UNIVERSITÀ DELLA CALABRIA

Dipartimento di Economia e Statistica
Ponte Pietro Bucci, Cubi 0/C-1/C
87036 Arcavacata di Rende (Cosenza)
Italy
http://www.ecostat.unical.it/

Working Paper n. 05 - 2009

CLASSIFICATION OF SHORT TIME SERIES

Agostino Tarsitano
Dipartimento di Economica e Statistica
Università della Calabria
Ponte Pietro Bucci, Cubo 1/C
Tel.: +39 0984 492465
Fax: +39 0984 492421
e-mail: agotar@unical.it

Febbraio 2009

Pubblicazione depositata ai sensi della L. 106 del 15-4-2004 e del DPR 252 del 3-5-2006

Classification of short time series

Agostino Tarsitano
Dipartimento di Economia e Statistica
Università della Calabria
agotar@unical.it
Via Pietro Bucci, cubo 1C, Rende (CS) - Italy
Tel.: +39-0984-492465, Fax:+39-0984-492421

february 2009

Classification of short time series

Abstract: Many time series are of short duration because data acquisition has, of necessity, proceeded for but a brief term. Such data have previously often been analyzed by methods that either do not explicitly take into account time related changes or that are designed for long time series. In this paper, we consider several ways of assigning a dissimilarity between univariate time series in short term behavior. In particular, we have defined a measure that works irrespective of different baselines and scaling factors and its effectiveness has been evaluated on real and synthetic data sets.

Keywords: Time trajectories; Distances; PAM clustering; Representative trends.

1 Introduction

Short time series may be all there is available when data are acquired by an infrequent survey due to experimental factors or high costs. For instance, many economic series are internally comparable for very few periods and statistical estimates from such short series tend to be biased. The same is true for micro array data since technical equipment and methods of measurement change from time to time. This type of data is obviously undersampled, and some important features of the temporal pattern can be obscured by the stochastic noise.

Similarity-based mining of time-varying data has attracted an increasing attention in recent years. Applications where short time series occur range from biomedicine through macroeconomics, psycology statistics, to scientific discipline such us archeology, seismology and astronomy (**?**). Classic problems in handling short time series involve the clustering of such series into similar categories and the classification of new observed series into two or more known categories. These two problems, of course, are very common and there exists a vast literature on methods of discriminant and cluster analysis as applied to time independent observations. The basic idea is to extract distinctive features from the data, compare them and perform the grouping of the units into distinct categories. The clustering is satisfactory if the distance between units within clusters is relatively small compared with distances between clusters. Once the structure and the required number of clusters have been established, the cluster representatives can be employed to classify the old and new units using, for example, the nearest-centroid method.

Clustering methods can identify meaningful patterns even in time dependent observations; however, they have some limitations if standard algorithms are blindly applied measuring the closeness of the observed values, but ignoring the temporal dimension. The question of distance between short time series has been addressed previously in **??**. In this case, one wants to assign a value to the distance between individual time series rather than quantify the strength of relationship between the stochastic processes that generate the observations. For instance, dimensionality reduction techniques such as autoregressive and spectral representations, decomposition methods, discrete wavelet transforms and so forth are not useful in case of modest length sequences.

It must be considered that cluster analysis heavily relies on the concept of distance to map each comparison into a numerical value that quantifies the degree of proximity between two units. To address this issue, we have developed a simple computational technique to compare and classify relatively short time series (less than 25 time points). In Section 2, we will propose a distance function that takes into account both the observed values and the proximity between the temporal behavior of the sequences. As an illustrative example, a partitioning around medoids method is used to compare and cluster the relative consumer price index of OECD countries and the results are discussed in Section 3. In Section 4 we have compared our method with that of literature known methods. The procedures are assessed by a Monte Carlo simulation study. Conclusions and future research are then presented in Section 5.

2 A new metric for short time series

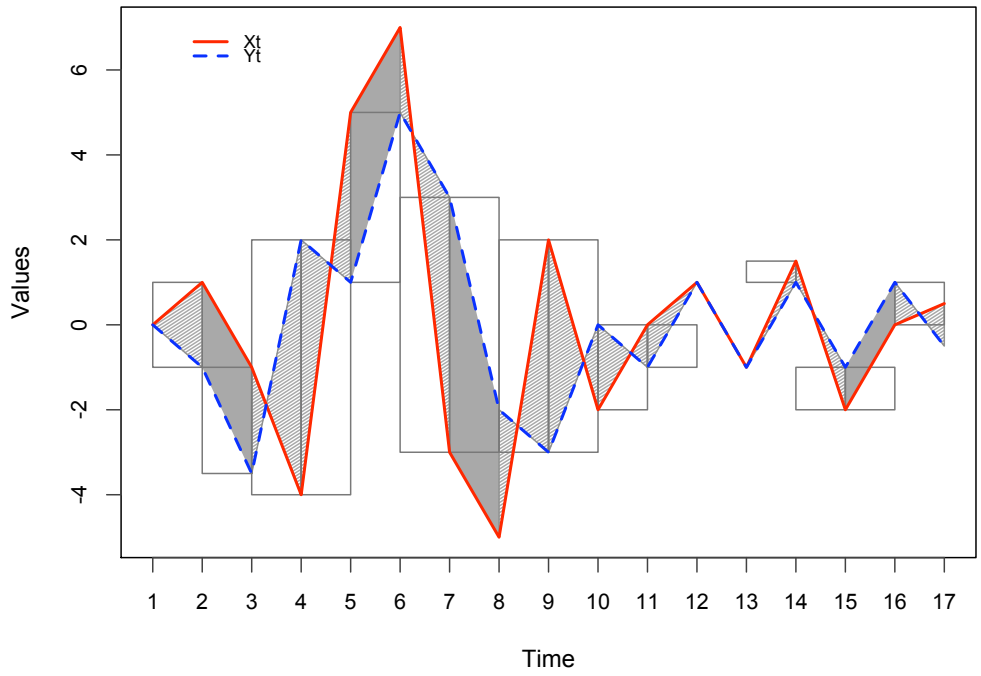
To analyze a set of time series $S = \left\{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p\right\}$ and to determine homology among them, we need to define an appropriate distance measure to describe quantitatively how closely the two sequences simulate each other. In this section we introduce a new method for performing distance search over time-varying data.

To present our approach, we consider time series in which the values have been standardized to have a mean of zero and a standard deviation of one. This type of invariance is useful for dealing with heterogeneity of scales and/or baseline shift when one considers meaningless to compare sequences with different levels and oscillations. Moreover, we restrict ourselves to deal with p different time series of equal length, that is, each element of $\mathbf{X}_i = (x_{i,t}, t = 1,2,\cdots,n)$ has a corresponding element in the matched sequence $\mathbf{X}_j = \left(x_{j,t}, t = 1,2,\cdots,n\right)$ for $i, j = 1,2,\cdots,p$ and the time points are equally spaced.

A time series $\mathbf{X}_i$ is represented as a continuous and non-self-intersecting polygonal curve linking n neighboring vertices $(t, x_{i,t}), t = 1,2,\cdots,n$. Our method is based on the observation that similar sequences will have a small area enclosed

between the polygonal curves representing them. Figure (1) illustrates the idea.

Figure 1: Area between two polygonals



The corresponding formula is

$$DA\left(\mathbf{X}_i,\mathbf{X}_j\right)=\frac{\sum_{t=1}^{n}A_i}{n} \tag{1}$$

where A_i is the area between the two polygons in the time interval $[t,t+1]$. The function essentially is an average of separation distances between two intervals. It could be seen that A_i is made of trapezoids or triangles. These are formed when the two polygons intersect and the intersection point belongs to $[t,t+1]$. In this case, the local contribution to the general distance is obtainted by summing the area of the two triangles (computed with the Erone's formula). On the other hand, the trapezoids are not necessarily cyclic so that their area can be computed with the Breithschneider's formula.

? shows that, since the lines forming the two polygons have their left and right end points on a vertical line, the area between them is a metric and this implies that DA satisfies the properties of reflexivity: $DA\left(\mathbf{X}_i,\mathbf{X}_j\right)=0$ iff $\mathbf{X}_i=\mathbf{X}_j$, and triangle inequality: $DA\left(\mathbf{X}_i,\mathbf{X}_j\right)+DA\left(\mathbf{X}_j,\mathbf{X}_k\right)\geq DA\left(\mathbf{X}_i,\mathbf{X}_k\right)$ for any $\mathbf{X}_k$.

The sum in (1) is commutative and, consequently, DA will not be able to distinguish between trajectories that have the same area distributed in a different way over the time domain (on this, see **?**). This is not always a drawback. For instance, there is no special reason to measure the resemblance between two time

series starting from the first vertex and ending to n-th, so that the distance should be the same both for start-to-finish sequences and from finish to start. The norm *DA* fulfils this requirement. However, for a distance function to be useful in the context of time series, it should be driven by the relative change of intensity in a given interval as well as the time order of observations (**?**).
To incorporate both, a time weighted distance function is described below

$$WDA\left(\mathbf{X}_i, \mathbf{X}_j\right) = \frac{\sum_{t=1}^{n} w_t A_t}{\sum_{t=1}^{n} w_t} \tag{2}$$

where

$$w_t = 0.5\left(\min\left\{|x_{i,t} - x_{j,t}|, |x_{i,t+1} - x_{j,t+1}|\right\} + \max\left\{|x_{i,t} - x_{j,t}|, |x_{i,t+1} - x_{j,t+1}|\right\}\right) \tag{3}$$

for $t = 1, 2, \cdots, n-1$. Each boxes in Figure (1) represents a time-varying weight assigned to what happens in the interval $[t, t+1]$; the larger is a box, the more important is the role of the corresponding local distance within the global distance function, while irrelevant differences have virtually zero weight. Of course, the weights in (2) are normalized by dividing them for the total area so that they are non negative and sum to one. The range of WDA spans from a minimum of zero to the max A_i. Practically, the smaller the value of WDA the greater is the similarity between the time sequences under comparison.

It can be easily shown that the function introduced in (2) is a metric by using the theorem given in the appendix of **?**.

3 Cluster analysis

The aim is to partition a set of objects into groups $C_1, C_2, \cdots, C_k$ in such a way that objects (in our case, time series) in the same cluster are near and similar, whereas objects in different clusters are distinct and distant. Thus, the distance function that quantifies the dissimilarity between units, is crucial in determining the outcome of any cluster analysis. A large majority of techniques for cluster analysis are designed for units that are described by means of numerical feature vectors. On the other hand, applications to time-varying data require the clustering of items that are defined by the relationships between individual data (distances).

Over the years, many methods have been devised to find groups within observed time series. We chose the partitioning around medoids method (PAM) proposed by **?** for several reasons. First, the typical representative of each group (the cluster medoid) is the most centrally located item in a cluster, that is, the item in the cluster whose average dissimilarity to all other items in the same cluster

is minimal. In practice, the medoid is one of the observed time series and, as such, it is not an artificial object created in the virtual environment of the algorithm. Second, it can operate directly on a distance matrix as long as it is metric in the sense of **?**. In fact, the computation of cluster medoids does not require the presence of feature vectors, but can be done for a distance matrix. Third, it is a partitional algorithm that does not impose a hierarchical structure, which is not necessarily present in the underlying hypothetical population. Fourth, rather than selecting starting centers at random, PAM evaluates all possible starting centers and chooses the best centers to start cluster building. This gives consistent results when clustering is repeated. Finally, PAM has been shown to be both more robust to inclusion of outliers than the popular k-means method (**?**) because it uses the most centrally located object in a cluster

The PAM procedure is based on the search of k representative time series (medoids) that minimize the overall sum of distances of the sample units to their closest representative unit **?** (in this method, the analyst has to decide in advance the number of clusters). More specifically, the method starts by choosing k medoids such that the total distance of all units to their nearest medoid is minimal, *i.e.*, the algorithm finds a subset $\{\mathbf{M}_1, \mathbf{M}_2, \cdots, \mathbf{M}_k\} \in S$ which minimizes the function

$$\sum_{i=1}^{p} \min_{j=1,\cdots,k} WDA\left(\mathbf{X}_i, \mathbf{M}_j\right) \tag{4}$$

Each $\mathbf{X}_i$is then assigned to the category corresponding to the nearest medoid. That is, the time trajectory $\mathbf{X}_i$ is assigned to cluster C_j whose associated medoid, $\mathbf{M}_j$, is nearest to $\mathbf{X}_i$

$$WDA\left(\mathbf{X}_i, \mathbf{M}_j\right) \leq WDA\left(\mathbf{X}_i, \mathbf{M}_r\right) \text{ for all } r = 1, 2, \cdots, k \tag{5}$$

The result of PAM is a partition of the original set in k categories each formed by time sequences with similar dynamics, and each guided by a leading sequence that explains how the group is shaped, how it evolves over time and what sequences we are to expect to be included in the group. In addition, PAM returns for each unit a silhouette width that reflects how well the particular unit is clustered.

3.1 Case study

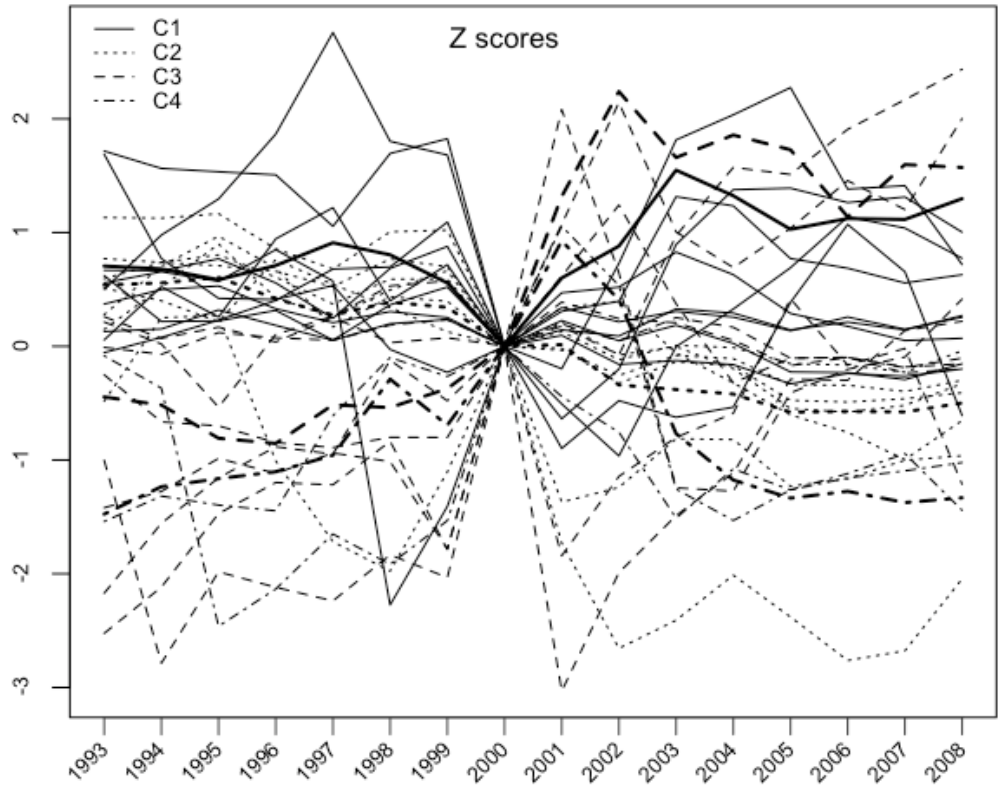
To demonstrate the efficiency of the proposed method in clustering of short-run dynamics, we performed experiments using a real data set. In particular, we used the relative consumer price index, 2000 = 100 (downloadable from http://webnet.oecd.org/wbos/index.aspx). This dataset is a collection of time series presenting the relative CPI of $p = 31$ states for the period of 1993–2008

($n = 16$). We have applied the metrics (2) combined with the PAM algorithm available in the R-package *cluster* to the standardized data. The number of clusters was selected based on the average silhouette width. The following groups have been established

Cluster 1 $= \{AUS, CAN, DNK, GRC, IRL, KOR, LUX, NLD, NZD, ESP, EUR\}$
Cluster 2 $= \{AUT, BEL, FIN, FRA, DEU, JAP, SWE, CHE\}$
Cluster 3 $= \{CZE, HUN, ICE, ITA, NOR, POL, PRT, SVK, TUR\}$
Cluster 4 $= \{MEX, GBR, USA\}$

Figure (2) shows the patterns

Figure 2: Temporal patterns of RCPI for OECD countries



Clusters 1 groups together sequences that oscillate steadily with time; cluster 2 consists mainly of the states showing a decreasing trend whereas most of the sequences belonging to cluster 3 exhibit an overall tendency to increase; cluster 4 is distinct from the others by grouping three states that have a slowly increasing

behavior over the full range of measurements. The leading indicators or representative trends of the various clusters are the time series of IRE (C1), AUT (C2), HUN (C3) and USA (C4) and are indicated with a more marked line.

4 Alternative metrics for short time series

In **?** it is advocated the comparison of a new proposal to other existing techniques. Many other measures of similarity between time series have have been introduced in literature. To undertake this task we have examined some distance functions that have assumed a definite and independent form.

The simplest distance function can be expressed by the Minkowsky formula

$$M_a\left(\mathbf{X}_i,\mathbf{X}_j\right)=\left\{\sum_{t=1}^{n}\left(x_{i,t}-x_{j,t}\right)^a\right\}^{\frac{1}{a}} \tag{6}$$

where $a>1$ is a parameter to be given (Manhattan for $a=1$, Euclidean for $a=2$, Tchebycheff for $a\rightarrow\infty$, putting different emphasis on large deviations). The function (6) has the advantage of being easy to compute and interpret. However, the Minkowky metrics ignore the temporal structure of the values as resemblance is only based on the differences between the values, independently of the increase/decrease behavior before and after these values. As a consequence, if we believe that two profiles should be considered more similar if they have the same temporal pattern, then the use of $M_a\left(\mathbf{X}_i,\mathbf{X}_j\right)$ will waste some important information.

It can be noted that if the variables are standardized then the Euclidean metric, obtained from (6) for $a=2$, is functionally related to the correlation metric

$$M_2\left(\mathbf{X}_i,\mathbf{X}_j\right)=\sqrt{2n\left[1-r\left(\mathbf{X}_i,\mathbf{X}_j\right)\right]} \tag{7}$$

where $r\left(\mathbf{X}_i,\mathbf{X}_j\right)$ is the Pearson's cross-product correlation coefficient. The more series are linearly related, the smaller is the Euclidean distance between them after the standardization. The logcorrelation metric $-\log\left(0.5\left[1+r\left(\mathbf{X}_i,\mathbf{X}_j\right)\right]\right)$, common in biological studies, can be derived from (7). It is unlikely, however, that clusterings obtained by these methods will show readily observable differences between them.

The strength of monotonic association between two set of values can be measured by a rank-based measure of correlation: the Kendall's concordance coefficient

$$\tau\left(\mathbf{X}_i,\mathbf{X}_j\right)=1-\frac{2d_k\left(\mathbf{X}_i,\mathbf{X}_j\right)}{n\left(n-1\right)} \tag{8}$$

where $d_k\left(\mathbf{X}_i,\mathbf{X}_j\right)$ can be interpreted as the symmetric difference distance between the ordered values of $\mathbf{X}_i$ and $\mathbf{X}_j$ and used as a metric to classify time series. The Kendall's τ gives the probability that any two corresponding pairs of values in the two vectors are identically ordered. Of course, the order taken into account by τ is the ranking of values in relation to other values and clearly not in relation to the temporal characteristics. Whereas the Person's correlation coefficient is appropriate mainly for indicating linear association, the Kendall's τ is invariant under increasing monotone transformations.

To overcome the problems of the existing measure of association in different *a priori* unknown situations, **?** proposed the order statistics correlation coeffifient

$$2\rho_k\left(\mathbf{X}_i,\mathbf{X}_j\right)=\frac{\sum_{t=1}^{n}\left[x_{i,(t)}-x_{i,(n+1-t)}\right]x_{j,[t]}}{\sum_{t=1}^{n}\left[x_{i,(t)}-x_{i,(n+1-t)}\right]x_{j,(t)}}+\frac{\sum_{t=1}^{n}\left[x_{j,(t)}-x_{j,(n+1-t)}\right]x_{i,[t]}}{\sum_{t=1}^{n}\left[x_{j,(t)}-x_{j,(n+1-t)}\right]x_{i,(t)}} \tag{9}$$

where $x_{i,(t)}$ is the t-th order statistics of $\mathbf{X}_i$ and $x_{j,[t]}$ is the corresponding value of $\mathbf{X}_j$, that is, the value in the t-th position of $\mathbf{X}_j$ when the pairs $\left(\mathbf{X}_i,\mathbf{X}_j\right)$ is sorted with respect to the magnitude of $\mathbf{X}_i$. The coefficient $\rho_k\left(\mathbf{X}_i,\mathbf{X}_j\right)$ has the basic properties ofa correlation coefficient, but it is more robust against noise than the Pearson's, and more sensitive than the Kendall's (see **?**). A distance function based on (9) can be calculated using $\sqrt{2\left[1-\rho_k\left(\mathbf{X}_i,\mathbf{X}_j\right)\right]}$.

The resemblance between polygonals can also be measured by a discrete version of the Fréchet distance (**?**). The discrete Fréchet distance is similar to the dynamic time warping distance (DTW). It must be observed that, in our context, the Fréchet distance coincides with $M_\infty\left(\mathbf{X}_i,\mathbf{X}_j\right)$ so that we could expect results similar to the Euclidean metrics.

One shortcoming of (6) is that it ignores sequential or temporal characteristics. To improve its ability to assess the proximity between dynamics **??** added to $M_a\left(\mathbf{X},\mathbf{Y}\right)$ the distance between the first differences (velocities) and the distance between the second differences (accelerations). This idea can slightly be extended if we incorporate knowledge about higher differences

$$V_{a,q}\left(\mathbf{X}_i,\mathbf{X}_j\right)=\frac{1}{q}\sum_{k=1}^{q}\sum_{t=k}^{n}\left\{\frac{M_a\left[\mathbf{X}_i^{(k)},\mathbf{X}_j^{(k)}\right]}{M_{a,max}^{(k)}-M_{a,min}^{(k)}}\right\} \tag{10}$$

where $M_{a,q}\left[\mathbf{X}_i^{(k)},\mathbf{X}_j^{(k)}\right]$ for $k=1,2,\cdots,q$ is the Minkowky distance for the original values ($k=0$), first differences ($k=1$) of two consecutive time points, second difference ($k=2$), and so on. Symbols $M_{a,max}^{(k)}$ and $M_{a,min}^{(k)}$ denote, for each k, the Minkowky distance of the furthest and closest pair of sequences, respectively. Thus, $V_{a,q}$ has values between zero and one. For $a=2,q=3$ expression (10)

yields the distance function proposed by **?** and for $a = 2, q = 2$ the distance function used by **?**. The measure considered by **?** can easily be derived from (10).

? present a simple metric able to exploit the temporal structure of time series:

$$L_p\left(\mathbf{X}_i, \mathbf{X}_j\right) = \left[\sum_{t=1}^{n} \left(B_{t-1} + B_{t+1}\right)^p\right]^{\frac{1}{p}} \tag{11}$$

with

$$2B_{t-1} = \begin{cases} \Delta_{i,j}^{(t)} & if\ \Delta_{i,j}^{(t)}\Delta_{i,j}^{(t-1)} \leq 0 \\ \dfrac{\Delta_{i,j}^{(t)}}{\Delta_{i,j}^{(t)} + \Delta_{i,j}^{(t-1)}} & \text{otherwise} \end{cases} \tag{12}$$

$$2B_{t+1} = \begin{cases} \Delta_{i,j}^{(t)} & if\ \Delta_{i,j}^{(t)}\Delta_{i,j}^{(t+1)} \leq 0 \\ \dfrac{\Delta_{i,j}^{(t)}}{\Delta_{i,j}^{(t)} + \Delta_{i,j}^{(t+1)}} & \text{otherwise} \end{cases} \tag{13}$$

where $\Delta_{i,j}^{(t)} = x_{i,t} - x_{j,t}$. This norm involves the area of the triangles located on the left and right sides of each coordinate.

? observed that two time series that are non-linearly related to each other will result distant in terms of a correlation based distance measure, regardless of their similarity. In this sense they suggested using a qualitative difference between $\mathbf{X}_i$ and $\mathbf{X}_j$ by summing up the scores

$$QD = \frac{0.5}{n\left(n-1\right)} \sum_{r<s} \delta_{r,s}\left(\mathbf{X}, \mathbf{Y}\right) \tag{14}$$

where

$$\delta_{i,j}\left(\mathbf{X}, \mathbf{Y}\right) = \begin{cases} 0 & \text{if } w_{r,s}\left(\mathbf{X}\right) = w_{r,s}\left(\mathbf{Y}\right) \\ \left|w_{r,s}\left(\mathbf{X}\right)\right| + \left|w_{r,s}\left(\mathbf{X}\right)\right| & \text{otherwise} \end{cases} \tag{15}$$

with $w_{r,s}\left(\mathbf{X}\right) = sign\left(x_s - x_r\right)$ and $w_{r,s}\left(\mathbf{Y}\right) = sign\left(y_s - y_r\right)$.

4.1 Comparison of the procedures

We have evaluated the performance of the proposed technique by considering artificially generated sample time sequences. The test data has been designed specially for this purpose so as to include a variety of curves to demonstrate the effectiveness of the new distance function. In particular, the data sets were created using a random time series generator that produces $p = 48$ sequences belonging to a predefined number $k \in K = \{4, 8, 12\}$ clusters $C_1, C_2, \cdots, C_k$ of length $n \in N = \{11, 17, 23\}$. Each cluster includes a number of $48/k$ sequences given by the following expressions

$$
\begin{aligned}
1 &: (\phi_1 t + \phi_2)\sin(2\pi\mathbf{t}) + \mathbf{e} & 2 &: (\phi_1 t + \phi_2)\cos(2\pi\mathbf{t}) + \mathbf{e} \\
3 &: (\phi_1 t + \phi_2)\sin(2\pi\mathbf{t} - \pi/4) + \mathbf{e} & 4 &: (\phi_1 t + \phi_2)\cos(2\pi\mathbf{t} - \pi/4) + \mathbf{e} \\
5 &: (\phi_1 t + \phi_2)\sin(4\pi\mathbf{t}) + \mathbf{e} & 6 &: (\phi_1 t + \phi_2)\cos(4\pi\mathbf{t}) + \mathbf{e} \\
7 &: (\phi_1 t + \phi_2)\sin(2\pi\mathbf{t} - t) + \mathbf{e} & 8 &: (\phi_1 t + \phi_2)\cos(2\pi\mathbf{t} - t) + \mathbf{e} \\
9 &: (\phi_1 t + \phi_2)\sin(4\pi\mathbf{t} - t) + \mathbf{e} & 10 &: (\phi_1 t + \phi_2)\cos(4\pi\mathbf{t} - t) + \mathbf{e} \\
11 &: (\phi_1 t + \phi_2)\sin(4\pi\mathbf{t} - \pi/4) + \mathbf{e} & 12 &: (\phi_1 t + \phi_2)\cos(4\pi\mathbf{t} - \pi/4) + \mathbf{e}
\end{aligned}
$$

where $t \in [-0.4, 0.4, by\ 0.8/(n-1)]$. The parameter ϕ_1 is obtained randomly from $U(0.1, 0.5)$, ϕ_2 randomly from $U(1.1, 1.5)$, and $\mathbf{e}$ is a vector of random errors form $N(0, \sigma = 1.5)$. All the quantities ϕ_1, ϕ_2, and $\mathbf{e}$ varies across time series.

To compare the stability of the above algorithms, we repeated the data generation 1000 times with this organization for each k and for each n. Since we know the cluster that each sequence belongs to, we can use the true clustering membership and assess the quality of the agreement between the true and the resulting cluster membership. The Adjusted Rand Index (ARI) has been used as cluster validation measure (**??**). As it is well known, the ARI takes values on $[-1, 1]$ with one indicating a perfect concordance between the known and produced partition and values near zero or negative corresponding to cluster agreement found by chance. The numerical summaries are given in Table 1.

Each entry shows the average ARI of the true clustering with clusters using the various metrics. A high average adjusted Rand index means that clusters determined by the algorithm are in strict agreement with clusters from the original data. A low value of the standard deviation of ARI indicates stability across the simulations.

Table 1: Summary of ARI from 1000 simulations for some metrics

k	n		WAD	Eucl.	$V_{2,2}$	$V_{2,3}$	τ	ρ_k	QD	L_2
4	11	Mean	0.928	0.915	0.723	0.655	0.838	0.906	0.837	0.662
	11	Std.Dev.	0.062	0.068	0.123	0.134	0.091	0.074	0.092	0.116
	17	Mean	0.992	0.991	0.883	0.815	0.965	0.984	0.965	0.890
	17	Std.Dev.	0.020	0.022	0.086	0.117	0.046	0.032	0.046	0.078
	23	Mean	0.999	0.998	0.943	0.883	0.994	0.996	0.993	0.977
	23	Std.Dev.	0.009	0.010	0.060	0.089	0.020	0.015	0.020	0.037
8	11	Mean	0.662	0.656	0.589	0.573	0.621	0.638	0.621	0.506
	11	Std.Dev.	0.051	0.054	0.074	0.078	0.060	0.055	0.062	0.077
	17	Mean	0.716	0.707	0.681	0.662	0.694	0.700	0.695	0.633
	17	Std.Dev.	0.041	0.037	0.065	0.078	0.047	0.038	0.045	0.063
	23	Mean	0.745	0.726	0.732	0.725	0.721	0.722	0.721	0.695
	23	Std.Dev.	0.051	0.039	0.061	0.076	0.043	0.034	0.043	0.035
12	11	Mean	0.596	0.577	0.508	0.485	0.530	0.566	0.530	0.430
	11	Std.Dev.	0.059	0.060	0.073	0.076	0.066	0.056	0.066	0.073
	17	Mean	0.677	0.646	0.608	0.588	0.633	0.644	0.632	0.553
	17	Std.Dev.	0.053	0.045	0.062	0.072	0.052	0.048	0.051	0.060
	23	Mean	0.726	0.682	0.665	0.658	0.676	0.683	0.673	0.610
	23	Std.Dev.	0.059	0.046	0.058	0.067	0.051	0.046	0.050	0.037
All		Av. Mean	0.782	0.766	0.704	0.672	0.741	0.760	0.741	0.662
		Av. St.Dev.	0.045	0.042	0.074	0.087	0.053	0.044	0.053	0.064

All the calculations are carried out by the statistical language R, **?**, and the programs are available by the authors.

The following facts emerge from Table 1:

1. The mean ARI increases and the standard deviation of ARI decreases as the length of the sequences increases. This is in line with the principle that longer sequences allow progressively more detailed description of the underlying dynamics.

2. The mean ARI decreases as the standard deviation of ARI increases as the number of clusters increases. Such a result should not be considered as a surprising one; in fact, more groups imply higher chances of misclassification.

3. Our method, WDA (Weighted Area Distance) has the best performance among the eight alternatives; in fact, the average of the mean ARI is 0.782 (the highest) and the average of the standard deviations is 0.045 (the second lowest-level). However, its results are very similar to those of the Euclidean distance; in practice, WAD does not bring massive improvement over the current State of the Art. At the very least, one should wonder if it just makes sense to introduce a product if it is only slightly better than the current best.

4. Many distance functions currently in use, perform worse than the Euclidean distance. Perhaps, empirical evaluations in the past have often been inadequate (**?**); however, more discussions regarding the relative merits and demerits among of these approaches can be found in the referred articles.

5 Conclusions and Future Work

Dissimilarity search in short time series data is a rich and rapidly growing research field prevailingly devoted to the methods of cluster and discriminant analysis applied to dated information. The problem of distance quantification in time-varying data is extensively studied and various pairwise distance function have been proposed in the literature.

In this paper, a new metric was designed in the hope that it would be able to detect similarities in the dynamic of short time series where conventional statistical time series analysis methods are not applicable. The metric is sensible both to the observed values and to the rate of change. Group of sequences where searched by means of the PAM algorithm. A case study on relative consumer price index of OECD countries has been undertaken. Also, it has been verified by the help of test data that the proposed technique has a good performance with short time series, although, the quality of the results it is not significantly superior to that achievable by using the simple Euclidean distance between the observed data. We have also discussed different propositions for a distance measure between short-run sequences concluding that

In this paper we have assumed that a time series comprises of samples of a single measured variable against time. In future work, we intend to broaden its scope so that it can handle multivariable time sequences. Another desirable generalization of the approach taken in this paper is to allow a smooth change between dynamical systems considered similar. However, the construction of a reliable measure of distance with that property is still an open problem and subject to ongoing research

References

Dudoit S. and Fridlyand J. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, 97, 77–87.

D'Urso P. (2000) Dissimilarity measures for time trajectories, *Journal of the Italian Statistical Society*, 9, 53–83.

D'Urso P. and Vichi M. (1998) Dissimilarities between trajectories of a three-way longitudinal data set, in: *Advances in data science and classification.*, Rizzi A., Vichi M. and Bock H.H., eds., Springer, Berlin, 585–592.

Gower J.C. and Legendre P. (1986) Metric and Euclidean properties of dissimilarity coefficients, *Journal of Classification*, 3, 5–48.

Hubert L. and Arabie P. (1985) Comparing partitions, *Journal of Classification*, 2, 193–218.

Kaufman L. and Rousseeuw P. (1990) *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons, New York.

Keogh E. and Kasetty S. (2002) On the need for time series data mining benchmarks: a survey and empirical demonstration, in: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002,Edmonton, Alberta, Canada*, 102–111.

Kundu S. (2006) Conflating two polygonal lines, *Pattern Recognition*, 39, 363–372.

Lee J. and Verleysen M. (2005) Generalization of the l_p norm for time series and its application to self-organizing maps, in: *WSOM 2005, Workshop on Self-Organizing Maps*, Paris (France), 733–740.

Liao T.W. (2005) Clustering of time series data - a survey, *Pattern Recognition*, 38, 1857–1874.

Ljubič P., Todorovski L., Lavrač N. and Bullas J. (2002) Time-series analysis of UK traffic accident data, in: *Proceedings of Multi-conference Information Society, 14-18 Oct. 2002*, M. Bohanec M. Gams D.M.M.G., ed., Springer-Verlag, Ljubliana, Slovenija, 135–138.

Milligan G.W. and Cooper M.C. (1986) A study of the comparability of external criteria for hierarchical cluster analysis, *Multivariate Behavioral Research*, 21, 441–458.

Mosig A. and Clausen M. (2005) Approximately matching polygonal curves with respect to the Fréchet distance, *Comput. Geom. Theory Appl.*, 30, 113–127.

R Development Core Team (2008) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Sieburg H. and Müller-Sie C. (2004) Classification of short kinetics by shape, *In Silico Biology*, 4, 1–8.

Song J. (2008) A new dissimilarity measure in time-dependent experiments, *Journal of the Korean Statistical Society*, 37, 145–153.

Tsai G. and Qu A. (2008) Testing the significance of cell-cycle patterns in time-course microarray data using nonparametric quadratic inference functions, *Computational Statistics & Data Analysis*, 52, 1387–1398.

Xu W., Chang C., Hung Y. and Kwan S.K. F.P. (2006) Order statistic correlation coefficient and its application to association measurement of biosignals, in: *ICASSP 2006 Proceedings*, Toulouse (France), II 1068–II 1071.

Xu W., Chang C., Hung Y. and Kwan S.K. F.P. (2007) Order statistics correlation coefficient as a novel association measurement with applications to biosignal analysis, *IEEE Transactions on Signal Processing*, 55, 5552–5563.