IJoSI

DOI: 10.6977/IJoSI.202112_6(6).0005
N. K. Nguyen, M. Bui/ Int. J. Systematic Innovation, 6(6), 37-50 (2021)

Detecting Anomalies in the Dynamics of a Market Index with Topological Data Analysis

Ngoc-Kim-Khanh Nguyen[1,2*], and Marc Bui[2]
[1]Faculty of Basic Science, Van Lang University, Ho Chi Minh City, Vietnam
[2] CHArt Laboratory EA 4004, EPHE, PSL Research University, Paris, France
* Corresponding author E-mail: khanh.nnk@vlu.edu.vn
(Received 10 April 2020; final version received 22 February 2021 Accept September 10, 2021)

Abstract

We investigate the collective behavior of a stock market by studying the dynamics of its representative index's return, using the persistence diagram of the index return's time-delay embedding, an approach of the Topological Data Analysis (TDA) in time series analysis. While the time-delay embedding captures the state space of the index return's dynamics, the persistence diagram encodes the space's topological information under different spatial resolu-tions. Therefore, based on the changes in the point distribution of the persistence diagram over time, we propose a framework to detect its extraordinary movements. Our method provides a measure for the stability level of the mar-ket's collective behavior. After applying this method for the daily return of the S&P 500 index from 1970 to 2020, we demonstrate that the measure efficiently tracks the changes in topological information of the index re-turn. Furthermore, we can capture major American recessions when the measure exceeds a threshold. A continuous and rapid increase of the measure approaching the threshold is considered a warning of a crisis. Hence, our method provides a technical indicator for systematic risk management.

Keywords: anomalies detection, market index, persistence diagram, time-delay embedding.

1. Introduction

Dramatical changes of a financial market are often of great concern to many investors, managers, or policymakers because they have to make decisions such as taking profits, cutting losses, or making policies to avoid market crashes. However, as a complex system, the chaotic and collective behaviors of the market are really difficult to predict. Although there are a lot of mathematical tools to study such behaviors, in stock markets, market indexes are often used to gauge the markets' movements. In fact, since a market index is calculated from the prices of all or underlying shares, it is explicit and available to represent the market's state.

Since the market's movement is complicated and has noise, it is not easy to extract the necessary information within the time series data of its indexes. One way to deal with this problem is using TDA, an approach using new topological and geometric tools to infer information about the structure of point clouds in metric spaces (Edelsbrunner and Harer, 2010). This approach is suitable to deal with noises since it helps study the behavior of a system for a wide range of parameters (Carlsson, 2009). Also, the topological features are expected to reflect the qualitative changes in a time series' dynamics. Hence, TDA is recently used in many works to study the behaviors of time series such as detecting the periodicity of biological time series (Perea et al., 2015), understanding the global behavior of biological aggregations (Topaz et al., 2015), detecting early warning signals of imminent market crashes (Gidea and Katz, 2018), analyzing a bridge deterioration through its vibration data (Umeda et al., 2019), studying the classification problem of volatile time series (Umeda, 2017) …

In this study, we use the TDA approach to construct a method that can detect significant changes in a market index and give signals about financial crises. Similar to the work introduced by Gidea and Katz (2018), we investigate the behavior of the daily log price difference of a

market index, called the daily index return. However, for the purpose of getting the state space of the time series' behaviors, we use the time-delay embedding of the financial time series where parameters of time-delay and embedding dimension are chosen from the empirical data. In the context of time series analysis, the time-delay embedding method, which is first introduced by Packard et al. (1980) and Ruelle (1979), is a simple method to convert one-dimensional data into point clouds of another higher-dimensional space, called the state space. It is useful to analyze chaotic time series because, according to the Takens' embedding theorem (Takens, 1981), a chaotic series can be perfectly modeled by a smooth function when it is correctly embedded. Besides, with a suitable time-delay, the consequent series can be an efficient summary of the whole data (Sauer et al., 1991). As a result, an appropriate time-delay embedding of a time series helps reconstruct the original chaotic data such that we are able to capture the data's dynamics in different states. Especially, when combining this method with TDA, meaningful topological features, such as connected components, circles, holes, associated with the reconstructed data can be extracted by TDA's tools. Consequently, the dynamical characteristics of the time series are discovered, for example, its periodicity, its pattern, or the qualitative changes in its states. Some theoretical studies about this method can be listed such as (Fraser and Swinney, 1986), (Sauer et al., 1991), (Kennel et al., 1992), (Abarbanel et al., 1993), as well as practical studies such as (Umeda, 2017), (Brown and Knudson, 2009), (Seversky et al., 2016), (Ma, 2020). Moreover, instead of considering data in only a certain time window by using L^p-norms of the persistence landscape as Gidea and Katz (2018) and Ma (2020), in this work, we compare the topological characteristics of the data with its historical characteristics encoded in the persistence diagram of its time-delay embedding. Hence, our method is expected to find out the time series' anomalies more obviously. For this purpose, we only focus on persistence diagrams and use unsupervised machine learning models such as k-means clustering in comparing them. More details about our method are described in Section 2. In Section 3, we provide our empirical results when applying the method for investigating the dynamics of the daily return of the S&P 500 index in two cases: when the current dynamics of the time series is quite similar to its historical dynamics and when they are much different from each other. Next, in Section 4, we discuss the efficiency of our method in detecting strange fluctuations of the index return series by considering the relation between our calculation with recessions in the United State market. Finally, we give conclusions for our method of using TDA in examining the behavior of a market index and detecting its anomalies in Section 5.

2. Research Method

When we observe a time series data in a period, how can we recognize that its present behaviors are so different from its historical behaviors? The problem can be dissolved by comparing the topological features of the present data with the features of the historical data. The features were used to detect the qualitative changes in many studies such as (Donato, 2016; Umeda, 2019; Ma, 2020). In order to get the features, we firstly embed the observed time series into a higher dimensional space to construct a state space of the data's dynamics. The time-delay embedding method then combines with the persistence diagram, a powerful tool of TDA that helps encode the topological features of the underlying data's behavior. In addition, we use the k-means clustering algorithm to make the topological information's comparison of a given data and its historical data easier. More details are given in the following paragraphs.

2.1 Time-delay embedding

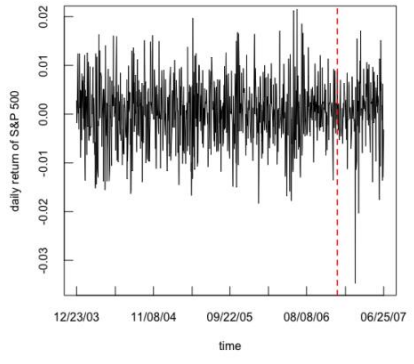
As discussed in (Sauer et al., 1991), an adequate embedding of a time series can define a state space or phase space of the system from which the data is acquired because it can capture the system dynamics in different states, preserve determinism and create a diffeomorphism for the attractors. So, in our work, before taking a topological analysis of our time series data, we embed the data into a suitable state space.

Let remind that a time-delay embedding of a time series (r_t) of length N is a set of vectors $X = (x_1, x_2, \dots, x_{N-(d-1)\tau})$ where each vector is obtained by gathering d adjacent values of the

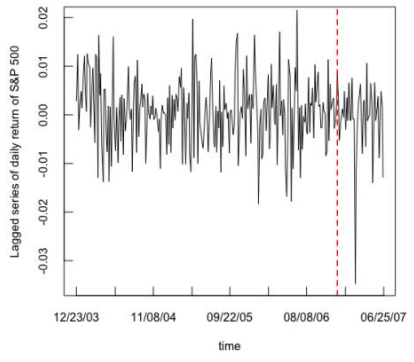
series that is delayed by τ, i.e. $x_t = (r_t, r_{t+\tau}, r_{t+2\tau}, \dots, r_{t+(d-1)\tau})$. The vectors are called reconstructed vectors, τ is called the time-delay and d is called the embedding dimension. Determining values of τ and d such that the corresponding reconstructed space can store the data's dynamical information is an attractive problem. In this work, we choose the lag τ by using the average mutual information (AMI) provided in (Gallager, 1968):

$$\mathrm{AMI}(\tau) = \sum_{t=1}^{t=N-\tau} \hat{p}(r_t, r_{t+\tau}) \log_2 \frac{\hat{p}(r_t, r_{t+\tau})}{\hat{p}(r_t)\hat{p}(r_{t+\tau})} \quad (1)$$

where $\hat{p}(r_t, r_{t+\tau})$ is the estimated joint probability distribution of the bivariate time series $(r_t, r_{t+\tau})$. This measurement tells us how much information about $r_{t+\tau}$ we can receive when r_t is known. As suggested in (Fraser and Swinney, 1986), the time-delay should be chosen where the first minimum of AMI occurs because we should not keep both r_t and $r_{t+\tau}$ when $\mathrm{AMI}(\tau)$ is large. Besides, a large τ makes much data lost nontrivially. Fig. 1 illustrates a sample data and its lagged version where the lag τ as suggested. In the figure, the dashed line is used to divide the training data and the test data. Moreover, to make the reader comfortable when following our method steps by steps, we use the same data in Fig. 1 – 4.



(a)



(b)

Fig. 1. Daily return of the S&P 500 index and its lagged version with $\tau = 3$ given by the first minimum of AMI.

Otherwise, for finding a suitable embedding dimension, one of the popular methods is the false nearest neighbors method proposed in (Kennel et al., 1992). The main idea of this method is that d is chosen as the smallest number such that, for any point, its nearest neighbor in dimension d is still close enough in dimension d + 1. So, the trouble of this method is verifying the threshold for the distance of a pair of points such that the two points are considered to be close to each other. To overcome this trouble, a new approach to the false nearest neighbors method was introduced in (Cao, 1997). The author defines:

$$a(t, d) = \frac{\|x_{t,d+1} - x_{t^*,d+1}\|}{\|x_t - x_{t^*}\|}, \qquad t = \overline{1, N - d\tau} \quad (2)$$

where x_{t^*} is the nearest neighbor of x_t in dimension d; $x_{t,d+1}$ and $x_{t^*,d+1}$ are the reconstructed vectors of x_t and x_{t^*} in dimension d + 1, respectively, i.e., $x_{t,d+1} = (r_t, r_{t+\tau}, \dots, r_{t+d\tau})$ and $x_{t^*,d+1} = (r_{t^*}, r_{t^*+\tau}, \dots, r_{t^*+d\tau})$; $\|.\|$ represents for the distance between the inside points. In this work, we use the Euclidian distance. Let E(d) be the mean value of a(t, d) over time, and E1(d) be the ratio of E(d + 1) to E(d). If E1(d) stops changing when d is greater than a number d0, it means that the time series comes from an attractor and d0 + 1 should be selected as the embedding dimension. In our implementation below with the financial time series, we choose d0 as the point where the ratio of E1(d) to E1(d + 1) is larger than 95% for any d > d0. Fig. 2 illustrates the result of this method when it is applied for the test data used in Fig. 1.

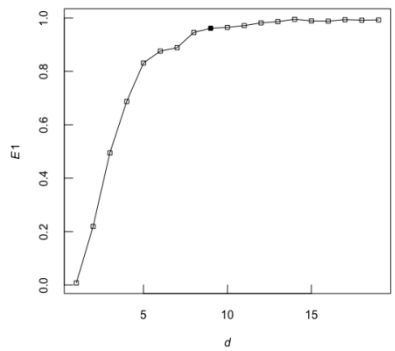


Fig. 2. An example of selecting the embedding dimension (filled point) by the idea of the false nearest neighbors.

We hope that the state space obtained from the time-delay embedding of our time series in previous periods can help recognize strange behaviors of the current data's dynamics. Therefore, we call this space the training space. With the same parameters of time-delay, we also embed the data in the current period into the same dimensional space, which is called the test space. The training space should be large enough and, by partitioning it into consecutive segments of length m, we have a set of sample state spaces having the same size to check whether the test space has different topology. Due to this reason, we suggest choosing m as the size of the test space, i.e., its number of vectors. This enables the observation of periodic property or timing pattern of our data in a period with a certain length.

2.2 Persistence diagram

In order to reveal abnormal behaviors of the observed time series in the present compared with itself in the past, we propose to compare the topological structure of the training space and the test space by using TDA. TDA is an approach that provides topological and geometrical tools to infer information about the structure of a point cloud of a metric space at different spatial resolutions. In particular, for a point cloud whose distribution is unknown, to highlight the point cloud's topology or geometry, TDA's approach is building a "continuous" shape on the points. The shape is often a simplicial complex. Then, the homology groups of the simplicial complex are studied to infer the point cloud's topology. Furthermore, to avoid perturbation or noise in the input data, the point cloud's structure is investigated through a filtration, i.e., a sequence of simplicial complexes ordered by inclusion. The homology groups of each simplicial complex of the filtration represent the point cloud's topology at a certain spatial resolution. So, the persistent homology of the filtration gives meaningful information about the point cloud's topology at different scales. One of TDA's main tools to study the persistent homology is the persistence diagram which encodes the topological information's change of the point cloud's structure through the filtration (Edelsbrunner and Harer, 2010). The diagram is a graph in the plane $\mathbb{R}^2$ such that it includes the diagonal $\{(x, y) \in \mathbb{R}^2 \mid x = y\}$ and points, whose x and y coordinates are the birth and death scales of topological features respectively through a filtration of the space.

In our study, the time-delay embedding of the index return series provides the state space of the data's dynamics. Since the state space is a point cloud of the Euclidean space $\mathbb{R}^d$, where d is the embedding dimension, we can use the TDA's tools to study the persistent homology of the state space, then draw a meaningful conclusion for the index return's movement. For example, the groups of dense 0-dimensional features on the persistence diagram help classify the index return's behaviors, while 1-dimensional features having high persistence values relate to the periodic trend of the system's dynamics. Especially, we expect that by comparing the persistent homology of the index return in a certain period with the one in previous periods, we can recognize strange behaviors in the index return's dynamics.

More specifically, the topological changes over all scales of the reconstructed data in each segment of the training space are tracked by its persistence diagram. In this study, for each segment of the training space, we have a set of m reconstructed vectors, which are embedded points $\{x_{(1)}, x_{(2)}, \ldots, x_{(m)}\}$ of the Euclidean space $\mathbb{R}^d$. We use the Vietoris-Rips complex filtration $(\mathrm{Rips}_\alpha(\mathbb{X}))_{\alpha\in\mathbb{R}}$, where the complex $\mathrm{Rips}_\alpha(\mathbb{X})$ is the set of simplices spanned by $\mathbb{X}$ such that $\|x_{(i)} - x_{(j)}\| \leq \alpha$ for all i, j (Edelsbrunner and Harer, 2010). Since the persistence of a feature is the difference between the scale where the feature appears and disappears, the persistence diagram briefly describes the evolution of the data's structure over scales.

Furthermore, we only pay attention to the 0-dimensional and 1-dimensional topological features. The reason is that 0-dimensional features, which are corresponding to connected components of the point cloud under the filtration $(\mathrm{Rips}_\alpha(\mathbb{X}))_{\alpha\in\mathbb{R}}$, give information about concentration and clustering patterns of the time series' dynamics, while 1-dimensional features, which are corresponding to holes, give information about the dynamics' periodicity. In addition, in persistence diagrams of our financial time series data in many different time windows,

we observe that the features having higher dimensions rarely appear or only appear at large scales with short persistence such that they can be considered noises. For example, in Fig. 3, we can observe the numbers and positions of the features whose dimensions are from 0 to 3 in persistence diagrams of state spaces constructed from the training data and test data used in Fig. 1. In Fig. 3, for the features whose dimensions are higher than 1, their number is too small in any persistence diagram. Also, the points corresponding to the features are very closed to the diagonal, so the features' death scales are approximately their birth scales. This means that the features' existences are not steady when the spatial resolution changes. As a result, we are only interested in the distribution of the points corresponding to 0-dimensional and 1-dimensional features.

Next, we merge persistence diagrams of all segments of the training space into one diagram, called the total diagram (ex. see Fig. 4b). This diagram helps get a general view of the "shape" of historical data in periods that are close and have the same length with the test data.

2.3 K-means algorithm

Giving the total diagram constructed from the segments of the training space, we would like to use it as a standard pattern to test the anomalies of the persistence diagram of the test space. Although there are some distance measures to compute the similarity between two persistence diagrams such as the bottle-neck distance and the Wasserstein distance that can be seen in (Edelsbrunner and Harer, 2010) for more details, these measures are not suitable for our comparison because they require examining all of the matchings between the two diagrams while the number of points outside the diagonal in the total diagram is extremely larger than the ones of the persistence diagram of the test space. So, we propose to compare the two diagrams by region. At first, we divide the points outside the diagonal of the total diagram into clusters. After that, we partition the plane $\mathbb{R}^2$ into many regions corresponding to the clusters and compute the regions' degree of commonalities, which will be the key to detect the strange topology of the test space.

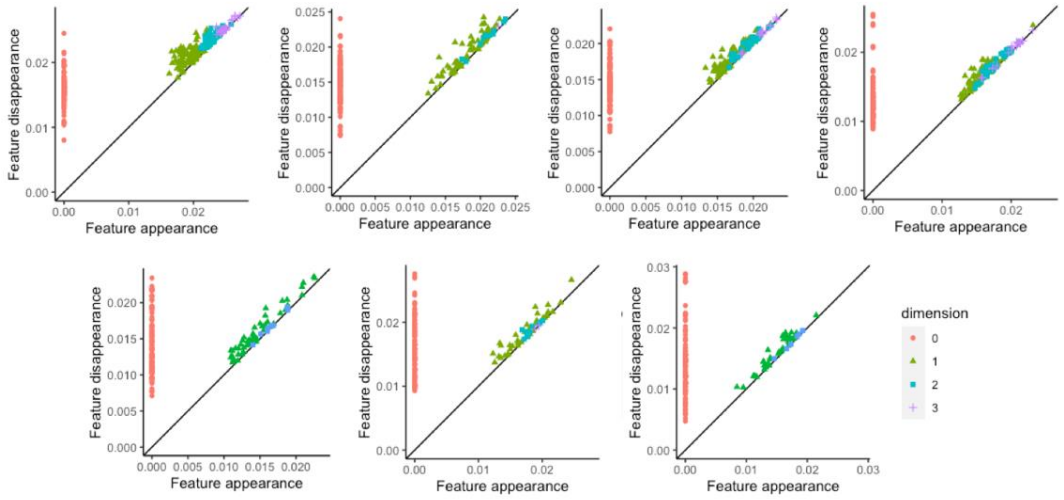


Fig. 3. Persistence diagrams of $(\boldsymbol{Rips}_{\alpha}(\mathbb{X}))_{\alpha\in\mathbb{R}}$ where $\mathbb{X}$ is respectively the state spaces constructed from the S&P 500 index's daily return in some periods; feature appearance and feature disappearance are the values of $\boldsymbol{\alpha}$ that the corresponding feature first appears and disappears in $(\boldsymbol{Rips}_{\alpha}(\mathbb{X}))_{\alpha\in\mathbb{R}}$, respectively.

Here, we use a simple but popular k-means clustering algorithm, proposed in (Hartigan and Wong, 1979), to divide our points in the total diagram into clusters. The idea of this algorithm is that, for a given number k, we classify data points into k clusters so that the total within-cluster variation, which equals the sum of square Euclidean distances between each data point assigned to the same cluster and the cluster center, is minimum. The first centers are chosen randomly from the data and are recalculated when one more point is assigned to a cluster. Since all

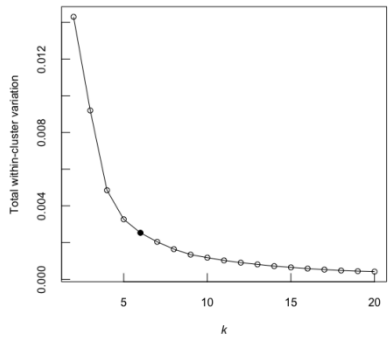
things we have to do is computing the distance between the points and cluster centers, this algorithm is easy to understand and fast. Moreover, it is good enough to classify the points in our financial persistence diagrams because of the points' uncomplicated arrangements. In fact, in the diagrams, the points represented for 0-dimensional features only lie on a line parallel to the vertical axis while the points represented for 1-dimensional features mostly concentrate on a band along the diagonal $\{(x, y) \in \mathbb{R}^2 | x = y\}$ (ex. see Fig. 3). Therefore, when we consider each point as a vector of three dimensions including the birth scale, death scale, and homological dimension of the corresponding feature, the clustering using the k-means algorithm run very fast and all points assigned to the same cluster have the same homological dimension.

However, because the algorithm's result is sensitive to its initial value, we should perform many iterations until the result converges. The main disadvantage of our clustering is that the number k of clusters must be verified before clustering the points. In order to solve this problem, we use the elbow method to find the ideal value for k. For more specific, we first apply k-means clustering with different values of k and draw the total within-cluster variation as a function of k. Next, we normalize the input and output values of this function to get the elbow point as the point with respect to the maximum curvature of the curve. Fig. 4 shows the result of selecting k by the elbow method when it is applied for the total diagram constructed from the training data used in Fig. 1.

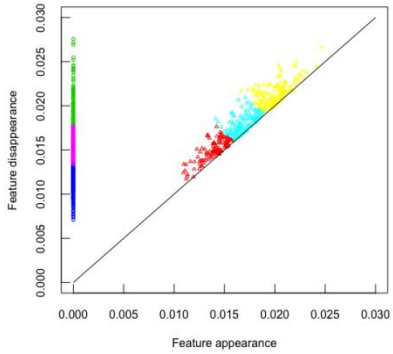
For any cluster i ($i = \overline{1,k}$), we calculate its degree of commonality as the following:

$$P_i = \left\langle \frac{n_{i,j}}{n_j} \right\rangle \quad (3)$$

where $n_{i,j}$ is the number of points assigned to cluster i in persistence diagram j, n_j is the number of points in persistence diagram j and $\langle \cdot \rangle$ is the average over all persistence diagrams of the training space's segments. By Eq. (3), Pi is just an estimator of the probability that a point in the persistence diagram of a state space of our historical data can belong to cluster i.



(a) The number of clusters (filled point) selected by the elbow method.



(b) The total diagram after dividing to clusters.

Fig. 4. An example of applying the k-means algorithm for a total diagram.

2.2 Detecting anomalies of the topological structure of a state space

Let's consider the test data. After constructing its state space with the same time-delay and embedding dimension parameters, we want to know whether there are anomalies in the topological information of its state space when comparing to the topological structure of the training data. For this purpose, the persistence diagram of 0 and 1-dimensional features of the test space are computed to compare with the features encoded in the total diagram.

Next, each point in the persistence diagram of the test space is assigned to cluster i of the total diagram if the cluster is the nearest ones having the same homological dimension with the point. Nevertheless, we need a condition to verify significant changes in this persistence diagram compared with the total diagram. We propose that a point in the persistence diagram of the test space will be assigned to a new cluster, cluster k + 1, if

and only if its persistence is larger than a threshold. Especially, we choose the threshold as the mean of the persistence of points belongs to cluster i of the total diagram added 3 times of its standard deviation. Remind that the total diagram is the coherence of many persistence diagrams constructed from disjointed segments of the training space, so we think that the three-sigma rule is valid enough to define the region of the $\mathbb{R}^2$ plane contained cluster i. Consequently, when the persistence of a point in the persistence diagram of the test space exceeds the threshold, the point shows a dramatical difference of topology from its nearest neighborhood. So, we have an acceptable reliability in verifying topological abnormalities of the test data. Besides, let's remember that the new cluster k + 1 has no points in the total diagram, so its degree of commonality equals zero.

Finally, we simply use the following measure to compute the deviation between the two diagrams through the difference in distributions of their clusters:

$$\delta = \sqrt{\sum_{i=1}^{k+1}(P_i - Q_i)^2} \quad (4)$$

where Qi is the fraction of points assigned to cluster i in the test space's persistence diagram. Clearly, the larger δ is, the more the deviation of the test data's dynamics and the training data's dynamics is. Thus, a large enough value of δ confirms that the index return's dynamics are much strange relative to its previous dynamics.

Some studies also use TDA's tools to investigate stress periods of a stock market, such as the works of Gidea and Katz (2018) and Ma (2020). The authors use the persistence landscape, another tool of TDA which is equivalent to the persistence diagram in encoding the topological information of a point cloud when the spatial resolution changes. However, their studies don't use the time-delay embedding method to convert a given time series to a point cloud because of the concern of the prior lack of an attractor and the intrinsic stochasticity of the time series of index returns. Instead, they use $\tau = 1$ and use more market indexes to create a point cloud reflecting the market's state, where the number of the indexes is considered the points' dimension. In our opinion, because index returns are usually stationary or closer to being stationary (see Fig. 1), there may be some deterministic properties of index returns' dynamics. This is confirmed by the relative stability of the persistence diagrams of 0-dimensional and 1-dimensional features in Fig. 3. Also, in our empirical study presented in the next section, we show that if the index return's behavior is significant strange in a certain period, the deviation δ between the point distribution of the persistence diagram got in the period and the point distributions of persistence diagrams got from previous periods will be considerably large. On the other hand, we think that by using more major indexes of a stock market to be data of different dimensions of a point to apply TDA, coordinates of a point are not independent of each other. Indeed, even if the indexes are composed of different components, their components must be driven by the same market factor. Meanwhile, by using the time-delay embedding method, we convert the 1-dimensional time series of an index return to a point cloud of a higher-dimensional space to capture different states of the index return's dynamics, where one coordinate of a point is nearly impossible to get from the point's other coordinates. This is certainly more meaningful in our detecting problem. Furthermore, since δ can measure the stability level of an index return's dynamics compared to its historical dynamics, we also show a threshold of δ concerning the dynamics' remarkable changes in the next section.

In summary, our anomalies detection method has two main parameters, the size of the test data and the size of the training data. According to the researching purpose, we suggest that the size of the test data must be large enough to capture the market's behaviors in a few months to discover a recession if it exists. In general, a recession is often verified by a negative economic growth for at least two successive quarters. Therefore, in our implementation with the daily return series of the S&P 500 index presented below, we consider the test data including 132 trading days, i.e., about 6 months. On the other hand, although the training data must be large enough as in many machine learning models but, when studying financial time series, the time window of the data should not be too large to avoid outdated information, which can affect the current market's analysis. Hence, in Section 3, we consider the training data including 750 trading days only, i.e., about 3 years. As a

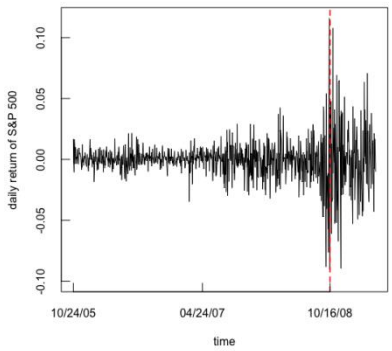
result, we think that our method can help quantify the differences of a market index's behaviors in a certain period from its behaviors in preceding periods through the value of δ. So, a large value of δ is expected to give a signal of serious fluctuations, which might change the market's current level of stability.

3. Empirical Results with the S&P 500 Index

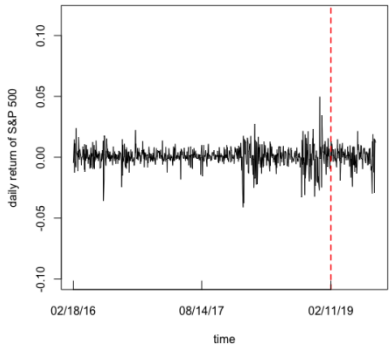
In this section, we study the daily return (rt) of the S&P 500 index where the time series is the log difference of the daily closing value of the index. At first, we study the information got from the topological features of the time series through the persistence diagrams of the corresponding reconstructed vectors in different time windows. Then, we analyze the changes of δ when the test data's dynamics is significantly different from the training data's dynamics and when it doesn't. The two cases are illustrated more clearly through the two example databases described below:

- Database 1: The training data is the return series in trading days from 10/24/2005 to 10/15/2008; the test data is the return series in trading days from 10/16/2008 to 04/27/2009.
- Database 2: The training data is the return series in trading days from 02/18/2016 to 02/10/2019; the test data is the return series in trading days from 02/11/2019 to 08/19/2019.

The values of r_t in the two databases are shown in Fig. 5 where the dashed line divides the training data (on the left) and the test data (on the right). Using the two databases, we compare the behaviors of (r_t) in 132 trading days with its historical behaviors in the 750 closest trading days before. The test data's size approximates the number of trading days in 6 months while the training data's size approximates the number of trading days in 3 years as mentioned at the end of Section 2.4. By tools given in Section 2.1, we found that $\tau = 1$, $d = 9$ and $m = 124$ for the training data in Database 1 while $\tau = 2$, $d = 7$ and $m = 120$ for the training data in Database 2.



(a) Database 1



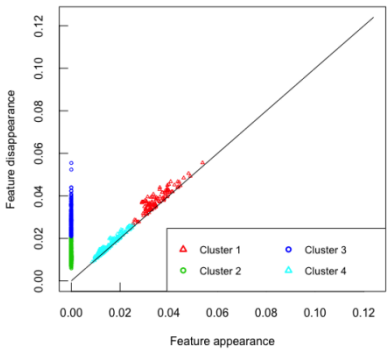
(b) Database 2

Fig. 5. The daily return of the S&P 500 index.

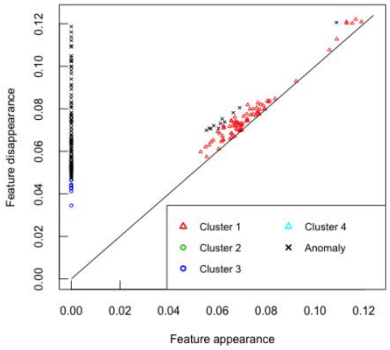
Fig. 6 and Fig. 7 show the total diagram combined from persistence diagrams of all segments of the training space and the persistence diagram of the test space for Database 1 and Database 2, respectively. From the figures as well as statistics in Table 1 and Table 2, some characteristics of our financial data can be found. Firstly, the 1-dimensional features, which are reflected by points plotted as triangles, are always near the diagonal. Hence, the features have low persistence, which implies the instability of periodicity or repetitive patterns in the dynamics of our data. Secondly, most of the 0-dimensional features, which are reflected by points plotted as circles, tend to have higher persistence than 1-dimensional features. The two phenomena are also observed when we study the daily return in other time windows having the same length. Otherwise, for the persistence diagram constructed by the test data in Database 1 (Fig. 6b), the high death scales of its 0-dimensional features imply the existence of some extreme

patterns in the data's dynamics. This means that there are some periods in which r_t's behavior is dramatically different from its normal behavior. Clearly, this result is compatible with the large fluctuation of r_t in test data as observed in Fig. 5a. By contrast, for Database 2, in the persistence diagram of the test space, we don't see any point whose position is dramatically different from points of the total diagram. Consequently, these observations confirm that we can recognize strange behaviors in the dynamics of our financial data through the changes in points' distribution in the persistence diagram.

As mentioned in Section 2.3 and 2.4, the strange behaviors of r_t are detected quantitatively by the measure of dissimilarity between the total diagram found from the training space and the test space's persistence diagram. After applying the k-means algorithm, we get $k = 4$ and $k = 6$ for the total diagram found from Database 1 and Database 2, respectively. The clustering results are illustrated in Fig. 6a and Fig. 7a. Table 1 and Table 2 give some fundamental statistics of the clusters' persistence and their degree of commonality defined by Eq. (3) for Database 1 and Database 2. The results of assigning each point of the persistence diagram of the test space to a cluster given by the total diagram in the two databases are given in Fig. 6b and Fig. 7b.



(a) The total diagram

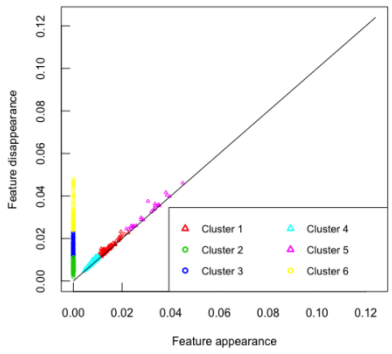


(b) The persistence diagram of the test data's state space

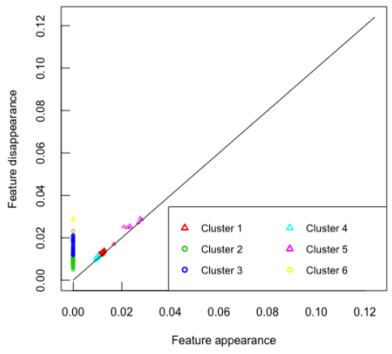
Fig. 6. Comparing the total diagram and the persistence diagram of the test space for Database 1. Circles represent for 0-dimensional features and triangles represent for 1-dimensional features. The black sign × denotes features that cannot be assigned to any clusters of the total diagram.

Table 1. Statistics of features' persistence by clusters in the total diagram for Database 1.

Cluster	Homological dimension	Average persistence (%)	Standard deviation of persistence (%)	Degree of commonality (%)
1	1	0.29	0.22	09.76
2	0	1.25	0.34	43.83
3	0	2.89	0.61	19.62
4	1	0.11	0.09	26.80



(a) The total diagram



(b) The persistence diagram of the test data's state space

Fig. 7. Comparing the total diagram and the persistence diagram of the test space for Database 2. Circles represent for 0-dimensional features and triangles represent for 1-dimensional features.

Table 2. Statistics of features' persistence by clusters in the total diagram for Database 2.

Cluster	Homological dimension	Average persistence (%)	Standard deviation of persistence (%)	Degree of commonality (%)
1	1	0.09	0.08	09.81
2	0	0.74	0.21	47.75
3	0	1.57	0.32	15.84
4	1	0.08	0.06	17.70
5	1	0.15	0.15	02.40
6	0	3.13	0.66	06.50

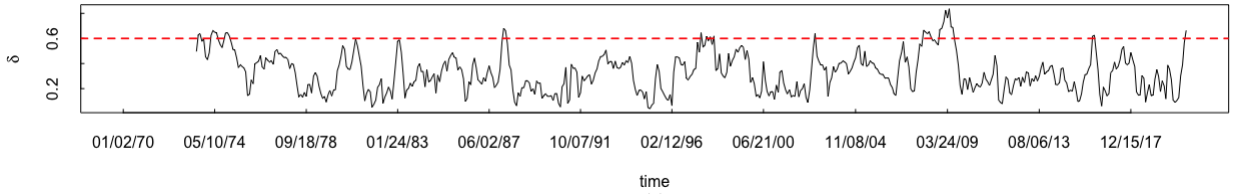
By Eq. (4), we compute that the deviation δ between the point distribution of the total diagram and the point distribution of the test space's persistence diagram is about 83.7% for Database 1 and 11.9% for Database 2. Clearly, the large value of δ for Database 1 is compatible with the extraordinary dynamics of the test data in this database, while the small value of δ for Database 2 is consistent with the indifference between the dynamics of the index return in the test period and the dynamics in the training period.

As a result, the persistence diagram of the time-delay embedding associated with the index return series can reflect the characteristics of the index return's dynamics through its topological information. Therefore, our framework can help detect strange attractors of the time series by recognizing dramatic changes in the point distribution of the diagram.

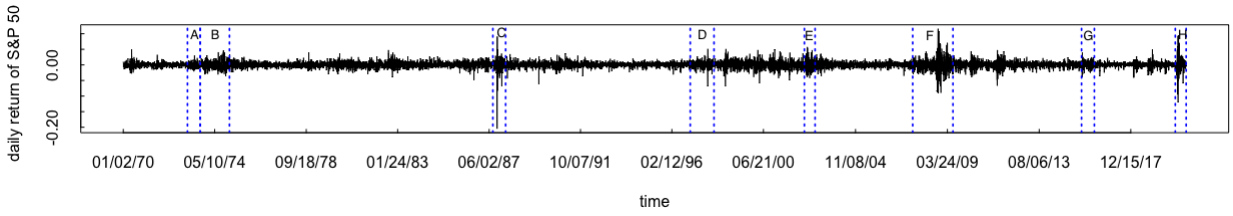
4. Discussions

An important question is how large the value of δ such that it can be considered as a signal of the market phases' switch. For finding the answer, we perform our anomalies detection framework for r_t in each time window, including 132 trading days with 22 rolling trading days, of the long period from 12/18/1972 to 08/04/2020 to get a general view about the value of δ in the U.S. market. Consequently, we get 541 time windows. The dynamics of r_t in each time window is compared to its historical dynamics in 750 preceding trading days. It means that we approximately compare the behaviors of r_t in 6 months to its behaviors in 3 preceding years with 1-month sliding. The value of δ is given in Fig. 8a as a function of time. The maximum of δ approximates 83.7%, which is corresponding to the time window from 10/16/2008 to 04/27/2009, the test period in Database 1. This period has attracted a lot of attention in literature since it is the time when the Great Recession of 2008 happened terribly after the shock of the bankruptcy of Lehman Brothers on 09/15/2008. The average and the standard deviation of δ in our test are about 32.7% and 15.4%, respectively.

Because δ is larger when the point distributions of the test data's persistence diagram and the total diagram of the training data are more different, we consider the values of δ which are on the left tail of its histogram in Fig. 9. Especially, in case that the deviation δ is larger than 60%, we found that the corresponding periods relate to recessions or market crashes. Moreover, this threshold is demonstrated empirically to be large enough to recognize significant changes of the market's state. In fact, there are 32 time windows satisfy this condition of δ in our test. Because the length of each time window is 132 while the length of the sliding period is only 22, some of those 32 time windows intersect with each other. Every pair of time windows which intersect with each other and time windows which lie between them are merged together. Consequently, we get 8 periods, named from A to H (Fig. 8b). The detailed information of these periods is provided in Table 3.



(a) Value of δ plotted at the corresponding test period's last day



(b) Daily return of the S&P 500 index

Fig. 8. Relation of δ and the behavior of the S&P 500 index's daily return.

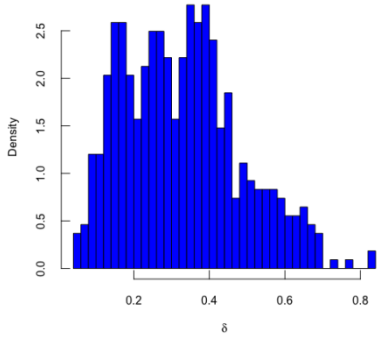


Fig. 9. Histogram of δ from 12/18/1972 to 08/04/2020.

Table 3. Statistics of features' persistence by clusters in the total diagram for Database 2.

Period	Time	Maximum of δ (%)	Average of δ (%)
A	01/22/1973 – 08/30/1973	63.7	63.3
B	10/03/1973 – 01/22/1975	66.3	61.2
C	08/06/1987 – 03/15/1988	68.2	67.1
D	12/27/1996 – 02/13/1998	64.5	58.8
E	06/03/2002 – 12/06/2002	64.0	64.0
F	07/30/2007 – 06/29/2009	83.7	67.0
G	08/12/2015 – 03/22/2016	62.5	62.1
H	01/28/2020 – 08/04/2020	66.4	66.4

We found that the strange dynamics of the S&P 500 index's daily return discovered in periods A and B are corresponding to the 1973 – 1974 stock market crash spreading from January 1973 to December 1974 (Davis, 2003). Besides, periods B also relates to the 1970s stagflation, where the OPEC oil embargo signed on 10/19/1973 is widely blamed for causing the stagflation (Merrill, 2007). Similarly, our framework indicates that there are anomalies in the dynamics of the index in period C because this period relates to the stock market crash of 1987, which was a rapid and severe downturn in the U.S. stock prices that occurred over several days in late October 1987. It is well-known with the name "Black Monday". In addition, this period is a sensitive time with the 1989 savings and loan crisis where more than 1000 of the country's savings and loans had failed. In fact, the crisis is an outcome of uncontrollable bad loans and losses for a long time, especially after the Federal Savings and Loan Insurance Corporation, an institution that administered deposit insurance for savings and loan institutions in the United States, had become insolvent by 1987 (Hanc, 1997; Pyle, 1995). How about period D? Although it doesn't really link to a recession since it just contains to a fast crash in October 1997, which is affected by an economic crisis in Asia started in July. However, the crash is considered as the beginning of the end of the 1990s economic boom in the U.S. (Schwert, 1998) because, after the crash, economic growth became slower in 1997 – 1998. Meanwhile, period E is corresponding to the stock market downturn of 2002, also known as the internet bubble bursting with a dramatic decline in July and September 2002 (Mishkin and White, 2005). In fact, the crash is just the

worst result of the dot-com crash 2000 – 2002. Especially, the longest period, period F, is clearly related to the 2008 financial crisis, the worst crisis in the U.S. from the Great Depression of 1929. The crisis officially lasted from December 2007 to Jun 2009, and the bankruptcy of the investment bank Lehman Brothers in September 2008 is often thought to play a major role in the unfolding of the crisis (Williams, 2010). Period G also relates to a substantial change of the market's collective behavior, a stock market selloff that occured between August 2015 to Jun 2016 (Wikipedia definition of 2015 – 2016 stock market selloff). Finally, the last period is corresponding to the COVID-19 recession, which started in February 2020 (Wikipedia definition of Covid-19 recession).

As a result, our method allows us to track anomalies of a stock market's behavior through observing its representative index's dynamics, especially when the market enters a dramatic downside movement. The method provides a new approach in financial analysis when using the changes of the topological structure of the index return's dynamics to capture the market's stability level. In fact, people frequently conjecture a stock market's stability based on analyzing macro factors that can directly affect all of the market's components and drive their movements in the same direction. These factors can be the political situation, the government's financial policies and procedures, the infrastructure, the import and export values, the monetary... Although macro factors can provide a valuable prediction of the market development, this method requires deep knowledge about the economy and take much time to analyze many statistics of different aspects. Furthermore, because the macro statistics are published periodically, it's not easy for investors to get this data to evaluate the market's current situation if there is suddenly any significant change, such as the occurrence of an epidemic or a disaster. Meanwhile, with our method, we can detect significant changes in the market's collective behaviors at any time because the index's value is updated in real-time when the stock exchange opens. In particular, because the value of δ helps measure the level of the market collective behavior's change, we propose that the value of δ can give a warning of systemic changes when it increases rapidly toward the threshold of 60%. When it exceeds the threshold, this indicates that the market is in, or is about to, a recession or simply has extraordinary movements that are difficult to be predicted by experience and historical data. With this extreme case, the value of δ helps gauge the severity of the recession; for instance, the largest value of δ in our test corresponds to the 2008 financial crisis, the worst crisis in the last 50 years.

5. Conclusions

Basing on the persistence diagram of the time-delay embedding associated with a market index's daily return, we can understand more about characteristics of the market's dynamics such as its concentration and periodicity through topological information encoded in the diagram. Furthermore, we also detect anomalies in the market's collective behavior through substantial changes in the point distribution of the diagram. For example, when applying our method for the daily return of the S&P 500 index in 6 months, we found that the index return lacks stable periodicity and repetitive patterns because of the small persistence of the 1-dimensional features. Besides, the extremely high death scales of the 0-dimensional features in some periods are the result of some abnormal patterns of the data's dynamics. In addition, after considering the change over time of the persistence diagram of 0-dimensional features and 1-dimensional features, we found that the deviation δ between the points' distribution of the persistence diagram computed from the return series in a certain time window and the ones computed from the preceding data can efficiently capture the changes of the return's dynamics. Especially when δ exceeds the threshold of 60%, the market has exceptional behaviors that are difficult to be predicted by experience and historical data. Evenly, it may be falling in a dramatic recession. Therefore,

the value of δ can be used to evaluate the market's stability level, and a continuous and rapid increase of δ approaching the threshold gives a warning of a recession or crisis. In order to apply our framework of detecting abnormalities to other markets, similar researches should be taken to find their own thresholds. As a result, our study not only demonstrates an application of TDA to time series analysis in the financial context but also provides an easy-to-implement method to evaluate the stability of a financial market at any time without studying many economic factors. Thus, parameter δ can be an efficient indicator of systematic risk management, especially for individual investors who are not easy to get a full analysis of the economy's operation.

References

Abarbanel, H.D.I., Brown, R., Sidorowich, J.J., and Tsimring, L.S., 1993. The analysis of observed chaotic data in physical systems, Reviews of Modern Physics, 65(4), 1331-1392.

Brown, K.A., and Knudson, K.P., 2009. Nonlinear statistics of human speech data, International Journal of Bifurcation and Chaos, 19(07), 2307-2319.

Cao, L., 1997. Practical method for determining the minimum embedding dimension of a scalar time series, Physica D: Nonlinear Phenomena, 110, 43-50.

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

Davis, E.P., 2003. Comparing bear markets – 1973 and 2000, National Institute Economic Review, 183(1), 78–89.

Donato, I., Gori, M., Pettini, M., Petri, G., Nigris, S.D., Franzosi, R., and Vaccarino, F., 2016. Persistent homology analysis of phase transitions, Physical Review E, 93, 052138.

Edelsbrunner, H., and Harer, J., 2010. Computational Topology: An Introduction. American Mathematical Society, U.S.A.

Fraser, A.M., and Swinney, H.L., 1986. Independent coordinates for strange attractors from mutual information, Physical Review A, 33, 1134-1140.

Gallager, R.G., 1968. Information Theory and Reliable Communication. Wiley, New York, U.S.A.

Gidea, M., and Katz, Y., 2018. Topological data analysis of financial time series: Landscapes of crashes, Physica A: Statistical Mechanics and its Applications, 491, 820-834.

Hanc, G., 1997. The banking crises of the 1980s and early 1990s: Summary and implications, History of the Eighties: Lessons for the Future, Vol. 1. (pp. 3–86). Federal Deposit Insurance Corporation, Washington DC, U.S.A.

Hartigan, J.A., and Wong, M.A., 1979. Algorithm AS 136: A k-means clustering algorithm, Applied Statistics, 28, 100-108.

Kennel, M.B., Brown, R., and Abarbanel, H.D.I., 1992. Determining embedding dimension for phase-space reconstruction using a geometrical construction, Physical Review A, 45, 3403-3411.

Ma, G., 2020. Using topological data analysis to process time-series data: A persistent homology way, Journal of Physics: Conference Series, 1550, 032082.

Merrill, K.R., 2007. The Oil Crisis of 1973–1974: A Brief History with Documents. Bedford/St. Martin's, U.S.A.

Mishkin, F.S., and White, E.N., 2005. U.S. stock market crashes and their aftermath: Implications for monetary policy, Asset Price Bubbles: The Implications for Monetary, Regulatory and International Policies, 1st Ed, vol. 1. (pp. 53–79). Cambridge MA: MIT Press, U.S.A.

Packard, N.H., Crutchfield, J.P., Farmer, J.D., and Shaw, R.S., 1980. Geometry from a time series, Physical Review Letters, 45(9), 712-716.

Perea, J.A., Deckard, A., Haase, S.B., and Harer, J., 2015. Sw1pers: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data, BMC Bioinformatics, 16(1), 257.

Pyle, D.H., 1995. The U.S. savings and loan crisis, Handbooks in Operations Research and Management Science, vol. 9 (pp. 1105–1125). Elsevier B.V.

Ruelle, D., 1979. Ergodic theory of differentiable dynamical systems, Publications Mathématiques de l'IHÉS, 50, 27-58.

Sauer, T., Yorke, J.A., and Casdagli, M., 1991. Embedology, Journal of Statistical Physics, 65, 579-616.

Seversky, L.M., Davis, S., and Berger, M., 2016. On time-series topological data analysis: New data and opportunities, IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 1014-1022), IEEE Press, Las Vegas, NV, U.S.A.

Schwert, G.W., 1998. Stock market volatility: Ten years after the crash, Brookings-Wharton Papers on Financial Services (pp. 65-99), The Brookings Institution. Washington DC, U.S.A.

Takens, F., 1981. Detecting strange attractors in turbulence, Dynamical Systems and Turbulence, Lecture Notes in Mathematics, vol. 898 (pp. 366-381). Springer, Berlin, Heidelberg.

Topaz, C.M., Ziegelmeier, L., and Halverson, T., 2015. Topological data analysis of biological aggregation models, PloS ONE, 10(5), e0126383.

Umeda, Y., 2017. Time series classification via topological data analysis, Information and Media Technologies, 12, 228–239.

Umeda, Y., Kaneko, J., and Kikuchi, H., 2019. Topological data analysis and its application to time-series data analysis, Fujitsu Scientific and Technical Journal, 55(2), 65–71.

Wikipedia, 2015 – 2016 stock market selloff. Available on line at:

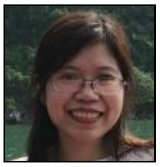
https://wikipedia.org/wiki/2015–2016_stock_market_selloff, accessed on 2021/01/11.

Wikipedia, COVID-19 recession. Available on line at:

https://wikipedia.org/wiki/ COVID-19_recession, accessed on 2021/02/19.

Williams, M.T., 2010. Uncontrolled Risk: Lessons of Lehman Brothers and How Systemic Risk Can Still Bring Down the World Financial System, 1st ed., McGraw-Hill Education, U.S.A.

AUTHOR BIOGRAPHIES



Ngoc-Kim-Khanh Nguyen is a researcher at Van Lang University in Vietnam since 2010. Nguyen received her master degree of Calculus from Can Tho University, Vietnam. She also got a master degree of Quantitative and Computational Finance from John Von Neumann Institute, Vietnam National University, Ho Chi Minh City, Vietnam. Nguyen is currently a doctoral student of Information, Mathematics and Application at École Pratique des Hautes Études, France. Her areas of interests include calculus, data science, machine learning, economics and complex system modelling.

Marc Bui is received his PhD in Computer Science in 1989 from University Paris 11. He is professor of Computer Science at University Paris 8 and Directeur d'études cumulant at EPHE. He was formely at the head of the LaISC/EPHE, where he has managed a research team on Complex Systems Modelling and Computer Science. Prior to this, he was professor at the University of Technology of Compiegne from 1995-1999. From 1991 to 1995, he was associate professor at the University of Paris 10. Marc Bui is now at the AOROC UMR8546 CNRS-ENS-EPHE laboratory and is working on Digital Humanities within the ERC Vietnamica project. His research interests include distributed computing, complex systems modelling and simulation, agent-based simulation, and digital&computational humanities. Specialities: Computer Science, Complex Systems Simulation, Distributed Computing, Information Technology.