Received: 8 July 2019 | Revised: 22 October 2020 | Accepted: 23 October 2020

DOI: 10.1002/eng2.12329

RESEARCH ARTICLE

WILEY

Harnessing the power of intersection for pattern recognition: a novel unsupervised learning method and its application to financial engineering

Michel Ferreira Cardia Haddad[1,2]

[1]University of Cambridge, Cambridge, UK

[2]The Alan Turing Institute, London, UK

Correspondence
Michel F. C. Haddad, Department of Land Economy, University of Cambridge, 19 Silver Street, Cambridge CB3 9EP, UK.
Email: mfch2@cam.ac.uk

Funding information
Cambridge Commonwealth, European and International Trust; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES): BEX 2220/15-6

Abstract

In the present paper a data-driven hard cluster analysis derived from a novel similarity measure is proposed to support financial investors in their portfolio management decision-making process. The main objective of the proposed method is to provide a less arbitrary learning procedure to quantify similarity levels between investment alternatives (pairwise) as well as revealing clustering patterns (whole sample). This is especially useful during periods of high volatility, when investment alternatives tend to become more similar and, therefore, harder to distinguish between themselves. The method dynamics may be readily interpreted through a clear data visualisation. The advantages and caveats of the proposed method is compared to the most popular class of cluster analysis, applied to the well-known Fisher's Iris dataset. Such results show a slightly superior performance of the proposed method but, most importantly, through remarkably different clustering allocation approaches. Moreover, further empirical results applied to daily data reflecting a period of 15 years of time series of economies/stock markets of the Group of Seven (G7) illustrate the potential practical usefulness of the proposed unsupervised learning method, particularly, for portfolio strategy, asset allocation, and investment diversification.

KEYWORDS

cluster analysis, pattern recognition, portfolio management, similarity measure, unsupervised learning

1 | INTRODUCTION

Financial crises impose negative impacts over equity markets and the real economy. Media news regarding the multiplication of financial crisis events are well covered, along with the devastating effects incurred by financial markets worldwide. There is a wide agreement that financial crises entail huge costs to society.[1-4] It is also widely recognised that during

turbulent times (e.g. times of high volatility, crisis periods) investment alternatives tend to become more similar and less distinguishable between themselves.[5,6]

Therefore, investment portfolio diversification tends to be more limited due to an overall higher level of volatility concomitantly with a larger similarity degree in most markets and geographies.[7,8] Cases exemplifying this phenomenon are abundant in the financial crisis literature, such as the Asian crisis of 1997, the dot-com bubble of 2000, the global financial crisis of 2007–2008 (GFC of 2007–08 henceforth), and the European sovereign debt crisis of 2011–2013.[9-11]

Considering the relevance of the problems and challenges aforementioned, the present paper then highlights the benefits of consistently disaggregating financial data into meaningful homogenous groups. Such a data disaggregation process aims to better understand similarity and clustering patterns between competing investment alternatives, particularly during periods of high volatility. Improvements in portfolio diversification are highly desirable due to the fact that unsystematic risk across individual assets, industries, and geographies may be unveiled and then further reduced.[12-15]

The main contribution of the present paper consists of providing an alternative method that is capable of disaggregating data into meaningful groups through a data-driven process. This paper designs, details, and proposes a novel unsupervised learning method named similarity-cluster analysis (SCA henceforth). The main objective of the proposed method in the context of this study is to support financial investors to precisely measure the level of similarity between competing investment alternatives through time. The SCA method is performed through four subsequent steps, in which each investment alternative* is represented as an 2-sphere in the three-dimensional Euclidean space $\mathbb{R}^3$.[16]

In the first step is performed an analysis over the (i) trajectory of each investment alternative through time based on three variables placed at the axes x, y, and z in $\mathbb{R}^3$, and (ii) the calculation of the spherical volume variation, termed† as $V_{i|t}$, in which the correspondent 2-sphere (simply sphere henceforth) dynamic radius‡ is found. Both (i) and (ii) are based on relevant§ variables used as data inputs to compare investment alternatives during a time span. In the second step, the spatial distance between distinct investment alternatives/spheres is graphically depicted and the intersection volume, termed as the common feature $CF_{i,j|t}$, between every pair of spheres, is calculated through analytical geometry.

Subsequently, based on the values found in previous steps, in the third step a measure bounded from zero to one, termed as the similarity factor $SF_{i,j|t}$, quantifies the degree of similarity between every pair of investment alternatives included in the sample. Lastly, in the fourth step, by averaging all similarity factors found in step three, an overall clustering measure, termed as $\mathscr{R}_t$, is then found. The $\mathscr{R}_t$ is an output proxy value designed to meaningfully assess the level of similarity between all investment alternatives in a particular sample, either at any point in time or through a period of time ($t = 1, \ldots, T$).

It is worth mentioning that the motivation to perform the SCA in $\mathbb{R}^3$ in the present paper relies on the fact that, in comparison to the two-dimensional space $\mathbb{R}^2$, one additional input variable is included, enriching the analysis by allowing more information being added as input data. Furthermore, although feasible, in the present study the proposed method is not performed in an n-dimensional Euclidean space $\mathbb{R}^n$ due to the fact that, although allowing the use of n variables instead of only three - which is usually highly desirable, there would not be a consistent data visualisation of the SCA outputs. Therefore, in order to preserve the data visualisation characteristic of the SCA method for instructional purposes, it is then opted to work only in the $\mathbb{R}^3$ in the present paper.

Additional motivations to create a novel unsupervised learning method - which is ultimately a nonhierarchical deterministic hard cluster analysis - among many existing ones, are based on an attempt to tackle some relevant challenges frequently reported in the clustering literature[17,18] and considering limitations of traditional methods, such as partitional clustering (e.g. k-means) and hierarchical clustering (e.g. divisive analysis or DIANA), addressing the following clustering challenges in a data-driven fashion[18,19]: (i) setting the number of clusters in a non-arbitrary manner, (ii) the possibility of a dataset forming zero cluster, and (iii) autonomously allocating outliers to no cluster.

*Throughout this paper, both terms (i.e. investment alternative and sphere) are used interchangeably.
†Regarding the terminology and index notation adopted in this paper, the indexes related to each of the objects in a dataset (such as i and j) and the time unit t are separated by a vertical bar - that is, |. It is worth noting that, in this paper, such a vertical bar does not refer to conditional probability, as commonly adopted in the Bayesian statistics terminology, being used instead as a neater manner to separate objects' identification from unit of time.
‡Dynamic radius refers to the fact the radius of each sphere varies through time.
§Relevant according to the point of view of the investor/researcher/analyst/user.

At some extent, the proposed method aims to tackle these three clustering problems at once through the adoption of the following linkage criterion strategy: imposing an artificial boundary on each object in a dataset by representing them as spheres instead of points in the Euclidean space. Based on such an artificial boundary, it is then calculated the degree of similarity between each pair of objects in the sample through a data-driven process. As the proposed method is based on a data-driven approach, once the input data (e.g. real-world variables) are included in the method's algorithm, then the proposed method may potentially provide interesting outputs and useful insights.

The SCA method linkage criteria of cluster member allocation consist of a nonprobabilistic data-driven process. In such a process, the number of clusters in a sample is learned only subsequently to the input data have being processed by the method's algorithm. This allows the number of clusters being "naturally" revealed, depending solely on the stochastic behavior of the input data, instead of being arbitrarily set by the analyst. Moreover, the SCA method permits the number of clusters in a dataset consisting of zero, in the case in which objects in a sample are highly dissimilar between themselves. Therefore, the SCA method does not artificially impose an arbitrary number of clusters a dataset supposedly should have, as it is the case of many established and popular cluster analyses.

In the present paper, the SCA method practical usefulness is illustrated through two empirical applications performed on distinct datasets. The first empirical analysis is based on the well-known Fisher's Iris dataset, and the second empirical analysis explores a larger dataset of economic/financial time series of the Group of Seven (G7) countries. In the first empirical analysis, all competing clustering methods (i.e. *k*-means, *k*-medians, *k*-medoids, and SCA) yield a very similar performance, although the SCA method provides a slightly larger precision. Nonetheless, much more importantly than such a similar performance it is the fact that the proposed unsupervised learning method reveals the number of clusters (i.e. $k = 2$) through a remarkably different approach, based on a data-driven fashion instead of being arbitrarily imposed by the analyst, as it is the case of the competing cluster analyses.

In the second empirical analysis, the overall cluster measure $\mathscr{R}_t$ is able to properly capture important negative events experienced during the GFC of 2007–08 and the European sovereign debt crisis of 2011–2013. The SCA outputs show that the similarity level between all G7 stock markets experienced a relevant increase from 2003 to the end of 2009, being particularly much steeper from 2006 to 2009, possibly, among other potential factors, as a consequence of major negative events and media news closely related to relevant financial crises, such as the Lehman Brothers bankruptcy in mid-September 2008. Such results suggest that G7 stock markets/economies seem to be more similar between themselves in times of turmoil, providing evidence of a more compact clustering structure of G7 stock markets/economies according to market circumstances. Such empirical results are supported by the related existing literature (e.g. references 14,20,21) and it may have relevant implications for portfolio management and investment diversification.

It is worth mentioning that the SCA method does not refer to a time series clustering method. In fact, the proposed method is performed using static datasets as input data. As the proposed similarity factor $SF_{i,j|t}$ and $\mathscr{R}_t$ measures may be calculated not only considering one point in time but also as a sequence of such points through time (i.e. performing the proposed cluster analysis individually and then repeatedly from 1 to T, in the sequence $t = 1, \ldots, T$), then it is possible, and highly recommended, to build and plot a time series to visualise the temporal clustering pattern progress of the sample. Therefore, it is worth noting that such SCA outputs are distinctively different in comparison with time series clustering methods, which, in general, consider each time series in the sample as analogous to a discrete object and then group such times series according to their similarity levels.[22]

Following this introduction, this paper is divided into four sections. Section 2 presents relevant related studies. Section 3 details the four steps of the SCA method. Section 4 contains empirical results and a discussion based on the outputs. Finally, Section 5 concludes and suggests extensions for further research.

2 | LITERATURE REVIEW

The adoption of clustering methods to address portfolio diversification dates back to Elton and Gruber,[23] in which the authors explore the use of cluster analysis to disaggregate economic data into meaningful groups. The authors discuss the importance of grouping data, develop a metric to potentially measure the degree of similarity between sample objects, and propose alternative procedures related to the group forming process.

The need to disaggregate economic and financial data into meaningful groups, but focused mainly in forecasting purposes as an operational tool to investment decision-making, is subsequently explored in Elton and Gruber.[24] The authors discuss forecasting improvements as a result of the application of clustering procedures on financial data - more specifically, to the problem of forecasting earnings per share. Such forecasted results, performed on grouped data, are then compared to traditional criteria and extrapolation techniques.

In a seminal paper, Panton et al.[25] employ cluster analysis to investigate the structure of comovements in the rate of return between 12 major international equity markets to identify the international equity market structure and structural changes through time. It is unveiled the following data patterns: (i) markets from countries generally described as relatively well developed and open to international capital flows (Belgium, Canada, Germany, the Netherlands, Switzerland, and the United States) have larger degree of similarity compared to other markets; (ii) it is confirmed an expected strong link between equity markets of the United States and Canada; (iii) less strong but still identifiable ties between equity markets of the following pairs of countries: Belgium and France, Germany and the Netherlands, and Australia and the United Kingdom; and that (iv) Austria and Italy tend to experience the lowest degree of similarity in the sample.

In influential study in Mantegna,[26] it is proposed a hierarchical arrangement of equity stocks by investigating the daily time series of the logarithm stock market price, demonstrating a method of selecting a topological space linking stocks with an associated meaningful economic taxonomy. The detected hierarchical structure aims at searching economic factors that affect specific groups of stocks. The result shows that equity stock price time series carry detectable and valuable economic information for investment portfolio strategy and, more specifically, for portfolio diversification purposes.

A high-frequency cross-correlation between pairs of stocks in a sample of 100 stocks traded in the equity market of the United States is analysed in Bonanno et al.[27] A hierarchical organisation of the stocks in the sample is obtained by determining a distance metric between stocks and also analysing properties of the subdominant ultrametric associated with it. In cases in which there is a change in the time horizon used to determine stock returns, it is also observed a clear modification in the hierarchical organisation of the stocks in the sample.

The problem of the statistical uncertainty of the correlation matrix in the optimisation of an investment portfolio is considered in Tola et al.[28] The authors demonstrate that the use of clustering algorithms potentially improve the reliability of portfolio investment in terms of the ratio between predicted and realised risk. It is also shown that several of the results obtained by assuming idealised conditions still hold under more realistic assumptions (i.e. no short selling and mean return as well as volatility forecasting based on historical data).

An evolution strategy-based solution to the investment portfolio optimisation problem - including basic, bounding, cardinality, and class constraints - is discussed and proposed in Pai and Michel.[29] The authors find that investment portfolios clustered by the *k*-means method provide efficient frontiers that are close to the classical efficient frontier for large asset portfolios and, therefore, suggesting that *k*-means clustering may potentially be an interesting alternative for investment diversification purposes of either medium or large portfolios.

In order to categorise stocks according to predetermined investment criteria, a data mining technique to classify equity stocks into clusters is applied in Nanda et al.[30] After the classification being performed, the stocks could then be allocated in order to build a portfolio that aims to minimise risk levels through diversification given a particular level of return. The empirical analysis performed for equity classification purposes shows that the *k*-means method forms the most compact clusters, compared to the self-organising map and fuzzy c-means clustering.

3 | THE SCA METHOD

The proposed method is performed through four subsequent steps, after the selection of the input variables and respective data collection, as summarised in Figure 1. Before the first step, it is performed an exploratory data analysis (EDA) in order to learn and analyse the main characteristics of the selected input variables. In step 1, the spatial trajectory is individually (i.e. each object separately) analysed and calculated. Subsequently, in step 2, the spatial interaction between each pair of objects is analysed, being calculated the volumes of each nontrivial intersection between every possible pair of spheres in $\mathbb{R}^3$.

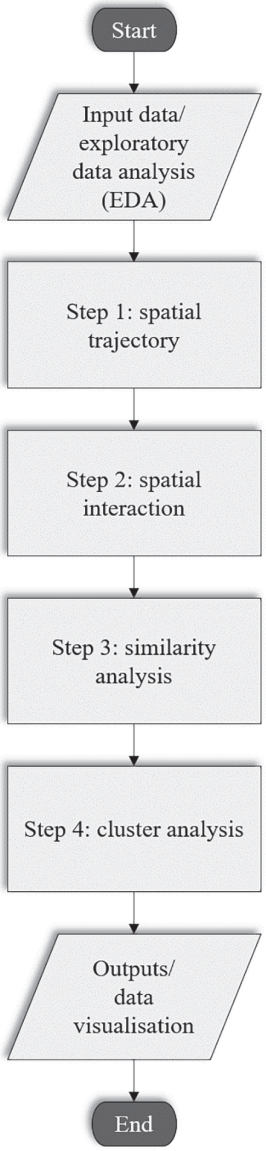

FIGURE 1 Flowchart of the proposed similarity-cluster analysis (SCA) method

The still noncomparable intersection volume found in step 2 is then transformed into a comparable value (i.e. similarity factor) in step 3. In step 4, an overall clustering measure is found by averaging all similarity factors calculated in the previous step. Moreover, in step 4 is also performed the cluster member allocation process, being revealed the number of clusters in the sample, following the method's linkage criteria. Every step in the SCA method is effectively dependent upon the respective immediate previous step, as detailed in the next subsections.

3.1 | Step 1: Spatial trajectory

After the selection of the input variables and data collection process, the SCA effectively begins by verifying the three-dimensional spatial trajectory of each investment alternative and the temporal progress of their respective spherical volumes. There are four values to be used as data input: the first, second, and third values are placed at the x-axis, y-axis, and z-axis, respectively; and the fourth value is reflected in the volume of each sphere. Each of four real-world input variables refers to the respective rate of variation between time t and the immediate previous time $t-1$, which in the finance literature is commonly called either simple one-period net return or simple return[31], as follows:

$$
\begin{aligned}
x_{q|t} &= \frac{X_{q|t}}{X_{q|t-1}} - 1 = \frac{X_{q|t} - X_{q|t-1}}{X_{q|t}}, \quad \forall x_{q|t}, X_{q|t} \in \mathbf{R} \\
y_{q|t} &= \frac{Y_{q|t}}{Y_{q|t-1}} - 1 = \frac{Y_{q|t} - Y_{q|t-1}}{Y_{q|t}}, \quad \forall y_{q|t}, Y_{q|t} \in \mathbf{R} \\
z_{q|t} &= \frac{Z_{q|t}}{Z_{q|t-1}} - 1 = \frac{Z_{q|t} - Z_{q|t-1}}{Z_{q|t}}, \quad \forall z_{q|t}, Z_{q|t} \in \mathbf{R} \\
V_{q|t} &= \left| \frac{v_{q|t}}{v_{q|t-1}} - 1 \right| = \left| \frac{v_{q|t} - v_{q|t-1}}{v_{q|t}} \right|, \quad \forall V_{q|t} \in \mathbf{R}^{+}, \forall v_{q|t} \in \mathbf{R}
\end{aligned} \tag{1}
$$

where $q = (1, 2, \ldots, i, j, \ldots, Q-1, Q)$ is the number of investment alternatives in the sample, $t = (1, 2, \ldots, T-1, T)$ refers to units of time, $\mathbf{R}$ indicates the real numbers and $\mathbf{R}^{+}$ the real positive numbers. Throughout this paper, the objects i and j are used to illustrate two generic investment alternatives in a given sample. It is worth mentioning that, as the fourth variable refers to the volume of each investment alternative's sphere in $\mathbb{R}^3$, it must then be a positive value (i.e. value in modulus).

Each sphere represents a distinct investment alternative in $\mathbb{R}^3$ and the volume of each sphere is analytically calculated and graphically depicted in the same three-dimensional Euclidean space. The spherical volume reflects a variable that is considered as the most relevant one in the context of a particular similarity and cluster analysis. This variable should be relevant and meaningful according to the point of view of a domain expert or the analyst who is performing the analysis. For example, if the focus of the analysis is on stock market returns, then the volume of the sphere might reflect stock market return in absolute value, and the remaining three variables might be the ones which the domain expert/analyst/investor considers as the most important variables that drive and/or influence stock market returns.

Ubiquitously to any quantitative method, it is of utmost importance that the analyst selects coherently each variable to be used as input data, specially the fourth and most relevant variable reflected in the volume of the respective sphere $V_{i|t}$. The discussion on selecting the most appropriate variables is out of the scope of this paper, which may be based on a number of factors, such as risk tolerance level, market consensus, domain expert opinion, individual preferences, among many others.

In the case that there is more than a single variable as a candidate to be the fourth and most relevant one, a combination of $\binom{n}{k}$ nonrepeated rounds of analysis might then be performed and the respective outputs combined by calculating the means of each input variable observed in each round. For instance, in the $\mathbb{R}^3$ case (i.e. four input variables), a total of four rounds of analysis are initially performed, being each variable placed in all three axes as well as volume, one round at a time. Subsequently, such four resulting pairwise similarities and cluster analyses are then averaged in order to obtain the respective final SCA results. It is worth mentioning that such an averaging procedure would lead to a potentially slightly different clustering formation, in the sense that the objects in the sample would tend to be more similar to each other than otherwise.

3.1.1 | Finding the spherical volume $V_{i|t}$ through the time-varying radius $r_{i|t}$

As shown in Figure 2, each investment alternative is graphically represented by its respective sphere in $\mathbb{R}^3$, and their spherical volume $V_{i|t}$ is based on a real-world variable as shown in Equation (1), being calculated through the following equation[32]:

$$
V_{i|t} = {r_{i|t}}^3 \frac{4}{3}\pi, \quad V_{i|t} \in \mathbf{R}^{+}, i \in \mathbb{N}, t \in \mathbb{Z}^{+}, \tag{2}
$$

where $r_{i|t}$ is the time-varying radius of the sphere, π is the ratio of the sphere's circumference to its diameter, and all variables refer to the investment alternative $S_{i|t}$ at time t.

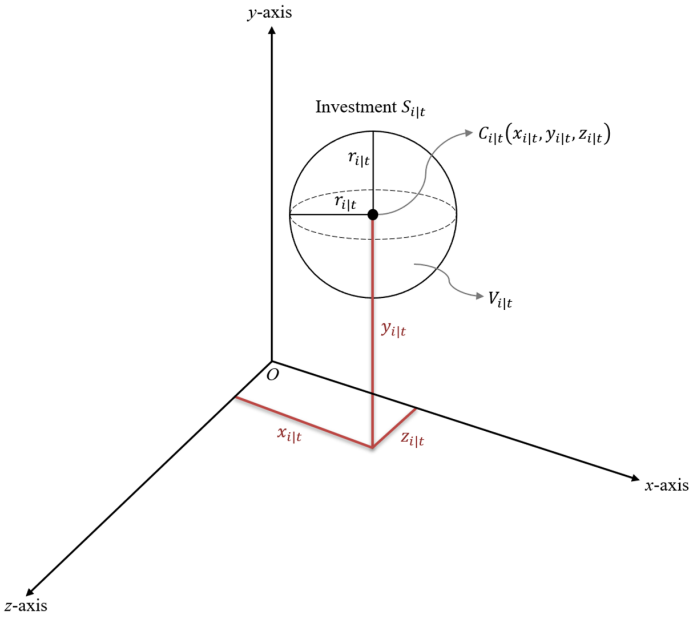

FIGURE 2 Investment alternative $S_{i|t}$ depicted as a sphere in $\mathbb{R}^3$, with volume $V_{i|t}$, center at $C_{i|t}$ $(x_{i|t}, y_{i|t}, z_{i|t})$, and radius $r_{i|t}$

Equation (2) refers to the classic formula to calculate the volume of a sphere in $\mathbb{R}^3$ but it is adapted in the present paper to incorporate the temporal variation reflected in the spherical volume as a consequence of the respective time-varying radius. In fact, as π refers to a constant value and as the spherical volume precisely reflects a real-world data, the $V_{i|t}$ is a given value which, therefore, it is known in advance by the analyst. Thus, the only unknown value in Equation (2) is the time-varying spherical radius $r_{i|t}$, which may be found as follows:

$$\frac{V_{i|t}}{{r_{i|t}}^3} = \frac{4}{3}\pi \tag{3}$$

Solving for $r_{i|t}$ yields:

$$\begin{aligned} \frac{1}{{r_{i|t}}^3} &= \frac{(4/3)\,\pi}{V_{i|t}} \\ \frac{1}{r_{i|t}} &= \sqrt[3]{\frac{(4/3)\,\pi}{V_{i|t}}} \\ r_{i|t} &= \frac{1}{\left[\pi\,(4/3)\,{V_{i|t}}^{-1}\right]^{1/3}} \end{aligned} \tag{4}$$

Therefore, as new real-world events succeed through time - such as official communications made by policymakers, changes in regulation, media news, political issues, among many more - spheres, representing investment alternatives, perform a spatial trajectory and experience spherical volume variation, either positive or negative, through time. Thus, real-world exogenous events, either beneficial or detrimental ones, are captured by the proposed method and then graphically represented as perfectly round geometrical objects (i.e. spheres) interacting in $\mathbb{R}^3$.

3.2 | Step 2: Spatial interaction

After placing the three input variables at the three respective spatial axes and calculating the spherical volumes of each investment alternative in the sample in step 1, then the proposed method proceeds by performing a visual inspection in $\mathbb{R}^3$ and, more importantly, calculating all interactions between each pair of spheres. The respective output may be either a trivial intersection between a pair of investment alternatives (i.e. no volume shared between a particular pair of spheres) or a nontrivial intersection between a pair of investment alternatives (i.e. a volume greater than zero shared between a particular pair of spheres), as illustrated in Figure 3.

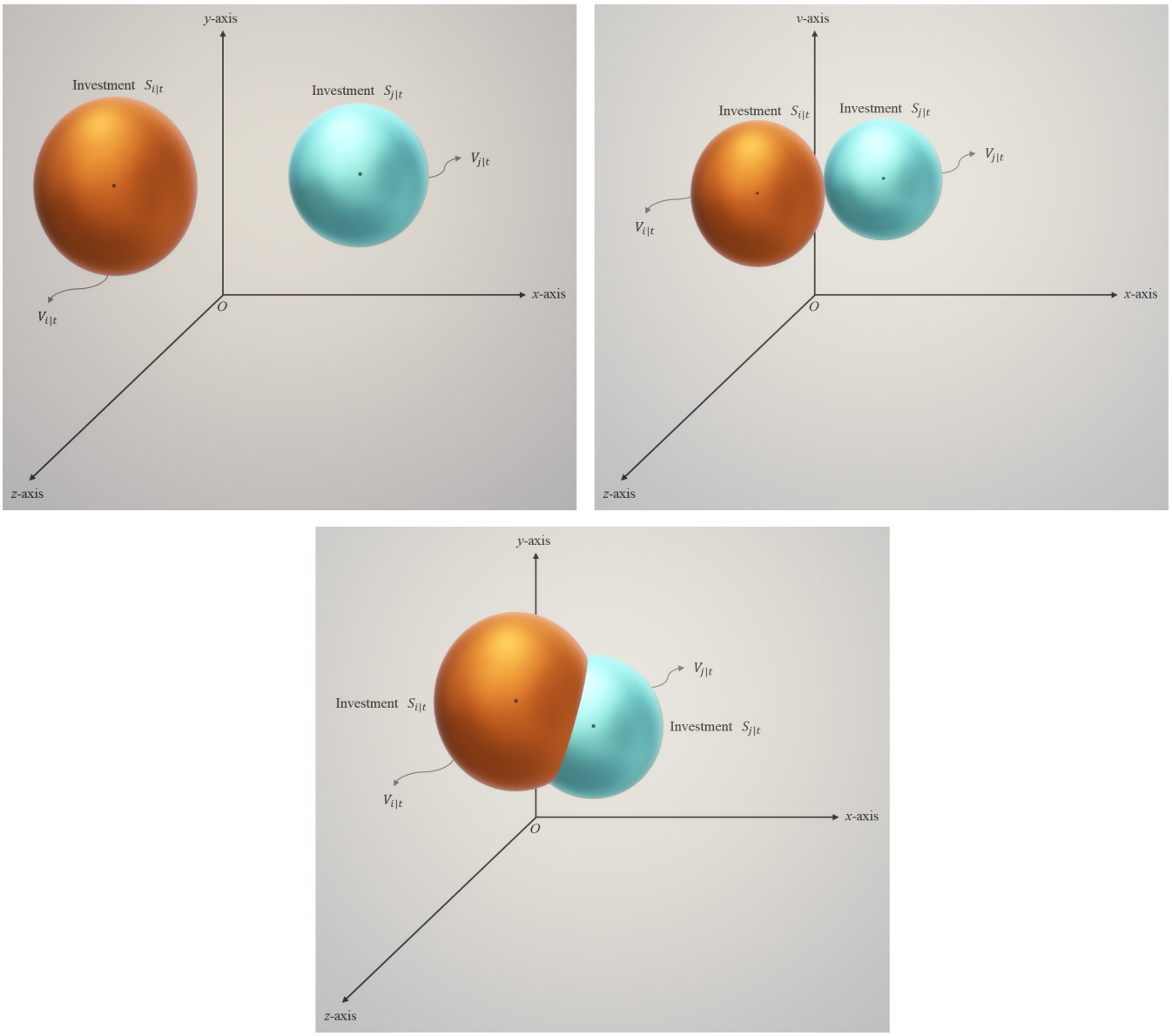

FIGURE 3 Two nonvolume sharing situations between investment alternatives $S_{i|t}$ and $S_{j|t}$ in $\mathbb{R}^3$ at time t, represented as trivial intersections $S_{i|t} \cap S_{j|t} = \emptyset$ (top left-hand side and right-hand side), and the respective volume sharing situation represented as the nontrivial intersection $S_{i|t} \cap S_{j|t} \neq \emptyset$ (bottom centre)

The measurement of each of the spherical volumes is found in step 1 by calculating the individual volume of each sphere that represents a single investment alternative at each point in time t. If there is an intersection between two investment alternatives $S_{i|t}$ and $S_{j|t}$ in $\mathbb{R}^3$, then the respective shared value between those two investment alternatives may be calculated. This shared value (i.e. volume greater than zero resulting from a nontrivial intersection between a pair of spheres) occurs due to (i) greater spatial proximity of the spheres representing investment alternatives, (ii) increased individual spherical volume of either both or even only one of the investment alternatives, or (iii) processes (i) and (ii) combined.

3.2.1 | The proposed method's key measure: the common feature $CF_{i,j|t}$

The shared volume between investment alternatives $S_{i|t}$ and $S_{j|t}$ is termed as the common feature $CF_{i,j|t}$ due to the fact that this volume/value is attributed to individual features of $V_{i|t}$ and $V_{j|t}$ that are shared between each pair of investment

alternatives $S_{i|t}$ and $S_{j|t}$, as follows:

$$CF_{i,j|t} := S_{i|t} \cap S_{j|t} \tag{5}$$

The common feature consists of a central measure in the SCA method because it is the value in which the level of similarity between every pair of objects in a sample is based upon (calculated in step 3) and it also refers to the linkage criterion adopted by the proposed cluster analysis (performed in step 4) to automatically identify the existence of "natural" clusters, allocating each object as a member of its respective cluster in the case there is at least one cluster in such a dataset. The analytic geometry solution to precisely calculate the volume shared between two spheres intersecting in $\mathbb{R}^3$ is detailed in Court[33] and Weisstein.[34]

3.3 | Step 3: Similarity analysis (pairwise)

The output values found in previous steps are used as input values to calculate a meaningful measure, termed as the similarity factor $SF_{i,j|t}$. This measure aims to, in an easily interpretable and comparable manner, indicate how similar (or dissimilar) are each pair of objects/investment alternatives in a dataset at each point in time. Hence, after finding the individual volumes $V_{i|t}$ and $V_{j|t}$ in step 1, and the common feature $CF_{i,j|t}$ between every pair of distinct spheres in the sample in step 2, the sum of the respective intersecting volumes are placed in the numerator and the sum of the individual volumes of both spheres are placed in the denominator of the following proposed similarity measure:

$$SF_{i,j|t} := \frac{\left(V_{i|t} \cap V_{j|t}\right) + \left(V_{j|t} \cap V_{i|t}\right)}{V_{i|t} + V_{j|t}} \tag{6}$$

which may be rewritten as:

$$SF_{i,j|t} := \frac{CF_{i,j|t} + CF_{j,i|t}}{V_{i|t} + V_{j|t}} \tag{7}$$

where:

$0 \leq SF_{i,j|t} \leq 1, \quad \forall SF_{i,j|t} \in \mathbb{Q}^+$

$0 \leq V_{1|t}, \ldots, V_{Q|t} < +\infty, \quad \forall V_{1|t}, \ldots, V_{Q|t} \in \mathbf{R}^+$

$0 \leq (CF_{1,2|t}) + \ldots + (CF_{Q,Q-1|t}) \leq (V_{1|t} + \ldots + V_{Q|t}),\ \forall (V_{1|t} + \ldots + V_{Q|t}) \in \mathbf{R}^+$ and $\forall (CF_{1,2|t}) + \ldots + (CF_{Q,Q-1|t}) \in \mathbf{R}^+$.

Schematically, the similarity factor $SF_{i,j|t}$ refers to a novel similarity measure that reflects the ratio between the intersection volume counted twice (two spherical caps in magenta) and the sum between the volumes of spheres $S_{i|t}$ (blue) and $S_{j|t}$ (red) – that is, $V_{i|t}$ and $V_{j|t}$, respectively, as depicted in Figure 4.

In the case in which there is a trivial intersection (i.e. $S_{i|t} \cap S_{j|t} = \emptyset$) the numerator would then be zero, which would yield $SF_{i,j|t} = 0$. Conversely, in the opposite extreme case in which both spheres are placed in the exact same three spatial coordinates and contain the precise same volume, then $SF_{i,j|t} = 1$. Thus, on one hand, the closer $SF_{i,j|t}$ gets to one, the more similar both investment alternatives are in a particular point in time. On the other hand, the closer $SF_{i,j|t}$ gets to zero, the more dissimilar, and separated apart in $\mathbb{R}^3$, are both investment alternatives compared to each other.

While performing the analysis applying real-world variables as data input, would then be highly unlikely, although not impossible, to reach the precise maximum value of one. It is worth mentioning that care must be taken in order to interpret outputs resulted from intersecting spherical volume values before translating them properly into meaningful similarity and clustering measures that would support the investment decision-making process.

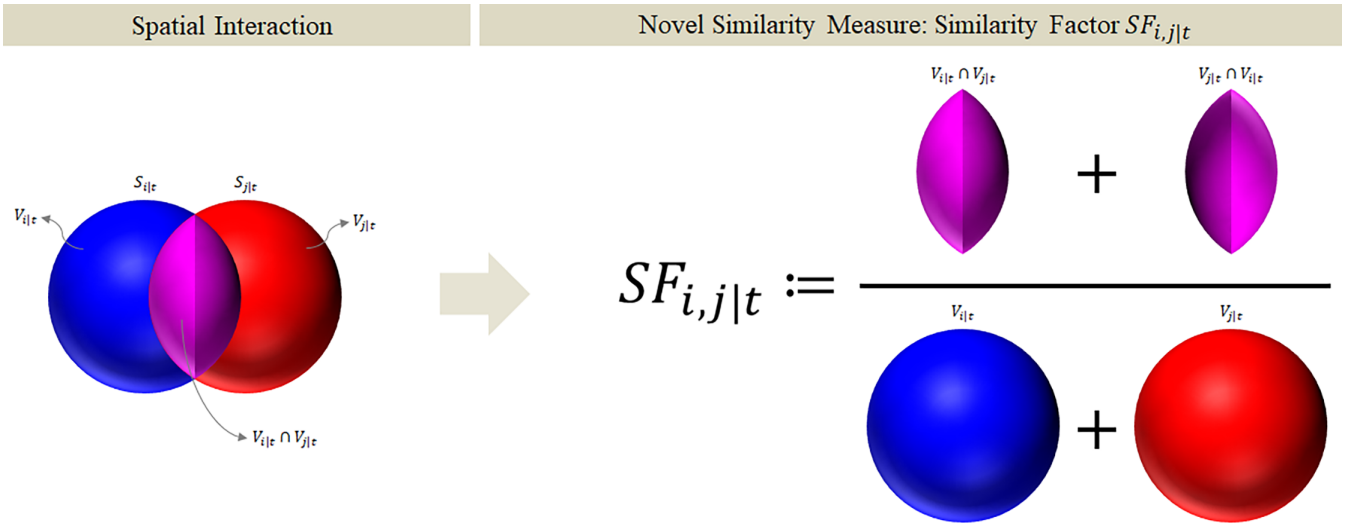

FIGURE 4 Graphical illustration of step 3 (right-hand side), which is performed subsequently after step 2 (left-hand side)

3.4 | Step 4: Cluster analysis (whole sample)

The proposed cluster analysis is performed in two phases. In phase 1, it is calculated the mean of all similarity factors $SF_{i,j|t}$ found in step 3, resulting in a value that measures the overall similarity level in the whole sample at once. In phase 2, the linkage criterion adopted sets the rule to, in the case there is at least one cluster, deterministically assign each object to its single respective cluster and then identify the number of clusters in the dataset.

3.4.1 | Phase 1. Overall similarity measure: the $\mathscr{R}_t$

The output values found in step 3 are used as input values to an overall clustering measure, termed as $\mathscr{R}_t$. This measure computes the precise level of similarity of all sample objects (e.g. investment alternatives) in a particular point in time as well as through time, by averaging the similarity factors $SF_{i,j|t}$ of all pairs of sample objects at once. Thus, the following times series clustering ratio of all similarity factors is calculated in step 4:

$$\mathscr{R}_t := \frac{1}{N_{SF|t}} \left(SF_{1,2|t} + SF_{2,1|t} + \ldots + SF_{Q-1,Q|t} + SF_{Q,Q-1|t} \right), \tag{8}$$

where:

$0 \leq \mathscr{R}_t \leq 1, \quad \forall \mathscr{R}_t \in \mathbb{Q}^+$

$2 \leq N_{SF|t} < +\infty, \quad \forall N_{SF|t} \in \mathbb{Z}^+$

$0 \leq SF_{i,j|t} \leq 1, \quad \forall SF_{i,j|t} \in \mathbb{Q}^+.$

The term $N_{SF|t}$ represents the total number of existing similarity factors at time t, as calculated in step 3. In the case there is at least one non-trivial intersection (i.e. $S_{i|t} \cap S_{j|t} \neq \emptyset$), both the numerator and denominator in Equation (8) would consist of a value greater than zero, which would result in $\mathscr{R}_t > 0$. Moreover, in the highly unlikely extreme case in which all spheres in the sample are placed in the exact same three coordinates and contain the precise same volume, then $\mathscr{R}_t = 1$. Conversely, in the case there is no intersection volume between any pair of spheres in a dataset, then $\mathscr{R}_t = 0$.

Therefore, analogously to the similarity factor $SF_{i,j|t}$ measure, the closer $\mathscr{R}_t$ gets to the value of one, the more similar and grouped the investment alternatives are in a sample as a whole. On the other hand, the closer $\mathscr{R}_t$ gets to zero, the

more dissimilar are the investment alternatives from each other as a whole (i.e. considering all objects in the sample at once). Once real-world variables are used as data input, it is expected that most, if not all, results would lie within those two extreme values in the interval [0,1], which does not mean whatsoever that the number of clusters in a sample is expected to be one, as detailed in the Section 3.4.2 below.

3.4.2 | Phase 2. Cluster linkage criterion

The last phase of the last step of the proposed method is to identify the existence of clusters in the dataset. In the case there is at least one cluster in such a dataset, as the proposed method refers to a deterministic hard cluster analysis, each object is then assigned as a cluster member of its respective single cluster at each point in time. The rule adopted in the proposed method to recognise clusters in a dataset is conditional upon to the existence of nontrivial volume intersection between two or more spheres, which cluster members would be either directly or indirectly linked/connected through intersecting spherical volumes.

The direct link/connection exists in the case two or more spheres have an intersecting volume shared between themselves. An indirect link/connection refers to a situation in which two spheres have no intersecting volume between each other, but both share an intersecting volume with a common third sphere. There is a nontrivial volume intersection between two spheres $S_{i|t}$ and $S_{j|t}$ in the case that, if and only if:

$$d_{\boldsymbol{i,j|t}} = d_{\boldsymbol{j,i|t}} = \left\| C_{i|t} - C_{j|t} \right\| < r_{i|t} + r_{j|t} \quad \therefore \quad S_{i|t} \cap S_{j|t} \neq \emptyset, \tag{9}$$

where $d_{\boldsymbol{i,j|t}}$ or $d_{\boldsymbol{j,i|t}}$ is the Euclidean distance between $C_{i|t}$ and $C_{j|t}$, which refer to the centres of spheres $S_{i|t}$ and $S_{j|t}$, respectively, and $\|\cdot\|$ is the Euclidean norm; all variables are at time t. The calculation in Equation (9) involving only the spheres $S_{i|t}$ and $S_{j|t}$ is, in fact, performed for the whole sample, performed $Q(Q-1)$ times, from object $S_{1|t}$ to $S_{Q|t}$ from $q = (1, \ \ldots, i, j, \ \ldots, Q)$, which corresponds to the total number of meaningful (i.e. excluding self-interaction cases, such as $d_{\boldsymbol{i,i|t}} = 0$) pairwise interactions in a sample with Q objects. In fact, the number of unique meaningful pairwise interactions is $2^{-1}[Q(Q-1)]$ due to the fact that every pairwise interaction is generated twice, being one related to the $S_{i|t} \cap S_{j|t} = a_t$ and the other related to its mirror value $S_{j|t} \cap S_{i|t} = b_t$, being $a_t = b_t$.

Moreover, following the definition that a cluster consists of a group formed by at least two members, then the number of clusters in a dataset according to the proposed method lies in the range $\{0, \ \ldots, Q2^{-1}\}$, $Q \in \mathbb{Z}^+$. It is worth noticing that following this definition (i.e. a cluster must be composed by a group of two or more members), it is excluded then the unreasonable possibility of a cluster being formed by a single member, in which case such a single cluster member should be considered as either a potential multivariate outlier[**] or even an actual multivariate outlier - depending on which arbitrary threshold would be adopted to classify a sample object as multivariate outlier.

4 | EMPIRICAL RESULTS

The present section contains empirical analyses based on two distinct datasets used as input data. Firstly, in order to evaluate the SCA method against competing alternative methods, the best-known publicly available dataset in the existing pattern recognition literature, namely the Fisher's Iris dataset,[35,36] is used as input data. In this first empirical analysis, clustering outputs of the SCA method are compared to outputs of the k-means, k-medians, and k-medoids methods. These competing clustering methods compose the benchmarking and most popular family of cluster analysis.[19,37]

In the second empirical analysis, in order to illustrate the potential practical usefulness of the proposed method to tackle relevant real-world problems, stock market returns and macroeconomic variables of G7 economies are used as input data to the SCA method. This empirical analysis contains results that unveils the temporal dynamic of pairwise similarity levels between international investment alternatives (i.e. each possible pair composed by G7 members) and the structure of market clustering patterns through time (i.e. the G7 as a whole).

[**] In order to be as precise as possible, in this section it is used the word "multivariate outlier" instead of only outlier due to the fact that each sample object in the SCA method in the present paper is, in fact, a combination of values from four input variables ($x_{i|t}$, $y_{i|t}$, $z_{i|t}$, and $V_{i|t}$) rather than an extreme single data point (i.e. the classical definition of outlier). However, for instructional purposes at the expense of terminology preciseness, it is still possible to use both terms in the SCA context.

4.1 | Fisher's Iris data

The Fisher's Iris dataset is a multivariate dataset collected under, presumably, equal conditions in order to quantify the morphologic variation of the following three related flower species: Iris-setosa, Iris-versicolor, and Iris-virginica.[38] This dataset contains a total of 600 records, reflecting a sample of 50 observations of each of these three Iris species, and for each observation the following four variables is measured in centimetres: length of sepals, width of sepals, length of petals, and width of petals. A few descriptive statistics of this dataset is reported in Table 1.

TABLE 1 Summary statistics of the Fisher's Iris dataset

Class/characteristics (in cm)	Mean	Median	Variance	SD	Smallest	Largest	Range	Obs
Iris-setosa								
Sepal length	5.01	5.00	0.12	0.35	4.30	5.80	1.50	50
Sepal width	3.42	3.40	0.15	0.38	2.30	4.40	2.10	50
Petal length	1.46	1.50	0.03	0.17	1.00	1.90	0.90	50
Petal width	0.24	0.20	0.01	0.11	0.10	0.60	0.50	50
Iris-versicolor								
Sepal length	5.94	5.90	0.27	0.52	4.90	7.00	2.10	50
Sepal width	2.77	2.80	0.10	0.31	2.00	3.40	1.40	50
Petal length	4.26	4.35	0.22	0.47	3.00	5.10	2.10	50
Petal width	1.33	1.30	0.04	0.20	1.00	1.80	0.80	50
Iris-virginica								
Sepal length	6.59	6.50	0.40	0.64	4.90	7.90	3.00	50
Sepal width	2.97	3.00	0.10	0.32	2.20	3.80	1.60	50
Petal length	5.55	5.55	0.30	0.55	4.50	6.90	2.40	50
Petal width	2.03	2.00	0.08	0.27	1.40	2.50	1.10	50

Notes: "SD" means standard deviation and "Obs" stands for the number of observations in the sample. Moreover, except for "Obs", all values are in centimetres.

The dataset summarised in Table 1, used in the empirical analysis of this subsection, is collected from the Machine Learning Repository of the University of California, Irvine.[39]

4.1.1 | Fisher's Iris dataset empirical analysis

From a number of previous studies using the Fisher's Iris dataset to tackle various research problems in different fields of knowledge (e.g. references 40-42), it is well-known that this dataset contains two clearly segregated clusters. One cluster is composed only by Iris-setosa observations, and the remaining cluster contains Iris-versicolor and Iris-virginica observations. Therefore, although the first empirical cluster analysis of the present paper is performed setting the number of clusters as three (i.e. $k = 3$) using the k-means method, this is only to explore and detail such a stylised fact of this popular dataset. However, all empirical clusters comparisons in this subsection are based on clustering methods considering the number of clusters set as two (i.e. $k = 2$).

Following the previously known classification of the input dataset (i.e. three species of Iris flowers) and only for illustrative purposes, the k-means clustering is then firstly set to allocate all sample observations into three distinct clusters (i.e. $k = 3$). The outputs of the analysis with three clusters confirm previous studies in the sense that the Iris-setosa (i.e. cluster 3) clearly forms a separate cluster, while clusters 1 and 2 - composed by Iris-versicolor and Iris-virginica, respectively - are not distinctly segregated, suggesting that, due to their respective comparatively more similar characteristics, they potentially form a single cluster, as detailed in Table 2, Figure 5, and Figure 6.

TABLE 2 Summary statistics of each cluster formed by the k-means method, with $k = 3$

Class/characteristics (in cm)	Mean	Smallest	Largest	Range
Cluster 1 ($n = 61$)				
Sepal length	5.88	4.90	6.80	1.90
Sepal width	2.74	2.00	3.40	1.40
Petal length	4.39	3.00	5.10	2.10
Petal width	1.43	1.00	2.40	1.40
Cluster 2 ($n = 39$)				
Sepal length	6.85	6.10	7.90	1.80
Sepal width	3.08	2.50	3.80	1.30
Petal length	5.72	4.70	6.90	2.20
Petal width	2.05	1.40	2.50	1.10
Cluster 3 ($n = 50$)				
Sepal length	5.01	4.30	5.80	1.50
Sepal width	3.42	2.30	4.40	2.10
Petal length	1.46	1.00	1.90	0.90
Petal width	0.24	0.10	0.60	0.50

Notes: All values are in centimetres.

Out of the 61 observations allocated to cluster 1, Iris-versicolor totalises 47 observations (77%) and 14 observations (23%) refer to Iris-virginica. On the other hand, from the 39 observations that are members of cluster 2, Iris-virginica accounts to 36 observations (92%), and Iris-versicolor totalises three (8%) observations. Moreover, all members of cluster 3 consist of Iris-setosa.

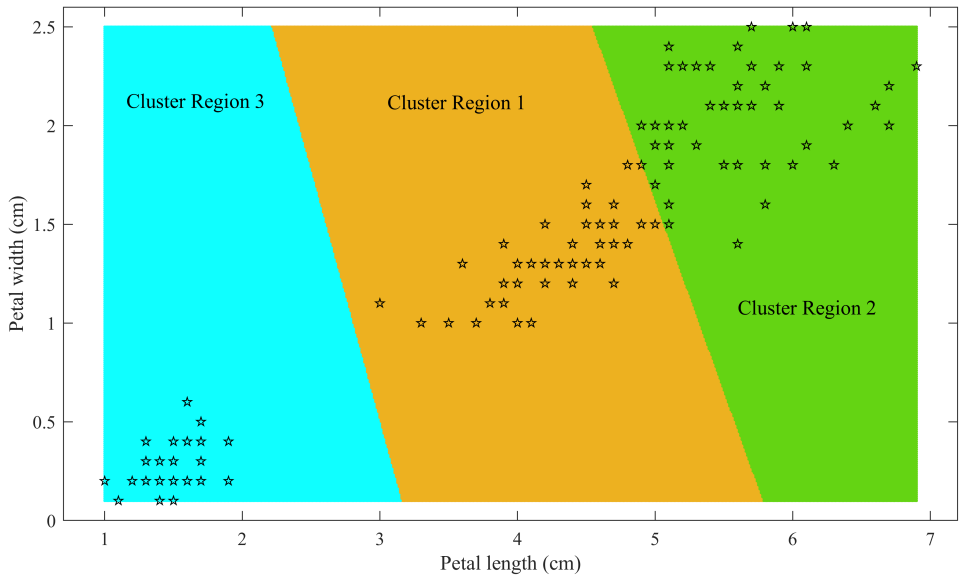

FIGURE 5 Two-dimensional plot (petal width and length) highlighting the three cluster regions according to the k-means method performed on the Fisher's Iris dataset

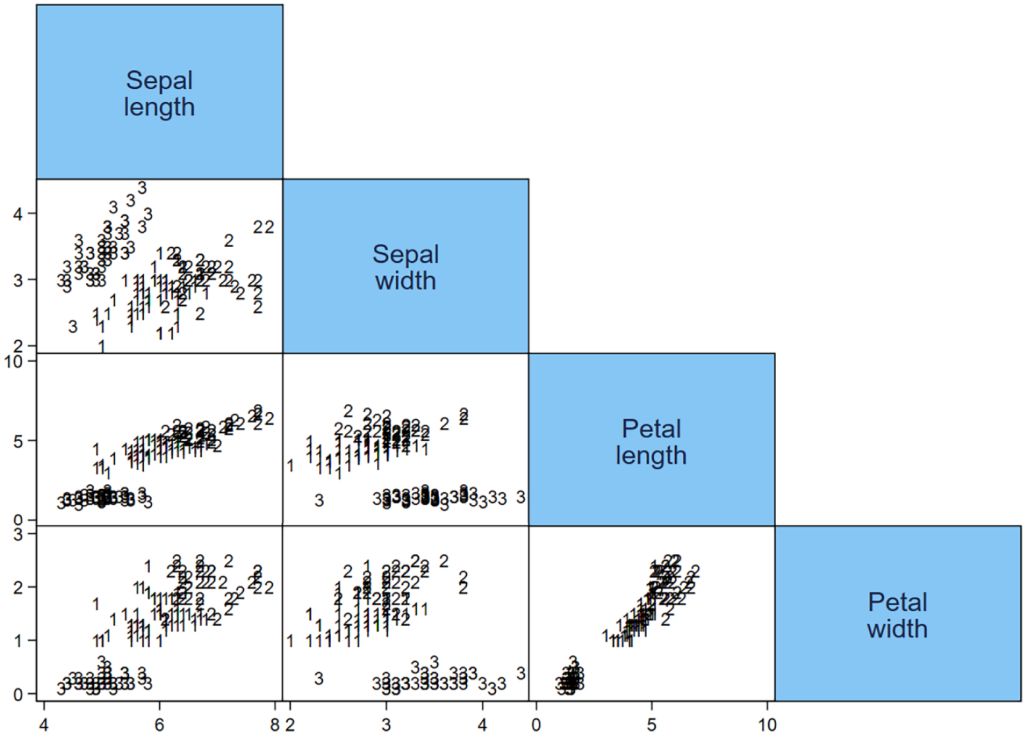

FIGURE 6 Matrix chart detailing cluster members allocated according to the *k*-means method, in two-dimensional charts with relationship between each pair of characteristics

In the matrix chart in Figure 6 it is possible to visualise that clusters 1 and 2 are, very frequently, in the same area of each of the charts, while cluster 3 is almost always clearly separated from the remaining two clusters. Such *k*-means results lead to the intuition that, possibly, the "natural" number of clusters in the Fisher's Iris dataset is, in fact, two instead of three, despite the fact that there are three species in such a sample.

Established clustering methods outputs (k = 2)

In order to reflect the clustering structure of this dataset in a more realistic manner based on the empirical analysis detailed in the previous subsection, the *k*-means, *k*-medians, and *k*-medoids clustering methods are then performed considering the number of clusters being set as two (i.e. $k = 2$). More specifically, the *k*-means and *k*-medoids clustering methods are based on the squared Euclidean distance metric, using the *k*-means++ algorithm for choosing initial cluster medoid positions.[43] The empirical results of these three clustering methods are reported in Table 3.

As reported in Table 3, the outputs of these three clustering methods are very similar, being identical the outputs of *k*-means and *k*-medoids. The pattern of these clusters, being $k = 2$, consists of allocating all 50 Iris-virginica and almost all Iris-versicolor observations into cluster 1, while all 50 Iris-setosa observations and between one (*k*-means and *k*-medoids) and two (*k*-medians) Iris-versicolor observations form cluster 2.

TABLE 3 Summary statistics of each cluster formed by *k*-means, *k*-medians, and *k*-medoids methods, with $k = 2$

Cluster/ characteristics	*k*-means				Cluster/ characteristics	*k*-medians				Cluster/ characteristics	*k*-medoids			
	Mean	**Smallest**	**Largest**	**Range**		**Mean**	**Smallest**	**Largest**	**Range**		**Mean**	**Smallest**	**Largest**	**Range**
Cluster 1 ($n = 99$)					Cluster 1 ($n = 98$)					Cluster 1 ($n = 99$)				
Sepal length	6.27	4.90	7.90	3.00	Sepal length	6.29	4.90	7.90	3.00	Sepal length	6.27	4.90	7.90	3.00
Sepal width	2.88	2.00	3.80	1.80	Sepal width	2.88	2.00	3.80	1.80	Sepal width	2.88	2.00	3.80	1.80
Petal length	4.93	3.30	6.90	3.60	Petal length	4.94	3.30	6.90	3.60	Petal length	4.93	3.30	6.90	3.60
Petal width	1.68	1.00	2.50	1.50	Petal width	1.69	1.00	2.50	1.50	Petal width	1.68	1.00	2.50	1.50
Cluster 2 ($n = 51$)					Cluster 2 ($n = 52$)					Cluster 2 ($n = 51$)				
Sepal length	5.01	4.30	5.80	1.50	Sepal length	5.01	4.30	5.80	1.50	Sepal length	5.01	4.30	5.80	1.50
Sepal width	3.40	2.30	4.40	2.10	Sepal width	3.38	2.30	4.40	2.10	Sepal width	3.40	2.30	4.40	2.10
Petal length	1.49	1.00	3.00	2.00	Petal length	1.53	1.00	3.30	2.30	Petal length	1.49	1.00	3.00	2.00
Petal width	0.26	0.10	1.10	1.00	Petal width	0.28	0.10	1.10	1.00	Petal width	0.26	0.10	1.10	1.00

Notes: All values are in centimetres

SCA outputs (data-driven k)

The proposed cluster analysis is then also performed using the Fisher's Iris dataset, placing the four input variables as follows: sepal length, sepal width, and petal length into the axes x, y, and z, and the petal width as the spherical volume. The respective outputs are reported in Table 4.

TABLE 4 Summary statistics of each cluster formed while performing the SCA method

Class/characteristics (in cm)	Mean	Smallest	Largest	Range
Cluster 1 ($n = 100$)				
Sepal length	6.26	4.90	7.90	3.00
Sepal width	2.87	2.00	3.80	1.80
Petal length	4.91	3.00	6.90	3.90
Petal width	1.68	1.00	2.50	1.50
Cluster 2 ($n = 50$)				
Sepal length	5.00	4.30	5.80	1.50
Sepal width	3.42	2.30	4.40	2.10
Petal length	1.46	1.00	1.90	0.90
Petal width	0.24	0.10	0.60	0.50

Notes: All values are in centimetres

After placing all four input variables in their respective axis or spherical volume, the proposed cluster analysis then splits, in a data-driven fashion, the data into two clusters. Cluster 1 is composed by all Iris-virginica and Iris-versicolor observations, totalling 100 observations, while cluster 2 is composed only by the 50 observations of the Iris-setosa. These results are graphically depicted in the three-dimensional plots in Figure 7.

Concomitantly to the performance of the proposed cluster analysis, the pairwise similarity factor $SF_{i,j|t}$ is also precisely calculated for each pair of objects in the sample. As the resulting 150×150 symmetric matrix would not fit in the space of this paper, then the respective heat map reflecting all similarity factors is used instead, as shown in Figure 8.

In such a heat map, the darker the colour the lower the value, in which the black colour refers to the value of zero. Conversely, the closer to the white colour the greater the value, in which the white colour represents the value of one. As depicted in Figure 8, the upper left part of the heat map corresponds to the 50 observations of the Iris-setosa, which shows a high similarity level within its group but zero spherical volume intersection with the remaining 100 observations of the two distinct flower species. This is the cause of the large black rectangular parts in this heat map.

Moreover, it is worth mentioning that the axial and volume configuration (i.e. x, y, z, and V) of the input values considered in this particular toy example worked well for the Fisher's Iris dataset, although variations of it yields similar but still distinct results, which refers to a caveat yet to be addressed regarding the proposed cluster analysis.

Empirical comparison between clustering methods

In order to objectively compare the four cluster analyses considered in this subsection, the following five established classification measures are applied on each of the clustering outputs: sensitivity or true positive rate (TPR), specificity or true negative rate (TNR), precision or positive predictive value (PPV), negative predictive value (NPV), and F-measure.

As detailed in Table 5, all four clustering methods yield very similar performances, in which the proposed cluster analysis (SCA) results in a slightly larger precision (i.e. 0.5) and F-measure (i.e. 0.571) values. The precision measure corresponds to the ratio between true positives (TP) and the sum of TP with false positives (FP), while the F-measure refers to the harmonic mean of precision and sensitivity.[44,45]

Still according Table 5, all four clustering methods report a sensitivity value of 0.67, which measure is the ratio between TP and the sum between TP and false negatives (FN). In terms of specificity, all methods result in a value of 0.83; such a measure calculates the ratio between true negatives (TN) and the sum between TN and FP. Moreover, the largest measure value identically shown by all four cluster analyses refers to the NPV of 0.89. This last measure corresponds to the value of TN over the sum between TN and FN.[46]

One common underlying assumption of the well-established cluster analyses used in this subsection is that all of them require that the number of clusters must be known in advance. In addition, a common general challenging problem in the clustering literature refers to the presence of outliers, and even methods have been developed specifically to detect and remove outliers in partitional and hierarchical clustering analysis.[47] However, in many instances such assumptions and *ad hoc* methods are largely invalid, potentially leading to poor and unrealistic results.

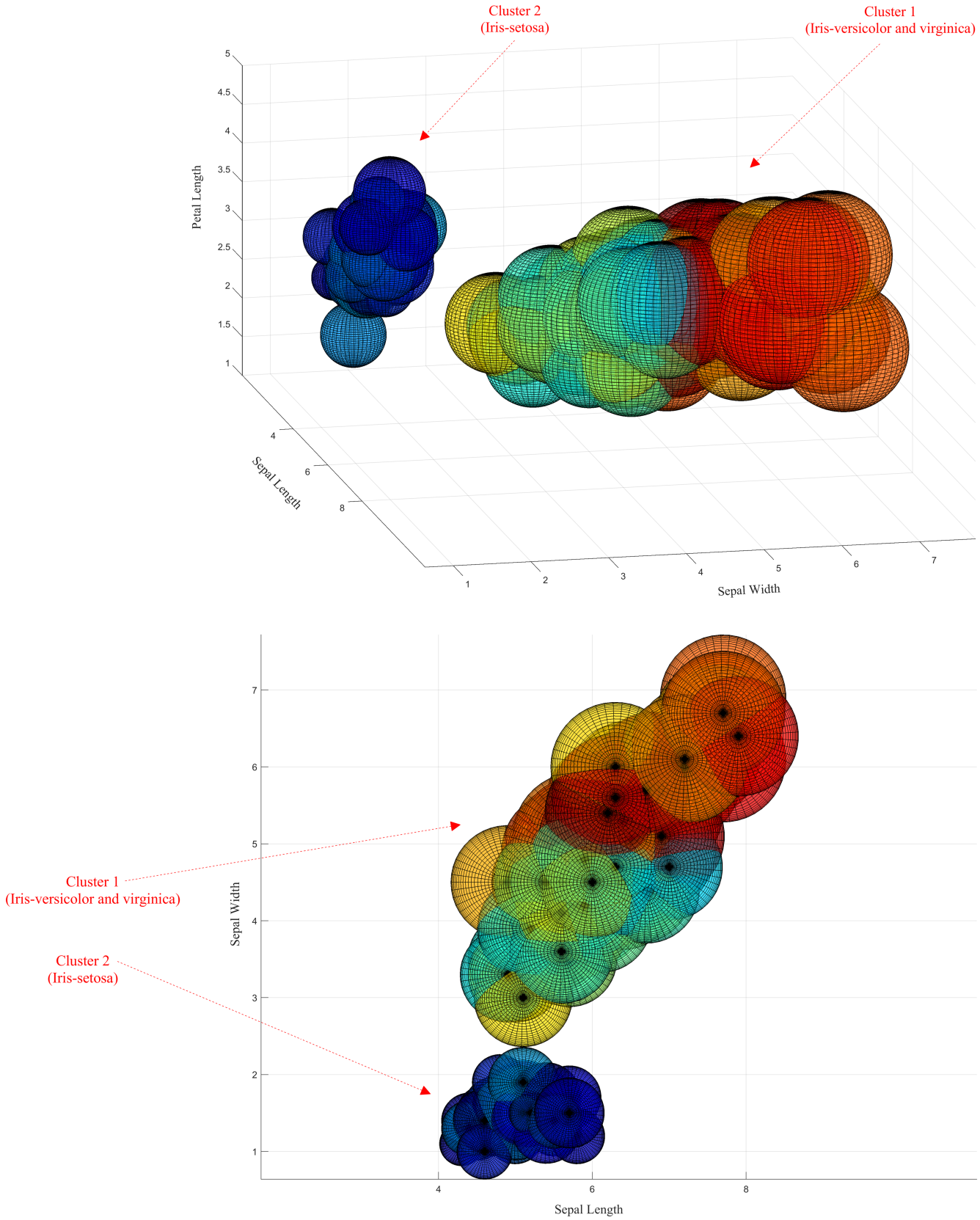

FIGURE 7 Three-dimensional plots (lateral and aerial view, placed at the upper and bottom side, respectively) of the proposed cluster analysis performed on the Fisher's Iris dataset used as input data, which variables are in centimetres

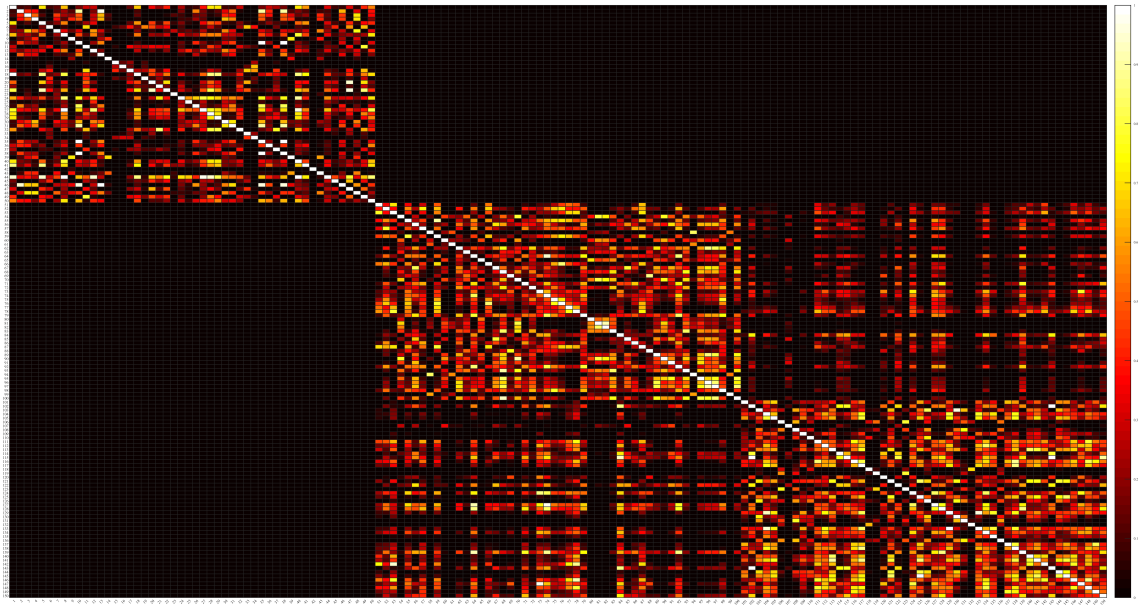

FIGURE 8 Similarity factor matrix represented as a heat map

TABLE 5 Clustering methods compared according to five established classification measures

Clustering method	Sensitivity (TPR)	Specificity (TNR)	Precision (PPV)	Negative Predictive Value (NPV)	*F*-measure
SCA	0.667	0.833	0.500	0.889	0.571
k-means	0.667	0.833	0.495	0.889	0.568
k-medians	0.667	0.833	0.491	0.889	0.565
k-medoids	0.667	0.833	0.495	0.889	0.568

Therefore, considering the limitations of the established methods aforementioned, much more importantly than this marginally superior but, in fact, virtually inexistent performance shown by the proposed cluster analysis, it is the fact that, differently from all competing established clustering methods, the number of clusters in the SCA is provided in a data-driven manner instead of being arbitrarily set/imposed/forced by the analyst. Such a data-driven mechanism may potentially find and reveal interesting clustering patterns and insights extracted from further real-world datasets as, for example, demonstrated in the subsequent empirical analysis involving economic/financial data. Potentially, due to methodological and respective algorithmic limitations, a large part of such clustering patterns and insights revealed by the SCA method may not be properly captured by traditional clustering methods.

4.2 | Stock market and macroeconomic data

The data used as input variables in this subsection are the daily stock market return and three macroeconomic variables of G7 economies, one at national level and two global variables, from 3 October 2003 to 27 December 2017. More specifically, the following four time series refer to the variables used as data input, applied as daily simple return: stock market index, foreign exchange rate, oil price, and gold price. The former two variables refer to national variables of each of the G7 economies, and the last two ones consist of relevant global macroeconomic variables, totalling 170,268 data points. All data are collected from Bloomberg and Thomson Reuters' Eikon.

More specifically about the main input variable of this study (i.e. stock market return), it refers to the daily (closing position) variation of the following main stock market indices: S&P/TSX Composite (Canada), CAC 40 (France), DAX (Germany), FTSE MIB (Italy), Nikkei 225 (Japan), FTSE 100 (United Kingdom), and S&P 500 (United States).

4.2.1 | Pre-analysis data treatments

Before start performing the SCA analysis itself, it is necessary to treat the datasets twice. The first data treatment refers to the fulfillment of missing values through a multiple imputation technique,[48] in which the Expectation-Maximization with Bootstrapping (EMB[††]) algorithm is applied to the variables. The second data treatment is performed due to the fact that, as Japan is the only Eastern country in the sample, its stock market operates with no overlapping trading hours related to the remaining G7 stock markets and, therefore, it is operationally one trading day ahead. Thus, following a procedure commonly adopted in the finance literature (e.g. references 49-52) and as a matter of trading day synchronisation, all Japanese time series actually reflect the date of $t+1$.

4.2.2 | Input data descriptive statistics

The data distribution of the daily returns of stock market, foreign exchange, oil price, and gold price follow stylised facts of the finance literature.[53] As shown in Figure 9 and Table 6, such time series commonly present negative skewness (third central moment), excess kurtosis (fourth central moment), volatility clustering, and are mean-reverting processes.

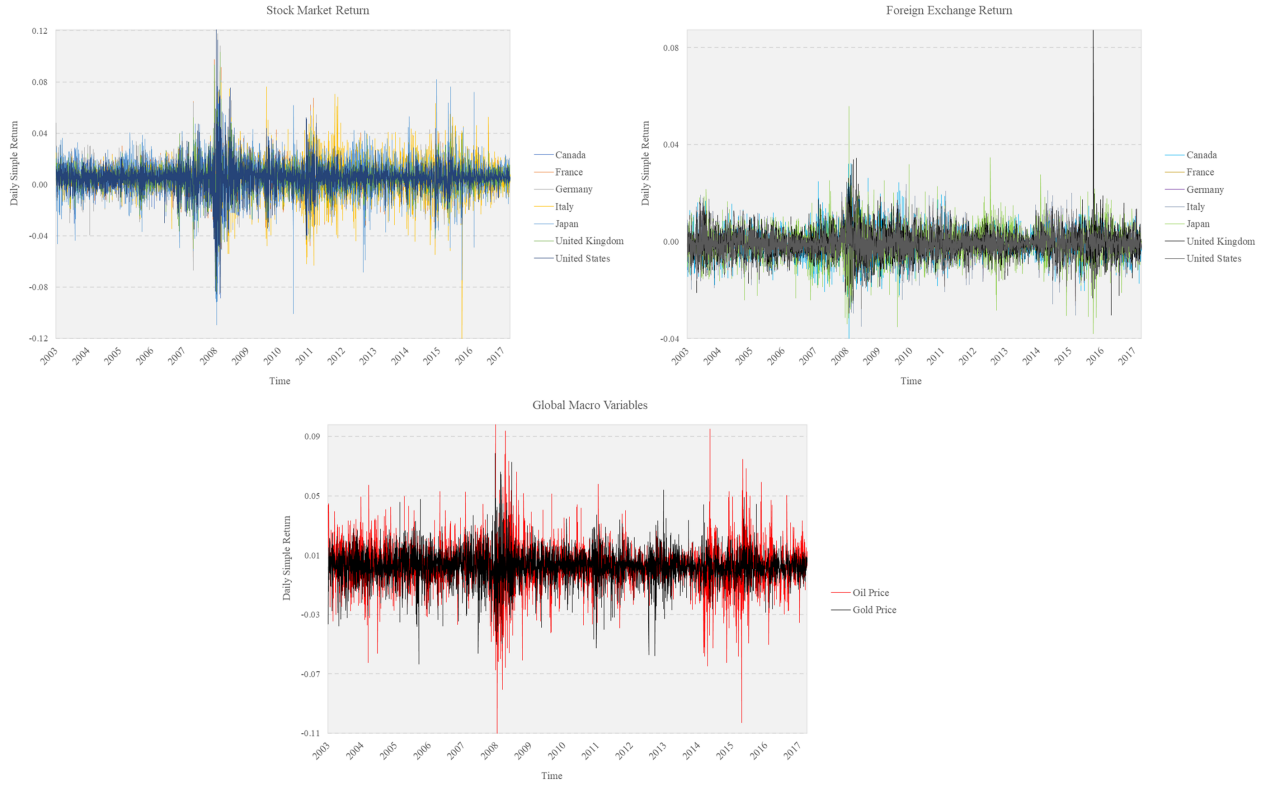

FIGURE 9 Line charts of time series of G7 stock market daily returns (left-hand side), foreign exchange daily returns (right-hand side), and oil and gold price daily variation (bottom)

Considering that the SCA method does not require the data to follow the Gaussian distribution, or at least a distribution similar to a normal one, then it is not required to transform stock market returns into their respective natural logarithm, as commonly adopted in the finance literature.

[††] The R package Amelia II is used to apply the EMB algorithm for missing values fulfillment through multiple imputation, in which $m = 5$ is adopted in the present study, as usually recommended in the respective literature.

TABLE 6 Descriptive statistics of G7 daily stock market return and macroeconomic variables

Variable/Geography	Mean	Median	Variance	SD	SE	Skewness	Kurtosis	Smallest	Largest	Range	Obs
Stock Market Return											
Canada	0.0002	0.0007	0.0001	0.0099	0.0002	−0.6585	11.8871	−0.0932	0.0720	0.1652	4,054
France	0.0002	0.0004	0.0002	0.0127	0.0002	0.1112	7.9598	−0.0904	0.1118	0.2022	4,054
Germany	0.0004	0.0009	0.0002	0.0127	0.0002	0.1464	7.2940	−0.0716	0.1140	0.1856	4,054
Italy	0.0001	0.0005	0.0002	0.0143	0.0002	−0.0905	6.2903	−0.1248	0.1149	0.2397	4,054
Japan	0.0003	0.0005	0.0002	0.0139	0.0002	−0.4197	7.7387	−0.1141	0.1108	0.2249	4,054
United Kingdom	0.0002	0.0003	0.0001	0.0105	0.0002	0.0402	10.4689	−0.0885	0.0984	0.1869	4,054
United States	0.0003	0.0006	0.0001	0.0108	0.0002	−0.0183	14.2802	−0.0903	0.1158	0.2061	4,054
Foreign Exchange Return											
Canada	0.0000	0.0000	0.0000	0.0058	0.0001	0.1210	2.9122	−0.0392	0.0331	0.0723	4,054
France	0.0000	0.0000	0.0000	0.0058	0.0001	−0.0802	2.1469	−0.0342	0.0250	0.0592	4,054
Germany	0.0000	0.0000	0.0000	0.0058	0.0001	−0.0802	2.1469	−0.0342	0.0250	0.0592	4,054
Italy	0.0000	0.0000	0.0000	0.0058	0.0001	−0.0802	2.1469	−0.0342	0.0250	0.0592	4,054
Japan	0.0000	0.0000	0.0000	0.0061	0.0001	−0.0115	4.9824	−0.0371	0.0566	0.0937	4,054
United Kingdom	0.0001	0.0000	0.0000	0.0057	0.0001	1.1254	15.8476	−0.0295	0.0881	0.1176	4,054
United States	0.0000	0.0000	0.0000	0.0048	0.0001	−0.0228	2.1722	−0.0269	0.0255	0.0524	4,054
Global Macroeconomic Variables											
Oil Price	0.0003	0.0002	0.0002	0.0144	0.0002	−0.1077	5.8859	−0.1121	0.0960	0.2081	4,054
Gold Price	0.0004	0.0005	0.0001	0.0111	0.0002	−0.0875	4.2967	−0.0654	0.0767	0.1421	4,054

Notes: "SD" means standard deviation, "SE" refers to standard error, and "Obs" reflects the number of observations in the sample.

As all series behaved as expected, following results frequently reported in the literature, further time series analysis (e.g. autocorrelation, partial autocorrelation, unit root tests) are then omitted in this paper.

4.2.3 | Similarity analysis outputs of G7

The first task of the similarity analysis is computing every possible combination of intersection volume between every pair of spheres placed in the three-dimensional Euclidean space, totaling 21 nonduplicated common feature outputs for each of the 4,054 trading days. Subsequently, these common features and the respective volumes of each sphere are placed in the numerator and denominator of the similarity factor measure in each trading day, respectively. After calculating the mean of the similarity factor measure for the time period from October 2003 to December 2017, a similarity factor matrix is built, as reported in Table 7.

TABLE 7 Similarity factor lower triangular matrix (mean values)

$SF_{i,j}$	Canada	France	Germany	Italy	Japan	United Kingdom	United States
Canada	1.000						
France	0.562	1.000					
Germany	0.550	0.696	1.000				
Italy	0.548	0.665	0.642	1.000			
Japan	0.517	0.536	0.532	0.534	1.000		
United Kingdom	0.574	0.644	0.629	0.604	0.523	1.000	
United States	0.592	0.573	0.566	0.553	0.524	0.581	1.000

The similarity factor matrix in Table 7 consists of a symmetric matrix with its upper triangular part omitted only to avoid showing the respective transposed duplicated values. In addition, in order to highlight some interesting data patterns, Table 8 ranks all similarity factors, except for the self-similarity factor cases - which always yields one.

TABLE 8 Rank of the similarity factor matrix

Rank	G7 Pair (i, j)	$SF_{i,j}$ Mean
1	France and Germany[a]	0.696
2	France and Italy[a]	0.665
3	France and United Kingdom[a]	0.644
4	Germany and Italy[a]	0.642
5	Germany and United Kingdom[a]	0.629
6	Italy and United Kingdom[a]	0.604
7	Canada and United States	0.592
8	United Kingdom and United States	0.581
9	Canada and United Kingdom	0.574
10	France and United States	0.573
11	Germany and United States	0.566
12	Canada and France	0.562
13	Italy and United States	0.553
14	Canada and Germany	0.550
15	Canada and Italy	0.548
16	Japan and France[b]	0.536
17	Japan and Italy[b]	0.534
18	Japan and Germany[b]	0.532
19	Japan and United States[b]	0.524
20	Japan and United Kingdom[b]	0.523
21	Japan and Canada[b]	0.517
	Overall Mean	0.578
	Overall SD	0.051

[a] It indicates pairs involving only European countries.
[b] It indicates pairs involving Japan.

A number of interesting data characteristics and patterns - some expected and also a few unusual ones - are revealed by the similarity factor matrix and respective ranking in Tables 7 and 8. Pairs composed of two European countries have the largest similarity factors, being on average 0.647, which is 12% above the overall mean of 0.578. Moreover, the France and Germany pair produces the largest similarity factor (0.696) of the sample. Potential reasons to explain this may be related to the fact that most of these economies adopt the same currency (i.e. Euro), they are traditional trade partners, and their financial systems are highly interconnected. Conversely, Japan is the country with the six smallest similar factors in the sample, being 0.528 on average. Possible explanations may rely on the fact that Japan is the only Eastern country in the sample, which trading day has no overlapping hours with any other remaining G7 stock market.

In addition, France is more similar to Germany and Italy in comparison to the United Kingdom. A possible explanation for this may rely on the fact that the United Kingdom is the only economy among these four countries that does not adopt the Euro as its currency but the British pound instead. Despite the fact that France and the United Kingdom adopt different currencies (i.e. Euro and British pound, respectively), this pair of European countries are slightly more similar if compared to the pair Germany and Italy, which countries adopt the same currency (i.e. Euro). Moreover, Canada and United States consist of the highest similarity factor (0.592) just after all pairs involving only European countries. Potential reasons to explain this may be related to the fact these two economies are traditional trade partners, their financial systems are highly interconnected, and they tend to experience similar fluctuations in their macroeconomic and financial indicators, mainly Canada following movements/changes in US markets.

The charts in Figure 10 illustrate data patterns described above and summarised in Table 9. However, it is worth noticing the dynamic of these pairwise relations through time, in which, for instance, despite the fact that almost through

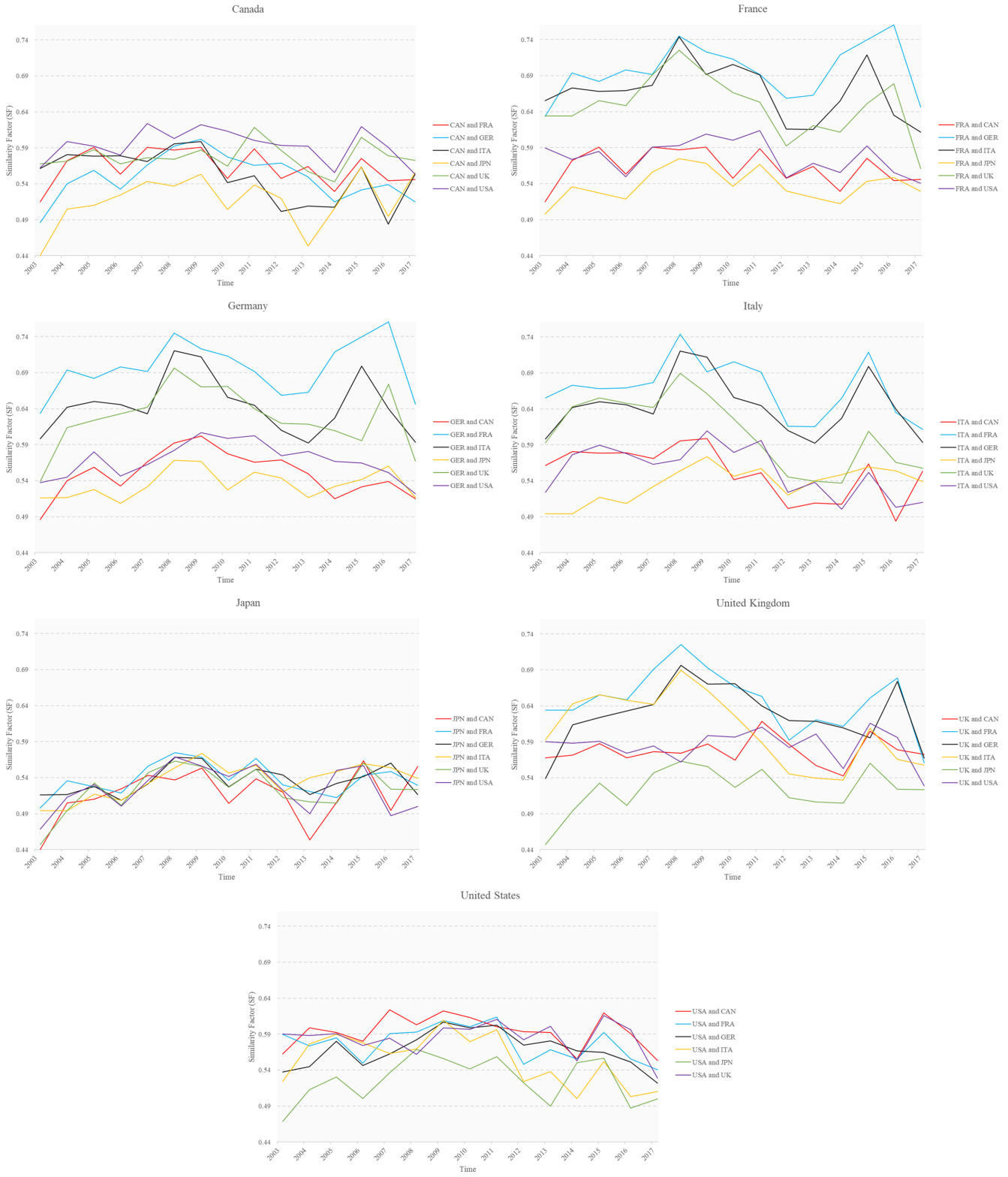

FIGURE 10 Annual time series of each similarity factor between G7 countries, considering all possible pairs in the sample.
Notes. CAN, FRA, GER, ITA, JPN, UK, and USA correspond to Canada, France, Germany, Italy, Japan, the United Kingdom, and the United States, respectively

the whole sample period the United States is the country that forms the pair with the largest similarity value among all Canadian pairwise series, for a brief period of time between 2011 and 2012, the United Kingdom becomes more similar to Canada compared to the United States.

In addition, an interesting pattern revealed by the time series in Figure 10 is that virtually all similarity factors experienced an increase from 2007 until 2009. This two-year period refers to the most complicated time window of the worst financial crisis in 80 years (i.e. the GFC of 2007–08), as detailed in Table 9. The largest annual mean value of each country is reached only in the worst two years of the GFC of 2007–08, which results are in accordance with the related literature. Moreover, Table 10 summarises the statistics of the similarity factor for each country throughout the sample period.

TABLE 9 Similarity factor $SF_{i,j}$ smoothed as annual mean

Year	Canada	France	Germany	Italy	Japan	United Kingdom	United States
2003	0.518	0.584	0.548	0.567	0.473	0.558	0.542
2004	0.557	0.610	0.588	0.598	0.506	0.587	0.562
2005	0.566	0.614	0.600	0.606	0.520	0.604	0.574
2006	0.552	0.602	0.590	0.601	0.507	0.592	0.551
2007	0.575	0.629	0.601	0.599	0.537	0.610	0.573
2008	0.578	0.658[a]	0.647[a]	0.642[a]	0.557	0.631[a]	0.576
2009	0.589[a]	0.642	0.643	0.637	0.558[a]	0.624	0.597[a]
2010	0.554	0.625	0.620	0.605	0.527	0.605	0.584
2011	0.573	0.630	0.612	0.601	0.550	0.607	0.593
2012	0.549	0.578	0.592	0.549	0.521	0.569	0.554
2013	0.534	0.588	0.583	0.552	0.501	0.570	0.558
2014	0.522	0.593	0.591	0.559	0.522	0.556	0.543
2015	0.573	0.633	0.608	0.613	0.550	0.602	0.580
2016	0.535	0.617	0.617	0.560	0.524	0.599	0.544
2017	0.545	0.569	0.556	0.557	0.524	0.548	0.522

[a] It indicates the largest annual value of each country in the sample period.

France is the country with the largest mean, median, kurtosis, variance and, consequently, also the largest standard deviation and standard error among all G7 countries; however, France reports the smallest skewness value. Conversely, Japan is the country with the smallest mean, median, kurtosis, variance and, consequently, also the smallest standard deviation and standard error among all G7 countries; the largest skewness value is produced by Japan.

Considering a hypothetical equal level of return observed in all G7 countries, *ceteris paribus*, and also hypothetically considering the sample of the present study as the universe of investment alternatives which a typical risk-averse financial investor would be interested in, then such an investor seeking for portfolio diversification might potentially draw insightful conclusions through a visual inspection on Figure 10 and reading Tables 9 and 10, such as but not limited to: (i) combinations between only European countries would contribute to a relevant decrease in the level of portfolio diversification, resulting in a less diversified portfolio; (ii) the combination between the United States and Canada decreases the level of diversification of the portfolio; (iii) Japan is the best alternative to increase the level of diversification of the portfolio; (iv) a portfolio including either Canada or the United States combined with countries from the Eurozone (i.e. France, Germany, and Italy) would slightly increase the level of diversification compared to a portfolio composed by only European countries, resulting in a more diversified portfolio; and (v) including the combination either between the United States and Canada or the United States and the United Kingdom, to any portfolio composed by European countries, would virtually not affect its level of diversification, by either slightly increasing or decreasing it, respectively.

Therefore, the findings and insights provided through the similarity analysis described in this subsection might be useful for investment portfolio management purposes. Such results may shed light on and reveal the precise level of similarity between each pair of investment alternatives, at each point in time as well as comparing them in a pairwise fashion.

TABLE 10 Similarity factor $SF_{i,j}$ statistics based on daily frequency data

Statistic	Canada	France	Germany	Italy	Japan	United Kingdom	United States
Mean	0.557	0.613	0.602	0.591	0.528	0.593	0.565
Median	0.579	0.644	0.630	0.616	0.554	0.619	0.588
Variance	0.065	0.065	0.065	0.065	0.063	0.065	0.065
SD	0.255	0.256	0.255	0.254	0.250	0.255	0.255
SE	0.002	0.002	0.002	0.002	0.002	0.002	0.002
Skewness	−0.304	−0.497	−0.446	−0.396	−0.277	−0.434	−0.324
Kurtosis	−0.814	−0.642	−0.679	−0.717	−0.884	−0.684	−0.776
Smallest	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Largest	0.999	1.000	1.000	1.000	0.992	0.999	0.999
Range	0.999	1.000	1.000	1.000	0.992	0.999	0.999
Obs	24,324	24,324	24,324	24,324	24,324	24,324	24,324

Notes: "SD" means standard deviation, "SE" refers to standard error, and "Obs" reflects the number of observations in the sample.

4.2.4 | Cluster analysis outputs of G7

An empirical illustrative example of how the proposed cluster analysis works is depicted in Figure 11 by slicing the spheres that represent each of the G7 members in $\mathbb{R}^3$. It is possible to visualise that, for instance, Italy (green sphere) is more similar to France (blue sphere), and vice versa, in comparison with Japan (the smallest sphere in yellow in the centre). This type of difference is solely based on their respective input data, which impacts the value reflected in each of the pairwise similarity factors (detailed in Section 3.3) and, consequently, in the overall clustering measure $\mathscr{R}_t$ (detailed in Section 3.4).

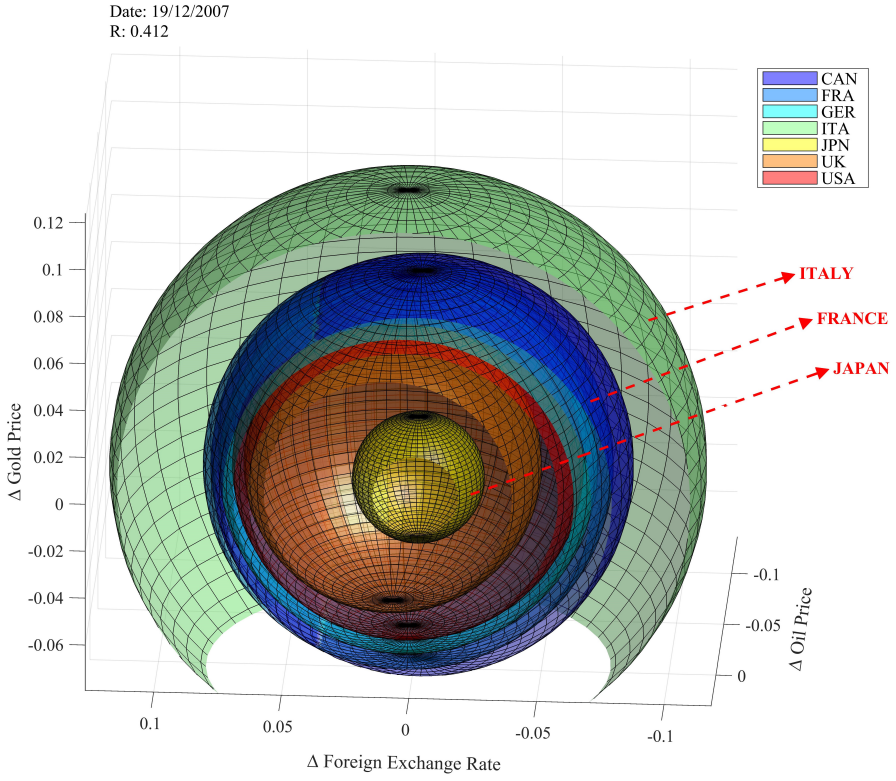

FIGURE 11 Sliced spheres in $\mathbb{R}^3$ representing G7 members for illustrative purposes. Italy, France, and Japan are highlighted as examples

Also for illustrative purposes, the daily progress of the cluster analysis performed involving G7 stock markets and macroeconomic variables is illustrated in Figure 12, in which is depicted one example per year of the last trading day[‡‡] of each of the 15 years in the sample. Through a data visualisation of such a very limited number of data points

[‡‡] It refers to the last trading day of each year in the sample, in which all G7 stock markets are operating in the same date.

(i.e. only 15 cases out of 4,054 trading days) it is possible to realise that this group of seven objects (i.e. countries/members/economies/stock markets) potentially forms only one cluster, although each object is distinct between its peers as well as compared to itself in different points in time.

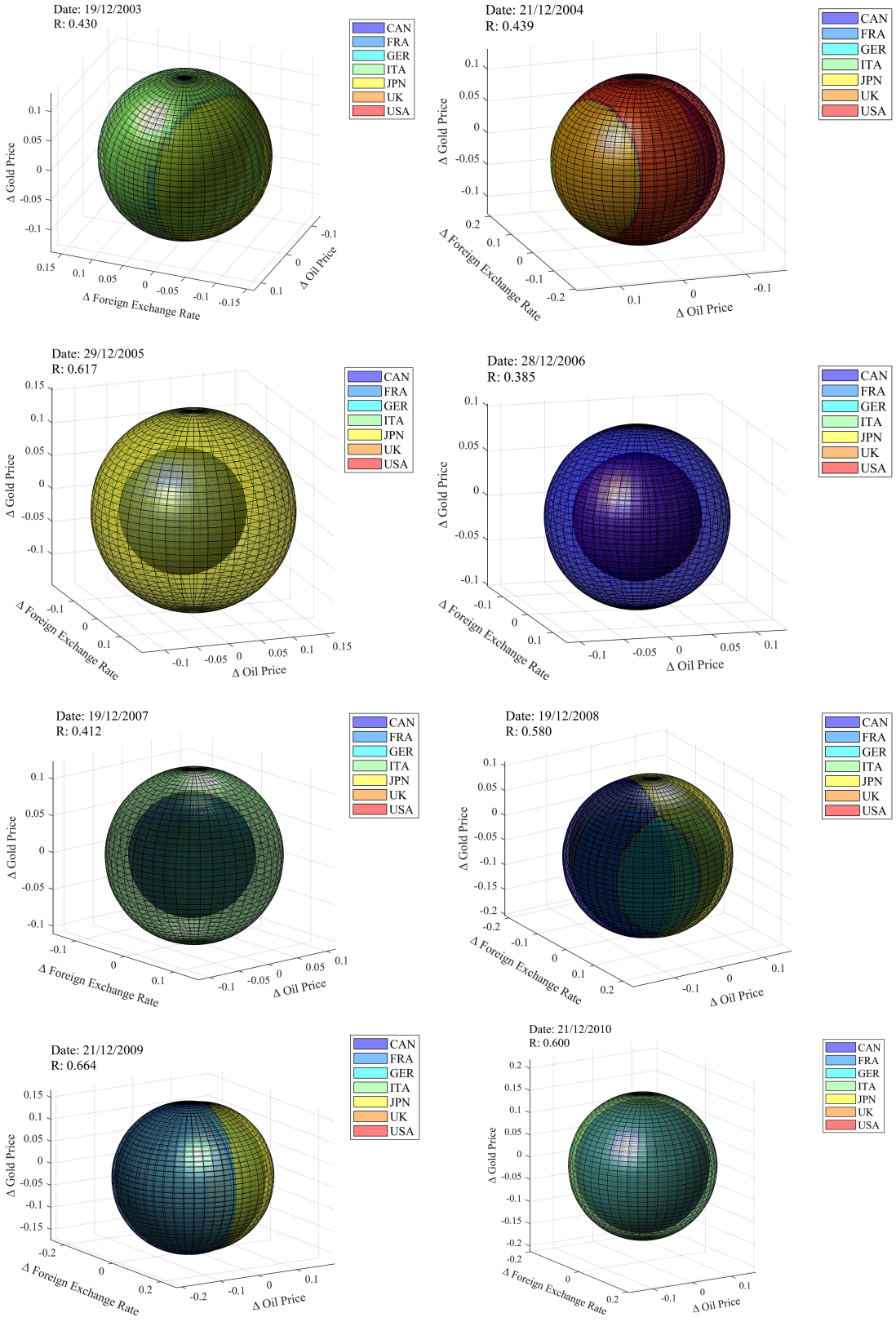

FIGURE 12 (Continues)

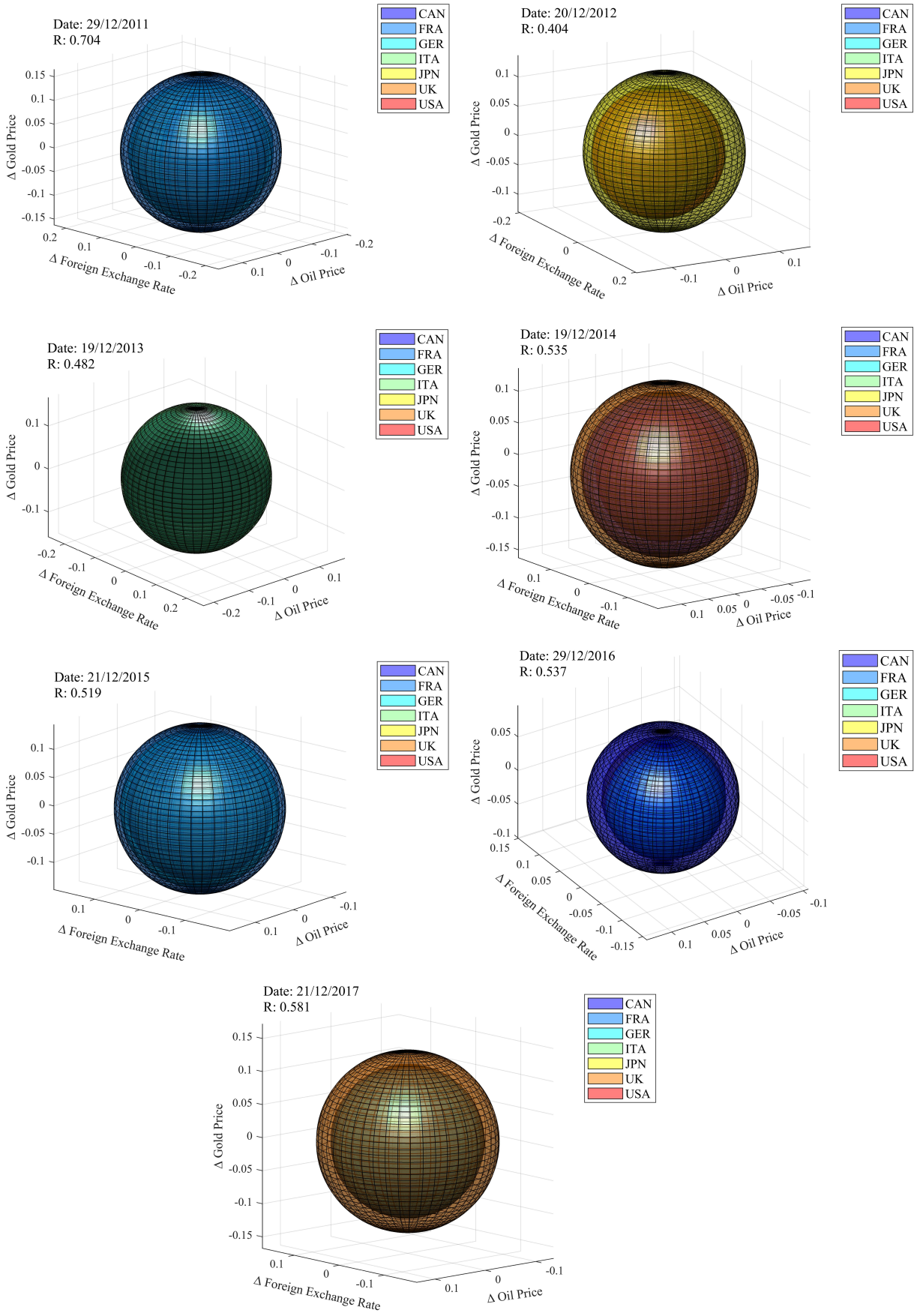

FIGURE 12 Daily progress of 15 cases (i.e. trading days/dates) of the cluster analysis performed involving G7 stock markets and macroeconomic variables

As further explored in the next subsection, based on the input variables chosen in the present study, the SCA method reveals that this dataset is composed by a single cluster throughout the whole sample period.

Number of clusters

Differently from the Fisher's Iris dataset, in the analysis performed using the G7 economic/financial dataset, the number of clusters "naturally" (i.e. according to the data-driven linkage criterion adopted, as detailed in Section 3.4.2) revealed by the SCA method is constantly one through the whole sample period. One possible reason behind this particular G7 dataset has revealed only one cluster might be related to the fact that in the present study two out of the four input variables refer to global variables (i.e. oil and gold price), which are the same for virtually[§§] all countries in the sample. Another possible justification might be related to the fact that G7 stock markets reflect a highly integrated global system/network, and the respective variables included in the present analysis experience a similar behaviour through the time span.

It is worth noticing, however, that the focus of the present paper is on introducing the SCA method and detailing its inner workings. Therefore, it is out of scope of the present study exploring further economic, social, political, or any other motivations that might have contributed to, or even caused, a larger or smaller similarity level, neither between any pair of objects (i.e. countries/economies/stock markets) nor regarding the whole sample at once (for instance, detailing economic/social/political explanations to justify the number of clusters revealed in this empirical analysis).

Clustering patterns through time

In order to highlight further interesting time series patterns through a data visualisation approach, the overall daily mean of all G7 similarity factors is then smoothed into a quarterly $\mathscr{R}_t$ time series, as shown in Figure 13, which also highlights (in gray) periods of economic recession[***54] experienced by at least one G7 economy at each point in time. In general, during recessions the $\mathscr{R}_t$ tends to rise and, more specifically, there is an increase in the $\mathscr{R}_t$ from the 2004-Q4 until 2009-Q4, which consists of a period of major negative events of the GFC of 2007–08.

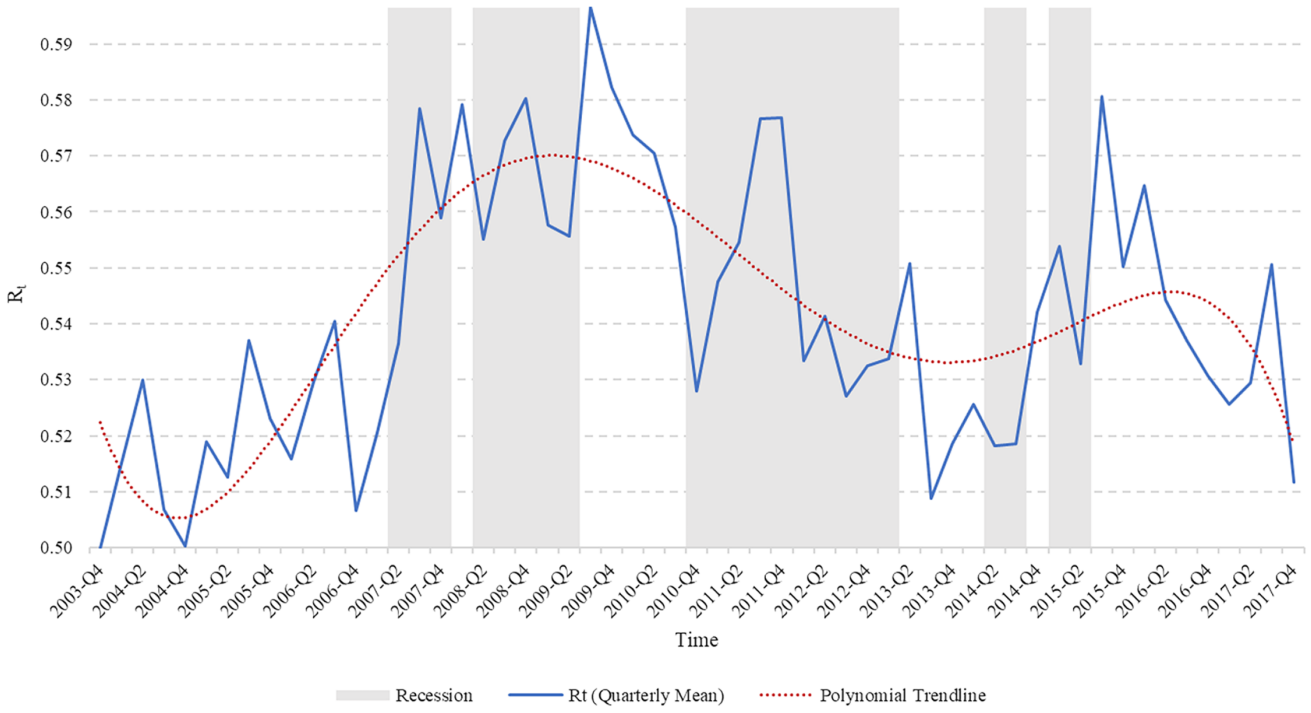

FIGURE 13 Quarterly time series of the overall clustering measure $\mathscr{R}_t$ through the sample period, depicted along with recession periods (in grey)

[§§] A trading day synchronisation is performed on only one country (i.e. Japan), as detailed in Section 4.2.1.

[***] The term "economic recession" used in this paper merely means that a particular economy experienced at least two consecutive quarters of decline in its real GDP growth rate. Therefore, this simplistic definition, frequently found in the literature, is adopted in this paper only to highlight crisis periods/times of high volatility levels involving G7 economies and, therefore, it does not refer to the respective definition adopted by the National Bureau of Economic Research (NBER).

As detailed in Figure 13, there is an increase of 14% in the $\mathscr{R}_t$ in a period of only three quarters (from 2006-Q4 to 2007-Q3), which coincides with the beginning of the GFC of 2007–08. Moreover, from the 2006-Q4 to the peak quarterly mean of the $\mathscr{R}_t$ (i.e. 2009-Q3), this measure experiences a major increase of almost 18% in a period of 11 quarters. From 2010-Q4 to 2011-Q4 there is an increase of 9%, from 0.531 to 0.580, respectively. This rise in the $\mathscr{R}_t$ coincides with important moments of the European sovereign debt crisis of 2011–2013, along with the aftermath of the GFC of 2007–08.

The $\mathscr{R}_t$ time series is then further smoothed into a yearly frequency in order to investigate if the data reveal any additional potential insights and/or more historical trends. The annual $\mathscr{R}_t$ data is then detailed in Table 11, in which there is the confirmation of the patterns and trends already informed by the respective quarterly data. Once more, it is possible to visualise a relevant increase in the highly turbulent years of 2008 and 2009 for the world economy, which specially hit developed economies and respective stock markets, reaching its annual peak in the respective years and experiencing an increase of 14% compared with 2003.

TABLE 11 The $\mathscr{R}_t$ measure smoothed into annual frequency

Year	$\mathscr{R}_t$	YoY	Yo2003
2003	0.503	n/a	n/a
2004	0.516	2%	2%
2005	0.526	2%	5%
2006	0.526	0%	5%
2007	0.551	5%[a]	10%
2008	0.575[a]	4%	14%[a]
2009	0.575[a]	0%	14%[a]
2010	0.560	−3%	11%
2011	0.567	1%	13%
2012	0.537	−5%	7%
2013	0.532	−1%	6%
2014	0.530	−1%	5%
2015	0.558	5%	11%
2016	0.547	−2%	9%
2017	0.533	−3%	6%

Notes: *n/a* stands for not applicable.
[a]It indicates the largest value(s) of each column.

The chart in Figure 14 reflects the $\mathscr{R}_t$ values of Table 11 along with the G7 averaged absolute stock market return, mentioning important negative events experienced during the GFC of 2007–08 and the European sovereign debt crisis of 2011–2013. It is possible to realise that there is a potential positive relationship between major negative events and an increase in the overall similarity level of the whole sample, which is in accordance with the related existing literature.

In summary, by interpreting these outputs one might conclude that the similarity level between all G7 stock markets experienced a relevant increase from 2003 until the end of the year of 2009, possibly, among other potential factors, as a consequence of major negative events and media news closely related to two relevant crises, namely the GFC of 2007–08 and the European sovereign debt crisis of 2011–2013.

It is worth mentioning that one of the main objectives of the proposed method is to measure and compare, as impartially as possible, the level of similarity between every pair of objects in a dataset. Therefore, regarding the specific empirical case used in this paper, it is out of the scope of the proposed method judging input variables either as beneficial or detrimental for the overall performance of a particular investment portfolio. As the proposed method follows a data-driven process, such a judgment must be made only by the investor/analyst/decision-maker, who would merely benefit from impartial outputs and potential insights generated by the proposed SCA method.

FIGURE 14 Annual time series of the overall clustering measure $\mathscr{R}_t$ and the G7 absolute stock market return through the sample period, being highlighted a few major negative events

Notes. The vertical axis on the left-hand side reflects $\mathscr{R}_t$ values while the vertical axis on the right-hand side refers to absolute stock market returns

5 | CONCLUSION AND FUTURE RESEARCH

In this paper an explainable novel data-driven unsupervised learning method is proposed, termed as the SCA method, which consists of a nonparametric and deterministic hard cluster analysis, derived from a novel similarity measure (i.e. similarity factor). Firstly, in order to test and highlight conceptual benefits and caveats of the proposed method against established competing alternatives, the most popular dataset in the pattern recognition literature (i.e. Fisher's Iris dataset) is used as input data to the SCA, *k*-means, *k*-medians, and *k*-medoid methods.

The results show a slightly superior performance of the SCA method regarding the allocation of sample objects to each cluster. However, much more importantly than such a virtually inexistent superior performance, it is the fact that such similar results are achieved through a remarkably distinct clustering approach. In the case of the established competing clustering methods, the number of clusters in the dataset is set arbitrarily by the analyst (i.e. $k = 2$), while in the case of the SCA method the two clusters in the dataset is revealed in a data-driven manner and, therefore, through a much less arbitrary clustering allocation process.

In the second empirical analysis, although a number of potential limitations and caveats may apply, insightful results based on real-world economic data indicate that G7 stock markets seem to be more similar between themselves in times of turmoil, providing evidence of a more compact clustering structure of the G7 stock markets/economies according to market circumstances (e.g. financial crisis). Such empirical results may have relevant implications regarding necessary portfolio diversification adjustments according to market conditions, being the conclusions derived from these results potentially useful to provide further guidance to investors and policymakers, mainly during turbulent periods.

A suggestion of future research could consist of testing the explanatory power of the measures introduced in the present study (i.e. $SF_{i,j|t}$ and $\mathscr{R}_t$) to explain past stock market returns and forecasting respective future values. Furthermore, it is worth noting that the SCA method may potentially be applied to a wide range of fields of knowledge and domain-specific questions - as diverse as astronomy, biology, business, engineering, geography, geopolitics, medicine, sports, among many more, in order to compare similarity levels between sample objects and, therefore, not limited to applications on financial and economic data.

Moreover, it would be interesting to perform the SCA method using as input data additional popular publicly available machine learning datasets (e.g. census income dataset, car evaluation data set, heart disease data set) and also comparing its performance against competing density clustering methods (e.g. DBSCAN[†††], OPTICS[‡‡‡]). Needless to mention that insights based on outputs to be obtained by the application of the SCA method on distinct datasets remain an empirical question. Finally, the development of an additional algorithm programmed in an open source programming language (e.g. Python) to perform all steps of the SCA method is an ongoing project.

ACKNOWLEDGEMENTS

This paper benefited from stimulating conversations with Alfred Hero, Andrew Harvey, Philip Arestis, Ludovic Cesbron, Andrew Harrison, Xiao-Li Meng, Christian Hennig, Mark Gross, Alexei Kovalev, Mardi Dungey (*in memoriam*), three anonymous referees, and the journal editor, Giampiero Accardo, for their careful reading and suggestions. All errors remain exclusively the author's responsibility. The present research was supported by the Coordination for the Improvement of Higher Education Personnel of Brazil (CAPES) and Cambridge Commonwealth, European & International Trust (under grant BEX 2220/15-6).

PEER REVIEW INFORMATION

Engineering Reports thanks the anonymous reviewers for their contribution to the peer review of this work.

DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available in the Appendix S1 of this article.

CONFLICT OF INTEREST

The author declares no competing interests that might be perceived to influence the results and/or discussion reported in this study.

ORCID

Michel Ferreira Cardia Haddad https://orcid.org/0000-0002-0978-9525

REFERENCES

1. Campello M, Graham JR, Harvey CR. The real effects of financial constraints: evidence from a financial crisis. *J Financ Econ*. 2010;97:470-487.
2. Foster JB, Magdoff F. *The Great Financial Crisis: Causes and Consequences*. New York, NYU Press; 2009.
3. Hamilton JD. Causes and Consequences of the Oil Shock of 2007-08. *National Bureau of Economic Research*. 2009. https://www.nber.org/papers/w15002.
4. Johnson S, Boone P, Breach A, Friedman E. Corporate governance in the Asian financial crisis. *J Financ Econ*. 2000;58:141-186.
5. Andersen TG, Bollerslev T, Diebold FX, Ebens H. The distribution of realized stock return volatility. *J Financ Econ*. 2001;61:43-76.
6. Schwert GW. Why does stock market volatility change over time? *J Financ*. 1989;44:1115-1153.
7. Forbes KJ, Rigobon R. No contagion, only interdependence: measuring stock market comovements. *J Financ*. 2002;57:2223-2261.
8. King MA, Wadhwani S. Transmission of volatility between stock markets. *Rev Financ Stud*. 1990;3:5-33.
9. Arestis P, Karakitsos E. *Financial Stability in the Aftermath of the 'great Recession'*. Basingstoke, UK: Springer; 2013.
10. Reinhart CM, Rogoff KS. *This Time is Different: Eight Centuries of Financial Folly*. Princeton, Princeton University Press; 2009.
11. Shiller RJ. *The Subprime Solution: how today's Global Financial Crisis Happened, and What to Do about it*. Princeton, USA: Princeton University Press; 2012.
12. Cappiello L, Engle RF, Sheppard K. Asymmetric dynamics in the correlations of global equity and bond returns. *J Financ Econom*. 2006;4:537-572.
13. Dungey M, Fry R, González-Hermosillo B, Martin VL. Empirical modelling of contagion: a review of methodologies. *Quant Financ*. 2005;5:9-24.
14. Ramchand L, Susmel R. Volatility and cross correlation across major stock markets. *J Empir Financ*. 1998;5:397-416.
15. Statman M. How many stocks make a diversified portfolio? *J Financ Quant Anal*. 1987;22:353-363.
16. Haddad MFC. Sphere-sphere intersection for investment portfolio diversification-a new data-driven cluster analysis. *MethodsX*. 2019;6:1261-1278. https://www.sciencedirect.com/science/article/pii/S2215016119301451.
17. Everitt BS. Unresolved problems in cluster analysis. *Biometrics*. 1979;35:169-181.
18. Jain AK. Data clustering: 50 years beyond K-means. *Pattern Recogn Lett*. 2010;31:651-666.
19. Hennig C, Meila M, Murtagh F, Rocci R. *Handbook of Cluster Analysis*. Boca Raton, CRC Press; 2015.

[†††] DBSCAN refers to the density-based spatial clustering of applications with noise method.
[‡‡‡] OPTICS refers to the ordering points to identify the clustering structure method.

20. Longin F, Solnik B. Is the correlation in international equity returns constant: 1960–1990? *J Int Money Financ*. 1995;14:3-26.
21. Solnik B, Boucrelle C, Le Fur Y. International market correlation and volatility. *Financ Anal J*. 1996;52:17-34.
22. Keogh E, Lin J. Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowl Inform Syst*. 2005;8:154-177.
23. Elton EJ, Gruber MJ. Homogeneous groups and the testing of economic hypotheses. *J Financ Quant Anal*. 1970;4:581-602.
24. Elton EJ, Gruber MJ. Improved forecasting through the design of homogeneous groups. *J Bus*. 1971;44:432-450.
25. Panton DB, Lessig VP, Joy OM. Comovement of international equity markets: a taxonomic approach. *J Financ Quant Anal*. 1976;11:415-432.
26. Mantegna RN. Hierarchical structure in financial markets. *Eur Phys J B Condens Matter Complex Syst*. 1999;11:193-197.
27. Bonanno G, Lillo F, Mantegna RN. High-frequency cross-correlation in a set of stocks. *Quant Financ*. 2001;1:96-104.
28. Tola V, Lillo F, Gallegati M, Mantegna RN. Cluster analysis for portfolio optimization. *J Econ Dyn Control*. 2008;32:235-258.
29. Pai GV, Michel T. Evolutionary optimization of constrained $ k $-means clustered assets for diversification in small portfolios. *IEEE Trans Evol Comput*. 2009;13:1030-1053.
30. Nanda S, Mahanty B, Tiwari M. Clustering Indian stock market data for portfolio management. *Expert Syst Appl*. 2010;37:8793-8798.
31. Tsay RS. *Analysis of Financial Time Series*. New York, NY: Wiley; 2005:2005.
32. Sommerville D. *An Introduction to the Geometry of n Dimensions*. London, UK: Methuen & Co. Ltd.; 1958:1929.
33. Court N. Four intersecting spheres. *Amer Math Month*. 1960;67:241-248.
34. Weisstein EW. Sphere-sphere intersection. 2007
35. Devasena CL, Sumathi T, Gomathi V, Hemalatha M. Effectiveness evaluation of rule based classifiers for the classification of iris data set. *Bonfring Int J Man Mach Interf*. 2011;1:05-09.
36. Moon K, Hero A. Multivariate f-divergence estimation with confidence. *Adv Neural Inform Process Syst*. 2014;27:2420-2428.
37. Ahmad A, Dey L. A k-mean clustering algorithm for mixed numeric and categorical data. *Data Knowl Eng*. 2007;63:503-527.
38. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Eugen*. 1936;7:179-188.
39. Bache K Lichman M. UCI machine learning repository. 2013.
40. Fred AL, Jain AK. Data clustering using evidence accumulation. Paper presented at: Object Recognition Supported by User Interaction for Service Robots; IEEE; 2002; Quebec City, Canada: 276–280.
41. Krzanowski WJ, Lai Y. A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*. 1988;44:23-34.
42. Timmis J, Neal M, Hunt J. An artificial immune system for data analysis. *Biosystems*. 2000;55:143-150.
43. Arthur D, Vassilvitskii S. *K-Means++: The Advantages of Careful Seeding*. 2006. http://ilpubs.stanford.edu:8090/778/.
44. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. *J Am Med Inform Assoc*. 2005;12:296-298.
45. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *J Mach Learn Technol*. 2011;2:37-63.
46. Olson DL, Delen D. *Advanced Data Mining Techniques*. Heidelberg, Germany: Springer Science & Business Media; 2008.
47. Deb AB, Dey L. Outlier detection and removal algorithm in K-means and hierarchical clustering. *World J Comput Appl Technol*. 2017;5:24-29.
48. Honaker J, King G, Blackwell M. Amelia II: a program for missing data. *J Stat Softw*. 2011;45:1-47.
49. Lin W-L, Engle RF, Ito T. Do bulls and bears move across borders? International transmission of stock returns and volatility. *Rev Financ Stud*. 1994;7:507-538.
50. Maslov S. Measures of globalization based on cross-correlations of world financial indices. *Phys A Stat Mech Appl*. 2001;301:397-406.
51. Mian GM, Adam CM. Does more market-wide information originate while an exchange is open: some anomalous evidence from the ASX. *Aust J Manage*. 2000;25:339-352.
52. Sandoval Jr L. To lag or not to lag? How to compare indices of stock markets that operate on different times. *Phys A Stat Mech Appl*. 2014;403:227-243.
53. Ghysels E, Harvey AC, Renault E. 5 stochastic volatility. *Handb Stat*. 1996;14:119-191.
54. Leamer EE. What's a Recession, Anyway?. *National Bureau of Economic Research*. 2008. https://www.nber.org/papers/w14221.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Haddad MFC. Harnessing the power of intersection for pattern recognition: a novel unsupervised learning method and its application to financial engineering. *Engineering Reports*. 2021;3:e12329. https://doi.org/10.1002/eng2.12329