Soft Computing (2021) 25:7813–7823
https://doi.org/10.1007/s00500-021-05639-y

FOCUS

A fuzzy method for evaluating similar behavior between assets

Soheyla Mirshahi[1] · **Vilém Novák**[1]

Published online: 9 March 2021

Abstract
In this paper, we propose a fuzzy method to investigate the interconnection between equity markets in the form of similar behavior. It has been proved before that the trend cycle of time series can be well estimated using the fuzzy transform. In the suggested method, first, we approximate the local behavior of stocks as a sequence of their trend cycles. Then we measure the distance between these local trend cycles conducting similar practices between different assets. Two experiments are performed to demonstrate the advantages of the suggested method. This method is easy to calculate, well interpretable, and in addition to statistical co-relation, the measure can assist investors in gaining more intuition about the behavior of their assets.

Keywords Stock interconnection · Stock markets similarity · Portfolio selection · Fuzzy transform

1 Introduction

Analyzing the interconnection among different assets has been the subject of interest of many researchers. Today, large datasets of multivariate time series are available in many fields such as business, finance, and economics, providing comprehensive data to mine some information deeply.

One of the critical applications of data mining in time series (see Mining 2006; Keogh and Kasetty 2003; Fu 2011; Liao 2005; Han et al. 2011) is mining the data in stock markets. Assessing time series similarity, i.e., the degree to which a given time series resembles another one is a core to many mining, retrieval, clustering, and classification tasks (Serra and Arcos 2014). There is no straightforward approach, known as the best measure for assessing the similarities in time series. Surprisingly, many simple tools like euclidean distance can outperform the most complicated methods (Serra and Arcos 2014). Wang et al. in Wang et al. (2013) perform an extensive comparison between nine measurements across 38 data sets from various scientific domains. One of their findings is that the euclidean distance remains an entirely accurate, robust, simple, and efficient way of measuring the similarity between two time series in general. However, in finance, the principle step in many crucial applications, including understanding the interconnection among financial time series (e.g., stocks, bonds), remains to be the dynamics of their correlation (Martens and Poon 2001). Due to the fact that diversification, which conveys investing in a variety of assets, is a key to reduce the risk of a chosen financial portfolio, and for that matter, correlation is the principal indicator for the goodness of it Statman and Scheid (2008). Thus it was the center of attention for many researchers see(Hamao et al. 1990; Hilliard 1979; Wu and Su 1998; Cha and Oh 2000; Bekaert and Harvey 2003).

For instance, in Wang et al. (2010), the cross-correlations between the Chinese A-share and B-share market are examined. Kulman et al. in Kullmann et al. (2002) use correlation as the function of the time shift between pairs of stock return time series and investigate the time-dependent cross-correlations between them. Bernanke, in 2016, analyzed the relation between oil prices and stock markets by correlation (Bernanke 2016). Nevertheless, there are specific problems when using correlation in stock analysis alone. Firstly, some researchers report that using correlation to analyze interrelations among international stock markets is low on average and differ rather fiercely across countries see(Roll 1992) and Eun and Shim (1989).

Secondly, similar to euclidean distance measurement, the Pearson correlation is also very sensitive to outliers (Devlin

Communicated by Vladik Kreinovich.

✉ Soheyla Mirshahi
soheyla.mirshahi@osu.cz

Vilém Novák
vilem.novak@osu.cz

[1] Institute for Research and Applications of Fuzzy Modeling, NSC IT4Innovations, University of Ostrava, Ostrava, Czech Republic

et al. 1975). Note that stock markets have some specific properties. For instance, stocks react to a lot of exogenous factors such as news (see, e.g., Chan (2003)); thus, the presence of outliers in them is inevitable. Analyzing the similarity for the price values of stock markets is critical; however, in practice, investors tend to maximize the overall return of their portfolio; therefore, the interaction between the return of each investment is also vital. Therefore, developing a similarity method that is capable of reacting to the nature of stock markets for both their price values as well their returns seems essential.

A very effective technique for the representation and consequently analysis of time series is the fuzzy transform. Using it, we can extract trend cycle (a low-frequency trend component) of the time series with high fidelity. The fuzzy transform provides not only the computed trend cycle but also its analytic formula (cf. Novák et al. (2014), Novák et al. (2010)). In this paper, using fuzzy transform, we first assign to each financial time series an adjoint one that consists of its local trend cycle. Then we measure the distance between these approximate time series by a suggested formula.

There are several reasons to employ our fuzzy estimation of the trend cycle for analyzing the interaction between stocks: Firstly, the trend cycle in stocks tends to smoothen the price value and describes the behavior of the market concerning the changes in price values. Thus, it is more intuitive for experts than price values themself. It has been proven that we can successfully reach this goal using the fuzzy transform. Secondly, stock markets can be boisterous with outliers. Consequently, assessing similarities among them based on actual price values without any preprocessing can lead to unrealistic results. Using our method, we can easily "wipe out" the outliers without harming the essential characteristics of the time series. Finally, Our method is flexible and can answer the question of how we can find stocks that behave similarly at zero-lag or shifted-lags and in various time slots. For instance, experts can measure the similarity between stocks that behave similarly in a short to long term (e.g., one to several weeks) at the same time or with delay.

This paper aims to provide a mathematical method that is powerful for measuring similarity among stocks, not sensitive to outliers, and can detect the stocks which behave similarly at the same time moment or with lag(s). It extends our previous paper (Mirshahi and Novák 2020). The suggested method can be considered as a powerful tool complementary to Pearson correlation in analyzing the relationship between assets.

The structure of this paper is as follows. After Introduction, we describe the preliminaries of our method in Sect. 2. Section 3 is dedicated to describing the suggested method and its illustration together with the evaluation of the results.

2 Preliminaries

2.1 Time series decomposition

Our techniques stem from the following characterization of a time series. It is understood as a stochastic process (see, e.g., Anděl (1976), Hamilton (1994)) $X : \mathbb{T} \times \Omega \to \mathbb{R}$ where Ω is a set of elementary random events and $\mathbb{T} = \{0, \ldots, p\} \subset \mathbb{N}$ is a finite set of numbers interpreted as time moments. Since financial time series typically posses no seasonality, we assume that they can be decomposed into components as follows:

$$X(t, \omega) = TC(t) + R(t, \omega), \qquad t \in \mathbb{T}, \tag{1}$$

where $TC(t) = Tr(t) + C(t)$ called *trend cycle* and R is a random *noise*, i.e., a sequence of (possibly independent) random variables $R(t)$ such that for each $t \in \mathbb{T}$, the $R(t)$ has zero mean and finite variance.

2.2 Fuzzy transform

Fuzzy transform (F-transform) is the fundamental theoretical tool for the suggested similarity measurement. Because of the lack of space, we will only briefly outline the main principles of the F-transform and refer the reader to the extensive literature, e.g., Novák et al. (2016), Novák et al. (2014) and many others.

The F-transform is a procedure applied, in general, to a bounded real continuous function $f : [a, b] \to [c, d]$ where $a, b, c, d \in \mathbb{R}$. It is based on the concept of a *fuzzy partition* that is a set $\mathcal{A} = \{A_0, \ldots, A_n\}$, $n \geq 2$, of fuzzy sets fulfilling special axioms. The fuzzy sets are defined over nodes $a = c_0, \ldots, c_n = b$ in such a way that for each $k = 0, \ldots, n$, $A(c_k) = 1$ and $\text{supp}(A_k) = [c_{k-1}, c_{k+1}]$[1]. The nodes are usually (but not necessarily) uniformly distributed, i.e., $c_{k+1} = c_k + h$ where $h > 0$ is a given value. To emphasize that the fuzzy partition is formed using the distance h, we will write $\mathcal{A}_h$. The fuzzy sets $A_k \in \mathcal{A}$ are often called *basic functions*.

The F-transform has two phases: direct and inverse. The *direct* F-transform assigns to each $A_k \in \mathcal{A}_h$ a component $F_k[f|\mathcal{A}_h]$. We distinguish *zero degree* F-transform whose components $F_k^0[f|\mathcal{A}]$ are numbers, and *first degree*[2] F-transform whose components have the form $F_k^1[f|\mathcal{A}_h](x) = \beta_k^0[f] + \beta_k^1[f](x - c_k)$. The coefficient $\beta_k^1[f]$ provides estimation of an average value of the tangent (slope) of f over the area characterized by the fuzzy set $A_k \in \mathcal{A}_h$.

[1] Of course, certain formal requirements must be fulfilled. They are omitted here and can be found in the cited literature.

[2] In general, higher degree F-transform.

From the direct F-transform of f

$$\mathbf{F}[f|\mathcal{A}_h] = (F_0[f|\mathcal{A}_h], \ldots, F_n[f|\mathcal{A}_h])$$

we can form a function $\mathbf{I}[f|\mathcal{A}_h] : [a,b] \to [c,d]$ using the formula

$$\mathbf{I}[f|\mathcal{A}_h](x) = \sum_{k=0}^{n} (F_k[f|\mathcal{A}_h] \cdot A_k(x)), \qquad x \in [a,b]. \quad (2)$$

The function (2) is called the *inverse F-transform* of f w.r.t. the fuzzy partition $\mathcal{A}_h$, and it approximates the original function f. It can be proved that this approximation is universal.

2.3 Application of the F-transform to the analysis of time series

The application of the F-transform to the time series analysis is based on the following result (cf. Novák et al. (2014), Nguyen and Novák (2018)). Let us now assume (without loss of generality) that the time series (1) contains periodic subcomponents with frequencies $\lambda_1 < \cdots < \lambda_r$. These frequencies correspond to periodicities

$$T_1 > \cdots > T_r, \quad (3)$$

respectively (via the equality $T = 2\pi/\lambda$).

Theorem 1 *Let $\{X(t) \mid t \in \mathbb{T}\}$ be a realization of the time series (1) with the trend cycle TC. Let us assume that all subcomponents with frequencies λ lower than λ_q are contained in the trend cycle TC. If we construct a fuzzy partition $\mathcal{A}_h$ over the set of equidistant nodes with the distance $h = d\,T_q$ where $d \in \mathbb{N}$ and T_q is a periodicity corresponding to λ_q, then the corresponding inverse F-transform $\mathbf{I}[X|\mathcal{A}_h]$ of $X(t)$ gives the following estimation of the trend cycle:*

$$|\mathbf{I}[X|\mathcal{A}_h](t) - TC(t)| \leq 2\omega(h, TC) + D \quad (4)$$

for $t \in [c_1, c_{n-1}]$, where D is a certain small *number and $\omega(h, TC)$ is a modulus of continuity of TC w.r.t. h.*[3]

The precise form of D and the detailed proof of this theorem can be found in Novák et al. (2014), Nguyen and Novák (2015). It follows from this theorem that the F-transform makes it possible to filter out frequencies higher than a given threshold and also to reduce the noise R. Consequently, we have a tool for separation of the trend cycle or trend. Theorem 1 tells us how the distance between nodes of the fuzzy partition should be set. This choice enables us to detect trend cycles for different time frames of interest. Of course, the estimation depends on the course of TC, and it is the better the smaller is the modulus of continuity $\omega(h, TC)$ (which in case of the trend cycle or trend is a natural assumption). The periodicities (3) can be found using the classical technique of *periodogram* — see (Anděl 1976; Hamilton 1994).

Selection of T_q in Theorem 1 can be based on the following general OECD specification: *Trend (tendency)* is the component of a time series that represents variations of low frequency in a time series, the high and medium frequency fluctuations having been filtered out. *Trend cycle* is the component of the time series that represents variations of low frequency, the high frequency fluctuations having been filtered out. Hence, in the sequel for a given time series X we will work with estimation $\widetilde{TC}$ of its trend cycle given by

$$\widetilde{TC} = \mathbf{I}[X|\mathcal{A}_h] \quad (5)$$

for a suitable fuzzy partition $\mathcal{A}_h$ determined on the basis of Theorem 1.

3 The suggested method

3.1 Similarity between time series

In this section, we describe how our suggested method evaluates the pairwise similarity between stocks. Our main concern is to detect stock that behave similarly at lag-zero as well as at shifted-lag. We will measure similarity using a binary *fuzzy relation* on a given set Z, which is a function $S : Z \times Z \to [0,1]$. The following properties of S can be considered (for all $z, u, v \in Z$):

(i) $S(z,z) = 1$, (reflexivity)
(ii) $S(z,u) = S(u,z)$, (symmetry)
(iii) $S(z,u) \otimes S(u,v) \leq S(z,v)$ (transitivity)
where $\otimes : [0,1] \times [0,1] \to [0,1]$ is a certain t-norm (cf. Novák et al. (1999))[4]. In this paper, we will consider the *Łukasiewicz t-norm* defined by $a \otimes b = \max\{0, a+b-1\}$ for all $a, b \in [0,1]$.

A reflexive and symmetric fuzzy relation is called a *fuzzy symmetry*. If it is, moreover, transitive, then it is called a *fuzzy equality*.

The fuzzy symmetry S is *separated* if

$$S(z,u) = 1 \text{ iff } z = u$$

holds for all $z, u \in Z$.

[3] Modulus of continuity is in our case defined as $\omega(h, TC) = \max_{\substack{|x-y|<h \\ x,y \in [c_1, c_{n-1}]}} |TC(x) - TC(y)|$.

[4] A t-norm is a special operation that in fuzzy logic models logical conjunction.

Let us consider realizations of two time series $\{F_i(t) \mid t = 1, \ldots, n\}$, $i = 1, 2$ and let $S : \mathbb{R}^n \times \mathbb{R}^n \to [0, 1]$ be a fuzzy relation defined by

$$S(F_1, F_2) = 1 - \frac{1}{n} \sum_{t=1}^{n} \frac{|F_1(t) - F_2(t)|}{|F_1(t)| + |F_2(t)|}. \tag{6}$$

It is easy to show that $S(F_1, F_2) \in [0, 1]$.

Theorem 2 *The fuzzy relation S given in (6) is a separated fuzzy symmetry. Let F_1, F_2, F_3 be realizations of time series of the length n. If $|F_2(t)| \leq \min\{|F_1(t)|, |F_3(t)|\}$ for all $t = 1, \ldots, n$ then S is a fuzzy equality w.r.t. Łukasiewicz t-norm $\otimes$.*

Proof (a) The reflexivity $S(F_1, F_1) = 1$ is immediate. The symmetry follows from the properties of absolute value.

(b) Separateness: If $F_1 = F_2$ then $S(F_1, F_2) = 1$ by reflexivity. Conversely, let $S(F_1, F_2) = 1$. Then

$$\frac{|F_1(t) - F_2(t)|}{|F_1(t)| + |F_2(t)|} = 0$$

for all t, which holds only if $F_1 = F_2$.

(c) The transitivity requires $S(F_1, F_2) \otimes S(F_2, F_3) \leq S(F_1, F_3)$. This holds if

$$\frac{|F_1(t) - F_3(t)|}{|F_1(t)| + |F_3(t)|} \leq \frac{|F_1(t) - F_2(t)|}{|F_1(t)| + |F_2(t)|} + \frac{|F_2(t) - F_3(t)|}{|F_2(t)| + |F_3(t)|},$$

for $t = 1, \ldots, n$. This inequality is fulfilled if both $|F_2(t)| \leq |F_1(t)|$ as well as $|F_2(t)| \leq |F_3(t)|$ hold for all $t = 1, \ldots, n$. □

Definition 1 Let $X = \{X(t) \mid t = 1, \ldots, n\}$ and $Y = \{Y(t) \mid t = 1, \ldots, n\}$ be two time series of the length n and $\widetilde{TC}_X$ and $\widetilde{TC}_Y$ be estimations of trend cycles of X and Y respectively,[5] calculated using equation (5) for a suitable fuzzy partition $\mathcal{A}_h$. Then we define the similarity between these two time series as follows:

$$S(\widetilde{TC}_X(t) - \mathbf{E}(\widetilde{TC}_X), \widetilde{TC}_Y(t) - \mathbf{E}(\widetilde{TC}_Y)) = 1 - \frac{1}{n} \sum_{t=1}^{n} \frac{|\widetilde{TC}_X(t) - \mathbf{E}(\widetilde{TC}_X) - (\widetilde{TC}_Y(t) - \mathbf{E}(\widetilde{TC}_Y))|}{|\widetilde{TC}_X(t) - \mathbf{E}(\widetilde{TC}_X)| + |\widetilde{TC}_Y(t) - \mathbf{E}(\widetilde{TC}_Y)|}, \tag{7}$$

where $\mathbf{E}(\widetilde{TC}_X)$ and $\mathbf{E}(\widetilde{TC}_Y)$ are mean values (averages) of $\widetilde{TC}_X$ and $\widetilde{TC}_Y$, respectively.

It follows from Theorem 2 that the similarity (7) is a fuzzy symmetry, or sometimes even fuzzy equality. For simplicity, in the sequel we will write (7) simply as $S(X, Y)$.

[5] It is necessary to emphasize, that we can work with estimations $\widetilde{TC}_X$ and $\widetilde{TC}_Y$ of the trend cycle only, because we do not know the real ones.

Remark 1 If we take X and Y as simple linear functions $X = \{k_1 t + q_1 \mid t = 1, \ldots, n\}$, $Y = \{k_2 t + q_2 \mid t = 1, \ldots, n\}$ then, by simple computation, we obtain that the similarity $S(X, Y) = 1 - \frac{|k_1 - k_2|}{|k_1| + |k_2|}$. Hence, if these lines are (almost) parallel then $S(X, Y) \approx 1$.

Stock price, can be seen as a time series $X = \{X(t) \mid t = 1, \ldots, n\}$ where $X(t)$ is a closing price at time $t \in \{t = 1, \ldots, n\}$. For instance, let us consider closing price of a stock from Nasdaq INC,[6] from 05.10.2008 to 30.09.2018 (522 weeks). In order to estimate its local trend cycle, we first build a uniform fuzzy partition $\mathcal{A}_h$ such that the length of each basic function $A_2, \ldots, A_m \in \mathcal{A}_h$ is equal to a proper time slot. In our case, by setting the length $h \in \{2, 3\}$, we obtain the approximation of the trend cycle for one month. In other terms, the monthly behavior of this stock is our concern here. Figure 1 depicts the mentioned weekly stock and the fuzzy approximation of its local trend cycle. The first and the last components of F-transform are subject to big error (because the corresponding basic functions A_1 and A_m are incomplete. Regardless of this, it is clear that the F-transform approximates the local trend cycles of the stock successfully. As we mentioned before, stock markets react to many exogenous factors; thus, the presence of outliers is unavoidable. A red square in Fig. 1 shows one of these outliers for the mentioned stock. It can be seen that the F-transform has successfully wiped out the outlier while preserving the core behavior of the stock.

The similarity from Definition 1 can be used for measuring similarity for any number of stocks based on their local behavior.

In the next section, we will demonstrate how our suggested method works with a relatively large data set of stock prices in conjunction with its comparison to standard the euclidean distance. The goal is to demonstrate the performance of the method in comparison to one of the most known similarity measurements. Further, we demonstrate how our method allows us to asses the lead-lag relation among returns of different assets in complementary to statistical correlation analysis.

3.2 Illustration

Our first data set consists of a closing price of 92 stocks over 522 weeks obtained from Nasdaq INC.[7] An example of twenty stocks from the mentioned data set is depicted in Fig. 2, where the x-axis and y-axis represent price values in dollars and number of weeks, respectively. From this figure, it is clear that any decision about the similarity between time

[6] https://www.nasdaq.com/Second footnote

[7] https://www.nasdaq.com/Second footnote.

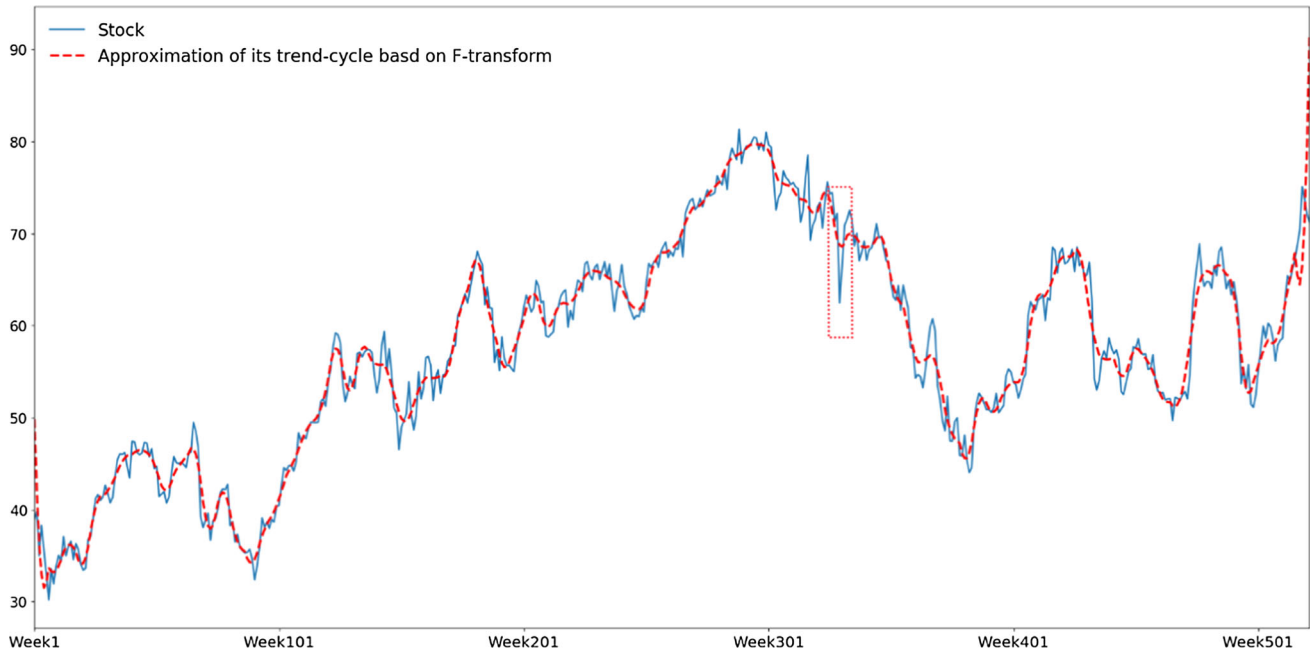

Fig. 1 A Stock and its TC approximation based on F-transform

series is impossible. Therefore it seems necessary to consider similarity between time series.

3.3 Evaluation

One possible way to evaluate the competency of any new similarity measurement (distance measurement) is to apply it to data clustering. The quality of clustering based on the new and current similarities can validate the competency of the suggested method (Morse and Patel 2007; Vlachos et al. 2006). Therefore, we will below apply clustering of time series and compare the behavior of our similarity with the euclidean one. However, let us emphasize that time series clustering is not the primary goal of this research since our focus is on discovering the most similar pairs of stocks available in the database. As we mentioned before, the euclidean distance is an accurate, robust, simple, and efficient way to measure the similarity between two time series and, surprisingly, can outperform most of the more complex approaches (see Serra and Arcos 2014, Wang (2001)). Therefore we will compare our method with the euclidean distance by means of the quality of hierarchical clustering on a dataset. Hierarchical clustering is a method of cluster analysis which attempts at building a hierarchy of similar groups in data (Kaufman and Rousseeuw 2009). In this case, one problem to consider is the optimal number of clusters in a dataset. Overall, none of the methods for determining the optimal numbers of clusters is flawless, and none of the suggested similarities are fully satisfactory. Hierarchical clustering does not reveal an adequate number of clusters and estimation of the proper number of clusters is rather intuitive. Hence, there is a fair amount of subjectivity in determination of separate clusters. Figures 3 and 4 demonstrate the dendrogram of hierarchical clustering of the 92 stocks based on the suggested and euclidean similarity, respectively. The proper number of clusters for both similarities is equal to six. In these figures, the 92 stocks are represented in the x-axis, and their distances are depicted on the y-axis accordingly. Since the stocks are from various industries, they have different scales, and in the case of the clustering with the euclidean distance, we will eliminate the different scaling by normalizing the data. Nevertheless, this step is not demanded by the suggested method since the scale does not influence it.

Red dashed squares in 3 and 4 represent the most similar stock pairs, determined according to each method. Interestingly, both methods selected the same stock pairs; (38 and 84) and (52 and 53) as the most similar stocks. However, the suggested method, primarily determines stock pair (38 and 84) as the most similar stocks, following by stock pair (52 and 53) while the euclidean method suggests otherwise. Figure 5 and 6 shows the behavior of theses stock pairs.

To measure the quality of clustering, we apply the Davies–Bouldin index, which is usually used in clustering. This measure evaluates intra-cluster similarity and inter-cluster differences (Davies and Bouldin 1979). Therefore, it can be a proper metric for clustering evaluation.

Table 1 demonstrates the Davies–Bouldin index for a different number of clusters based on the both similarities. Since the lower score indicates better quality of clustering, the results reveal that not only is our method reasonably compa-

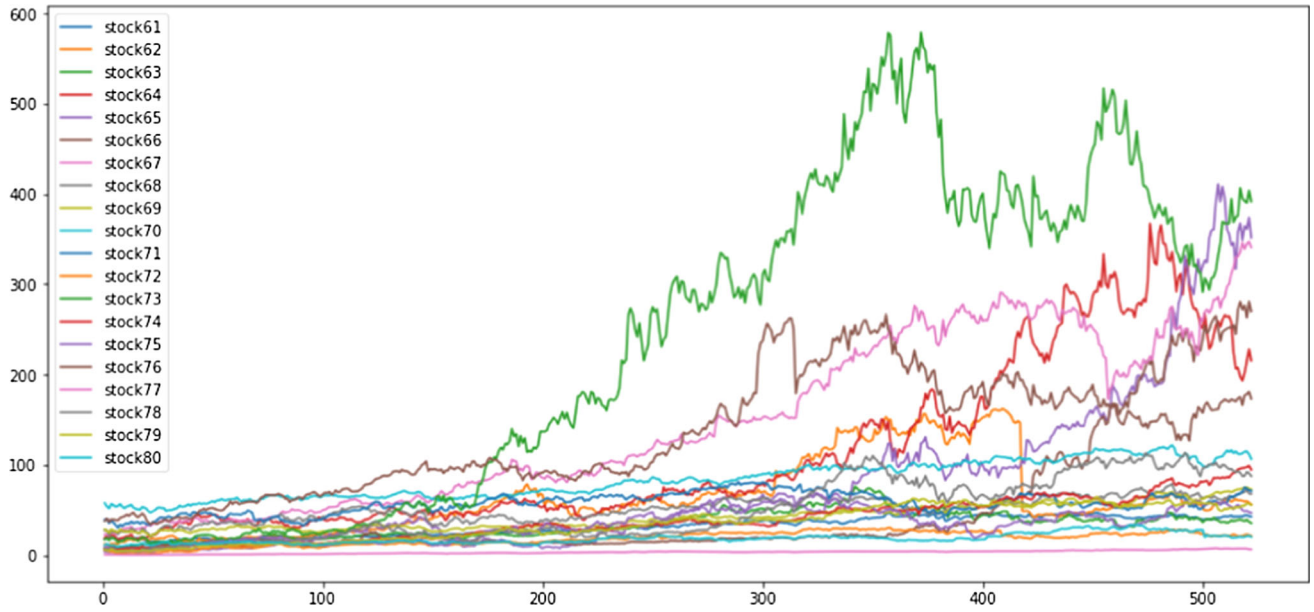

Fig. 2 Depiction of 20 stocks from the dataset for 522 weeks

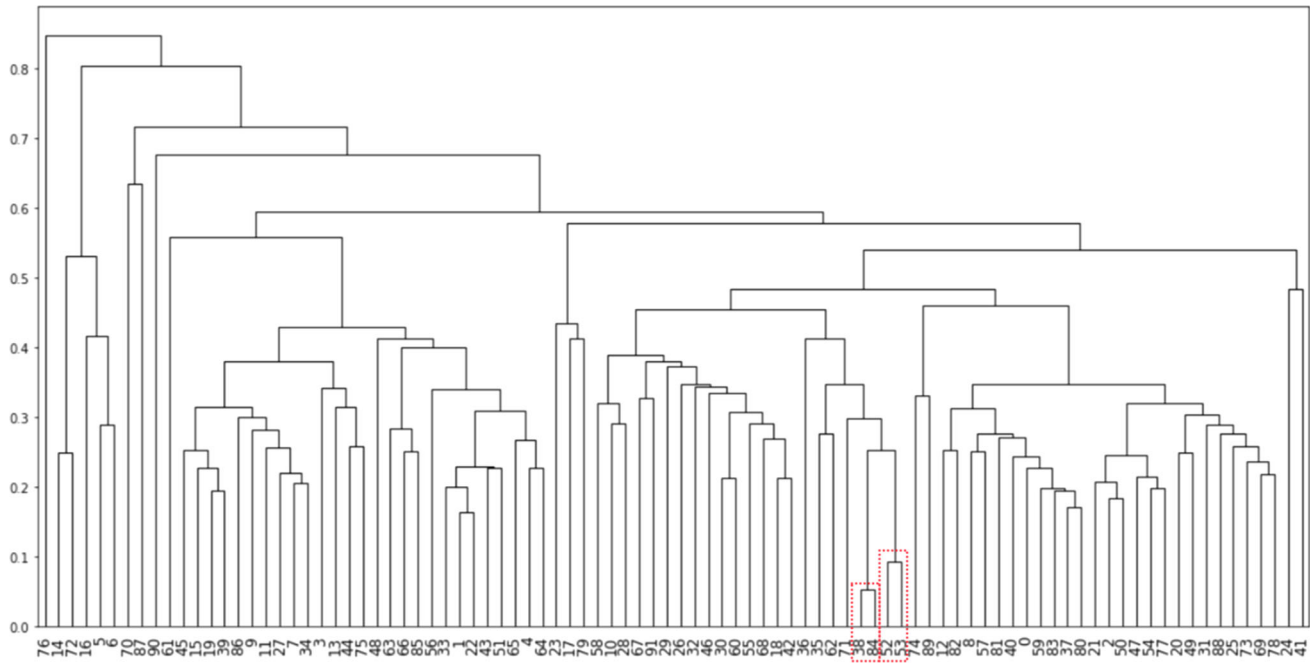

Fig. 3 Hierarchical clustering based on the suggested method

rable to the euclidean method, but also that it provides more efficient clustering for these examples.

Furthermore, as we mentioned before, stock markets are prone to exogenous factors such as bad or good news (see e.g., Chan (2003)). If a method pairs two stocks as similar, one can expect that after the occurrence of an outlier(s), the method would still evaluate these stocks alike. Hence, we will compare the performance of our method, and the euclidean distance metric for the stocks containing outliers. Recall from

Table 1 The Davies–Bouldin index for clustering based on the proposed method and euclidean method

Method	6 Clusters	8 Clusters	10 Clusters
The suggested method	0.61	0.64	0.72
The euclidean method	0.71	0.85	0.82

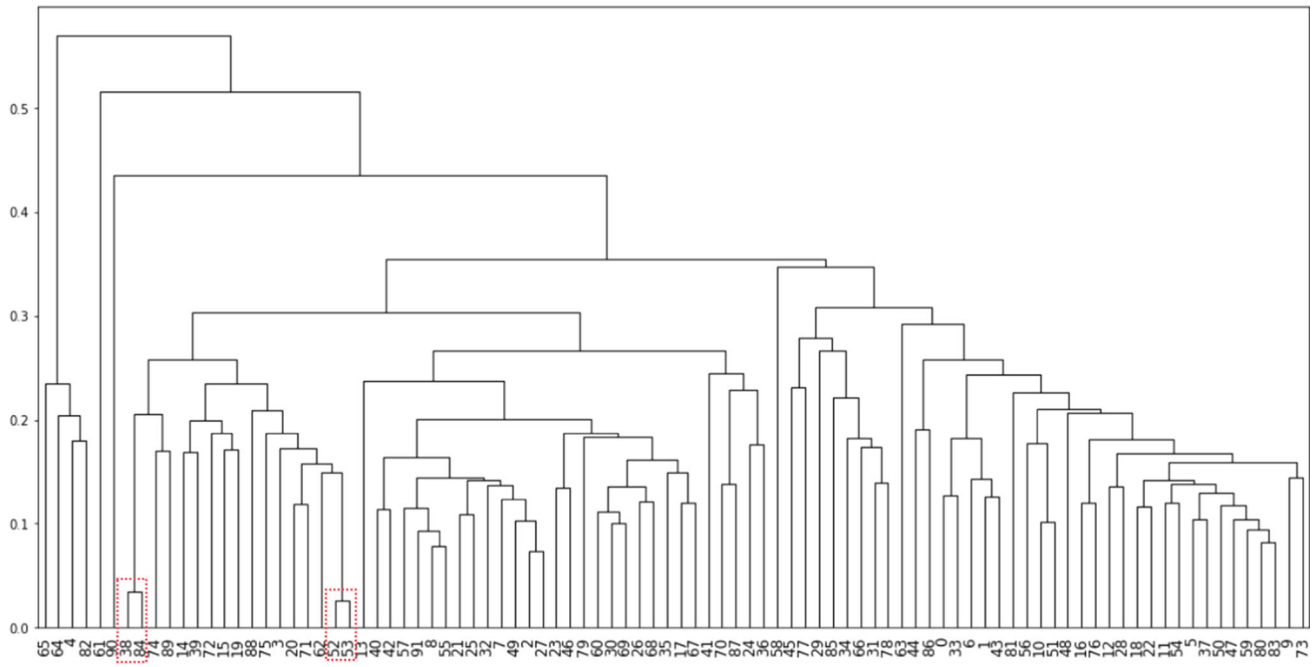

Fig. 4 Hierarchical clustering based on the Euclidean method

Fig. 5 Stock pair (38 and 84)

the previous section that based on both methods, stocks 52 and 53 are very similar to each other since their distance is minimal. Therefore, first, we will add some random artificial outliers to the stock 52, but we do not alter the stock 53 as shown in Fig. 7. Subsequently, we apply both methods to re-evaluate the similarity between these stocks.

Table 2 demonstrates the results. It is apparent, after including artificial outliers, that the euclidean distance has a dramatic jump (around 1800% increase). At the same time, the purposed method shows a minimal increase in distance (33%), which means that the suggested method is much less sensitive to the presence of outliers. Considering that the suggested method is based on the F-transform, it evaluates the similarity between the stocks concerning their local trend cycles; therefore, it does not have the drawbacks of raw-data-based approaches such as the euclidean distance. The

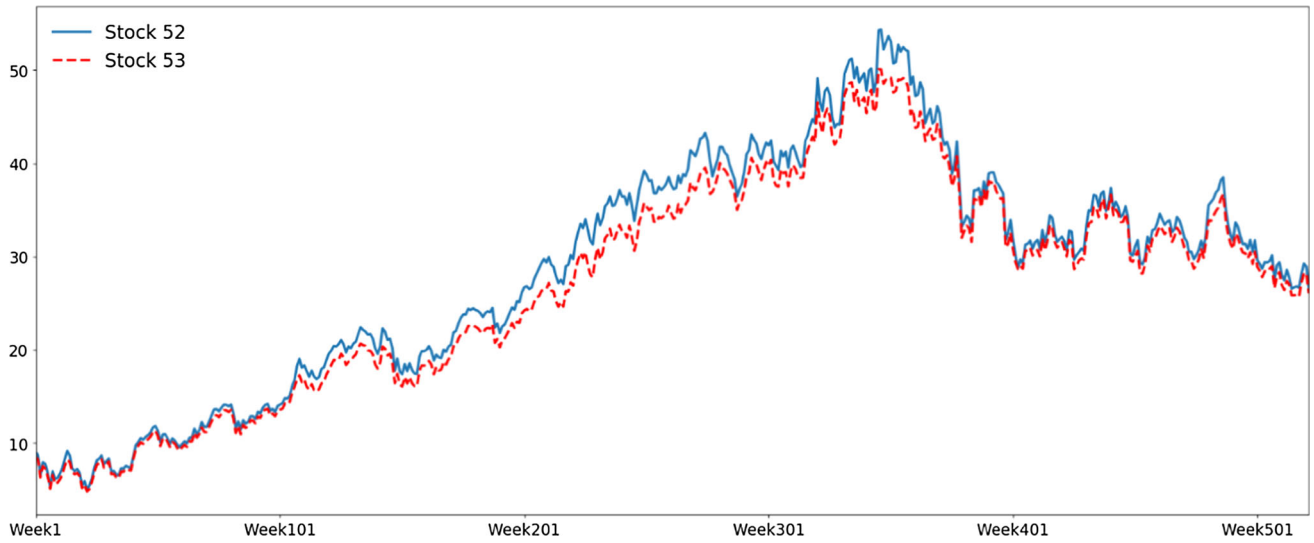

Fig. 6 Stock pair (52 and 53)

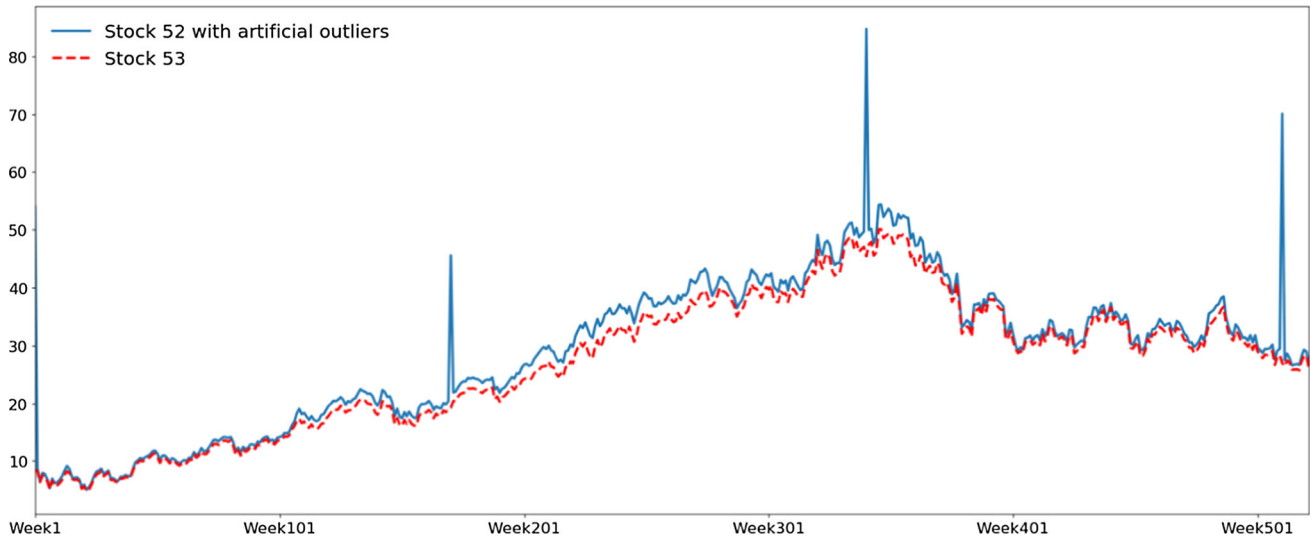

Fig. 7 Stock pair (52 and 53) containing artificial outliers

Table 2 The distance between stock 52 and 53, before and after outliers

Method	Distance before outliers	Distance after outliers
The suggested method	0.09	0.12
The euclidean method	0.17	3.33

latter methods are sensitive to noisy data(Zervas and Ruger 1999). One advantage of the euclidean method is its simplicity; however, the suggested method is also relatively simple since it has only one parameter to set (the length of the basic functions). Moreover, experts are able to adjust the suggested similarity measure, according to their time slot of interest.

3.4 Similarity at shifted-lag or lead-lag relation

The examples we provided earlier demonstrate the applicability and strength of the suggested method in finding similar behavior between stocks at lag-zero. However, there exist situations that two stocks might not be significantly similar at lag-zero, but they are more similar in shifted-lag(s). A condition where one (leading) variable is cross-correlated with the values of another (lagging) variable at other times is characterized as a lead-lag effect. The existence of the lead-lag effect between markets and its causes has been authenticated by many researchers (Roll 1992; Herbst et al. 1987; Mech 1993; McQueen et al. 1996). Generally, in practice, investors are interested in the relation among the return of the market and not actual price values. Lo and MacKinlay were among

Table 3 Cross-similarity and cross-correlation between daily return of DAX and AEX

Returns	i	Similarity	Correlation
AEX, DAX(i)	0	0.46	0.35
AEX, DAX(i)	−1	0.60	0.70

the first pioneers who showed how the return of small firms correlate with past returns of big firms(Lo and MacKinlay 1990), and more recently, Kewei Hou argues that there are strong intra-industry lead-lag effect (Hou 2007). The conventional method to evaluate this lead-lag relation among international stock indices is by cross-correlation. In this paper, in order to assess the lead-lag relation with the suggested method, we will move the stock returns against each other in different lags. Since the suggested method evaluates the strength of similarity between stocks, if the similarity degree in shifted-lag is considerably higher, we can assume that there is a lead-lag relation between the return of stocks. To illustrate the method in practice, here we demonstrate the relation between the daily return of two international stock indices, Germany (DAX) and Netherlands (AEX). We obtained their daily closing prices from yahoo finance[8] from 03/01/2018 to 29/03/2018 and calculate their return for that period.

The behavior of these daily returns, as well as their trend estimation , is represented in Fig. 8. The blue line demonstrates the data, and the dashed orange line is their estimation by F-transform. Note that unlike the prices in the previous example, here, we do not seek exact estimation for daily returns.

Data shows that lag one is a proper choice for shifting, meaning that we measure the similarity among the returns with one shift. Table 3 demonstrates the degree of their similarity at lag zero and lag one. These results suggest that the return of AEX follows a similar behavior as DAX after one day. The suggested similarity measure shows that the similarity between DAX and AEX(-1) is higher than their association at lag-zero. Seemingly, cross-correlation confirms this conclusion as well.

As shown in Fig. 9, by shifting the AEX for one lag, its similarity to DAX increases. Therefore, arguably for this period, DAX has a leading effect on AEX. Note that this relationship should not be considered as a causal relation.

However, it is possible to examine if this lead-lag relation founded by the suggested method can be causal. By causal relation, we mean the so-called Granger causality (Granger 1969). This concept is defined in terms of the predictability of a variable from its own past or the past of another variable. A time series X is said to Granger cause a time series Y if the available information apart from X provides statistically significant information about future values of Y.

In practice, to model the causality between two variables, it is imperative to determine the direction of the causality and its lag. To measure the causality, in Granger (1969), Granger proposed to compute the causal lag and causal strength (concerning two distinguish directions) based on the coherence and the phase functions defined with the help of the cross-spectrum of two stationary processes (Mandel and Wolf 1976).

Granger originally proposed a test based on comparing the mean square error of the forecasts of a variable with and without using the past of another variable. This work is then generalized in Granger (1980), where he assumes that, at time t, the value $Y(t+1)$ is a random variable which can be characterized by probability statement of the form $\text{Prob}(Y(t+1) \in A)$, for some given set A. Then, a general Granger causality definition is the following.

Definition 2 A time series X is said to cause Y if

$$\text{Prob}(Y(t+1) \in A \mid \mathcal{F}_t) \neq \text{Prob}(Y(t+1) \in A \mid \mathcal{F}_{-X}(t)),$$

where A is a universe in which $X(t)$ and $Y(t)$ are measured at specific time points $t \in \{1, \ldots, t\}$. Furthermore, $\mathcal{F}(t)$ represents information available at time t in the entire universe, and $\mathcal{F}_{-X}(t)$ is this information after X being excluded.

Hence, we tested the Granger causality between the pair (DAX, AEX) at lag 1. Table 4 depicts the results. Here, we cannot reject the hypothesis that DAX does not Granger cause AEX, but we do reject the hypothesis that AEX does not Granger cause DAX. Therefore it appears that Granger causality runs one-way from DAX to AEX at lag 1, and not the other way. This finding can be used later on for improving the prediction of AEX based on the past values of DAX. Thereby, the lead-lag relation that we found earlier can be regarded as a Granger-causal relation.

4 Conclusion

In this paper, we extended our previous paper (Mirshahi and Novák 2020) in two directions. First, we gave a proof for the suggested similarity method. Second, we showed that the suggested method could help to find the led-lag relation between the international stock market's return. We showed that the Granger causality confirms our findings. The method is based on the application of the fuzzy transform and a customized metric. The idea relies on the estimation of local trends using an inverse fuzzy transform. The financial time series can then be paired together according to the similarity of the adjoint time series consisting of the local trends. First, we demonstrated the application of the suggested method as

[8] www.finance.yahoo.com.

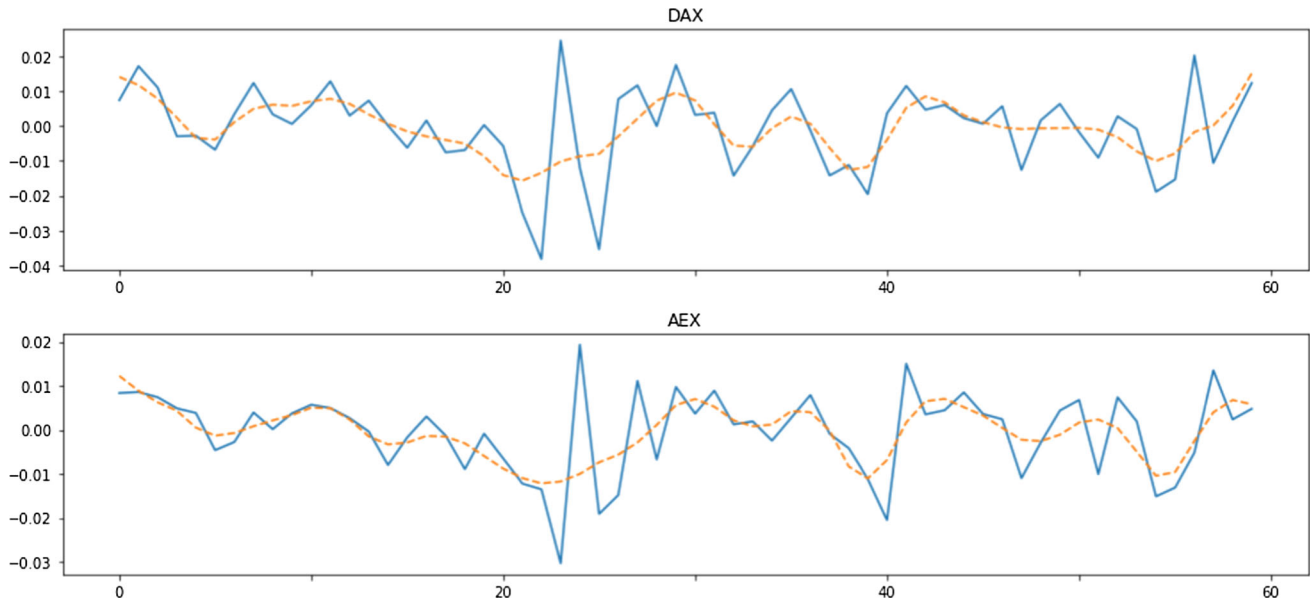

Fig. 8 Indices and their trend

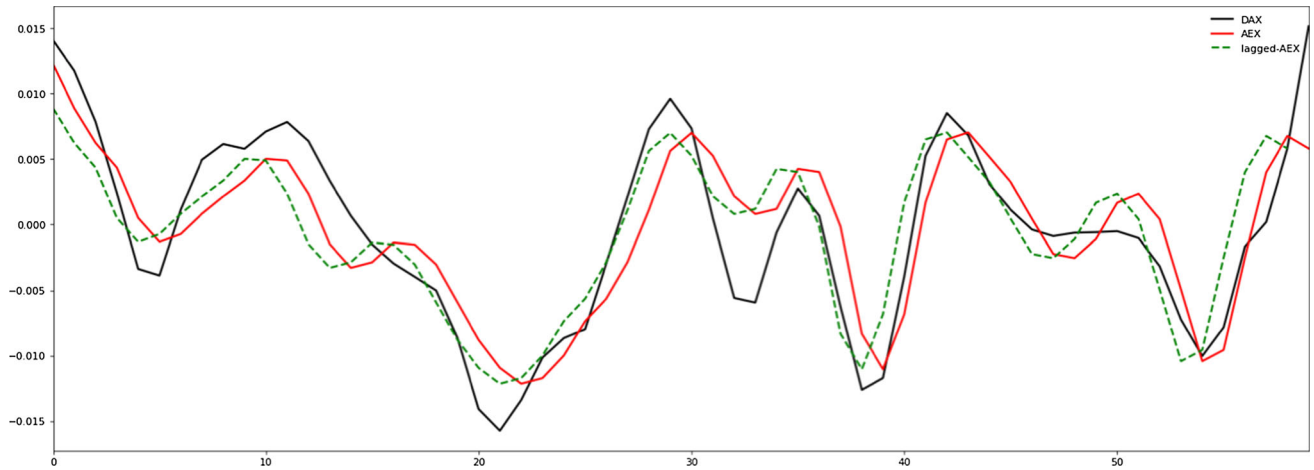

Fig. 9 DAX, AEX, shifted-AEX

Table 4 Granger causality between pair DAX- AEX provided by the software EViews10

Null hupothesis	Observations	F-statistic	Prob.
DAX does not Granger cause AEX	59	68.5058	3.E-11
AEX does not Granger cause DAX		1.01926	0.3170

a similarity measure on stock's price values in addition to its comparison with the euclidean distance.

Future work will focus on applying this method in portfolio management and evaluating its profitability in finance. Another addition to this work can be using the findings in multivariate forecasting.

Acknowledgements The paper has been supported by the grant 18-13951S of GAČR, Czech Republic.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Soheyla Mirshahi and Vilém Novák. The first draft of the manuscript was written by both authors, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Human and animal rights This article does not contain any studies with human participants or animals performed by the authors.

References

Anděl J (1976) Statistical analysis of time series. SNTL, Praha ((in Czech))

Bekaert G, Harvey CR (2003) Market integration and contagion. Tech. rep. National Bureau of Economic Research

Bernanke B (2016) The relationship between stocks and oil prices. Ben Bernanke's Blog on Brookings posted on February, vol 19

Cha B, Oh S (2000) The relationship between developed equity markets and the pacific basin's emerging equity markets. Int Rev Econ Finance 9(4):299–322

Chan WS (2003) Stock price reaction to news and no-news: drift and reversal after headlines. J Financ Econ 70(2):223–260

Davies DL, Bouldin DW (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 2:224–227

Devlin SJ, Gnanadesikan R, Kettenring JR (1975) Robust estimation and outlier detection with correlation coefficients. Biometrika 62(3):531–545

Eun CS, Shim S (1989) International transmission of stock market movements. J Financ Quantitat Anal 24(2):241–256

Fu T-C (2011) A review on time series data mining. Eng Appl Artif Intell 24(1):164–181

Granger CW (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica J Econ Soc, pp 424–438

Granger CW (1980) Testing for causality: a personal viewpoint. J Econ Dyn Control 2:329–352

Hamao Y, Masulis RW, Ng V (1990) Correlations in price changes and volatility across international stock markets. Rev Financial Stud 3(2):281–307

Hamilton J (1994) Time series analysis. Princeton University Press, Princeton

Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier, Amsterdam

Herbst AF, McCormack JP, West EN (1987) Investigation of a lead-lag relationship between spot stock indices and their futures contracts. J Futures Markets 7(4):373–381

Hilliard JE (1979) The relationship between equity indices on world exchanges. J Finance 34(1):103–114

Hou K (2007) Industry information diffusion and the lead-lag effect in stock returns. Rev Financial Stud 20(4):1113–1138

Kaufman L, Rousseeuw PJ (2009) Finding groups in data: an introduction to cluster analysis, vol 344. Wiley, Hoboken

Keogh E, Kasetty S (2003) On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Mining knowl Discovery 7(4):349–371

Kullmann L, Kertész J, Kaski K (2002) Time-dependent cross-correlations between different stock returns: a directed network of influence. Phys Rev E 66(2):026125

Liao TW (2005) Clustering of time series data-a survey. Pattern Recognit 38(11):1857–1874

Lo AW, MacKinlay AC (1990) When are contrarian profits due to stock market overreaction? Rev Financial Stud 3(2):175–205

Mandel L, Wolf E (1976) Spectral coherence and the concept of cross-spectral purity. JOSA 66(6):529–535

Martens M, Poon S-H (2001) Returns synchronization and daily correlation dynamics between international stock markets. J Bank Finance 25(10):1805–1827

McQueen G, Pinegar M, Thorley S (1996) Delayed reaction to good news and the cross-autocorrelation of portfolio returns. J Finance 51(3):889–919

Mech TS (1993) Portfolio return autocorrelation. J Financial Econ 34(3):307–344

Mining WID (2006) Data mining: Concepts and techniques. Morgan Kaufinann

Mirshahi S, Novák V (2020) A fuzzy approach for similarity measurement in time series, case study for stocks. In: International conference on information processing and management of uncertainty in knowledge-based systems, Springer, pp 567–577

Morse MD, Patel JM (2007) An efficient and accurate method for evaluating time series similarity. In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data, ACM, pp 569–580

Nguyen L, Novák V (2015) Filtering out high frequencies in time series using F-transform with respect to raised cosine generalized uniform fuzzy partition. In: Proc. Int. Conference FUZZ-IEEE, Istanbul. IEEE Computer Society, CPS, p 2015

Nguyen L, Novák V (2018) Forecasting seasonal time series based on fuzzy techniques. Fuzzy Sets and Systems, (to appear)

Novák V, Perfilieva I, Močkoř J (1999) Mathematical principles of fuzzy logic. Kluwer, Boston

Novák V, Štěpnička M, Dvořák A, Perfilieva I, Pavliska V, Vavříčková L (2010) Analysis of seasonal time series using fuzzy approach. Int J General Syst 39(3):305–328

Novák V, Perfilieva I, Holčapek M, Kreinovich V (2014) Filtering out high frequencies in time series using f-transform. Inf Sci 274:192–209

Novák V, Perfilieva I, Dvořák A (2016) Insight into fuzzy modeling. Wiley, Hoboken

Roll R (1992) Industrial structure and the comparative behavior of international stock market indices. J Finance 47(1):3–41

Serra J, Arcos JL (2014) An empirical evaluation of similarity measures for time series classification. Knowl Based Syst 67:305–314

Statman M, Scheid J (2008) Correlation, return gaps, and the benefits of diversification. J Portfolio Manage 34(3):132–139

Vlachos M, Hadjieleftheriou M, Gunopulos D, Keogh E (2006) Indexing multidimensional time-series. VLDB J 15(1):1–20

Wang PE (ed) (2001) Computing with Words. J. Wiley, New York

Wang Y, Wei Y, Wu C (2010) Cross-correlations between chinese a-share and b-share markets. Physica A Stat Mech Appl 389(23):5468–5478

Wang X, Mueen A, Ding H, Trajcevski G, Scheuermann P, Keogh E (2013) Experimental comparison of representation methods and distance measures for time series data. Data Mining Knowl Discovery 26(2):275–309

Wu C, Su Y-C (1998) Dynamic relations among international stock markets. Int Rev Econ Finance 7(1):63–84

Zervas G, Ruger SM (1999) The curse of dimensionality and document clustering