Topological data analysis for unsupervised anomaly detection in time series

Alexandre Bois, Brian Tervil, Laurent Oudre
Université Paris Saclay, Université Paris Cité, ENS Paris Saclay, CNRS, SSA, INSERM, Centre Borelli
Gif-sur-Yvette, France
{firstname.name}@ens-paris-saclay.fr

***Abstract*—In this article, we propose a new algorithm for unsupervised anomaly detection in univariate time series, based on topological data analysis. It relies on delay embeddings and on the extraction of persistent cycles from the 1-dimensional persistent homology module constructed from the distance to measure Rips filtration. This filtration makes it possible to identify 1-cycles (i.e. loops) corresponding to recurrent patterns by leveraging density information. Points in those cycles are considered as normal, and the algorithm can then assign an anomaly score to any point which is its distance to the normal set. In this paper, we describe the algorithm and test it on several real-world datasets, showing that it is competitive with state-of-the-art anomaly detection methods.**

***Index Terms*—Anomaly detection, topological data analysis, time series analysis**

I. Introduction

Anomaly detection in time series is an important problem in data science, with applications in many fields such as healthcare and engineering. A time series is a sequence of real numbers $\mathbf{x} = (\mathbf{x}_i)_{1 \leq i \leq n}$ (n will always denote the length of the time series). In the context of anomaly detection, $\mathbf{x}$ is assumed to be composed of a normal behavior and anomalies, i.e. points or sequences of points that differ from the normal behavior. More precisely, in several application contexts such as industrial monitoring or healthcare, the time series is usually assumed to be composed of some repetitive/frequent patterns, possibly of varying lengths (think for instance of an heartbeat in ECG data) among which some occur a large number of times (the *normal* ones) and some have significantly fewer occurrences (the *abnormal* ones) : see Figure 1 for an illustration.

Over the past years, several unsupervised anomaly detection algorithms have been developed from different research areas (see [1] and [2] for a comprehensive review). Among them, some rely on a model and use the prediction error as an anomaly detector and some are based of clustering techniques applied on the subsequences in order to detect outliers. For instance, LOF [3] transforms the time series into a point cloud and studies the density of each point to assess whether or not they correspond to normal or abnormal behaviors. The fact that the patterns can have different lengths, that there can be multiple normal patterns and multiple occurrences of an anomaly, and noise make it difficult to build a universal anomaly detection algorithm [1]. The main differences between the approaches actually lie in the implied definition given to the notion of normality.

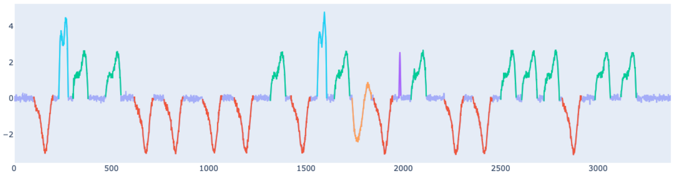

Fig. 1. An example time series, with normal patterns in green and red (and the null parts in dark blue) and three abnormal ones (in blue, orange and purple), with Gaussian noise of amplitude 0.1.

Topological data analysis (TDA), and more specifically persistent homology [4] is a set of techniques derived from algebraic topology, which allows to analyze the structure of data by constructing a sequence of simplicial complexes (a *filtration*). The *persistence diagram* sums up when connected components, loops or higher-dimensional simplices appear and disappear when going through the filtration. TDA has been applied to many fields including time series analysis. It is particularly adapted to study structured data such as time series with a periodic behavior [5], [6], [7], [8], and methods from TDA benefit from stability theorems that guarantee a certain robustness to noise [9], [10].

In the case of periodic functions, the importance of 1-dimensional persistent homology (the study of loops in point clouds) was theoretically studied in [5] and it was applied to time series in [6] by transforming the data into a point cloud and considering the most important loop. By extension, 1-dimensional persistent homology is also relevant to study time series with repetitive patterns. Our method consists in transforming the time series into a point cloud and extracting 1-cycles (i.e. loops) that are considered to correspond to normal patterns of the time series. Those cycles are identified on the persistence diagram because density information is used to construct the filtration. Once ’normal cycles’ have been extracted, an anomaly score is defined for each point of the embedding as its distance to the normal cycles.

Section II gives the required background on TDA to understand the algorithm described in Section III. We show the experimental results and discuss them in Section IV.

II. Topological data analysis background

In this section, we introduce the objects from topological data analysis (TDA) (mostly from [10]) that will be used in the rest of the paper. See [4], [10] for more complete background.

A. Simplicial homology

Definition 1. *A k-simplex on a set $X \subset \mathbb{R}^d$ is an unordered tuple $\sigma = [x_0, ..., x_k]$ of $k+1$ distinct elements of X. The elements $x_0, ..., x_k$ are called the vertices of σ. If each vertex of a simplex ρ is also a vertex of σ, then ρ is called a face of σ. A simplicial complex K is a set of simplices such that any face of a simplex of K is a simplex of K.*

Definition 2. *A filtration of a simplicial complex K is a family of simplicial complexes $(K_\alpha)_{\alpha \geq 0}$ such that $K_0 = \emptyset$, $\alpha < \alpha' \Rightarrow K_\alpha \subset K_{\alpha'}$ and $\bigcup_{\alpha \geq 0} K_\alpha = K$. The filtration value of a simplex σ is the lowest α such that $\sigma \in K_\alpha$.*

This definition is useful for theory, but in practice the number of simplices is finite, in which case we will only use a finite set of indices α_i such that $\emptyset = K^{\alpha_0} \subset K^{\alpha_1} \subset \cdots \subset K^{\alpha_N} = K$ and $\alpha_i \leq \alpha < \alpha_{i+1} \Rightarrow K_{\alpha_i} = K_\alpha$. Without loss of generality, we can also assume that for all i there exists a simplex $\sigma_{i+1} \in K$ such that $K^{\alpha_{i+1}} = K^{\alpha_i} \cup \{\sigma_{i+1}\}$.

Let K be a simplicial complex on a set $X \subset \mathbb{R}^d$, $\mathbb{F} = \mathbb{Z}/2\mathbb{Z}$ and $0 \leq k \leq d$.

Definition 3. *The space $C_k(K)$ of k-chains is defined as the set of formal sums of k-simplices of K with coefficients in $\mathbb{F}$, that is to say, if all the k-simplices of K are $\sigma_1, \ldots, \sigma_{n_k}$, all the elements of the form: $c = \sum_{i=1}^{n_k} a_i \sigma_i$ with $a_i \in \mathbb{F}$.*

$C_k(K)$ is a vector space whose addition and scalar multiplication are naturally defined.

Definition 4. *Let $\sigma = [v_1, \ldots, v_k]$ be a k-simplex with vertices $v_1, \ldots, v_k$, and $[v_1, \ldots, \hat{v_i}, \ldots, v_k]$ be the $(k-1)$-simplex spanned by those points minus v_i. The boundary operator ∂ is defined as:*

$$\partial : \left\{ \begin{array}{lll} C_k(K) & \longrightarrow & C_{k-1}(K) \\ \sigma & \longmapsto & \partial\sigma = \sum_{i=1}^{k} (-1)^i [v_1, \ldots, \hat{v_i}, \ldots, v_k]. \end{array} \right.$$

We have the following sequence of linear maps:

$$\{0\} \rightarrow C_d(K) \xrightarrow{\partial} C_{d-1}(K) \xrightarrow{\partial} \ldots \xrightarrow{\partial} C_0(K) \xrightarrow{\partial} \{0\}.$$

They satisfy $\partial \circ \partial = 0$: we call such a sequence of maps a *chain complex*. This constitutes the setup for homology. We can now define *cycles*, *boundaries* and *homology groups*.

Definition 5. *Let ∂_k denote the boundary operator $\partial : C_k(K) \rightarrow C_{k-1}(K)$. We define the set $Z_k(K)$ of k-cycles of K as $Z_k(K) = \mathsf{Ker}(\partial_k)$ and the set $B_k(K)$ of k-boundaries of K as $B_k(K) = \mathsf{Im}(\partial_k)$.*
We have $B_k(K) \subset Z_k(K) \subset C_k(K)$ so we can define the k^{th} homology group as: $H_k(K) = Z_k(K)/B_k(K)$.

The elements of $H_k(K)$ represent the of k-dimensional "holes". For example, elements of $H_1(K)$ are loops.

B. Persistent homology

The main idea of TDA is to build a filtration on top of the data and study how the structure of the simplicial complexes evolves while increasing the filtration parameter α using *persistent homology*.

Let $(K^{\alpha_i})_{0 \leq i \leq N}$ be a filtration such that for each index i, $K^{\alpha_{i+1}} = K^{\alpha_i} \cup \{\sigma_{i+1}\}$. We call $C_k^i, Z_k^i, B_k^i, H_k^i$ the respective spaces of k-chains, k-cycles, k-boundaries and k^{th} homology group of K^i. The goal is to follow the evolution of H_k^i as i increases. It can be shown [4] that when a k-simplex σ_{i+1} ($k > 0$) is added, it either creates a new homology class in H_k^{i+1} (i.e. a new k-cycle that is independent of those of H_k^i) or it closes a $k-1$-dimensional hole of H_{k-1}^{i-1}, so H_{k-1}^i has one less homology class than H_{k-1}^{i-1}, in that case we say that σ_{i+1} killed a homology class (by convention, we always consider that when two classes merge, the younger class gets killed). If $k = 0$, each new vertex creates a homology class in H_0.

The final result of persistent homology is the set of all so-called **persistent pairs** $(\sigma_{l(j)}, \sigma_j)$ such that for each j, $\sigma_{l(j)}$ creates a component and σ_j kills it. We say that the **persistence** of such a pair is $j - l(j) - 1$. The algorithms to compute them are described in detail in [4]. The k-dimensional **persistence diagram** is the set of points of coordinates $(\alpha_{l(j)}, \alpha_j)$ such that $\sigma_{l(j)}$ is a k-simplex (counted with multiplicity). A persistence diagram is showed on Figure 3.

III. Method

In this section, we describe our algorithm for unsupervised anomaly detection. The algorithms has four main steps that are described below: transform the time series into a *delay embedding*, compute the *Distance-to-measure (DTM) Rips filtration*, extract the normal 1-cycles, compute the anomaly scores.

The idea behind our algorithm is that, if we consider a time series such as the one from Figure 1, each pattern should correspond to at least one loop in the embedded space (see Figure 2) that can be detected using 1-dimensional persistent homology. As normal patterns have more occurrences than abnormal ones, their points should have a higher density and thus and thus a lower filtration value. This would make it possible to discriminate the corresponding 1-cycles on the persistence diagram, as their birth date will be lower. Our set of normal points is then the set of 1-cycles detected as normal.

A. Delay embeddings

A *delay embedding* is a way of transforming a time series into a point cloud of chosen dimension d. We will study the structure of this point cloud using the tools from TDA described in Section II to detect important loops.

Definition 6. *The delay embedding of a signal $\mathbf{x}$ with dimension $d \geq 2$ and delay $\tau \in \mathbb{N}$ is the following point cloud in $\mathbb{R}^d$: $X_{d,\tau} = ((\mathbf{x}_i, \mathbf{x}_{i+\tau}, \ldots, \mathbf{x}_{i+(d-1)\tau}))_{1 \leq i \leq n-(d-1)\tau}$.*

In what follows, $X_{d,\tau}$ will always denote the delay embedding associated to $\mathbf{x}$ with dimension d and delay τ: subscript

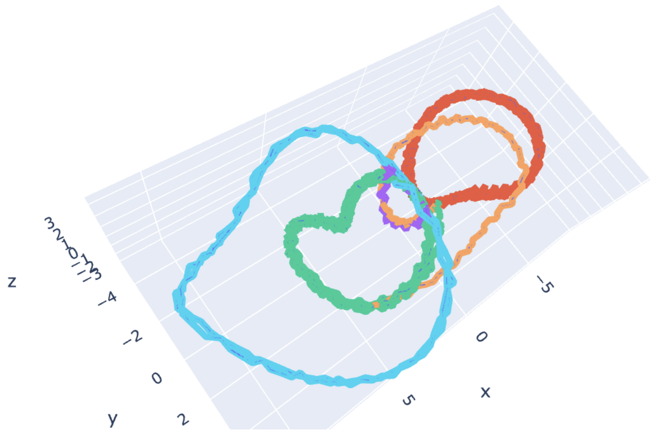

Fig. 2. Delay embedding of the time series from Figure 1 (PCA in 3D), with $d = 10$ and $\tau = 4$. Colors correspond to the colors in Figure 1.

terms will be omitted where the context is obvious. Figure 2 shows a delay embedding of the time series from Figure 1.

B. Distance-to-measure (DTM) Rips filtration

Here, we describe the distance-to-measure (DTM) Rips filtration used in our algorithm. See [10] for a more general introduction and stability results.

Let X be a finite subset of $\mathbb{R}^d$ (a point cloud) and $f : X \to \mathbb{R}^+$. For all $x \in X$ and $\alpha \in \mathbb{R}^+$, we define a radius $r_x(\alpha)$ as:

$$r_x(\alpha) = \begin{cases} -\infty & \text{if } \alpha < f(x) \\ \alpha - f(x) & \text{otherwise} \end{cases}$$

Now, let us denote by $\bar{B}_f(x, \alpha)$ the closed Euclidean ball $\bar{B}(x, r_x(\alpha))$ (by convention, a ball is empty if its radius is $-\infty$).

Definition 7. *With the above notations, the weighted Rips filtration with parameters (X, f), $\mathsf{Rips}[X, f]$ is the filtered simplicial complex such that each vertex (or 0-simplex) $x \in X$ has filtration value $f(x)$, and such that for $2 \le i \le d + 1$: $[x_1, .., x_i] \in \mathsf{Rips}[X, f]_\alpha \iff \forall (j, k),\ \bar{B}_f(x_j, \alpha) \cap \bar{B}_f(x_k, \alpha) \neq \emptyset$.*

As we only look at intersections of pairs of balls, Rips filtrations are characterized by the filtration values of 0 and 1 simplices. The following proposition from [10] gives those values.

Proposition 1. *Let $x, y \in X$. The filtration value of $[x]$ in $\mathsf{Rips}[X, f]$ is $f(x)$, and the filtration value of $[x, y]$ in $\mathsf{Rips}[X, f]$ is:*

$$\begin{cases} \max(f(x), f(y)) & \textit{if } ||x - y|| < |f(x) - f(y)| \\ \frac{||x-y||+f(x)+f(y)}{2} & \textit{otherwise.} \end{cases}$$

Distance-to-measure (DTM) functions were introduced in [10] to make weighted filtrations robust to outliers.

Definition 8. *If X is a finite subset of $\mathbb{R}^d$, we denote by μ_X the empirical measure on X. Let $q \le \mathsf{Card}(X) \in \mathbb{N}$ and $m = \frac{q}{\mathsf{Card}(X)}$. The DTM $d_{\mu_X,m}$ is defined as:*

$$\forall x \in \mathbb{R}^d,\ d_{\mu_X,m}(x) = \sqrt{\frac{1}{q}\sum_{i=1}^{q} ||x - NN^{(i)}(x)||^2}$$

where $NN^{(i)}(x)$ is the i^{th} nearest neighbor to x.

The DTM Rips filtration with parameter m is finally defined as $\mathsf{Rips}[X, d_{\mu_X,m}]$.

Our algorithm computes the DTM filtration on a delay embedding $X_{d,\tau}$ and the corresponding persistence diagram, and extracts 1-cycles. As point clouds can be very large (they have $n - (d - 1)\tau$ points), those steps can be too long for the algorithm to be used in practice ($O(n^3)$ in the worst case for 1D persistent homology and cycle extraction [4], [11]). To solve this problem, we compute persistent homology on a subsampled set $\tilde{X}_{d,\tau}$ of $X_{d,\tau}$, so the filtration is $\mathsf{Rips}[\tilde{X}_{d,\tau}, d_{\mu_{X_{d,\tau}},m}]$. Note that the filtration values of points in $\tilde{X}_{d,\tau}$ are their original values from $d_{\mu_{X_{d,\tau}},m}$, which insures that the subsampling does not change the fact that $\mathsf{Rips}[\tilde{X}_{d,\tau}, d_{\mu_{X_{d,\tau}},m}]$ is a filtration. This is important to keep the density information from the whole point cloud when subsampling (otherwise, the effect of the number of occurrences of normal points would disappear). To compute $\tilde{X}_{d,\tau}$, we choose a number of points n_{points} and use a greedy method: we start with a random point and, until we have n_{points} points, add the furthest one to the set of already chosen points.

C. Extraction of normal 1-cycles

Here, we describe the step that consists on identifying normal 1-cycles by reading the persistence diagram, and extracting those cycles.

Let D be the persistence diagram corresponding to the filtration $\mathsf{Rips}[\tilde{X}_{d,\tau}, d_{\mu_{X_{d,\tau}},m}]$. We propose an algorithm that relies on the choice of two thresholds: one on persistence (we focus on the most persistent points, which describe important structures), and one on the birth date (among those points, we consider those with a birth date above the cycle to be abnormal). To choose the persistence threshold, we sort the persistence of all points by decreasing order in a list l, find the index i such that $l[i] - l[i+1]$ is maximal, and keep points corresponding to indices from 1 to $i + n_{add}$, where n_{add} is a parameter. We will use $n_{add} = 2$ in practice, to keep at least three points and thus to be able to compare at least two differences in birth dates. Figure 3 illustrates our algorithm applied to the time series from Figure 1 (with delay embedding from Figure 2). We used $q = 10$, $n_{add} = 2$ and $n_{points} = 200$. The persistence threshold is represented as a line parallel to the diagonal as the persistence of a point is proportional to its distance to the diagonal. Here, the persistence threshold on the diagram is around 0.5 (5 points are above: 3 are above the largest gap, and we add two more).

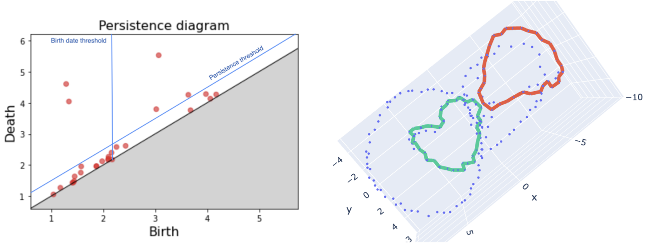

Fig. 3. Extraction of normal 1-cycles step. **Left:** persistence diagram of the DTM-filtration of the delay embedding from Figure 2, with persistence and birth date thresholds in blue. **Right:** subsampling of the delay embedding with 200 points, and cycles detected as normal (in green and red).

We choose the birth date threshold as follows: we sort the points that are above the persistence threshold by increasing birth date and add the point with minimal birth date to the list, then keep all the points until the maximal difference is reached and remove the one we added. If the point of maximal persistence is not kept, add all points until it is (we assume that the most persistent cycle should be normal, due to the DTM filtration). On the example from Figure 3, the birth date threshold is around 2.5, so the two points in the upper left corner are detected as normal (among the 5 most persistent ones).

If there are no 1-cycles in D, we take an arbitrary "cycle" as normal. In this case, one should look for different parameters d and τ to make the point cloud less dense.

The cycle extraction step takes as input the list of normal points on the diagram D and the filtration, and outputs a list of cycles $C = (c_1, \dots, c_{n_{cycles}})$, each one representing the homology class of a point from the list (a cycle is a list of points). We use the matrix reduction algorithm from [11], which consists in column additions on the matrix of the boundary operator on the simplices of the filtration. The idea is to form groups of 1-simplices (edges) that form a cycle with the desired birth and death dates. The green and red cycles on the point cloud $\tilde{X}_{d,\tau}$ from Figure 3 are the two normal cycles extracted with this method, corresponding to the two points evoked above (notice that we found the green and red cycle from Figure 2).

D. Anomaly scores

Once the list C of normal cycles has been computed, an anomaly score is given to each point $x \in X_{d,\tau}$, which is its Euclidean distance to C: $d(x, C)$.

Finally, we get an anomaly score for $\mathbf{x}_i$ by averaging the scores of all points in $X_{d,\tau}$ of which $\mathbf{x}_i$ is a coordinate: $\text{Score}(\mathbf{x})_i \leftarrow \text{mean}(\{d(X_j, C) \mid \text{max}(0, i-(d-1)\tau) \leq j \leq i\})$.

One can choose a threshold to the anomaly score to get a binary answer. Typically, one can chose to keep score only above a certain quantile. In section IV, we will compare algorithms using the AUC-ROC curve of each anomaly score not to be biased by an arbitrary choice of threshold. We do not give a specific method to choose the threshold, as it depends on the application.

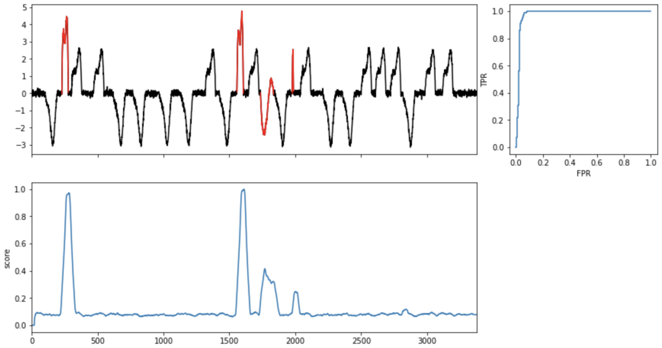

Fig. 4. Anomaly score step. **Top left:** signal from Figure 1 with anomalies in red. **Bottom:** Anomaly score obtained with our algorithm (distance to the cycles from Figure 3, normalized). **Top right:** ROC curve (AUC = 0.98).

Figure 4 shows the results and ROC curve of our algorithm applied to the signal from Figure 1 (with delay embedding from Figure 2 and normal cycles from Figure 3). The two cycles corresponding to the normal patterns have been extracted so those pattern have an anomaly score close to zero.

IV. Results and discussion

This section shows the result obtained with our algorithm on the 18 public datasets provided by the TSB-UAD benchmark suite [1], for a total of 1980 univariate time series with labeled anomalies.

For each time series, we compute the area under the ROC curve (AUC-ROC) obtained by looking at all the possible thresholds on the anomaly score. The same method is applied to 12 anomaly detection algorithms in [1]. Using the AUC-ROC, the evaluation does not depend on the choice of a threshold for each algorithm. Table III-D shows the average AUC-ROC obtained on each dataset with our method (TDA) with the above parameters, and the results of the TSB-UAD benchmark [1]. Figure 5 shows the critical diagrams comparing the average rank of each method using the Friedman test followed by the Wilcoxon or Nemenyi test with $\alpha = 0.05$.

The parameters are chosen the same way for all time series. We estimate a period L for the time series using the first maximum of the autocorrelation function (we used the find_length function from TSB-UAD, which was used for all other methods using a delay embedding). We empirically chose $\tau = 6$ (in practice, if τ is too small, the embedding will stay close to the line spanned by $(1, \dots, 1)$ and it will be harder to detect cycles). In [5] Perea and Harer show that in the case of trigonometric functions, $d\tau$ should be a multiple of the period to maximize persistence in 1D homology. With this in mind, we empirically set $d = \text{max}(40, \text{min}(120, \lfloor \frac{L}{3} \rfloor))$. We chose $q = \lfloor \frac{n}{L} \rfloor$ as this value would approximate the number of normal occurrences in the case where there is one normal atom. We take $n_{add} = 2$ for the reasons explained in section III-C and $n_{points} = 400$ for computation time reasons.

The above results show that our method is competitive with the state-of-the-art in anomaly detection on 18 standard datasets. It has the best score on 4 of them, the best average

	TDA	IForest	IForest1	LOF	MP	PCA	NORMA	HBOS	POLY	OCSVM	AE	CNN	LSTM
Dodgers	**0.79**	**0.79**	0.64	0.54	0.52	0.77	**0.79**	0.3	0.69	0.64	0.73	0.68	0.39
ECG	0.88	0.75	0.61	0.56	0.58	0.71	**0.95**	0.68	0.70	0.64	0.73	0.52	0.54
IOPS	**0.82**	0.54	0.78	0.50	0.72	0.74	0.76	0.64	0.68	0.71	0.63	0.61	0.61
MGAB	0.58	0.57	0.58	**0.96**	0.91	0.54	0.55	0.54	0.51	0.52	0.71	0.58	0.56
NAB	**0.76**	0.45	0.56	0.48	0.49	0.69	0.58	0.68	0.75	0.61	0.54	0.52	0.50
NASA-MSL	0.64	0.57	0.69	0.52	0.52	0.75	0.55	0.77	**0.81**	0.64	0.70	0.57	0.57
NASA-SMAP	**0.83**	0.72	0.68	0.68	0.62	0.74	0.80	0.77	0.80	0.65	0.77	0.68	0.64
SensorScope	0.52	0.56	0.56	0.55	0.50	0.54	0.59	0.56	**0.62**	0.51	0.52	0.52	0.53
YAHOO	0.64	0.62	0.81	0.86	0.86	0.57	0.92	0.57	0.76	0.50	0.79	**0.96**	0.94
KDD21	0.75	0.65	0.57	0.78	**0.90**	0.58	0.88	0.60	0.58	0.60	0.79	0.74	0.66
Daphnet	0.70	0.74	0.68	**0.78**	0.44	0.69	0.46	0.69	0.77	0.45	0.44	0.47	0.44
GHL	0.86	**0.94**	**0.94**	0.54	0.42	0.91	0.64	0.92	0.76	0.45	0.63	0.47	0.47
Genesis	0.84	0.78	0.66	0.68	0.35	0.85	0.6	0.59	**0.87**	0.70	0.72	0.73	0.53
MITDB	0.71	0.70	0.61	0.61	0.69	0.67	**0.86**	0.70	0.68	0.65	0.80	0.58	0.51
OPP	0.48	0.49	0.52	0.45	**0.82**	0.52	0.65	0.54	0.28	0.38	0.70	0.47	0.57
Occupancy	0.53	0.86	0.78	0.53	0.32	0.78	0.53	**0.89**	0.80	0.66	0.69	0.79	0.71
SMD	0.77	0.85	0.73	0.69	0.51	0.80	0.61	0.77	**0.87**	0.61	0.63	0.61	0.58
SVDB	0.77	0.72	0.58	0.59	0.74	0.68	**0.92**	0.71	0.67	0.68	0.79	0.58	0.55

TABLE I
AVERAGE AUC-ROC ON EACH DATASET.

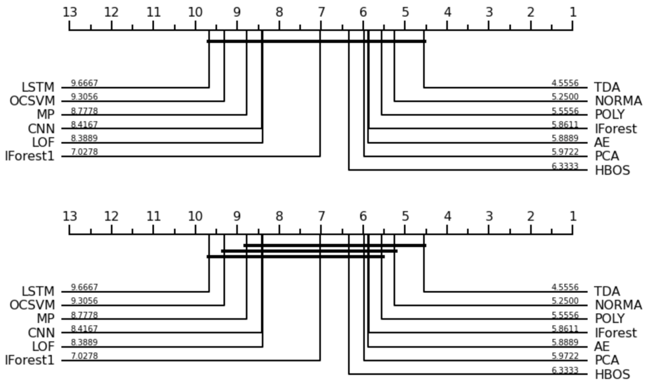

Fig. 5. Critical diagrams for $\alpha = 0.05$. **Top:** Friedman test + Wilcoxon test. **Bottom:** Friedman test + Nemenyi test.

rank (though the difference with the best methods is not significant as shown on Figure 5), and it is in the top 5 on 13 datasets. Our algorithm was built for a certain model of time series, with repeating patterns, so it should be used to study data that present such a behavior.

One could argue that using 1-cycles is useless because points could be studied individually using their filtration value (i.e. the birth dates on the 0-dimensional persistence diagram) which is the opposite of a measure of the density of $X_{d,\tau}$ around each point. This approach would then be similar to the LOF algorithm [3]. However, in practice, the filtration values can take a range of values due to noise or differences between occurrences of the same pattern and choosing a threshold would be hard. Moreover, an abnormal sequence with slow variations will give dense points in $X_{d,\tau}$ which would make the 0D approach or LOF fail. Considering 1D persistent homology and considering only cycles with a high persistence makes the choice of normal points easier by focusing on a few cycles corresponding to relevant components of $X_{d,\tau}$. It also makes it possible to eliminate sequences with slow variations (because the corresponding cycles have low persistence).

V. CONCLUSION

This paper shows how tools from topological data analysis can be used to detect structure in time series, and how the choice of a relevant filtration can highlight the desired structure (in this case, normal cycles corresponding to repetitive patterns).

The DTM filtration makes it possible to compute persistent homology on a subset of points while keeping density information from the whole point cloud, which makes the algorithm usable in practice on large datasets. An improvement perspective could be to use more efficient algorithms or smaller filtrations to be able to consider larger point clouds.

REFERENCES

[1] J. Paparrizos, Y. Kang, P. Boniol, R. S. Tsay, T. Palpanas, and M. J. Franklin, "Tsb-uad: an end-to-end benchmark suite for univariate time-series anomaly detection," *Proceedings of the VLDB Endowment*, vol. 15, no. 8, pp. 1697–1711, 2022.

[2] S. Schmidl, P. Wenig, and T. Papenbrock, "Anomaly detection in time series: a comprehensive evaluation," *Proceedings of the VLDB Endowment*, vol. 15, no. 9, pp. 1779–1797, 2022.

[3] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.

[4] J.-D. Boissonnat, F. Chazal, and M. Yvinec, *Geometric and topological inference*. Cambridge University Press, 2018, vol. 57.

[5] J. A. Perea and J. Harer, "Sliding windows and persistence: An application of topological methods to signal analysis," *Foundations of Computational Mathematics*, vol. 15, no. 3, pp. 799–838, 2015.

[6] S. Emrani, T. Gentimis, and H. Krim, "Persistent homology of delay embeddings and its application to wheeze detection," *IEEE Signal Processing Letters*, vol. 21, no. 4, pp. 459–463, 2014.

[7] T. Bonis, F. Chazal, B. Michel, and W. Reise, "Topological phase estimation method for reparameterized periodic functions," *arXiv preprint arXiv:2205.14390*, 2022.

[8] A. Bois, B. Tervil, A. Moreau, A. Vienne-Jumeau, D. Ricard, and L. Oudre, "A topological data analysis-based method for gait signals with an application to the study of multiple sclerosis," *Plos one*, vol. 17, no. 5, p. e0268475, 2022.

[9] F. Chazal, V. De Silva, and S. Oudot, "Persistence stability for geometric complexes," *Geometriae Dedicata*, vol. 173, no. 1, pp. 193–214, 2014.

[10] H. Anai, F. Chazal, M. Glisse, Y. Ike, H. Inakoshi, R. Tinarrage, and Y. Umeda, "Dtm-based filtrations," in *Topological Data Analysis: The Abel Symposium 2018*. Springer, 2020, pp. 33–66.

[11] Edelsbrunner, Letscher, and Zomorodian, "Topological persistence and simplification," *Discrete & Computational Geometry*, vol. 28, pp. 511–533, 2002.