, ..., \bar{Y}_{k2}\}$ est calculée par *DTW*, $k1$ et $k2$ étant respectivement le nombre de segments dans les séries X et Y.

III. EXPÉRIENCES

A. Signaux

Nous présentons dans cette section l'influence des distances présentées en section II-B dans les algorithmes de classification *k-Means* et *HC*. Nous considérons les méthodes présentées sur des signaux synthétiques composés de bruits de fond suivant une loi normale (de différentes variances et moyennes) et de motifs de formes, de tailles et de niveaux de bruits différents. Pour chaque type de bruit, trois réalisations de bruit sont générées pour chaque signal. Pour un des signaux, la moyenne du bruit de fond évolue sinusoïdalement. Les motifs sont obtenus via un logiciel de surveillance radiologique environnementale appartenant au CEA. Ces signaux sont construits d'après nos observations des signaux réels. Les amplitudes, les longueurs et les formes des motifs varient. Des exemples de signaux sont en figure 3. Environ 300 motifs sont extraits des signaux par la méthode présentée en partie II-A. Cette étape d'extraction des motifs introduit une erreur sur un horizon temporel de quelques points aux extrémités du motif. [10].

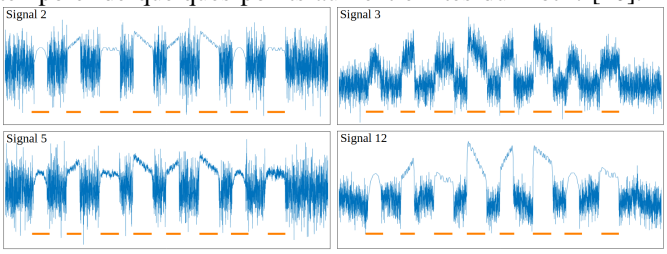

FIGURE 3 – Exemples de signaux synthétiques et positions de leurs motifs (orange).

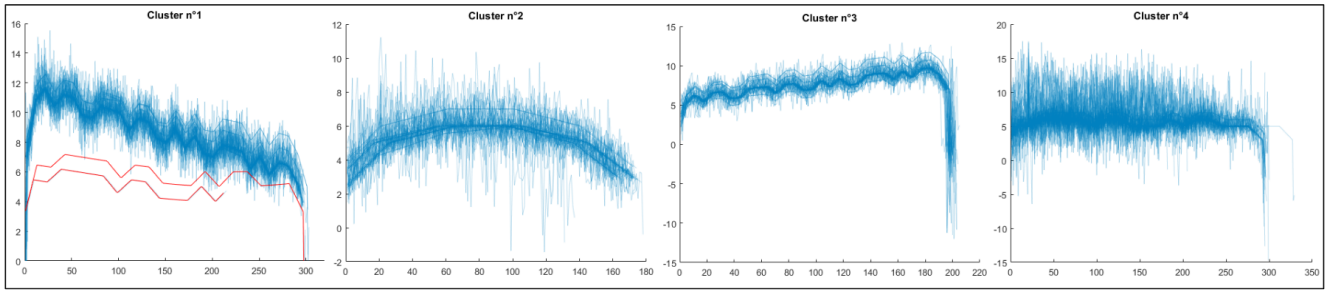

FIGURE 5 – Visualisation des clusters estimés par la méthode de classification *k-Means* en utilisant *MPdist* avec le meilleur score *RI* = 0.92. Superposés dans chaque cluster, les motifs bien classés sont en bleu et les motifs mal classés en rouge.

B. Métriques d'évaluation

Utilisant des algorithmes de classification supervisés, nous fournissons la valeur du nombre de clusters à estimer (dans notre cas, cette valeur est égale à 4). La performance des méthodes est évaluée par le *Rand Index (RI)* [12] :

$$RI = \frac{a+d}{a+b+c+d},$$

où a est le nombre de paires de motifs qui appartiennent à la même classe et au même cluster estimé, b est le nombre de paires de motifs qui appartiennent à la même classe et à différents clusters estimés, c est le nombre de paires de motifs qui appartiennent à différentes classes et au même cluster estimé, et d est le nombre de paires de motifs qui appartiennent à différentes classes et à différents clusters estimés.

C. Résultats

Les scores *RI* sont présentés dans le tableau I. Même si les améliorations de lissage et d'estimation L_1 apportées à *DTW* réhaussent légèrement les scores, *DTW* présente globalement de moins bons résultats que la distance *area-based shape*. Cependant, les performances avec *area-based shape* et le filtrage L_1 sont variables et nécessitent le réglage du paramètre de régularisation λ. Il est choisi ici par maximisation du score *RI* calculé à partir de la connaissance préalable des classes (voir figure 4). *MPdist* en non-paramétrique obtient les meilleures performances puisqu'elle présente les *RI* les plus élevés et les plus stables. En figure 5 sont représentés les clusters estimés par l'algorithme *k-Means* en utilisant la distance *MPdist* (score *RI*=0.92, voir le tableau I). On observe que les motifs sont tous bien classés à l'exception de deux motifs (figure 5). Ces deux motifs ont été mal classés du fait de la variation de la moyenne d'un des bruits de fond.

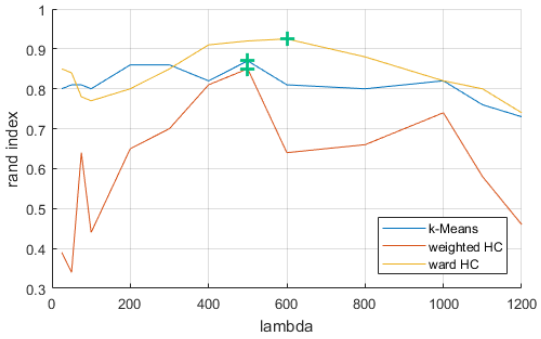

FIGURE 4 – Scores *RI* en fonction du paramètre de λ pour les méthodes *HC* et *k-Means* en utilisant *area-based shape*. Les croix vertes indiquent les λ retenus.

	k-Means	weighted HC	ward HC
DTW	0.85	0.82	0.85
lissage + DTW	0.85	0.84	0.87
filtrage L_1 trend + DTW	0.86	0.69	0.87
area-based shape	0.87	0.85	**0.92**
MPdist	**0.92**	**0.93**	0.91

TABLE I – Scores de *Rand Index RI* pour différents algorithmes de classification et mesures de distances. *k-Means* utilisant une initialisation aléatoire du centre des classes, les *RI* sont les moyennes de 10 réalisations.

Pour un ensemble d'environ 300 motifs, les distances *DTW*, *MPdist* et *area-based shape* ont respectivement des temps de calcul d'environ 20s, 60s et 15min. On utilise un ordinateur avec un processeur Intel(R) Core(TM) i7-10875H CPU 2.30 GHz, avec une mémoire RAM de 64.0 Go.

D. Discussion

Cette étude est réalisée dans le but de classer des séries temporelles de surveillance radiologique environnementale. Pour les distances *area-based shape* et *DTW* avec filtrage L_1, le réglage préalable du paramètre λ est problématique pour des données réelles bruitées et de tailles non-égales, puisque pour le définir, il faut d'abord connaître les clusters à estimer à priori (voir le tableau I). Une mesure de distance complètement non-supervisée comme notre adaptation de *MPdist* est plus appropriée sur ce point. De plus, *MPdist* a un temps de calcul très bas comparé à *area-based shape*. On estime alors que *MPdist* en non-paramétrique est le meilleur candidat pour classer nos données réelles. Comme première expérimentation sur nos données réelles, nous choisissons d'appliquer la méthode de classification *k-Means* avec un nombre de clusters choisi à 5 et en utilisant la distance *MPdist* sur les motifs extraits des données réelles par la méthode d'extraction de motifs présentée en partie II-A (exemple en figure 6). Les clusters trouvés sont illustrés en figure 7. Ces premiers résultats font apparaître des formes bien distinctes de motifs pour chaque cluster. Ce résultat est intéressant puisque l'algorithme a su classer des formes de signaux de manière automatique malgré

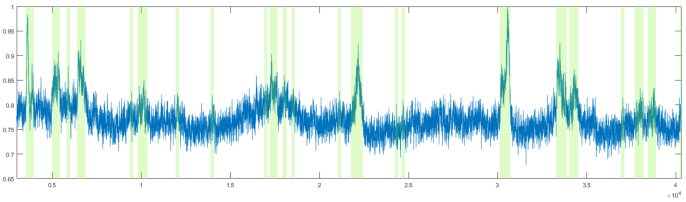

FIGURE 6 – Exemple de signal réel collecté sur un mois. Les marqueurs verts définissent les motifs extraits avec la méthode présentée en partie II-A. Le signal représenté est normalisé.

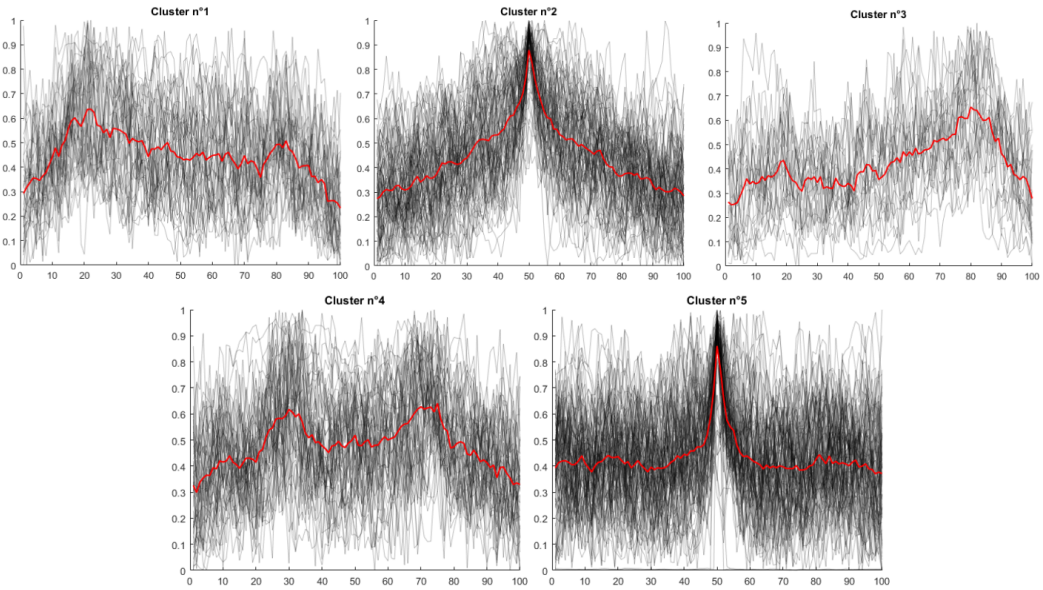

FIGURE 7 – Classification *k-Means* en utilisant *MPdist*, appliquée aux données réelles. Les motifs représentés sont normalisés en amplitude et interpolés pour obtenir des signaux de même durée pour une meilleure visualisation. Les signaux rouges sont les moyennes des signaux.

la diversité des longueurs, du bruit et des formes de signaux dans ce jeu de données réelles.

IV. CONCLUSION

Dans cette étude, nous proposons de classer des motifs de longueurs, de bruits et d'amplitudes variables, selon leurs formes. Ces motifs sont au préalable extraits de signaux de surveillance radiologique de l'environnement par une méthode d'extraction de motifs par seuillage avec écart-type glissant. Notre adaptation non-paramétrique de la mesure de distance *MPdist* présente de bonnes performances avec les algorithmes de classification testés : *k-Means* et la classification hiérarchique. De plus, son temps de calcul réduit et son aspect non-supervisé permet une classification adaptée aux données réelles. En effet, la méthode couplant notre adaptation de *MPdist* avec *k-Means* montre des résultats cohérents et encourageants avec l'apparition de formes bien distinctes pour chaque cluster, dans les données réelles. Cette méthode de data mining s'avère donc très intéressante pour le domaine de la surveillance radiologique de l'environnement.

Pour la suite de nos travaux, nous souhaitons explorer les algorithmes de classification non-supervisés afin de supprimer la limitation principale qui est la nécessité de définir un nombre de clusters au préalable.

RÉFÉRENCES

[1] S. Aghabozorgi, A. S. Shirkhorshidi, and T. Y. Wah, "Time-series clustering–a decade review," *Information Systems*, vol. 53, pp. 16–38, 2015.

[2] 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.

[3] C. Keskin, A. T. Cemgil, and L. Akarun, "Dtw based clustering to improve hand gesture recognition," in *International Workshop on Human Behavior Understanding*, pp. 72–81, Springer, 2011.

[4] P. Jančovič, 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.

[5] J. M. Landmesser *et al.*, "The use of the dynamic time warping (dtw) method to describe the covid-19 dynamics in poland," *Oeconomia Copernicana*, vol. 12, no. 3, pp. 539–556, 2021.

[6] X. Wang, F. Yu, and W. Pedrycz, "An area-based shape distance measure of time series," *Applied Soft Computing*, vol. 48, pp. 650–659, 2016.

[7] S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky, "l_1 trend filtering," *SIAM review*, vol. 51, no. 2, pp. 339–360, 2009.

[8] X. Wang, F. Yu, W. Pedrycz, and J. Wang, "Hierarchical clustering of unequal-length time series with area-based shape distance," *Soft Computing*, vol. 23, no. 15, pp. 6331–6343, 2019.

[9] S. Gharghabi, S. Imani, A. Bagnall, A. Darvishzadeh, and E. Keogh, "Matrix profile xii : Mpdist : a novel time series distance measure to allow data mining in more challenging scenarios," in *IEEE International Conference on Data Mining (ICDM)*, pp. 965–970, IEEE, 2018.

[10] L. Poirier-Herbeck, E. Lahalle, N. Saurel, and S. Marcos, "Unknown-length motif discovery methods in environmental monitoring time series," in *2022 International COnference on Electrical, Computer and Energy Technologies (ICECET)*, 20-22 juillet 2022, Prague.

[11] 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 *2016 IEEE 16th international conference on data mining (ICDM)*, pp. 1317–1322, Ieee, 2016.

[12] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.