Applications of the Topological Data Analysis in Real Life

S. Z. RIDA[1], ALAA HASSAN NORELDEEN[2], FATEN. R. KARAR[2]
[1]Mathematics Department, Faculty of Science, South Valley University, Qena, EGYPT
[2]Mathematics Department, Faculty of Science, Aswan University, Aswan, EGYPT

Abstract: Statistical topology inference is a branch of algebraic topology that analyzes the geometric structure's global topological properties underlying a point cloud dataset. There is an increasing need to analyze massive data sets and screen large databases to address real-world problems. A central challenge to modern applied mathematics is the need to generate tools to simplify the data in high dimensional order to extract the important features or the relationships while performing the analysis. A growing field of study at the intersection of algebraic topology, computational geometry, and statistics is topological data analysis (TDA) inference. This study applies TDA tools to test hypothesis between two high-dimensional data sets. Hypothesis testing is one of the most important topics of statistical topology inference. A proposed test was created, which was built on the nearest-neighbor function.

Three tests such as (Hypothesis testing based on persistent homology, hypothesis testing based on persistent landscapes, and hypothesis testing based on density estimation) based on TDA, are discussed. Moreover, a modification of these tests was proposed. Monte Carlo simulation was conducted to compare the power of the previous tests. We displayed the use of TDA tools in hypothesis testing. It was proposed that this test might be established based on the nearest neighbor distance function. Furthermore, a suggested modification for the present tests based on TDA was introduced. Finally, the tests specified in the vignette were enabled by two empirical applications within the biology field. We demonstrated the efficacy of the above tests on the heart disease dataset from Statlog and the Wisconsin breast cancer dataset.

Keywords: Persistent homology; Topological data; Density estimation; Betti numbers; Persistent landscapes; Hypothesis testing

2010 Mathematics Subject Classification: 55N35, 55U05, 62H99.

Received: August 31, 2022. Revised: October 1, 2022. Accepted: October 20, 2022. Available online: November 28, 2022.

1 Introduction

Topology, a mathematics field that arises in an attempt to describe global features of space using point cloud data, can provide new insights and tools for finding and quantifying interclass relationships. Computational topology is particularly useful for understanding non-practical data using standard statistical methods (e.g., canonical correlation, principal component analysis, and hierarchical clustering).

Topology data analysis combines techniques and tools that allow academics to discover and analyze topological data for invariant structures, [1].

Those processes often use point cloud data as an input, commonly represented as a huge finite dataset in an n-dimensional metric space taken from a geometrical object, perhaps with some noise. The result is a set of data analyses and diagrams required to evaluate the statistical properties of the data accurately.

2 Simplicial Complexes

Simplicial complexes are used as the prime data structure to represent topological spaces. Graphs are commonly employed in many data analysis applications since they store relationships between data points. Simplicial complexes generalize the notion of graphs by allowing for 2, 3, and higher dimensional building blocks, called simplices.

2.1 Definition

E^n is an n-dimensional Euclidean space. Point cloud data (PCD) is an unordered sequence of points $S = \{x_1, \dots, x_4\}$ embedded in E^n.

A simplicial complex on PCD is defined by considering each point in the metric space as a vertex of an approximation. An edge connects two vertices based on their proximity. Higher-dimensional simplices can then be defined on the approximations in different ways. One of the most commonly used complexes is the Vietoris-Rips complex. To convert data from PCD into a metric

space (X, d), we use the point cloud (PC) as vertices of the approximation whose edges are determined by proximity using vertices within a specified $\varepsilon > 0$ through a distance metric d, which satisfies the following conditions, for all data points x, y, and z we have:

$$d(x,y) \geq 0, \quad d(x,y) = d(y,x)$$

and

$$d(x,z) \leq d(x,y) + d(y,z).$$

A simplicial complex is a finite collection of simplices K such that for any face $\sigma \in K$ and $\tau < \sigma$ implies $\tau \in K$, and $\forall \sigma_i$ and $\sigma_j \in K$ implies $\sigma_i \bigcap \sigma_j$ is either empty or a face of both. For example, 0-simplex consists of a point, a 1-simplex consists of a line segment, a 2-simplex consists of a triangle, and a 3-simplex consists of a tetrahedron (Fig. 1). A lot more can be extracted about the simplex, but for our need for the simplicial complex, this will be satisfactory, [2].

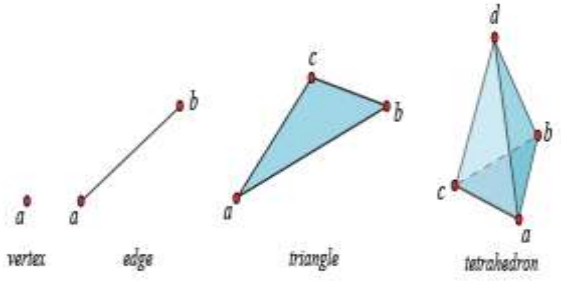

Fig. 1: Shapes of K-simplices for $k = 0,1,2,3$, respectively.

Indeed, different ways can be employed to filter the simplicial complex, such as Čech Lazy Witness. The typical difficulty associated with any filter method is choosing the suitable ε to give a decent approximation to the structure underlying the point cloud, as for ε sufficiently small, the complex is a discrete set; for ε sufficiently large, the complex is a single high-dimensional simplex. There are many notions of distance functions that one can reasonably use to obtain the Vietoris-Rips, [3], such as:

$$d_p(x,y) = \sqrt[p]{\sum |x-y|^p}\ .$$

3 Homology and Betti Numbers

Homology groups identify holes and loops indirectly by examining the space around them, but Betti numbers allow counting the number of different loops and holes. We begin building the homology groups by examining structured sums of simplices, creating an abelian group, [4].

3.1 Definition

The boundary $\partial_P \sigma$ of a p-simplex $\sigma = [u_0, u_1, .., u_p]$ is defined as the formal sum of its $(p-1)$ dimensional faces:

$$\partial_p \sigma = \sum_{j=0}^{p} (-1)^j [u_0, .., \widehat{u_j}, u_p], \qquad (1)$$

where $\widehat{u_j}$ represents a point that is not included in the simplex. That is for any verticesx_j: $\partial_0[x_j] = 0$ for any 1-simplex: $\partial_1[x_0, x_1] = [x_1] - [x_0]$, for any 2-simplex: $: \partial_2[x_0, x_1, x_2] = [x_1, x_2] - [x_0, x_2] + [x_0, x_1]$. In general, $\partial_p: \Delta_p \rightarrow \Delta_{p-1}$.

We may naturally extend the above definition to p-chains by specifying p -chain's boundary:

$c = \sum_{i \in I} a_i \sigma_i$ as $\partial c = \sum_{i \in I} a_i \partial \sigma_i$. We can establish a family of boundary homomorphisms. ∂_P connecting the various groups of -chains of a simplicial complex by mapping -simplices to their boundaries:

$$\cdots \xrightarrow{\partial_{p+2}} C_{p+1} \xrightarrow{\partial_{p+1}} C_p \xrightarrow{\partial_p} C_{p-1} \xrightarrow{\partial_{p-1}} \cdots \xrightarrow{\partial_1} C_0.$$

By construction, we have the property,

$$\partial_P (C + \acute{C}) = \partial_P C + \partial_P \acute{C}$$

Thus, ∂_P is indeed a homomorphism. This type of chain sequence C_P and homomorphisms ∂_P is defined as a chain complex, denoted by $C = (C_P, \partial_P)$.

We define the boundary $\partial_k \sigma$ of a *K*-simplex as the sum of its *(k-1)*-dimensional faces, which can be expressed as

$$\partial_k \sigma = \sum_{j=0}^{k} (-1)^j [x_0, x_1, \ldots, x_j, \ldots x_k],$$

where, x_j represents a point that is not included in the simplex, which means that for any vertices x_j: $\partial_0[x_j] = 0$, for any 1-simplex:$\partial_1[x_0, x_1] = [x_1] - [x_0]$, for any 2-simplex: $\partial_1[x_0, x_1, x_2] = [x_1, x_2] - [x_0, x_2] + [x_0, x_2]$. In general, $\partial_k \sigma: \Delta_k \longrightarrow \Delta_{k-1}$.

Hence, the *k-th* homology group $\mathcal{H}_k$ can be formulated as follows:

$$\mathcal{H}_k = \frac{\mathrm{Ker}\partial_k}{\mathrm{im}\partial_{k+1}},$$

where, $Ker\partial_k$ and $im\partial_{k+1}$ are donated to the kernel and the image of the boundary operator, respectively. One can easily prove that $\partial_k \partial_{k+1} = 0$, and $\mathrm{im}\,\partial_{k+1} \subset \mathrm{Ker}\partial_k$. Betti numbers are an important feather linked with the homology group because they convey relevant information about the complex. The k^{th} Betti number $\beta_k$represents the number of k^{th} dimensional independent holes in $\mathcal{H}_k$, so the number of connected components of

$\mathcal{H}_0$ is denoted as β_0, the number of loops is denoted as β_1, the number of enclosed voids is denoted as β_2, (Fig. 2). Generally, β_k can be computed as follows:
$\beta_k = \text{rank}(\mathcal{H}_k) = \text{rank}(\text{Ker}\partial_k) - \text{rank}(\text{im}\partial_{k+1})$,
since $\text{im}\partial_{k+1} \subset \text{Ker}\partial_k$, thus $\beta_k \geq 0\ \forall k > 0$.

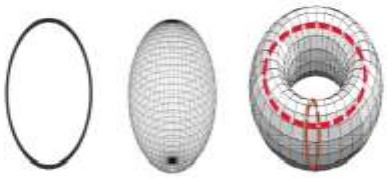

Fig. 2: The circle has $\beta_0 = 1, \beta_1 = 1$,
The sphere has$\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$,
The torus has $\beta_0 = 1, \beta_1 = 2\ , \beta_2 = 1$.

4 Persistent Homology

The concept of persistence homology was developed, [5]. The main idea of persistence indicates the topological characteristics, which persist over a considerable parameter range to be a signal feature. Short-lived characteristics, on the other hand, can be ignored as noise. Using the persistence homology, one can avoid choosing a single ε. Alternatively, we have to define the interval of ε for which that feather occurs. In other words, persistence homology is the method for studying homology at multiple scales simultaneously.

To realize how the persistence homology works, assume that we have a sequence for Vietoris-Rips complexes $V_i(\varepsilon)\,_{i=1}^{T}$ corresponding to the rising sequence of parameters $\varepsilon_i\,_{i=1}^{T}$. A chain of inclusion maps exists as follows:

$$V_1(\varepsilon) \subset V_2(\varepsilon) \subset \cdots \subset V_T(\varepsilon)$$

Instead of examining the homology of the individual terms $V_i(\varepsilon)$, one examines the homology of the iterated inclusions $V_i(\varepsilon) \hookrightarrow V_j(\varepsilon)\ 1 < i < j < T$. These chains reveal which features have long persistence intervals. As the $\mathcal{H}_{k,i}$ is born at a time i if $\mathcal{H}_{k,j}$ isn't in the inclusion image before the time i, whereas $\mathcal{H}_{k,j}$ dies entering time j if $\mathcal{H}_{k,j}$ isn't supported by the inclusion map $V_j(\varepsilon) \hookrightarrow V_{j+1}(\varepsilon)$. The birth at i and death at j of $\mathcal{H}_{k,i\hookrightarrow j}$ are recorded in the persistence diagram as ordered tuples points (i, j).

The 18 uniform randomly generated points are shown in (Fig. 3). We can observe from the figure that at $\varepsilon = 5$, ten points are only connected. However, at $\varepsilon = 7$, almost all the points are connected, giving birth to a circular hole. At $\varepsilon = 9$, the Vietoris-Rips complexes are finished.

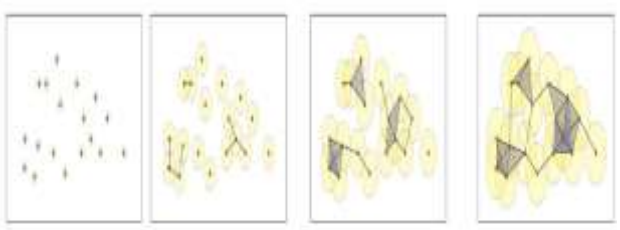

Fig. 3: Example of the persistent homology using a Vietoris-Rips complex at $\varepsilon = 0, \varepsilon = 5,\ \varepsilon = 7, \varepsilon = 9$, respectively.

5 Persistent Landscapes

Persistent Landscapes, produced by Bubenik, [6], can be considered as a diagram summarizing the data contained differently on the persistence diagram. The basic usage of the persistent landscapes enables us to summarize and compute data and use traditional statistics indicators, such as averages, median, and variance, instead of a barcode plot or a persistence diagram. Persistent landscapes may be considered a rotational version of a barcode diagram. To formulate the persistent landscapes diagram, begin by building a triangle whose base relates to a generalized persistence interval (i, j) and with a top vertex at the intersection of the vertical line going through the midpoint $(\frac{i+j}{2}, 0)$ and the circle passing through the endpoints, centered at the midpoint. Consequently, an isosceles right triangle is formed with the catheter intersecting at $(\frac{i+j}{2}, \frac{j-i}{2})$. Furthermore, Bubenik proposed that the persistent landscapes descriptor $\lambda_s(\varepsilon)$ is dependable during the statistical analysis and comparisons study. To obtain $\lambda_s(\varepsilon)$, it is required first to compute $\Lambda_s(\varepsilon)$, according to the following formula:

$$\Lambda_s(\varepsilon) = \begin{cases} \varepsilon - i & \varepsilon \in [i, \frac{i+j}{2}] \\ j - \varepsilon & \varepsilon \in (\frac{i+j}{2}, j] \\ 0 & otherwise \end{cases},$$

where s from *1: n,* and n is represented as a number of the points for the persistent shape. It is important to emphasize that $\Lambda_s(\varepsilon)$ is produced individually for each k-dimension. Then, $\lambda_s(\varepsilon)$ is the s^{th} biggest value of $\Lambda_s(\varepsilon)$ when the homology dimension is considered. At $s = 1$, then, $\lambda_s(\varepsilon)$ may be understood as the greatest possible distance of an interval centered about ε. We may assume that persistence landscapes represent an effective data analysis tool

in statistical topology with the above definition, [7]. Fig. 4 reveals the persistent landscapes for certain points.

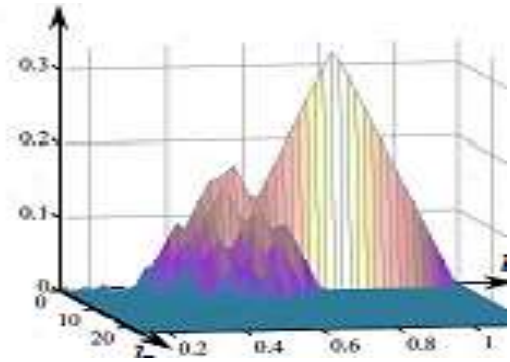

Fig. 4: The persistent landscapes diagram for random data.

6 Hypothesis Testing based on TDA

TDA inference is an emerging area of research at the intersection of statistics, computational geometry, and algebraic topology. The persistent homology framework has been used to construct statistical foundations for inference in the latest work, e.g., [8], [9], and [10]. There are many references about TDA, such as [11], [12], and [13]. Yet, this study focuses on the usage of TDA tools in testing hypothesis between two high-dimensional data sets.

a) Hypothesis Testing Based On Persistent Homology

The authors in [14] designed a test that is reliable for comparing two sets of persistent homology diagrams; each of them has a finite number of persistent homology diagrams. We can call this situation a multivariate persistent homology test. Since this topic is beyond our scope, we confine ourselves to converting that test into the un-invariant case.

The test statistic that may be used to compare two persistent homology diagrams based on [15] may be expressed as:

$$T_R = W(\hat{P}_1, \hat{P}_2) \tag{2}$$

where $W(\hat{P}_1, \hat{P}_2)$ is the Wasserstein distance between $\hat{P}_1$ and $\hat{P}_2$. By calculating the Wasserstein distance using the Hungarian algorithm. Let $\hat{p}_{1,1}, \hat{p}_{2,1} \dots \hat{p}_{n_1,1}$ and $\hat{p}_{1,2}, \hat{p}_{2,2} \dots \hat{p}_{n_2,2}$. These are points matching to $\hat{P}_1$ and $\hat{P}_2$.

The Hungarian algorithm included two samples, equal in size, which is accomplished by giving n_1 points to the second sample and n_2 points to the first sample, producing $n_1 + n_2$ points for both samples. The additional points are perpendicular distances that are a duplicate of a diagonal. The cost matrix is then constructed, whose entries are the square of Euclidean distances. Then, the optimum column (with the least distance in that column) gets for each row. Lastly, the Wasserstein distance (determined independently for points of dimensions zero, one, two, and so on) is the total of the optimal distances, implying that the Hungarian method provides the lowest cost value.

Because the sample distribution to T_R is undefined, various nonparametric methods, such as jackknifing, a permutation test, or bootstrapping, can be taken to obtain an empirical distribution for the statistical test. For applying this technique, the samples taken should reflect their populations. Hence, Robinson and Turner, [16], preferred to use the null hypothesis significance test or the permutation technique to determine the relevance of T_R. The nonparametric tool as the permutation approach entails permuting the data in the sample by shuffling their labels and computing T_R to every permutation. The null distribution is constructed by collecting T_R from permuted data. If we compare the two groups as statistically identical, random permutations applied to the observational data have no effect. In this situation, the observed test statistic falls within the range of permutations. The following steps are for getting a P-value using a permutation test:

Data: $\hat{P}_1$ and $\hat{P}_2$ with two sample sizes m_1 and m_2, respectively. The number of permutation samples is denoted as N.

Results: P-value for T_R

Calculate T_R from the original sample data.

for $I = 1{:}N$

Split the labels for the group at random into distinct groups of size n_1 and n_2;

Calculate T_R each permutation, take a sample, and record the results in E_i;

End.

P-value is the # of times that E_i is bigger than dividing T_R by N.

The main drawback of the T_R test is that it depends on the whole points in the persistence diagrams without eliminating the noisy observations, [2]. Thus, our first proposed test is operating T_R on the signal points of the persistence diagram only.

b) Hypothesis Testing Based On Persistent Landscapes

The average of $\lambda_s(\varepsilon)$ was used to create two new statistical tests that may be functioned to evaluate the difference between two provided samples in the case of high dimensional, [7]. To construct the first test, define the average persistent landscape over all persistent points taken the dimension of the points into account:

$$\eta_s = \frac{\sum_{\varepsilon=1}^{T} \lambda_s(\varepsilon)}{T} \ , \qquad s = 1 \dots n \ ,$$

which yields to $\eta = [\eta_1, \eta_2, \dots \eta_n]$. Hence, we may represent the test statistics as follows:

$$T_{sB} = \frac{|\eta_{s1} - \eta_{s2}|}{\sqrt{Var(\eta_{s1}) + Var(\eta_{s2})}} \qquad (3),$$

where η_{st} is the average of the $\lambda_s(\varepsilon)$ corresponding to the sample t. Although the test statistics T_{sB} has unknown sample distribution as we do not have such knowledge about T_{sB}, Bubenik, [6], proved that for large T, it is possible to consider the standard normal distribution as an asymptotic distribution for T_{sB} using the central limit theorem and the law of large numbers.

In addition, he suggested that one could conduct T_{sB} at all s simultaneously using multivariate T-test or Hotelling's T-square test. The second test statistics proposed by Bubenik in, [6], can be expressed as follows:

$$T_B = |\eta^1 - \eta^2|^t \left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1} |\eta^1 - \eta^2| \qquad (4),$$

where $\eta^t = [\eta_{1t}, \eta_{2t}, \dots \eta_{nt}]$ and S_t is the variance-covariance matrix of order $n \times n$ corresponding to the sample t of size n_t. The main drawback that may be thrown at T_B and T_{sB} tests is that it relies on the whole points in the persistence diagram, which leads to all the landscape values being used. Thus, our second and third proposed tests are operating T_B and T_{sB} only on the significant points of the persistence landscapes.

c) Hypothesis Testing Based On Density Estimation

A density estimate approach is a nonparametric approach used to estimate the underlying continuous distribution over a finite point set applied and studied in various contexts. Versatile methods can be adopted to estimate the density of the studied data, [4]; however, h-Nearest Neighbors (hNN) or k-NN are adopted to estimate the density points for our point cloud data. Although a broad range of authors uses hNN in classification and clustering, we decided to utilize hNN in testing that the given two groups of point cloud data are similar or not based on the following hNN density estimator:

$$\hat{f}(x_i) = \frac{h}{N \, v_d r_h^d(x_i)} \qquad (5)$$

where N is the total number of points in the dataset, h is the number of points we want in our neighborhood, often equals 10% of N, x_i is a vector of our given point, $r_h^d(x_i)$ is the Euclidean distance to the h^{th} nearest point, and v_d is the volume of the unit sphere in the dimension d of the data taking the following expression:

$$v_d = \frac{2\pi^{d/2}}{d \, \Gamma(d/2)}. \qquad (6)$$

Herein, we can summarize our main idea concerning our fourth test as first compute the $\hat{f}(x_i)$ for the two groups of the data, then calculate the following test statistic:

$$T_h = \frac{\left|\bar{\hat{f}}_1 - \bar{\hat{f}}_2\right|}{\sqrt{Var(\bar{\hat{f}}_1) + Var(\bar{\hat{f}}_2)}}, \qquad (7)$$

where $\bar{\hat{f}}_1$ and $\bar{\hat{f}}_2$ are the average of the density estimates for each group separately. Since the sample distribution for T_h is unknown, the critical values can be obtained via operating the permutation test, [4].

7 Simulation Study

The practical performance of the tests above is studied in this section. We compare the suggested tests, i.e., T_{MR},T_{MsB},T_{MB}, and T_h, to these current tests T_R,T_{sB}, and T_B. The first three proposed tests are computed based on the Bottleneck distance, not on the Hausdorff distance, as the latter has no noteworthy effect on the tests. To implement the comparison, we performed the tests above on standard geometric objects, which may be produced using the Geozoo Package, then documented the *p-values* for every test by applying the TDA technique, [4], [15]. When the two groups are formed from identical geometric objects, the *p-value* is indicated as the size of the test. Otherwise, the power of the test is determined by the p-value.

Because it may be difficult to make theoretical comparisons concerning the performance of past

experiments, one may turn to Monte Carlo simulation, which is currently a widely employed scientific technique for solving mathematically insoluble problems and high-cost experiments. Even still, simulation has downsides: It may consume a large amount of computational power and cannot provide perfect results, and the model and inputs are employed to determine the quality of the output.

The comparison between statistical tests should be conducted in several contexts, which may be stated as follows:

1. Various sample sizes: For our simulation, we run two distinct sample sizes, i.e., 50 and 100.
2. Data from various dimensional point clouds: In this study, we decide to perform the simulation at different dimensions of the generated data, such as:

a. At the $\mathbb{R}^2$: The comparison is between a circle with a radius equal to one and a normalized square.
b. At the $\mathbb{R}^3$: The comparison is between a sphere with a radius equal to two, and a torus with a radius from the center equal to two.
c. At the $\mathbb{R}^4$: The comparison is between a flat torus with a radius equal to one, and a Klein bottle with an inner radius equal to one.

3. Different dimension holes: At β_0, β_1 and β_2 the power and size for every test, unless T_h, are determined, allowing us to illustrate which dimensions the tests can completely represent the object's topological properties.
4. Different levels of the peak: Concerning T_{sB} and T_{MsB}, the power and size for every test are computed at peak equal to one, two, and three. Thus, the tests T_B and T_{MB} are operated simultaneously for η_{1t}, η_{2t} and η_{3t}.

Under the aforesaid parameters, the outputs from the Monte Carlo simulation are applied using 100 replicates (increasing the replicates will not change the finals results) and 100 permutations at 99 percent confidence intervals, equivalent to the Vietoris-Rips complex, and they are compatible with previous works, [8], [17]. Table 1 summarizes and organizes the results. The total results lead to a number of conclusions, which are discussed in detail:

1) The study shows that increasing the sample size has a significant impact on the simulated power. The size of these tests yields an increase in the power of the test and reduces the size of the tests. Consequently, using these tests with high sample numbers is suggested.

2) The Betti dimension is a factor or impact on the performance of overall tests. In general, with high levels of Betti dimension, the final decision is most likely to be correct. Conversely, the dimensional point cloud data have no sequential impact on these presented results.

3) The tests based on the persistent landscapes may have some difficulties that might arise in real life, especially at low d, that in some situations, one cannot compute the tests T_{sB} and T_{MsB} at high levels of s, which leads not to compute the tests T_B and T_{MB}.

4) Another problematic issue associated with the tests based on the persistent landscapes is that at different levels of s yields a wide range of p values, which may cause some confusion for the researchers during decision-making. The results reveal that the second and the third level of s can be recommended at the Betti dimension equal to zero, whereas $s = 1$ otherwise.

5) We observed the superiority of the power of T_h to the remaining tests in almost all the simulated cases. In contrast, its size is relatively far from the ideal nominal level of 1%.

6) It is observed that the size of the tests based on persistent homology is close to the nominal level of 1% compared to the tests based on persistent landscapes and density estimation. In contrast, the latter's power is superior to the first one.

This phenomenon refers to the fact that the tests based on persistent homology accept the null hypothesis. In contrast, the tests based on either persistent landscapes or density estimation reject the null hypothesis. Consequently, one can surely depend on the tests based on persistent homology in the case of rejection and the tests based on persistent landscapes or density estimation in the case of acceptance.

7) Inherently, removing the points with short lifetimes from the simulated persistent homology improves, in most cases, the performance of the tests. Therefore, one needs to implement the other methods, [2], to study the effect of applying the remaining methods on the performance of the tests.

Table 1. A simulation of the study's size and power of test statistics.

			Tests										
Sample Size		Dimension	T_R	T_{MR}	T_{sB}			T_{MsB}			T_B	T_{MB}	T_h
					1	2	3	1	2	3			
50	Circle V.S. Circle	0	.13	.05	.01	.25	.25	.01	.20	.27	.30	.31	.05
		1	.10	.07	.28	.25	.15	.20	--	--	--	--	
	Circle V.S. Square	0	.20	.30	.10	.70	.60	.10	.73	.62	.80	.85	.60
		1	.40	.45	.50	.30	.20	.60	--	--	.45	--	
	Square V.S. Square	0	.10	.03	.01	.15	.10	.01	.15	.11	.25	.27	.10
		1	.04	.03	.22	.20	.15	.20	--	--	--	--	
	Sphere V.S. Sphere	0	.10	.02	.01	.30	.35	.02	.31	.39	.25	.35	.20
		1	.06	.05	.38	.20	.24	.30	.24	--	--	--	
		2	.10	.09	.40	--	--	.41	--	--	--	--	
	Sphere V.S. Torus	0	.50	.53	.01	.75	.65	.01	.77	.66	.77	.80	.90
		1	.20	.30	.62	.60	.41	.70	.66	--	--	--	
		2	.35	.38	.90	--	--	1.0	--	--	--	--	
	Torus V.S. Torus	0	.08	.05	.01	.22	.18	.01	.20	.20	.25	.27	.22
		1	.05	.05	.28	.30	.31	.30	.24	--	--	--	
		2	.15	.10	.20	--	--	.16	--	--	--	--	
	Flat Torus V.S. Flat Torus	0	.08	.05	.01	.20	.15	.01	.18	.16	.30	.25	.23
		1	.05	.04	.20	.25	.36	.21	.20	.35	.50	.55	
		2	.20	.16	.35	.30	--	.38	.26	--	--	--	
	Flat Torus V.S. Klein Bottle	0	.90	1.0	.03	.93	.90	.01	1.0	1.0	1.0	1.0	1.0
		1	.95	1.0	.75	.60	.66	.85	.75	.70	.90	1.0	
		2	.25	.15	.30	.45	--	.30	.45	--	--	--	
	Klein Bottle V.S. Klein Bottle	0	.10	.07	.01	.14	.13	.01	.10	.10	.25	.24	.25
		1	.06	.04	.21	.30	.32	.21	.22	.30	.47	.45	
		2	.10	.05	.20	.15	--	.19	.10	--	--	--	
100	Circle V.S. Circle	0	.10	.04	.02	.08	.10	.01	.10	.07	.10	.11	.02
		1	.04	.04	.08	.20	.13	.05	--	--	--	--	
	Circle V.S. Square	0	.55	.65	.13	.90	.90	.05	1.0	1.0	1.0	1.0	.95
		1	.60	.65	1.0	.42	.30	1.0	--	--	1.0	--	
	Square V.S. Square	0	.07	.01	.01	.09	.09	.01	.10	.11	.11	.14	.04
		1	.05	.03	.03	.13	.06	.04	--	--	--	--	
	Sphere V.S. Sphere	0	.08	.06	.02	.08	.10	.02	.05	.11	.15	.15	.10
		1	.07	.05	.20	.16	.18	.22	.14	--	--	--	
		2	.13	.10	.30	--	--	.29	--	--	--	--	
	Sphere V.S. Torus	0	.80	.90	.03	1.0	.90	.02	1.0	1.0	1.0	1.0	1.0
		1	.30	.35	.65	.62	.45	.55	.69	--	--	--	
		2	.50	.57	1.0	--	--	1.0	--	--	--	--	
	Torus V.S. Torus	0	.06	.03	.01	.12	.10	.01	.20	.20	.15	.17	.11
		1	.05	.02	.20	.15	.12	.25	.14	--	--	--	
		2	.06	.06	.10	--	--	.09	--	--	--	--	
	Flat Torus V.S. Flat Torus	0	.02	.01	.01	.10	.07	.01	.12	.10	.13	.15	.12
		1	.02	.02	.10	.20	.19	.11	.20	.20	.20	.25	
		2	.03	.04	.14	.10	.09	.17	.10	.10	.20	.19	
	Flat Torus V.S. Klein Bottle	0	1.0	1.0	.05	1.0	1.0	.02	1.0	1.0	1.0	1.0	1.0
		1	.90	1.0	.86	.90	.90	.90	.92	.95	.95	.99	
		2	.80	.85	.85	.80	.55	.90	.82	.55	1.0	1.0	
	Klein Bottle V.S. Klein Bottle	0	.04	.01	.01	.06	.05	.01	.08	.00	.15	.14	.17
		1	.02	.03	.08	.15	.12	.11	.11	.10	.17	.20	
		2	.02	.01	.10	.07	.04	.11	.09	.10	.25	.22	

8 Real-Life Applications

TDA can be employed in a broad range of fields since it is an extremely useful tool for evaluating and analyzing massive amounts of data. TDA has been widely used in the biology field over recent decades. This work analyzes two empirical world datasets on the Wisconsin breast cancer dataset from the UCI repository (683 patterns) and the heart disease dataset from Statlog (270 patterns). These data were extensively cited and studied by other

researchers for different purposes. Therefore, it may be desirable to implement the above tests to examine whether the tests can successfully distinguish between patients with benign or malignant breast cancer, those who suffer from heart disease, and those who are not suffering from heart disease.

The Wisconsin breast cancer dataset consists of one predicted variable (benign-malignant), and nine continuous attributes ranging from 1 to 10, which can depend on our analysis. Furthermore, it is noted that the points concerning benign patients take too much time during the analysis; therefore, we select a random sample equal to 240, running the sequential maximum landmark method used by JPLEX. However, the heart disease dataset includes one predicted variable (absence -presence) of heart disease, six continuous attributes, and seven dichotomous variables. Thus, the seven dichotomous variables will be omitted from the analysis, and the six continuous attributes are only utilized.

Fig. 5 reveals the persistent homology, the barcode plots, and the probability distribution of the hNN density for the two datasets according to each predicted variable.

However, Fig. 6 presents the persistent landscape diagrams in different dimensions and different λ. Based on these figures, one can visually compare the patients with benign breast cancer and those without. We explored the type of breast cancer of the patient (benign-malignant) that substantially affects the topological features' shapes. Concurrently, figures illustrate that the patients with heart disease or without had a weak effect on the topological features' shapes. AT zero dimensions, all topological characteristics are identical, and the two sampling distributions of the hNN density are slightly different.

Conversely, (Tables 2 and Tables 3) display the *p-values* associated with the whole tests under study. Characteristically, all the tests stated a statistical difference among patients with benign breast cancer and those without at a nominal level of 1%. In contradiction concerning the heart disease dataset, all the tests failed to reject the null hypothesis at zero and almost two dimensions. Still, in one dimension, all the tests successfully rejected the null hypothesis that there is a significant difference among patients with heart disease and those without at a nominal level of 1%. One can notice surprisingly that despite the similarity of the two sampling distributions of the hNN density in the case of the heart disease dataset, T_h perfectly distinguished between the two groups. Consequently, in our view, T_h can be recommended in practical life.

Table 2. The empirical *p*-values for the statistics tests corresponding to breast cancer.

	Dimension		0	1	2
Breast Cancer Dataset	T_R		<.001	<.001	<.001
	T_{MR}		<.001	<.001	<.001
	T_{sB}	1	<.001	.004	<.001
		2	<.001	<.001	<.001
		3	<.001	<.001	<.001
	T_{MsB}	1	<.001	<.001	<.001
		2	<.001	.008	<.001
		3	<.001	<.001	.008
	T_B		<.001	<.001	<.001
	T_{MB}		<.001	.009	<.001
	T_h			<.001	

Table 3. The empirical *p-values* for the statistics tests corresponding to heart disease databases.

Heart Disease Dataset	T_R		.40	<.001	.80
	T_{MR}		.50	<.001	.77
	T_{sB}	1	.90	.002	<.001
		2	.90	<.001	.30
		3	.95	.008	.77
	T_{MsB}	1	.95	<.001	<.001
		2	.99	<.001	.40
		3	.99	<.001	.10
	T_B		.90	<.001	.006
	T_{MB}		.99	<.001	.004
	T_h			<.001	

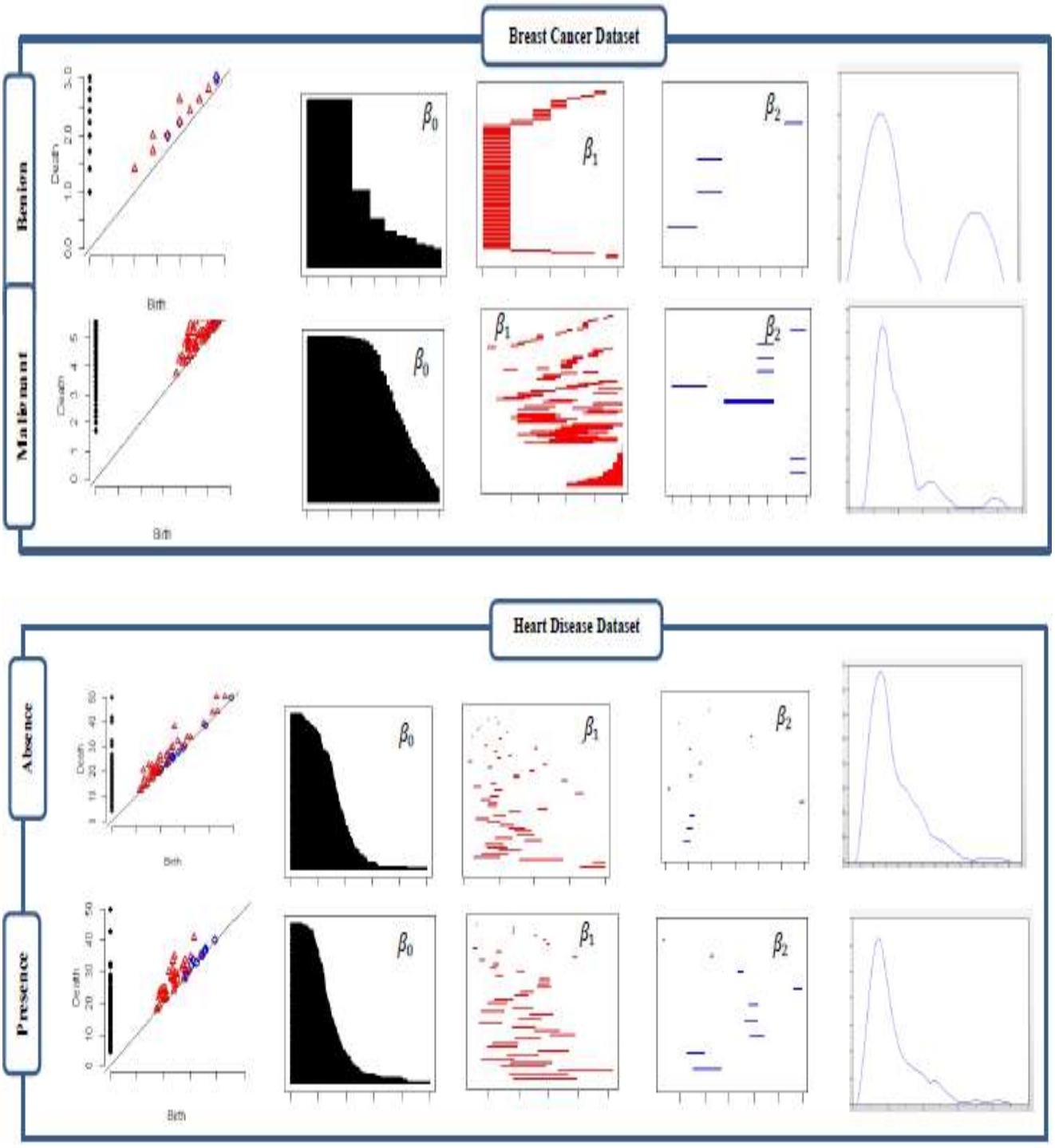

Fig. 5: The topological features for the breast cancer, and heart disease databases, respectively.

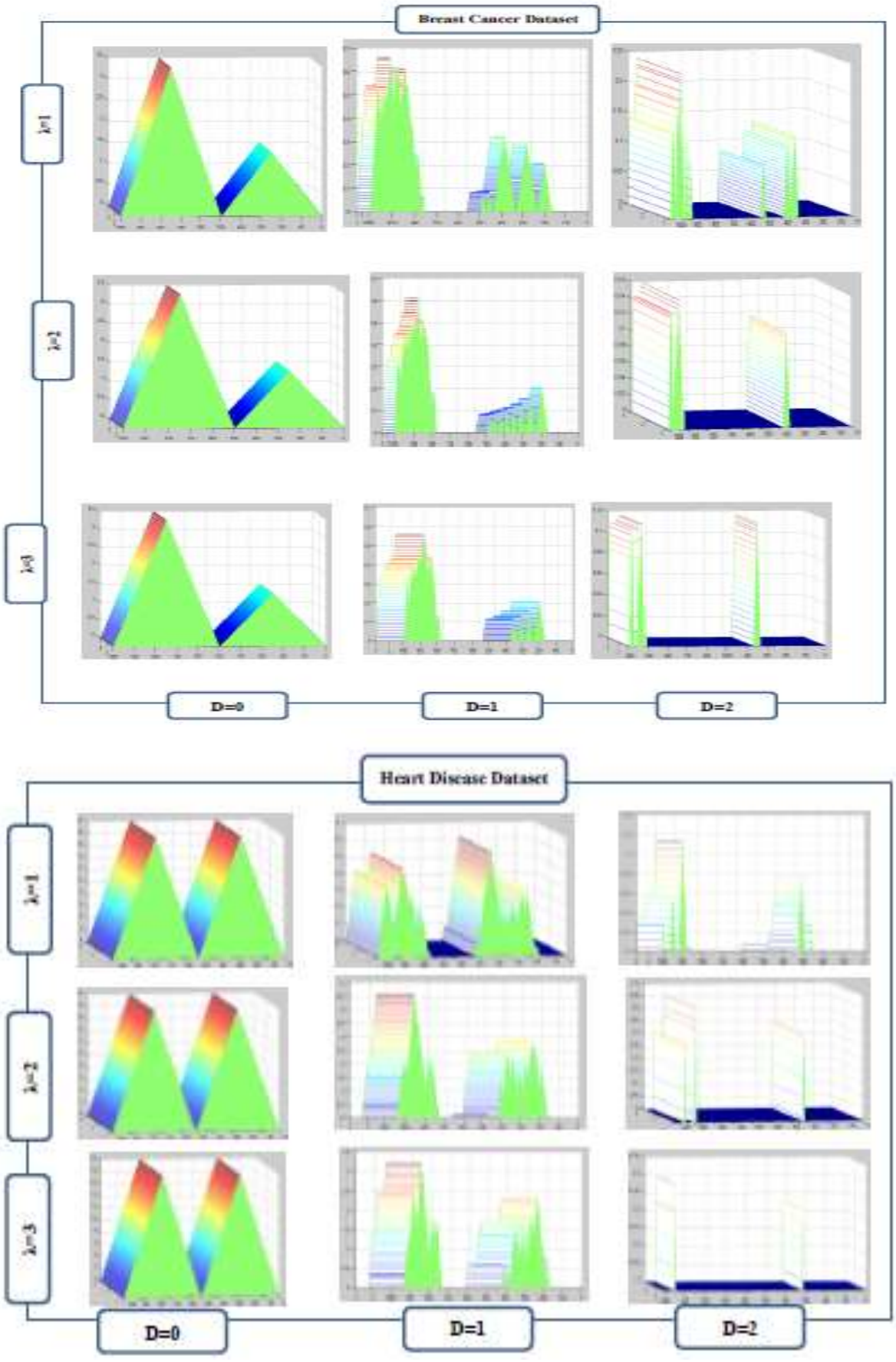

Fig. 6: The persistent landscape for the breast cancer, and heart disease databases, respectively.

9 Conclusion

In this paper, we displayed TDA techniques in hypothesis testing. It is proposed that this test is based on the nearest neighbour distance function. In addition, a suggested modification for presented tests depending on TDA is proposed. At different patterns, a comparison depending on persistent homology is performed among tests. These tests are based on a distance function, and other tests depend on a persistent landscape with two requirements: the test's size and power. According to our observations, tests that depend on persistent homology are much more appropriate. However, tests based on persistent landscape or distance function have higher power than the remaining. Generally, all TDA-based tests have fulfilled properties at dimension one; if the sample size for the point cloud data increases, it will positively impact the overall number of tests. We demonstrated the efficacy of the above tests on Statlog's heart disease dataset and Wisconsin breast cancer dataset. There is still much work to be conducted in future studies. For instance, in generalizing the preceding tests to more than two groups, comparing the various methods, [2], in-depth clustering analysis depends on TDA and evaluating it to another statistically existing method. As TDA techniques improve, we expect many researchers to apply topological analysis in their studies.

Acknowledgments:
The authors would like to thank the team of TDA, namely Brittany Terese Fasy, Jisu Kim, and Clement Maria, for their help, advice, vast expertise, and willingness to share their time freely. Moreover, a lot of gratitude to Dr. Fabrizio Lecci for her help in sending her thesis.

References:

[1] Bubenik, Peter, and Nikola Milićević, "*Homological algebra for persistence modules*", Foundations of Computational Mathematics 21.5 (2021): 1233-1278. https://doi.org/10.1007/s10208-020-09482-9.

[2] Fasy, Brittany Terese, et al. "*Confidence sets for persistence diagrams*", The Annals of Statistics (2014): 2301-2339. https://doi.org/10.1214/14-AOS1252.

[3] Alaa, H. N., and S. A. Mohamed. "*On the topological data analysis extensions and comparisons*", *Journal of the Egyptian Mathematical Society* 25.4 (2017): 406-413. https://doi.org/10.1016/j.joems.2017.07.001.

[4] Fasy, Brittany Terese, et al. "*Introduction to the R package TDA*", arXiv preprint arXiv:1411.1830 (2014). https://doi.org/10.48550/arXiv.1411.1830.

[5] Edelsbrunner, Herbert, David Letscher, and Afra Zomorodian. "*Topological persistence and simplification*", *Proceedings 41st annual symposium on foundations of computer science*. IEEE, 2000. **DOI:**10.1109/SFCS.2000.892133.

[6] Bubenik, Peter. "*Statistical topological data analysis using persistence landscapes*", *J. Mach. Learn. Res.*16.1 (2015): 77-102. https://doi.org/10.48550/arXiv.1207.6437

[7] Balchin, Scott, and Etienne Pillin. "*Comparing Metrics on Arbitrary Spaces using Topological Data Analysis*", *arXiv preprint arXiv:1503.04619* (2015). https://doi.org/10.48550/arXiv.1503.04619

[8] Carlsson, Gunnar. "*Topology and data*", *Bulletin of the American Mathematical Society* 46.2 (2009): 255-308. DOI:10.1090/S0273-0979-09-01249-X.

[9] Emmett, Kevin, et al. "*Parametric inference using persistence diagrams: A case study in population genetics*", *arXiv preprint arXiv:1406.4582* (2014). https://doi.org/10.48550/arXiv.1406.4582

[10] Gamble, Jennifer, and Giseon Heo. "*Exploring uses of persistent homology for statistical analysis of landmark-based shape data*", *Journal of Multivariate Analysis* 101.9 (2010): 2184-2199. https://doi.org/10.1016/j.jmva.2010.04.016.

[11] Kim, Wonse, et al. "*Investigation of flash crash via topological data analysis*", *Topology and its Applications* 301 (2021): 107523. https://doi.org/10.1016/j.topol.2020.107523.

[12] Dłotko, Paweł, and Thomas Wanner. "*Topological microstructure analysis using persistence landscapes*", *Physica D: Nonlinear Phenomena* 334 (2016): 60-81. https://doi.org/10.1016/j.physd.2016.04.015.

[13] Sizemore, Ann E., et al. "*The importance of the whole: topological data analysis for the network neuroscientist*", *Network Neuroscience* 3.3 (2019): 656-673. https://doi.org/10.1162/netn_a_00073.

[14] Artamonov, Oleg. "*Topological methods for the representation and analysis of exploration data in oil industry"*, Diss. Technische Universität Kaiserslautern, 2010. urn:nbn:de:hbz:386-kluedo-25456.

[15] Curran, James, and Maintainer James M. Curran. "Package 'Hotelling'", *R package version* (2017): 1-0. https://github.com/jmcurran/Hotelling.

[16] Robinson, Andrew, and Katharine Turner. "*Hypothesis testing for topological data analysis*", *Journal of Applied and Computational Topology* 1.2 (2017): 241-261. https://doi.org/10.1007/s41468-017-0008-7

[17] Edelsbrunner, Herbert, and John L. Harer. "*Computational topology: an introduction*", American Mathematical Society, 2010. https://doi.org/10.1007/978-3-540-33259-6_7.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors completed all aspects of the study through diligent work and an analysis of numerous sources and achievements in the field of mathematics. The final version has been read and accepted by all authors.

-S. Z. Rida, Mathematics Department, Faculty of Science, South Valley University, Qena, Egypt
-Alaa Hassan Noreldeen, Mathematics Department, Faculty of Science, Aswan University, Aswan, Egypt.
-Faten. R. Karar, Mathematics Department, Faculty of Science, Aswan University, Aswan, Egypt.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received.

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.