Journal of Multidisciplinary Engineering Science and Technology (JMEST)
ISSN: 3159-0040
Vol. 2 Issue 6, June - 2015

A Complexity Invariant Distance Modification To Discover Similarity In Time Series

Eralda GJIKA (DHAMO)[1]

Department of Applied Mathematics, Faculty of Natural Sciences, University of Tirana

Tirana, Albania

eralda.dhamo@fshn.edu.al

Llukan Puka[2]

Department of Applied Mathematics, Faculty of Natural Sciences, University of Tirana

Tirana, Albania

llukan.puka@fshn.edu.al

Eglantina KALLUÇI (XHAJA)[3]

Department of Applied Mathematics,

Faculty of Natural Sciences, University of Tirana

Tirana, Albania

eglantina.xhaja@fshn.edu.al

***Abstract*—Time series similarity is a recent theme in research field. Many algorithm which use a wide range of similarity measure are introduced. In our previous work we have built an algorithm to find similar subsequences in time series. The proposed algorithm has been subjected to numerous tests to understand his kindness in discover similar subsequence. A considerable number of time series is tested. We have tried to improve the effectiveness of our algorithm by increasing efficiency in quality, number of detected subsequence as well as reducing the time of execution. Here we introduce a modification of CID distance which was first proposed by Batista et al. The complexity of classic CID and modified CID is analyzed. The modification is applied as a similarity measure in Chouakria index proposed by D. Chouakria et al. The algorithm was modified to support the proposed similarity measure. Tests on execution time and on number of similar subsequence were performed and are presented at the end of this work.**

Keywords—time series, distance, Chouakria index, CID, algorithm, R

I. INTRODUCTION AND MOTIVATION

Techniques for detecting particles similarities in the time series are of interest for many researchers in recent years. These techniques are based in measurements of similarity or dissimilarity. Some of these similarity measures are widely known in the literature such as the classic Euclidean distance. Euclidean distance is at the basis of most of the algorithms built with the purpose of clustering similarity subsequences. A similarity measure gives a numerical value that indicates how similar the two sequences are; on other hand a dissimilarity measure gives a numerical value that indicates how much the two sequences differ. In the first case we request to maximize the value and in the second case to minimize it. Many similarity measures are proposed in recent years and also algorithms for detecting similar subsequence are numerous in time series literature [1][2][3][4][5][6]. In a previous work we have presented a new algorithm for detecting similar subsequences in time series with presence of invariance complexity [7]. We have compared the efficiency of CID distance (Complexity Invariant Distance) with the Euclidean distance in time series with complexity. The algorithm was tested on simulated and real time series data. In both cases CID provide satisfactory results compared to the Euclidean distance. Latter on we have tested CID efficiency with Chouakria index [8][9] and obtained satisfactory numerical and graphical results. A large data set of time series of different nature (engineering, meteorology, medicine, demography, and many others) were tested and we realize the advantages of Chouakria index compared to CID distance. We also arrive at the conclusion that: combination of the properties of CID with Chouakria index (with CID distance) provides more satisfactory results than CID alone.

In this work we have modified our algorithm [7] considering a modification in CID and combining it with Chouakria index.

II. DEFINITIONS ON SIMILARITY AND DISTANCES SELECTING A TEMPLATE (HEADING 2)

Some useful definitions on time series and similarity measure are listed below.

*Definition 1*A time series Q of length n is an ordered sequence of real numbers $\{q_1, q_2, q_3, \ldots, q_n\}$

*Definition 2*A time series subsequence Qi, j = $\{q_i, q_{i+1}, q_{i+2}, \ldots, q_{i+m-1}\}$ is a continuous subsequence of Q which start at position i and has a length m.

*Definition 3*A motif (Qi, m; Q j, m) of length m is the most repeated subsequence along the time series which has significant meanings in a time series.

*Definition 4*ε –Range Query: Given a time series query T of length m, a time series database (DataTS), a (dis)similarity measure and a threshold ε, find the set of subsequences that are within distance ε from T.

Definition 5 In this work we will refer as the first subsequence that subsequence having the highest number of repetitions along the time series.

*Definition 6*Temporal correlation coefficient (CORT) between two subsequences Q and P of length m is defined as:

$$CORT(Q,P)=\frac{\sum_{i=1}^{m}(q_{i+1}-q_i)(p_{i+1}-p_i)}{\sqrt{\sum_{i=1}^{m}(q_{i+1}-q_i)^2}\sqrt{\sum_{i=1}^{m}(p_{i+1}-p_i)^2}} \quad (1)$$

Correlation coefficient is a standard statistical measure of similarity. it takes values between -1 and 1. A value of it close to -1 indicates that the two subsequences exhibit opposite behavior, and a value close to 1 indicates that the two subsequences exhibit similar behavior.

It is important to note that in our algorithm the time series which are compared are standardized to achieve scale and offset invariance. Given that modified CID gave comparable results with classic CID we tested the performance of them in Chouakria index. Tests were done on 88 time series and many executions on different length of subsequences. Some of the time series, are: cmort, AirPassengers, milk, unemp, souvenir, tea, birth, etc.

III. CLASSIC CID, MODIFIED CID AND CHOUAKRIA INDEX WITH MODIFIED CID

A. Classic CID

CID was first proposed by Batista and Keogh [6]. It was presented as a distance which can deal with the complexity of a time series. The classic CID distance is based on Euclidean distance so it can be applied only on subsequences with equal length. An adjustment factor (CF) and a complexity measure (CE) are used to calculate the CID between two subsequences P and Q:

$$CF(Q,P)=\frac{\max(CE(Q),CE(P))}{\min(CE(Q),CE(P))} \quad (2)$$

The complexity measure is calculated for each subsequence:

$$CE(Q)=\sqrt{\sum_{i=2}^{m}(q_i-q_{i-1})^2}$$

$$CE(P)=\sqrt{\sum_{i=2}^{m}(p_i-p_{i-1})^2}$$

and classic CID is calculated by:

$$CID(Q,P)=d_E(Q,P)*CF(Q,P) \quad (3)$$

The adjustment factor (CF) is obviously larger than 1. When the two time series have the same value of complexity measure then CF=1 and CID is the Euclidean distance.

If CF(Q,P) tends to have large values, this is due to high complexity between the time series. This might be a sign to stop further test with these time series. Additional tests should be performed to determine whether CID is suitable for comparison of these time series.

B. Modified CID

In classic CID the complexity measure of a time series is calculated as the square root of the sum of the differences between consecutive observations squared. In our proposal we take into consideration the square root of the sum of differences between observation and the mean of the time series. So, we calculate the complexity measure for each subsequence (respectively P and Q with length m) using the following formula:

$$CE_M(P)=\sqrt{\sum_{i=1}^{m}\left(p_i-\bar{P}\right)^2}$$

and

$$CE_M(Q)=\sqrt{\sum_{i=1}^{m}\left(q_i-\bar{Q}\right)^2} \quad (4)$$

Letter M as the index of CE stands for Modification. Another way to obtain the proposed complexity measure is by using the standard deviation of the subsequence P multiplying by m or (m-1). After finding these complexity measure for each of the two subsequences the correction factor (CF(P,Q)) may be calculated by the formula:

$$CF_M(P,Q)=\frac{\max[sd(P),sd(Q)]}{\max[sd(P),sd(Q)]} \quad (5)$$

and modifed CID is calculated by:

$$CID_M(P,Q)=d_E(P,Q)*CF_M(P,Q) \quad (6)$$

It is easy to show that modified CID does not satisfy all conditions of being a distance. It satisfies only the property of transition and symmetry. It can be seen as a similarity measure when used in algorithms.

C. C.Chouakria index

Chouakria proximity measure (also referred in this paper as Chouakria index) proposed by *A. D. Chouakria et al* in [7] uses the *cost* function and is calculated as below:

$$Chou(P,Q) = \frac{2}{1+e^{k*CORT(P,Q)}} * \sqrt{\sum_{i=1}^{m}(P_i - Q_i)^2}, k = 0,1,2,... \quad (7)$$

For different values of *k* we get different behavior of the distance. According to the importance that we give to the form and the importance that we give to the behavior, in our tests we decided to use *k=2*.

Replacing CID and modified CID respectively in Chouakria index we obtain the following formula:

$$Chou_{CID}(P,Q) = \frac{2}{1+e^{2*CORT(P,Q)}} * CID(P,Q) \quad (8)$$

$$Chou_{CID\_M}(P,Q) = \frac{2}{1+e^{2*CORT(P,Q)}} * CID_M(P,Q) \quad (9)$$

D. Complexity of classic CID, modified CID and Chouakria index with CID and modified CID

Our algorithm was created in R environment and all calculation and tests with time series were made in R. Below are the R code for the classic CID, modified CID and Chouakria index with modified CID.

Classic CID code in R

```
CID=function(Q,P){
CE_Q=sqrt(sum(diff(Q)^2))
CE_P= sqrt(sum(diff(P)^2))
CID=sqrt(sum((Q-
P)^2)*(max(CE_Q,CE_P)/min(CE_Q,CE_P))
print(CID)}
```

Modified CID code in R

```
CID_mod=function(Q,P){
CE_Q=sqrt((m-1)*var(Q))
CE_P= sqrt((m-1)*var(P))
CID=sqrt(sum((Q-
P)^2)*(max(sd(P),sd(Q))/min(sd(P),sd(Q))
print(CID_mod)}
```

Chouakria index with modified CID code in R

```
CID_mod_Chou=function(P,Q,K)
{
#P, Q two subsequences
#K a parameter defined by the user, recommended
K=2
f=sum(diff(P)*(diff(Q))/sqrt((sum(diff(P)^2)*sum(diff
(Q)^2))))
CE_Q=sqrt((m-1)*var(Q))
CE_P=sqrt((m-1)*var(P))
d=sqrt(sum((P-
Q)^2))*max(CE_Q,CE_P)/min(CE_P,CE_Q)
dist=2*d/(1+exp(K*f))
return(dist)
}
```

The algorithm proposed by Dhamo (Gjika) et al. in [7] and modified latter in [10] is tested in R environment, an outline of the algorithm in general steps is shown in Table 1.

TABLE I. AN OUTLINE OF THE ALGORITHM

Algorithm Finding Subsequence of length **m** in a time series
1.Declare the time series vector (T) and subsequence length (m) 2.Calculate the fluctuations in time series 3. Decide on the similarity measure (CID, modified CID or other) 4.Standardize all subsequences of length m 5.Calculate the confidence interval for similarity measure 6.Slide along the time series and find subsequences who have distance within the confidence interval (store information about all the subsequences that meet the conditions) 7.Display numerically and graphically the results for the first subsequence with the larger number of repetitions.

To understand if the distance that will be used for the detection of similarities may be CID, CID modified or other, the algorithm calculates in step 2 the fluctuations of the time series. A formula to calculate the fluctuations of a time series is proposed by [11].

$$Fluctuation(T) = \frac{1}{n-1}\sum_{i=1}^{n-1}(T_{i+1} - T_i)^2 \quad (10)$$

As it can be seen this is a variant of the complexity coefficient in CID distance. A large value of fluctuations means that CID or its modification may not be an appropriate distance so the user may choose between other known distances (such as, Euclidean or Chouakria with Euclidean distance).

In step 4, all subsequences are standardized with mean zero and standard deviation 1.

In step 5, we decide to consider the best confidence interval the one that shows satisfactory results in terms of overcoming the trends difficulties. Which means that the algorithm finds similar subsequences independently of the location. Also we have removed from classification as similar those subsequences extending in positions less than (m-1) units apart from each other (where m-motif length).

The confidence interval for the similarity between two subsequences is based on the minimum distance between two subsequences and an error $\varepsilon = 1.96 * \frac{sd(Q)}{\sqrt{m}}$, where Q is the query subsequence

with length m, sd(Q) is the standard deviation of the subsequence Q (which is equal to 1 because we have standardized the subsequences).

At this point of modifications we have calculated the complexity of the modified distances.

For the calculation of the classic CID (3) for a sequence of m elements we need $(6m-4)$ operations (among them we have 2 comparisons, which we have calculate as a common operation).

For the calculation of modified CID (6) for a sequence of m elements we need $8m$ operation.

These distances are implemented to the Chouakria index (Eq.7), producing respectively Eq.8 which needs $(13m-3)$ operations plus one function evaluation (exponential of a value), and for Eq.9 we need $(15m+1)$ plus one function evaluation.

As given from the results we observe that the distance calucated by Eq.9 increases the number of operations, but remaining linear with respect to m.

IV. Experimental results between classic CID, Modified CID and their involvement in Chouakria index

Here we present the goodness of modified CID. Many tests where conducted in a considerable number of time series with different nature, obtaining interesting results. Further we tested the use of the modified CID in Chouakria index to increase the quality of the algorithm on finding similar subsequences. The results obtained in this case were also impressive.

Figure 1 shows the results of our algorithm in *tea* time series using the classic CID and modified CID.

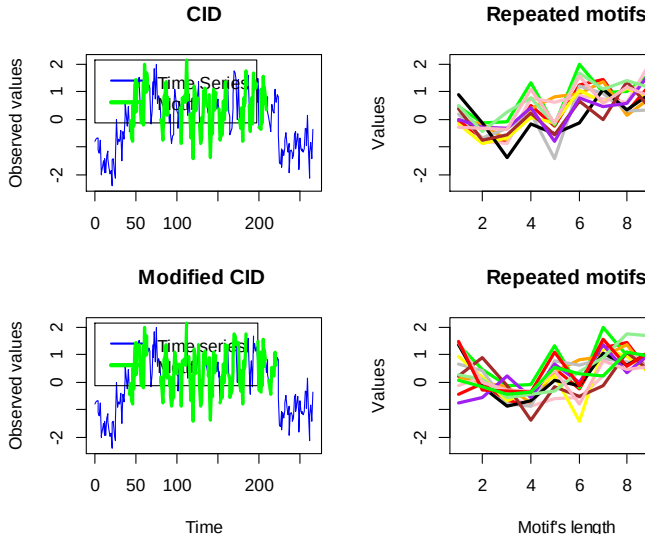
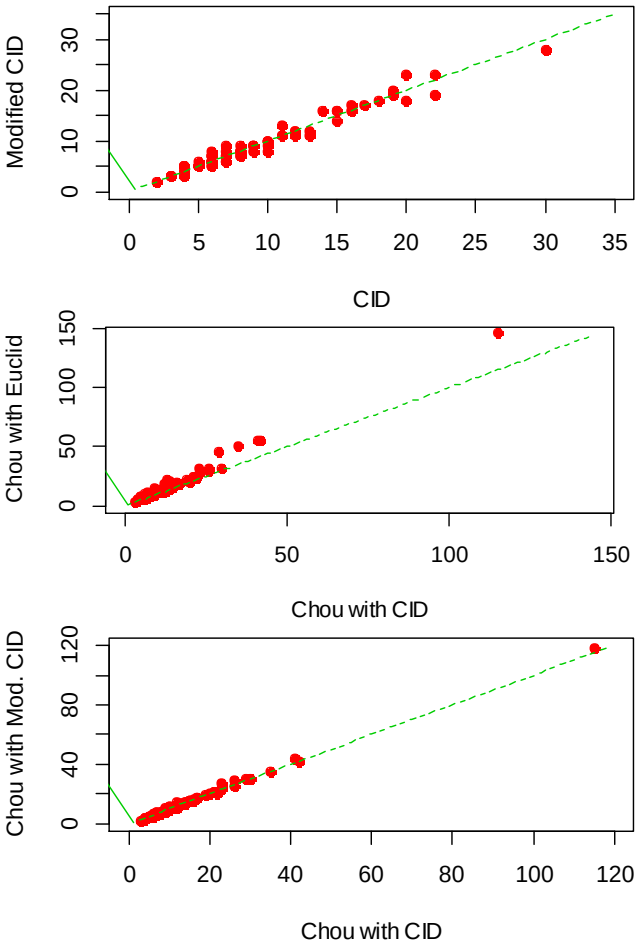


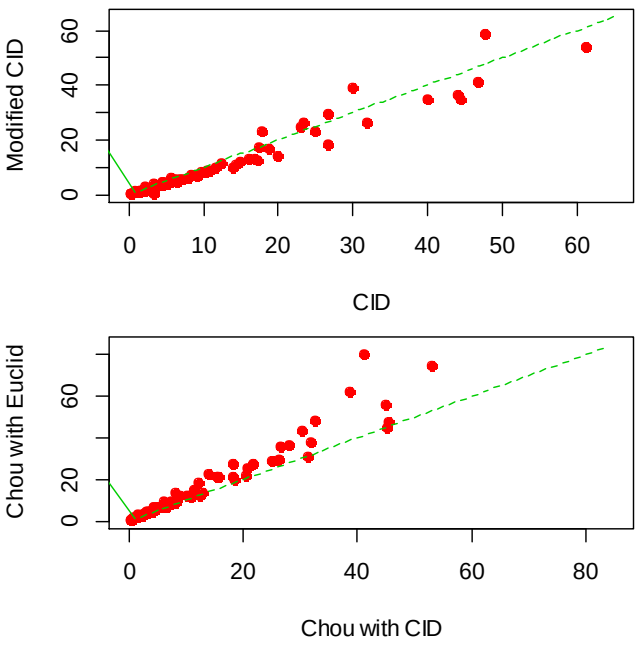
Fig. 1. *Classic CID and modified CID results (offices time serie)*

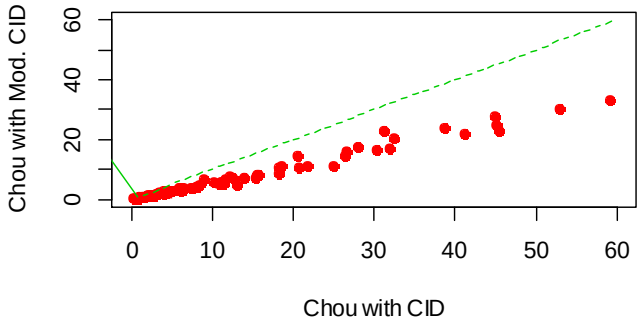
The time series has an apparent complexity and as seen from the graphical results the modified CID reveals more motifs and also it expand his search in a wider area than classic CID.

Figure 2 below shows the performance of classic CID, modified CID based on (a) the number of similar subsequences discovered and (b) average time of finding the first subsequence.



a) Number of similar subsequences





b) Average time (in sec) for finding similar subsequences with first subsequence

Fig. 2. *Efficiency of classic CID, modified CID and use in Chouakria index*

Each point on the graph shows one database which has as coordinates the number of subsequence similar to first subsequence detected by each similarity measure. The part of the graph below the line y=x shows the area where the similarity corresponding to x-axis performs better than the similarity corresponding to y-axis. It is the same for the section on the upper part of the line y = x which indicates the area where the distance represented in the y axis perform better.

In Figures 1. (b) the interpretation of points in the area is the opposite, the area that has more points spends more time in detecting similar subsequence with first subsequence.

Observing Figure 2 (a.1) it is evident that the performance of modified CID and classic CID is almost the same. Here we must emphases that modified CID has priority on classic CID in cases when the series appeared to have high complexity.

Comparing the average time of detection of the most repeated motif (Figure 2, b.1), even here in the case of time series with high complexity, modified CID has priority on classic CID. Chouakria index with Euclidean distance spends more time on finding similar subsequences compared to Chouakria index with CID. Also applied on Chouakria index with classic CID the average time is higher than the average time of Chouakria index with modified CID. Which makes the modified CID a better similarity measure used together with the Chouakria index.

We have also compared the elapsed time of classic CID, modified CID taking into consideration the length (*n*) of time series and length of subsequence (*m*).

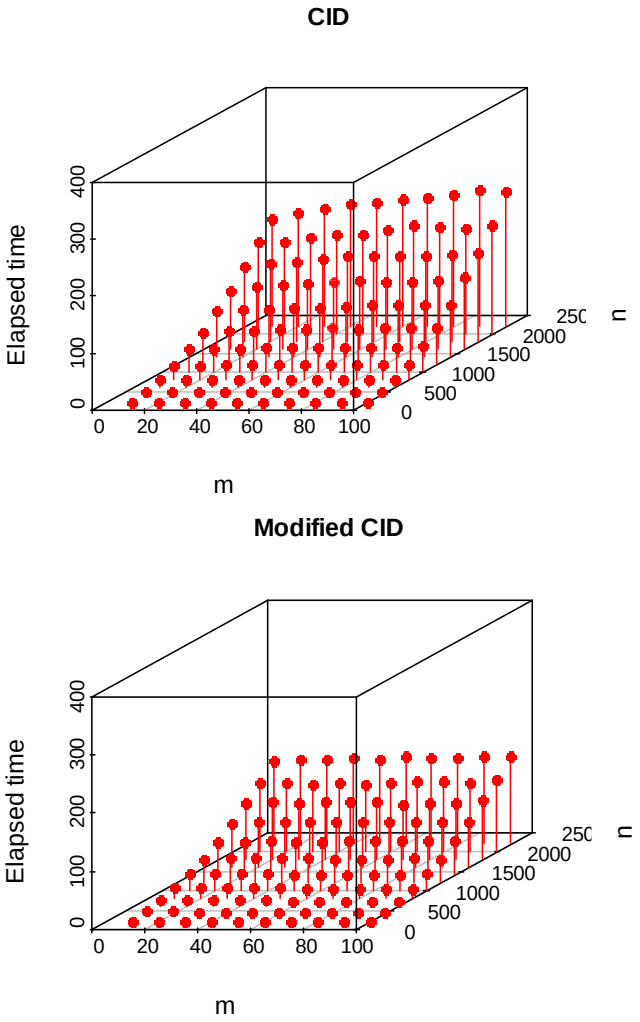


Fig. 3. *Execution time of classic CID and modified CID in motif discovery algorithm*

The average elapsed time of classic CID and modified CID differ clearly. The modified CID has an elapsed time nearly 20% lower than the classic CID.

V. RESULTS

In this work we have presented a modification of the Complexity Invariant Distance (CID) and tested its efficiency in a wide set of time series (engineering, meteorology, medicine, demography). Using the algorithm proposed by [7] which was modified in the work of [10] and in this work was adopted to the modified CID. The algorithm provides detailed information on the subsequence with higher repetition, including information on starting and ending indices of subsequence similar to test subsequence, numerical values of each of them, confidence interval for similarity calculated based on previously defined criteria. Graphical presentations are not missing, a graphic of all similar subsequence with the test subsequence along the time series and a special graphic with all subsequence similar to the first subsequence to understand more clearly the effectiveness of the algorithm and distance used.

Except the numerical results and graphical view in each case we also use as a comparative measure between these distances the number of motifs discovered by each of them and the execution time. We want to emphasize that these results are obtained from the work we have done with time series that have been available to us.

Tests show that modified CID offers satisfied results if it is combined with Chouakria index compared to modified CID.

VI. ACKNOWLEDGEMENT:

We want to thank Nertila Ismailaja for testing some of the results. Also we want to mention that many of the time series are taken from the following sites: http://robjhyndman.com/forecasting/data/, http://www.instat.gov.al/,

http://vincentarelbundock.github.io/Rdatasets/datasets.html

VII. REFERENCES

[1] Faloutsos C., Ranganathan M., Manolopoulos M., (1994). Fast subsequence matching in time-series databases, ACM SIGMOD Record 23 (2), 419-429

[2] Batista G., Wang X., Keogh E., (2011) A Complexity-Invariant Distance Measure for Time Series. SDM 2011

[3] Chiu, B. Keogh, E., & Lonardi, S. (2003). Probabilistic Discovery of Time Series Motifs. In the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 24 - 27, 2003. Washington, DC, USA. pp 493-498

[4] Mueen A., Keogh E., Zhu Q., Cash S., Westover B., (2009) Exact Discovery of Time Series Motifs, *In the Proceedings of SIAM International Conference on Data Mining, pp. 473-484,* SDM 2009.

[5] Mueen A., (2013) Enumeration of Time Series Motifs of All Lengths, In the Proceedings of IEEE International Conference on Data Mining, pp. 547-556, ICDM 2013. (94/809)

[6] Batista G., Keogh E., Tataw O. M, Vinícius M. A. de Souza (2014) CID: an efficient complexity-invariant distance for time series, Data Mining and Knowledge Discovery, May 2014, Volume 28, Issue 3, pp 634-669

[7] Dhamo (Gjika) E., Puka Ll. (2014) An algorithm for discovering similar subsequeces in time series data using CID (Complexity Invariant Distance), In the proceedings of Statistics Probability and Numerical Analysis International Conference, SPNA 2014, pp 82-88, ISBN 978-9928-4252-4-9

[8] Chouakria,A. D., Diallo A., Giroud F., (2007) Adaptive clustering of time series. International Association for Statistical Computing (IASC), Statistics for Data Mining, Learning and Knowledge Extraction, 2007, Aveiro, Portugal

[9] Chouakria A. D., Nagabhushan P. N. Adaptive dissimilarity index for measuring time series proximity, Published online: 31 January 2007, © Springer-Verlag 2007

[10] Dhamo (Gjika) E., Ismailaja N., Kalluçi E. (2015) Comparing the efficiency of CID distance and CORT coefficient for finding similar subsequences in time series, In the proceedings of 6th International Conference Information Systems and Technology Innovations: inducting modern business solutions. Tirana, June 5-6, 2015,

[11] C. Yan et al.(2013) An Approach of Time Series Piecewise Linear Representation Based on Local Maximum Minimum and Extremum , Journal of Information & Computational Science 10:9 (2013) 2747–2756