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

March 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 value to the dissimilarity between univariate time series in short term behavior. In particular, we define a distance function that works irrespective of different baselines and scaling factors and its effectiveness is evaluated on real and synthetic data sets.

2000 Mathematics Subject Classification: 62H30, 62M10, 32F45.
Keywords: Time trajectories; Distances; PAM clustering; Representative trends.

1 Introduction

Short sequences (less than $20/25$ time points) may be all there is available when data are acquired by an infrequent survey due to experimental factors and/or high costs or due to methodological changes, or interruption of survey administration or data access. This type of data is obviously undersampled, and some important features of the temporal pattern can be obscured or hidden by the stochastic noise. 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.

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, psychology, demography to scientific discipline such us archeology, seismology and astronomy (Sieburg and Müller-Sie [2004]). Classic problems involve clustering of short-run time series into similar groups and classification of new observed series into two or more known groups. These two problems, of course, are very common and there exists a vast literature on methods of cluster and discriminant analysis as applied to time independent observations. The basic idea is to extract distinctive features from data, compare them and perform the grouping of the objects into distinct categories in such a way that objects in the same cluster are near and similar, whereas objects in different clusters are distinct and distant; once the structure and the required number

of clusters have been established, the cluster representatives can be employed to classify old and new objects.

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 D'Urso [2000] and Liao [2005]. In this case, one wants more assigning a value to the distance between observed sequences rather than quantifying the strength of relationships between the stochastic processes that generate observations. Consequently, techniques such as autoregressive and spectral representations, decomposition or subsequence matching methods, discrete wavelet transforms and so forth are not useful in case of short length sequences.

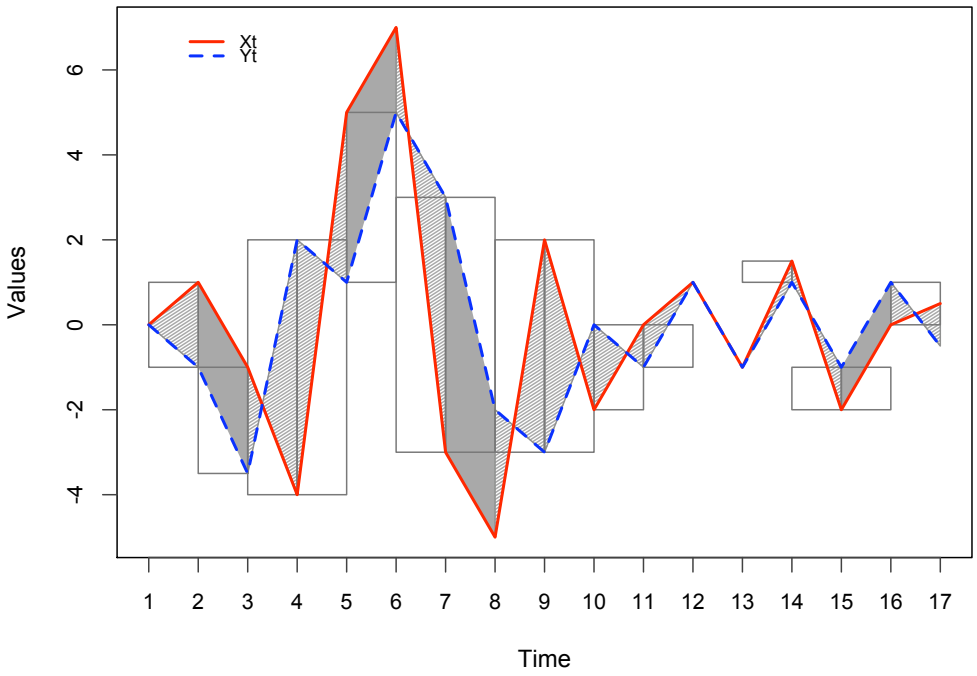
The distance function that quantifies the proximity between objects (in our case, short-run time series) is crucial in determining the outcome of any cluster analysis. To address this issue, we have developed a simple computational technique, based on geometrical concepts, to compare and classify relatively short time series. In Section 2, we will propose a distance function that takes into account both the observed values and the temporal behaviors of the sequences. As an illustrative example, a partitioning around medoids method is used to cluster the relative consumer price index of OECD countries and the results are discussed in Section 3. In the next section we have compared our method with some well-known classification methods of time series. 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 p different time series $S = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p\}$ and to determine homology among them, we need to define an appropriate distance measure to describe quantitatively how closely two of them simulate each other. To present our approach, we consider time series of equal length n, that is, each element of $\mathbf{X}_i = (x_{i,t}, t = 1,2,\cdots,n;\ n > 2)$ 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. Each time series was standardized to have a mean of 0 and a standard deviation of 1 because this type of invariance is useful for discarding heterogeneity of scales and/or baseline shifts.

A time series $\mathbf{X}_i$ can be represented as a continuous and non-self-intersecting polygonal curve linking n neighboring vertices $\left(t, x_{i,t}\right), t = 1,2,\cdots,n$. 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



To a first approximation, we define a distance function using the simple average of the separations between intervals:

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

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

Kundu [2006] shows that, since the lines forming the polygons have their left and right end points on a vertical line, the area between them is a metric and this, in turn, implies that the sum in (1) satisfies the properties of

reflexivity $DA\left(\mathbf{X}_i, \mathbf{X}_j\right) = 0 \quad$ iff $\quad \mathbf{X}_i = \mathbf{X}_j$

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 $\mathbf{X}_k \in S$.

Formula (1) is commutative and, consequently, $DA\left(\mathbf{X}_i, \mathbf{X}_j\right)$ 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 Lee and Verleysen [2005]). This is not always a drawback. For instance, there is no special reason to measure the resemblance between 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 metric $DA\left(\mathbf{X}_i, \mathbf{X}_j\right)$ fulfils this requirement.

However, to be useful and appropriate in the context of time series, a metric should be driven by the relative change of intensity in a given interval as well as the time order of observations (Liao [2005]). To incorporate both factors, we defined

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

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 the 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 to the range of zero to one so that WAD spans from a minimum of zero to max $A_t, t = 1, 2, \cdots, n$. Practically, the smaller the value of WAD, the greater the similarity between the time series under comparison.

It can be easily shown that even the function introduced in (2) is a metric by using the theorem given in the appendix of Kundu [2006].

3 Cluster analysis

A large majority of techniques for cluster analysis is designed for objects described by a numerical feature vector. On the other hand, applications to time-varying data require a clustering determined by the relationships between individual objects, that is, distances and, more precisely, distances depending on the temporal characteristics of the objects.

Over the years, many methods have been used to find groups within time series. We chose the partitioning around medoids method (PAM) proposed by Kaufman and Rousseeuw [1990] for several reasons. First, the typical representative of each group (the medoid) is the most centrally located object *i.e.* the object in the cluster whose average dissimilarity to all other objects in the same cluster is minimal. In effect, 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 the matrix of the distances is metric in the sense of Gower and Legendre [1986]. Third, it is a partitional algorithm that does not impose a hierarchical structure, which could not be necessarily present in $S = \{\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_p\}$. Fourth, rather than selecting starting centers at random, PAM evaluates all possible starting centers and chooses the best centers to start cluster building and 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 each cluster (see Dudoit and Fridlyand [2002]).

The PAM procedure is based on the search of k representative objects that minimize the overall sum of distances of the sample units to their closest representative object Kaufman and Rousseeuw [1990] belonging to one of $C_1, C_2, \cdots, C_k$ (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 objects 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

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

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

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

The result of PAM is a partition of the original set S of p sequences in k categories, each formed by 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 yields for each sequence a silhouette width that reflects how well the particular sequence is clustered.

3.1 Case study

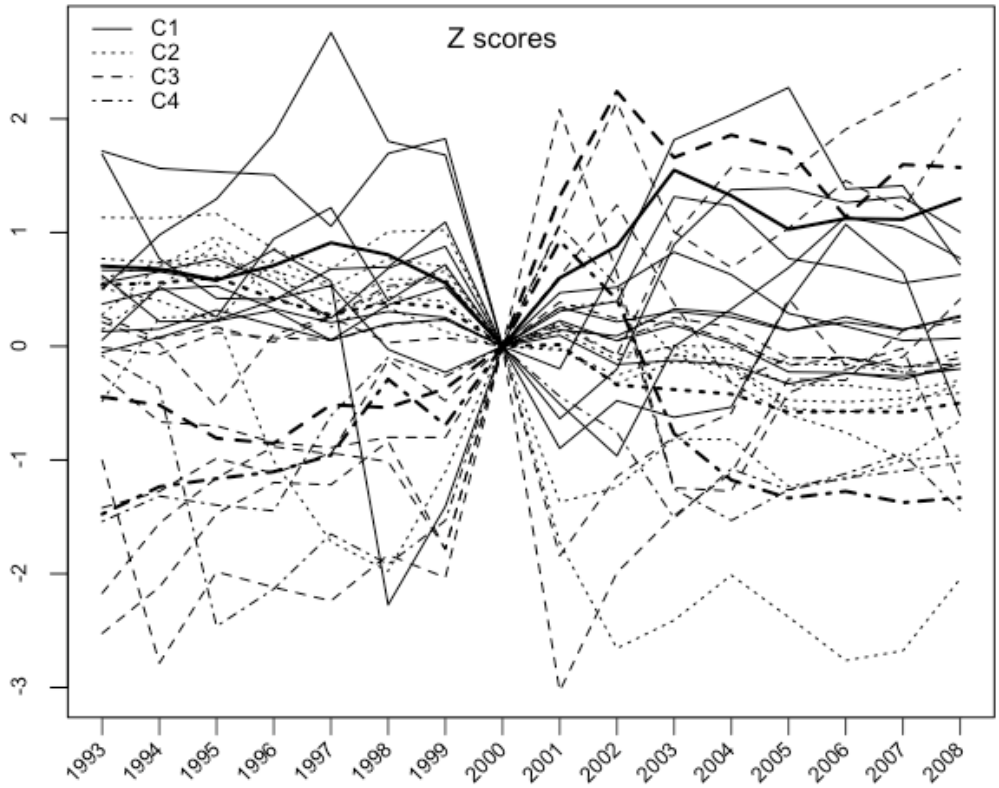
To demonstrate the efficiency of the proposed method in clustering of short-run sequences, we performed an experiment using a real data set. In particular, we used the relative consumer price index, 2000 = 100 (downloadable from `http://webnet.oecd.org/wbos/index.aspx`). This dataset is a collection of time series presenting the relative CPI of $p = 31$ states for the period of 1993–2008 ($n = 16$). We have applied (2) combined with the PAM algorithm 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 found by the algorithm.

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

Many other measures of dissimilarity between time series have been introduced in literature. In Keogh and Kasetty [2002] it is wisely advocated the comparison of each new proposal to other existing techniques. To undertake this task we have examined some distance functions that had already assumed a definite and independent form.

4.1 Minkowsky and correlation metrics

The simplest distance function between time series is the Minkowsky metric

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

where $a>1$ is a parameter to be fixed according to the emphasis to be put on large deviations. The function (6) has the advantage of being easy to compute and interpret: the classical Euclidean distance is found for $a=2$ and the Manhattan metric for $a=1$; the discrete Fréchet distance, often invoked in this context (see Mosig and Clausen [2005]), coincides with $M_\infty\left(\mathbf{X}_i,\mathbf{X}_j\right)$ so that we could expect results similar to the Euclidean metric.

It can be noted that when the variables are standardized then the Euclidean metric, 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 the sequences are linearly related, the smaller is the Euclidean distance between them. The logcorrelation metric $-\log\left(0.5\left[1+r\left(\mathbf{X}_i,\mathbf{X}_j\right)\right]\right)$, common in biological studies, is a monotone transformation of (7).

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

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

where $d_\tau\left(\mathbf{X}_i,\mathbf{X}_j\right)$ 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. Whereas the Person's correlation coefficient is appropriate mainly for indicating linear association, the Kendall's τ is invariant under increasing monotone transformations. Of course, the order taken into account by $\tau\left(\mathbf{X}_i,\mathbf{X}_j\right)$ is the ranking

of values in relation to other values and clearly not in relation to the temporal characteristic of the sequences.
Xu [2006] proposed the order statistics correlation coefficient

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

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 $\rho\left(\mathbf{X}_i,\mathbf{X}_j\right)$ has the basic properties of a correlation coefficient, but it is more robust against noise than the Pearson's, and more sensitive than the Kendall's (see Xu [2007]). A distance function based on (9) can be calculated using $d_\rho\left(\mathbf{X}_i,\mathbf{X}_j\right)=\sqrt{2\left[1-\rho\left(\mathbf{X}_i,\mathbf{X}_j\right)\right]}$.

Ljubič et al. [2002] 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 particular, they suggested using a qualitative difference between $\mathbf{X}_i$ and $\mathbf{X}_j$ by summing up a special set of scores:

$$QD = \frac{0.5}{n(n-1)}\sum_{r<s}\delta_{r,s}(\mathbf{X},\mathbf{Y}) \quad (10)$$

$$\delta_{i,j}\left(\mathbf{X}_i,\mathbf{X}_j\right) = \begin{cases} 0 & \text{if } w_{r,s}\left(\mathbf{X}_i\right)=w_{r,s}\left(\mathbf{X}_j\right) \\ \left|w_{r,s}\left(\mathbf{X}_i\right)\right|+\left|w_{r,s}\left(\mathbf{X}_i\right)\right| & \text{otherwise} \end{cases} \quad (11)$$

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

4.2 Evolutionary metrics

Minkowky and correlation metrics ignore the sequential structure of the values as resemblance is only based on the difference between the values at simultaneous measurements, independently of the increase/decrease behavior before and after these values. As a consequence, if we believe that two time paths should be considered more similar because of an affine temporal profile, then M_a or d_τ and d_ρ should not be used because they neglect an important aspect of the task.

To improve the ability of (6) to assess the proximity between trajectories D'Urso and Vichi [1998] and D'Urso [2000] added to $M_a\left(\mathbf{X}_i,\mathbf{X}_j\right)$ the distance between the first differences (velocities) and the distance between the second differences (accelerations). This idea can be further extended if we incorporate knowledge about jerks (third differences) and higher differences

$$V_{a,q}\left(\mathbf{X}_i,\mathbf{X}_j\right)=\sum_{k=1}^{q}w_q\left\{\frac{M_{a,k}\left[\mathbf{X}_i^{(k)},\mathbf{X}_j^{(k)}\right]}{M_{a,max}^{(k)}}\right\} \quad (12)$$

where $M_{a,k}\left[\mathbf{X}_i^{(k)}, \mathbf{X}_j^{(k)}\right]$ for $k = 0, 1, \cdots, q$ 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 until $k = q = \lceil (n-1)/2 \rceil$ (a bound is necessary to compute differences on enough data).

The weights $\mathbf{w} = \left(w_1, w_2, \cdots, w_q\right)$ in (12) are non negative and sum to one. $M_{a,max}^{(k)}$ denotes, for each k, the Minkowky distance of the furthest pair of sequences; thus, $V_{a,q}$ has values between zero and one. Each of a, q and $\mathbf{w}$ should be determined on a case-by-case basis. For $a = 2, q = 3, \mathbf{w} = (1/q, 1/q, \cdots, 1/q)$ expression (12) yields the distance function proposed by D'Urso and Vichi [1998] and for $a = 2, q = 2, \mathbf{w} = (1/q, 1/q, \cdots, 1/q)$ the distance function used by Tsai and Qu [2008]; the measure of Song [2008] can easily be derived from (12).

In some cases, the sequences are fairly short and it becomes necessary to exploit the full potential of the time series to assure good generalization ability and better classification accuracy. A reasonable approach is to extract both slowly and quickly varying features in order to provide a comprehensive view of the evolutionary behavior of a time series. We can devise an alternative to (12) by fragmenting the change from one period to another in many submovements each of which provides a separate piece of information. More specifically, the rate of change between two given times $r < s,\ r, s \in (1, 2, \cdots, n)$ is measured by

$$v_{r,s} = \begin{cases} \dfrac{(x_s - x_r)/(s-r)}{|x_u - x_t|/|u-t|} & \text{if } s \neq r \text{ and } x_u \neq x_t \\ 0 & \text{if } s = r \text{ or } x_u = x_t \end{cases} \tag{13}$$

The numerator of (13) is the relative variation between two data points, both in the positive and in the negative direction; the denominator is the maximal relative deviation achieved in a period-to-period change in $\mathbf{X}_i$ for any particular time interval. The values of $v_{r,s}$ range between -1 and 1 and are stable in linear transformation of intensity and insensitive to scale variation. From another point of view, formula [13] expresses the slope of the chord joining two time points.

To obtain an accurate representation of a time series we calculate the totality of its growth movements

$$\mathbf{V}_\mathbf{X} = \begin{pmatrix} 0 & v_{1,2} & \cdots & v_{1,n} \\ v_{2,1} & 0 & \cdots & v_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n,1} & v_{n,2} & \cdots & 0 \end{pmatrix} \tag{14}$$

Since $v_{r,s} = -v_{s,r}$ for all r and s, then (14) is antisymmetric, that is the growth matrix $\mathbf{V}_\mathbf{X}$ is equal to the additive inverse of its transpose. If the matrices $\mathbf{V}_{\mathbf{X}_i}$ and $\mathbf{V}_{\mathbf{X}_j}$ are a good summary of the fluctuations observed in $\mathbf{X}_i$ and $\mathbf{X}_j$ respectively,

then the problem of measuring the distance between two time series becomes a problem of measuring the distance between the two corresponding growth matrices. In other words, if the magnitude of $\left(\mathbf{V}_{\mathbf{X}_i} - \mathbf{V}_{\mathbf{X}_j}\right)$ does not depart too far from a zero matrix, then the two sequences have to be considered similar.

A matrix norm is a scalar measure which expresses numerically the magnitude of $\left(\mathbf{V}_{\mathbf{X}_i} - \mathbf{V}_{\mathbf{X}_j}\right)$. There are numerous useful matrix norms. We choose the weighted spectral norm defined as

$$C_W\left(\mathbf{X}_i, \mathbf{X}_j\right) = \sqrt{\left(\mathbf{V}_{\mathbf{X}_i} - \mathbf{V}_{\mathbf{X}_j}\right)^t \mathbf{W}^{-1} \left(\mathbf{V}_{\mathbf{X}_i} - \mathbf{V}_{\mathbf{X}_j}\right)} \tag{15}$$

because it captures the subspace where the difference matrix is most active. The matrix $\mathbf{W}$ is a (n,n) non negative matrix of weights that can be useful, for example, in comparing time series when it is more effective to assign a greater weight to the simultaneous presence of a peak or valley.

Lee and Verleysen [2005] present a simple metric that includes the sequential structure of time series:

$$L_p\left(\mathbf{X}_i, \mathbf{X}_j\right) = \left[\sum_{t=1}^{n} \left(B_{t-1} + B_{t+1}\right)^p\right]^{\frac{1}{p}} \tag{16}$$

with

$$2B_{t-1} = \begin{cases} \Delta_{i,j}^{(t)} & if\ \Delta_{i,j}^{(t)}\Delta_{i,j}^{(t-1)} \leq 0 \\ \dfrac{\Delta_{i,j}^{(t)}}{\Delta_{i,j}^{(t)} + \Delta_{i,j}^{(t-1)}} & \text{otherwise} \end{cases} \tag{17}$$

$$2B_{t+1} = \begin{cases} \Delta_{i,j}^{(t)} & if\ \Delta_{i,j}^{(t)}\Delta_{i,j}^{(t+1)} \leq 0 \\ \dfrac{\Delta_{i,j}^{(t)}}{\Delta_{i,j}^{(t)} + \Delta_{i,j}^{(t+1)}} & \text{otherwise} \end{cases} \tag{18}$$

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.

4.3 Comparison of the procedures

We have evaluated the performance of the new technique by considering artificially generated sample time sequences. 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 $N(0, \sigma = 1)$. All the quantities ϕ_1, ϕ_2, and $\mathbf{e}$ varies across time series.

To compare the stability of the above algorithms, we repeat 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 (see Hubert and Arabie [1985] and Milligan and Cooper [1986]). 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 m of the 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 s of ARI indicates stability across the simulations.

The following facts emerge from Table 1:

1. The mean ARI increases and the standard deviation 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.

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

k	n		WAD	M_2	d_τ	d_ρ	QD	$V_{2,2}$	$V_{2,3}$	$C_{\mathbf{I}}$	L_2
4	11	m	0.928	0.915	0.838	0.906	0.837	0.723	0.655	0.622	0.662
	11	s	0.062	0.068	0.091	0.074	0.092	0.123	0.134	0.120	0.116
	17	m	0.992	0.991	0.965	0.984	0.965	0.883	0.815	0.776	0.890
	17	s	0.020	0.022	0.046	0.032	0.046	0.086	0.117	0.109	0.078
	23	m	0.999	0.998	0.994	0.996	0.993	0.943	0.883	0.849	0.977
	23	s	0.009	0.010	0.020	0.015	0.020	0.060	0.089	0.098	0.037
8	11	m	0.662	0.656	0.621	0.638	0.621	0.589	0.573	0.514	0.506
	11	s	0.051	0.054	0.060	0.055	0.062	0.074	0.078	0.072	0.077
	17	m	0.716	0.707	0.694	0.700	0.695	0.681	0.662	0.576	0.633
	17	s	0.041	0.037	0.047	0.038	0.045	0.065	0.078	0.067	0.063
	23	m	0.745	0.726	0.721	0.722	0.721	0.732	0.725	0.618	0.695
	23	s	0.051	0.039	0.043	0.034	0.043	0.061	0.076	0.060	0.035
12	11	m	0.596	0.577	0.530	0.566	0.530	0.508	0.485	0.449	0.430
	11	s	0.059	0.060	0.066	0.056	0.066	0.073	0.076	0.072	0.073
	17	m	0.677	0.646	0.633	0.644	0.632	0.608	0.588	0.522	0.553
	17	s	0.053	0.045	0.052	0.048	0.051	0.062	0.072	0.062	0.060
	23	m	0.726	0.682	0.676	0.683	0.673	0.665	0.658	0.560	0.610
	23	s	0.059	0.046	0.051	0.046	0.050	0.058	0.067	0.055	0.037
		av.m	0.782	0.766	0.741	0.760	0.741	0.704	0.672	0.610	0.662
		av.s	0.045	0.042	0.053	0.044	0.053	0.074	0.087	0.079	0.064

All the calculations are carried out by the statistical language R, R Development Core Team [2008], and the programs are available by the authors.

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

3. Minkowsky and correlation metrics tend to achieve higher levels of ARI than evolutionary metrics.

4. Our method, WAD (Weighted Area Distance), has the best performance among the nine alternatives presented; 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 achievements are very similar to those of the Euclidean distance

5. Although the WAD has obtained the most satisfactory results, it does not bring massive improvement over the easily accessible M_2. At the very least, one should wonder if it just makes sense to introduce a new product if it is only slightly better than the one that it replaces.

5 Conclusions and future work

Dissimilarity search in short-run 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 several pairwise distance functions have been proposed in the literature. Nonetheless, the construction of a reliable measure of distance with optimal properties is still an open problem and subject to ongoing research.

In this paper we have designed a new metric, WAD, in the hope that it would be able to detect similarities over trajectories of short-run time series when conventional statistical analysis methods are not applicable. The proposed metric is sensible both to observed values and to rates of change. Cluster of sequences where determined by means of the PAM algorithm. A case study on relative consumer price index of OECD countries has been undertaken and the findings show that the new technique has acceptable results. Also, it has been verified by the help of test data that WAD has good recovery properties of the true group structure.

We have also discussed different propositions for a distance measure between short-run sequences concluding that many distance functions currently in use perform worse than the simple Euclidean metric. Perhaps, empirical evaluations in the past have been inadequate and a more thorough analysis is now required for definite conclusions (Keogh and Kasetty [2002]); however, more discussions regarding the relative merits and demerits of the various approaches can be found in the cited articles.

Finally, 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.

References

Dudoit, S. and J. Fridlyand, 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 M. Vichi, 1998. Dissimilarities between trajectories of a three-way longitudinal data set. In A. Rizzi, M. Vichi and H.-H. Bock, editors, *Advances in data science and classification.*, pp. 585–592. Berlin: Springer.

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

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

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

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

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

Lee, J.A. and M. Verleysen, 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*, pp. 733–740. Paris (France).

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

Ljubič, P., L. Todorovski, N. Lavrač and J.C. Bullas, 2002. Time-series analysis of UK traffic accident data. In *et al.* M. Bohanec, editor, *Proceedings of Multi-conference Information Society, 14-18 Oct. 2002*, pp. 135–138. Ljubliana, Slovenija: Springer-Verlag.

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

Mosig, A. and M. Clausen, 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.
URL http://www.R-project.org

Sieburg, H.B. and C.E. Müller-Sie, 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 A. Qu, 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. *et al.*, 2006. Order statistic correlation coefficient and its application to association measurement of biosignals. In *ICASSP 2006 Proceedings*, pp. II 1068–II 1071. Toulouse (France).

Xu, W. *et al.*, 2007. Order statistics correlation coefficient as a novel association measurement with applications to biosignal analysis. *IEEE Transactions on Signal Processing*, 55:5552–5563.