Vers des mesures d'évaluation interprétables pour la segmentation de séries temporelles

Félix Chavelli*, Paul Boniol*
Michaël Thomazo*

*Inria, ENS, CNRS, Université PSL
45 rue d'Ulm, 75005 Paris, France
felix.chavelli@inria.fr

Résumé. La segmentation de séries temporelles est cruciale dans de nombreux domaines. L'évaluation des méthodes état de l'art est toutefois limitée : les mesures actuelles (ARI, points de changement) ignorent la qualité des segments, la nature des erreurs et ne sont pas interprétables. Nous proposons **WARI** (Weighted ARI), sensible à la position des erreurs, et **SMS** (State Matching Score), évaluant quatre types d'erreurs de segmentation. Validées empiriquement, elles offrent une évaluation plus fidèle et des diagnostics inédits. Ce travail a été accepté à NeurIPS 2025 sous le titre "Toward Interpretable Evaluation Measures for Time Series Segmentation".

1 Introduction

Les séries temporelles sont omniprésentes dans les domaines scientifiques et industriels Bagnall et al. (2019); Palpanas (2015). Leur analyse inclut la segmentation (ou détection de points de changement/d'états), visant à identifier les états sous-jacents et leurs transitions, correspondant par exemple à une activité humaine Reiss (2012) ou une consommation électrique Petralia et al. (2023). Bien que de nombreux algorithmes existent Ermshaus et al. (2023); Gharghabi et al. (2017); Nagano et al. (2018); Lai et al. (2024); Wang et al. (2023); Carpentier et al. (2024), leur évaluation souffre de limites majeures : (1) ignorance de la qualité globale par les mesures de points de changement ; (2) traitement uniforme des erreurs par l'ARI ; (3) manque d'interprétabilité. Nous introduisons **WARI** (**W**eighted **A**djusted **R**and **I**ndex) et **SMS** (**S**tate **M**atching **S**core). WARI étend l'ARI en intégrant la position temporelle des erreurs. SMS identifie et évalue quatre types d'erreurs, permettant une pondération personnalisée et une meilleure transparence. Nous validons empiriquement ces mesures et montrons leur apport pour l'évaluation de l'état de l'art. Nos contributions incluent : (1) une typologie d'erreurs (Sec. 2.1-2.2) ; (2) l'analyse des limites (Sec. 2.3) ; (3) WARI et SMS (Sec. 3) ; (4) validation empirique (Sec. 4.1) ; et (5) l'impact sur l'évaluation des méthodes existantes (Sec. 4.2).

2 Contexte et Fondamentaux

Une **série temporelle** de longueur N et dimension D est une séquence $T = [t_1, \ldots, t_N]$ où $t_i \in \mathbb{R}^D$. Une sous-séquence est notée $T_{[i,j]}$. Une **séquence d'états** associée $S = [s_1, \ldots, s_N]$ contient des étiquettes discrètes $s_i \in \mathcal{S}$. Un indice i est un **point de changement** si $s_i \neq s_{i+1}$.

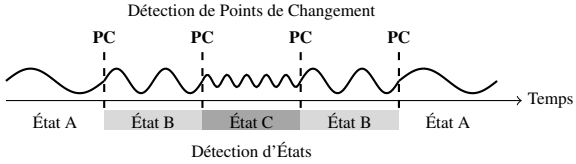

FIG. 1 – *Illustration de la détection de points de changement vs. détection d'états.*

2.1 Segmentation de Séries Temporelles : Un Problème à Deux Visages

La segmentation divise une série en segments homogènes. Deux approches principales existent (Fig. 1) : la **détection de points de changement** et la **détection d'états**.

2.1.1 Détection de Points de Changement

Elle produit une séquence de points de changement $(c_1, \ldots, c_M)$ marquant les transitions entre régimes stables. Les méthodes incluent les approches basées sur des profils (ClaSP Ermshaus et al. (2023), FLUSS Gharghabi et al. (2017)), statistiques (BinSeg Bai (1997), PELT Killick et al. (2012)) ou bayésiennes (BOCD Adams et MacKay (2007)).

2.1.2 Détection d'États

Elle vise à prédire une séquence d'états $P = (p_1, \ldots, p_N)$ où chaque état correspond à un régime récurrent. Les méthodes incluent les encodeurs (E2USD Lai et al. (2024), Time2State Wang et al. (2023)), les CNN (RP-mask Mirmomeni et al. (2021)), les graphes (GRAB Lu et al. (2021)), les modèles probabilistes (HDP-HSMM Nagano et al. (2018), TICC Hallac et al. (2018)) ou les règles (PaTSS Carpentier et al. (2024)). La détection de points de changement étant un sous-problème de la détection d'états, nous nous concentrons sur cette dernière.

2.2 Évaluation de la Segmentation de Séries Temporelles : Typologie des Erreurs et Propriétés Souhaitées

Afin d'évaluer avec précision la qualité d'une segmentation lorsqu'elle est comparée à une vérité terrain, nous devons définir formellement les types d'erreurs de segmentation. Soit $\mathcal{S} = \{s_1, \ldots, s_M\}$ un ensemble fini d'états. Soit $R = (r_1, r_2, \ldots, r_N) \in \mathcal{S}^N$ la séquence d'états *réelle*, ou *vérité terrain*, et soit $P = (p_1, p_2, \ldots, p_N) \in \mathcal{S}^N$ la séquence d'états *prédite*. Nous définissons un *bloc d'erreur* comme un intervalle d'indices contigus maximal $[i, j]$ qui ne peut être étendu sans inclure un point correctement classifié tel que $\forall k, l \in [i, j],\ p_k = p_l$ et $\forall k \in [i, j],\ p_k \neq r_k$. Pour un bloc d'erreur $[i, j]$ dans P, nous définissons l'*atomicité* $A_{[i,j]} = \big|\{\, r_k : k \in [i, j] \,\}\big|$ comme le nombre d'états distincts au sein de $R_{[i,j]}$.

Cette typologie est essentielle car la sévérité des erreurs varie. Les transitions entre états sont souvent graduelles, mais les annotations sont généralement binaires, créant une ambiguïté sur la position exacte des frontières. Bien que des approches comme l'étiquetage graduel existent Carpentier et al. (2024), une alternative consiste à utiliser des mesures robustes à cette ambiguïté. Ainsi, une erreur proche d'une frontière (retard, transition) est supposée moins sévère qu'une erreur au sein d'un segment stable (manquant, isolé).

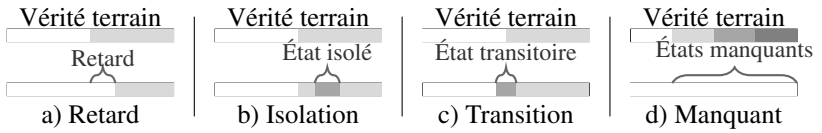

FIG. 2 – *Vérité terrain (haut) et exemples d'erreurs (bas) : retard, isolation, transition, manquant.*

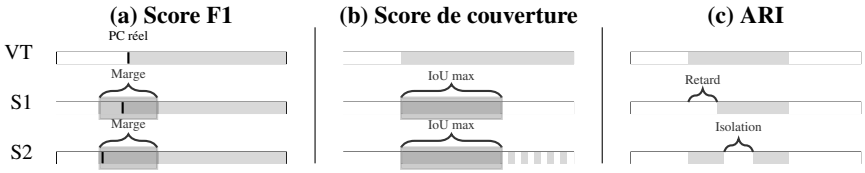

FIG. 3 – *Limitations des scores (a) F1, (b) couverture et (c) ARI. Pour deux segmentations différentes (S1 meilleure que S2), les scores sont identiques.*

Propriétés Souhaitées Cette typologie motive quatre propriétés clés pour une mesure d'évaluation : **P1** : *Sensibilité à la **longueur** des erreurs.* **P2** : *Prise en compte de la **position** temporelle des erreurs.* **P3** : *Pénalisation différenciée selon le **type** d'erreur.* **P4** : ***Interprétabilité** du score quant à la qualité de la segmentation.*

Ces propriétés sont des lignes directrices et non des axiomes stricts. Elles peuvent interagir : par exemple, l'impact de la longueur d'une erreur (**P1**) peut être modulé par sa position (**P2**) ou son type (**P3**).

2.3 Mesures Existantes et Limitations

Bien que plusieurs mesures aient été adoptées dans la littérature, chacune présente un ensemble d'hypothèses et d'inconvénients, ne parvenant pas à capturer certaines des propriétés souhaitées mentionnées ci-dessus. Cette section passe en revue ces mesures couramment utilisées, en discutant de leurs forces et limitations.

2.3.1 Mesures de Détection de Points de Changement

Parmi les mesures couramment utilisées pour la détection de points de changement, le **score F1** est une mesure harmonique qui combine la précision et le rappel. Suivant les travaux précédents, nous évaluons un score F1 "tolérant", identifiant les détections correctes en appariant les points de changement prédits aux annotations de vérité terrain dans une *marge* donnée.

Cependant, la sélection de la *marge* appropriée est difficile. L'exemple de la Fig. 3(a) illustre deux scénarios S1 et S2 dans lesquels le score F1 est de 1, bien que S1 contienne moins d'erreurs que S2, ne satisfaisant donc pas la **Propriété P1**. Un paramètre qui est une fonction de la longueur de la série temporelle est préférable, comme proposé dans Ermshaus et al. (2023), où il est fixé à 1% de la longueur de la série temporelle.

Le **score de couverture**, une autre mesure couramment utilisée, capture la similarité au niveau des segments plutôt que la correspondance exacte des points de changement. Contrairement au score F1, qui traite les points de changement comme des événements discrets, la couverture tient compte du chevauchement des segments. Il est défini comme la moyenne des

scores d'intersection sur union (IoU) pour chaque segment de la vérité terrain, normalisée par le nombre de segments. Avec R et P représentant les séquences d'états réelles et prédites, le score de couverture est calculé comme suit : $C = \frac{1}{N} \sum_{r \in R} |r| \max_{p \in P} \frac{|r \cap p|}{|r \cup p|}$. Cependant, le score de couverture peut attribuer des scores identiques à des segmentations qualitativement différentes, échouant ainsi à satisfaire la **Propriété P1**. La Fig. 3(b) illustre cette limitation avec deux segmentations prédites qui obtiennent le même score de couverture malgré des différences significatives. Cela souligne la nécessité de mesures complémentaires qui capturent mieux la qualité de la segmentation.

2.3.2 Mesures de Détection d'États

La performance de détection d'états est le plus souvent évaluée avec des mesures basées sur le clustering (Wang et al. (2023); Lai et al. (2024)), telles que l'**Indice de Rand Ajusté** (ARI), l'**Information Mutuelle Normalisée** (NMI) et l'**Information Mutuelle Ajustée** (AMI). Dans le reste de l'article, nous nous concentrerons sur l'ARI, mais les observations et propositions faites s'appliquent également à la NMI et à l'AMI.

Le calcul de l'ARI est basé sur l'Indice de Rand (RI) qui calcule la fraction de paires concordantes (c'est-à-dire, des paires qui sont soit groupées soit séparées ensemble) sur le nombre total de paires. Formellement, avec R et P représentant les séquences d'états réelles et prédites et $U_R = \{r_i : r_i \in R\}$ et $U_P = \{p_i : p_i \in P\}$ les ensembles uniques d'états, nous définissons la *matrice de contingence* $C = [n_{ij}]$ de taille $|U_R| \times |U_P|$ avec $n_{ij} = \sum_{k=1}^{N} \mathbf{1}\{r_k = U_R[i] \wedge p_k = U_P[j]\}$, c'est-à-dire le nombre d'observations au pas de temps k qui appartiennent à l'état $U_R[i]$ dans la première séquence d'états R et $U_P[j]$ dans la seconde séquence d'états P. Enfin, avec $\mathbb{E}[\mathrm{RI}]$ l'Indice de Rand attendu sous un modèle aléatoire, l'ARI est calculé comme suit : $\mathrm{RI} = \frac{\sum_{i,j} \binom{n_{ij}}{2}}{\binom{N}{2}}, \quad \mathrm{ARI} = \frac{\mathrm{RI} - \mathbb{E}[\mathrm{RI}]}{1 - \mathbb{E}[\mathrm{RI}]}$.

3 WARI et SMS : Nos Mesures Proposées

Comme exposé dans la section précédente, les mesures d'évaluation existantes présentent des limitations importantes (ne satisfaisant pas la **Propriété P1**, **P2** ou **P3**). De plus, ces mesures offrent une interprétabilité limitée, rendant difficile pour les utilisateurs de comprendre les scores de précision ou d'identifier les faiblesses spécifiques et les axes d'amélioration (ne satisfaisant pas la **Propriété P4**). Pour remédier à ces lacunes, nous proposons deux mesures de détection d'états. La première, **WARI**, consiste en une version modifiée (*pondérée*) de l'ARI standard, les rendant sensibles à la *distance aux frontières*. La seconde est une nouvelle mesure, nommée **SMS** (**S**tate **M**atching **S**core), qui identifie une correspondance entre les états prédits et la vérité terrain, laquelle est ensuite utilisée pour calculer un score basé sur les types d'erreurs rencontrées selon la taxonomie définie dans la section précédente.

3.1 Vers une Sensibilité à la Position : L'Indice de Rand Ajusté Pondéré

L'ARI traitant toutes les erreurs uniformément, nous proposons l'**Indice de Rand Ajusté Pondéré** (WARI) pour intégrer la sensibilité à la position (**P2**). WARI pondère les erreurs selon leur distance aux frontières. Soit d_i la distance temporelle de l'instant i au point de changement

réel le plus proche. Nous définissons un poids $w_i = 1 + \alpha\, d_i$, où $\alpha \geq 0$ est un paramètre (défaut 0,1). Ainsi, les erreurs au cœur des segments (grand d_i) sont davantage pénalisées.

Matrice de Contingence Pondérée : La matrice de contingence est adaptée en remplaçant les comptages par des sommes pondérées : $\widetilde{n}_{ij} = \sum_{x_k \in U_i \cap V_j} w_k$, utilisées ensuite dans la formule de l'ARI pour calculer WARI. Cette pondération est applicable à d'autres mesures (NMI, AMI).

Propriétés : WARI étend l'ARI en introduisant une sensibilité à la position (**P2**). Pour $\alpha = 0$, WARI équivaut à l'ARI. WARI conserve l'échelle et l'interprétation de l'ARI. Le score est robuste aux variations de α (différence linéairement bornée).

3.2 Améliorer l'Interprétabilité : Le State Matching Score (SMS)

Pour satisfaire **P3** et **P4**, nous introduisons le **State Matching Score** (SMS). SMS aligne les séquences d'états pour évaluer les erreurs selon leur type et gravité. Le calcul (les algorithmes complets sont donnés dans l'article publié) se fait en deux étapes : (1) correspondance optimale des états (via l'algorithme hongrois Kuhn (1955)) maximisant le chevauchement ; (2) identification et pénalisation des blocs d'erreurs selon leur type, longueur et contexte. SMS permet une personnalisation via des poids de pénalité, sans altérer la robustesse du score, déterminé principalement par le volume d'erreurs.

Propriétés formelles. SMS est interprétable et robuste. Avec des poids nuls, $\mathrm{SMS} = 1 - E/N$ (E : longueur totale des erreurs). Avec des poids non nuls, le score reste borné : $1 - \frac{(1+w_{\max})\,E}{N} \leq \mathrm{SMS} \leq 1 - \frac{E}{N}$. La sensibilité aux poids est bornée : $|\mathrm{SMS}(w) - \mathrm{SMS}(w')| \leq \|w - w'\|_\infty \frac{E}{N}$.

4 Évaluation Expérimentale

Nous évaluons WARI et SMS face à l'ARI sur 6 méthodes (E2USD Lai et al. (2024), Time2State Wang et al. (2023), HDP-HSMM Nagano et al. (2018), TICC Hallac et al. (2018), ClaSP Ermshaus et al. (2023), PaTSS Carpentier et al. (2024)) et 5 jeux de données (PAMAP2 Reiss (2012), USC-HAD Zhang et Sawchuk (2012), UCR-SEG Dau et al. (2018), ActRecTut Bulling et al. (2014), MoCap noa). Le code est disponible en open-source [1].

4.1 Évaluation des Mesures d'Évaluation

Un exemple qualitatif, issu du jeu de données MoCap (Fig. 4), montre que l'ARI favorise E2USD malgré des erreurs isolées, tandis que WARI et SMS préfèrent Time2State, plus cohérent. SMS offre de plus un diagnostic des types d'erreurs.

4.2 Impact sur l'État de l'Art

SMS révèle (Fig. 5) que les méthodes basées sur des encodeurs-décodeurs (Time2State, E2USD) produisent souvent des erreurs isolées, contrairement à ClaSP ou TICC (erreurs manquantes/retards). Cette analyse fine guide l'optimisation des algorithmes (ex : ajustement de la sensibilité de clustering pour Time2State).

1. Dépôt Public : `https://github.com/fchavelli/tsseg-eval`

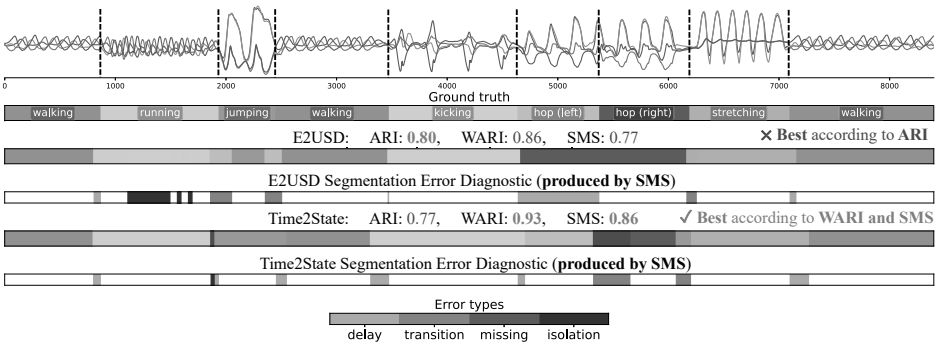

FIG. 4 – *Segmentation d'une série temporelle du jeu de données MoCap utilisant E2USD et Time2State.*

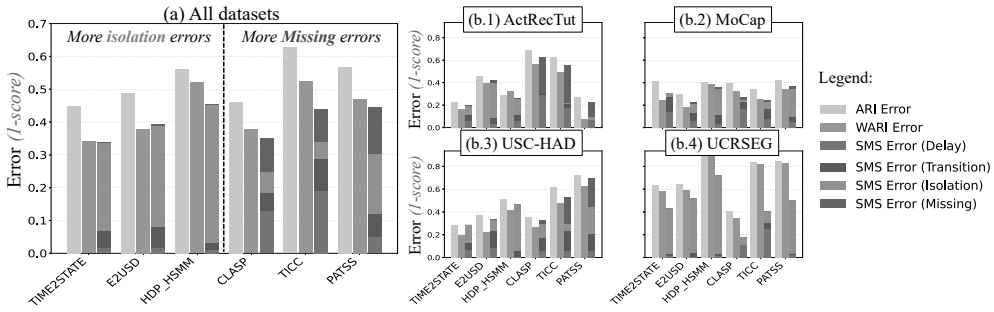

FIG. 5 – *Taux d'erreur et contribution des types d'erreurs pour (a) tous les jeux de données, et (b) par jeux de données.*

5 Conclusion

Nous abordons une lacune clé dans l'évaluation de la segmentation de séries temporelles en formalisant une typologie de quatre types d'erreurs distincts, et en proposant un ensemble de propriétés souhaitables pour les mesures d'évaluation. Nous introduisons deux nouvelles mesures d'évaluation, **WARI** et **SMS**, qui dépassent des limitations des approches existantes et fournissent de nouvelles informations. De telles informations ouvrent des directions prometteuses pour la sélection de modèles sensible aux erreurs, le développement, la paramétrisation et l'ensembling. Dans l'ensemble, ce travail contribue à des outils interprétables, robustes et personnalisables pour faire progresser l'évaluation et la conception d'algorithmes de segmentation.

Références

Carnegie Mellon University - CMU Graphics Lab - motion capture library.

Adams, R. P. et D. J. C. MacKay (2007). Bayesian online changepoint detection.

Bagnall, A. J., R. L. Cole, T. Palpanas, et K. Zoumpatianos (2019). Data series management (dagstuhl seminar 19282). *Dagstuhl Reports 9*(7), 24–39, doi: 10.4230/DAGREP.9.7.24.

Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric Theory 13*(3), 315–352, doi: 10.1017/S0266466600005831.

Bulling, A., U. Blanke, et B. Schiele (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys 46*(3), 1–33, doi: 10.1145/2499621. Publisher : Association for Computing Machinery (ACM).

Carpentier, L., L. Feremans, W. Meert, et M. Verbeke (2024). Pattern-based time series semantic segmentation with gradual state transitions. In S. Shekhar, V. Papalexakis, J. Gao, Z. Jiang, et M. Riondato (Eds.), *Proceedings of the 2024 SIAM International Conference on Data Mining, SDM 2024, Houston, TX, USA, April 18-20, 2024*, pp. 316–324. SIAM, doi: 10.1137/1.9781611978032.36.

Dau, H. A., E. Keogh, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, Yanping, B. Hu, N. Begum, A. Bagnall, A. Mueen, G. Batista, et Hexagon-ML (2018). The UCR Time Series Classification Archive.

Ermshaus, A., P. Schäfer, et U. Leser (2023). ClaSP : parameter-free time series segmentation. *Data Mining and Knowledge Discovery 37*(3), 1262–1300, doi: 10.1007/s10618-023-00923-x. Publisher : Springer Science and Business Media LLC.

Gharghabi, S., Y. Ding, C.-C. M. Yeh, K. Kamgar, L. Ulanova, et E. Keogh (2017). Matrix Profile VIII : Domain Agnostic Online Semantic Segmentation at Superhuman Performance Levels. In *2017 IEEE International Conference on Data Mining (ICDM)*, New Orleans, LA, pp. 117–126. IEEE, doi: 10.1109/icdm.2017.21.

Hallac, D., S. Vare, S. Boyd, et J. Leskovec (2018). Toeplitz Inverse Covariance-Based Clustering of Multivariate Time Series Data. arXiv :1706.03161 [cs].

Killick, R., P. Fearnhead, et I. A. Eckley (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association 107*(500), 1590–1598, doi: 10.1080/01621459.2012.737745. arXiv :1101.1438 [stat].

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly 2*(1-2), 83–97, doi: https://doi.org/10.1002/nav.3800020109.

Lai, Z., H. Li, D. Zhang, Y. Zhao, W. Qian, et C. S. Jensen (2024). E2Usd : Efficient-yet-effective Unsupervised State Detection for Multivariate Time Series. In *Proceedings of the ACM Web Conference 2024*, Singapore Singapore, pp. 3010–3021. ACM, doi: 10.1145/3589334.3645593. E2USD.

Lu, Y., P. Wang, B. Tang, S. Liang, C. Wang, W. Wang, et J. Wang (2021). GRAB : Finding Time Series Natural Structures via A Novel Graph-based Scheme. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, Chania, Greece, pp. 2267–2272. IEEE, doi: 10.1109/icde51399.2021.00235.

Mirmomeni, M., L. Kulik, et J. Bailey (2021). A Transferable Technique for Detecting and Localising Segments of Repeating Patterns in Time series. In *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China, pp. 1–10. IEEE, doi: 10.1109/ijcnn52387.2021.9534157.

Nagano, M., T. Nakamura, T. Nagai, D. Mochihashi, I. Kobayashi, et M. Kaneko (2018). Sequence Pattern Extraction by Segmenting Time Series Data Using GP-HSMM with Hierarchical Dirichlet Process. In *2018 IEEE/RSJ International Conference on Intelligent Robots*

and Systems (IROS), Madrid, pp. 4067–4074. IEEE, doi: 10.1109/iros.2018.8594029.

Palpanas, T. (2015). Data series management : The road to big sequence analytics. *SIGMOD Rec. 44*(2), 47–52, doi: 10.1145/2814710.2814719.

Petralia, A., P. Charpentier, P. Boniol, et T. Palpanas (2023). Appliance detection using very low-frequency smart meter time series. In *Proceedings of the 14th ACM International Conference on Future Energy Systems*, e-Energy '23, New York, NY, USA, pp. 214–225. Association for Computing Machinery, doi: 10.1145/3575813.3595198.

Reiss, A. (2012). PAMAP2 Physical Activity Monitoring.

Wang, C., K. Wu, T. Zhou, et Z. Cai (2023). Time2State : An Unsupervised Framework for Inferring the Latent States in Time Series Data. *Proceedings of the ACM on Management of Data 1*(1), 1–18, doi: 10.1145/3588697. Publisher : Association for Computing Machinery (ACM).

Zhang, M. et A. A. Sawchuk (2012). USC-HAD : a daily activity dataset for ubiquitous activity recognition using wearable sensors. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, Pittsburgh Pennsylvania, pp. 1036–1043. ACM, doi: 10.1145/2370216.2370438.

Summary

Time series segmentation is crucial across domains. Current evaluation measures (e.g., ARI), however, fail to capture segment quality and error nature. We introduce **WARI** (Weighted ARI), accounting for error positions, and **SMS** (State Matching Score), scoring four customizable error types. Empirically validated, they offer accurate assessment and novel insights.