CENTRO PER LA FORMAZIONE IN ECONOMIA E POLITICA DELLO SVILUPPO RURALE
UNIVERSITÀ DI NAPOLI FEDERICO II - DIPARTIMENTO DI SCIENZE STATISTICHE
UNIVERSITÀ DI SALERNO - DIPARTIMENTO DI SCIENZE ECONOMICHE E STATISTICHE

Quaderni di STATISTICA

VOLUME 9 - 2007

LIGUORI EDITORE

Volume 9, anno 2007

ISSN 1594-3739 (edizione a stampa)

Registrazione al n. 5264 del 6/12/2001 presso il Tribunale di Napoli
ISBN-13 **978 - 88 - 207 - 4209 - 6**

Direttore responsabile: Gennaro Piccolo

© 2007 by Liguori Editore
Tutti i diritti sono riservati
Prima edizione italiana Dicembre 2007
Finito di stampare in Italia nel mese di Dicembre 2007 da OGL - Napoli

Questa opera è protetta dalla Legge sul diritto d'autore (Legge n. 633/1941).
Tutti i diritti, in particolare quelli relativi alla traduzione, alla citazione, alla riproduzione in qualsiasi forma, all'uso delle illustrazioni, delle tabelle e del materiale software a corredo, alla trasmissione radiofonica o televisiva, alla registrazione analogica o digitale, alla pubblicazione e diffusione attraverso la rete Internet sono riservati, anche nel caso di utilizzo parziale.
La riproduzione di questa opera, anche se parziale o in copia digitale, è ammessa solo ed esclusivamente nei limiti stabiliti dalla Legge ed è soggetta all'autorizzazione scritta dell'Editore. La violazione delle norme comporta le sanzioni previste dalla legge.
Il regolamento per l'uso dei contenuti e dei servizi presenti sul sito della Casa Editrice Liguori è disponibile al seguente indirizzo: http://www.liguori.it/politiche_contatti/default.asp?c=legal
L'utilizzo in questa pubblicazione di denominazioni generiche, nomi commerciali e marchi registrati, anche se non specificamente identificati, non implica che tali denominazioni o marchi non siano protetti dalle relative leggi o regolamenti.

Il C.F.E.P.S.R. si avvale per la stampa dei Quaderni di Statistica del contributo dell'Istituto Banco di Napoli - Fondazione.

La carta utilizzata per la stampa di questo volume è inalterabile, priva di acidi, a PH neutro, conforme alle norme UNI EN ISO 9706 ×, realizzata con materie prime fibrose vergini provenienti da piantagioni rinnovabili e prodotti ausiliari assolutamente naturali, non inquinanti e totalmente biodegradabili.

Indice

Quaderni di Statistica
Vol. 9, 2007

Dissimilarity criteria for time series data mining

Marcella Corduas
Dipartimento di Scienze Statistiche, Università degli Studi di Napoli Federico II
E-mail: corduas@unina.it

Summary: In the last decade there has been an increasing interest in mining time series data and many distance measures and representations have been proposed for this purpose. This paper illustrates some of the dissimilarity measures introduced in literature to index time series and discusses their importance and critical aspects.

Keywords: Time Series, Data mining, Dissimilarity measures.

1. Introduction

In recent years the mining of time series data has attracted considerable interest stimulated by the progress in computer technology. Data mining, however, has not a clear definition. It has been viewed as a part of the larger process of knowledge discovery in databases (usually denoted as KDD). It includes all the operational steps necessary for extracting "knowledge" from observed data organized in very large archives. This process requires both a preliminary management of information, aimed at organizing and cleaning data, and the subsequent statistical analysis; it consists of the following steps: data-warehousing, target data selection, data cleaning, data transformation and reduction, data mining, model identification, estimation and interpretation, use of the extracted knowledge (Fayyad, 1997). Besides, data mining has been tied to pattern or structure discovery in large databases and to the construction of useful knowledge for predicting and controlling complex systems (Hand, 1998).

CENTRO PER LA FORMAZIONE IN ECONOMIA E POLITICA DELLO SVILUPPO RURALE
UNIVERSITÀ DI NAPOLI FEDERICO II - DIPARTIMENTO DI SCIENZE STATISTICHE
UNIVERSITÀ DI SALERNO - DIPARTIMENTO DI SCIENZE ECONOMICHE E STATISTICHE

Quaderni di STATISTICA

VOLUME 9 - 2007

LIGUORI EDITORE

Volume 9, anno 2007

ISSN 1594-3739 (edizione a stampa)

Registrazione al n. 5264 del 6/12/2001 presso il Tribunale di Napoli
ISBN-13 **978 - 88 - 207 - 4209 - 6**

Direttore responsabile: Gennaro Piccolo

© 2007 by Liguori Editore
Tutti i diritti sono riservati
Prima edizione italiana Dicembre 2007
Finito di stampare in Italia nel mese di Dicembre 2007 da OGL - Napoli

Questa opera è protetta dalla Legge sul diritto d'autore (Legge n. 633/1941).
Tutti i diritti, in particolare quelli relativi alla traduzione, alla citazione, alla riproduzione in qualsiasi forma, all'uso delle illustrazioni, delle tabelle e del materiale software a corredo, alla trasmissione radiofonica o televisiva, alla registrazione analogica o digitale, alla pubblicazione e diffusione attraverso la rete Internet sono riservati, anche nel caso di utilizzo parziale.
La riproduzione di questa opera, anche se parziale o in copia digitale, è ammessa solo ed esclusivamente nei limiti stabiliti dalla Legge ed è soggetta all'autorizzazione scritta dell'Editore. La violazione delle norme comporta le sanzioni previste dalla legge.
Il regolamento per l'uso dei contenuti e dei servizi presenti sul sito della Casa Editrice Liguori è disponibile al seguente indirizzo: http://www.liguori.it/politiche_contatti/default.asp?c=legal
L'utilizzo in questa pubblicazione di denominazioni generiche, nomi commerciali e marchi registrati, anche se non specificamente identificati, non implica che tali denominazioni o marchi non siano protetti dalle relative leggi o regolamenti.

Il C.F.E.P.S.R. si avvale per la stampa dei Quaderni di Statistica del contributo dell'Istituto Banco di Napoli - Fondazione.

La carta utilizzata per la stampa di questo volume è inalterabile, priva di acidi, a PH neutro, conforme alle norme UNI EN ISO 9706 ×, realizzata con materie prime fibrose vergini provenienti da piantagioni rinnovabili e prodotti ausiliari assolutamente naturali, non inquinanti e totalmente biodegradabili.

Indice

Both the concepts of data mining imply a strong role of data processing and for this reason the related research field has been significantly occupied by researchers working on database management and machine learning who have often rediscovered known statistical techniques (see the interesting review by Keogh and Kasetty, 2003). Furthermore, it is relevant, in our opinion, to point out that the inferential problem is rarely considered and almost all the contributions remain at a descriptive level and concentrate on the developing of algorithms. The inferential problem, instead, seems to be fundamental: data mining, in fact, aims at finding unknown patterns in the information recorded in a database, but the discovered patterns will be really useful only if they reflect a 'general' truth.

Prediction and control are typical objectives of time series analysis and many applications in real-life involve the study of massive data archive and require some kind of data mining technique. Similarly, the search of common patterns, such as trend, cycle, seasonality, is a well known problem in time series analysis. Therefore, the real challenge is the large amount of data available which makes any traditional 'ad hoc' procedure useless.

With respect to time series data, attention has been generally focused on four main problems: a) *indexing*, which, given a time series (a query sequence), finds the nearest matching time series in a database; b) *clustering*; c) *classification* and *discrimination*; d) *segmentation*, which represents a time series through a piecewise model in order to use such a representation for more complex tasks. In this article we will discuss the first of the above mentioned problems with special reference to dissimilarity criteria.

In order to make the comparison of time series meaningful, one important question is to decide what similarity means and what features have to be extracted from a time series (Corduas, 2007). This question leads to the fundamental dichotomy: a) similarity can be based solely on time series *shape*; b) similarity can be measured by looking at time series dynamic *structure*.

In this respect, the leading idea is that data mining has to discover objects that move similarly or closely follow certain given pattern. The

final match should be consistent with human intuition. This concept is typical of shape based dissimilarity measures. However, as this article will discuss, the final objective of a statistical analysis, such as forecasting and control, may lead to different approaches where time series modelling assumes a definite role.

Finally, data mining involves complex computational problems due to the large size of data archives (Scepi and Milone, 2007). For this reason, an effective dissimilarity measure should allow the user to achieve an efficient implementation at a low cost. This is in general accomplished by a pre-processing step where, on the basis of a fast criterion, the majority of distant series with respect to the 'query' are ruled out. For this purpose, a lower-bound for the chosen distance criterion is needed. This bound underestimates the true distance but it only uses a reduced set of information. This strategy is computationally advantageous although it may generate a risk of false dismissals. Once the possible matching candidates have been selected using the approximated dissimilarity criterion, the final comparison is made on the basis of the exact distance measure.

This paper presents some of the dissimilarity measures introduced in literature to index time series and discusses their importance and critical aspects.

2. Shape (dis)similarity

Given the objectives of data mining that we have just illustrated, it is not surprising to find out that the Euclidean distance is one of the most common device used in practice for data mining purposes.

The comparison is simply referred to the observations:

$$D_E(x_t, y_t) = \{\sum_{t=1}^{n} [x_t - y_t]^2\}^{1/2}, \tag{1}$$

where x_t and y_t, $t = 1, 2, ...n$, are zero mean time series. The distance may be referred to standardized time series $\widetilde{x}_t$ and $\widetilde{y}_t$ leading to a more meaningful criterion which is invariant for linear transformation of data. In such a case, D_E^2 is just a linear transformation of the correlation coefficient of the two series, being $D_E(\widetilde{x}_t, \widetilde{y}_t) = \{2n(1 - r_{xy}\}^{1/2}$.

This is a well known descriptive criterion which has been used in statistics since the beginning of eighties when the interest for time series classification arose (Piccolo, 1987). In this respect, among others, Zani (1981) considered this kind of distance and extended the idea in order to take lagged relationships between time series into account. In fact, he introduced the following distance:

$$D_s(\widetilde{x}_t, \widetilde{y}_{t-s}) = \{\sum_{t=1}^{n} [\widetilde{x}_t - \widetilde{y}_{t-s}]^2\}^{1/2} = \{2(n-s)(1-r_{xy}(s)\}^{1/2}, \quad (2)$$

where $r_{xy}(s)$ is the s-lag cross-correlation coefficient between the two series. In such a way, time series, characterized by similar patterns which are simply shifted on the time-axis, are recognized as similar.

Indexing is solved exploiting the non decreasing property of the illustrated metrics with respect to added terms in the summation. The approach for finding the best matching time series is sequential. Given a time series a possible matching candidate is dismissed as soon as the distance between the first k observations is larger than a fixed threshold.

Time series length can affect the quality of the results. The distance (1) requires that the time series being compared have the same length; this aspect may become critical when a large number of time series is involved in the comparison since it may lead to an inefficient use of available information. Moreover, when the number of observations is very large, local changes or structural changes in time series pattern are more likely and, consequently, the distance (1) becomes less informative. Finally, the measure is very sensitive to outliers.

Figure 1 illustrates the plot of the ECG's sine wave of two healthy people. This is a typical example of time series which show a remarkable similarity though they are locally out-phase. As a matter of fact, the heart rate differs from one individual to another. Then, the peaks in the series may not be aligned and, for this reason, the Euclidean distance produces meaningless results.

Moreover, special attention has to be paid at the choice of the clustering algorithm. Keogh and Li (2005) showed in fact that k-means clustering provides meaningless results when time series subsequences grouping is considered. In particular, the cluster centers tend to appear as simple

trigonometric functions due to the obvious Slutsky effect caused by the computation of the average time series (Slutsky, 1937). Besides, the patterns of the identified cluster centers have in general no reference with the behaviour of the observed time series.

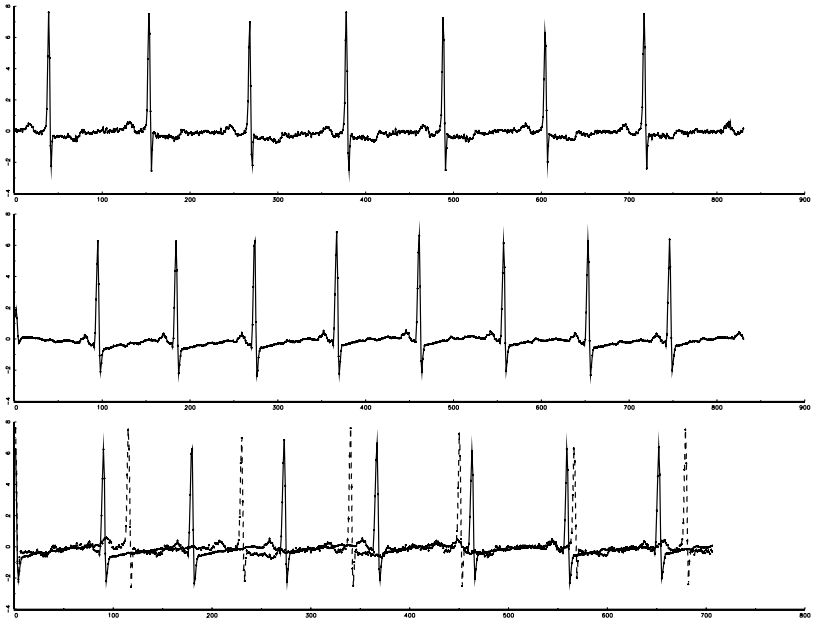

Figure 1. a-b) ECG's standardized time series c) time series aligned with respect to the first heartbeat

Finally, it is worth mentioning a novel problem which has recently raised attention: the search of time series discords. This term refers to subsequences of longer time series which are maximally different to all the rest of the time series, in other words, their pattern is very far from the observed temporal behaviour of the rest of the time series. The search of anomalous pattern is of course of interest for applications in medicine for detecting specific illness from ECG or ECC recordings, but other fields such as financial data analysis, text processing, quality control, seismic wave analysis may also benefit of such techniques. However, the approach requires the solution of a rather complex algorithmic prob-

lem since a strategy for the selection of the best subsequence length is needed (see Keogh *et al.*, 2005).

3. The Dynamic Time Warping

The Euclidean distance is very sensitive to distortion in time axis and may lead to poor results for sequences which are similar, but locally out of phase. For this reason, the Dynamic Time Warping (DTW), originally introduced for speech processing (Sakoe and Ciba, 1978; Berndt and Clifford, 1994; Wang and Gasser, 1997), has been reconsidered for data mining purposes.

DTW allows non-linear alignments between time series. Specifically, given two data sequences $x = \{x_i, i = 1, 2, ..., m\}$, and $y = \{y_j, j = 1, 2, ..., n\}$, the procedure starts by constructing the $m \times n$ matrix Δ where the (i, j) element is the distance (or dissimilarity) $\delta(x_i, y_j)$ between two points x_i and y_j. The best matching is found by searching a path through this matrix such that the total cumulative distance between the aligned elements of the two time series is minimized.

For this purpose, we denote by $w = \{(i(k), j(k)), k = 1, ..., K, i(1) = j(1) = 1, i(K) = M, j(K) = N\}$ with $\max(m, n) \leq K \leq m + n - 1$, a warping path connecting $(1, 1)$ and (m, n). The alignment between the time series is obtained by searching for the path through the matrix Δ which minimizes a cost function such as:

$$\mathcal{C}(x, y, w) = \sum_{k=1}^{K} \delta(x_{i(k)}, y_{j(k)}) r(k), \tag{3}$$

where $r(k)$ is an appropriate non negative weighting function (this is often set to $1/k$). Of course, the choice of the cost function determines the warping result.

Some constraints are imposed in order to reduce the number of paths considered:

- *boundary*: $i(1) = j(1) = 1$, $i(K) = M$, $j(K) = N$; as mentioned above, the optimal path has to connect the elements $(1, 1)$ and (m, n).

- *monotonicity*: $i(k) \leq i(k+1)$ and $j(k) \leq j(k+1)$, that is both index can never decrease;

- *continuity*: $i(k+1) - i(k) \leq 1$ and $j(k+1) - j(k) \leq 1$; which implies that indices can only increase by 0 or 1 going from k to $k+1$;

- *window*: the path is allowed to move within a definite region around the matrix diagonal. This region is usually defined as a rectangular band (assuming $|i(k) - j(k)| \leq h$ where h is a given positive integer, see Sakoe and Chiba, 1978) or a parallelogram (Itakura, 1975);

- *slope*: the path should be neither too steep nor too shallow.

At the end of the optimizing process, the optimal path also provides a measure of the *dynamic warping distance* between the two time series:

$$DTW(x, y) = \inf_{w} \mathcal{C}(x, y, w). \tag{4}$$

For instance, assuming $\delta(x_i, y_j) = (x_i - y_j)^2$, that is the squared Euclidean distance between two data points, Ratanamahatana and Keogh (2004) uses the following cost function:

$$\mathcal{C}(x, y, w) = \sqrt{\sum_{k=1}^{K} \delta(x_{i(k)}, y_{j(k)})} = \sqrt{\sum_{k=1}^{K} (x_{i(k)} - y_{j(k)})^2}. \tag{5}$$

Note that the Euclidean distance (1) is simply (5) under the constraint that the warping path w satisfies $i(k) = j(k) = k$.

It is generally recognized that the performance of DTW deteriorates for noisy data since the search for an optimal alignment tends to privilege very extreme data by accommodating outliers in one of the time series with extreme values of the other.

Moreover, the indexing requires the identification of a lower bound for DTW. In this respect, Keogh (2002) introduced a technique to produce such a bound for rectangular and parallelogram bands which computationally is very efficient. Later, a general approach which allows for

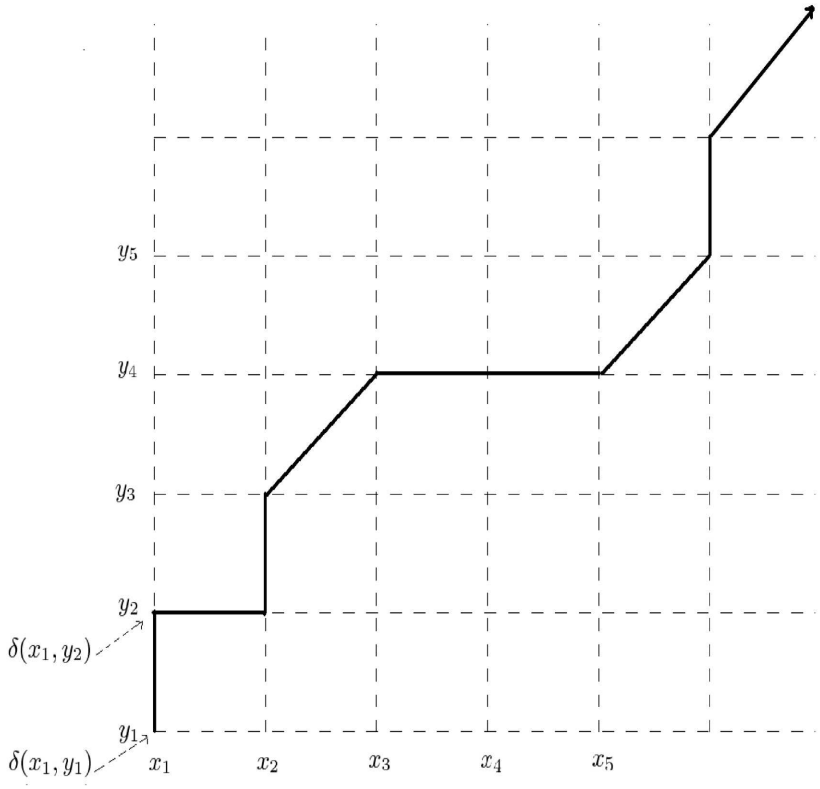

Figure 2. Optimal path search

an arbitrary shape of the window has been studied by Ratanamahatana and Keogh (2004).

Various developments of this technique have been proposed such as, among the others, the extension to multidimensional data (Vlachos *et al.* 2006), the use of smoothing for noisy data (Morlini, 2005), the joint use of DTW and *S*elf *O*rganizing *M*ap (SOM) algorithm for improving time series clustering (Romano and Scepi, 2006), the study of new techniques for approximating DTW (Chu *et al.* 2002).

Regardless of the computational complexity, DTW has been used in several fields in order to compare processes which evolve at varying rates.

For instance, it has been applied in bio-informatics for gene expression time series (Aach and Church, 2001), for monitoring chemical and industrial processes (Kassidas *et al.* 1998), for classifying ECG data (Caiani *et al.* 1998) and for comparing time series extracted from video or classifying handwritten texts.

4. Measuring dissimilarity by feature extraction

The idea that the components underlying the dynamic structure of the phenomena under investigation could help in detecting similarity was already present in the earlier studies.

Agrawal *et al.* (1994) exploited the Parseval's theorem in order to transpose $D_E^2(x_t, y_t)$ in the frequency domain. They considered the *D*iscrete *F*ourier *T*ransform (DFT) of the data:

$$x(\omega_j) = n^{-1/2} \sum_{t=0}^{n-1} x_t \exp(-\imath \omega_j t), \tag{6}$$

where $\omega_j = 2\pi j/n$, $j = 0, 1, ..., (n-1)$ and introduced the criterion:

$$D_{A,n}^2 = \sum_{j=0}^{n-1} |x(\omega_j) - y(\omega_j)|^2. \tag{7}$$

In particular, they proposed linking each time series in the database with the first k coefficients of the respective DFT so that a selection of potential candidates for the final matching was simply identified by means of a bound over the Euclidean distance: $D_{A,k} < \epsilon$. Standardizing the data first will allow for differences in level and scale: although this step is not clearly stated in many papers, as previously mentioned, it is a necessary preliminary requirement for meaningful results.

A dimensionality reduction of the original problem is achieved by keeping a collection of selective Fourier coefficients, since it is reasonable to expect that those coefficients will summarize prominent features of the time series which are object of comparison.

However, the mentioned indexing strategy has two critical issues: the selection of the threshold (which is data dependent) and, above all, the assumption that low frequencies will be in general more informative about the temporal dynamics. This is a very restrictive assumption which strongly affects the effectiveness of the proposed approach in practice.

In general, $D_{A,k}$ can be used as a measure of diversity over a specific band of frequency by varying ω_j in a sub-interval of $(-\pi, \pi]$. Ng and Huang (1999) applied this technique for classifying stars light curves. They discussed the problems of working with large database (the whole database contained light curves of 20 million stars). The example that they provided referred to only 20383 time series, but this is already a number of subjects which many statistical approaches cannot handle.

The procedure based on DFT was improved by Rafei *et al.* (1998) who exploited the symmetry of the Fourier coefficients in order to provide a tighter lower bound for the Euclidean distance using the same number of coefficients. In the same vein, for periodic data, Vlachos *et al.* (2005) considered the Euclidean distance between the periodograms of the standardized series as a measure of dissimilarity between time series. In order to compress information, they suggested recording for each time series the largest Fourier coefficients (i.e. the coefficients corresponding to the highest peaks of the periodogram), their related frequencies and a measure of the approximation error due to the compression. In such a way a lowering bound for the proposed distance can be easily derived.

A fast preprocessing step can also be combined with computationally more demanding dissimilarity criteria, such as the Euclidean distance between the (global and partial) autocorrelation functions or the smoothed periodogram (Wang and Wang, 2000).

This type of dissimilarity measures also are not new in statistics: Bohte *et al.* (1980), Kovačić (1996) proposed several descriptive dissimilarity measures based on the comparison of auto or cross-correlation functions; Mélard and Roy (1984) investigated a test for assessing the equality of global autocovariance functions; Galeano and Peña (2000) considered the Mahalanobis distance between autocorrelation coefficients. Besides, in the frequency domain, Diggle and Fisher (1991) introduced a non parametric approach to compare the spectrum of two time series through the

underlying cumulative periodograms; Anderson (1993) compared the cumulative spectral functions, and, more recently, Caiado *et al.* (2006) considered the Euclidean distance between normalized periodograms as a measure to discriminate between stationary and non stationary time series.

The assumption of stationarity of the data generating process is, in our opinion, a critical issue for all methods which rely on time series features such as periodogram, spectrum or autocorrelation functions. Of course, those techniques can be still applied when the non stationarity is removed from each time series by means of the same differencing operator or detrending technique. This approach introduces a subjective judgment concerning the way of achieving the stationarity. However, in many papers the problem is not clearly stated and the dissimilarity measures based on those features are evaluated both for stationary and non stationary time series.

4.1. Wavelets

Some contributions have explored the use of wavelet analysis for data mining in order to find a valid alternative to the traditional Fourier analysis (see Li et al. 2002 for an extensive review, and Priestley, 1996; Percival and Walden, 2000, for a comprehensive illustration of wavelet analysis and related properties).

Wavelets can be viewed as orthonormal basis functions for various function spaces (though, in general, orthonormal property is not strictly required). The set of basis functions are obtained by dilations and translations of a single function, a mother wavelet $\psi(t)$. Specifically, this is simply defined by:

$$\{\psi_{j,k}(t) = 2^{(j/2)}\psi(2^j t - k),\ j,k \in Z\}, \tag{8}$$

where j is the dilation factor and k the translation factor. A suitable choice of $\psi(t)$ can generate a set of functions $\{\psi_{j,k}\}$ which is an orthonormal basis for $L^2(\Re)$. Any function $f(t) \in L^2(\Re)$ can be written as a wavelet

series expansion:

$$f(t) = \sum_{k=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} w_{j,k}\psi_{j,k}(t) \tag{9}$$

where the coefficients: $w_{j,k} = \int_{-\infty}^{\infty} f(t)\psi_{j,k}(t)dx$.

The above expressions show the obvious analogy between wavelet analysis and Fourier analysis. Both techniques are aimed at representing a function using orthonormal basis functions, but the former is capable of providing both time and frequency localization (via translations and dilations respectively) whereas standard Fourier sine and cosine series only provide frequencies representations.

In this respect, Priestley (1996) discusses the relationship between wavelet analysis and time-dependent spectral analysis. The former in fact is often referred as an effective tool for time-frequency decomposition of non stationary signals. In general, wavelets are in fact designed to have a good frequency resolution at low frequencies and a very poor resolution at high frequencies; the reverse applies to time resolution.

The main technique used for data mining purposes relies on the *D*iscrete *W*avelet *T*ransform (DWT) and, specifically, the most commonly used mother wavelet is the Haar wavelet:

$$\psi_H(x) = \begin{cases} 1, & 0 \leq x \leq 0.5; \\ -1, & -0.5 \leq x < 0; \\ 0, & \text{otherwise.} \end{cases}$$

Wavelets have several properties, but two of them are relevant for data mining, under the assumptions stated above: i) Parseval's theorem still holds; ii) DWT is computationally easier with respect to DFT: it requires only $O(n)$ multiplications whereas the best Fast Fourier Transform needs $O(n \log n)$ multiplications. Besides, the Haar wavelets are very simple to compute and to understand since, in practice, the transform is given by a recursive pairwise averaging and differencing of data.

Chan and Fu (1999) exploited the former property and showed the preservation of Euclidean distance in both time and Haar domain. This

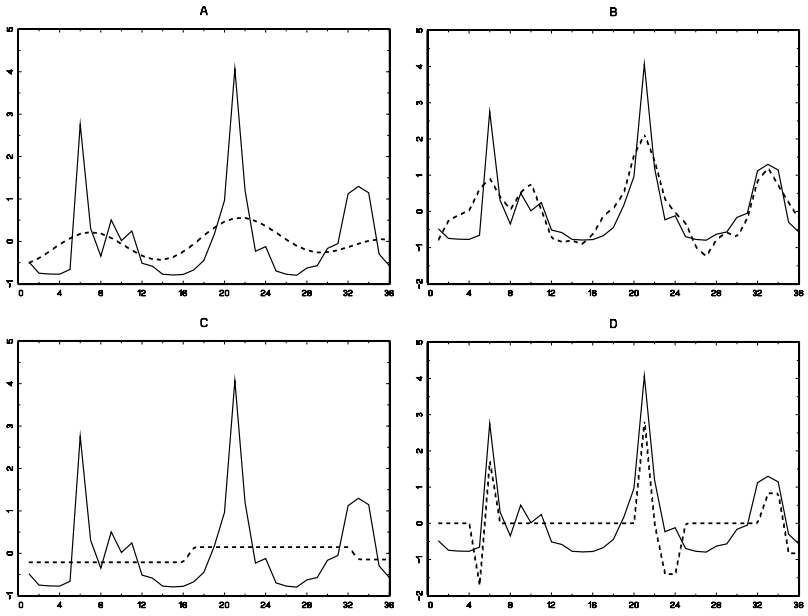

Figure 3. Time series reconstructed by Fourier (A-B) and Haar (C-D) transform coefficients: A-C first 15 coefficients; B-D largest 15 coefficients

fact ensures that the Haar transform can be used, by analogy to Fourier transform, as a tool for extracting characteristic features of a time series. Specifically, they proposed to use the first few coefficients of the transformed sequences for indexing purposes in order to perform a similarity search.

A further use of Haar wavelet transform was suggested by Struzik and Siebes (1999) who considered some special representations derived from it: a) the sign based representation, which uses only the sign of the wavelet coefficients; b) the difference of the logarithms of values of the wavelet coefficient at the highest scale and at the working scale. As a similarity measure they introduced the step-wise correlations between these special representations for the sequences to be compared.

Figure 3 reports the results obtained from the Fourier and Haar transform of a monthly streamflow series ($n = 256$). Specifically, the graph

illustrates the first 3 years of the observed time series together with the reconstructed series obtained by means of the first or "best" 15 coefficients. The improvement obtained by using the best coefficients (see panel B and D) is immediately clear arising from the special nature of the data which exhibit a strong seasonal pattern. Besides, the pattern of the reconstructed series enhances the fact that Haar transform is capable to capture local behaviour (and singularities) of the time series.

5. Structural dissimilarity

It is interesting to note that data mining literature generally considers time series as a recorded geometric trajectory and it often seems to ignore the theory of stochastic processes. Only recently, the attention for the dynamic structure has inevitably conveyed the investigation to the stochastic generating process that has originated the observed trajectory.

In this respect, the class of Gaussian $ARIMA$ processes provides a useful parsimonious representation (Box and Jenkins, 1976) for linear time series. Specifically, $Z_t \sim ARIMA(p, d, q)$ is defined by:

$$\phi(B)\nabla^d Z_t = \theta(B)a_t, \tag{10}$$

where a_t is a Gaussian White Noise (WN) process with constant variance σ_a^2 , B is the backshift operator such that $B^k Z_t = Z_{t-k}$, $\forall k = 0, \pm 1, \ldots$, the polynomials $\phi(B) = 1 - \phi_1 B - ... - \phi_p B^p$ and $\theta(B) = 1 - \theta_1 B - ... - \theta_q B^q$, have no common factors, and all the roots of $\phi(B)\theta(B) = 0$ lie outside the unit circle. Moreover, we assume that the time series has been preliminary transformed in order to improve Gaussianity, to deal with non-linearities, to reduce asymmetry, and to remove any outlier or deterministic component (such as deterministic seasonality, trading days, calendar effects, mean level, etc.).

First of all, we will introduce a distance criterion based on *cepstral coefficients*, $c_{x,j}$, of zero mean stationary series. These are determined by the following Fourier expansion:

$$\ln f_x(\omega) = \sum_{j=-\infty}^{\infty} c_{x,j} \exp(-\imath\omega j), \tag{11}$$

where $f_X(\omega)$, $\omega \in (-\pi, \pi]$ is the spectrum of the process X_t (Bogert *et al.*, 1962).

Assuming that the generating process is well approximated by a pure stationary $AR(p)$ model, a simple expression of the cepstral coefficients in terms of the AR parameters can be derived:

$$c_1 = -\phi_1;$$

$$c_k = -\phi_k - \sum_{j=1}^{k-1} \frac{1}{k}\,\phi_{k-j}\,c_j, \qquad 1 < k \leq p;$$

$$c_k = -\sum_{j=1}^{p} (1 - \frac{1}{k})\,\phi_j\,c_{k-j}, \qquad k > p.$$

Thus, for several decades, the cepstral distance:

$$D_{C,k} = \sqrt{\sum_{j=1}^{k} [c_{x,j} - c_{y,j}]^2} \tag{12}$$

had been widely applied to signal processing both for speech recognition and biomedical signal classification (see for instance Gray and Markel, 1976; Kang *et al.* 1995). More recently, Kalpakis *et al.* (2001) investigated the use of the Euclidean distance between cepstral coefficients for data mining purposes comparing it with other dissimilarity criteria.

Note that, in the expression (12), the term $(c_{x,0} - c_{y,0})^2 = \ln(\sigma_x^2/\sigma_y^2)$ is omitted since it is the log of the White Noise variance ratio and hence it simply represents a scale factor. Moreover, the cepstral coefficients quickly decay to zeros, and then, by analogy to previous methods, just a few number of cepstral coefficients, M, have to be stored for indexing purposes so that the Euclidean distance will be computed on the truncated series of cepstral coefficients.

Finally, an interesting interpretation can be given by considering that:

$$2D_{C,\infty}^2 + \ln(\sigma_x^2/\sigma_y^2) = \int_{-\pi}^{\pi} |\ln f_x(\omega) - \ln f_y(\omega)|^2 \frac{d\omega}{2\pi}. \tag{13}$$

When the WN variances are equal, the cepstral distance DC, ∞ is related to the root mean square distance between the log spectra.

Several improvements have been proposed for speech recognition purposes such as the use of Mahalanobis distance or the introduction of a weighted Euclidean distance in which each coefficient is simply weighted by the inverse of its variance in order to enhance the contribution of weights with lower variability (Tohkura, 1987).

5.1. The AR metric

As mentioned before, the strong premise of data mining, which evaluates similarity of time series patterns, limits the role of structural dissimilarity measures. As a matter of fact, when the data generating processes become the terms of the comparison, the fact that, according to some metric, two stochastic processes are very close does not imply that the patterns of the specific observed trajectories look 'visually' close. However, some attempts have been made in order to move the research focus on structural dissimilarity.

Piccolo (1984, 1990) proposed a distance criterion which compares the forecasting functions of two $ARIMA$ models given a set of initial values. In particular, assuming that Z_t is a zero mean invertible process which admits the $AR(\infty)$ representations: $\pi(B)Z_t = a_t$, the π-weights sequence and the WN variance completely characterize Z_t (given the initial values).

Exploiting this result the following distance can be introduced as a measure of structural diversity between two $ARIMA$ processes with given orders, X_t and Y_t:

$$D_{AR} = \sqrt{\sum_{j=1}^{\infty} (\pi_{xj} - \pi_{yj})^2}. \tag{14}$$

As before, the WN variances are not included in the distance formulation since they depend on the units of measurement. The criterion has been widely experimented (see Piccolo, 2007 for a review) and the asymptotic distribution of ML estimator of D_{AR} has been derived under general assumptions (Corduas, 2000; Corduas and Piccolo, 2007). The

approach developed is particularly interesting since it reduces the problem of comparing time series to an hypotheses testing problem and provides a more objective strategy for assessing the similarity of two time series.

It is worth noting that the distance between $ARIMA$ processes has been re-discovered by several contributing authors in data mining literature. For instance, Deng *et al.* (1997) suggested to compress information concerning a time series by retaining the estimated parameters of the corresponding $ARMA$ model. However, the metric that they proposed was not effective since they measured the distance between two time series by comparing the AR and MA coefficients separately and it is well known that those components have a very different role in determining the dynamics of a time series. Moreover, the presence of redundancy in one of the models may originate a very large distance between time series which are instead very close.

Recently, Bagnall and Janaceck (2005) proposed to translate a time series into binary sequences considering the median of the data as reference and to exploit the relationships between Gaussian $ARIMA$ processes and binary time series (see Kedem, 1980) in order to reduce the amount of storage and computational resources for time series comparison. They applied the AR metric (14) to cluster the transformed binary time series. The approach proved to be useful when data were affected by outliers; in addition the technique achieved a clustering accuracy which was equivalent to that obtained by cepstral distance.

Moreover, Baragona and Vitrano (2007) compared the performance of the AR metric with a criterion based on cross-correlations for data mining purposes.

In this respect, although one of the objective of data mining is reducing computing time and resources, we still believe that the AR metric is a valid alternative to other measure dissimilarity. It can be directly applied to the original observed time series by using an automatic modelling procedure. In this respect, Liu *et al.* (2001) provides a relevant case study which shows that preliminary modelling is not an obstacle to analyses involving very large time series databases. Specifically, they investigate the use of well known techniques for automatic identification of

time series models in order to solve a forecasting problem of a worldwide multi-brand fast-food restaurant chain at store and corporate level.

In the same vein, for instance, the Mahalanobis distance between Autoregressive (AR) models was widely applied to speech recognition. Again, the approach was developed in an hypotheses testing framework (see, Thomson and De Souza, 1985, and references therein reported). Xiong and Yeung (2004), instead, introduced a model-based clustering approach based on mixtures of $ARMA$ models.

6. Final remarks

Many other measures were introduced in literature such as piecewise normalization (Indyk *et al.* 2000), piecewise aggregate approximation, piecewise probabilistic measures (Keogh and Smyth, 1979), cosine wavelets (Huntala *et al.* 1999), the characteristic-based clustering (Wang *et al.* 2006) etc. This article only considered the main approaches which have assumed some importance in this research area.

Much of the work on time series data mining was developed by the database community that has often rediscovered criteria well known in statistics (see Corduas, 2003). In this respect the recent review by Keogh and Kasetty (2003) is illuminating: hundreds of papers have introduced new dissimilarity measures and algorithms to index and classify time series but only a few of them have really proved to be effective with respect to other existing criteria. Moreover, the comparison is confined to a mere description of observed time series patterns and no attempt to introduce a testing hypotheses framework has been done.

This fact proves that, despite several recommendations to become more involved with data mining problems (Hand, 1998), the contribution of statistics and statisticians to this research area is still scarce and a systematic cooperation with other disciplines is required.

Acknowledgements: This research was supported by Dipartimento di Scienze Statistiche, University of Naples Federico II, and C.F.E.P.S.R. (Portici). The author thanks the referees for helpful comments.

References

Aach J., Church G. (2001), Aligning gene expression time series with time warping algorithms, *Bioinformatics*, 17, 495-508.

Agrawal R., Faloutsos C., Swami, A. (1994), Efficient similarity search in sequence databases, *4th Proceedings of F.O.D.O.*, Lecture notes in Computer Science, 730, Springer Verlag, New York, 69-84.

Anderson T.W. (1993), Goodness of fit tests for spectral distributions, *Annals of Statistics*, 21, 830-847.

Bagnall A.J., Janaceck G.J. (2005), Clustering time series from ARMA models with clipped data, *Machine Learning*, 58, 151-178.

Baragona R., Vitrano S. (2007), Statistical and numerical algorithms for time series classification, *Proceedings of CLADAG 2007*, EUM, Macerata, 65-68.

Berndt D., Clifford J. (1994), Using dynamic time warping to find patterns in time series, *Proceedings of AAAI-94 workshop of SIGKDD*, 229-248.

Bogert B.P., Healy M.J., Tukey J.W. (1962), The quefrequency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking, in Rosenblatt M., ed.: *Proceed. Symposium on Time Series Analysis*, J. Wiley, New York, 209-263.

Bohte Z., Cepar D., Kosmelij K. (1980), Clustering of time series, *Compstat 80*, 587-593.

Box G.E.P., Jenkins G.M. (1976), *Time series analysis: forecasting and control*, rev. edition, Holden-Day, San Francisco.

Caiado J., Crato N., Peña D. (2006), A periodogram-based metric for time series classification, *Computational Statistics & Data Analysis*, 50, 2668-2684.

Caiani E.G., Porta A., Baselli G., Turiel M., Muzzupappa S., Pieruzzi F., Crema C., Malliani A., Cerutti S. (1998), Warped-average template technique to track on a cycle-by-cycle basis the cardiac filling phases on left ventricular volume, *IEEE Computers in Cardiology* , 25, 73-76.

Chan K., Fu A.W. (1999), Efficient time series matching by wavelets, *ICDE*, 126-133.

Chu S., Keogh E., Hart D., Pazzani M. (2002), Iterative deepining dynamic time warping for time series, *Proceedings of SIAM KDD*, electronic edition.

Corduas M. (2000), La metrica Autoregressiva tra modelli ARIMA: una procedura in linguaggio *G*AUSS, *Quaderni di Statistica*, 2, 1-37.

Corduas M. (2003), Il confronto tra serie storiche nell'analisi statistica di dati dinamici, *Atti della Riunione SIS*, R.Curtò, Napoli, 213-224.

Corduas M. (2007), Comparing time series: shape-based or structural dissimilarity?, *Proceedings of CLADAG 2007*, EUM, Macerata, 69-72.

Corduas M., Piccolo D. (2007), Time series clustering and classification by the Autoregressive Metric, *Computational Statistics & Data Analysis*, doi: 10.1016/j.csda.2007.06.001.

Deng K., Moore A., Nechyba M.C. (1997), Learning to recognize time series: combining ARMA models with memory-based learning, *Proceeding of the International Symposium on Computational Intelligence in Robotics and Automation*, 246-251.

Diggle P.J., Fisher N.I. (1991), Nonparametric comparison of cumulative periodograms, *Applied Statistics*, 40, 423-434.

Fayyad U. (1997), Editorial, *Data Mining and Knowledge Discovery*, 1, 5-10.

Galeano P., Peña D. (2000), Multivariate analysis in vector time series. *Resenhas*, 4, 383-404.

Gray A.H., Markel J.D. (1976), Distance measures for speech recognition, *IEEE Transactions on Acoustic, Speech, Signal Processing*, ASSP-24, 380-391.

Hand D.J. (1998), Data mining: Statistics and more?, *American Statistician*, 52, 112-117.

Huntala Y., Karkkainen J., Toivonen H. (1999), Mining for similarities in aligned time series using wavelts, *Data Mining and Knowledge Discovery: theory, tools and technology*.

Indyk P., Koudas N., Muthukrishnan S. (2000), Identifying representative trends in massive time series data sets using sketches, *26th Int. Conference on Very Large Data Bases*, 363-372.

Itakura F. (1975), Minimum prediction residual principle applied to speech recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-23, 67-72.

Kalpakis K., Gada D., Puttagunta V. (2001), Distance measures for effective clustering of ARIMA time series, *IEEE Int. Conference on Data Mining*, 273-280.

Kang W., Shiu J., Cheng C., Lai J., Tsao H., Kuo T. (1995), The application of cepstral coefficients and maximum likelihood method in EGM pattern recognition, *IEEE Transactions on Biomedical Engineering*, 42, 777-785.

Kassidas A., MacGregor J.F., Taylor P. (1998), Synchronization of batch trajectories using dynamic time warping, *American Institute of Chemical Engineers*, 44, 864-874.

Kedem B. (1980), Estimation of the parameters in stationary autoregressive

processes after hard limiting, *Journal of the American Statistical Association*, 75, 146-153.

Keogh E. (2002), Exact indexing of dynamic time warping, *28th International Conference on Very Large Data Bases*, Hong Kong, 406-417.

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

Keogh E., Lin J. (2005), Clustering of time series subsequences is meaningless: implications for previous and future research, *Knowledge and Information Systems*, 8, 154-177.

Keogh E., Lin J., Fu A., Van Herle H. (2005), Finding unusual medical time series: algorithms and applications, *CBMS2005-IEEE Transactions on Information Technology in Biomedicine*, 1-9.

Keogh E., Smyth P. (1997), A probabilistic approach to fast pattern matching in time series databases, *3rd Int. Conference on Data Mining and Knowledge Discovery*, 24-30.

Kovačić Z. J. (1996), Classification of time series with application to leading indicator selection, *Proceedings of IFCS-96*, 2, 204-207.

Li T., Li Q., Zhu S., Ogihara M. (2002), A survey on wavelet applications in data mining, *SIGKDD Explorations*, 4, 49-68.

Liu Lon-Mu, Bhatthacharyya S., Sclove S., Chen R., Lattyak W.J. (2001), Data mining on time series: an illustration using fast-food restaurant franchise data, *Computational Statistics & Data Analysis*, 37, 455-476.

Mélard G., Roy R. (1984), Sur un test d'égalité des autocovariances de deux séries chronologiques, *The Canadian Journal of Statistics*, 12, 333-342.

Morlini I. (2005), On the dynamic time warping for computing the dissimilarity between curves, in Vichi M., Monari P., Mignani S., Montanari A., eds., *New Developments in Classification and Data Analysis*, Springer, Berlin, 63-70.

Ng M.K., Huang Z. (1999), Data mining massive time series astronomical data: challenges, problems and solutions, *Information and Software Technology*, 41, 545-556.

Percival D., Walden A.T. (2000), *Wavelet methods for time series analysis*, Cambridge University Press, Cambridge.

Piccolo D. (1984), Una topologia per la classe dei processi ARIMA, *Statistica*, XLIV, 47-59.

Piccolo D. (1987), Problemi di confronto in rappresentazioni alternative di fenomeni dinamici, *Quaderni di Statistica e Econometria*, IX, 1-10.

Piccolo D. (1990), A distance measure for classifying ARIMA models, *Journal of Time Series Analysis*, 11, 153-164.

Piccolo D. (2007), Statistical issues on the AR metric in time series analysis, *Proceedings of the SIS Intermediate Conference*, Cleup, Padova, 221-232.

Priestley M.B. (1996), Wavelets and time-dependent spectral analysis, *Journal of Time Series Analysis*, 17, 85-103.

Rafei D., Mendelzon A.O. (1998), Efficient retrieval of similar time series sequences using DFT, *Proceedings of the 5-th Int. Conference on Foundation of Data Organization and Algorithms*, Kobe.

Ratanamahatana C.A., Keogh E. (2004), Making time-series classification more accurate using learned constraints, *4-th SIAM Int. Conference on Data Mining*, 15, 1-20.

Romano E., Scepi G. (2006), Integrating time alignment and Self-Organizing Maps for classifying curves, *Proceedings of KNEMO COMPSTAT 2006 Satellite Workshop*, Capri.

Sakoe, H., Chiba S. (1978), Dynamic programming algorithm optimization for spoken word recognition, *IEEE Transactions on Acoustic, Speech and Signal Processing*, 26, 143-165.

Scepi G., Milone G. (2007), Temporal data mining: clustering methods and algorithms, *Proceedings of CLADAG 2007*, EUM, Macerata, 73-76.

Slutsky E. (1937), The summation of random causes as the source of cyclic processes, *Econometrica*, 5, 105-146, (original russian paper, 1927).

Struzik Z.R., Siebes A. (1999), The Haar wavelet in the time series similarity paradigm, *3rd European Conference on Principles of Data Mining and Knowledge Discovery*, Springer Verlag, Prague, 12-22.

Thomson P.J., De Souza P. (1985), Speech recognition using LPC distance measures, in Hannan E.J., Krishnaiah P.R., Rao M.M., eds., *Handbook of Statistics*, Elsevier Science Publishers, 5, 389-412.

Tohkura Y. (1987), A weighted cepstral distance measure for speech recognition, *IEEE Transactions on Acoustic, Speech and Signal Processing*, ASSP-35, 1414-1422.

Vlachos M., Yu P., Castelli V., Meek C. (2005), Structural periodic measures for time series data, *Data Mining and Knowledge Discovery*, 12, 1-28.

Vlachos M., Hadejieleftheriou M., Gunopulos D., Keogh E. (2006), Indexing multidimensional time series, *The VBDL Journal*, 15, 1-20.

Zani S. (1981), Osservazioni sulle serie storiche multiple e l'analisi dei gruppi, in D. Piccolo, ed., *Analisi moderna delle serie storiche*, F. Angeli, Milano, 263-274.

Wang C., Wang X.S. (2000), Supporting content-based searches on time series via approximations, *12th Int. Conference on Scientific and Statistical Database Management*, 69-81.

Wang K., Gasser T. (1997), Alignment of curves by dynamic time warping, *Annals of Statistics*, 25, 1251-1276.

Wang X., Smith K. and Hyndman B. (2006), Characteristic-based clustering for time series data, *Data Mining and Knowledge Discovery*, 13, 335-364.

Xiong Y., Yeung D. (2004), Time series clustering with ARMA mixtures, *Pattern Recognition*, 37, 1675-1689.