Stock Price Behaviour Analysis Based on The Topological Structure of Time Series, Application on Colombo Stock Exchange A Machine Learning Approach

K.K.R. Udayanga[1], R. Niranjan[1], R.P. Somawansa[1], E.G.T.S. Elaulla[1], K.P.P.J. Chandrasiri[1], and U.P. Liyanage[2*]

[1]Department of Mathematics, University of Colombo, Sri Lanka

[2]Department of Statistics & Computer Science, University of Kelaniya, Sri Lanka

[*]2019s17633@stu.ac.lk, [†]2019s17469@stu.ac.lk

December 2023

Abstract

The Colombo Stock Exchange (CSE) is where investors can interact in buying and selling financial instruments, allowing them to make profits through capital gain. The CSE includes many operations in 291 companies across 19 sectors. The stock market itself has an inherent nature of volatility and uncertainty. Due to this unpredictability, it's difficult to predict or model the stock price behavior. Throughout research many attempts were made to grasp this behavior, however, no general mechanism accurately models the volatility of these data and identifies similarities between them. This study aims to model stock prices using topological structures to find a mechanism to model these volatile data. Topological data analysis (TDA) is evolving in numerous domains and applied to various fields with distinct disciplines. In this paper, we perform behavioral analysis on Colombo stock exchange datasets since stock behavior is complex and volatile frequently, thereby investigating their pictorial characteristics. No studies were conducted in CSE using topological data analysis. In this research, we employed the Fourier series to decompose the complex signals into sinusoidal components, to analyze them and understand the time series. Persistent homology plots are the context of the TDA. These are used to capture and visualize the topological features of datasets in multiple scales and construct solid topological structures for different sectors. Then Principal component analysis (PCA) was used to reduce the dimensions of the large data set. Hierarchical clustering is used to show the multiple scales of data and helps identify the persistence features and distinguish significant structures from the noises. Therefore, It was used since we wanted to excavate deeper into the topological

structure to identify the sub-topological structures within these sectors and find the unique behavior. Finally, the results were compared using Wasserstein distance to find similarities between topological structures. From the results, we discovered specific stock behavior, and patterns and obtained, that the banking and finance sectors had more similar characteristics.

Keywords: Colombo stock exchange, Topological data analysis, Time series analysis, Behavioral analysis of stock data, Fourier series, Principal component analysis

1 Introduction

Topological data analysis (TDA) is a recent and fast-growing field providing a set of new topological and geometric tools to infer relevant features for possibly complex data. It proposes new well-founded mathematical theories and computational tools that can be used independently or in combination with other data analysis and statistical learning techniques [4]. Under TDA, persistence homology plays a major role in our study. Our utmost goal is to read time series behavior through topological structures in a quantitative manner. Thus, we used persistent homology, the idea behind persistent homology is finding the core topological features of our data that are hopefully robust to noise. For example, we know that the simplest polygon we can construct is the triangle, using triangles we can define any other polygon like a square consisting of two triangles, and a pentagon can be constructed using four triangles so on and so forth. So, if we analyze this idea more deeply, we can break down such polygons into several triangles, which is the core of that particular shape, likewise, this is what persistence homology does, however, most data are inherent with multidimensional properties in a complex manner so two-dimensional triangles are not enough. To overcome this issue mathematicians came up with a generalized version of a triangle to any number of dimensions, namely "simplexes", and the collection of simplexes is called "simplicial complex" and this is the key idea in persistent homology, the required proofs and mathematics behind this technique are presented by many authors in many books [1].

The stock market is a crucial platform for companies and investors to buy and sell financial instruments, making it essential for profit-making and dividend generation. Major stock markets worldwide, including the Colombo Stock Exchange (CSE) in Sri Lanka, record data on company stocks and financial bonds. It has 323 listed companies representing The CSE has 290 listed companies representing 20 business sectors as of 30 June 2019, with a market capitalization of 30,941 million US dollars. The CSE also oversees compliance through a set of rules, promotes standards of corporate governance among listed companies, and is actively involved in educating investors. In the course of its operations, the CSE interacts with many customers and stakeholders which include issuers (such as companies, corporations, and unit trusts), commercial banks, invest-

ment banks, fund managers, stockbrokers, financial advisers, market data vendors, and investors [3].

This research uses TDA and machine learning to analyze the behavior of stocks in the CSE, focusing on five sectors: banking, finance, hospital, hotel, and insurance. Identifying topological structures that can be used to analyze data behavior within the stock market, and comparing their pictorial characteristics are the objectives that are sought to achieve through this research. To exploit the complex topological and geometric structures underlying data often represented as point clouds in Euclidean or more general metric spaces [4]. As in the explanation of the given example, we found some core topological structures of the CSE for the selected sectors. further explanation is given in the methodology. This analysis employs persistent homology, Principal component analysis (PCA), hierarchical clustering, and Wasserstein distance to provide an in-depth analysis of stock market behavior and find the unique patterns of the stocks, aiding companies and investors in data-driven decision-making.

2 Methodology

The novel approach of employing topological data analysis (TDA) for stock price behavior analysis necessitates a comprehensive methodology to ensure the accuracy, reliability, and relevance of the results. The following sections outline the systematic procedure adopted for this study, encompassing data collection, preprocessing, topological feature extraction, model development, and validation. The first step in our methodology involves gathering extensive historical stock price data from reliable sources such as financial market databases, stock exchanges, or financial APIs. The data set should include various stocks across different sectors and time frames to ensure diversity and comprehensiveness. So, we collected data from CSE. After that, we did data preprocessing. It is crucial to cleanse the data and prepare it for analysis. This stage involves handling missing values, removing outliers, and ensuring data consistency. Additionally, we normalize the stock prices to mitigate the impact of scale differences across various stocks.

Then we apply TDA, and we need to transform the time series data into a point cloud in a high-dimensional space. We utilize Taken's embedding theorem to reconstruct the state space, ensuring that the topological features of the time series are preserved in the point cloud representation. With the point cloud-ready, we proceed to extract topological features using persistent homology—a central tool in TDA. Persistent homology captures the birth and death of topological features (such as connected components, loops, and voids) across multiple scales, providing a multi-scale description of the data's topology. We compute the persistence diagrams, a summary representation of the topological features, and the lifespan of features to quantify the topological structures in the stock price data. To make more analysis and find the behavior of the stock

prices, we used the hierarchical clustering method to cluster out each point in TDA plots. Then we listed similar clusters in 4 types, plotted time series for each plot, and did the TDA for each cluster. By that, we got more unique patterns for each period in the time series. It helps to find the similarity between sectors. We used Wasserstein distance to provide our results on whether sectors are similar or not [10]. Finally, we approached a method called Principal component analysis (PCA) that explained the variation used to find out the accuracy of our results.

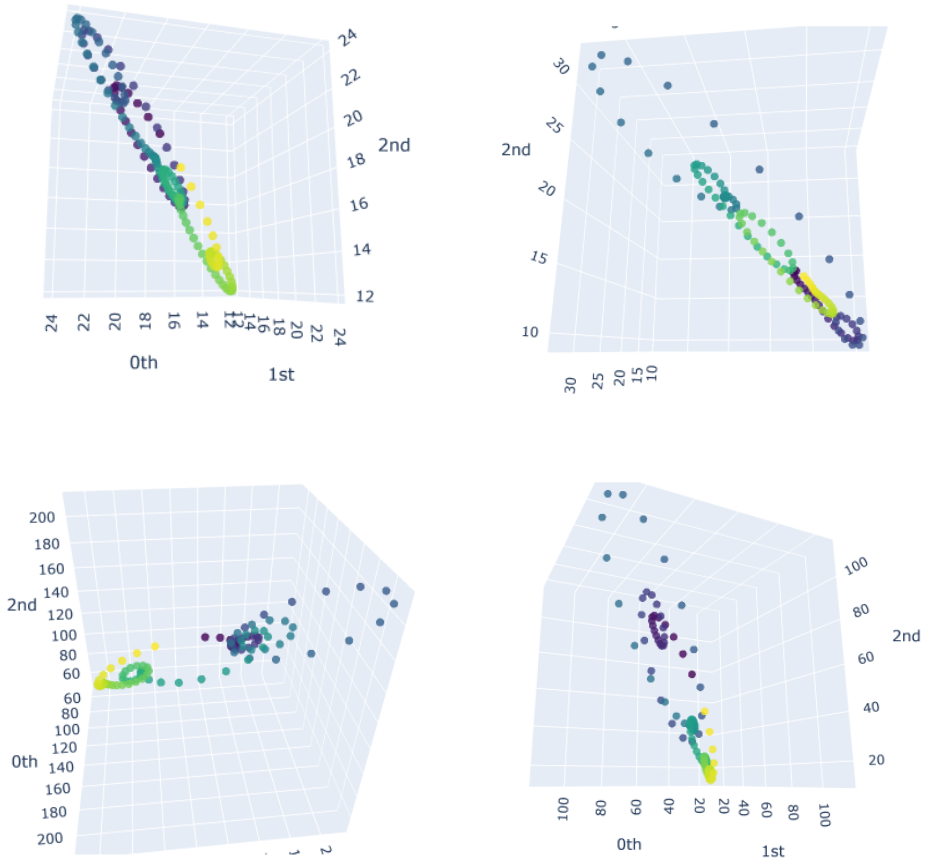


Figure 1: The resulting topological structures seen in the different sectors

2.1 Time series transformation of stock prices

As the initial analysis step, we plotted the time series for the stock market data. Then we selected a suitable time horizon for the time series such as capturing anomalies in the time series. To clear the noises and unnecessary oscillations

we approximate the selected time series by using the Fourier series to get the time series as a combination of cyclical patterns of various frequencies [9]. The main benefit of Fourier series approximation is that very little information is lost during the transformation [9]. The Taylor series approximation is a very famous concept in mathematics and is a local approximation i.e. it shows the behavior around a data point (neighborhood). However in our study, we deal with various time intervals, and the Fourier series shows the global properties of the function over the full-time interval, thus we developed our methodology based on Fourier series transformation. We can get topological structures within the data by applying the persistent homology. We could have a more fitted curve by increasing the Fourier components, but it might not give topological shapes. If we use over-fitted the topological shape is scattered in nature, if we use under-fit the topological shape does not give a complete image, we confined ourselves to using average Fourier components.

2.2 Projection of the topological structure of different stock prices

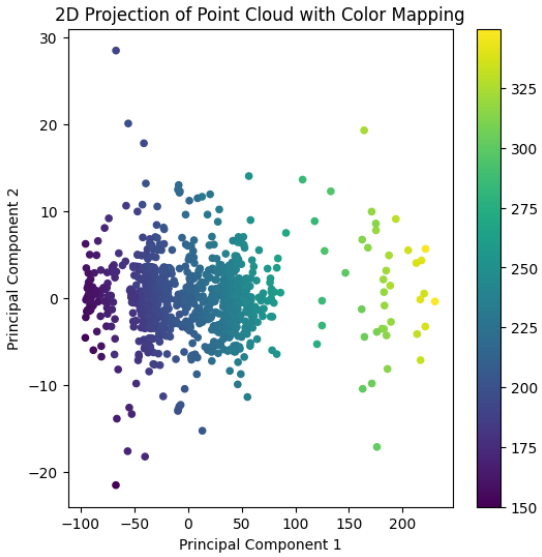


Figure 2: 2D Projection of Point Cloud with Color Mapping

By analyzing the Persistent homology plots we can observe that noises(h1) increase when increasing the embedding dimensions. Then we projected the topological graph into a two-dimensional plot determining pictorial characteristics. To achieve our goal we use PCA (principal component analysis) [11]. The reason to use PCA specifically for this part is it does not change or reduce the original information that is contained in the time series when it is mapped into

2D dimension. Since we have used PCA, we have to check the explained variation is at its satisfying range for each case. When we used this in our code implementation, the execution time was much higher when compared to other steps, we also observed that low volatility, and oscillations in the time series, resulted in much faster outputs.

2.3 Cluster classification

Under this step, we used Hierarchical clustering to identify the number of clusters from the projected topological structure. The key benefit of Hierarchical clustering is that there is no need to pre-specify the number of clusters. Instead, the dendrogram can be cut at the appropriate level to obtain the desired number of clusters [2]. However, in some cases, we manually defined the number of clusters to identify specific patterns. Then we observed that, If the number of clusters is below four or exceeds four, we cannot get a proper shape for the sub-topological structure (ST). In general, we got four clusters for each one of the data point clouds by using hierarchical clustering.

2.4 Formation of the sub-topological structure (ST)

For those clusters, we extracted time series, and we subsequently repeated our methodology to extract the sub-topological structure (STs) However, we couldn't obtain core topological shapes that were like the initial data set since we excavated deeper into the topological structure, therefore the amount of data left for analysis was small. To encounter that we used cubic spline interpolation, which is preferred over polynomial interpolation because the interpolation error can be made small even when using low-degree polynomials for the spline, it is the lowest degree that allows separate control on the two endpoints and two ends derivatives and it is also the lowest degree that allows inflection points [12]. These results represent the core of each sector's topological structures which we used to identify, examine their uniqueness, and perform behavior analysis.

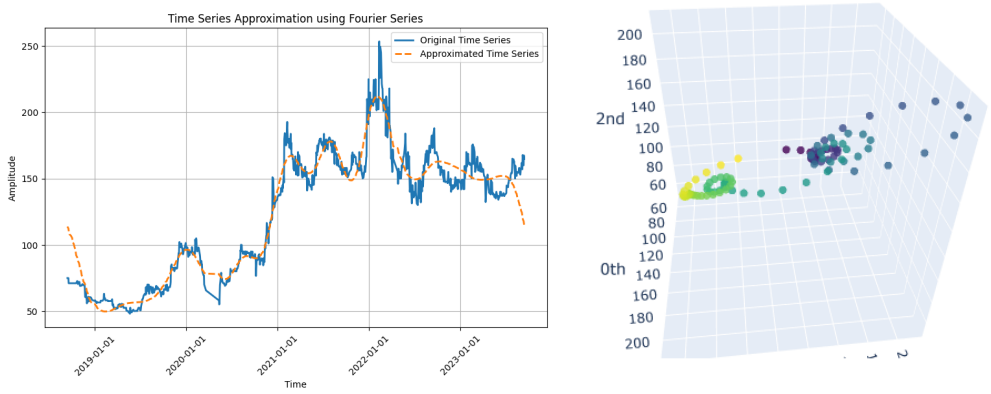
2.5 Identifying similarities and differences in stock prices through STs

The Wasserstein distance between two persistence diagrams can be used to quantify the dissimilarity or similarity between their topological features [5]. This distance metric provides a way to compare the persistence diagrams and, consequently, the underlying topological structures of different data sets. When the Wasserstein distances between two STs are low, it indicates that the similarities between two STs are high and vice versa. We also examined that the sectors that showed similarities in sub-topological structures, also inherent its main topology structure with compared one. Although this could have been done using the persistence diagram to the full stock data, our utmost aim was to capture the core topology, from this, we can assume that if the sub and main

topological structure are the same, then it provides a much stronger relationship between two companies customer behavior dynamics.[8]

3 Results

By examining the outputs of the core sub-topological structures, we could see that each sector indicated somewhat unique patterns, also industry expertise emphasized that the reason behind this could be the customer's preference, how they buy and sell financial instruments, and how they interact with the stock market. In a way, this gives the behavior dynamics of the stockholders in the Colombo Stock Exchange. The results were obtained by following the methodology step by step. Finally, the topological structures for the clusters were obtained as shown in the plots. In our study, we used 20–25 data sets purchased from the Colombo Stock Exchange and have been carried out for four different sectors. We considered five sectors, the Banking Sector, Hotel Sector, Insurance Sector, Finance Sector, and Hospital Sector. From each sector, 5 companies were analyzed. We've identified that each sector tends to emerge with similar sub-topological features(ST), which depict their unique nature and behavior over time. From the results, we discovered specific stock behavior, and patterns and obtained, that the banking and finance sectors had more similar characteristics.



(a) (b)

Figure 3: 3D embedding plot and a time series graph comparing a dataset with its Fourier series approximation

To showcase our methodology through an example we used Abans company's stock data, where Abans provides Electrical and Electronics Products and home Appliances, air-conditioning machines, Refrigerators, and freezers, and Abans Group has enjoyed more than 50 years as Sri Lanka's most trusted household name. As in methodology, we initiate our first step using Fourier series approx-

imation. Then we plotted the Embedding Visualization with Color Gradient, which allows the reader to distinguish the density of the price behavior like we used in the PCA process.

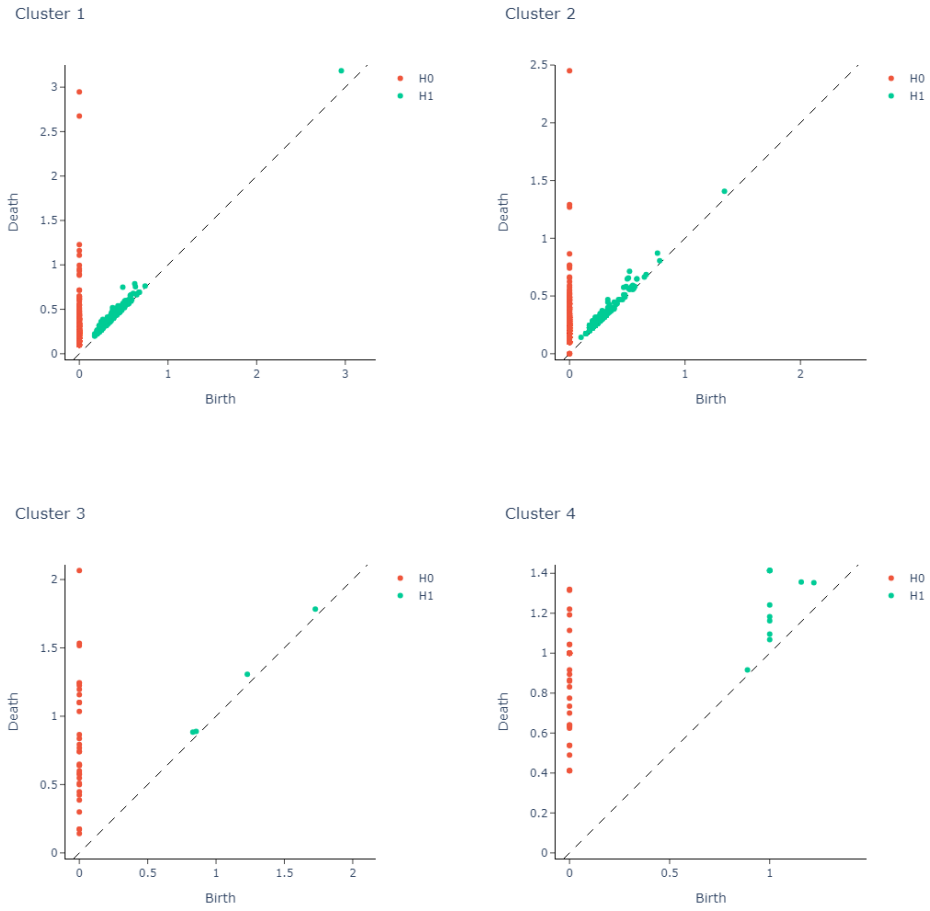


Figure 4: Topological Data Analysis of Multiple Clusters

Topological Data Analysis (TDA) is a method used to understand the structure and relationships within complex data sets, focusing on multiple clusters. It aims to uncover underlying topological features that may not be apparent through traditional methods. Clusters are groups of data points similar to each other based on certain criteria, and TDA examines how these groups are organized, interconnected, and differentiated in a high-dimensional space. Techniques like persistent homology and the Mapper algorithm help preserve the data's inherent topological properties, allowing visualization of high-dimensional data in a more comprehensible form. TDA provides insights into global and lo-

cal data structure, identifying patterns, trends, and subtle nuances within each cluster. This approach is particularly valuable in genomics and customer segmentation. TDA offers a deeper, geometrically informed understanding of data, providing a unique perspective that can reveal novel insights and drive more informed decision-making processes.

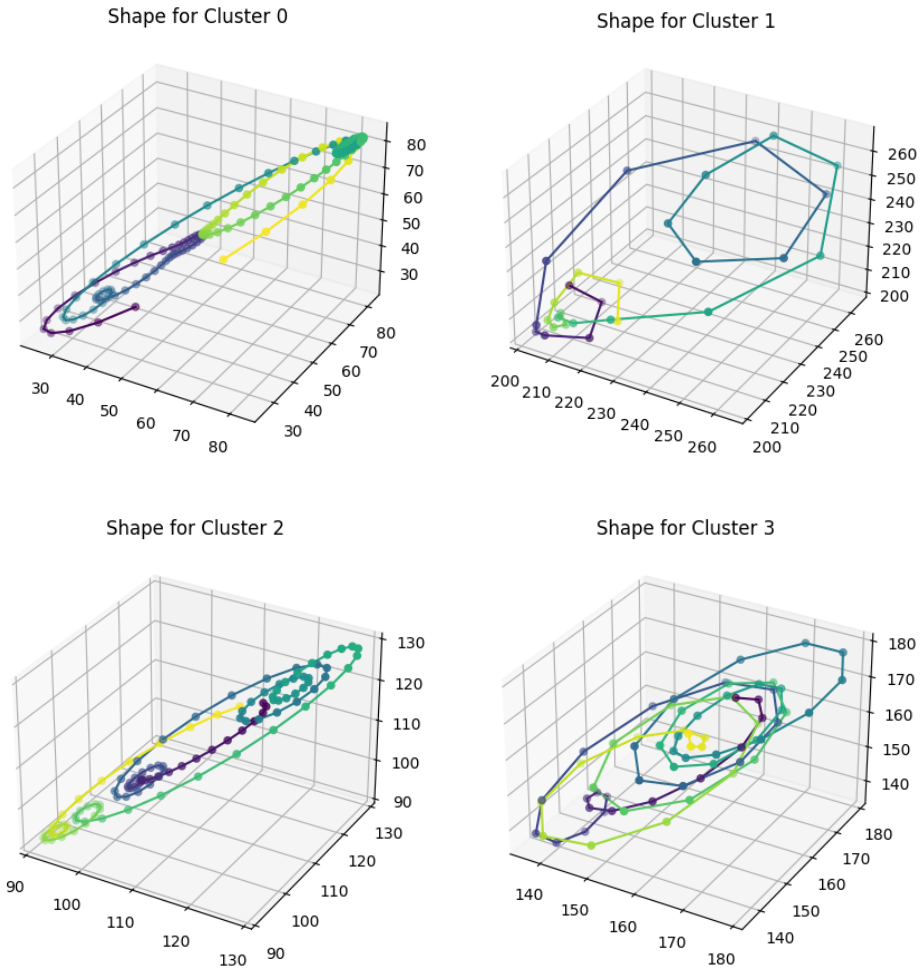


Figure 5: Cluster Configuration via Cubic Spline Interpolation

One technique used in Topological Data Analysis (TDA) to display complicated data structures is cubic spline interpolation. The resulting smooth, continuous curves reflect the underlying topological features as they traverse across the data points in a cluster. This method preserves the organic form and organization of the data, enabling a more sophisticated investigation of data clusters. The linkages and patterns that can be overlooked by linear or

less complex approaches might be highlighted by the smooth curves. It also improves the study of data discontinuity and continuity within clusters, providing information on possible outliers and anomalies as well as the density and distribution of data points. All things considered, cubic spline interpolation is an effective technique for the in-depth investigation of complicated data sets, permitting a greater comprehension of intrinsic patterns and relationships.

3.1 Financial Sector

Economics and finance are interrelated, informing and influencing each other. Investors care about economic data because they also influence the markets to a great degree [6]. When we consider the finance sector, there are three different types, personal, corporate, and public finance, each dealing with various types of policies and regulations directly linked with customers' behavior dynamics in the Colombo Stock Exchange. Through our analysis, the companies in the financial sector show more sophisticated topological shapes compared to other sectors.

3.1.1 Similarities Among Companies

John Keells Holdings (JKH) is one of the largest listed companies on the Colombo Stock Exchange, with business interests primarily in transportation, leisure, property, consumer foods and retail, financial services, and information technology. at the same time, Cargills Food City is the largest retailer on the island in all categories. Pursuing innovation and food safety its manufacturing brands Cargills Supremo and Cargills Finest (processed meats) Cargills Kist (processed fruits and vegetables) and Cargills Magic (ice cream and dairy products) lead sectoral growth.

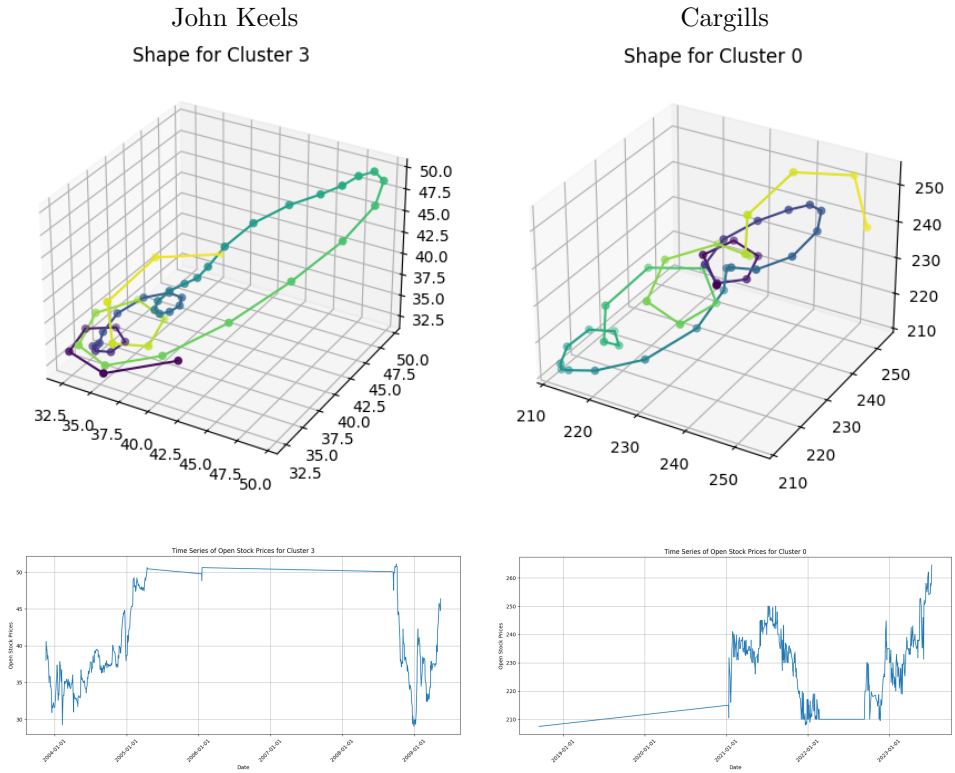


Figure 6: Topological shapes of John Keels and Cargills

Clusters that showed similar characteristics from the John Keells and Cargills companies have more loops in the shapes. As mentioned in the methodology, Wasserstein distance was used to measure the similarity with these topological graphs of clusters. Accordingly, it recorded a Wasserstein distance of 26.15. As shown below (Figure 4), even when the topological shapes seem similar the respective time series graphs are different.

As per the results, an observation was made that Wasserstein distances were smaller in between the cluster topological graphs of financial sector companies compared to some other sectors that were studied. Similarities between the shapes are more significant among John Keells and Cargills.

3.2 Banking Sector

The Banking sector plays an important role in the economy by helping the flow of money and providing financial services. It is a complex and highly regulated industry. Mainly they accept deposits, lend money, and provide various financial services to individuals, businesses, and the government. Key players in this sector are commercial banks, investment banks, and Central banks (responsible for monetary policy, issuing currency, and regulating other banks) [7]. The

customer base of the banking sector is more diverse than other sectors. It includes individuals, businesses, and the government. Each of these customer segments has unique needs and the banks they choose may vary according to their needs.

3.2.1 Similarities Among Companies

Sampath Bank and Commercial Bank are two prominent Sri Lankan banks that offer a variety of banking and financial services. Sampath Bank offers investment advisory and brokerage services, while Commercial Bank provides investment, corporate, and personal banking. Both banks are crucial for TDA studies due to their broad network and significant position in the banking industry. Fitch Ratings defines credit ratings for banks, and Sampath Bank and Commercial Bank are ranked equally in Fitch's report. The study aims to compare their results to see if TDA can provide the same results.

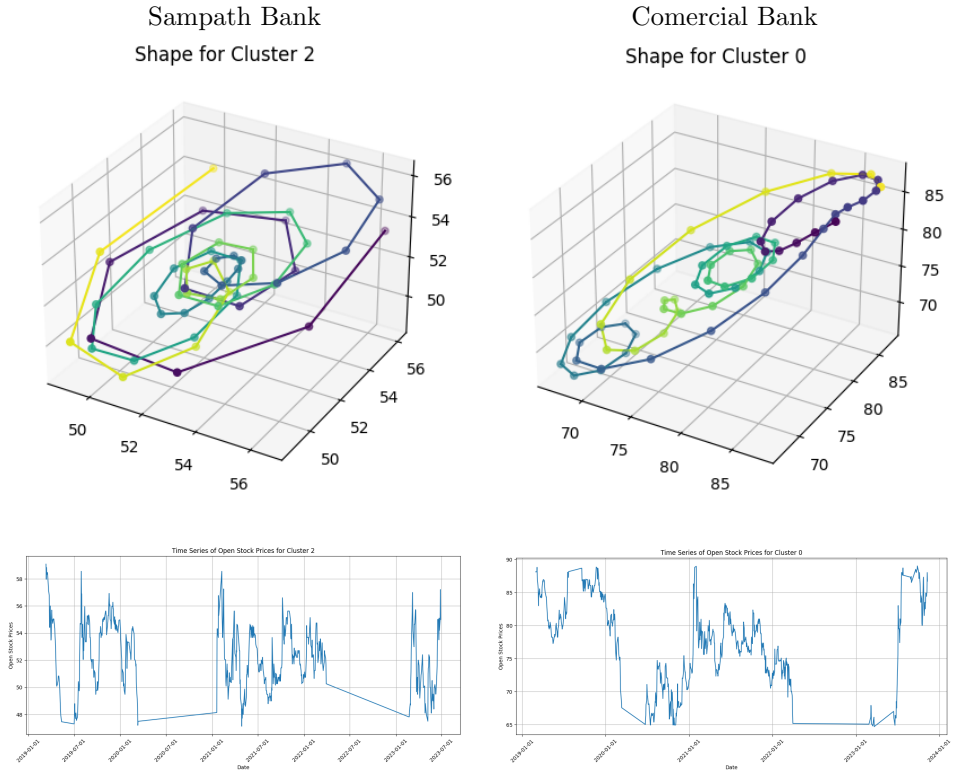


Figure 7: Comparison of Commercial and Sampath Banks

Among the five banks analyzed under the banking sector, we selected Sampath Bank and Commercial Bank. Commercial and Sampath banks show the lowest Wasserstein distances, which is 41.403. This is a high value compared to

some other sectors. Even Though the topological structures show similarities, their relevant time series show some differences compared to each other. Apart from above mentioned banks, clusters of other banks also generated well-defined topological graphs. Most of the clusters had a fair amount of loops in the topological structures whereas few of them had little number of loops. Another observation is that most similar topological graphs show values between 40 and 80 for Wasserstein distances.

3.3 Insurance Sector

We know that the terminologies of insurance and finance go parallelly as well as insurance and banks, but they have different business models and face different risks. While both are subject to interest rate risk, banks have more of a systemic linkage and are more susceptible to runs by depositors. While insurance companies' liabilities are more long-term and don't tend to face the risk of a run on their funds [3] i.e. less risk implies less volatility because of that unlike in the finance and banking sectors, the insurance sector depicted rather weak sub-topological structures.

3.3.1 Similarities Among Companies

People's Insurance PLC provides travel, property, marine, and auto insurance, and is an openly traded business that engages in active stock market trading and investments to maintain financial stability and pay claims. Janashakthi Insurance PLC offers life, health, and general insurance, and engages in stock market activity by issuing stocks and drawing capital. Both companies have significant influence on the insurance industry, participation in the stock market, and investment methods. The study found that insurance companies have weak topological shapes, with high dissimilarities, despite observable structures in Janaskthi and People's.

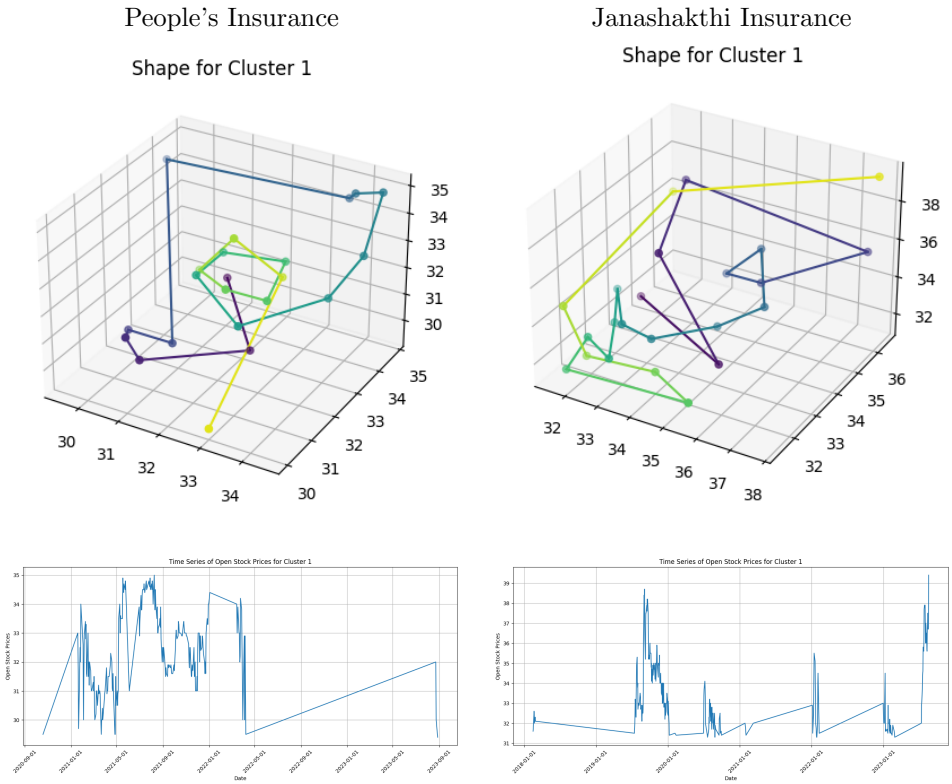


Figure 8: Comparison of People's and Janashakthi Insurance

3.4 Hotel Sector and Hospital Sector

It was clear from a detailed analysis of the data that the stock performance graphs of the hotel and hospital sectors both showed poor topological forms. This result implies that there are no discernible, recurring patterns, which demonstrates that stock behaviors in these industries are very variable and diverse. Remarkably, when the sectors were taken as a whole, the pattern of weak topological structures became more apparent. But when we looked more closely at individual companies in these industries, like large hospital groups with a reputation for cutting-edge medical facilities and all-encompassing healthcare services, or well-known hotel chains with a strong brand and global reach, we saw clearer, stronger structural trends in their stock performance. In the case of the hotel and hospital industries, these visible structures in particular organizations provide insight into potential underlying elements impacting stock performance, such as operational efficiency, brand reputation, and flexibility in response to changing market conditions. The intricacy and diversity of stock performance behavior in various businesses are shown by the disparity between the overall sector trend and particular entities.

4 Conclusions

The study's conclusion stressed that a promising method for examining stock market activities is to integrate topological data analysis (TDA) with machine learning. The distinct and intricate behaviors of the several industries represented on the Colombo Stock Exchange, especially the banking, finance, healthcare, hotel, and insurance sectors, were well captured by this approach. Different topological patterns were identified by the investigation in each sector, indicating that each has different market dynamics. Potential market interactions were suggested by the patterns that were comparable across the banking and finance industries. The paper highlights the potential of TDA in financial market analysis and suggests employing these approaches to construct sophisticated image classifiers for predictive analysis in the future. By offering a more sophisticated knowledge of market behaviors, this method has the potential to completely transform how market analysts perceive and anticipate stock market movements.

References

[1] Coreduction homology algorithm for inclusions and persistent homology. *Computers Mathematics With Applications*, 2010.

[2] Francesc Barberá Flichi. Micromechanics in additively manufactured metals using electron beam-based powder bed fusion. B.S. thesis, Universitat Politècnica de Catalunya, 2023.

[3] Bartleet Religare. Bartleet Religare Securities: FAQs, Year.

[4] F Chazal and B Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. front. *Artif. Intell*, 4:667963, 2021.

[5] Moo K Chung, Camille Garcia Ramos, Felipe Branco De Paiva, Jedidiah Mathis, Vivek Prabhakaran, Veena A Nair, Mary E Meyerand, Bruce P Hermann, Jeffrey R Binder, and Aaron F Struck. Unified topological inference for brain networks in temporal lobe epilepsy using the wasserstein distance. *NeuroImage*, 284:120436, 2023.

[6] Investopedia. Investopedia: Finance, Year.

[7] Investopedia. Investopedia: What Are the Major Categories of Financial Institutions and What Are Their Primary Roles?, Year.

[8] Meinard Müller. *Information retrieval for music and motion*, volume 2. Springer, 2007.

[9] Stuart Parsons, Arjan M Boonman, and Martin K Obrist. Advantages and disadvantages of techniques for transforming and analyzing chiropteran echolocation calls. *Journal of Mammalogy*, 81(4):927–938, 2000.

[10] Max Sommerfeld. *Wasserstein distance on finite spaces: Statistical inference and algorithms*. PhD thesis, Niedersächsische Staats-und Universitätsbibliothek Göttingen, 2017.

[11] Xin T Tong, Wanjie Wang, and Yuguan Wang. Pca matrix denoising is uniform. *arXiv preprint arXiv:2306.12690*, 2023.

[12] Kai Wang. A study of cubic spline interpolation. *InSight: Rivier Academic Journal*, 9(2), 2013.