Expert Systems with Applications 42 (2015) 6026–6038

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# Persistent homology for time series and spatial data clustering

Cássio M.M. Pereira [*], Rodrigo F. de Mello

*Institute of Mathematical and Computer Sciences, São Carlos, SP 13566-590, Brazil*

ARTICLE INFO

*Article history:*
Available online 16 April 2015

*Keywords:*
Data clustering
Spatial data
Time series
Persistent homology
Topological data analysis

ABSTRACT

Topology is the branch of mathematics that studies how objects relate to one another for their qualitative structural properties, such as connectivity and shape. In this paper, we present an approach for data clustering based on topological features computed over the persistence diagram, estimated using the theory of persistent homology. The features indicate topological properties such as Betti numbers, i.e., the number of $n$-dimensional holes in the discretized data space. The main contribution of our approach is enabling the clustering of time series that have similar recurrent behavior characterized by their attractors in phase space and spatial data that have similar scale-invariant spatial distributions, as traditional clustering techniques ignore that information as they rely on point-to-point dissimilarity measures such as Euclidean distance or elastic measures. We present experiments that confirm the usefulness of our approach with time series and spatial data applications in the fields of biology, medicine and ecology.

## 1. Introduction

Topological data analysis (TDA) (Carlsson, 2009; Lum et al., 2013) is a new research field that bridges computational methods with the mathematical theory of topology (Zomorodian, 2005). One of the motivations for the development of this field is the realization that while data analysis is useful in providing quantitative information, it could also benefit from the knowledge gained by analyzing more basic, or qualitative structures, especially in the presence of noise (Chang, Bacallado, Pande, & Carlsson, 2013). For instance, while traditional data analysis has excellent tools to quantitatively learn discriminating functions or hyperplanes from labeled examples, it has fewer tools for analyzing structural, or qualitative information, such as shape. That type of information can be very important to fields in which shape implies functional properties, such as in protein analysis (Sacan, Ozturk, Ferhatosmanoglu, & Wang, 2007).

While traditional data analysis is currently performed by comparing pairwise similarities between objects, commonly using a metric such as the Euclidean distance, it disregards information such as shapes, including loops and holes. For instance, a traditional clustering algorithm that connects geometrically close points will fail to find sets of points that are spatially distant but that present similar structure, or shape. Historically, topology was the field of mathematics created to study basic properties such as loops and holes.

Topology has its roots in the work of Leonhard Euler (Stillwell, 1993), when he examined the “Seven Bridges of Königsberg” problem, which aims to find a path through the city that crosses each bridge only once. Insights, such as the irrelevance of land masses to the problem led to the idea of vertices, while bridges to the idea of edges of a graph. That laid the foundations for what would become graph theory and later topology (Holzinger, 2014).

To realize the type of insights topology can provide, consider the mathematical objects of a square and a circle. From a purely geometric point of view, a square and a circle have very different properties. On the other hand, to topologists they are exactly the same object, as one can be continuously deformed into the shape of the other, without the need to cut or glue pieces together (Zomorodian, 2005).

Those points highlight a different perspective to approaching data analysis, namely of a more qualitative nature, which considers basic data properties. The approach we propose is suited to data sets in which shape has meaning. Applications abound in which that assumption is true. Scenarios such as protein analysis, spatial data and time series analysis are examples that benefit from a topological perspective. Our proposal is especially useful because it is more general and resilient than the simple application of distance measures in a metric space, which can often be corrupted by noise (Tan, Steinbach, & Kumar, 2005).

This idea that topological spaces are a generalization is illustrated in Fig. 1, in which a space endowed with connectivity information characterizes a topological space and the further addition

[*] Corresponding author.
*E-mail addresses:* cpereira@icmc.usp.br (C.M.M. Pereira), mello@icmc.usp.br (R.F. de Mello).

http://dx.doi.org/10.1016/j.eswa.2015.04.010
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

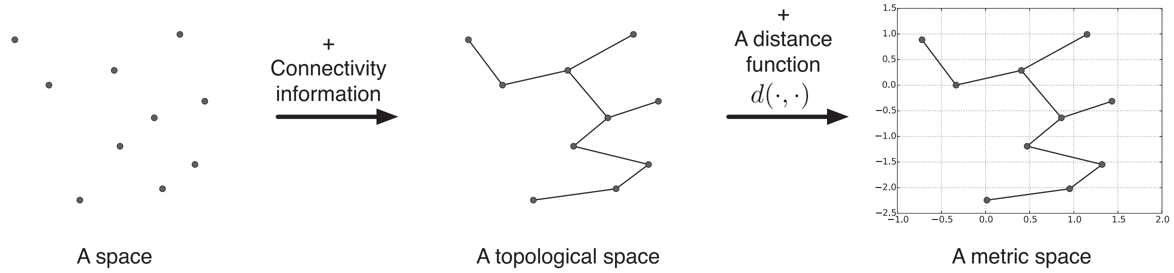


**Fig. 1.** A hierarchy of spaces. A topological space is characterized by the addition of connectivity information, while a metric space by the addition of a distance measure. Adapted from (Zomorodian, 2005).

of a specific distance function constitutes a metric space. One of the advantages of working only with connectivity information in topological spaces is that it does not suffer from the curse of dimensionality, i.e., points tend to be equidistant in high dimensional space (Bishop, 2006).

In this paper, we argue that topological features may help to identify interesting patterns in the data clustering of time series and spatial data. We present experiments in the fields of biology, medicine and ecology that support this hypothesis. Our application scenarios consist of GPS-tracking of birds, in which we assume the hypothesis that birds of the same species have similar movement behavior; EEG (Electroencephalography) time series analysis of alcoholic patients versus a control group; and time series analysis of the population growth of *Tribolium* flour beetles.

The basis of our topological analysis lies in the theory of persistent homology, which allows the constructions of persistence diagrams (PDs). Such diagrams can be viewed as summary statistics, which capture multi-scale topological features (Fasy et al., 2014). Homology can detect features such as connected components, tunnels and voids. A persistence diagram allows one to study for how long those features persist when a scaling parameter is varied. Thus, features that persist longer are said to be the most prominent ones. For instance, considering a point cloud of a 2-d circle will yield a long-persisting 2-d hole in the persistence diagram.

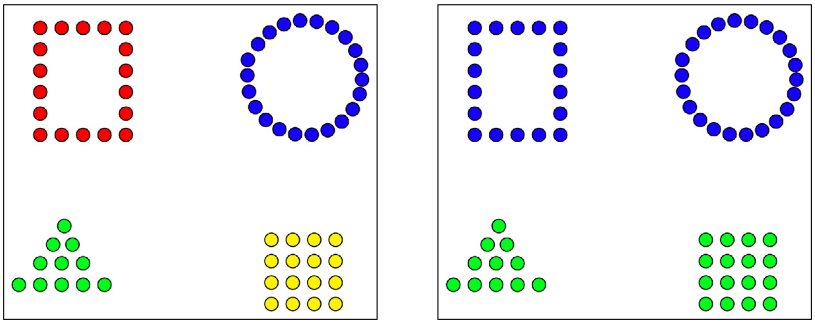
The proposed methodology can be summarized as follows. In order to compute topological features, we need to first discretize the data space. We do that by applying persistent homology, which first creates a triangulation of the space using simplicial complexes, to extract connectivity information. After that, topological features are computed over the persistence diagram, which indicates the birth and death of n-dimensional holes in the induced topological spaces. Finally, the computed topological features are used as input for a clustering algorithm, which allows us to map the extracted clusters to meaningful data properties.

To illustrate the basic premise of what our method aims to achieve consider Fig. 2. Traditional clustering techniques find the clusters shown in Fig. 2(a) by joining points according to a distance measure defined in a metric space. Our proposed topology-based approach takes as input distinct cloud points, such as those in Fig. 2(a) and merges clusters that have similar topological properties. In the example, shown in Fig. 2(b), clusters are merged according to the number of 2-d holes that they surround.

The main contribution of this paper is a framework for clustering spatial and time series data sets based on topological features. Insights can be gained by exploring basic qualitative properties of a data set, which are often ignored by traditional distance-based clustering approaches. For instance, analyzing how the topological properties of a time series attractor change over time, one can identify non-obvious recurrence relations, which can be missed if time series are clustered using traditional distance-based techniques with measures such as Euclidean distance or Dynamic Time Warping.

To illustrate our contribution consider three time series, $S_1, S_2, S_3$, defined according to Eqs. (1)–(3) and displayed in Fig. 3((a)–(c)). Series $S_1$ and $S_3$ are periodic sine waves, while series $S_2$ is a damped sine wave. If we use Dynamic Time Warping (Warren Liao, 2005) to compute normalized distances that already correct for time-axis misalignments we obtain the dissimilarity matrix $D$ displayed in Eq. (4). We observe that, according to DTW distance, $S_1$ and $S_2$ are closer together than $S_1$ and $S_3$, even though $S_1$ and $S_3$ are both periodic functions, while $S_2$ is not. This is due to the fact that while computing the distance, DTW focus on quantitative aspects and the higher frequency of $S_3$ contributes to the increase in distance.

Our proposed framework allows for a more qualitative analysis of the series. Consider the reconstructions in phase space using Takens embedding theorem (Takens, 1981) with embedding dimension $m = 2$ and separation $\tau = 100$, as illustrated in



(a) Traditional clustering.

(b) Topology-based clustering.

**Fig. 2.** Traditional clustering (a) groups points based on their distance in a metric space. Our proposed topology-based approach takes as input the clusters shown in (a) and then groups the clusters themselves according to their topological features. In the example, clusters are joined according to the number of 2-d holes they surround.

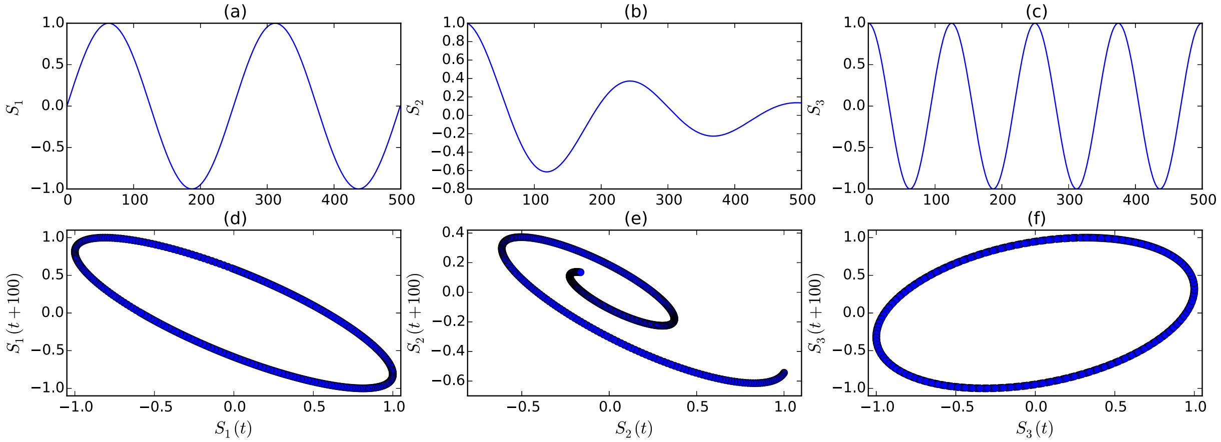


**Fig. 3.** Time series $S_1$ (a) and $S_3$ (c) are both sampled from periodic functions, while $S_2$ (b) is sampled from a damped sine wave. Traditional time series clustering techniques, based on DTW for instance, identify $S_1$ and $S_2$ as closer together, even though the behavior of their attractors are clearly different in phase space (d–f). Our proposed approach can identify those qualitative differences through the use of persistent homology and cluster $S_1$ and $S_3$ together.

Fig. 3((d)–(f)). One can clearly see the differences in behavior of the attractors, highlighting the recurrent nature of $S_1$ and $S_3$. Persistent homology allows one to quantify those qualitative aspects by computing the number of persistent $d$-dimensional holes of each attractor. Such computation yields the following Betti numbers: $\beta_{1_1} = 1, \beta_{2_1} = 0, \beta_{3_1} = 1$, such that $\beta_{i_j}$ indicates the number of $j$-dimensional holes of the $i$th attractor. Thus, persistent homology correctly identifies that $S_1$ and $S_3$ have more in common than $S_1$ and $S_2$. We explore in this paper the idea that such relationships can be meaningful and provide new insight.

$$S_1(t) = \sin(2\pi t) \tag{1}$$

$$S_2(t) = e^{-t}\cos(2\pi t) \tag{2}$$

$$S_3(t) = \sin(4\pi t + \pi/2) \tag{3}$$

$$D = \begin{array}{c} \begin{array}{ccc} S_1 & S_2 & S_3 \end{array} \\ \left( \begin{array}{ccc} 0 & 0.1426 & 0.1653 \\ & 0 & 0.2371 \\ & & 0 \end{array} \right) \end{array} \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \end{array} \tag{4}$$

In the next section, we provide a literature review of related research. In Section 3, we introduce background concepts and explain our proposed methodology. In Section 4, we present experimental results that support our hypothesis. Finally, in Section 5, we give our final remarks and conclusions.

## 2. Literature review

Chiang (2007) proposes a topology-based method to transform text documents into a hierarchical semantic space. Term associations form simplicial complexes, which as a connected component can identify a concept contained in a Web page. The technique consists in first constructing an undirected connected graph in which extracted features from the data are nodes. Edges are added if two features co-occurred in some instances. Next, concepts are recursively generated from the graph by separating the simplicial complexes into documents that have one of its faces but not the other. Finally, the concepts are clustered together. The authors evaluated their method on several category-based text collections, obtaining better results than other distance-based techniques. It is interesting to note that the authors did not use shape information as computed by persistent homology, but only the idea of co-occurrences in order to create the simplices. In our proposed approach, topology information is gathered from features computed over the persistence diagram, which has well-established mathematical theory to support it in the field of persistent homology.

Chazal, Guibas, Oudot, and Skraba (2013) proposed a persistence-based clustering scheme that combines a mode-seeking phase with a cluster merging phase in the density map. The first step, mode-seeking, consists in finding local peaks of the density function $f$, from which points are drawn from, to use as cluster centers. The problem with that approach is that the gradient and extreme points of a density function are unstable, so the approximation can lead to poor results. The authors propose to overcome this problem by using persistent homology to detect and merge unstable clusters. Each peak in the density function $f$ is mapped to a point in the persistence diagram, in which its importance is measured as the vertical distance from the point to the diagonal $y = x$. The main difference between their approach and ours is that they have used topology to make better estimates of the density functions that generate the data set, as to find the correct number of clusters. Our approach differs in that we are not interested solely in the clusters themselves, but in using topological properties of each cluster to derive relationships among them, i.e., we compare and group clusters based on their own topological properties.

Lum et al. (2013) propose a topological pipeline to study high dimensional data sets by extracting shapes and obtaining insights from them. The method works by using a filter function to partition data into overlapping bins. Next, single-linkage clustering (Jain & Dubes, 1988) is applied in each bin. Finally, a network is built by connecting the previously found clusters with an edge if they share one or more points in common. The authors applied the technique to analyze gene expression data, NBA players performances and the behavior of the US House of Representatives. Interesting insights were drawn, although the authors did not explore the use of persistent homology to find clusters with similar topological properties. The difficulty in their work is in finding the correct filter functions to appropriately bin the data.

Holzinger (2014) provides a good survey to the area and motivates the idea of topological data mining, especially in scenarios of complex, high-dimensional, heterogeneous and noisy data, in which structural and temporal patterns are often hidden. The author argues that computational geometry and algebraic topology may be of great help in facing those challenges. For instance, meaningful relationship networks of biological concepts, such as genes, proteins and drugs can be constructed with topological text data mining, such as in Chen and Sharp (2004).

Fasy et al. (2014) present an assessment of the uncertainty of persistence diagrams. The authors derive a confidence set for the persistence diagram $\mathcal{P}$ which corresponds to the distance function to a topological space $T$ using a sample from $\mathcal{P}$ supported by $T$. More concretely, the method adds a band around the diagonal of the persistent diagram, such that points in the band are considered noise. In our work we have not considered confidence sets, but compute features such as the maximum life-time of a point in $\mathcal{P}$, which already selects a point distant enough from the diagonal to be outside a confidence band. However, it would be interesting to add more features which take into account confidence bands in the future.

Glisse et al. (2014) determine convergence rates of persistent diagrams given certain conditions. They consider the persistent homology of different simplicial complexes built over a data set sampled from a probability distribution supported on an unknown compact metric space. They study the rate of convergence of the persistence diagrams to some well-defined persistence diagram associated to the support of the probability distribution. It is an important theoretical work that results in a bound for the probability of divergence between two persistent diagrams, such that it diminishes as the number of points in the metric space increase.

Li, Ovsjanikov, and Chazal (2014) present a framework for object recognition using persistent homology. They propose the use of persistence diagrams to serve as compact and informative descriptors for images and shapes. The authors conduct experiments on 3D shape retrieval, hand gesture recognition and text classification, obtaining good results in comparison to state-of-the-art methods. They consider the persistence diagrams computed over 3D meshes of objects, such as horses, dinosaurs and pliers. The PDs allow to distinguish the three classes since the diagrams of horses are more similar to each other than to those of other objects. Their idea is similar to ours, however they compute PD similarity by constructing a bipartite graph considering two PDs and computing the bottleneck distance, minimizing the weight of the edges connecting the graph. In our approach, we compute several features of the PDs and then cluster them using a standard clustering algorithm, such as K-Means, which gains in performance over their approach and can highlight different features of the PD, making it easier to find similarities.

Wei, Ge, Chattopadhyay, and Lobaton (2014) propose the use of persistence analysis to segment obstacles from disparity maps obtained from a stereo-vision system. Current state-of-the-art methods for obstacle detection rely on parameter values that have to be carefully selected to obtain good accuracy. Often, small changes in the parameters lead to large variations in the segmentation of images. The authors introduce a methodology that characterizes the sensitivity of a segmentation result based on persistent homology. The authors analyze how different parameter values lead to the birth and death of regions in the image by using persistent topology to quantify how long those regions last. Thus, the longer-lasting regions indicate with higher probability the obstacles. Their work is similar to ours, although they did not explore the various features we compute over the persistence diagram.

In the next section, we introduce background concepts from computational topology and our proposed approach.

## 3. Methodology

### 3.1. Background concepts on computational topology

We give a brief introduction to the notation and concepts surrounding computational topology and, in particular, persistent homology (Carlsson, 2009; Zhu, 2013). Formally, a topology on a set $X$ is a subset $T \subseteq 2^X$ such that: 1) if $S_1, S_2 \in T$, then $S_1 \cap S_2 \in T$; 2) if $\{S_j \mid j \in J\} \subseteq T$, then $\cup_{j \in J} S_j \in T$; and 3) $\emptyset, X \in T$. This implies that topology is simply a system of subsets that describe the connectivity of a set.

The pair $(X, T)$ of a set $X$ and a topology $T$ is a *topological space*. $\mathbb{X}$ is used as notation for a topological space $X$, with $T$ implicit. In data mining, we are used to metric spaces, specially Euclidean space. A metric space is a topological space, which is given by a set $X$ with a metric function $d$. Euclidean space can thus be defined as the Cartesian product of n copies of $\mathbb{R}$ along with the Euclidean metric: $d(x, y) = \sqrt{\sum_{i=1}^{n} (u_i(x) - u_i(y))^2}$, where $u_i$ is the $i$th Cartesian coordinate function, thus forming the $n$-dimensional Euclidean space $\mathbb{R}^n$.

In persistent homology, we ultimately want to compare topological spaces based on the characteristic holes that they encompass. Because we usually operate with finite point clouds in data analysis, we first need to discretize the space in order to add the notion of connectivity. That is done through the creation of simplicial complexes.

A $p$-simplex $\sigma$ is the convex hull of $p + 1$ linearly independent points $x_0, x_1, \ldots, x_p \in \mathbb{R}^d$ (Zhu, 2013). More intuitively, a 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle, a 3-simplex is a tetrahedron, and so forth. A simplicial complex $K$ is a finite set of simplices such that for $\sigma \in K$, all of its faces are also in $K$.

The core idea in persistent homology is to analyze how holes appear and disappear, as simplicial complexes are created. To do that, a filtration is constructed. An increasing sequence of $\epsilon$ values, i.e., distance values, produces a filtration, such that a simplex enters the sequence no earlier than all its faces. A common way to do this is using Vietoris–Rips complexes, that are only added to the filtration at $\epsilon = \epsilon'$ if the distance between two points in $\sigma$ is less than or equal to $\epsilon'$. This is useful because all simplexes that are present in $\epsilon'$ will also be contained in $\epsilon''$, if $\epsilon'' \geqslant \epsilon'$.

Since the number of generated simplicial complexes can be exponential in the number of input data points for a Vietoris–Rips filtration (Holzinger, 2014), we used in this study the Lazy Witness algorithm (Silva & Carlsson, 2004), which provides similar results. The advantage of the lazy version is that it does not use all points to compute the complexes, but only a smaller set of landmark points, while the remaining ones are used as witnesses to the edges or simplices spanned by combinations of landmarks. This subsampling gives enormous speed gains, while maintaining a good filtration quality, provided a suitable strategy is used to choose the landmark points. We used the max/min strategy proposed by Silva and Carlsson (2004) in our experiments, which has empirical evidence for its quality.

Consider the geometric shapes illustrated in Fig. 4. While the two circles (b) and (c) have the same basic equation and are geometrically closer together, to a topologist the empty square (a) and the empty circle (c) are the most similar objects, as one can be deformed into the other without gluing or cutting pieces. This

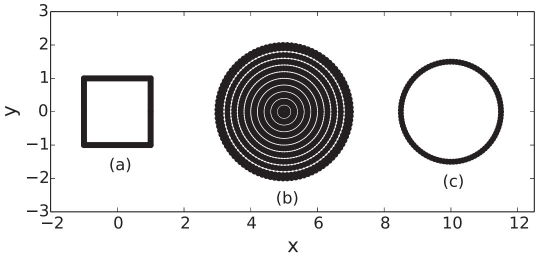


**Fig. 4.** Sample geometric shapes for a filtration analysis. While the two circles are geometrically closer, the empty square and the empty circle are actually the two closer objects in terms of their topological features, as one can be deformed into the other.

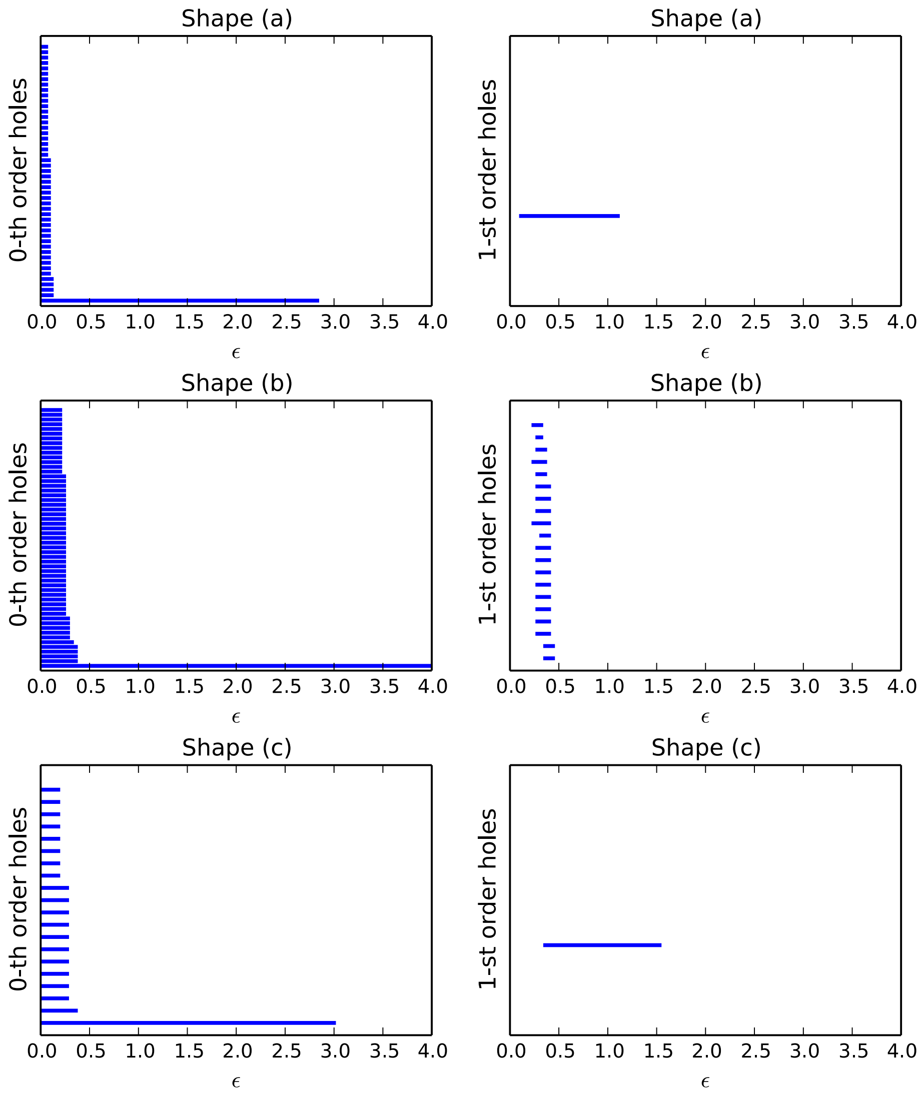


**Fig. 5.** Barcode plots for the geometric shapes of Fig. 4. The left plots indicate the output of the 0-th order holes filtrations, while the right ones of the 1-st order holes filtrations. By analyzing 1-st order holes, we clearly observe that (a) and (c) have more in common than with (b).

idea is clearly captured in persistent homology by computing a filtration for each object and displaying it as a barcode plot. Fig. 5 illustrates the 0th and 1st order barcode plots for the geometric shapes of Fig. 4. It is clearly visible that when examining 1st order holes, shapes (a) and (c) have a distinct hole that persists for a long time (a large range of $\epsilon$ values) in the filtration. On the other hand, the filled circle (b) has many small holes that are quickly filled. Analyzing the output of the filtrations allows us to find equivalence classes of those objects, in terms of their topological properties.

Another way to represent topological holes is through the persistence diagram (PD). Fig. 6 illustrates a hypothetical PD. Each point indicates the birth ($x$ axis) and death ($y$ axis) of an $n$-dimensional hole. Because we are interested in ignoring noise and concentrating on the most salient features of a data set, in persistent homology we look for holes that persist for a long time, or possibly for the entire filtration. These are easy to see in the PD, as they stand far away from the diagonal. Points that lie close to the diagonal indicate holes that appeared and then were soon closed, thus they may not represent important shape features of the data.

In this paper, we compute features based on information from the persistence diagram over various dimensions. Next, we briefly review time series analysis concepts that were also used in this research.

### *3.2. Background concepts on time series analysis*

Since we analyze time series data in our experiments, we briefly review concepts we used in this research. Because we want to extract topological information that is not readily available in a time series in its standard form, we use Takens' embedding theorem to extract the attractor of the series and perform the analysis on it.

Takens' embedding theorem (Takens, 1981) states that we can reconstruct a time series considering time delays. A series $x_0, x_1, \dots, x_{n-1}$ can be reconstructed in phase space as $x_n(m, \tau) = (x_n, x_{n+\tau}, \dots, x_{n+(m-1)\tau})$, in which $m$ is the embedding dimension and $\tau$ the time delay. Thus, instead of analyzing the series along time, we analyze its trajectory as a set of visited states in an $m$-dimensional Euclidean space. This is particularly useful to extract the topological behavior we are interested in.

Consider, for instance, the Henon map (Eqs. (5) and (6)), illustrated in Fig. 7(a) and the logistic map (Eq. (7)) in Fig. 7(b). Both present somewhat similar behaviors upon visual inspection. However, if we reconstruct their trajectories in phase space, we can clearly see that their behavior is very different, as shown in Fig. 8(a) and (b), respectively. The analysis of those attractors in phase space is particularly suited to our topological approach.

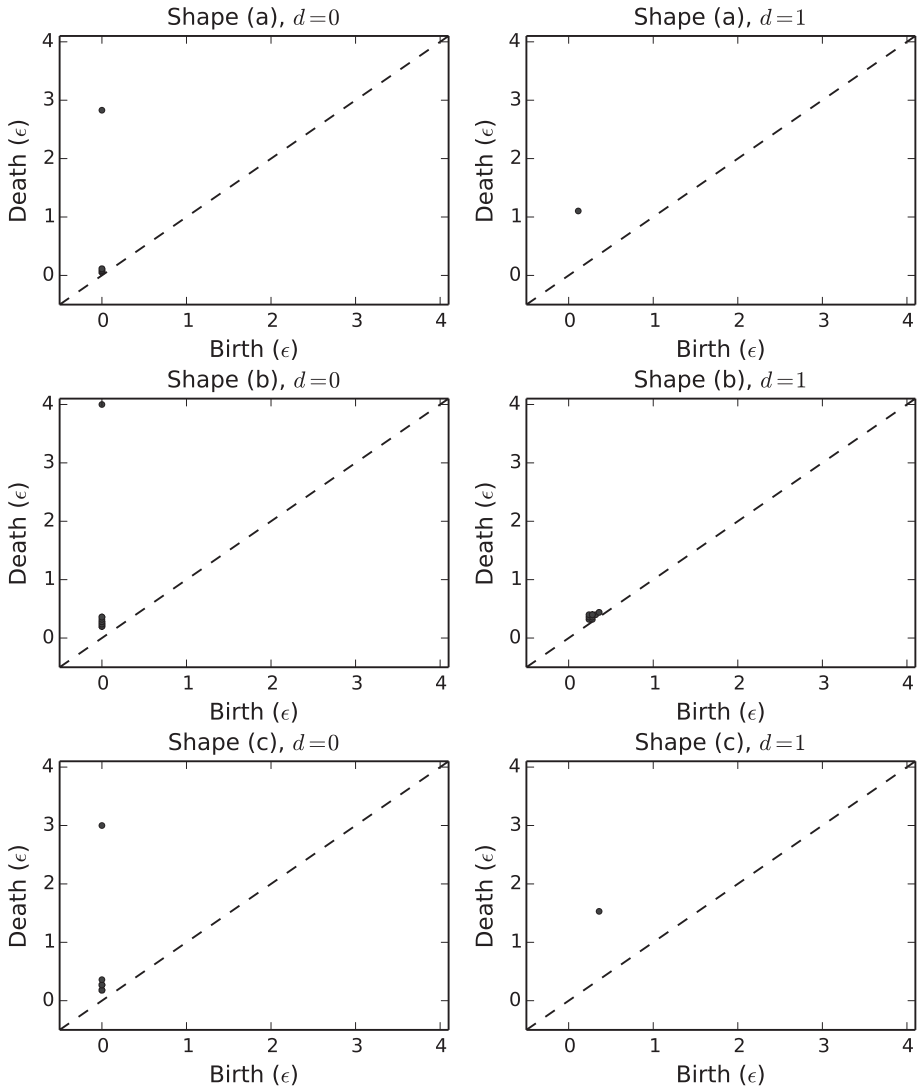


**Fig. 6.** Persistence diagrams (PDs) for the shapes of Fig. 4. Each point indicates the pair of birth and death $\epsilon$ values at which the hole appeared and was filled in the filtration. Points far away from the diagonal indicate more important shape-based features of the data set, as they persisted longer in the filtration.

$$x_{t+1} = 1.0 - ax_t^2 + y_t \tag{5}$$

$$y_{t+1} = bx_t \tag{6}$$

$$z_{t+1} = rz_t(1 - z_t) \tag{7}$$

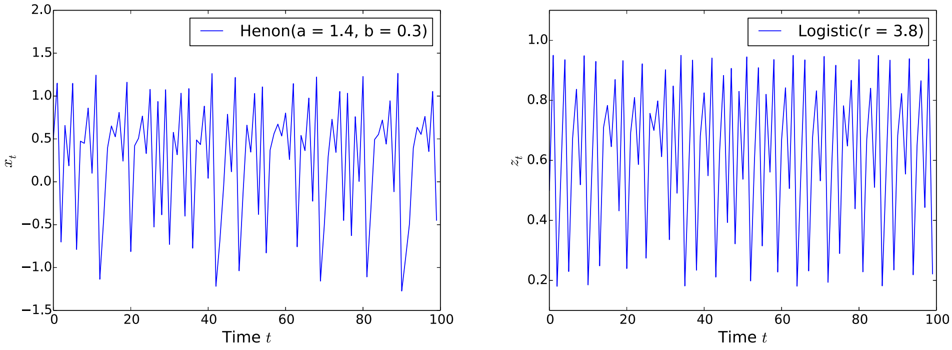
Another concept we use from time series analysis is that of the Empirical Mode Decomposition (EMD) (Huang et al., 1998). We use it in the context of noise removal, as it is able to separate the more stochastic components from the deterministic ones in a time series (Rios & de Mello, 2013). Empirical Mode Decomposition works by decomposing a signal into its intrinsic mode functions (IMFs), which can be summed to recover the original signal. To obtain the IMFs, first local extrema and zero crossings are identified. Next, lower and upper envelopes are computed by interpolating local extrema. Afterwards, the mean of the envelopes is calculated, producing a low frequency component, which is subtracted from the original signal. The resulting signal is the first IMF after certain criteria are met, as detailed in Huang et al. (1998).

To exemplify EMD, consider the sinusoidal signal shown in Fig. 9(a) and the same signal with white noise added in Fig. 9(b). Running EMD on the noisy signal (b), we obtain a sequence of IMFs and a residue series. Fig. 9(c) illustrates the sum of the last five IMFs plus the residue series. The way EMD works, the more stochastic components are contained in the first IMFs and the more deterministic ones in the last. One notes that high frequency components are removed in (c). It is important to observe that the use of EMD in this way can be useful if it is expected that high frequency components are the result of noise in the signal. However, for signals that are high-frequency in nature, using EMD as a low-pass filter is not appropriate.

### 3.3. Proposed methodology

We propose in this paper the use of topological features for data clustering. The first step in this process is in preprocessing the data set. In case it is composed of time series, we need to first analyze the level of noise or stochastic components it contains. If noise removal is warranted, then we can use the EMD technique to obtain a series composed of the more deterministic components.

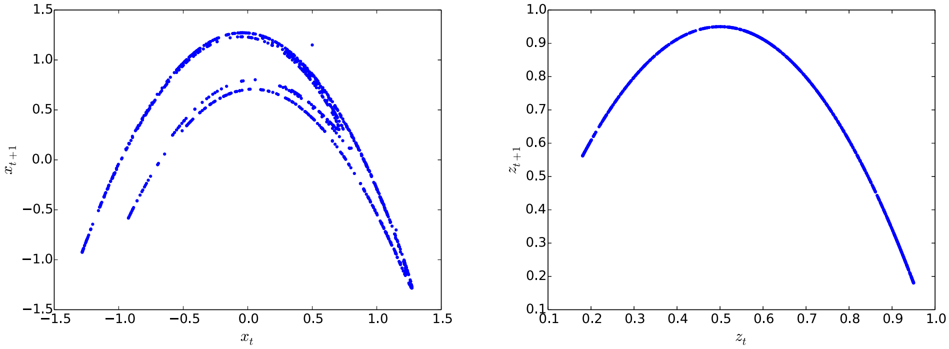
Next, we reconstruct the trajectory using Takens' theorem, to obtain the series attractor in phase space. This facilitates topological analysis because the series behavior is made explicit in the reconstruction, as states visited along time. In case we are dealing with spatial data, we do not need the EMD and reconstruction steps, as the trajectory data is already available.



(a) Henon time series.

(b) Logistic time series.

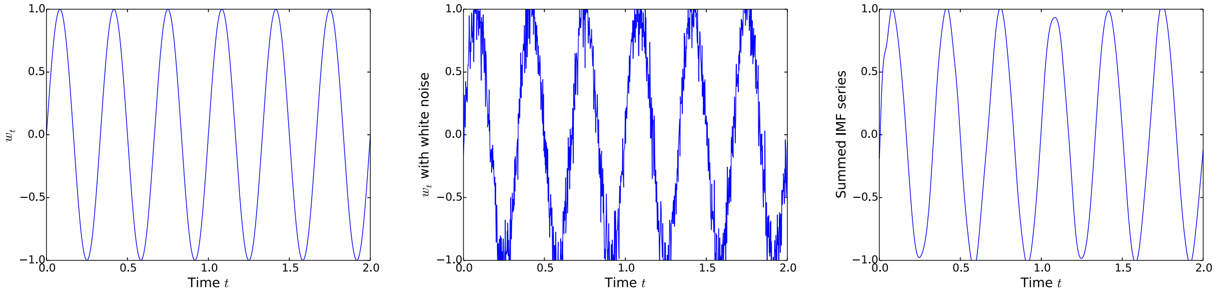
**Fig. 7.** Original time series.



(a) Reconstruction for Henon time series.

(b) Reconstruction for logistic time series.

**Fig. 8.** Reconstructed trajectories using $m = 2$ and $\tau = 1$.



(a) Original sinusoidal signal.

(b) The same signal as in (a) but with white noise added.

(c) Output series from EMD analysis on the noisy signal in (b). It is the sum of the last five IMFs plus residue.

**Fig. 9.** A sample time series and output of using EMD for noise removal.

After preprocessing steps have been carried out, the next phase is discretizing the space to obtain connectivity information. We carry out that task by computing a filtration for a sequence of $\epsilon$ values using Lazy Witness complexes (Silva & Carlsson, 2004). We select the maximum $\epsilon$ value using a heuristic that computes the maximum pointwise distance of a subsample of the original data set, which makes the estimation very fast.

Once the filtration is computed, which can be summarized in the persistence diagram, we compute a set of topological features to characterize the PD. First, we define the PD as a set $\Pi = \{\pi_0, \pi_1, \ldots, \pi_n\}, \pi_i = (b_i, d_i)$, in which $b_i$ indicates the birth $\epsilon$ value of point $i$ and $d_i$ the death value.

The first feature we compute is the *number of holes* for each dimension. This information is readily available from the filtration. We note that in the 0th dimension, the holes actually indicate connected components, i.e., vertices that were connected along the filtration.

The second feature is the *maximum hole lifetime* in each dimension $d$, as displayed in Eq. (8). This helps to identify the most significant hole in each dimension, i.e., the point farthest away from the diagonal in the PD. That is an important point, as it is the one that lasted the longest in the filtration it indicates a prominent shape-based feature of the data set.

$$max_d = max_{\pi_i \in \Pi}(d_i - b_i) \tag{8}$$

The third feature is the *number of relevant holes*, which we measure as the fraction of points that are at a similar distance as the farthest point from the diagonal. This is formalized in Eqs. (9) and (10).

Those are also relevant points, as they are also further away from the diagonal of the PD, indicating more prominent features of the data set.

$$nrel_d = \sum_{\pi_i \in \Pi} f(d_i - b_i, max_d, \text{ratio}) \quad (9)$$

$$f(\text{life}, max_d, \text{ratio}) = \begin{cases} 1, & \text{if life} \geqslant \text{ratio} \cdot max_d \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

The fourth feature is *the average lifetime of all holes* in dimension *d* (Eq. (11)). A small average indicates that for that dimension, the data set has mostly short-lived holes, i.e., small holes that were quickly filled during the filtration. The fifth feature is *the sum of all lifetimes* (Eq. (12)), which intends to compute the integral of the PD graph. A data set that sums close to zero in a particular dimension indicates that it has practically no holes in that dimension.

$$avg_d = \frac{\sum_{i=1}^{n}(d_i - b_i)}{n} \quad (11)$$

$$sum_d = \sum_{i=1}^{n}(d_i - b_i) \quad (12)$$

After all topological features are computed, we create a new data set composed of those shape-based features and use a standard clustering algorithm to obtain the final partition. The entire procedure is summarized in the diagram in Fig. 10.

## 4. Results

For our experiments, we have used the Lazy Witness Stream (Silva & Carlsson, 2004) as implemented in the JavaPlex library (Tausz, Vejdemo-Johansson, & Adams, 2011). For the step parameter in the filtration, which controls the amount $\epsilon$ increases in each iteration, we set it at 0.001. Our implementation was done in Java. For comparison purposes, we used as baseline the K-Means clustering algorithm implementation from Weka (Hall et al., 2009) on the raw data sets. We used K-Means++ as the initialization strategy to avoid poor initial centers selection.

Since the data sets we analyzed had two ground truth clusters, we present results as a confusion matrix composed of two clusters: a positive and a negative, as illustrated in Eq. (13), in which TP stands for true positives, FP for false positives, TN for true negatives and FN for false negatives.

$$CFM_{\text{experiment, technique}} = \begin{array}{ccl} \text{Cluster +} & \text{Cluster -} & \\ TP & FN & \text{Ground cluster +} \\ FP & TN & \text{Ground cluster -} \end{array} \quad (13)$$

We used the following scoring measures to determine agreement with the ground truth. See (Fawcett, 2006) for a discussion of those measures. *Sensitivity*, or *true positive rate* or *recall* (Eq. (14)), which measures the fraction of correct positive identifications from all positive cluster identifications that should have been made. Higher is better.

$$\text{Recall} = TP/(TP + FN) \quad (14)$$

*Specificity* or *true negative rate* (Eq. (15)), which measures the amount of correct negative identifications from all negative cluster identifications that should have been made. Higher is better.

$$\text{Specificity} = TN/(TN + FP) \quad (15)$$

*Precision*, or *positive predictive value* (Eq. (16)), which measures the amount of correct positive identifications from the total of positive identifications made. Higher is better.

$$\text{Precision} = TP/(TP + FP) \quad (16)$$

*Negative predictive value* (Eq. (17)), which measures the amount of correct negative identifications from all negative identifications made. Higher is better.

$$\text{Negative predictive value} = TP/(TP + FP) \quad (17)$$

*Fall-out* or *false positive rate* (Eq. (18)), which measures the amount of false positive identifications made from all negative identifications that should have been made, i.e., 1 − Specificity. Lower is better.

$$\text{Fall-out} = FP/(FP + TN) \quad (18)$$

*False discovery rate* (Eq. (19)) measures the amount of false positive identifications from all positive identifications that were made, i.e., 1 − Precision. Lower is better.

$$\text{False discovery rate} = FP/(FP + TP) \quad (19)$$

*Accuracy* (Eq. (20)), which measures the overall amount of correct identifications from all predictions that were made. Higher is better.

$$\text{Accuracy} = (TP + TN)/(TP + TN + FP + FN) \quad (20)$$

*F1 score* (Eq. (21)), which is a harmonic mean of precision and recall, tending to the lower of the two. Higher is better.

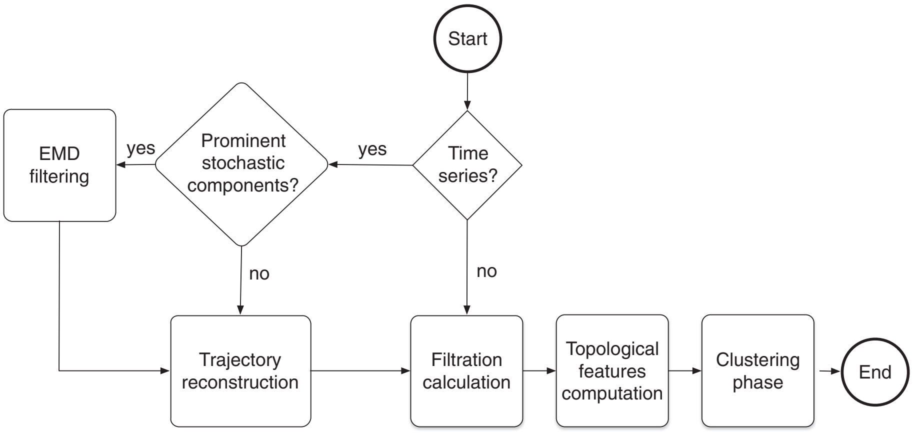
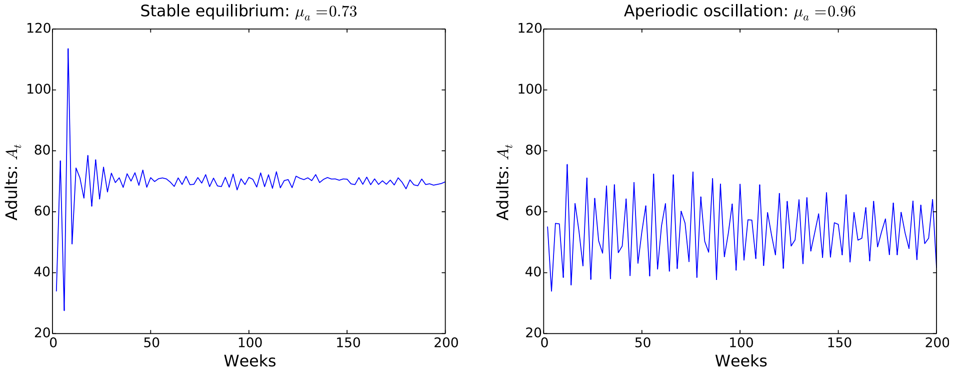

Start
Time series?
yes
Prominent stochastic components?
yes
EMD filtering
no
no
Trajectory reconstruction
Filtration calculation
Topological features computation
Clustering phase
End

**Fig. 10.** The entire proposed procedure, considering both time series and spatial data.



(a) Stable regime.

(b) Aperiodic oscillations regime.

**Fig. 11.** Examples of *Tribolium* SS strain adult population growth.

$$\text{F1 score} = 2TP/(2TP + FP + FN) \tag{21}$$

*Matthews correlation coefficient* (MCC) (Eq. (22)), which measures the overall amount of agreement between the predictions and the ground truth. A value of −1 indicates complete disagreement with the ground truth, while 0 indicates the equivalent of a random prediction and 1 a perfect agreement, thus higher positive values are better.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{22}$$

### *4.1. Population growth analysis of Tribolium flour beetle*

Flour beetles can cause significant costs to the food industry. Learning more about the population dynamics of those insects can provide insight into identifying explosive growth behaviors, which in turn can be used to preemptively act and avoid financial losses.

Researchers have modeled the population dynamics of the *Tribolium* flour beetle (Costantino, Cushing, Dennis, & Desharnais, 1995), which has a life-cycle starting as a larva for two weeks, then it enters a pupa stage lasting another two weeks before finally becoming an adult. *Tribolium* has a characteristic that, while under a regime of overpopulation, cannibalism happens, with adults eating pupae and unhatched eggs.

The population growth was modeled as a system of three difference equations (Eqs. (23)–(25)), in which $L_t$ is the number of feeding larvae, $P_t$ is the number of non-feeding larvae, pupae and callow adults, while $A_t$ is the number of mature adults at time $t$. The time unit is taken to be two weeks. The value $b > 0$ indicates the number of larval recruits per adult per unit of time when there is no cannibalism. Terms $\mu_\ell$ and $\mu_a$ are the probabilities of larvae and adults, respectively, of dying in the absence of cannibalism. Values $\exp(-c_{ea}A_t)$ and $\exp(-c_{e\ell}L_t)$ are the probabilities that an egg is not eaten in the presence of $A_t$ adults and $L_t$ larvae. Term $\exp(-c_{pa})$ indicates the survival probability of a pupa with $A_t$ adults. The terms $E_{1t}, E_{2t}$ and $E_{3t}$ model noise, due to environmental and other causes, as a multivariate normal distribution with means equal to zero.

$$L_{t+1} = b \cdot A_t \cdot \exp(-c_{ea}A_t - c_{e\ell}L_t + E_{1t}) \tag{23}$$

$$P_{t+1} = L_t(1 - \mu_\ell)\exp(E_{2t}) \tag{24}$$

$$A_{t+1} = (P_t \exp(-c_{pa}A_t) + A_t(1 - \mu_a))\exp(E_{3t}) \tag{25}$$

In our experiments, we have simulated the SS genetic strain of *Tribolium*, such that parameters were set as $b = 7.48, c_{ea} = 0.009, c_{pa} = 0.004, c_{e\ell} = 0.012, \mu_p = 0$ and $\mu_\ell = 0.267$, as modeled in Costantino et al. (1995). We varied the number of initial populations for larvae, pupae and adults by sampling an integer from 2 to 100. Since we simulated low level noise, EMD was not used. The lazy witness algorithm was set with a landmarks ratio of 1 landmark for 5 data set points.

Our intent with this study was to evaluate two regimes of adult *Tribolium* population growth under stable equilibrium at $\mu_a = 0.73$ and aperiodic oscillations at $\mu_a = 0.96$, as modeled in Costantino et al. (1995). Oscillatory behavior can be more dangerous, as overpopulation may be harder to predict, leading to worse financial losses. Fig. 11(a) and (b) present examples of stable and aperiodic oscillatory behaviors, respectively.

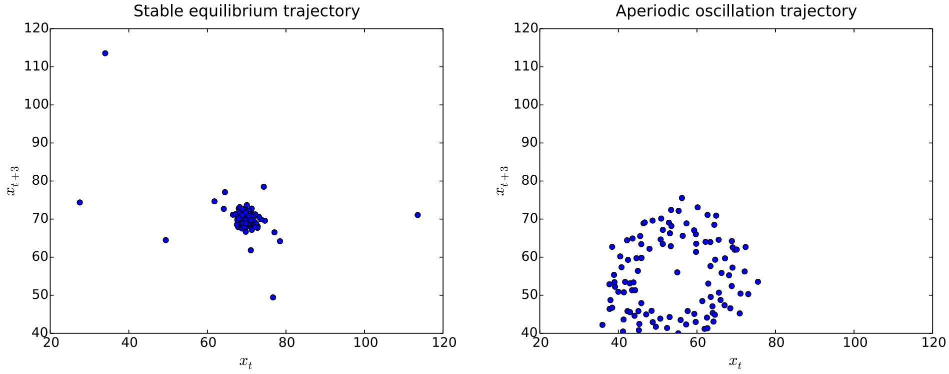
A total of 400 time series were simulated for a period of 240 weeks (2 weeks per time unit), with 200 series following the stable equilibrium regime and 200 following the aperiodic oscillation one. For baseline comparisons, we present results of running K-Means with $K = 2$ on the raw time series. The cluster-to-class assignments are presented in the confusion matrix of Eq. (26). We observe that correctly identifying aperiodic oscillatory behavior is not very successful using a traditional approach, with high number of false positives encountered, i.e., high recall but low precision.

Our hypothesis, that by reconstructing the trajectories of the time series in phase space and then computing topological properties of the attractors, was tested with K-Means clustering the final extracted topology features. Reconstruction in phase space was done with $\tau = 3$ and $m = 2$. The separation parameter $\tau$ was determined using the first minimum of the auto mutual information, as the recommended estimation strategy (Fraser & Swinney, 1986). The embedding dimension $m$ was set at 2 as using 3 dimensions did not substantially improve results. Fig. 12(a) and (b) present the trajectory reconstructions for the sample time series presented in Fig. 11(a) and (b).

The obtained confusion matrix is displayed in Eq. (27). A comparative overview with the baseline approach is listed in Table 1. One observes that our proposed approach maintains high recall while also providing high precision. The Matthews correlation coefficient indicates near 90% agreement with the external ground truth.

$$CFM_{Tribolium,\,\text{baseline}} = \begin{array}{cc} \text{Cluster +} & \text{Cluster -} & \\ \left(\begin{array}{c} 200 \\ 113 \end{array}\right. & \left.\begin{array}{c} 0 \\ 87 \end{array}\right) & \begin{array}{l} \text{Stable} \\ \text{Aperiodic} \end{array} \end{array} \tag{26}$$

$$CFM_{Tribolium,\,\text{proposed}} = \begin{array}{cc} \text{Cluster +} & \text{Cluster -} & \\ \left(\begin{array}{c} 200 \\ 25 \end{array}\right. & \left.\begin{array}{c} 0 \\ 175 \end{array}\right) & \begin{array}{l} \text{Stable} \\ \text{Aperiodic} \end{array} \end{array} \tag{27}$$



(a) Stable regime.

(b) Aperiodic oscillations regime.

**Fig. 12.** Trajectory reconstructions of the time series in Fig. 11. We observe that the aperiodic trajectory tends to be more dispersed and surround a one-dimensional hole. The stable regime, on the other hand, tends to be more densely packed with a few outliers, which correspond to points before equilibrium was reached.

**Table 1**
Comparative results for the *Tribolium* study.

| Measure | Score | |
|---|---|---|
| | Proposed | Baseline |
| Recall (or sensitivity or true positive rate) | 1.00 | 1.00 |
| Specificity (true negative rate) | 0.88 | 0.43 |
| Precision (positive predictive value) | 0.89 | 0.64 |
| Negative predictive value | 1.00 | 1.00 |
| Fall-out (false positive rate) | 0.12 | 0.56 |
| False discovery rate | 0.11 | 0.36 |
| Accuracy | 0.94 | 0.72 |
| F-1 score | 0.94 | 0.78 |
| Matthews correlation coefficient (+1 perfect, 0 random, −1 total disagreement) | 0.88 | 0.53 |

### 4.2. Animal tracking

In this experiment, we used GPS latitude/longitude data gathered from spatial tracking of birds, specifically vultures (Dodge et al., 2014) and pigeons (Santos, Neupert, Lipp, Wikelski, & Dechmann, 2014). The data was obtained from the Movebank Data Repository.[1]

We wanted to evaluate in this experiment the hypothesis that animals of the same species have closer movement behavior than with animals from other species. We used spatial data from 18 turkey vultures and 10 pigeons, in a total of 3200 latitude/longitude GPS coordinates per animal. Fig. 13(a) and (b) present sample tracking data for one vulture and one pigeon, respectively. We can observe that the behavior for those two examples is indeed different. Since we want to test for behavior, the GPS coordinates were normalized in the $[0, 1]$ range, otherwise we would obtain a clustering simply based on the animals geographical location.

Running K-Means with $K = 2$ on the raw coordinates results in the confusion matrix shown in Eq. (28). We observe that there is almost an even spread of animals assigned per cluster per class.

For our proposed approach we set the lazy witness algorithm to use a ratio of 1 landmark per 100 points, since there is a large number of coordinates per animal we can use a smaller fraction of the original points. The results are presented in the confusion matrix in Eq. (29). We see that there is a clearer distinction among clusters, with no pigeons identified as vultures. Table 2 presents a comparative overview of the accuracy of our proposed approach with the baseline. We observe that while the baseline has a Matthews correlation coefficient close to what would be achieved by chance (0.22), our proposed approach has much better agreement with the actual external ground truth (0.65).

In this experiment trajectory reconstruction was not needed, since we already had the spatial trajectories visited by the animals. The experiment also illustrates the point that trajectory reconstruction is not sufficient to obtain a good clustering, as shown by the results of the baseline approach that clustered the GPS coordinates directly. The topology step had a direct impact in improving the results, as it can better extract morphological, or qualitative structure from the data set.

$$CFM_{\text{animal tracking, baseline}} = \begin{array}{cl} \begin{array}{cc} \text{Cluster +} & \text{Cluster -} \end{array} & \\ \begin{pmatrix} 5 & 5 \\ 5 & 13 \end{pmatrix} & \begin{array}{l} \text{Pigeons} \\ \text{Vultures} \end{array} \end{array} \tag{28}$$

$$CFM_{\text{animal tracking, proposed}} = \begin{array}{cl} \begin{array}{cc} \text{Cluster +} & \text{Cluster -} \end{array} & \\ \begin{pmatrix} 10 & 0 \\ 6 & 12 \end{pmatrix} & \begin{array}{l} \text{Pigeons} \\ \text{Vultures} \end{array} \end{array} \tag{29}$$

### 4.3. Electroencephalography of alcoholic subjects

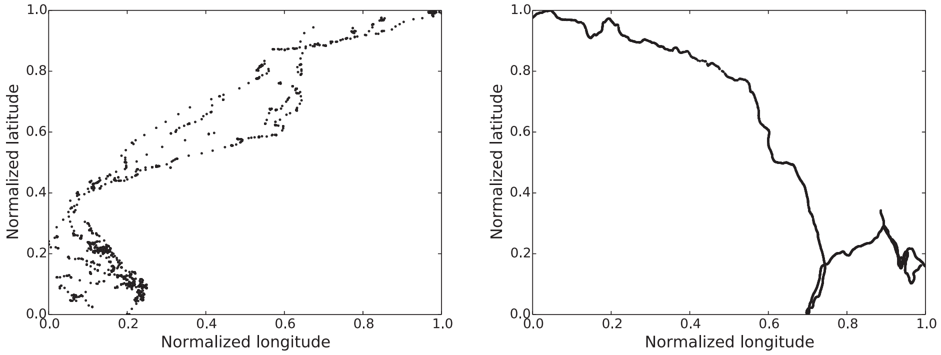
In this experiment we used electroencephalography data collected by Henri Begleiter at the Neurodynamics Laboratory at the State University of New York Health Center at Brooklyn and made available at the UCI machine learning repository.[2]

The data comes from a study of EEG correlates of genetic predisposition to alcoholism. It contains measurements from 64 electrodes placed on subject's scalps with a sampling rate of 256 Hz. There were two groups of subjects: alcoholic and control. Each subject was given a single stimulus (S1) or two stimuli (S1 and S2), which were pictures of objects chosen from the 1980 Snodgrass and Vanderwart picture set. When two stimuli were shown, they were presented in either a matched condition where S1 was identical to S2 or in a non-matched condition where S1 differed from S2.

We used the data set containing 10 alcoholic and 10 control subjects, resulting in a total of 600 time series, with 300 from alcoholic subjects and 300 from control subjects. Upon inspection, we selected time series gathered from the CPZ sensor, which corresponds to the midline sensor between the central and parietal lobes. Fig. 14(a) and (b) present sample time series from one control and one alcoholic subjects respectively, while Fig. 15(a) and (b) present the trajectory reconstructions for those time series. In this experiment we used EMD filtering by summing the last three

[1] Available at https://www.movebank.org

[2] Available at https://archive.ics.uci.edu/ml/datasets/EEG+Database



(a) Vulture.

(b) Pigeon.

**Fig. 13.** Sample trajectories of a vulture and a pigeon for the animal tracking experiment.

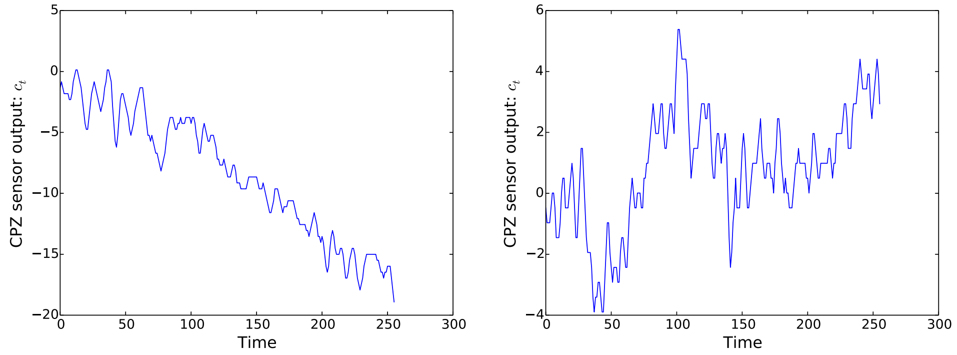
**Table 2**
Comparative results for the animal tracking study.

| Measure | Score | |
|---|---|---|
| | Proposed | Baseline |
| Recall (or sensitivity or true positive rate) | 1.00 | 0.5 |
| Specificity (true negative rate) | 0.67 | 0.72 |
| Precision (positive predictive value) | 0.62 | 0.5 |
| Negative predictive value | 1.00 | 0.72 |
| Fall-out (false positive rate) | 0.33 | 0.28 |
| False discovery rate | 0.38 | 0.5 |
| Accuracy | 0.79 | 0.64 |
| F-1 score | 0.77 | 0.5 |
| Matthews correlation coefficient (+1 perfect, 0 random, $-1$ total disagreement) | 0.65 | 0.22 |

intrinsic mode functions with the residue. The embedding was done using $m = 2$ and $\tau = 10$, determined empirically.

The results of running baseline K-means with $K = 2$ on the raw time series are presented in the confusion matrix of Eq. (30). We observe that there is a large number of incorrectly clustered instances for both alcoholic and control subjects.

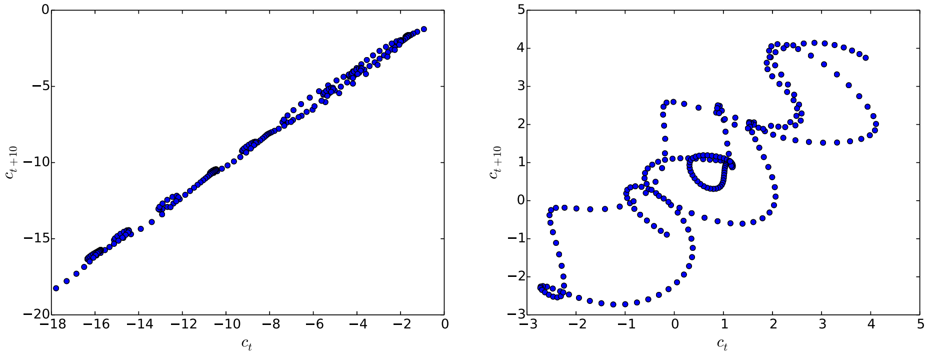
We set the number of landmarks for the lazy witness algorithm at the ratio of 1:10. The results for our proposed approach are presented in the confusion matrix of Eq. (31) and the comparative Table 3. We observe that while the baseline approach achieves exactly what would be expected by random guessing, according to the Matthews correlation coefficient, our proposed approach achieves a score of 0.9, or near perfect agreement with the external ground truth.



(a) Control subject.

(b) Alcoholic subject.

**Fig. 14.** Sample CPZ time series from one control and one alcoholic subjects.



(a) Control subject.

(b) Alcoholic subject.

**Fig. 15.** Trajectory reconstructions for the time series of Fig. 14, after EMD filtering was applied.

**Table 3**
Comparative results for the electroencephalography study.

| Measure | Score | |
|---|---|---|
| | Proposed | Baseline |
| Recall (or sensitivity or true positive rate) | 0.99 | 0.68 |
| Specificity (true negative rate) | 0.91 | 0.32 |
| Precision (positive predictive value) | 0.91 | 0.5 |
| Negative predictive value | 0.99 | 0.5 |
| Fall-out (false positive rate) | 0.09 | 0.68 |
| False discovery rate | 0.09 | 0.5 |
| Accuracy | 0.95 | 0.5 |
| F-1 score | 0.95 | 0.58 |
| Matthews correlation coefficient (+1 perfect, 0 random, −1 total disagreement) | 0.90 | 0.00 |

$$CFM_{\text{EEG, baseline}} = \begin{array}{cc} \text{Cluster +} & \text{Cluster -} & \\ \left(\begin{matrix} 203 \\ 203 \end{matrix}\right. & \left.\begin{matrix} 97 \\ 97 \end{matrix}\right) & \begin{matrix} \text{Control} \\ \text{Alcoholic} \end{matrix} \end{array} \tag{30}$$

$$CFM_{\text{EEG, proposed}} = \begin{array}{cc} \text{Cluster +} & \text{Cluster -} & \\ \left(\begin{matrix} 296 \\ 28 \end{matrix}\right. & \left.\begin{matrix} 4 \\ 272 \end{matrix}\right) & \begin{matrix} \text{Control} \\ \text{Alcoholic} \end{matrix} \end{array} \tag{31}$$

## 5. Discussion and conclusions

We have proposed in this paper an approach to clustering time series and spatial data based on topological features computed over a discretization of the space. The main idea is that persistent homology allows for a very flexible and multi-resolution analysis of the most prominent features of a data set via filtration. The most persistent features, in this case $n$-dimensional holes, are indicated by points that lie far away from the diagonal in the persistence diagram. Features computed over those points can be used for obtaining a clustering based on more intrinsic, or qualitative aspects of a data set.

The main contribution of this paper is a framework for clustering time-series and spatial data based on topological properties, which can correctly identify qualitative aspects of a data set currently missed by traditional distance-based techniques. The main advantages are that our technique can detect similarities in recurrent behavior for time series data sets and spatial structures in spatial data sets. This is especially important for fields in which spatial data has meaning, such as in protein analysis.

We showed that the proposed approach had successful results in three different application scenarios: population growth of flour beetles, spatial tracking of animals and EEG analysis of alcoholic subjects. This indicates that for various fields, the analysis of qualitative structure, or shape information, in the form of n-dimensional holes as studied in persistent homology, can provide interesting insights with the extraction of meaningful clusters.

The main lesson learned from this research is that qualitative features, such as spatial arrangements or recurrent behavior in time series attractors, can be successfully extracted via persistent homology and applied for clustering purposes. Our applications and examples showed that traditional distance-based techniques disregard qualitative information, as they do not focus not on the shape of data. While many techniques can, for instance, successfully distinguish two types of cancer on an imaging data set, we argue that it may be more useful at first to identify qualitative aspects that can lead an expert to detect a new type of disease.

Our technique has the additional advantage of being very fast to run for low dimensional spaces, in terms of seconds for the data sets we analyzed. The bottleneck is in computing the filtration, which needs to discretize the space to obtain connectivity information. In this sense, one limitation resides in the number of simplicial complexes that may be generated, which can be exponential in the number of points in the data set. The lazy witness algorithm, which we used in this study, is already a stride in the direction of solving this problem, as it uses the concept of landmarks that have strong data support (neighboring witnesses) to reduce the number of input points required for the discretization step. More research is still needed for efficient computation when the dimensionality is $d > 3$.

We believe more research into applying mathematical topology concepts can be very useful for several areas of data mining, in which shape has functional meaning. One particular field of study that we believe would benefit from this approach is the analysis of time series attractors, as we showed in this paper, since recurrent behavior can be captured well with persistent homology analysis. Future work could also focus on using confidence bands as proposed by Fasy et al. (2014) in order to compute topological features that ignore noise points in the persistence diagram. Another possibility is to modify our proposal to work on data streams. We are currently working on such approach, which considers objects changing over time. Concept drift may be thus detected by analyzing how topological properties evolve over time, using the tools of persistent homology to compute persistence diagrams and then evaluating how they change.

## Acknowledgment

This work has been funded by FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) under Grants #2013/04453-0 (Cássio), #2014/13323-5 (Rodrigo) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) under Grant #303051/2014-0 (Rodrigo).

## References

Bishop, C. M. (2006). *Pattern recognition and machine learning (information science and statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society, 46*(2), 255–308.

Chang, H.-W., Bacallado, S., Pande, V. S., & Carlsson, G. E. (2013). Persistent topology and metastable state in conformational dynamics. *PloS One, 8*(4), e58699.

Chazal, F., Guibas, L. J., Oudot, S. Y., & Skraba, P. (2013). Persistence-based clustering in riemannian manifolds. *Journal of the ACM, 60*(6), 1–38.

Chen, H., & Sharp, B. M. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics, 5*, 147.

Chiang, I.-J. (2007). Discover the semantic topology in high-dimensional data. *Expert Systems with Applications, 33*(1), 256–262.

Costantino, R. F., Cushing, J. M., Dennis, B., & Desharnais, R. A. (1995). Experimentally induced transitions in the dynamic behaviour of insect populations. *Nature, 375*(6528), 227–230.

Dodge, S., Bohrer, G., Bildstein, K., Davidson, S. C., Weinzierl, R., Bechard, M. J., et al. (2014). Environmental drivers of variability in the movement ecology of turkey vultures (Cathartes aura) in North and South America. *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1643), 20130195.

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., & Singh, A. (2014). Confidence sets for persistence diagrams. *The Annals of Statistics, 42*, 2301–2339.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Fraser, A. M., & Swinney, H. L. (1986). Independent coordinates for strange attractors from mutual information. *Physical Review A, 33*(2), 1134–1140.

Glisse, M., Labru, C., Bourgogne, D., Michel, B., Chazal, F., Fr, I., & Fr, U. (2014). Convergence rates for persistence diagram estimation in topological data analysis. In *31st international conference on machine learning* (Vol. 32, pp. 163–171).

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2009). The WEKA data mining software: an update. *SIGKDD Exploration Newsletter, 11*(1), 10–18.

Holzinger, A. (2014). *Interactive knowledge discovery and data mining in biomedical informatics. Lecture Notes in Computer Science* (vol. 8401). Berlin, Heidelberg: Springer Berlin Heidelberg.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., et al. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 454*(1971), 903–995.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Li, C., Ovsjanikov, M., & Chazal, F. (2014). Persistence-based structural recognition. In *2014 IEEE conference on computer vision and pattern recognition* (pp. 2003–2010).

Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., et al. (2013). Extracting insights from the shape of complex data using topology. *Scientific Reports, 3*, 1236.

Rios, R. A., & de Mello, R. F. (2013). Improving time series modeling by decomposing and analyzing stochastic and deterministic influences. *Signal Processing, 93*(11), 3001–3013.

Sacan, A., Ozturk, O., Ferhatosmanoglu, H., & Wang, Y. (2007). LFM-Pro: A tool for detecting significant local structural sites in proteins. *Bioinformatics (Oxford, England), 23*(6), 709–716.

Santos, C. D., Neupert, S., Lipp, H.-P., Wikelski, M., & Dechmann, D. K. N. (2014). Temporal and contextual consistency of leadership in homing pigeon flocks. *PloS One, 9*(7), e102771.

Silva, V. D. & Carlsson, G. (2004). Topological estimation using witness complexes. In *Proceedings of the first eurographics conference on point-based graphics* (pp. 157–166). Aire-la-Ville, Switzerland. Eurographics Association.

Stillwell, J. (1993). *Classical topology and combinatorial group theory*. *Graduate Texts in Mathematics* (Vol. 72). New York, NY: Springer New York.

Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980* (pp. 366–381). Springer.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Tausz, A., Vejdemo-Johansson, M., & Adams, H. (2011). JavaPlex: A research software package for persistent (co)homology. *Software*.

Warren Liao, T. (2005). Clustering of time series data-a survey. *Pattern Recognition, 38*(11), 1857–1874.

Wei, C., Ge, Q., Chattopadhyay, S., & Lobaton, E. (2014). Robust obstacle segmentation based on topological persistence in outdoor traffic scenes. In*2014 IEEE symposium on computational intelligence in vehicles and transportation systems (CIVTS)*, number c (pp. 92–99).

Zhu, X. (2013). Persistent homology: An introduction and a new text representation for natural language processing. In *Proceedings of the 23rd IJCAI, IJCAI'13* (pp. 1953–1959). AAAI Press.

Zomorodian, A. J. (2005). *Topology for computing (cambridge monographs on applied and computational mathematics)* (1 ed.). Cambridge University Press.