Quantitative Finance and Economics

https://www.aimspress.com/journal/QFE

QFE, 8(4): 705–732.
DOI:10.3934/QFE.2024027
Received: 06 August 2024
Revised: 08 October 2024
Accepted: 24 October 2024
Published: 30 October 2024

Research article

A comparative study of symbolic aggregate approximation and topological data analysis

Fredrik Hobbelhagen and Ioannis Diamantis[*]

School of Business and Economics, Maastricht University, P.O.BOX 616, 6200 MD, Maastricht, The Netherlands

[*] **Correspondence:** i.diamantis@maastrichtuniversity.nl.

Abstract: The movement of stocks is often perceived as random due to the complex interactions between different stocks and the inherently chaotic nature of the market. This study investigated the similarity in stock movements across multiple industry sectors in Europe. Specifically, we applied topological data analysis (TDA) to analyze stock time series data and compared the results with those obtained using an expanded form of a more classical time series analysis method, symbolic aggregate approximation (SAX). Our findings indicated that while TDA offered detailed insights into "local" stock movements, SAX was more effective in capturing broader trends in financial markets, where less detail was required, making it suitable for portfolio optimization. We also presented an extension of SAX that incorporated volatility measures, improving its performance in highly volatile markets.

Keywords: topological data analysis; symbolic aggregate approximation; comparison, time series, stock markets

JEL Codes: C63; C65; C53; C18

1. Introduction

The behavior of stock prices is often assumed to be random because of the intricate interactions between various stocks and the inherently chaotic nature of the market. However, gaining a deeper understanding of which stocks move similarly, can offer significant insights into economic dynamics and it could potentially lead to the the creation of portfolios with less volatility and less risk. Similarity in this context refers to how closely one stock follows the behavior of another.

There are multiple classical approaches to this problem, most of which rely on aggregation. For example, one effective method is the symbolic aggregate approximation (SAX). The SAX method transforms time series data into a symbolic representation, making it easier to identify patterns and similarities between different stocks. By reducing the dimensionality of the data and discretizing it into a sequence of symbols, SAX allows for efficient comparison and analysis of stock movement patterns, enabling more informed decision-making in portfolio management. A newer and more complex approach to analyze data is topological data analysis (TDA). TDA focuses on analyzing the entire structure of data without distorting or manipulating it, producing more accurate results. It is often regarded as a particularly challenging type of analysis due to its requirement for a deeper mathematical understanding and the ability to work with complex high-dimensional methods and mathematical theorems.

Both SAX and TDA have already been applied in a financial context with promising results (see for example Liu and Shao, 2009; Gidea and Katz, 2018). However, there is a lack of comprehensive studies that compare the effectiveness of these methods specifically in the context of financial markets. Although TDA is typically used to detect black swan events, there has not been much research on continuous stock market analysis Majumdar and Laha (2020). As suggested in Wasserman (2018), there is an increasing interest in comparing TDA to other more classical and simplex methods in order to analyze data. This was the motivation behind this study. In particular, we believe that by providing more information on the applicability of TDA to specific financial problems, we may gain a deeper understanding of the chaotic behavior of the market.

TDA has been shown to outperform classical approaches in various types of analyses Majumdar and Laha (2020), but it is, however, computationally complex and it requires more time and resources. The purpose of this study is to compare SAX to TDA and identify differences between the two methods regarding the long-term movements of stocks. In particular, understanding the strengths and weaknesses of these two approaches is crucial and with this study we aim to:

- Identify which method provides more accurate and insightful analyses of financial time series data.
- Reveal opportunities for combining these methods.

The paper is organized as follows: In § 2 we discuss the process of data collection and in § 3 we present the SAX method and discuss the conclusions that can be drawn using SAX for our analysis. § 4 starts by recalling basic results on TDA, and we apply this method to the same dataset used for SAX. Finally, in § 5 we compare the results obtained from both methods and in the concluding section § 6 we discuss the results of our analysis and its limitations.

2. Data Collection

In this section we discuss the process of data collection. The data used for this study consists of the daily closing prices of stocks from 60 European companies, spanning from May 1, 2023, to May 1, 2024. The data was gathered using a Yahoo finance API 7 available for Python 3 Van Rossum and Drake (2009). Only European companies were included in the data to create a more manageable and concise analysis of a single economic area and to ensure a comprehensive analysis of the wider economy. In order to ensure a heterogeneous sample of companies, five stocks were selected from the majority of classes within the International Standard Industrial Classification of All Economic Activities (ISIC) United Nations (2008). This classification was created by the United Nations to classify all economic activities including a clear separation into different industries (see Table 1). For our purposes we considered the classes A to L and we did not include the other classes, since they either involve no companies at all (see, for example, class T), or they do not contain activities relevant to our studies (see, for example, classes U and S).

Table 1. A list of all economic activities with a short description according to ISIC.

Letter	Class
A	Agriculture, forestry, and fishing
B	Mining and quarrying
C	Manufacturing
D	Electricity, gas, steam, and air conditioning supply
E	Water supply; sewerage, waste management, and remediation activities
F	Construction
G	Wholesale and retail trade; repair of motor vehicles and motorcycles
H	Transportation and storage
I	Accommodation and food service activities
J	Information and communication
K	Financial and insurance activities
L	Real estate activities
M	Professional, scientific, and technical activities
N	Administrative and support service activities
O	Public administration and defence; compulsory social security
P	Education
Q	Human health and social work activities
R	Arts, entertainment, and recreation
S	Other service activities
T	Activities of households as employers; undifferentiated goods- and services-producing activities of households for own use
U	Activities of extraterritorial organizations and bodies

Note that in order to simplify the interpretation of the results of our study, we fixed the amount of companies within each sector. By a sector we mean the selected parts of an economic activity. We believe that this is helpful when comparing sectors within and between different economic activities. Note also that since our analysis uses stock data, smaller or privately owned companies are excluded, potentially introducing bias. The exact list of the companies included in our analysis is presented in appendix table.

Remark 1. While we selected an equal number of companies across industries to ensure uniform representation, we recognize that different industries vary significantly in size and market influence. An alternative could be selecting companies based on market capitalization or contribution to the sector's market dynamics. This would allow for a more weighted analysis of industry-specific trends. We chose to equalize the number of companies to maintain a balanced comparative approach between sectors for clustering purposes, as unequal representation might skew the clustering results.

As mentioned before, our aim is to identify the similarities of the movements of stocks using SAX and TDA. Note that this is a challenging task, since stocks may move in a complex way but also may have different scales. Figure 1 illustrates such an example, where the time series presentation of Skanska, H&M and Allianz stocks are considered.

3. SAX analysis

In this section we introduce SAX, a classical time series analysis method, first introduced in *"Experiencing SAX: a novel symbolic representation of time series"* (Lin et al., 2007). SAX simplifies a time series by reducing the number of observations through the aggregation of equal sections. These aggregated values are then binned and represented by a letter. The advantage of this approach is that it reduces the complexity of a complete time series into an easy-to-use and informative ordered collection of letters called *words*. This also simplifies the handling of time series and makes comparing multiple time series computationally easier with the major drawback being loss of information. The efficacy of SAX in financial data analysis has been well-documented in Liu and Shao (2009).

Sections 3.1 and 3.2 provide an overview of well-known estimators and stratified randomization methods. While these are foundational techniques, we include them here to ensure the paper is self-contained and accessible to a broader audience, including those who may not have a strong background in these mathematical concepts. This review also serves as the necessary groundwork for the novel contributions introduced later, where we present new insights into the integration of covariate adjustment with stratified randomization.

3.1. SAX algorithm

We start by providing a step-by-step algorithm to construct SAX from a given time series:

1. We begin with normalizing the time series to have a mean of 0 and a standard deviation of 1. This step ensures that the data is on the same scale and helps in comparing different time series. There are several methods to normalize time series data and for our purposes we will be using the z-score normalization (Lin et al., 2007).
2. We then divide the time series into equal sized time sections called time frames. The choice of the amount of time frames is arbitrary.
3. We proceed by calculating the piecewise aggregate approximation (PAA) for each time frame. To do so, we first evaluate the mean within each time frame and the PAA transform is obtained by averaging the points within each segment:

$$\bar{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j \tag{1}$$

where n is the length of the time series and w is the number of time frames.

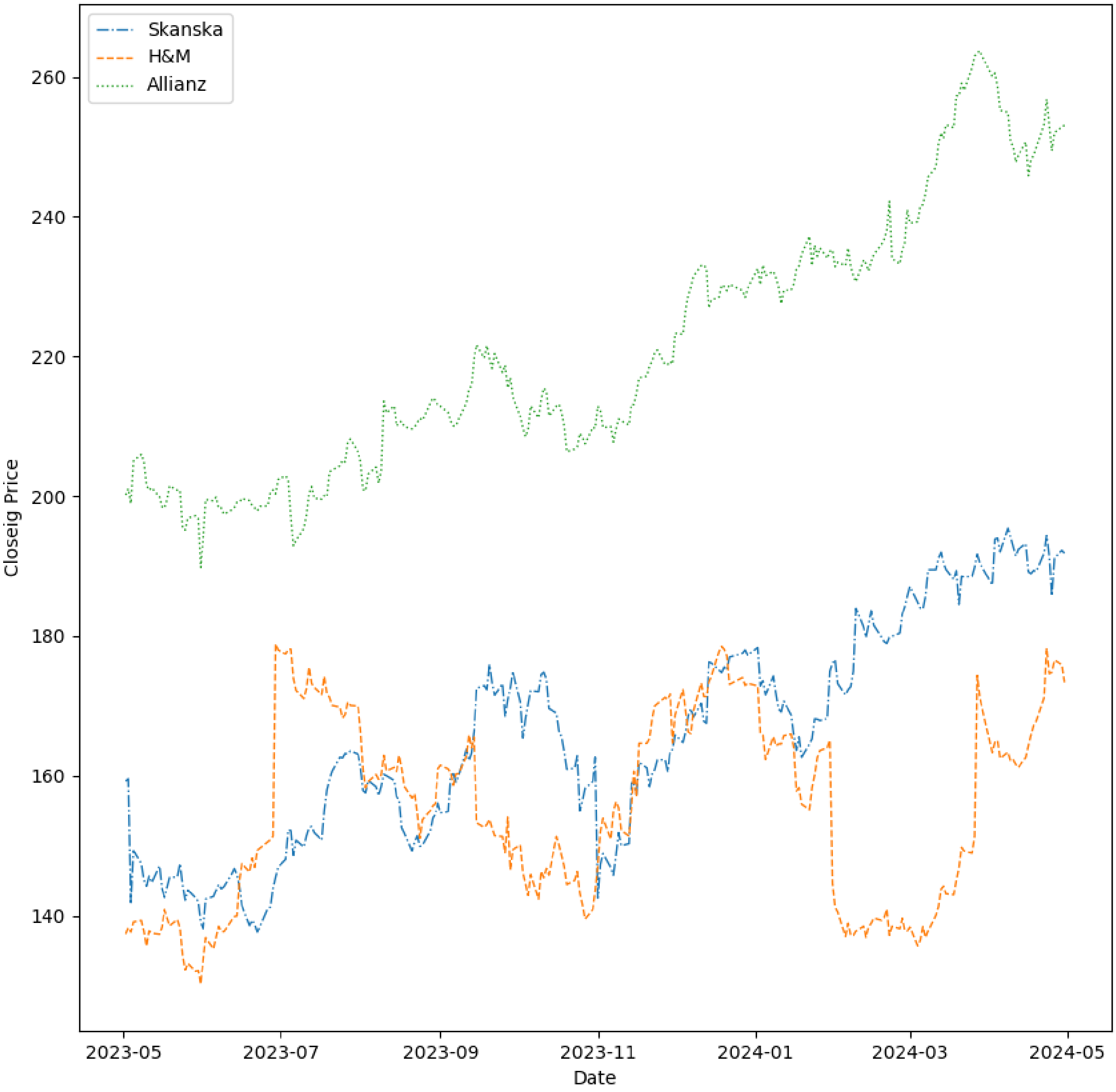

Figure 1. Time series presentation of the Skanska, H&M and Allianz stocks.

4. We now consider breakpoints that correspond to the cumulative distribution function (CDF) of a standard normal distribution. The breakpoints partition the range of PAA values into α intervals. We then map each PAA value to a letter based on which interval it falls into. Note that the number of intervals, i.e. α, determines the alphabet size. This procedure transforms the PAA values into a sequence of symbols $\hat{C} = \{\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_w\}$, i.e. it creates a symbolic representation of the time series as a *word* as follows: Let α_i denote the i-th element of the alphabet, i.e., $\alpha_1 = a$ and $\alpha_2 = b$. Then the mapping from a PAA approximation $\bar{C}$ to a word $\hat{C}$ is obtained as follows Lin et al. (2007):
$$\hat{c}_i = \alpha_j \quad \text{if} \quad \beta_{j-1} \leq \bar{c}_i < \beta_j \tag{2}$$
5. The final SAX representation is the sequence of symbols $\hat{C}$. Each symbol represents the normalized value of a segment of the time series, discretized according to the chosen alphabet. In this paper the alphabet consists of the lower integer part of the mean value for each time frame.

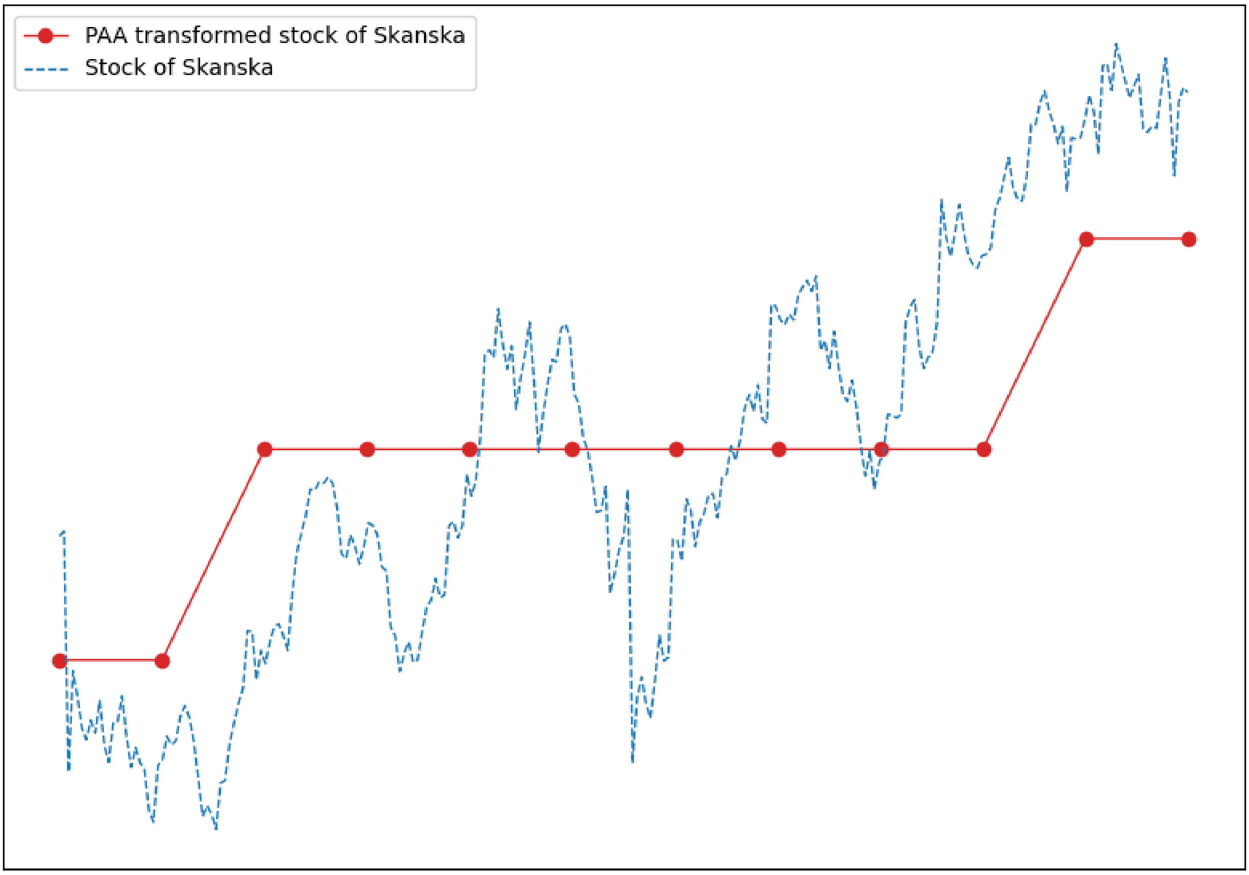

Figure 2. An example of the PAA method on a randomly selected time series.

Note that SAX results in the loss of information in favor of a simplified representation of the time series. An example of this process is presented in Figure 2. The blue line shows the movement of the Skanska stock before transformation and the red line shows the PAA transformation of this stock. We observe that the movement of the stock is still captured in the PAA representation of the time series, despite the simplifications made. Note also that the simplified representation of the time series makes determining distances between time series easier and allows for a variety of clustering techniques to be applied.

Remark 2. In Canelas et al. (2013) and Leitão et al. (2016) SAX's utility in reducing the dimensionality of financial stock data is explored, integrating it with advanced techniques like genetic algorithms. More precisely, in Canelas et al. (2013) SAX parameters were fine-tuned to identify pattern-rich areas in financial data, and in Leitão et al. (2016) the authors adapt SAX by focusing on perceptually important points (PIPs) in each time step, selected based on significant amplitude differences.

An enhanced SAX method that incorporates various measures for each time step, including maximum and minimum values, has been introduced in Lkhagva et al. (2006). This extended approach proves particularly effective for financial time series, where extreme points often convey more significant patterns than mean values. The difference between this enhanced SAX method and the classical SAX approach lies in the third step of the algorithm (PAA). The enhanced SAX method is presented in the next paragraph.

3.2. Extended SAX Method

The SAX method implemented in this study is the extended SAX, introduced in Lkhagva et al. (2006). This approach uses the mean, minimum, and maximum values to aggregate each time frame. The result of this method combines the output of the three aggregations into a single word for each time series. To do so, Equation 3 is used, where the distance between one letter of a word from a stock to the corresponding letter of a different stock is calculated. This is applied over all SAX methods and scaled using $\sqrt{\frac{n}{N}}$, where n corresponds to the length of the initial time series and N corresponds to the number of the reduced dimensions, i.e the number of time frames used (for more details the reader is referred to Keogh et al. (2001)).

$$\sqrt{\frac{n}{N}}\sqrt{\sum_{i=1}^{N}(\bar{x}_i - \bar{y}_i)^2} \tag{3}$$

Since this version of SAX is designed to be used on financial data, it is a natural choice for this analysis. Note that, for reasons of homogeneity, we are using the same PAA processing steps for each descriptive statistic as described in the algorithm above. This allows us to focus more on the movements of stocks rather than their respective scale.

We set the number of time frames to 12 (corresponding to the months of the year), in order to effectively apply SAX, but also the TDA method (for more details, see Section 4). Figure 3 illustrates an example of the extended SAX transformation applied to the same stocks as in Figure 1. Note that the piecewise linear segments represent the mean of the corresponding time series, while the shaded areas represent the minimum and maximum values, capturing the variance of each time frame. We view the mean value of the time series as the blueprint of the variance captured, corresponding to the shaded areas.

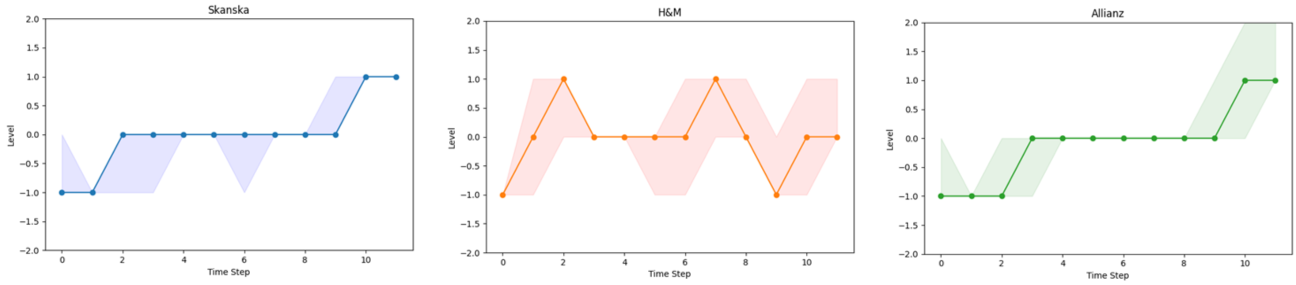

Figure 3. Application of the extended SAX on example stocks from Figure 1.

3.3. Results & Conclusions

In this subsection we present the results of the SAX analysis on the normalized stock data that was considered in this study.

To create clusters of stocks, a simple hierarchical clustering approach is applied using the distances between the stocks. The resulting dendrogram can be seen in Figure 4 with a threshold of 34.3. This dendrogram splits the stocks into three clusters with six unclassified stocks. The yellow cluster consists of 2 stocks, the green cluster consists of 39 stocks, and the red cluster consists of 13 stocks. Note that the blue cluster represents the six stocks not being classified to any cluster. Although the clusters are different, the distance between them is very small, meaning there is not much confidence of which stock belongs to which cluster. Additionally, one cluster contains the majority of all observations from the dataset, resulting in a less informative result as the majority is classified as similar.

Despite these issues, the SAX analysis does still capture some similarities in the movement of stocks, as illustrated in Figure 5. These visualized stocks were all classified in the red cluster and had some common movements. All stocks have an initial decrease in price with a minimum around November of 2023, followed by a brief increase for a few months, before returning to a lower price level. However, the exact movement of the stocks is not captured with DHL, a German logistics company, having a steady stock price after the bump, while the other examples have a sharp increase in the later months. DHL also has an earlier increase around August which is not observed in the movements of the other stocks.

4. TDA

TDA is an emerging field that applies techniques from algebraic topology to study the shape of data. This method is widely considered to be a particularly complex type of analysis focusing more on the mathematical structure of data, while ensuring no loss of information (Wasserman, 2018). This is the main reason why there are very few examples of TDA being applied in a financial context, despite some promising results in fields like biology (Skaf and Laubenbacher, 2022) or computer science (Bendich et al., 2010). In this section we provide a gentle introduction to some foundational concepts of TDA, assuming only an elementary background in linear and abstract algebra, while focusing on concepts that we apply in our analysis. It is worth mentioning that TDA comprises multiple techniques that share the underlying principle of preserving the topological structure of the data, and thus avoiding the simplification of the data, which implies that the outcome is more precise. Finally, note that while methods such as regression adjustment and re-randomization have been well studied in stratified

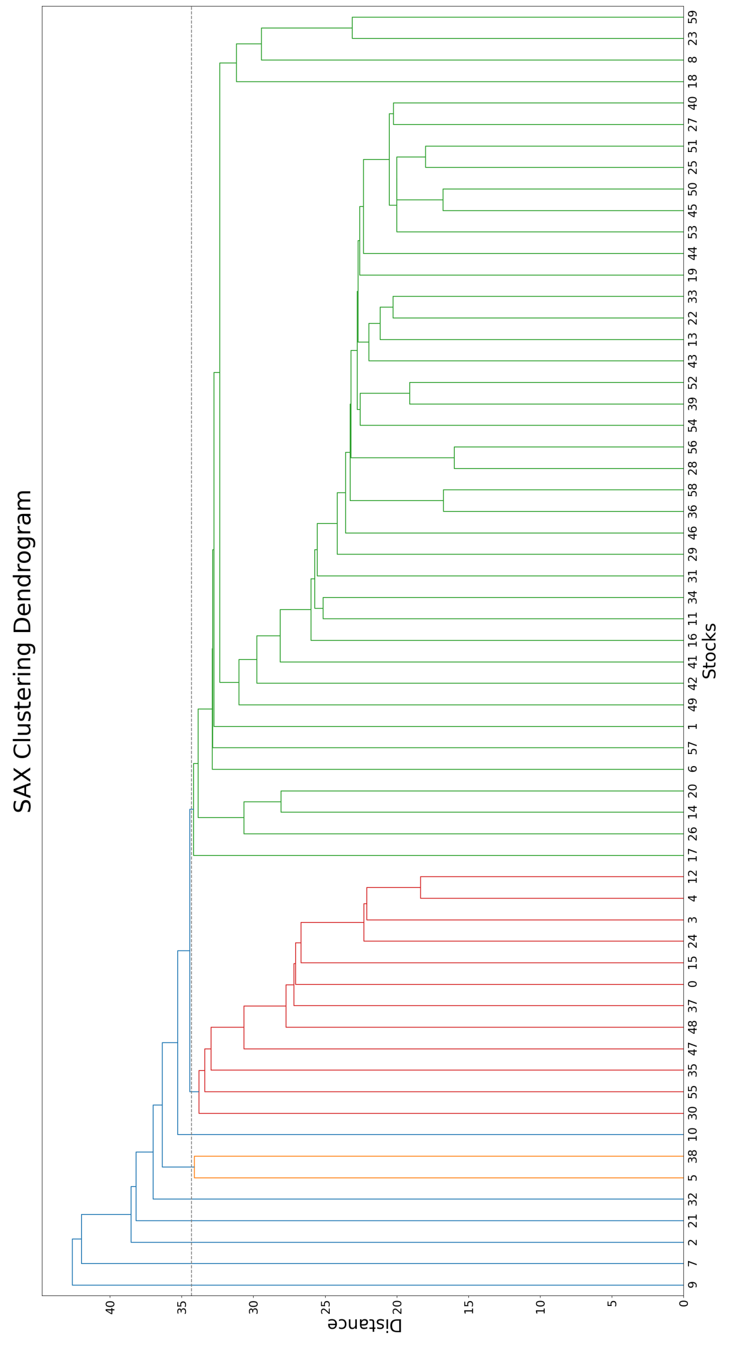

Figure 4. A dendrogram showing the result of the hierarchical clustering.

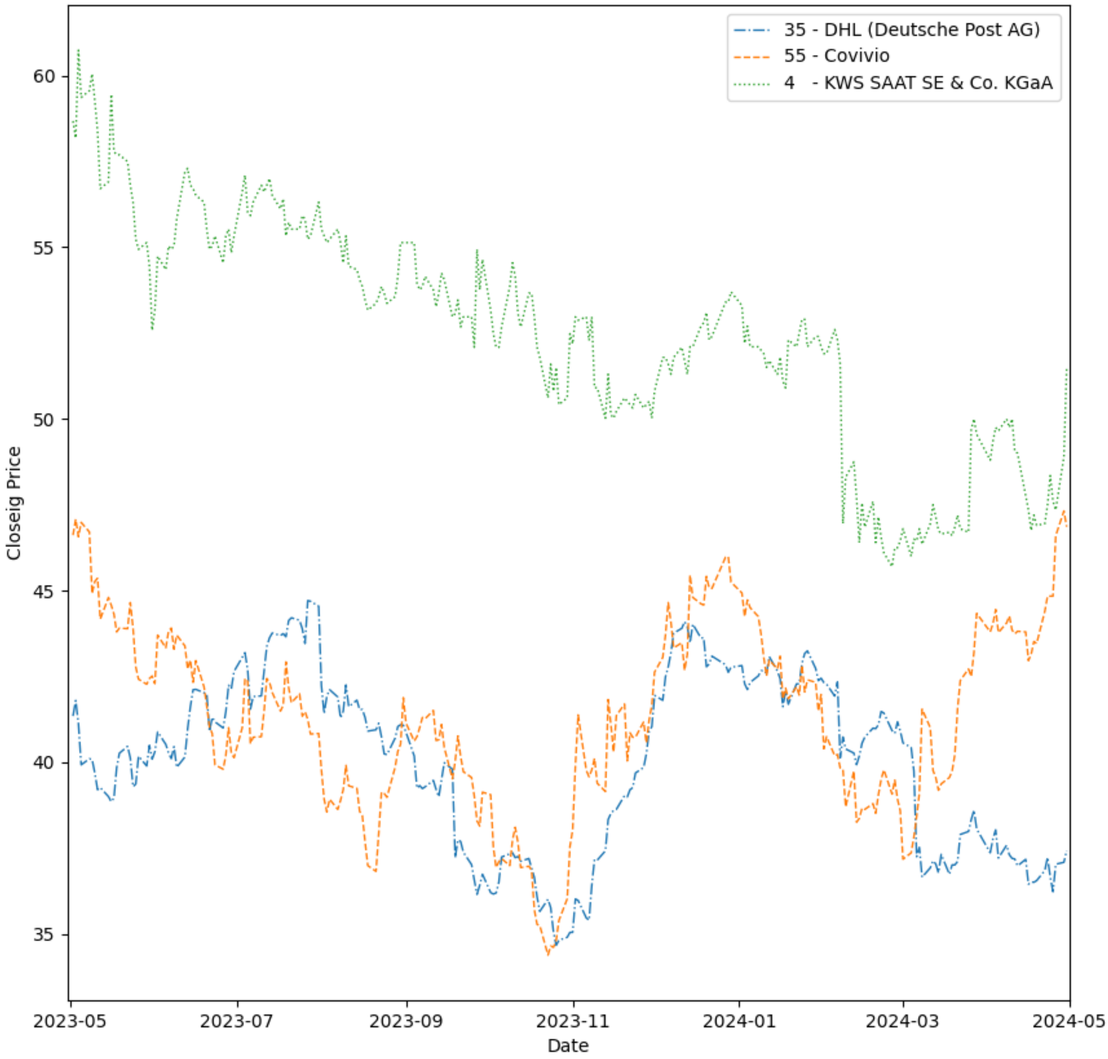

Figure 5. An example of stocks from the red cluster with differences in smaller sections and common trajectories.

randomized experiments (Liu and Yang, 2020; Li and Ding, 2020), our application of TDA provides a more comprehensive analysis of stock market data by capturing topological features that traditional regression models might overlook.

4.1. Preliminaries

In Skaf and Laubenbacher (2022), two common approaches of TDA are discussed, that is, persistent homology and mapper. Persistent homology is a TDA technique to discover the topological features in the data. Topological features can be thought of as holes and loops within the data and can be captured by the structure of the *homology groups*. Homology groups provide a way to categorize and quantify the topological features of data at different dimensions. They offer a powerful tool for understanding the underlying structure of complex datasets by breaking them down into their essential topological components.

At the simplest level, homology groups count the number of connected pieces in a space. For example, if there are two separate clusters of data points, the zeroth homology group H_0 would capture this by having two generators, one for each cluster. Moving to the next level, homology groups identify holes or loops in the data. For example, if the data forms a shape with a single loop (like a circle), the first homology group H_1 would capture this loop, since each loop or hole in the data corresponds to a generator in the homology group. Homology groups can also capture higher-dimensional voids. For instance, a hollow sphere has a 2-dimensional void inside it, which would be identified by the second homology group H_2.

To identify these features, persistent homology continuously connects observations that are close to each other, gradually forming a network. In this process, new topological features, also referred to as *births*, emerge as connected components. As more connections are made, some of these features, such as holes or loops, will eventually merge or fill in, leading to their *death*. The birth and death of topological features can be visualized using tools such as *barcodes, persistence landscapes*, or *persistence diagrams* (for an illustration, see Figure 6 (Karan and Kaygun, 2021)).

- A barcode is a visual representation of the persistence of topological features across different scales in a dataset, and it provides a clear way to see which features persist over a range of scales and which are short-lived, helping to distinguish noise from significant structures in the data. A barcode consists of horizontal line segments (bars) that represent the lifespan of each feature. Each bar starts at the value where a feature appears (birth) and ends where it disappears (death).
- A Betti curve is a plot that shows the number of topological features (such as connected components, loops, or voids) that are present in the data at each scale. They summarize the topological features in a more condensed form compared to barcodes, providing an overall picture of their persistence. Betti curves plot Betti numbers, which count the number of features for each dimension at a given scale. By plotting the Betti numbers against the scale parameter, Betti curves give a global view of how the topological complexity of the data changes as the scale varies.
- A persistence landscape is a way to represent the persistence of topological features that combines the information from barcodes into a smooth, continuous function. They are useful for statistical analysis and comparison because they convert the discrete information of barcodes into a continuous form that can be more easily averaged and compared across multiple datasets. Each feature contributes a ”tent” function to the landscape, where the peak corresponds to the midpoint

of the feature's lifespan, and the height decreases linearly to zero at the birth and death points. The landscape is formed by summing these tent functions, creating a series of landscapes for each dimension of features.

These methods provide a simple overview of the topological features in a dataset and they are useful in making comparisons between different datasets.

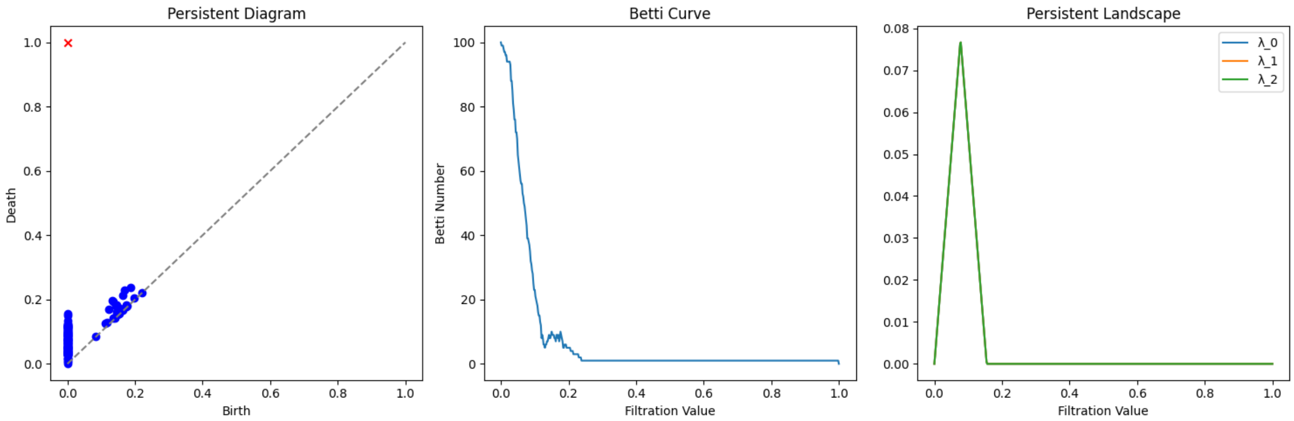

Figure 6. (a) A persistent diagram with the axis representing the birth and death of topological features, presented as dots. (b) A barcode which visualizes the length of topological features via the length of the bar with each new feature created getting a new bar. (c) A persistent landscape which is comparable to the persistent diagram.

As discussed in Karan and Kaygun (2021) and Perea and Harer (2015), in order to apply TDA to time series data, adjustments should be made. Both studies utilize a similar approach by first dividing a time series into equal-sized segments called windows. Since these windows can overlap, they are called sliding windows as one window also contains information from another. To use this data in persistent homology, the data within a window is embedded from a univariate time series into a higher dimensional space using a method called *"time-delayed embedding"*. Time-delayed embedding is used to transform a univariate time series into a higher dimensional point cloud while preserving the topology of the time series according to Takens theorem (Khasawneh and Munch , 2017). Takens theorem states that for a generic smooth dynamical system, the dynamics of the system can be reconstructed from the time series data of a single observable. Specifically, if the original system has dimension d, then the state space of the system can be embedded into a higher-dimensional space, typically $\mathbb{R}^{2d+1}$, using time-delayed coordinates. For an illustration of this process, see Figure 7 (Karan and Kaygun, 2021). The time series (a) and (b) are embedded into a higher dimensional space illustrated in (c), (d), (e), and (f). A time series in higher dimensions is comparable to dynamical systems and as can be seen the choice of embedding dimension has a large effect on the shape of the data.

A time series is embedded into higher dimensions by transforming it into multidimensional vectors, with the length of the vector m corresponding to the dimensions of the embedding. Using a time delay parameter τ, values from the time series are sampled to form these vectors. For example, with $\tau = 2$ every second value from the time series is included in the vectors. Although both m and τ significantly affect the shape of the time series in higher dimensions, there is no universally optimal way to choose these values. Therefore, discretion must be used when applying this method to find an

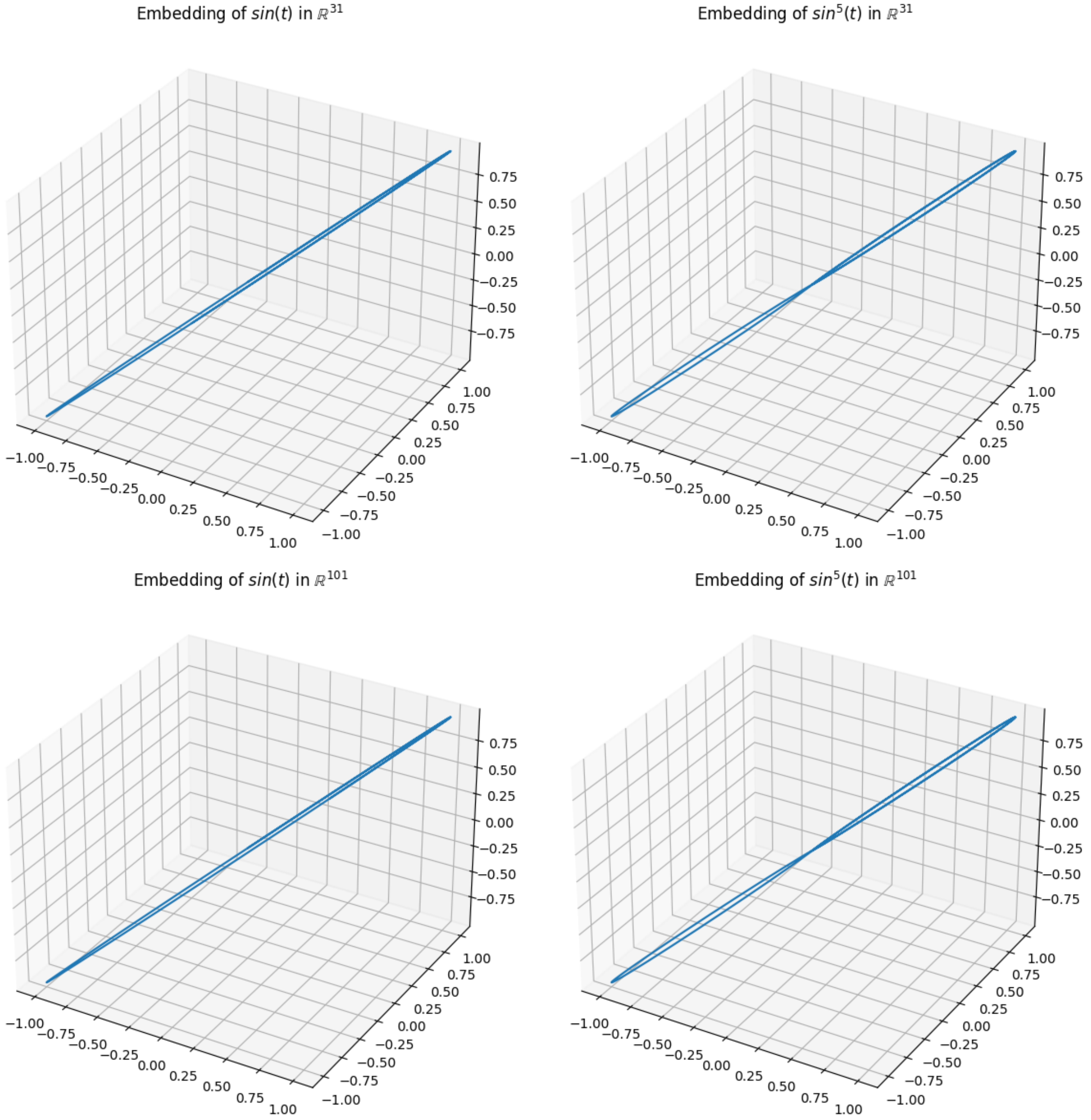

Figure 7. Embedding of two time series (a) and (b) into higher dimensions.

appropriate range for both m and τ. Toward that end, we will be using the *false nearest neighbour* (FNN) method, as discussed in Kennel et al. (1992). False neighbors appear when embedding; these are points that seem close in lower-dimensional space but move apart in higher dimensions. A false neighbor is identified by the following conditions:

$$\frac{R_{d+1}(n)}{R_d(n)} > R_{tol} \tag{4}$$

$$\frac{R_{d+1}(n)}{R_A} > A_{tol} \tag{5}$$

These formulas describe the conditions under which a neighbor is classified as a false nearest neighbor. If either of these conditions is met, a point is classified as an FNN. For this analysis, a lower-dimensional point cloud is only compared to its next higher dimension. In Eq. (4) a ratio is considered between the distance to the nearest neighbor in the lower dimension, $R_d(n)$, and the distance between the same points in the next higher dimension, $R_{d+1}(n)$. Therefore, a point is classified as a false nearest neighbor if the ratio is more than R_{tol} percentage points away, when in a higher dimension. R_{tol} is an arbitrary threshold that is typically set at 10%. Eq. (5) is a supplementary condition to Eq. (4), as it is only necessary when there are only a few data points that are close together. The same distance to the nearest neighbor in a higher dimension is used, divided now by the value of the attractor in a dynamical system R_A. Since this is not always known, R_A can be approximated with the standard deviation of the distances in the lower dimension Kennel et al. (1992). A_{tol} is also a threshold typically set at 2. Hence, using Eq. (5), a point is classified as false neighbor if the distance is significantly different from the mean distance of the lower embedding.

Both equations are used to calculate an FNN ratio for every increasing dimension to a set point and all feasible values for τ. Ideally, the FNN ratio would be 0. When this is not feasible, selecting the dimension with the fewest FNN and the lowest dimension is the best choice. Note that other methods work in similar ways (see for example (Liebert et al., 1991)), however, the FNN approach is the most commonly used in the literature, and it is also the one used in our analysis.

It is worth mentioning that in Karan and Kaygun (2021), the authors employ sub-windowing, where the sliding windows are further split to mitigate noise and reduce computational time.

One of the key properties of persistent homology is *stability*. Small changes in the input data result in small changes in the persistence diagram. This makes persistent homology robust to noise and ensures that significant topological features are consistently detected. In order to understand stability of data, we first need to introduce the concept of *distance* in TDA. As mentioned before, in the context of persistent homology, persistence diagrams capture the birth and death of features such as connected components, loops, and voids. Comparing these diagrams helps us understand the similarity between different datasets or different filtrations of the same dataset. The Wasserstein distance provides a meaningful and flexible way to quantify this similarity.

Definition 1. *Given two persistence diagrams D_1 and D_2, each consisting of points in the plane representing birth and death times of features, the p-Wasserstein distance is defined as follows:*

$$W_p(D_1, D_2) = \left(\inf_{\gamma : D_1 \to D_2} \sum_{x \in D_1} ||x - \gamma(x)||^p \right)^{\frac{1}{p}} \tag{6}$$

where γ ranges over all bijections between D_1 and D_2 (augmented with points on the diagonal if necessary), and $\|\cdot\|$ denotes the L^p norm (typically the L^2 norm for practical purposes).

Remark 3. The bottleneck distance is another metric used to compare persistence diagrams. It is defined as:

$$d_B(D_1, D_2) = \inf_{\gamma:D_1\to D_2} \sup_{x\in D_1} \|x - \gamma(x)\|. \tag{7}$$

This distance measures the maximum distance any point in D_1 that needs to be moved to match a point in D_2. It is closely related to the ∞-Wasserstein distance, which can be written as:

$$W_\infty(D_1, D_2) = \lim_{p\to\infty} W_p(D_1, D_2) = \inf_{\gamma:D_1\to D_2} \sup_{x\in D_1} \|x - \gamma(x)\|. \tag{8}$$

Thus, the bottleneck distance is a special case of the Wasserstein distance where $p \to \infty$.

Example 1. Consider two persistence diagrams D_1 and D_2 representing the topological features of two datasets. Using the Wasserstein distance, we can quantify how similar these datasets are in terms of their topological structure as follows:

- For $p = 1$, the Wasserstein distance measures the average amount of work needed to match the points in D_1 to those in D_2.
- For $p = 2$, the distance gives more weight to larger discrepancies, highlighting significant differences between the diagrams.
- For $p \to \infty$, the Wasserstein distance converges to the bottleneck distance, focusing on the maximum difference.

By comparing the Wasserstein distances for different values of p, we can gain insights into the nature of the differences between the datasets and the significance of their topological features.

In Truong (2017), the abovementioned methods have been translated to financial time series and have been applied to foreign exchange data in nanosecond intervals. More precisely, the authors use the techniques of sliding windows and time-delayed embedding to get insights into financial time series as well as the effects of quantum noise on the TDA analysis. However, an additional step is used to lessen the computational burden, namely, principal component analysis (PCA). PCA identifies the components that explain the most variance in a system and they can reduce the complexity of the embedding cloud by multiple dimensions Gao et al. (2016). Although there is a loss of information involved, this is sometimes inevitable in order to achieve feasible computational times in an analysis.

4.2. Method

To cluster stock data using TDA, we follow the same approach as in (Karan and Kaygun, 2021). In particular, we separated all time series into 12 windows, each window roughly corresponding to a month of the year. Note that this is similar to the number of time frames we considered in the SAX method. However, since the time series is shorter than 365 days, each window consists of 21 observations. In this way we guarantee that there are enough observations in a window to be informative, as well as being granular enough to allow for a fair comparison with the results from the SAX method.

To justify the comparison between SAX and TDA, we do not use sliding windows, that is, we only consider nonoverlapping windows, and thus, the information from a window is independent from the information of another window. These windows are then embedded using time-delayed embedding into five dimensions, $\mu = 5$, and a delay of three, i.e. $\tau = 3$. These parameters were set using the FNN approach as described in Kennel et al. (1992). We start from a 1-dimensional time series and add the next observation with a delay, until the appropriate dimension is reached. If there are no more observations in the window available, a 0 is added. Note that this approach allows for more observations in a single window without using information from the next window, which would be possible with a sliding window. As a result, each embedded window consists of a 7-dimensional point cloud containing 21 points. For each of the point clouds (windows) in the time series, persistent homology was applied using the ripser library (Tralie et al., 2018). The difference between the two persistent homologies was calculated using the Wasserstein distance, resulting in a similar outcome to SAX, as each time series is represented by 12 separate structures. We used the Wasserstein distance for its sensitivity to all features, since each window has a small number of close-together observations, and the Wasserstein distance is more sensitive to the small changes within each window. As for SAX, hierarchical clustering was used to create a dendrogram with a threshold of 4000.

4.3. Results & Conclusions

In this subsection we present the results of the TDA analysis. In Figure 8, the FNN analysis is illustrated, with the different lines showcasing the effect of different values for the delay. The shaded area visualizes the standard deviation. For this analysis, an embedding dimension of $m = 5$ with a delay of $\tau = 3$ was selected. The values are aggregated over all windows in all stocks, with the lines representing the mean value and the shaded areas visualizing the standard deviation. Each line shows the embedding with a different delay (τ). The number of dimensions investigated for each FNN was limited by $\frac{m}{\tau}$, since any value after that would result in all vectors being embedded with a 0 value for the higher dimension, meaning no distance change.

It is well-known that it is ideal for the dimension to converge to zero. In our case, only two delays converge to 0, possibly due to randomness and noise of the data used. For our purposes, we decided to use $m = 7$ for the dimensions, with a delay of $\tau = 3$. However, since there is no specific method for determining these values, other parameter combinations are also viable. Note also that empirical tests showed that small changes to these parameters do not meaningfully impact the result of the dendrogram illustrated in Figure 9. This dendrogram splits the stocks into three broad clusters, which lie relatively far away from each other, and 10 outliers. The first cluster is colored yellow, the second and largest cluster is colored green, and the third is colored red. Finally, we color blue a stock not being classified to any cluster. Moreover, even though the green cluster makes up the majority of the observations, it is straightforward that the dendrogram is made up of at least three more determinable clusters. Considering these clusters in our analysis, we have a total of five clusters with a similar number of stocks in each.

When looking at the individual stocks within a cluster, we observe that the movements of the stocks are similar (for an illustration see Figure 10). These stocks were selected from the right side of the green cluster. The price of each stock initially decreases until July 2023, after which it remains relatively stable until March 2024, when it increases again, forming a rough U-shaped pattern. Although the overall movement of the stocks is reasonably comparable, the smaller movements of the stocks are

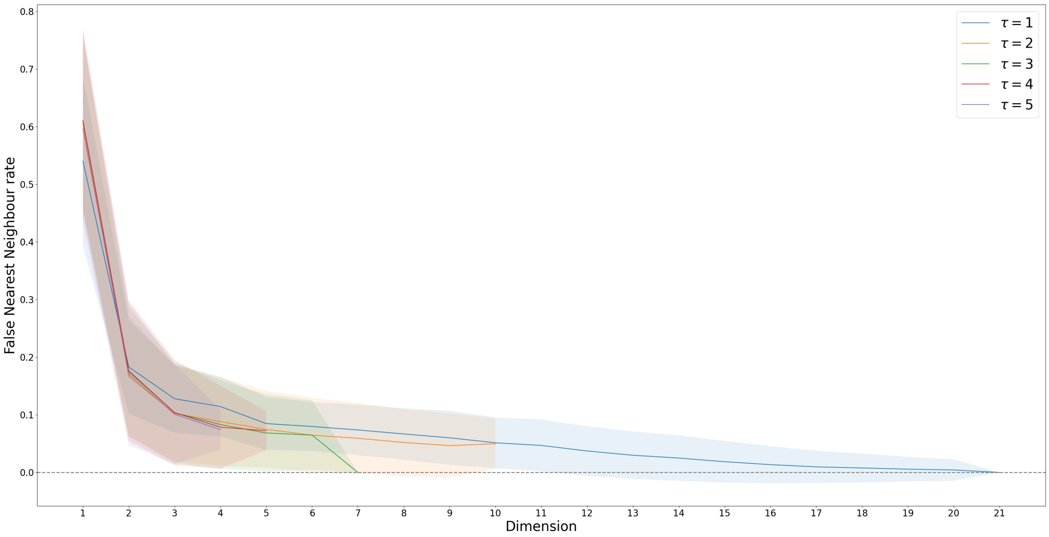

Figure 8. The average FNN rate per embedding dimension.

very similar, signaling *a focus on detail in TDA*. It is important to note the fact that these stocks appear to be closer compared to the one presented in the SAX example.

5. Method comparison

In this section, we present a comparative analysis of the SAX and TDA methods applied to stock prices. Both methods offer distinct approaches to clustering and analyzing time series data, each with its own strengths and weaknesses. By examining the dendrograms generated by each method, we aim to highlight the key differences in clustering efficacy and the handling of outliers. Additionally, we delve into the intra-cluster characteristics to further understand how each method segments and interprets stock behavior within identified clusters. This analysis provides valuable insights into the suitability of SAX and TDA for various financial data applications.

1. **Differences in clustering:** The SAX dendrogram indicates that the distances between individual stocks are “minimal”, making it challenging to identify clear clusters, as shown in Figure 4. Conversely, the TDA method produces clusters that are more distinctly separated with greater distances between them, leading to improved clustering outcomes as illustrated in Figure 9. Additionally, TDA handles outliers more effectively, isolating them from the main clusters.
2. **Intra-cluster Differences:** Beyond the overall clustering, there are notable differences within the clusters themselves that the dendrograms do not fully capture. Specifically, how clusters differ and how stocks behave within each cluster is not immediately apparent. To explore this, the mean of the normalized stock prices was calculated for each cluster derived from both methods. Normalized values were used to facilitate easier comparison and to standardize all values around zero. Additionally, the market mean (the mean of all stocks) was subtracted from the cluster means to isolate the effects of individual clusters. These results are presented in Figure 11. It

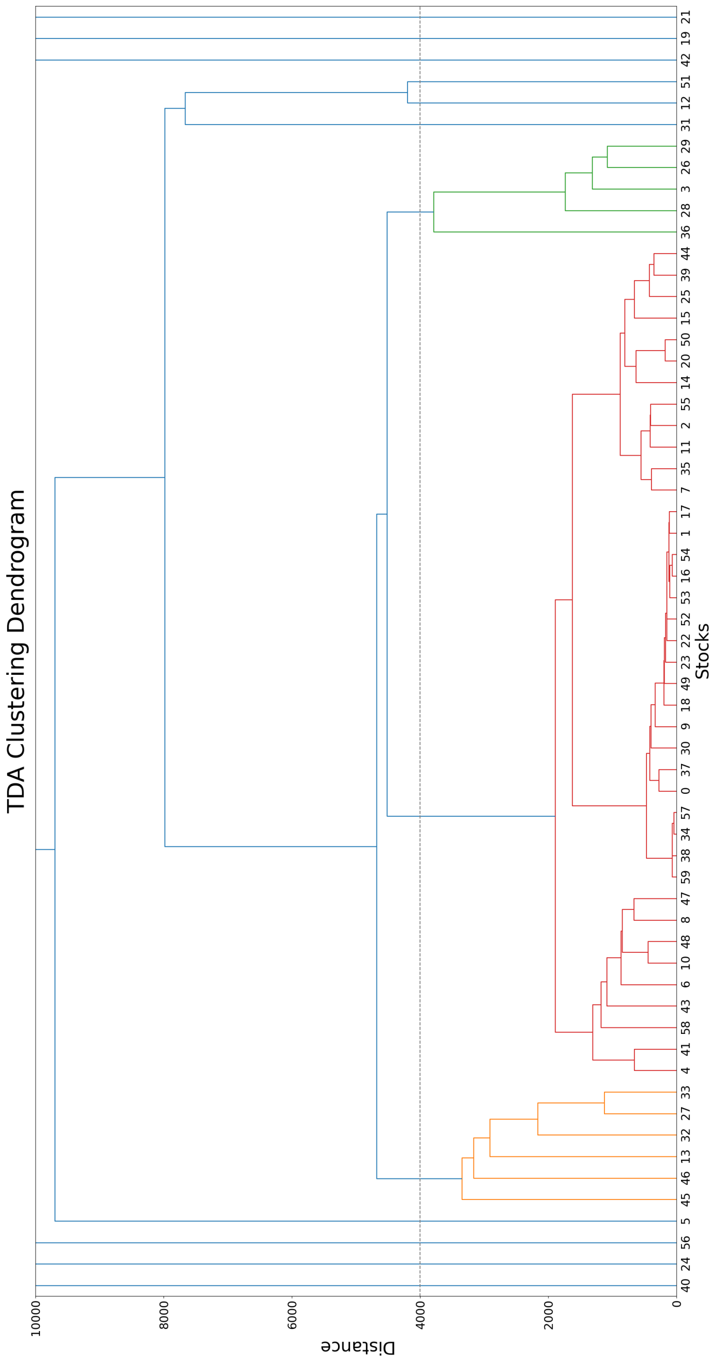

Figure 9. A dendrogram showing the result of the hierarchical clustering based on TDA.

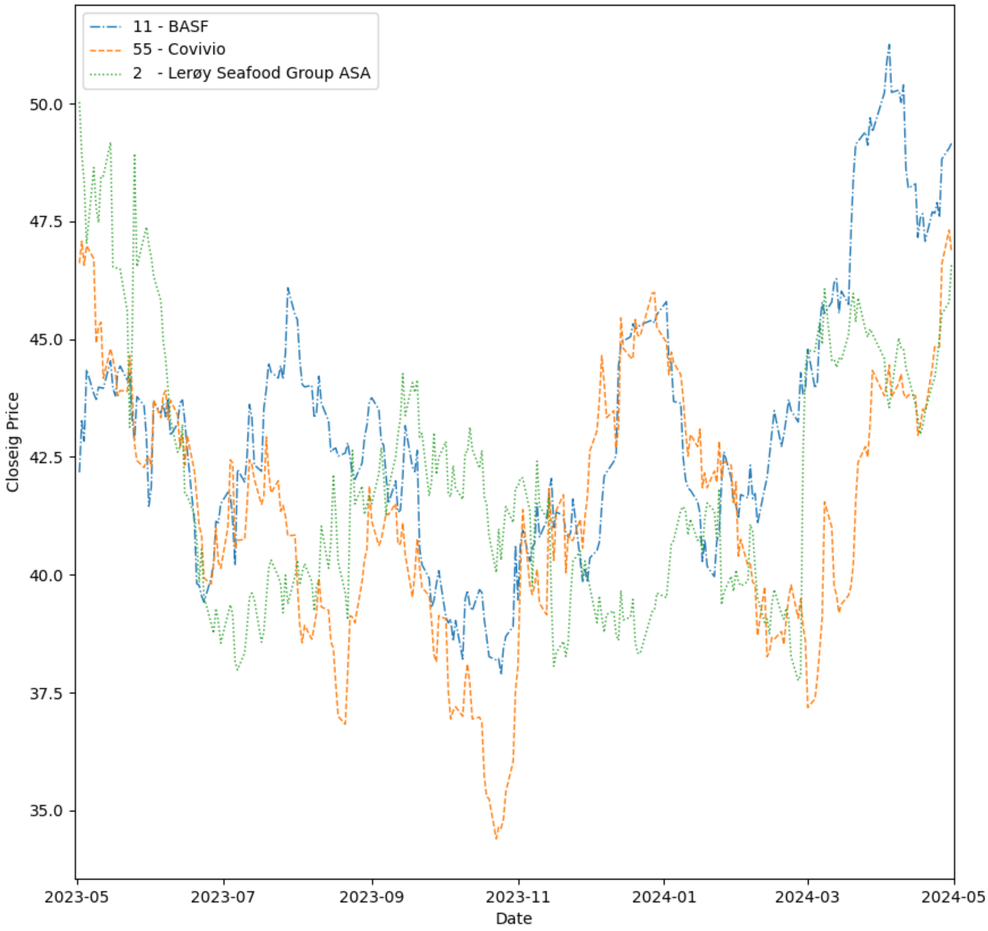

Figure 10. An example of three stocks from the green cluster.

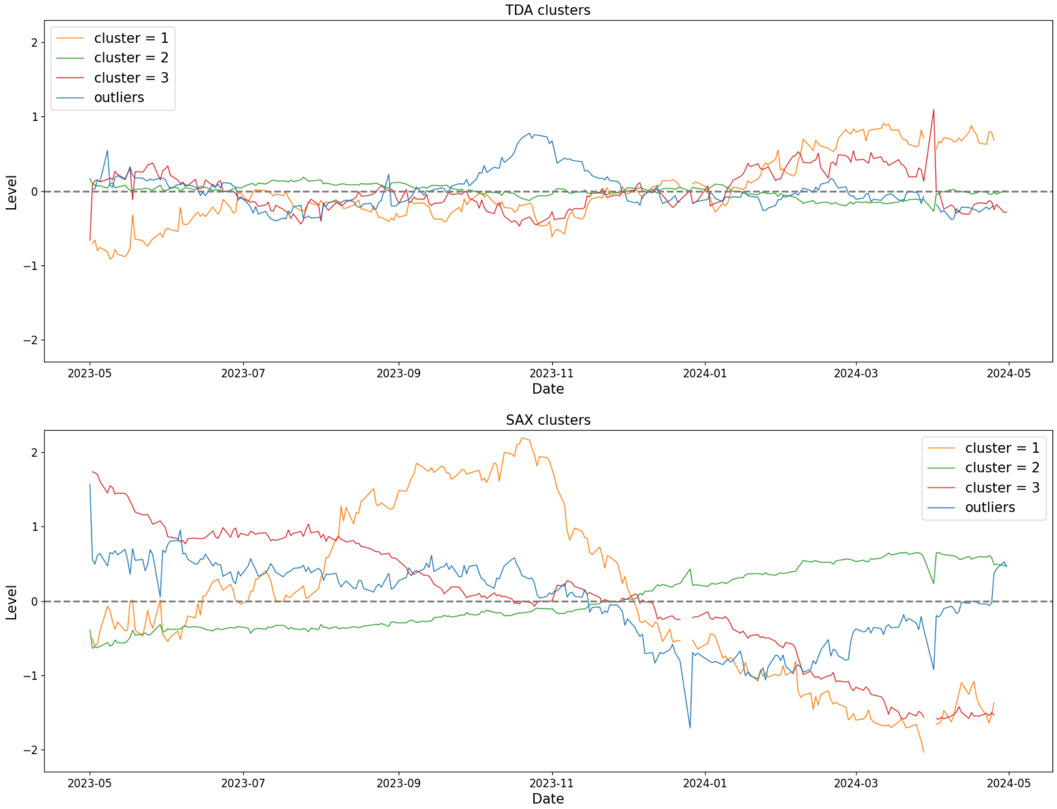

Figure 11. The average normalized stock price of each cluster for TDA (upper) and SAX (lower).

is important to note that the sharp spikes and drops on specific days are due to differences in trading days between stock markets, and these anomalies do not represent the overall clusters accurately. As can be observed, SAX differentiates stocks more distinctly than TDA, suggesting it may be more effective for broader analyses. Notably, in Figure 11, the line for cluster 1 is not fully representative as it comprises only two stocks.

3. **Comparison of SAX and TDA Clustering:** SAX appears to differentiate stocks more effectively than TDA. TDA clusters show values mostly centered around zero with relatively smaller differences between clusters. In contrast, SAX clusters exhibit more distinguishable time series patterns. This is particularly evident when comparing clusters 1 and 3 from the TDA analysis. These clusters follow similar paths, with differentiation primarily at the beginning and end. Cluster 2 from TDA, although different, appears to track the market mean closely, indicating stocks that move in line with overall market trends. SAX, however, identifies more distinct patterns: cluster 2 starts below the mean but rises steadily above it, while cluster 3 aligns with outlier stocks on a downward trajectory.

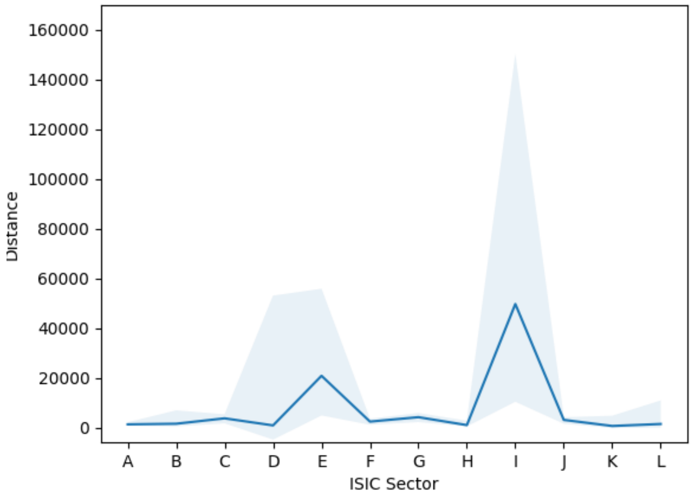

Figure 12. The average distance between stocks of the same economic sector based on TDA.

5.1. *Overview of differences between industry sectors*

Understanding how different industry sectors are represented in SAX and TDA analyses provides insights into the clustering performance of each method. To measure this, we calculated the median distance within each economic activity sector along with its standard deviation. Smaller distances indicate that stock prices within an industry move similarly, suggesting higher similarity within the sector.

Figure 12 illustrates the within-industry distances for TDA, while Figure 13 illustrates the within-industry distances for SAX. Note that the line segments represent the mean value and the shaded areas represent the standard deviation. It is evident that three industries exhibit particularly high distances according to TDA: electricity, gas, steam, and air conditioning supply (D); water supply, sewerage, waste management, and remediation activities (E); and accommodation and food service activities (I) United Nations (2008). Both analyzes agree that the construction and financial sectors are closely connected. However, the difference in the accommodation sector is particularly interesting and warrants further investigation.

While both analyzes suggest that the construction and financial sectors are closely linked, the disparity observed in the accommodation sector is particularly noteworthy.

5.2. *Proposed Objective Evaluation Criteria*

In addition to visual comparisons of the clusters, future studies could introduce objective metrics to quantitatively evaluate the performance of SAX and TDA. For instance, the *silhouette score* measures the cohesion and separation of clusters by comparing how close data points are within a cluster versus other clusters. A higher silhouette score suggests well-defined clusters, making it an ideal metric for assessing the quality of both SAX and TDA clustering results.

Another metric that could be utilized is the *Davies-Bouldin index*, which evaluates the ratio of intracluster distances to intercluster distances. A lower Davies-Bouldin index indicates better clustering

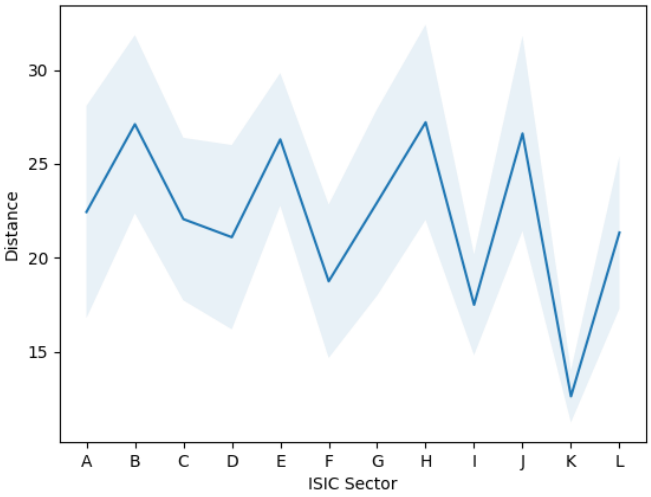

Figure 13. The average distance between stocks of the same economic sector based on SAX.

performance. Given the differences in how SAX aggregates data and TDA preserves finer details, these metrics could highlight which method is more suited for various types of financial time series analysis.

Furthermore, *runtime complexity* is an important factor when considering practical applications of these methods. SAX's dimensionality reduction leads to faster processing times, making it a more scalable option for large datasets, while TDA requires more computational resources due to its high-dimensional nature.

In future research, the inclusion of such quantitative metrics would provide a more comprehensive evaluation of the methods beyond visual comparison.

5.3. *Practical Applications of SAX and TDA*

Beyond just measuring stock similarity, SAX and TDA each have distinct strengths that lend themselves to different applications in the financial industry:

SAX method is well-suited for applications that require simplified representations of large datasets, such as portfolio management and trend-following strategies. Because SAX reduces time series data into symbolic representations, it allows for the detection of broad, long-term patterns across many stocks. For example, portfolio managers could use SAX to identify long-term trends and create portfolios that minimize risk by selecting stocks that follow favorable aggregated patterns. Additionally, SAX could help detect market regime shifts, where a portfolio might be rebalanced when a large change in the symbolic representation indicates a significant market event.

On the other hand, TDA is more sensitive to small fluctuations and can identify subtle topological features like loops or voids in time series data. This makes TDA particularly useful for high-frequency trading and anomaly detection, where the ability to detect short-term, localized trends is critical. For instance, TDA could uncover small, recurring patterns in high-frequency stock price movements, which could be leveraged in high-frequency trading algorithms to exploit small arbitrage opportunities. Additionally, TDA's ability to capture complex topological structures in data allows for the identification of market anomalies, such as black swan events or persistent cycles that are not easily detected by other methods.

In future work, we plan to further explore these specific applications by applying SAX and TDA in different market environments, demonstrating how each method's strengths can be leveraged based on the specific financial problem at hand.

6. Discussion & Conclusions

The main aim of this report was to investigate the similarity in the movement of stocks and identify the strengths and shortcomings of different analytical methods. To provide a comprehensive market view, five stocks were selected for each economic activity sector in Europe United Nations (2008). TDA methods and a variant of SAX were applied to these stocks, resulting in hierarchical clustering based on similarity.

This study finds relevant differences in the distances between stocks using the two methods. Specifically, TDA prioritizes smaller day-to-day changes, while SAX's aggregation provides a broader overview of total stock movements. These results are counterintuitive, as TDA uses all available information rather than aggregated data like SAX. However, the aggregation in SAX may lose irrelevant information, helping differentiate small stock price changes from larger movements.

TDA, on the other hand, seems to over-represent comparatively minor day-to-day price differences. In practice, this means that while stock movements appear similar graphically, TDA loses track of larger movements by focusing too much on daily changes. This is especially apparent when comparing the average movements of clusters identified by both methods. SAX identifies clusters that, on average, move drastically differently from one another, while TDA clusters exhibit relatively similar movements (for an illustration see Figure 11). Because of its focus on small changes, TDA is better suited for environments where small data perturbations are crucial, such as high-frequency trading Truong (2017). In such scenarios, minor stock price changes are vital for profit-making. Conversely, SAX is more useful for broader, more holistic economic analyses, where larger trends are more important, and the aggregation captures the most significant variance for each stock. Thus, SAX is more appropriate for optimizing portfolios or conducting industry-wide analyses that require an overview rather than exact details.

When examining the findings, it is interesting to note where the two methods agree and where they diverge. For instance, highly separated industries like ”electricity and gas” and ”water and waste” indicate a strong connection to other sectors rather than within themselves. The separation in the electricity and gas industry may stem from its significant influence on other industries like manufacturing, which heavily relies on energy prices Hankinson and Rhys (1983). The water and waste sector, however, may exhibit such behavior due to local economic shifts rather than Europe-wide trends. Although both analyses agree on the close connection between the construction and financial sectors, the accommodation sector's differences are noteworthy. This discrepancy can be explained by the overall trend of the stocks being similar, while small daily perturbations do not resemble each other. Figure 14 illustrates this point, showing (a) the entire movement of accommodation stocks and (b) a zoomed-in version. The zoomed-in panel shows no clear relationship, while the full view reveals a clear connection, validating previous observations. Indeed, the larger movements of the sector are reasonably uniform, however when zoomed in further, the day-to-day changes are very different from one another.

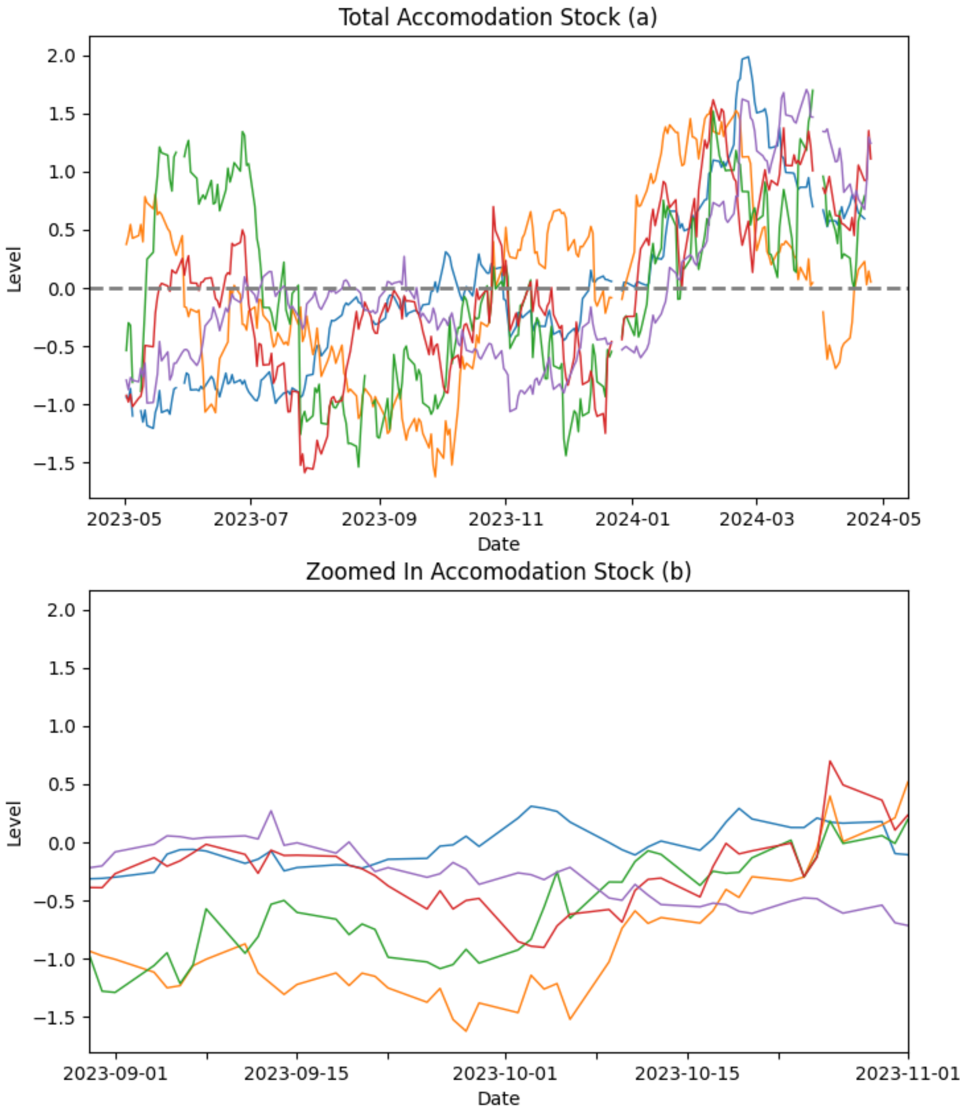

Figure 14. Movement of stocks in the Accommodation and Food Service Activities sector (I), with (a) showing the overall trend and (b) a zoomed-in view for detailed stock movements.

As the stocks in this sector move relatively similarly, SAX identifies it as having a small average distance. However, TDA estimates a much larger difference, which is accurate when considering small stock movements in this sector. Thus, SAX is better suited for tasks requiring similarity in stock movements. Despite using less data, SAX's aggregation provides a clearer overview of overall stock movements, whereas TDA emphasizes smaller daily differences, missing the similarity in overall movement.

We conclude the following:

- SAX effectively identified broader stock movements, whereas TDA failed to differentiate clusters meaningfully from the mean.
- Despite SAX losing information through aggregation, it provided a clearer overview for wider scope analyses. TDA's focus on small differences makes it unsuitable for this type of analysis.
- The interconnectivity of different economic sectors differed between the analysis, with SAX finding more sectors connected to others and TDA finding most sectors more interconnected.
- The practical application of SAX versus TDA depends on the specific needs of the analysis. SAX is more suited for scenarios where computational simplicity and broad pattern recognition are important, such as portfolio management and trend-following strategies in long-term investment. On the other hand, TDA excels in detecting local anomalies and high-frequency movements, making it ideal for high-frequency trading and market microstructure analysis. In practice, the choice between these methods should be guided by the data's frequency and the level of detail required in the analysis.

In conclusion, TDA is a more complex analysis requiring deeper knowledge to apply and interpret. However, its granular nature makes it suitable for high-frequency trading, where sensitive analysis of complex patterns is crucial. Conversely, SAX is better for identifying larger trends, making it more appropriate for broader analyses.

6.1. Recommendations for Enterprises and Regulatory Bodies

Based on our findings, enterprises, especially those involved in high-frequency trading, could benefit from incorporating TDA into their data analysis frameworks to detect localized trends and anomalies that may otherwise go unnoticed. For portfolio managers and long-term investors, SAX offers a simplified method for recognizing broader market trends across sectors. Regulatory bodies might consider encouraging the use of advanced analytical techniques like TDA in monitoring market anomalies that could signal manipulative trading practices or systemic risks.

6.2. Applications in Financial Analysis

Our findings suggest that SAX's ability to aggregate large-scale trends makes it ideal for tasks such as long-term investment strategies, portfolio optimization, and sectoral analysis. TDA, with its focus on local structures and topological features, is better suited for high-frequency trading and the detection of market anomalies. Future studies could benefit from exploring these applications in more depth, with a focus on the specific financial use cases for each method.

6.3. Limitations & future research

The current analysis is limited in scope and computational power. TDA, being computationally intensive, is challenging to apply to very large datasets, which could provide more insight into TDA's classification capabilities. More data might reveal broader applications and hidden strengths of TDA. Moreover, this analysis focused only on the European stock market to provide a clearer overview. Including stock data from different markets with varying exchange rates and dynamics could add interesting and relevant insights for future studies. Additionally, both SAX and TDA can handle multidimensional data. Providing additional information like mean market value, industry value, currency exchange, or debt to the time series data could yield interesting results.

In a future study we will consider merging SAX and TDA by applying TDA techniques to data aggregated using PAA. This approach could balance TDA's granularity with SAX's broad overview, resulting in a more efficient and comprehensive analysis. Additionally, since PAA reduces dimensionality, the resulting analysis should be more computationally efficient than the current TDA analysis.

Moreover, while this study provides insights into the differences between SAX and TDA through visual cluster analysis, future research could benefit from the use of quantitative clustering metrics such as the silhouette score and Davies-Bouldin index. These metrics would allow for a more objective comparison by evaluating cluster cohesion and separation in a formalized manner. Additionally, runtime complexity should be considered, especially for large-scale applications in financial time series, where SAX's efficiency might be favored over TDA's granularity. By introducing these metrics in future studies, we may better understand the trade-offs between these methods and select the appropriate tool based on the size, complexity, and goals of the analysis. This would provide a more robust evaluation and help guide the choice of method in various financial contexts.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Author contributions

The authors contributed equally to this work, all authors have read and agreed to the published version of the manuscript.

Acknowledgments

This work was supported by Maastricht University.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

Bendich P, Edelsbrunner H, Kerber M (2010) Computing robustness and persistence for images. *IEEE Transactions on Visualization and Computer Graphics* 16: 1251–1260. https://doi.org/10.1109/TVCG.2010.139

Canelas A, Neves R, Horta N (2013) A SAX-GA approach to evolve investment strategies on financial markets based on pattern discovery techniques. *Expert Syst Appl* 40: 1579–1590. https://doi.org/10.1016/j.eswa.2012.09.002

Gao T, Li X, Chai Y, et al. (2016) Deep learning with stock indicators and two-dimensional principal component analysis for closing price prediction system. In: *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, IEEE, 166–169. https://doi.org/10.1109/ICSESS.2016.7883040

Gidea M, Katz Y (2018) Topological data analysis of financial time series: Landscapes of crashes. *Physica A* 491: 820–834. https://doi.org/10.1016/j.physa.2017.09.028

Hankinson GA, Rhys JMW (1983) Electricity consumption, electricity intensity and industrial structure. *Energy Econ* 5: 146–152. https://doi.org/10.1016/0140-9883(83)90054-3

Karan A, Kaygun A (2021) Time series classification via topological data analysis. *Expert Syst Appl* 183: 115326. https://doi.org/10.1016/j.eswa.2021.115326

Kennel MB, Brown R, Abarbanel HDI (1992) Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys Rev A* 45: 3403. https://doi.org/10.1103/PhysRevA.45.3403

Keogh E, Chakrabarti K, Pazzani M, et al. (2001) Dimensionality reduction for fast similarity search in large time series databases. *Knowl Inf Syst* 3: 263–286. https://doi-org.mu.idm.oclc.org/10.1007/PL00011669

Khasawneh FA, Munch E (2017) Utilizing topological data analysis for studying signals of time-delay systems, In: Insperger, T., Ersal, T., Orosz, G. (eds), *Time Delay Systems: Theory, Numerics, Applications, and Experiments*, Springer, 93–106. https://doi-org.mu.idm.oclc.org/10.1007/978-3-319-53426-8$_7$

Leitão J, Neves RF, Horta N (2016) Combining rules between PIPs and SAX to identify patterns in financial markets. *Expert Syst Appl* 65: 242–254. https://doi.org/10.1016/j.eswa.2016.08.032

Liebert W, Pawelzik K, Schuster HG (1991) Optimal embeddings of chaotic attractors from topological considerations. *Europhys Lett* 14: 521. https://doi.org/10.1209/0295-5075/14/6/004

Liu H, Yang Y (2020) Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika* 107: 935–948. https://doi-org.mu.idm.oclc.org/10.1093/biomet/asaa038

Li X, Ding P (2020) Rerandomization and regression adjustment. *J R Stat Soc Ser B-Stat Methodol* 82: 241–268. https://doi-org.mu.idm.oclc.org/10.1111/rssb.12353

Lin J, Keogh E, Wei L, et al. (2007) Experiencing SAX: a novel symbolic representation of time series. *Data Min Knowl Discov* 15: 107–144. https://doi-org.mu.idm.oclc.org/10.1007/s10618-007-0064-z

Liu W, Shao L (2009) Research of SAX in distance measuring for financial time series data. In: *2009 First International Conference on Information Science and Engineering*, IEEE, 935–937. https://doi.org/10.1109/ICISE.2009.924

Lkhagva B, Suzuki Y, Kawagoe K (2006) Extended SAX: Extension of symbolic aggregate approximation for financial time series data representation.

Majumdar S, Laha AK (2020) Clustering and classification of time series using topological data analysis with applications to finance. *Expert Syst Appl* 162: 113868. https://doi.org/10.1016/j.eswa.2020.113868

United Nations (2008) International standard industrial classification of all economic activities. Rev. 4. *United Nations*. Available from: https://unstats.un.org/unsd/publication/seriesm/seriesm_4rev4e.pdf.

Perea JA, Harer J (2015) Sliding windows and persistence: An application of topological methods to signal analysis. *Found Comput Math* 15: 799–838. https://doi-org.mu.idm.oclc.org/10.1007/s10208-014-9206-z

Skaf Y, Laubenbacher R (2022) Topological data analysis in biomedicine: A review. *J Biomed Inform* 130: 104082. https://doi.org/10.1016/j.jbi.2022.104082

Tralie C, Saul N, Bar-On R (2018) Ripser.py: A lean persistent homology library for Python. *J Open Source Softw* 3: 925. https://doi.org/10.21105/joss.00925

Truong P (2017) An exploration of topological properties of high-frequency one-dimensional financial time series data using TDA. Preprint.

Van Rossum G, Drake FL (2009) Python 3 Reference Manual. *CreateSpace, Scotts Valley, CA*.

Wasserman L (2018) Topological data analysis. *Annu Rev Stat Application* 5: 501–532. http://dx.doi.org/10.1146/annurev-statistics-031017-100045

Van Rossum G, Drake FL (2009) Python 3 Reference Manual. *CreateSpace*, Scotts Valley, CA. ISBN: 1441412697.