A Fuzzy Approach for Similarity Measurement in Time Series, Case Study for Stocks

Soheyla Mirshahi(✉) and Vilém Novák

Institute for Research and Applications of Fuzzy Modeling, NSC IT4Innovations, University of Ostrava, 30. dubna 22, 701 03 Ostrava 1, Czech Republic
{soheyla.mirshahi,vilem.novak}@osu.cz

Abstract. In this paper, we tackle the issue of assessing similarity among time series under the assumption that a time series can be additively decomposed into a trend-cycle and an irregular fluctuation. It has been proved before that the former can be well estimated using the fuzzy transform. In the suggested method, first, we assign to each time series an adjoint one that consists of a sequence of trend-cycle of a time series estimated using fuzzy transform. Then we measure the distance between local trend-cycles. An experiment is conducted to demonstrate the advantages of the suggested method. This method is easy to calculate, well interpretable, and unlike standard euclidean distance, it is robust to outliers.

Keywords: Similarity measurements · Stock markets similarity · Time series analysis · Time series data mining

1 Introduction

Time series is a feasible way of representing data in many fields, including the finance sector. Financial crises in the 19th and early 20th caused a challenging situation for economies, and it led to a massive interest in economic and financial analysis. In this situation, any information that provides a better understanding to the behavior of markets is highly critical. Among many types of research concerning data mining in time series (see-[4,7,9,10]); One of the key applications in this field [11] is stock data mining. 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 [18]. In the construction of financial portfolios (see [5]), diversification, which conveys investing in a variety of assets, is a key to reduce the risk of a chosen portfolio. Thus, identifying stocks that share similar behavior is vital. There is no straightforward approach, known as the best measure for assessing the similarities in time series. Surprisingly, many simple approaches like simple euclidean distance can outperform the most complicated approaches [18]. Wang et al., in 2013, perform an extensive comparison

M.-J. Lesot et al. (Eds.): IPMU 2020, CCIS 1239, pp. 567–577, 2020.
https://doi.org/10.1007/978-3-030-50153-2_42

between nine measurements across 38 data sets from various scientific domains (see [21]). 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. However, stock markets have some properties which make the current similarity measures unfavorable. For instance, stocks react to a lot of exogenous factors such as news (see, e.g., [2]); thus, the presence of outliers in them is inevitable. Therefore, developing a measure that can react to the nature of stock markets seems essential.

A very effective technique in the 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. [16,17]). In this paper, using fuzzy transform, we first assign to each 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 similarity measurement: 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 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 basic characteristics of the time series. Finally, Our method is flexible and can answer the question of how we can find stocks that behave similarly 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).

The paper is structured as follows. After Introduction, we describe our method in Sects. 2 and 3. Section 4 is dedicated to an illustration of the purposed method and 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., [1,6]) $X : \mathbb{T} \times \Omega \to \mathbb{R}$ 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., [15,16] 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 $Supp(A_k) = [c_{k-1}, c_{k+1}]$[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 F-transform has two phases: direct and inverse. The *direct* F-transform assigns to each $A_k \in \mathcal{A}$ a component $F_k[f|\mathcal{A}]$. 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 $F_k^1[f|\mathcal{A}](x) = \beta_k^0[f] + \beta_k^1[f](x - c_k)$. 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 $A_k \in \mathcal{A}$.

From the direct F-transform of f

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

we can form a function $\mathbf{I}[f|\mathcal{A}] : [a, b] \to [c, d]$ using the formula $\mathbf{I}[f|\mathcal{A}](x) = \sum_{k=0}^{n}(F_k[f|\mathcal{A}] \cdot A_k(x))$, $x \in [a, b]$. The function $\mathbf{I}[f|\mathcal{A}]$ is called the *inverse F-transform* of f 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. [14,16]). 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

$$T_1 > \cdots > T_r, \tag{2}$$

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

[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.

Theorem 1. *Let $\{X(t) \mid t \in \mathbb{T}\}$ be a realization of the time series (1). 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 $h = dT_q$ where $d \in \mathbb{N}$ and T_q is a periodicity corresponding to λ_q then the corresponding inverse F-transform $\mathbf{I}[X|\mathcal{A}]$ of $X(t)$ gives the following estimation of the trend-cycle:*

$$|\,\mathbf{I}[X|\mathcal{A}](t) - TC(t)| \leq 2\omega(h, TC) + D \tag{3}$$

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.*

The precise form of D and the detailed proof of this theorem can be found in [13,16]. 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 (2) can be found using the classical technique of periodogram—see [1,6].

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.

3 The Suggested Similarity Measurement

In this section, we will describe how our suggested method evaluates the pairwise similarity between time series.

Definition 1. *Let $X = \{X(t)|t = 1, \ldots, n\}$ and $Y = \{Y(t)|t = 1, \ldots, n\}$ be two time series of the length n and TC_X and TC_Y be estimations of trend cycles of X and Y respectively calculated based on Eq. (3). Then we define the similarity between these two time series as follows:*

$$S(X,Y) = 1 - \frac{1}{n}\sum_{t=1}^{n} \frac{|TC_X(t) - E(TC_X) - (TC_Y(t) - E(TC_Y))|}{|TC_X(t) - E(TC_X)| + |TC_Y(t) - E(TC_Y)|}, \tag{4}$$

where $E(TC_X)$ and $E(TC_Y)$ are mean values (averages) of TC_X and TC_Y, respectively and $|.|$ denotes absolute value. It is easy to show that $S(X,Y) \in [0,1]$ where it has certain features that is described on the following theorem and can be proved. In Definition 1, it is necessary to emphasize, that TC_X and TC_Y are estimation, not the real trend-cycles, since we do not know them (cf. formulas (1) and (3)).

Theorem 2. *$S(X,Y)$ is a fuzzy equality w.r.t. Łukasiewicz conjunction, i.e., it is: reflexive : $S(X,X)=1$, symmetric : $S(X,Y)=S(Y,X)$ and transitive : $S(X,Y)\otimes S(Y,Z)\leq S(X,Z)$ where $\otimes$ is the Łukasiewicz conjunction defined by $a\otimes b=\{\max 0, a+b-1\}$.*

A stock can be seen as a time series $\{X(t)|t=1,\ldots,n\}$ where $X(t)$ is closing price at time t within an interval $[0,T]$. For instance, let us consider closing price of a stock from Nasdaq INC[3], 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 such that the length of each basic functions $A_2;...;A_{n_1}$ is equal to a proper time slot. In our case, by setting the length of basic function to $four$, we obtain the approximation of the trend-cycles 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_n are incomplete). Regardless it is clear that F-transform has approximated 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 is clear to see that F-transform has successfully wiped out the outlier while preserved the core behavior of the stock.

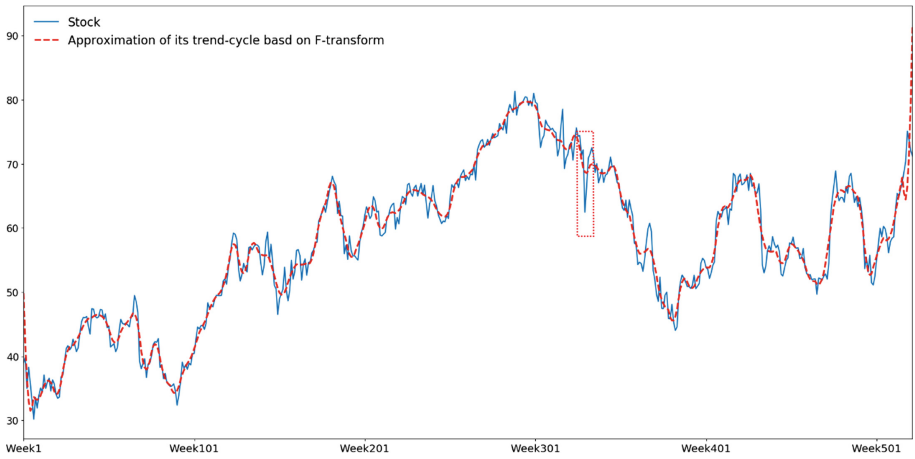

Fig. 1. A stock and its TC approximation based on F-transform.

The similarity from Definition 1 can be used for measuring the similarity between any number of stocks. We can measure using it also local behavior of them. In the next section, we will demonstrate how our suggested method works

[3] https://www.nasdaq.com/.

with a relatively large data set in conjunction with its comparison to standard the euclidean distance.

4 Illustration

4.1 Data Set

Our data set consists of a closing price of 92 stocks over 522 weeks obtained from Nasdaq INC. 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 series is impossible. Therefore it seems necessary to consider similarity between time series.

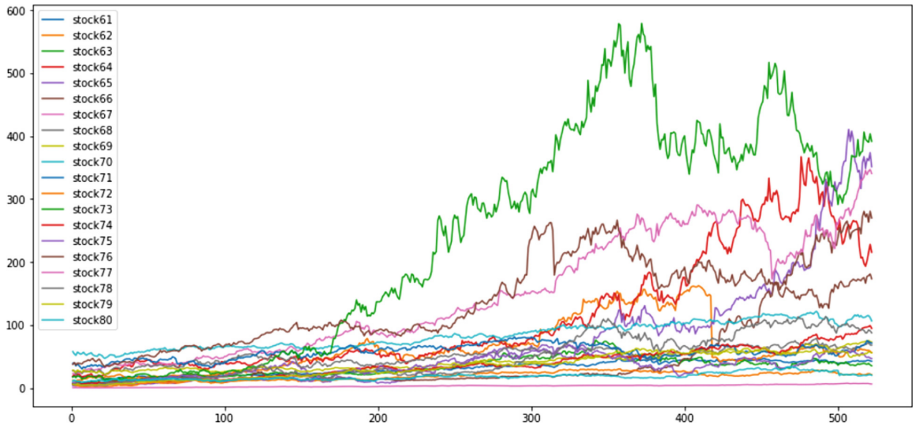

Fig. 2. Depiction of 20 stocks from the dataset for 522 weeks

4.2 Evaluation of the Suggested Method

One possible way to evaluate the competency of any new similarity measurement (distance measurement), is to apply it for data clustering. The quality of clustering based on the new and current similarities can validate the competency of the suggested method [12,19]. 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 [18,20]). 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 [8]. 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.

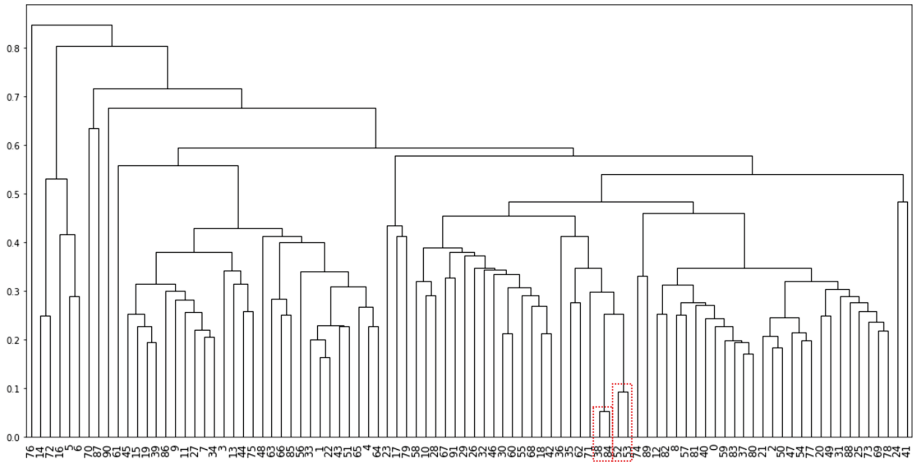

Fig. 3. Hierarchical clustering based on the suggested method (Color figure online)

Red dashed squares in 4.2 and 4.2 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 behaviour of theses stock pairs.

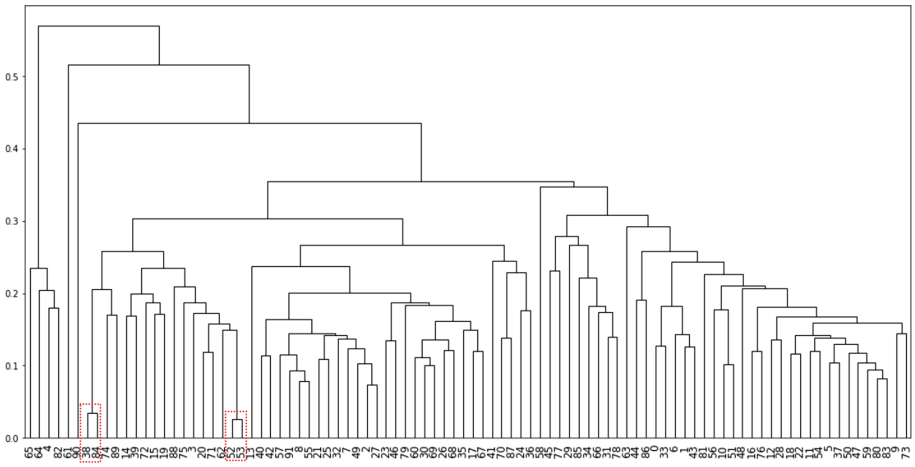

Fig. 4. Hierarchical clustering based on the Eucliden method

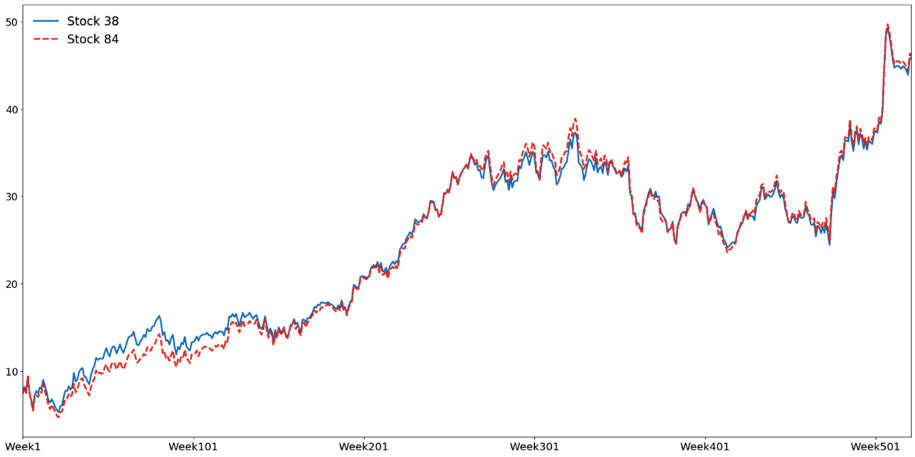

Fig. 5. Stock pair (38 and 84)

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 [3]. 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 comparable to the euclidean method, but also it has provided more efficient clustering for these examples.

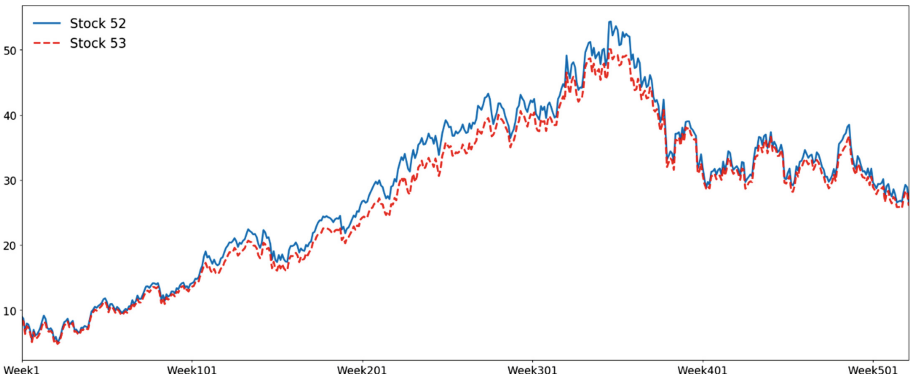

Fig. 6. Stock pair (52 and 53)

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

Furthermore, as we mentioned before, stock markets are prone to exogenous factors such as bad or good news (see e.g.,[2]). 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 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 latter methods are sensitive to noisy data [22]. 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.

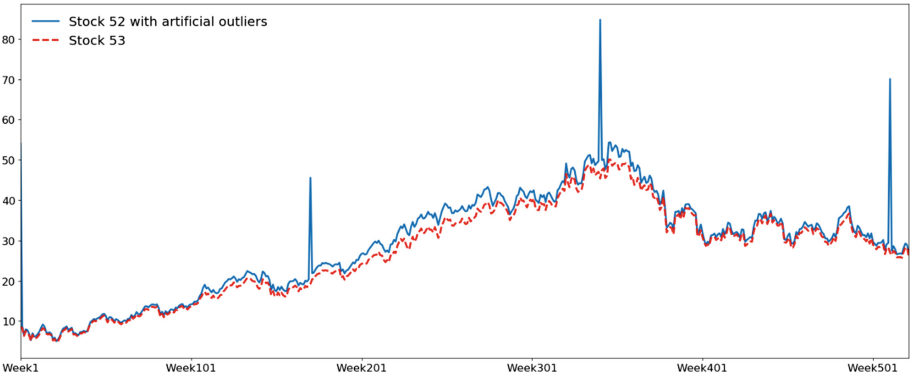

Fig. 7. Stock pair (52 and 53) containing artificial outliers

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

5 Conclusion

In this paper, we developed a new method for pairwise similarity measurement. The method is based on the application of the fuzzy transform and a customized metric. The idea is based on the estimation of local trends using inverse fuzzy transform. The time series can then be paired together according to the similarity of the adjoint time series consisting of the local trends. We demonstrated the application of the suggested method in real life in addition to its comparison with the euclidean distance. Experimental results verify the capability of the suggested method for measuring the similarity between time series.

Further work will be focused on the application of this method in portfolio management and evaluation of its profitability in finance. Another addition to this work can be extending the method for time series of various lengths and compare the result with the so-called dynamic time warping (DTW) method.

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

References

1. Anděl, J.: Statistical Analysis of Time Series. SNTL, Praha (1976). (in Czech)
2. Chan, W.S.: Stock price reaction to news and no-news: drift and reversal after headlines. J. Financ. Econ. **70**(2), 223–260 (2003)

3. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **2**, 224–227 (1979)
4. Fu, T.C.: A review on time series data mining. Eng. Appl. Artif. Intell. **24**(1), 164–181 (2011)
5. Gilli, M., Maringer, D., Schumann, E.: Numerical Methods and Optimization in Finance. Academic Press, Cambridge (2019)
6. Hamilton, J.: Time Series Analysis. Princeton University Press, Princeton (1994)
7. Han, J., Pei, J., Kamber, M.: Data mining: concepts and techniques. Elsevier(2011)
8. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis, vol. 344. Wiley, Hoboken (2009)
9. Keogh, E., Kasetty, S.: On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Min. Knowl. Disc. **7**(4), 349–371 (2003)
10. Liao, T.W.: Clustering of time series data–a survey. Pattern Recogn. **38**(11), 1857–1874 (2005)
11. Mining, W.I.D.: Data Mining: Concepts and Techniques. Morgan Kaufinann, Burlington (2006)
12. Morse, M.D., Patel, J.M.: An efficient and accurate method for evaluating time series similarity. In: Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data, pp. 569–580. ACM (2007)
13. Nguyen, L., Novák, V.: Filtering out high frequencies in time series using F-transform with respect to raised cosine generalized uniform fuzzy partition. In: Proceedings of International Conference on FUZZ-IEEE 2015. IEEE Computer Society, CPS, Istanbul (2015)
14. Nguyen, L., Novák, V.: Forecasting seasonal time series based on fuzzy techniques. Fuzzy Sets and Systems (to appear)
15. Novák, V., Perfilieva, I., Dvořák, A.: Insight into Fuzzy Modeling. Wiley, Hoboken (2016)
16. Novák, V., Perfilieva, I., Holčapek, M., Kreinovich, V.: Filtering out high frequencies in time series using F-transform. Inf. Sci. **274**, 192–209 (2014)
17. Novák, V., Štěpnička, M., Dvořák, A., Perfilieva, I., Pavliska, V., Vavříčková, L.: Analysis of seasonal time series using fuzzy approach. Int. J. Gen Syst **39**(3), 305–328 (2010)
18. Serra, J., Arcos, J.L.: An empirical evaluation of similarity measures for time series classification. Knowl.-Based Syst. **67**, 305–314 (2014)
19. Vlachos, M., Hadjieleftheriou, M., Gunopulos, D., Keogh, E.: Indexing multidimensional time-series. VLDB J. **15**(1), 1–20 (2006)
20. Wang, P.E. (ed.): Computing with Words. Wiley, New York (2001)
21. Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., Keogh, E.: Experimental comparison of representation methods and distance measures for time series data. Data Min. Knowl. Disc. **26**(2), 275–309 (2013)
22. Zervas, G., Ruger, S.M.: The curse of dimensionality and document clustering (1999)