TOPOLOGICAL DATA ANALYSIS VIA UNSUPERVISED MACHINE LEARNING FOR RECOGNIZING ATMOSPHERIC RIVERS CONDITIONS ON FLOOD DETECTION

OHANUBA FELIX OBI

UNIVERSITI SAINS MALAYSIA

2022

TOPOLOGICAL DATA ANALYSIS VIA UNSUPERVISED MACHINE LEARNING FOR RECOGNIZING ATMOSPHERIC RIVERS CONDITIONS ON FLOOD DETECTION

by

OHANUBA FELIX OBI

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

August 2022

ACKNOWLEDGEMENT

I hereby express my heartfelt gratitude to God for giving us life and good health; for keeping my parents, Mr Lawrence Ohanuba and Mrs. Cecilia Ohanuba, alive during my PhD study. I thank my elder brother Mr Emeka Ohanuba, for his financial support and encouragement. I thank my supervisor, Prof. Mohd Tahir Ismail, for his guidance and quick response to the numerous challenges I encountered in this research; he is the best supervisor I have encountered throughout my academic journey, a noble man with a golden heart. Thank you, sir, for the tremendously positive feedback from you regarding this study. My sincere gratitude goes to my wife, Mrs Amarachi Juliet Ohanuba, and my son, Master Chimeremeze Jared Ohanuba, for their prayers over my study in Malaysia. Special appreciation goes to Dr Majid Khan Majahar Ali, a senior lecturer in USM, for suggesting the research area; he also made many contributions that helped develop the manuscript for publication. Exceptional gratitude goes to all my TOG brethren for their prayers – Mr Tay Sew Chew and his wife Mrs Eunice Oo, Miss Sue Lin, Bro Edwin Ong for their financial support.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF SYMBOLS

E	Threshold (intersect)
$\mathbb{R}$	Real numbers
$\mathbb{R}^d$	Dimension of real number
$\in$	Epsilon
β_K	Rank
Inf	Infimum
dist	Distance/space
d	Space
e	Exponent
$\mathbb{R}_{2R}$	Rips complex (Filtered simplicial complex)
Z_2	Integers modulo 2
N	Natural numbers
Z^+	Positive integers
$\subseteq$	Subset
$\mathbf{s^2}$	Variance
D	Dimension
f_x	Interpolation function
max	Maximum
min	Minimum
A'	Prime of A
$\in$	Member of

LIST OF ABBREVIATIONS

A.csv file	A comma-separated file
AR	Atmospheric River
Avg	Average
CSV	Comma-separated values
DE ECHO	Directorate-General for European Civil Protection and Humanitarian Aid Operations'
DEM	Digital Elevasion Model
EPD-RW	Random Work Based Method
FRM	Flood Risk Management
GEOSFM	Geospatial Stream Flow Model
GO	Gene Ontology
GS	Gap statistic
IPS	Institut Pengajian Siswazah
ML	Machine Learning
MSPC	Multivariate Statistical Process Control
N/A	Not Available
NC	Northcentral
NE	Northeast
NEMA	Nigeria's National Emergency Management Agency
NIHSA	Nigeria Hydrological Services Agency
NIMET	Nigeria Metrological Agency
NW	Northwest
PCA	Principal Component Analysis
PD	Persistent Diagram
PH	Persistent Homology

PI	Persistent Image
PLS	Partial least-squares
PPI	Protein-Protein Interaction
RST	Replicating Statistical Topology
SC	Silhouette coefficient
SDG	Sustainable Development Goal
SE	Southeast
SI	Silhouette Index
S.No	Serial Number
SS	South-south
SW	Southwest
SW	Silhouette width
SWAT	Soil Water Assessment Tool
TDA	Topological Data Analysis
TDA-uML	Topological Data Analysis and Unsupervised Machine Learning Method
TECA	Toolkit for Extreme Climate Analysis
USM	Universiti Sains Malaysia
Var	Variance

LIST OF APPENDICES

ANALISIS DATA TOPOLOGI MELALUI PEMBELAJARAN MESIN TANPA PENYELIAAN UNTUK MENGENAL PASTI KEADAAN ATMOSFERA SUNGAI BAGI PENGESANAN BANJIR

ABSTRAK

Banjir ialah bencana alam yang setiap tahun memusnahkan bangunan, tanah ladang, harta benda, dan kehidupan di banyak wilayah di dunia. Kurang daripada dua dekad yang lalu, analisis data Topologi (TDA) dan pembelajaran mesin (ML) telah digunakan dalam ramalan, yang mempunyai kelebihan berbanding kaedah biasa. Oleh itu, kerja ini memperkenalkan kaedah hibrid TDA dan ML tanpa pengawasan (TDA-uML) untuk pengurusan banjir. TDA-uML menggabungkan algebra topologi dengan sains komputer untuk menjadi bidang kajian baharu dalam statistik, mengendalikan bentuk dalam data raya. Tiga sifat menjadikan TDA berbeza daripada kaedah biasa; ia adalah ketakvarianan koordinat, ketakvarianan canggaan, dan perwakilan termampat. Kaedah ini melibatkan latihan, ujian, pengiraan, mendapatkan nilai optimum dan pengesahan nilai optimum. Beberapa model pengurusan banjir biasa seperti model Hidrologi, hidraulik dan statistik yang penyelidik telah gunakan adalah tidak tepat dalam ramalan, mahal, kurang pelaksanaan model hibrid dan tidak disahkan berbanding kaedah TDA-uML. Teknik ini bertujuan untuk membangunkan kaedah hibrid TDA-uML untuk ramalan banjir; menilai ketepatan kaedah hibrid (TDA-uML) dalam meramal banjir, memilih ujian kesahan terbaik untuk kajian, dan menentukan sama ada terdapat hubungan dalam corak ciri. TDA-uML terdiri daripada k-min berkelompok dan aspek homologi gigih TDA. Penemuan menunjukkan bahawa kaedah mengekstrak ciri topologi yang berkaitan daripada set data tetapi dengan cara yang berbeza, menghasilkan hasil yang cekap sebanyak 80%. Akhir sekali, TDA-uML

dapat mengesan banjir dan tiada banjir dalam set data untuk kawalan banjir; nilai terpencil telah dibetulkan dalam pelaksanaan prosedur ujian kesahan. Selain itu, aspek homologi gigih (PH) kaedah mengenal pasti banjir dan tiada bulan banjir dalam set data, mewujudkan hubungan yang sama dengan kaedah pengelompokan k-min baharu. PH boleh mengekstrak maklumat dari segi ringkasan (nilai) topologi dan mentafsir keputusan melalui kod bar dan plot gambar rajah gigih (PD); pengelompokan k-min boleh mengesan corak ciri luar biasa pada 4 negeri (Kogi, Oyo, Taraba dan Kano) daripada 14 negeri terpilih. Penyepaduan pengelas mesin vektor sokongan (SVM) dalam kajian ini, memberikan hasil yang berkelajuan tinggi dan tepat yang boleh meramalkan banjir dan mencegah bencana besar.

TOPOLOGICAL DATA ANALYSIS VIA UNSUPERVISED MACHINE LEARNING FOR RECOGNIZING ATMOSPHERIC RIVERS CONDITIONS ON FLOOD DETECTION

ABSTRACT

Flooding is a natural disaster that annually destroys buildings, farmland, properties, and life in many regions of the world. Less than two decades ago, Topological data analysis (TDA) and machine learning (ML) were used in predictions, which have advantages over the common method. Thus, the present work introduces a hybrid method of TDA and unsupervised ML (TDA-uML) for flood management. The TDA-uML blends topological algebra with computer science to become a new study area in statistics, handling shapes in big data. Three properties make TDA distinct from common methods; they are coordinate invariance, deformation invariance, and compressed representation. The method involves training, testing, computation, obtaining of optimal values and validation of optimal value. Some common flood management models such as Hydrologic, hydraulic, and statistical models that researchers had used are inaccurate in the prediction, costly, lack the implementation of hybrid models, and are not validated compared to the TDA-uML method. The technique is aimed at developing a hybrid method of TDA-uML for flood prediction; evaluating the accuracy of the hybrid method (TDA-uML) in predicting flood, choosing the best validity tests for the study, and determining whether there is a relationship in the feature patterns. The TDA-uML comprises k-means clustering and Persistent homology aspects of TDA. Findings showed that the methods extracted relevant topological features from datasets but in distinct ways, producing efficient outcome of 80%. Finally, TDA-uML could detect flood and no flood in the dataset for

flood control; outliers were fixed in the implementation of the validity test procesure. Moreover, persistent homology (PH) aspect of the method identified flood and no flood months in the datasets, establishing a similar relationship with the new k-means clustering method. The PH could extract information in terms of topological summaries (values) and interpret the results via barcodes and persistent diagram (PD) plots; the k-mean clustering could detect the unusual features pattern on 4 states (Kogi, Oyo, Taraba, and Kano) out of 14 selected states. The integration of the support vector machine (SVM) classifier in this study, provided high speed and accurate results that could predict flood and prevent eminent disaster.

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Topological data analysis (TDA) is a recent research area in statistics that uses algebraic topology tools to capture a dataset's shape and structure. According to Gholizadeh and Zadrozny (2018), TDA is a mathematical tool that captures the intrinsic structure of shapes in a dataset. It was reported in Biscio and Møller (2019) that the TDA as a method in statistics uses algebraic topological ideas to summarize and visualize complex datasets. The origin of TDA was traced to the 18th century by a Swiss Mathematician, Leonard Euler (Richeson, 2019). The study further revealed that the impact of Euler's formula in the field of geometry contributed to the historical development of TDA, and was applied in the study of shape. Topological data analysis gradually appeared over time, and its historical development can be traced to the work of these authors: Frosini (1990); Vanessa (1999); Letscher and Zomorodian (2002); Zomorodian and Carlsson (2005). The advent of the combined method of TDA and Machine Learning (ML) is more recent and has gained more applications as recorded in the studies of Carlsson (2014); Bubenik (2015).

The development of persistent homology (PH), k-mean clustering, and other tools in TDA has been driven by algorithm (Vejdemo-Johansson, 2012). In the 90s, the area started with computational geometry and with researchers (geometers) interested in studying the algorithmic aspects of the classical subjects in mathematics, such as algebraic topology. Topological Data Analysis (TDA) was discovered as a researchable area in statistics after many years of origin from Mathematical Topology. The area flourished in the 80s' and 90s' by addressing different practical problems and

enriching the area of discrete geometry in the course. Computational topology can address the area of shape and data analysis while drawing upon and perhaps developing further the area of topology in the discrete context (Edelsbrunner & Harer, 2010).

Since the discovery, some tools have been identified and used in the analysis and/or computation of TDA; among them are clustering, Persistence homology, Ridge estimation, and Manifold estimation. This study used two TDA's tools: clustering (i.e., k–mean) and persistence homology (that combines the Vietoris Rips and Persistent diagram functions). In both cases, the same datasets were used. The two TDA's tools integrated unsupervised machine learning (uML) procedure in their implementation; this made the method synthetic, automated, and unique in flood management compared to other flood models. There are three key properties that give TDA the power to analyze and understand shape. They are (a) Coordinate invariance (freeness); the principal idea is that it should not matter how the dataset was represented in terms of coordinate, provided the internal distances remain unchanged. Moreover, the measure of topological shape does not change if the shape is rotated. (b) Deformation invariance; the property of invariant deformation remains unchangeable even if there is stretching on the object. For example, the letter 'A' remains a loop with two legs and a close triangle if apply stretching to it. (c) Compressed representation; If closely observed, the illustration of important attributes in the letter 'A' can be observed as having three bounded angles with two stands (legs) irrespective of the compression of millions of data points with similar relationships in the object. More details on the three properties of TDA with illustrated examples can be found in Offroy and Duponchel (2016).

The implementation of the TDA-uML method invoved importation of important libraries, loading of datasets, training and testing, data preprocessing, and computation of cluster centroids follow. Finally, the support vector machines (SVM) classified the resultant feature into flooding and no flooding zones, establishing flood detection in the proposed method (TDA-uML method). The PH aspect also integrated ML procedure in its computation to obtain topological features. The outcome of the plots in PH could extract inputs that share similar features with the k-mean clustering – this was discovered in the bars of the barcode and persistent diagram (PH) plots between the months of January and September. The outcome could establish a relationship in the feature patterns of the analysis.

A few pieces of literature that applied the combined method of TDA in flood control are Muszynski et al. (2019); Zulkepli et al. (2020); and Musa et al. (2021). Nevertheless, many researchers have researched flood control without applying TDA to study pattern recognition using complex data (simulated data). These listed authors did not implement TDA in flood modeling and control (Olugunorisa, 2009; Olajuyigbe et al., 2012; Archfield et al., 2013; Nkwunonwo et al., 2015; Nkwunonwo, 2016; Yang et al., 2019). This study used actual dataset and is the first to apply TDA in studying flooding in Nigeria. The advantage of the method is that it uses classifiers in binary classification and can combine with other techniques. Next, it works well in large and noisy datasets via dimensionality reduction. The cluster validity test(s) is observed to produce a better feature pattern in the outcome regarding connectedness and compartment in the k = 2 clusters.

In this study, Excel is used to arrange our dataset and fix missing values, Python programming language in writing a suitable code for the analysis, and R programming language codes to evaluate our results' validity. Next, four rainfall

parameters (maximum and minimum temperature, precipitation, and wind speed) were used to fill a research gap, considering the finding in Riihimäki et al. (2020), which stated that the more the variable, the better the result. After that, various flooding patterns were identified in the fourteen selected regions in Nigeria. The results obtained in the subsequent tables and plots showed potential flooding, no flooding, and their performances.

The validity tests were conducted to appropriately standardize the distribution of the dataset considering the work of Tibshirani et al. (2001). The variances were used to identify the zone/state with the highest flood rate and the degree of flood disaster in a particular zone. The density features of the four variables used in this study were obtained. The Persistent homology (PH) was used to study our datasets in their low-dimension topological features. Finally, the findings provided answers to the research questions: to develop a hybrid method of TDA and unsupervised ML (TDA-uML) for flood detection, to evaluate the accuracy of the hybrid method (TDA-uML) in detecting flood, to measure the extent of spread of the resultant clusters from the centroid, to choose the best test that validated our method, and to determine whether there is a relationship in the feature patterns of our analysis.

1.2 Motivation of Study

In this recent era, all types of flood modelling (mathematical, statistical, and machine learning) have rapidly grown in number, value, and implementation procedure. They all use a working algorithm and computations via computer programs and developed software. This rapid progress has simultaneously reduced flood impact and increased confidence in flood prediction and potentially generated the essentiality of proper risk management in flood control. Moreover, the increase in technology and

industry brings about a corresponding increase in a data point, such that automated and hybrid methods are needed to meet the geometric rate of increase in datasets. Most recent studies are beginning to drift to a hybrid system of combining two methods in flood control; research in time series has started adopting and implementing the combined method of TDA, which makes it extract topological features of datasets (Lei, 2020).

The combined application of the TDA approach was used in Musa et al. (2021) in the theory of the Critical Slowing Down to produce a reliable Level Early Warning System on flooding. The Persistent homology applied in the model sequentially extracts two kinds of topological feature patterns (the components and the holes) from datasets. The study controlled the flood rate and minimized the fatality rate. The question here is this: can it be applied to a new environmental scenario?

Many researchers applied flood models to different flood zones for flood control. Among them are Archfield et al. (2013) predicted the flow of the stream in uncontrolled catchments, Nkwunonwu et al. (2016) estimated the risk levels of municipals affected by flood using weather instead, Olajuyigbe et al. (2012) used a secondary dataset, administered questionnaire to households, and obtained key information on flood in Lagos state, (Yang et al., 2019), used high-resolution atmospheric and hydrological model simulations in Arizona.

The menace of floods is an annual occurrence and has destroyed most environments near the flood zones. The havoc associated with flooding has cost lives, properties, and displaced families. Besides, the feature pattern of floods is rarely studied. Due to these mentioned problems, the researchers are motivated to fill the gaps and help minimize disasters from extreme weather and climate change by applying a hybrid method.

Besides, the proposed approach combines TDA and ML. It is meant to minimize outliers, noise in long-memory time series, error in the output and reduce dimensionality in a large dataset. Also, our method does not require threshold conditions in pattern recognition, and so it can use the current spatial dataset for recognition (prediction) of ARs (Muszynski et al., 2019). In a hybrid method (TDA-uML), the questions remain if more than one variable could be used in a spatial dataset; none of them considered that. Can TDA-uML method be developed for predicting flood? Can a hybrid method be evaluated? Can the degree of flood concentration be measured? None of the above research considered these questions. Therefore, it is time to see what happens if a hybrid method is implemented in flood management and control. Will these new models improve the results?

1.3 Problem statement

There have been many problems with the use of traditional models for flood prediction. The Nigerian Hydrological Services Agency (NIHSA) used the geospatial streamflow model (GEOSFM) and the Soil Water Assessment Tool (SWAT) for flood prediction. Munzimi et al. (2019) reported that they use hydrological and hydrogeological data, rainfall data, topographical data, soil, and water balance index with the Digital Elevation Model (DEM). Such models can help to establish information on the baseline flows throughout the river and provide a benchmark for assessing future hydrological changes associated with changes in land overflow and climate change. Despite their efforts, the agency has some challenges; the performance of their models has not been evaluated/validated for accurate prediction. Moreover, there are still problems ranging from inadequate use of satellite technology to inadequate training in labor development and computation.

The three traditional flood modelling types include 1) One dimension (1D) modeling approach that solves the 1D equation of river flow; 2) Two-dimension (2D) modeling approach that solves the 2D equation of river flow; 3) Link (1D channel and 2D floodplain) combines 1D, and 2D water flow models. The models were developed to permit vertical feature representation, describing fluid substances' motion (Teng et al. 2017). Many problems illustrated in this section were encountered in the common flood models. For instance, they are non-predictive, and have no direct linkage to hydrology. They are difficult to use. These have engineering, environmental, and processing limitations. They are computationally intensive, and the errors associated with the input can grow with time. Most of the flood modelling approaches, unlike TDA, lack inherent topological invariant and theoretical properties. The advent of computational power and ML procedure integrated with the proposed hybrid method has led to more robust advancements in the modelling of floods; this tackles most of the problems encountered in the common methods described here. Besides, the integration of the support vector machine (SVM) classifier in pattern detection in time-series dataset, enhances the high accuracy of flood prediction in our proposed method (TDA-uML). The result is more robust when machine learning procedure is combined according to Ambrosio et al. (2021). In networking, the whole framework can be trained and retrained in such a manner that a deeply supervised framework for the effective detection of structures in datasets could result in accurate information extraction (Ji et al., 2021).

1.4 Research questions

The question of how to detect flood feature patterns and minimize the negative effect on life (environment) is the main problem to be addressed in this study. This research study aims at answering these questions.

1. Can a hybrid method for flood detection be developed?
2. Can the accuracy of a hybrid method of flood detection be evaluated?
3. How can the extent of concentration or spread of the flood pattern be measured?
4. What is the best valid test for the research study?
5. Is there any relationship in the feature patterns obtained in this study?

1.5 Research objectives

The answers to the above-stated research questions formed the objectives of this study, and they are stated as follows:

1. To develop a hybrid method of TDA and unsupervised ML (TDA-uML) for flood detection.
2. To evaluate the accuracy of the hybrid method (TDA-uML) in detecting flood.
3. To measure the extent of spread of the resultant clusters from the centroid.
4. To choose the best test that validated our method.
5. To determine whether there is a relationship in the feature patterns of our analysis.

1.6 Significance of the Study

This study is significant to the existing literature in several ways. The main contribution of this research to the literature lies in the ability to develop a hybrid method of TDA and unsupervised ML (TDA-uML) for flood detection. This development resulted in the discovery of the potential flooding and no flooding partitions in each state. Based on the potential flooding and no flooding patterns discovered, relevant information on the eminent flooding (disaster) can be passed across to those dwelling in the regions to relocate; the awareness created will help avoid the menace resulting from flooding. The information on the flood detection will also provide vital information to the government on how to construct channels and where to channel the flood to minimize havoc.

The next significance is in the use of evaluating the accuracy of the hybrid method (TDA-uML) in detection flood. Here, the intra-cluster validity test(s) were used to measure the accuracy of the hybrid method. The similarity in the features, the outcome of the tests, and the percentage score are significant factors that provide excellent evaluation measure to the hybrid method.

In the literature on time series analysis like volatility and economic data, it is common to treat outliers and structural change separately using different procedures. This study will add to the literature by using the same procedure to detect outliers during the reanalysis process and automatically fix them, avoiding using separate procedures, saves time and money.

There has been a rare application of the technique to an extreme event like flooding. Few works of literature have applied it in the study related to flooding, and from our reviewed literature, the method has not been applied in any African country; this study is set to fill the gap and add to the literature. In its aspect, persistent

homology (PH) is a unique property that separates real features from noise. The remarkable aspect helps PH in the analysis of datasets at multiple resolutions. Our study extended the unique property to a new environmental scenario.

As the world is advancing in various aspects; in population growth; economic growth; industrial growth; production growth, and so on, more datasets are generated daily at a rapid rate; there is, therefore, a need for the adoption of a synthetic and robust technique that can meet up with the proportional rate at which dataset is being generated (Muszynski et al., 2019). The study will meet the demand of the recent growth in technology in terms of the speed and increase in a dataset: the procedure works well with both complex and simulated datasets.

1.7 Outline of the thesis

This study was carried out by employing a quantitative research technique detecting feature structure of datasets using a hybrid method of TDA and unsupervised ML (TDA-uML); meanwhile, validity was established, variances were obtained, and a comparison was made to establish the benchmark in the context. This research is structured as follows

•In Chapter 1, a brief discussion on the background and the motivation of this study, followed by the research problem, research questions, research objectives, and significance of the study.

•In Chapter 2, a summary of previous literature works related to this study and some reviews categories like the development of TDA, the limitation, and why TDA hybrids with other methods are described. The flood, mathematical, statistical, and machine learning modelling were portrayed. Persistent Homology and clustering are discussed for a theoretical and empirical understanding of the improved model.

•In Chapter 3, interpolation that was used to estimate the missing values were presented. The procedure for implementing the TDA-uML method are portrayed. Five steps that the study maintained to achieve results are also presented in this chapter.

•In Chapter 4, the results of the selected datasets considered in this result are presented. The new k-means cluster, the validity tests, the variance, and the Persistent homology were discussed. Tables of the results, the graphical plot, and views of the optimal results from the datasets are displayed. A concise discussion on the results and the reason for choosing the methods and the benchmarks are explained here.

•In Chapter 5, this is the final chapter containing a summary of the whole study carried out and a brief discussion on the research questions, contributions, and concept of the future works.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Some past pieces of literature have contributed significantly to the development of this study. It covers the goal, brief, and analytic outline of printed materials useful to the methodology used in this study. This thesis focused on Topological data analysis (TDA), flood modelling, and TDA with flood modelling which is relevant to the topic of this study. Therefore, this chapter critically evaluates the precious and relevant works in TDA to discover the vital gaps relevant to the topic as mentioned above.

This chapter is structured in the following sections and subsections to achieve the purpose: Section 2.1 Introduction, Section 2.2 Topological data analysis (TDA), 2.2.1 discusses the development of TDA, 2.2.2 is the limitation of TDA and why hybrid with other methods, 2.2.3 is the application of TDA, 2.2.4 is the critical review to find the gap. Section 2.3 is on Flood modelling, 2.3.1 is on mathematical modelling, 2.3.2 is on statistical modelling, 2.3.3 is on machine learning modelling, 2.3.4 is the critical review to find the gap. Section 2.4 TDA and Flood modelling, 2.4.1 is on the development of TDA and flood modelling, 2.4.2 is the critical review to find the gap. Finally, Chapter 2.5 is the summary.

2.2 Topological data analysis (TDA)

Topological data analysis (TDA) was recently discovered as a researchable area in statistics that uses algebraic topology tools to capture the shape and structure of data. TDA was established under some theories that data have shapes and patterns

that possess inherent topological summaries (Angarita et al., 2019). Gholizadeh and Zadrozny (2018) defined TDA as a collection of mathematical tools that capture the dataset's structure and that the primary purpose of TDA is to find the intrinsic structure of shapes in the dataset. Biscio and Møller (2019) defined TDA as a statistical method that utilises algebra in mathematical topology to compact and picture compounded datasets. The TDA blends topological algebra with computer sciences to become a new area of study in statistics.

The concept of TDA is also a technique that accumulates and analyzes datasets to identify the shapes in the dataset; an example includes cluster methods (k-means), persistent-homology, estimation of manifold, estimation of mode, and estimation of the ridge (Wasserman, 2018). The rigorous study of shape combines algebra and topological aspects of mathematics to provide a novel way of describing the structure of a dataset known as topological features. Many statisticians are not familiar with TDA, the concept, and the main ideas behind the tools since it is a recent research area in statistics.

2.2.1 Development of TDA

Topology can be described as a mathematical field that analyzes, synthesizes, and visualises the shapes and structures of datasets. Topology has been commonly used to study the shape and surface of abstract objects; applying topology knowledge in evaluating and visualizing high dimensional and complicated datasets has developed the Topological data analysis (Hwang et al., 2021). Topology has been recently (10-15 years ago) coded into a point cloud world, a world where you have a finite sample from a geometrical object (finite dataset). The formalism for measuring and representing shape has been pure mathematics since the 1700s and has recently pulled into the point cloud world to what topologists call TDA. The development of

the TDA and other TDA methods, like Persistent homology (PH), has been driven by algorithm development (Vejdemo-Johansson, 2012).

In the 90s, the area started with computational geometry and with researchers (geometers) interested in studying the algorithmic aspect of the classical subject of algebraic topology in mathematics. The area flourished in the 80s' and 90s' by addressing different practical problems and enriching the area of discrete geometry in the course. A small number of computational geometers felt that similar to this development, and computational topology can address the area of shape and data analysis while drawing upon and perhaps developing further the area of topology in the discrete context (Edelsbrunner & Harer, 2010).

The fundamental object in a topological space is an underlying set whose elements are called points. A topology on these points identifies connection by listing out the points that constitute a neighborhood. The expression "rubber-sheet topology" commonly associated with the term 'topology' exemplifies this idea of connectivity of neighborhoods, and more attention should be focused on the topological properties (Dey et al., 1999; Edelsbrunner, 2001). If we fold or apply force to enlarge the size of an expansible object like rubber, the structure will distort and transform to other shapes, but it still retains neighborhoods in terms of the points and manner of their connections. These ideas (notions) were first developed and formed the backbone in the study of properties in topology like manifold, isotopy, and other maps used later to study algorithms for Topological data analysis. Perhaps, it is more natural to understand the concept of topology in the presence of a metric because metric balls such as Euclidean space can be used to define neighborhoods. Topological spaces provide a way to abstract out this idea without a metric or point coordinates, so they are more general than metric spaces: The connectivity can be encoded in place of a

metric of a point set by supplying a list of all of the open sets (Delfinado & Edelsbrunner, 1995; Bern et al., 1999; Biasotti et al., 2011).

Richeson (2019) in his work, mentioned the impact of Euler's formula (Euler's polyhedron formula) in the field of geometry that led to the historical development of Topological data analysis (TDA). Euler's formula was named after Leonhard Euler, the founder. The first example of a topological invariant, which is one of the characteristics of TDA, was first established in the Euler characteristic; it is a quantity that can be calculated and gives back the same value on many different representations of the same topological shape. These illustrated equality below must be satisfied in Euler's polyhedron; the number of vertices (V), edges (E), and faces (F) are integrated into the equation, V – E = 2 – F. For instance, Figure 2.1 has twelve loops, six voids, and eight connected components, then 8 –12 = 2 – 6.

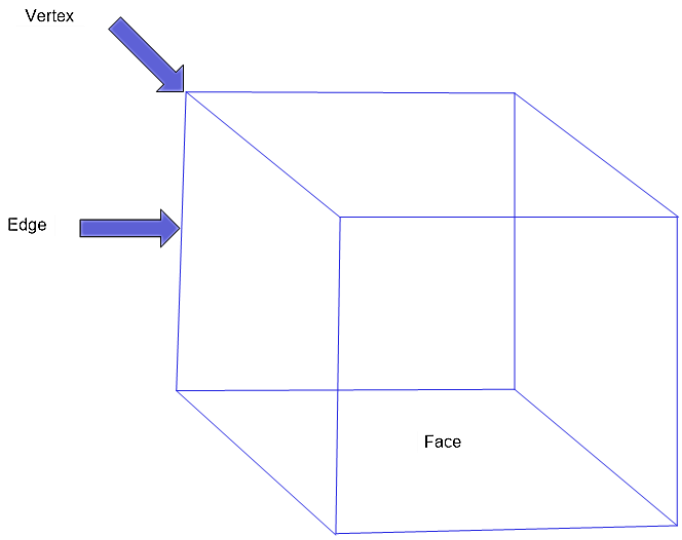

Figure 2.1 Diagram of a cube (hexahedron)

The formula is the same as the natural characteristics of a defined structure or object. Equivalently,

x = #(Connected Components/Vertices) − #(Loops/Edges) + #(Void/Face),

where # is equal to “the number of” the word inside the brackets.

All convex polyhedra process 1 connected component and 1 void, the Euler characteristic x is 1 − 0 + 1 = 2 (i.e., Vertices – Edges + Faces = 2). One can easily view the numbers in Euler’s equation in other objects. An example is a solid torus (a doughnut) shown in Figure 2.2 with one connected component, edge, and zero faces; The Euler’s formula is +1 − 1= 0 − 0. Euler’s formula transformed into Topology and was first applied to study the intrinsic shape (Richeson, 2019).

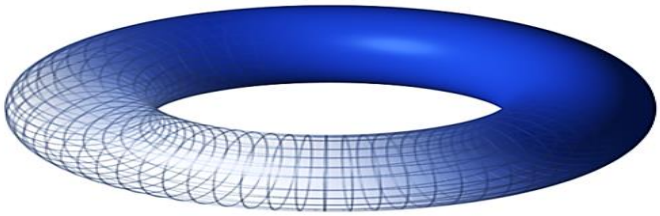

Figure 2.2 A Solid Torus

Source: https://imgbin.com/png/torus-mathematics-geometry-topology-shape-png

Suppose we apply the Euler characteristic to Topological data analysis (TDA) by taking note of the quantity of structures made up of simplexes. In that case, its topological features are incidentally summarized as recorded in Amézquita et al. (2020). The polyhedron formula propounded by Euler for the simplicial complex is expressed as

$$e^{ix} = \text{cosine}\, x + i\text{sine}\, x, \tag{2.1}$$

where x is a real number, d is the base of a natural logarithm, e is known as Euler’s number (i.e., approximately equal to 2.7182 and can be evaluated in many ways). The real number can also be evaluated using a metric dataset (Euclidean distance) and Gaussian density (Calinger, 1996; Sandifer, 2007; Offroy & Duponchel,

2016). Legendre (1752-1833) proved the Euler polyhedron equation to estimate a spherical object; the proof as recorded was mathematically correct. Legendre polynomials were equivalent to Bernoulli, Euler, and Bernstein polynomials. After that, Legendre discovered that his result could be larger polyhedra (Araci et al., 2013).

Feng and Porter (2020) and Vejdemo-Johansson and Skraba (2016) explained the steps in the formulation of TDA, and in the first step, called partition, the lens with the metric measurements are chosen. The circle file or lens can be a function in mathematics used to view datasets, and metric measures are the calculated distances or similarities between two or more points in a dataset (Vejdemo-Johansson & Skraba, 2016; Feng & Porter, 2020). Lenses can be seen in the statistical enclave (average, maximum, minimum, etc.), or mathematics; the lens derives the sections from datasets and further transforms them into sub-population (super-imposed circular files) (Offroy & Duponchel, 2016). A set is viewed sequentially using various kinds of the lens through the multiplication of outcomes they produce. The partition is analyzed in the second step (called cluster analysis); data are clustered within these bins so that the group will have aligned rows resembling one another. Because the datasets are divided into bins (circular files) overlapping, each row is oversampled and falls into more than one cluster (partition). In the third step, called network generation, data are reassembled to generate the final network. If two clusters in different bins share one or more rows, an edge is used between the two clusters to form the final network, as presented in Figure 2.3.

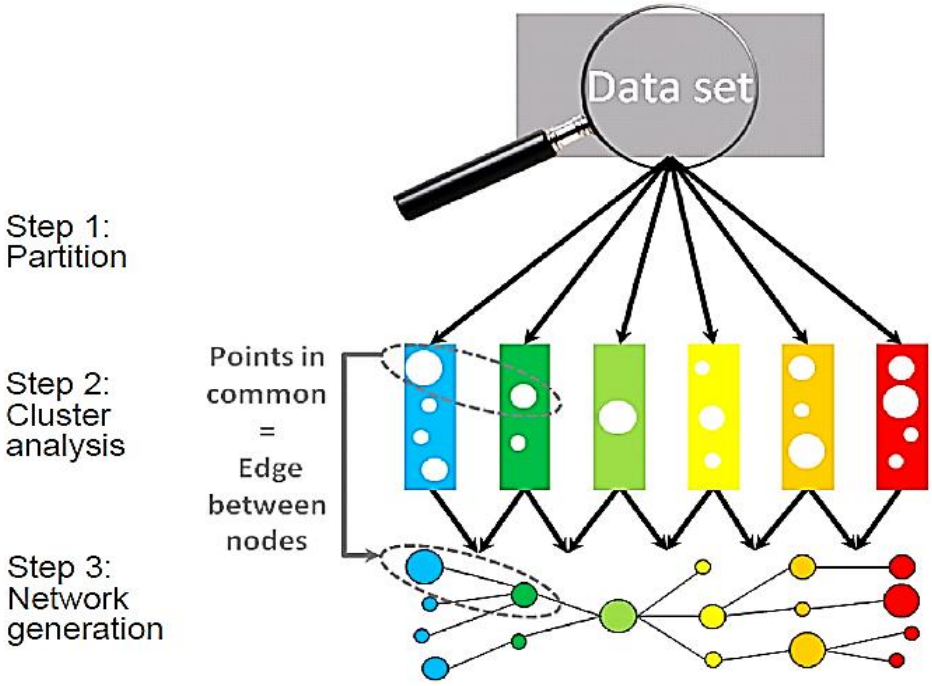

Figure 2.3 Visualization of general framework of TDA

Source: (Offroy & Duponchel, 2016)

2.2.2 The limitation of TDA and why hybrid with other methods

Some limitations are peculiar to the TDA methodology, including that the entire nature of what it explains is entirely not clear (complex to understand), especially when explained to those with no background knowledge in mathematics. The use of the software is also a limitation since not all software can be used to compute TDA. The computational procedure is also a limitation.

Topological data analysis (TDA) was shown to have limited usefulness for event-related functional magnetic resonance imaging (fMRI); the reason, as recorded, was that the fMRI dataset is too noisy to allow representation to be detected by TDA (Ellis et al., 2018). The same study was updated in 2019, and it specified that Persistent homology (PH) is the kind of TDA that has limited relevance for fMRI. It was summarized that PH is potentially important to the study of cognitive neuroscience in that it can recover topological signals from event-related fMRI datasets (Ellis et al.,

2019). Another reviewed work on human brain mapping revealed TDA patterns in fMRI; functional near-infrared spectroscopy, fNIRS; MEG, magnetoencephalography (Knyazeva et al., 2016; Bassett & Sporns, 2017). Using multiple fMRI datasets, the Topological data analysis approach detected structure within and between-task transitions at a much faster time scale (Saggar et al., 2018). Therefore, improper use of the TDA code and the wrong code are limitations to TDA.

Another limitation lies in the generation of simplexes which can discretize the spaces in data points by building a set of points, especially when the dimensionality $d > 3$ (Pereira & de Mello, 2015). Topological data analysis (TDA) also is limited to making scientific claims which can be statistically verifiable. Topological data analysis (TDA) has not used the statistical procedure as part of its approach, so there is an issue of statistical reliability despite being a recent research area in statistics (Adler et al., 2017).

A little close observation of the Euler characteristic formula in Equation (2.2) shows that *x* can represent a set of points in a dataset, where x_i is a row vector defining sample i for which a lens value is calculated and x_j, are all the other samples in the dataset. Drastic topological changes can happen at vertices x, when Euler characteristic $G(x)$ changes, based on the addition of critical points with non-zero indexes. The critical points of zero do not lead to Euler characteristics; therefore, the point $x_i,\ x_j = 0$ is beyond the limitation. Offroy and Duponchel (2016) reported that the lens uses a Gaussian kernel estimator over the data, and it is stated as

$$G(x) = \sum_{x_j \in data\ set} e^{-d^2(x_i, x_j)} \tag{2.2}$$

Therefore, this limitation may depend on the algorithm hybrid of TDA with other methods; it is also limited to the use of code (i.e., the use of package and the kind

of algorithm combination used in the TDA's analysis) procedure of the code. Ellis et al. (2019) had incorrectly used the TDA code in their publications where Persistent homology (PH) was employed to detect structures in fMRI datasets. The correct code was discovered in their 2020 research study, and its proper application to detect structures in fMRI dataset was also conducted according to Ellis et al. (2020).

Why TDA hybrid with other methods

Topological data analysis (TDA) hybrids with other methods produce more robust methods due to some characteristics listed in this section. It has joint topological and geometric attributes; and applies both methods to obtain a high dimensional dataset (Lei, 2020).

The Topological data analysis has dependent attributes of a dataset, which continue over various scales; it produces a firm description of the dependable structure of inputted dataset; it possesses high efficiency, especially during perturbation of input dataset. The implementation and coding of TDA have a free source that uses a compounded algorithm; the users always need to update and get the latest version since the newer version contains more current and efficient usage (Otter et al., 2017).

Topological data analysis (TDA) has theories upon which geometric morphometrics is formed; it starts with a set of vertices, the shape of manifolds, procrustean metrics, and complex projective spaces (Amézquita et al., 2020). Chevyrev *et al.* (2018) mentioned that embedding two methods minimizes the potency of intolerance that is not in line with the standard dataset in terms of combination with other algorithms. The flexibility implies that the TDA in a combined procedure will reduce the benchmark that might come from any kind of dataset in a particular study (Chevyrev et al., 2018).

Another reason is that the Topological data analysis (TDA) uses an idea that considers a dataset as a sample collection of numerous datasets in a high dimension of metric. Each process of superimposing sets of coordinates allows a metric space to measure the shape and representation of the same when hybrids with other methods (Chazal & Michel, 2017; Chen et al., 2021).

Offroy and Duponchel (2016) revealed that Topological data analysis (TDA) have topology properties that are good and useful in analyzing complex datasets in diverse fields of study: It was mentioned that simplices are constructed from the sample data, and those simplices develop intervals, which combined and provided a network approximately to a manifold as shown in Figure 2.4 (i).

Three key properties that give TDA the power for analyzing and understanding shape are:

(a) coordinate invariance (freeness); the principal idea says it should not matter how we represent the dataset in terms of coordinate provided we keep in track with the internal similarities (distances). The measure of topological shape does not change if you change the coordinate system of viewing the shape. The two A letters could constitute a set of data samples analyzed with two forms of analysis, and the topology constructed brings out the actual feature in it Figure 2.4 (ii).

(b) the property of invariant deformation remains unchangeable even if there is stretching of its; the letter A in figure 2.4 (iii) is a loop with two legs and a closed triangle, maintaining the key features retrieved in the topological representation. Different fonts of the letter A can be recognized in a topologically way due to how our brain captures it.

(c) compressed representation. If we observe more closely, the illustration of important attributes in the letter A can be observed as having three bounded angles with two stands (legs) see Figure 2.4 (iv). In consideration of this characteristic, the letter A has many connected data points. The TDA in the object is capable of producing a network of five edges and nodes.

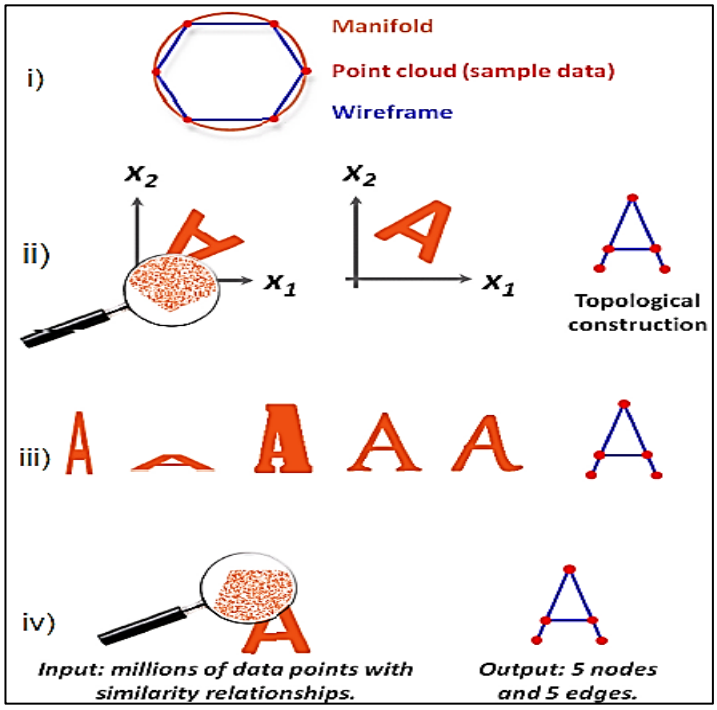

Figure 2.4 (i) Vital idea of TDA. The three key properties: (ii) coordinate invariance, (iii) deformation invariance and (iv) compression representation

Source: (Offroy & Duponchel, 2016)

2.2.3 Application of TDA

Topological data analysis (TDA) is fast growing. It has cut across various fields of study such as sciences, social science, natural sciences, biology, education, and financial econometrics, physics, chemistry, medicine, neurosciences, natural language classification, data sciences, engineering, remote sensing, natural imaging, climatology, epidemiology and health. Topological data analysis (TDA) is applied in the natural language classification, gene ontology, developmental biology, and other

biological systems to detect topological features and the timing of their formation; there is a problem with interpreting feature persistence (Ciocanel et al., 2021). Topological data analysis (TDA) also has been successfully applied in finance, financial econometric and cryptocurrencies: It has been successfully used to detect the early warning signals of potential market crashes, which help investors to make good decisions (Gidea & Katz, 2018; Gidea et al., 2018; Kim et al., 2018).

Topological data analysis (TDA) has been applied to time series analysis, dynamical systems, and signal to process; the same method is used in risk analysis and prediction of critical transitions in financial markets (Gholizadeh & Zadrozny, 2018). The TDA has been applied in Biology, medicine, and ecology; the combination of the theory of Persistent homology (PH) is summarized in persistent diagram and clustering technique. The approach reduces the number of input points required for the discretization step, but the traditional clustering techniques failed when applied as they rely on point-to-point dissimilarity measures such as Euclidean distance (Pereira & de Mello, 2015).

In medicine, for instance, it has been used to discover preclinical spinal cord injury and traumatic brain injury; it has also been used in molecular biology to assess skin function. It was revealed that TDA provides a lot of potential for decision-making in basic research and clinical concerns like outcome evaluation, neurocritical care, therapy planning, and rapid, precision diagnosis. But there is complexity in the implementation procedure of their method (Nielson et al., 2015; Koseki et al., 2020). Dindin et al. (2020) used a combined method, TDA and auto-encoder, to tackle the problem of individual differences in heartbeats. Their method achieved Arrhythmia detection and categorization techniques. Future work, as suggested is focused on using larger datasets since the benchmark did not use large datasets. Therefore, there is no

evaluation in the accuracy of Arrhythmia detection and classification in their TDA method. There is no validity of their method's performance and their study is yet to be applied to new environment to compare the performance.

In education, TDA has been applied in measuring IQ in the study of the gifted and the underlying results across various measurements of intelligence (Farrelly, 2017). The result produced using PH is robust and, it has close values compared with different TDA tools. The method is also robust in small samples; it needs to be applied in a large sample to see how well it will work in large sample sizes. Future studies may focus on these individuals with exceptional ability across tests (representing a fundamentally distinct population that appears across different psychometric measures). Their research lacks the application of TDA to a large sample data set as well as other environmental scenarios and no test was conducted to measure the validity of their study.

Bruno et al. (2017) applied Topological data analysis (TDA) methods in health (psychometry), where they identified two big divisions concerning the patterns of the structure of the images in the brain. The study stated that longitudinal TDA was chosen and used in separating individuals from Fragile X syndrome, compared to the cross-sectional TDA method. Other mental illnesses have not been studied using longitudinal TDA. They have not applied their method to other extreme events (floods) or environments. There is no validity test carried out to evaluate their study's method like the TDA-uML method in this study.

The Persistent homology, PH, has been used to analyze poetry data; the mapper algorithm (another TDA method) has been applied to analyze and visualize datasets in natural language classification (Lei, 2020). The study did not measure the performance of the two TDA methods used or explore to find the variation between