J. Ital. Statist. Soc. (2000)
1-3, pp. 53-83

DISSIMILARITY MEASURES FOR TIME TRAJECTORIES

Pierpaolo D'Urso*
Università degli Studi «La Sapienza», Roma

Abstract

In this paper, by considering one of the possible geometric representations of time arrays in the 'object space' (space of the units), we analyze different dissimilarity measures between multivariate time trajectories of the units, which are classified, systematically, in various approaches, by taking into account their features. In particular, we define three classes of dissimilarities: the 'geometric' class, in which dissimilarities are built according to the geometrical features of the trajectories (instantaneous position, slope of the inter-temporal segments (velocity), concavity and convexity of each pair of inter-temporal segments (acceleration), polygonal line (shape), portion of area between each pair of trajectories); the 'correlative' class, in which dissimilarity measures that analyze the autocorrelation and cross-correlation functions of the univariate components of each multivariate trajectory are classified; the 'structural' class, containing dissimilarities which consider the structural aspects of the trajectories, such as the linear or polynomial trends and the seasonality of each univariate component. An empirical comparison is also included.

Keywords: Time trajectory, Dissimilarity measures, Position, Velocity, Acceleration, Inertia, Polygonal line, Area, Autocorrelation, Cross-correlation, Polynomial trend, Linear spline, Seasonality coefficient, Cluster analysis.

1. Introduction

In the 'classical' multivariate statistical analysis, the data are, generally, displayed in a matrix of the type '*units* × *variables*': $\mathbf{X} \equiv \{x_{ij} : i = 1, I; j = 1, J\}$, where I

* *Address for correspondence*: Dipartimento di Statistica, Probabilità e Statistiche Applicate, Università degli Studi di Roma «La Sapienza», P.le A. Moro, 5 – I-00185, Roma, Italy. E-mail: pierpaolo.durso@uniroma1.it

denotes the units (objects) on which J variables are observed. However, in the real cases, the information about the phenomenon is often more complex. In this case, we have the so-called *multi-way data* (Carroll and Arabie, 1980; Kiers, 2000). The statistical analysis that allows to elaborate these data is just the *multi-way statistical analysis* which represents a recent development of the classical multivariate statistical analysis (Coppi and Bolasco, 1989).

In particular, following an exploratory approach, here we consider the *three-way data* of the type 'units × variables × occasions', in which the occasions are times, the so-called *time array*.

The aim of this paper is to analyze, from an algebraic and geometric point of view, the time array defining the so-called multivariate *time trajectory* and illustrate different kinds of dissimilarity measures between multivariate time trajectories, which can be used, for instance, to cluster the time trajectories into homogeneous classes.

The paper is structured as follows. In Section 2, we analyze the algebraic-geometric features of the time arrays. In Section 3, we briefly introduce the clustering problem for time trajectories. Then, in Section 4, we classify, systematically, different typologies of dissimilarities between time trajectories (analyzing also data pre-processing and weighting criteria, in Section 4.1) and, successively, examine the dissimilarities: geometrical (Section 5), correlative (Section 6) and structural (Section 7). The considered dissimilarity measures are compared, by analyzing an empirical dataset (Section 8). Some concluding remarks are finally provided (Section 9).

2. Multivariate time arrays: algebraic and geometric aspects

A multivariate time array can be represented in the following way:

$$\mathbf{X} \equiv \{x_{ijt} : i = 1, I; j = 1, J; t = 1, T\} \tag{1}$$

where i ($i = 1,I$) indicates the generic unit (object), j ($j = 1,J$) the variable, and t ($t = 1,T$) the generic time; x_{ijt} represents the j-th variable observed in the i-th unit at time t (Fig. 1).

In a time array, we have the *same* units in which we observe the *same* variables at different times. In Figure 1, the entities along the horizontal axis are indicated by the first index (i), those along the vertical axis by the second index (j) and those along the depth axis by the third index (t) (Kiers, 2000). «The three sets of entities define the three 'ways' or three 'dimensions' of the three-way array. Because the term 'way' is not very specific and the term 'dimension' may be confused with the concept of dimensions of a factor space in factor analysis [...], we use the term 'mode' to refer to a set of entities» (Kiers, 2000).

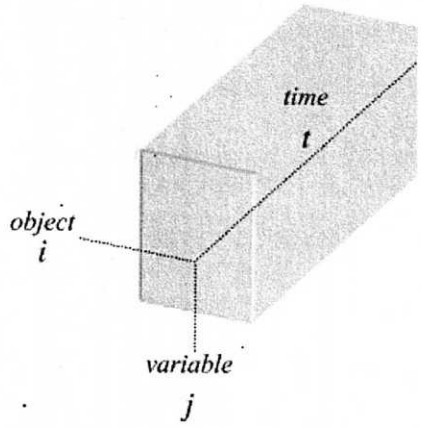


Fig. 1 – Representation of the time array.

In several real cases, the basic information presents a time three-way structure. For example, we can observe the same socio-economic variables on a panel of consumers during different years or the same medical indicators on a group of patients at subsequent times.

In this paper, we study the three-way time array $\mathbf{X}$ in which the variables are quantitative. Then, by considering an exploratory approach, different algebraic structures of time array are formalized.

The time array $\mathbf{X}$ can be represented by a bi-dimensional matrix (stacked matrix) by combining two of the three indices i, j, t on the rows and assigning the other index to the columns. The various types of stacked matrices are respectively: $\mathbf{X}_i \equiv \{x_{ijt} : j = 1, J; t = 1, T\}$, with dimension $T \times J$, $\mathbf{X}_j \equiv \{x_{ijt} : i = 1, J; t = 1, T\}$, with dimension $T \times I$ and $\mathbf{X}_t \equiv \{x_{ijt} : i = 1, J; j = 1, J\}$, with dimension $I \times J$. The $\mathbf{X}$ can be seen as a set of bi-dimensional matrices which represent the *slices* of the 'parallelepiped of data', which represents $\mathbf{X}$ (Fig. 2). In particular, we have I matrices $\mathbf{X}_i$ (case a), J matrices $\mathbf{X}_j$ (case b) and T matrices $\mathbf{X}_t$ (case c).

Let us now analyze in detail the different ways to represent $\mathbf{X}$ in a bi-dimensional form, matching in the different modes the indices i, j, t.

We denote by $\mathbf{X}(I, J, T)$ the time array when the three indices are not pooled in some way.

By matching the two indices i and t in the rows and putting the index j in the columns, we obtain the bi-dimensional matrices $\mathbf{X}(IT, J)$ and $\mathbf{X}(TI, J)$. In particular, in the stacked matrix $\mathbf{X}(IT, J)$ the index t varies within the index i; in this case i undertakes the part of a group index (and t the part of 'elementary unit'). Vice versa, in the stacked matrix $\mathbf{X}(TI, J)$ the index i varies within t and then, in this case, it plays the role of a group index. So $\mathbf{X}(IT, J)$ represents the group-units as a

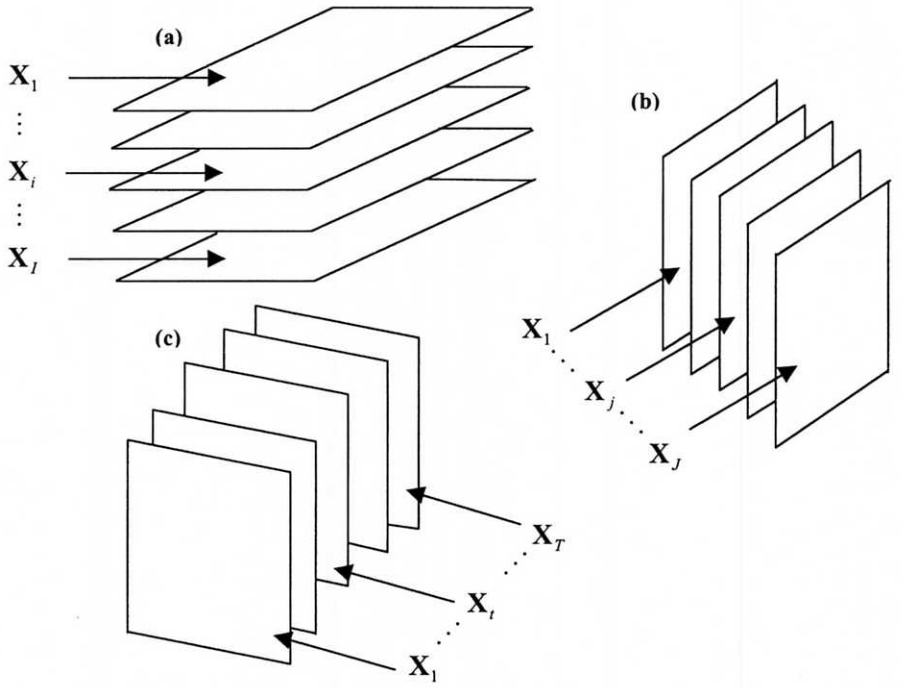


Fig. 2 – Representation of the slices $\mathbf{X}_i$ ($i = 1, I$), $\mathbf{X}_j$ ($j = 1, J$) and $\mathbf{X}_t$ ($t = 1, T$).

whole, and for each of them the realization of the J variables over T times is observed. $\mathbf{X}(T\,I, J)$ represents the group-times as a whole, and for each of them the realization of the J variables in the various I units is observed. Similar formulations can be obtained for $\mathbf{X}(I\,T, J), \mathbf{X}(T\,I, J)$ and also $\mathbf{X}(J\,T, I), \mathbf{X}(T\,J, I), \mathbf{X}(I\,J, T)$, $\mathbf{X}(J\,I, T), \mathbf{X}(I, J\,T), \mathbf{X}(I, T\,J), \mathbf{X}(J, I\,T), \mathbf{X}(J, T\,I), \mathbf{X}(T, I\,J), \mathbf{X}(T, J\,I)$ (Kiers, 2000).

The data time array $\mathbf{X}$ can be geometrically represented on a suitable vectorial space. By indicating the elements of one of the three possible classification modes as vectors of a vectorial space, defined with respect to the other ones, we have:

1) the *space of the units* $\Re^{J+1}$;
2) the *space of the variables* $\Re^{I+1}$;
3) the *space of the time occasions* $\Re^{I+J}$.

1) In the space $\Re^{J+1}$ (the first J dimensions correspond to the J variables and the last dimension is referred to the time), each unit i is represented, for each time t, by the vector $\mathbf{u}_{it} = (x_{i1t}, \ldots, x_{ijt}, \ldots, x_{iJt}, t)'$, $i = 1, I$; $t = 1, T$. For fixed t, the matrix $\mathbf{X}_t$ is represented by the scatter $N_I(t) \equiv \{\mathbf{u}_{it}\}_{i=1,J}$. Then $\{N_I(t) \equiv \{\mathbf{u}_{it}\}_{i=1,J}\}_{t=1,T}$ represents

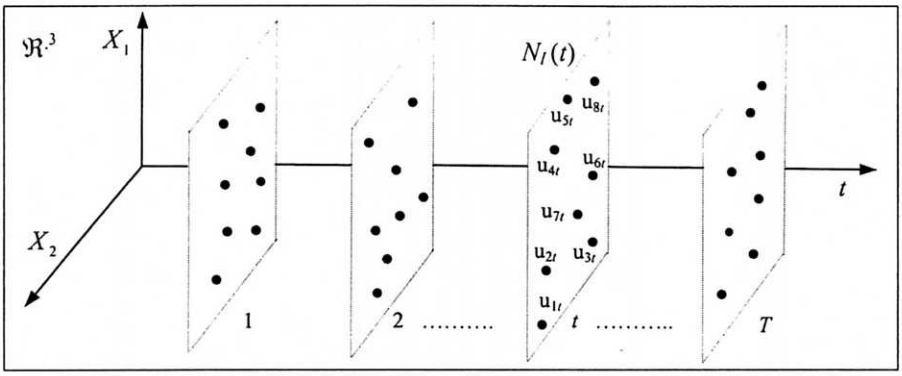


Fig. 3 – Geometric representation of the scatter $N_I(t)$ in the space $\Re^{J+1}$ for $I = 8$ and $J = 2$.

the set of scatters located on T hyperplanes parallel to the co-ordinate sub-space $\Re^J$ (Fig. 3).

Conversely, for fixed i, the matrix $\mathbf{X}_i$ is represented by the scatter $N_T(i) \equiv \{\mathbf{u}_{it}\}_{t=1,T}$ which describes the *time trajectory* of the i-th unit across the time. Then, the set of scatters $\{N_T(i) \equiv \{\mathbf{u}_{it}\}_{t=1,T}\}_{i=1,I}$ represents the set of the time trajectories of the I units. Each time trajectory crosses the T hyperplanes (see Figure 4) (Coppi and Zannella, 1978).

2) The space $\Re^{I+1}$ is a vectorial space in which the first I dimensions correspond to the I units and the last dimension is referred to the time. In this space $\Re^{I+1}$ the time trajectories of the variables can be defined.

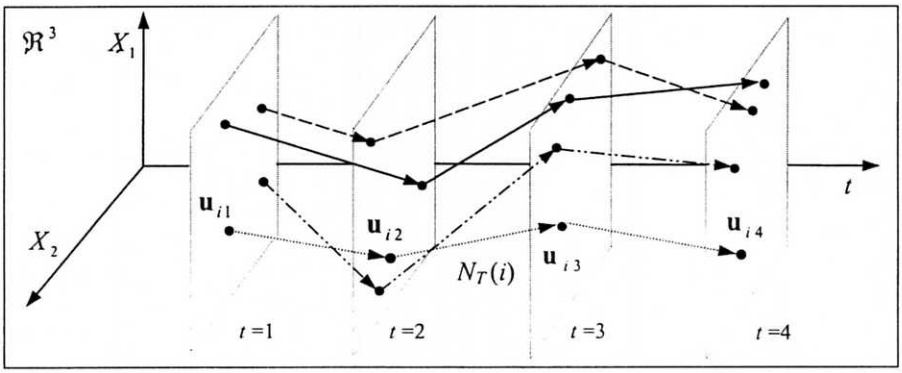


Fig. 4 – Geometric representation of the scatter $N_T(i)$ in the space $\Re^{J+1}$ for $J = 2, I = 4$ and $T = 4$.

3) The data space of the time occasions $\Re^{I \times J}$ is the space of the matrices in which the matrices units × variables, observed at different times, can be represented.

In this paper, we consider only the case in which the time array $\mathbf{X}$ is represented in the space of the units $\Re^{J+1}$ and then we analyze the behaviour of the time trajectories of units in this data space.

3. Clustering of multivariate time trajectories

We can identify different kinds of time trajectories: quantitative, qualitative, mixed; univariate and multivariate; discrete and continuous; synchronous and asynchronous (the variables are observed at the same or different times); with constant or variable time-support (with equidistant or non equidistant inter-temporal intervals).

In this paper, we analyze discrete multivariate synchronous quantitative time trajectories.

Notice that in the following the time-support is indicated by $\mathbf{T} \equiv \{t_1, \ldots, t_s, \ldots, t_T\}$. Therefore, the time array is $\mathbf{X} \equiv \{x_{ijt_s} : i = 1, I; j = 1, J; s = 1, T\}$.

Moreover, we observe that, in the following, in the formalization of a time trajectory we do not consider the co-ordinate t, because it does not affect the computation of dissimilarities. Thus, $\mathbf{X}_i \equiv \{x_{it_s} : s = 1, T\}$ represents the i-th time trajectory, where $\mathbf{x}_{it_s} \equiv (x_{i1t_s}, \ldots, x_{ijt_s}, \ldots x_{iJt_s})'$, $i=1,I$, $s=1,T$.

The main reasons for classifying time trajectories are the identification of trajectories with similar structural and dynamical characteristics, 'typical' time trajectories, each representing a consistent subset of the trajectories, and 'anomalous' time trajectories, which are characterized by a different behaviour with respect to the other ones (see Figure 5).

The clustering problem of multivariate time trajectories is of interest in several fields (Carlier, 1986; Costanzo and D'Urso, 2000; D'Urso and Vichi, 1998; Kakizawa *et al.*, 1998; Košmelj and Batagelj, 1990; Košmelj and Lachet, 1983; MacGregor *et al.*, 1994; Nomikos and MacGregor, 1995; Piccolo, 1990; Zani, 1983). For instance, in economics it is very interesting to classify the Countries according to whether they show similar or anomalous dynamics with respect to the principal economic aggregates. In marketing research, it is of great importance for the decisional process of the companies, to identify the consumers who, in a particular period, have similar preferences, perceptions, purchase behaviours of makes and products. Also in medicine, physics, engineering, hydrology the identification of similar trajectories is interesting (for instance, in the diagnosis of electrocardiograms or electroencephalograms, sound recognition, studies of seismic waves and hydrological variables and so on).

In order to apply a clustering procedure for classifying time trajectories, we must define suitable dissimilarity measures. For this reason, in the following Sections, we analyze, systematically, different types of these measures. Some of these

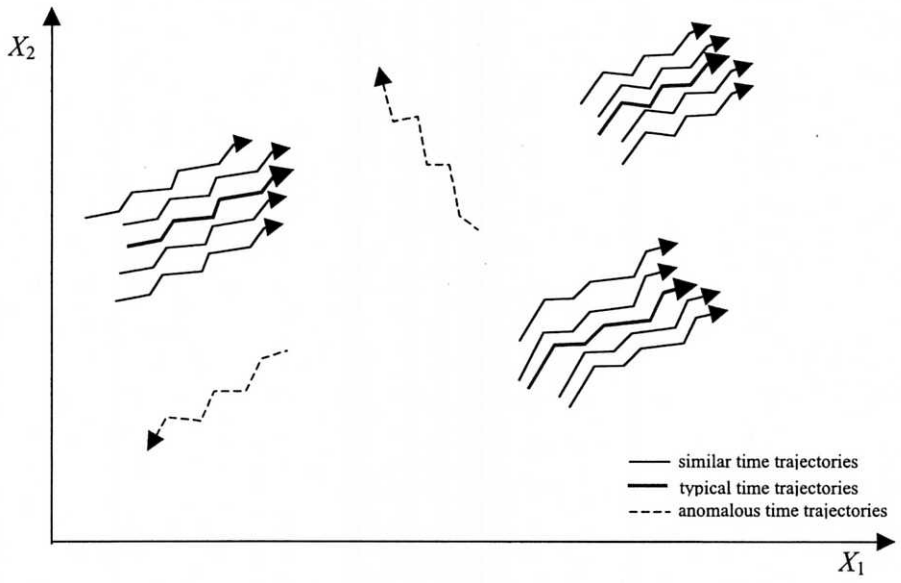


Fig. 5 – Example of similar, typical and anomalous time trajectories in $\Re^2$.

measures have already been suggested in literature (Bohte *et al.*, 1980; Carlier, 1986; Coppi and D'Urso, 2001; D'Urso and Vichi, 1998; Košmelj, 1986, Raveh, 1981; Raveh and Tapiero, 1980; Saporta and Lavallard, 1996; Zani, 1983); the other ones ('polygonal', 'areal' and 'trend' dissimilarities) are new proposals.

4. Dissimilarity measures between time trajectories: different approaches

We can define dissimilarity measures between time trajectories, following two formulations: the exploratory and the inferential ones.

In the former case, the dissimilarities are defined by analyzing the observed trajectories without probabilistic assumptions. In the latter case, we consider an inferential approach, by building stochastic models, for instance autoregressive, that describe the dynamics of the trajectories (see, for instance, Kakizawa *et al.*, 1998; Maharaj, 1996, 1999, 2000; Martin, 2000; Piccolo, 1990).

By considering the exploratory formulation, different kinds of dissimilarity between time trajectories can be defined. In particular, we classify systematically these measures in three different approaches:

1) *geometric approach*;

2) *correlative approach*;
3) *structural approach*.

Briefly, in the first case, we carry out a comparative analysis of the trajectories by considering dissimilarity measures the take into account their geometric aspects (position, slope, shape, etc.) in the vectorial space $\Re^{J+1}$. In the second case, the time dependence (in autocorrelative and cross correlative sense) that characterizes the trajectories is taken into account. In the latter case, we suppose that each time trajectory is decomposed in the usual common features (trend, cycle, seasonality (for sub-annual observations), signals which can be contaminated by random oscillations. In particular, in the following, we define different dissimilarities that compare the trajectories by considering the trend and seasonal components.

Before explaining the previous approaches, we analyze in the following Sub-Section some important observations concerning the data displayed in the time array **X**.

4.1. *Data pre-processing and weighting criteria*

In the utilization of dissimilarity measures and in clustering problems, attention should be paid to problems of heterogeneity, with particular reference to the variables (different variances and/or units of measurement) (*data pre-processing*) and of possible weighting of the units, variables and times (*weighting criteria*).

• *Data pre-processing:* An appropriate pre-processing of the data may be required, such as normalization/standardization. In this connection, we can consider some types of pre-processing procedures (Kiers, 2000; Harshman and Lundy, 1984): 1) *Centering* of the data, by taking into account the average of the data. For instance, we can use the following transforms: $\tilde{x}_{ijt_s} = x_{ijt_s} - \bar{x}_{.jt_s}$, $\tilde{x}_{ijt_s} = x_{ijt_s} - \bar{x}_{..t_s}$, where the subscript dot is used to indicate the mean across $i = 1,I$ ($\bar{x}_{.jt_s}$) and across $i = 1,I$ and $j = 1,J$ ($\bar{x}_{..t_s}$). 2) *Normalization* of the data, by dividing the data x_{ijt_s} by the normalization factor $\bar{x}_{.j.}$, obtaining: $\tilde{\tilde{x}}_{ijt_s} = \dfrac{x_{ijt_s}}{\bar{x}_{.j.}}$. We observe that the normalization of the data is particularly indicated for coping with problems of heterogeneity of units of measurement and/or of size of the variables. In fact, by means of the suggested transform of the data we get a sort of index numbers which allow us to keep the information due to the variability of the different variables (within and across the occasions), while getting rid of possible differences in their overall mean value or in their type of measurement. 3) *Standardization* of the data, by using, for instance,

$$x^{*}_{ijt_s} = \frac{\tilde{x}_{ijt_s}}{\frac{1}{T}\sqrt{\sum_{i=1}^{I} \tilde{x}^{2}_{ijt_s}}}.$$

Then, the corresponding transformed time arrays are:

1) $\tilde{\mathbf{X}} \equiv \{\tilde{x}_{ij t_s} : i = 1, I; j = 1, J; s = 1, T\}$ (centered time array)

2) $\tilde{\tilde{\mathbf{X}}} \equiv \{\tilde{\tilde{x}}_{ij t_s} : i = 1, I; j = 1, J; s = 1, T\}$ (normalized time array)

3) $\mathbf{X}^* \equiv \{x^*_{ij t_s} : i = 1, I; j = 1, J; s = 1, T\}$ (standardized time array).

Moreover, in order to stabilize the variability of the data, suitable transformations (logarithmic, etc.) for the original data can be also taken into account (Box and Jenkins, 1976).

In conclusion, when choosing a specific transform of the original data, we should consider the particular informational features we would like to keep in or eliminate from the analysis.

• *Weighting criteria*: By considering the time array $\mathbf{X} \equiv \mathbf{X}(I,J,T)$, in correspondence with the different indices I, J, T, we can suggest for the units i ($i = 1,I$) the weights w_i, $i = 1, I$, contained in the $(I \times I)$-diagonal matrix $\mathbf{W}$, for the variables j ($j = 1,J$) the weights $\tilde{w}_j$, $j = 1, J$, contained in the $(J \times J)$-diagonal matrix $\tilde{\mathbf{W}}$ and for the times t_s ($s = 1,T$) the weights $\tilde{\tilde{w}}_{t_s}$, $s = 1, T$, contained in the $(T \times T)$-diagonal matrix $\tilde{\tilde{\mathbf{W}}}$. Then, from the weight matrices $\mathbf{W}$, $\tilde{\mathbf{W}}$, and $\tilde{\tilde{\mathbf{W}}}$ we can define the weight array $\hat{\mathbf{W}}$ of the time array $\mathbf{X}$: $\hat{\mathbf{W}} = \tilde{\tilde{\mathbf{w}}} \otimes (\mathbf{w} \otimes \tilde{\mathbf{w}}')$ where $\mathbf{w} = (w_1,\ldots, w_i,\ldots, w_I)'$, $\tilde{\mathbf{w}} = (\tilde{w}_1,\ldots, \tilde{w}_j,\ldots, \tilde{w}_J)'$, $\tilde{\tilde{\mathbf{w}}} = (\tilde{\tilde{w}}_{t_1},\ldots, \tilde{\tilde{w}}_{t_s},\ldots, \tilde{\tilde{w}}_{t_T})'$ and $\otimes$ denotes the Kronecker product.

Notice that, we can also consider normalized nonnegative weights, by considering, for instance for the time weights $\tilde{\tilde{w}}_{t_s}$, $s = 1, T$, the constraints: $\tilde{\tilde{w}}_{t_s} \geq 0$ ($s = 1,T$), $\sum_{s=1}^{T} \tilde{\tilde{w}}_{t_s} = 1$. Moreover, we can define weighting systems for each pair of times; in this case the intertemporal weights are denoted by $\tilde{\tilde{w}}_{t_s t_{s+k}}$, $s = 1, T$; $k = \pm 1, \pm 2,\ldots$

5. The geometric approach

Following this approach, we compare the time trajectories by defining dissimilarity measures which take into account the geometric aspects of the time trajectories; in particular the instantaneous *position*, the *slope* of each segment between each pair of contiguous instantaneous positions, the *concavity* or *convexity* of each pair of contiguous segments, the *polygonal line* that characterizes the shape of the trajectory, the portion of *area* between each pair of trajectories.

5.1. *'Cross sectional-longitudinal' dissimilarities*

D'Urso and Vichi (1998) have proposed a cross sectional dissimilarity, which compares the instantaneous position of each pair of trajectories, and two longitudinal dissimilarities, based on the concepts of *velocity* and *acceleration* of a time

trajectory. Consequently, the authors have suggested a compromise dissimilarity, that is a conic combination of the three previous dissimilarities, where the weight of each dissimilarity-component is determined by considering an objective procedure.

To compare the trajectories by taking into account their *position* in the data space, the authors suggested the following cross sectional (instantaneous) dissimilarity:

$$_1d_{il}^2 = \sum_{s=1}^{T}\sum_{j=1}^{J}(x_{ijt_s} - x_{ljt_s})^2 = \sum_{s=1}^{T}\left\|\mathrm{x}_{it_s} - \mathrm{x}_{lt_s}\right\|^2. \qquad (2)$$

Comparative measures based on the evolutive behaviour of the time trajectories are defined by considering two characteristics of the time trajectories: the *velocity* and the *acceleration*.

Definition 1

The *velocity* of the i-th time trajectory, for the j-th variable, measures the variation of the i-th time trajectory in the interval $[t_{s-1}, t_s]$. Then, for each time trajectory i, the velocity in $[t_{s-1}, t_s]$, with respect to the j-th variable, is:

$$v_{ijt_s} = \frac{x_{ijt_s} - x_{ijt_{s-1}}}{t_s - t_{s-1}}, \qquad s = 2,T.$$

In particular, $v_{ijt_s} > 0$ if the i-th time trajectory, for the j-th variable, presents an increment of its position in the interval $[t_{s-1}, t_s]$; vice versa, $v_{ijt_s} < 0$ if the i-th time trajectory, for the j-th variable, presents a decrement of its position in the same interval; $v_{ijt_s} = 0$ if the variation is null.

Definition 2

The *acceleration* measures the variation of the velocity of a time trajectory in a time interval. In particular, for each time trajectory (and for the j-th variable), the acceleration in the interval $[t_{s-2}, t_s]$ is:

$$a_{ijt_s} = \frac{v_{ijt_s} - v_{ijt_{s-1}}}{t_s - t_{s-2}}, \qquad s = 3,T.$$

In particular, $a_{ijt_s} > 0$ if the i-th time trajectory, for the j-th variable, presents a positive variation of velocity in the interval $[t_{s-2}, t_s]$; vice versa, $a_{ijt_s} < 0$ if the i-th time trajectory, for the j-th variable, presents a negative variation of velocity in the same interval; $a_{ijt_s} = 0$ if the variation is null.

The velocity pertaining to each pair of successive time points $[t_{s-1}, t_s]$, represents geometrically the *slope* of the straight line passing through them: if the velocity is negative (positive) the slope will be negative (positive) and the angle made by each segment of the trajectory with the positive direction of the t-axis will be obtuse (acute). Geometrically, the acceleration of each pair of segments of a time trajectory represents in each interval $[t_{s-2}, t_s]$ its *convexity* or *concavity*: if the acceleration is positive (negative) the trajectory of the two segments is convex (concave).

Then, the evolutive dissimilarities, based on velocity and acceleration, are:

$$ {}_2d_{il}^2 = \sum_{s=2}^{T}\sum_{j=1}^{J}(v_{ijt_s} - v_{ljt_s})^2 = \sum_{s=2}^{T}\left\|\mathbf{v}_{it_s} - \mathbf{v}_{lt_s}\right\|^2, \tag{3} $$

where $\mathbf{v}_{it_s} = (v_{i1t_s}, \ldots, v_{ijt_s}, \ldots, v_{iJt_s})'$ (analogously $\mathbf{v}_{lt_s}$), called *longitudinal-velocity dissimilarity measure*;

$$ {}_3d_{il}^2 = \sum_{s=3}^{T}\sum_{j=1}^{J}(a_{ijt_s} - a_{ljt_s})^2 = \sum_{s=3}^{T}\left\|\mathbf{a}_{it_s} - \mathbf{a}_{lt_s}\right\|^2, \tag{4} $$

where $\mathbf{a}_{it_s} = (a_{i1t_s}, \ldots, a_{ijt_s}, \ldots, a_{iJt_s})'$ (analogously $\mathbf{a}_{lt_s}$), called *longitudinal-acceleration dissimilarity measure*.

Remark 1

Notice that if the time support is $\tilde{\boldsymbol{T}} \equiv \{1,\ldots,t,\ldots,T\}$, the dissimilarities (3) and (4) coincide with the dissimilarities proposed by Carlier (1986).

Let us consider the example shown in Figure 6.

In case (A), by applying, for instance, the longitudinal dissimilarity (3) we obtain that the two bivariate time trajectories have the same dynamic. But, in this way, we do not take into account their different positions on the plane. In case (B), by considering the cross sectional dissimilarity (2) we have that the trajectories present a similar position. However, the trajectories have a diverse evolution. Then, if we want to take into account simultaneously the position and the evolution, a good solution is the utilization of a compromise dissimilarity. For this reason D'Urso and Vichi have suggested a mixed dissimilarity that takes into account simultaneously the cross sectional and the two longitudinal dissimilarities:

$$ {}_{1M}d_{il}^2 = \sum_{r=1}^{3}\alpha_r d_{il}^2,\ \alpha_r \geq 0. \tag{5} $$

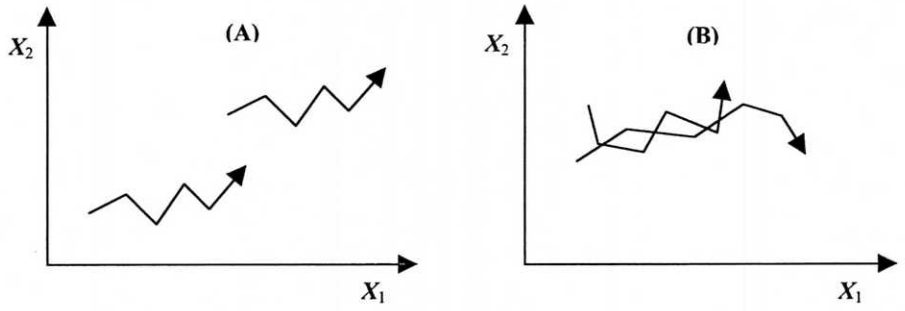


Fig. 6 – Example of time trajectories.

Remark 2

A normalized version of the dissimilarity ${}_r d_{il}^2$ ($r = 1,2,3$) is ${}_r \tilde{d}_{il}^2 = \frac{1}{\pi_r} {}_r d_{il}^2$ ($r = 1,2,3$), where $\pi_r = var({}_r d_{il}^2)$. Alternatively, we can consider the *range* of the respective dissimilarity.

In (5) the weights α_r ($r = 1,2,3$) are defined by considering an *objective* criterion, based on the maximization of the variance of the compromise dissimilarity (5) with the constraint $\alpha'\alpha = 1$, where $\alpha = (\alpha_1,\alpha_2,\alpha_3)'$ (see D'Urso and Vichi, 1998).

Remark 3

Notice that Carlier (1986) has proposed a compromise dissimilarity by taking into account only the cross sectional and the longitudinal-velocity dissimilarities (with time support $\tilde{T}$) in which the weights are fixed *subjectively*. Then, this mixed dissimilarity can be considered a particular case of (5).

Other mixed dissimilarities have been proposed in literature. Coppi and D'Urso (2000), by considering the concepts of position, velocity and acceleration of a trajectory (D'Urso and Vichi, 1998) have extended the procedure proposed by Saporta and Lavallard (1996), based on the concept of *inertia*.

Definition 3

We define *configuration* a set $\{a_{t_1},\ldots, a_{t_s},\ldots, a_{t_T}\}$ where $a_{t_1} = 0$, *i.e.*, the first co-ordinate, as in the Saporta-Lavallard procedure, is always a *position-co-ordinate* and, $\forall\ s = 2, T : a_{t_s} = 0$ if in the t_s-th instant we consider a *position-co-ordinate*; $a_{t_s} = 1$ if in t_s-th instant we consider an evolutive-co-ordinate (velocity); $a_{t_s} = 2$ if in the t_s-th instant we consider an evolutive-co-ordinate (acceleration).

Definition 4

For the configuration $\{a_{t_1},\ldots, a_{t_s},\ldots, a_{t_T}\}$, the total inertia I_T^* is decomposable as follows:

$$I_T^* = I_{T_1}^* + I_{T_2}^* + I_{T_3}^*,$$

where $I_{T_1}^* = \sum_{i=1}^{I} \sum_{\{t_s:a_{t_s}=0\}} \left\|\mathbf{x}_{it_s} - \bar{\mathbf{x}}_{t_s}\right\|^2$, defined by considering the positions of the objects with respect to the barycenters of the objects; $I_{T_2}^* = \sum_{i=1}^{I} \sum_{\{t_s:a_{t_s}=1\}} \left\|\mathbf{v}_{it_s} - \bar{\mathbf{v}}_{t_s}\right\|^2$, defined by considering the velocities of the objects with respect to the barycenters of the velocities; $I_{T_3}^* = \sum_{i=1}^{I} \sum_{\{t_s:a_{t_s}=2\}} \left\|\mathbf{a}_{it_s} - \bar{\mathbf{a}}_{t_s}\right\|^2$, defined by considering the accelerations of the objects with respect to the barycenters of the accelerations.

Then, in order to compute the mixed dissimilarity, analogously to Saporta-Lavallard, the procedure suggested by Coppi and D'Urso (2001) consists of two steps.

1) We determine the optimal configuration $\{a_{t_1},\ldots, a_{t_s},\ldots, a_{t_T}\}$ considering the criterion:

$$\min_{\{a_{t_1},\ldots,a_{t_s},\ldots,a_{t_T}\}} d^2(I_{T_1}^*, I_{T_2}^*, I_{T_3}^*)$$

where $d^2(I_{T_1}^*, I_{T_2}^*, I_{T_3}^*) = d^2(I_{T_1}^*, I_{T_2}^*) + d^2(I_{T_1}^*, I_{T_3}^*) + d^2(I_{T_2}^*, I_{T_3}^*)$ is the squared *three-way distance* among $I_{T_1}^*$, $I_{T_2}^*$ and $I_{T_3}^*$ (Joly and Le Calvè, 1995).

2) After having obtained the optimal configuration, we compute the following compromise dissimilarity:

$$_{2M}d_{il}^2 = \sum_{\{t_s:a_{t_s}=0\}} \left\|\mathbf{x}_{i\,t_s} - \mathbf{x}_{l\,t_s}\right\|^2 + \sum_{\{t_s:a_{t_s}=1\}} \left\|\mathbf{v}_{i\,t_s} - \mathbf{v}_{l\,t_s}\right\|^2 + \sum_{\{t_s:a_{t_s}=2\}} \left\|\mathbf{a}_{i\,t_s} - \mathbf{a}_{l\,t_s}\right\|^2. \quad (6)$$

The measure (6) takes into account simultaneously the cross sectional and evolutive (velocity and acceleration) information of the time trajectories by considering only T co-ordinates and presents objectively determined weights pertaining to the different components (instantaneous and longitudinal).

The differences between the Coppi-D'Urso dissimilarity and the Saporta-Lavallard dissimilarity are: 1) the first dissimilarity takes into account also the longitudinal-acceleration dissimilarity; the second one does not. 2) The criteria to de-

termine the optimal configuration are different. In fact Saporta and Lavallard have suggested the following criterion: min $|I^*_{T_1} - I^*_{T_2}|$, where $I^*_{T_1}$ and $I^*_{T_2}$ indicate, respectively, the so-called *cross sectional and longitudinal inertia*.

Remark 4

We underline that the different dissimilarities considered in this Section are based on the cross sectional (positional or instantaneous) and longitudinal (velocity and acceleration) features of the trajectories. Therefore, the two considered mixed dissimilarities represent suitable combinations of the previous characteristics, obtained by considering two criteria: the *maximization of the variance* of the compromise dissimilarity (5) (for the mixed dissimilarity suggested by D'Urso and Vichi (1998)) and the *minimization of the three-way distance* among the inertias $I^*_{T_1}$, $I^*_{T_2}$ and $I^*_{T_3}$ (for the mixed dissimilarity suggested by Coppi and D'Urso, 2000).

In conclusion, we can observe that the instantaneous and the longitudinal (velocity and acceleration) dissimilarity measures differ essentially because of the way they deal with the evolutive aspects of the trajectories. The former one looks at the instantaneous profile, thereby capturing the dissimilarity between the locations of the trajectories at the various time occasions. The latter measures, based on the velocity and acceleration features, emphasize, respectively the dissimilarity between the variations (geometrically, the slopes) observed for each pair of successive times (for the velocity) and the variations (geometrically, the convexity or concavity) observed for each pair of successive time intervals (for the acceleration). Obviously, the above aspects constitute two different pieces of information (cross sectional and longitudinal information) embodied in the data set. When both are considered important the mixed dissimilarity measure can be chosen for the clustering task. In this latter measure the acceleration, velocity and instantaneous components are weighted differently. The estimate of their respective weights is computed by considering the criteria suggested by D'Urso and Vichi (1998) and Coppi and D'Urso (2001).

5.2. *'Polygonal' dissimilarity*

In order to define a dissimilarity between time trajectories, another suggestion can be considered based on the *polygonal* representation, of each univariate component of the multivariate time trajectories. From a geometric point of view, a univariate time trajectory $\mathbf{x}_i = (x_{it_1}, \ldots, x_{it_s}, \ldots, x_{it_T})'$ can be represented exactly by the *polygonal* with vertex-points (t_s, x_{it_s}), $s = 1, T$ (see Figure 7).

The algebraic expression of the polygonal representation of a univariate time trajectory is:

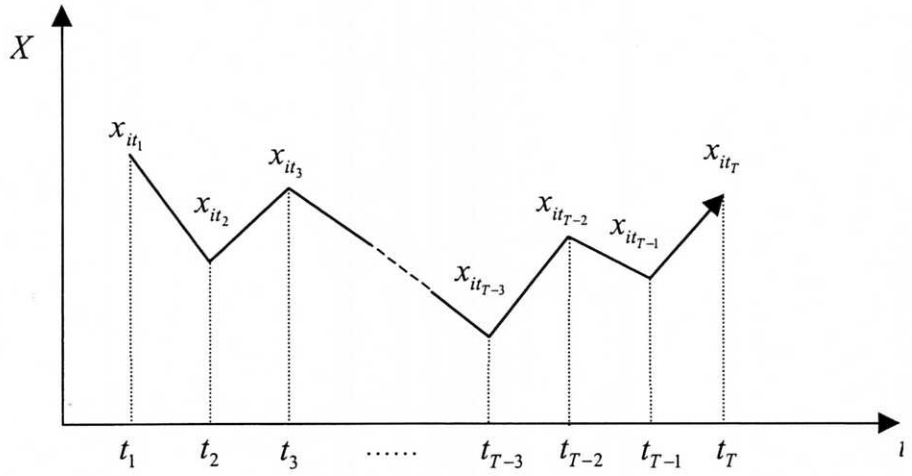


Fig. 7 – Polygonal representation of a univariate time trajectory.

$$x_{it_r} = \sum_{s=1}^{T} g_{it_s} |t_r - t_s|, (r = 1,T)$$

where g_{it_s} are the *polygonal coefficients* that characterize the dynamic of the trajectories.

The polygonal coefficients g_{it_s} are obtained by solving the previous system of equations. In fact, setting $G = \dfrac{x_{it_1} + x_{it_T}}{t_T - t_1}$, we have (Benedetti, 1972):

$$g_{it_1} = \frac{1}{2}\left(\frac{x_{it_2} - x_{it_1}}{t_2 - t_1} + G\right);$$

$$g_{it_{s+1}} = \frac{1}{2}\left(\frac{x_{it_{s+2}} - x_{it_{s+1}}}{t_{s+2} - t_{s+1}} - \frac{x_{it_{s+1}} - x_{it_s}}{t_{s+1} - t_s}\right), s = 1,T-2;$$

$$g_{it_T} = \frac{1}{2}\left(G - \frac{x_{it_T} - x_{it_{T-1}}}{t_T - t_{T-1}}\right).$$

The coefficients g_{it_s} represent the semi-difference of the velocities (defined in 5.1) of the time trajectories, calculated in the intervals $[t_s, t_{s+1}]$ and $[t_{s+1}, t_{s+2}]$. In

particular, the parameters g_{it_s} indicate the 'weights' of some time circumstances that determine the vertices of the polygonal line representing each univariate component of the trajectories. Moreover, the coefficients g_{it_s} supply indications on the 'oscillations' of each univariate component. In fact, since, by definition, in a trajectory there is an oscillation if a tern of consecutive values, $x_{it_{s-1}}$, x_{it_s}, $x_{it_{s+1}}$, such as $x_{it_{s-1}} < x_{it_s}$, $x_{it_s} > x_{it_{s+1}}$ or $x_{it_{s-1}} > x_{it_s}$, $x_{it_s} < x_{it_{s+1}}$, exists we have that the existence of vertices in the polygonal line (which depends on the non negative values of the polygonal coefficients g_{it_1}, $g_{it_{s+1}}$ and g_{it_T}) is the necessary condition for the existence of the oscillations. Then, the oscillations of the trajectories depend on the polygonal coefficients.

By considering a matrix formulation, the polygonal representation of the i-th multivariate time trajectory is:

$$\mathbf{X}_i' = \mathbf{C}\,\mathbf{G}_i \tag{7}$$

where $\mathbf{C}$ is a symmetrical and invertible matrix, defined as follows:

$$\mathbf{C} = \begin{pmatrix} 0 & |t_1 - t_2| & \cdots & |t_1 - t_T| \\ |t_2 - t_1| & 0 & \cdots & |t_2 - t_T| \\ \vdots & \vdots & \ddots & \vdots \\ |t_T - t_1| & |t_T - t_2| & \cdots & 0 \end{pmatrix}$$

and $\mathbf{G}_i$ the matrix of the polygonal coefficients:

$$\mathbf{G}_i = \begin{pmatrix} g_{i1t_1} & \cdots & g_{ijt_1} & \cdots & g_{iJt_1} \\ \vdots & & \vdots & & \vdots \\ g_{i1t_s} & \cdots & g_{ijt_s} & \cdots & g_{iJt_s} \\ \vdots & & \vdots & & \vdots \\ g_{i1t_T} & \cdots & g_{ijt_T} & \cdots & g_{iJt_T} \end{pmatrix} = \begin{pmatrix} \mathbf{g}_{it_1}' \\ \vdots \\ \mathbf{g}_{it_s}' \\ \vdots \\ \mathbf{g}_{it_T}' \end{pmatrix}$$

where $\mathbf{g}_i = (g_{it_1}, \ldots, g_{it_s}, \ldots, g_{it_T})'$.

From (7) we obtain the matrix of the polygonal coefficients of the i-th trajectory:

$$\mathbf{G}_i = \mathbf{C}^{-1}\mathbf{X}_i',$$

and then, by considering the polygonal representation of the trajectories, we can define the following dissimilarity:

$$ {}_P d_{il}^2 = \sum_{s=1}^{T}\sum_{j=1}^{J}(g_{ijt_s} - g_{ljt_s})^2 = \sum_{s=1}^{T}\left\|\mathbf{g}_{it_s} - \mathbf{g}_{lt_s}\right\|^2 = \left\|\mathbf{G}_i - \mathbf{G}_l\right\|^2. \qquad (8) $$

In conclusion, we can note that the 'polygonal' dissimilarity presents characteristics very similar to longitudinal-acceleration dissimilarity, therefore capturing an analogous informative contribution of the trajectories.

Furthermore, we observe that the polygonal dissimilarity performs very well with not very long time trajectories.

5.3. *'Areal' dissimilarity*

Another suggestion consists in comparing the trajectories by taking into account the *area* between each pair of univariate trajectories. In the multivariate case, the proposed dissimilarity is equal to the sum of the areas between the univariate components of the trajectories, represented on each of J co-ordinate bi-dimensional sub-spaces of the data-space $\Re^{J+1}$, where the axes are referred to time and each of the J variables.

By considering two univariate time trajectories i and l (for the j-th variable), an 'areal' dissimilarity measure can be defined taking into account two possible situations (see Figure 8).

In the first case, the area between the two trajectories in the interval $[t_s, t_{s+1}]$ is the area of a trapezium with bases $|b_{iljt_s}|$ and $|b_{iljt_{s+1}}|$, where $b_{iljt_s} = (x_{ijt_s} - x_{ljt_s})$ and $b_{iljt_{s+1}} = (x_{ijt_{s+1}} - x_{ljt_{s+1}})$, and height $(t_{s+1} - t_s)$. Then, the area is:

$$ \frac{\left(\left|b_{iljt_s}\right| + \left|b_{iljt_{s+1}}\right|\right)(t_{s+1} - t_s)}{2}. $$

In the second case, the area between the two trajectories in $[t_s, t_{s+1}]$ is the sum of the areas of two similar triangles, with bases $|b_{iljt_s}|$ and $|b_{iljt_{s+1}}|$ and sum of heights $(t_{s+1} - t_s)$:

$$ \frac{(b_{iljt_s}^2 + b_{iljt_{s+1}}^2)(t_{s+1} - t_s)}{2\left(\left|b_{iljt_s}\right| + \left|b_{iljt_{s+1}}\right|\right)}. $$

In fact, since the two mentioned triangles with bases $|b_{iljt_s}|$ and $|b_{iljt_{s+1}}|$ and sum of

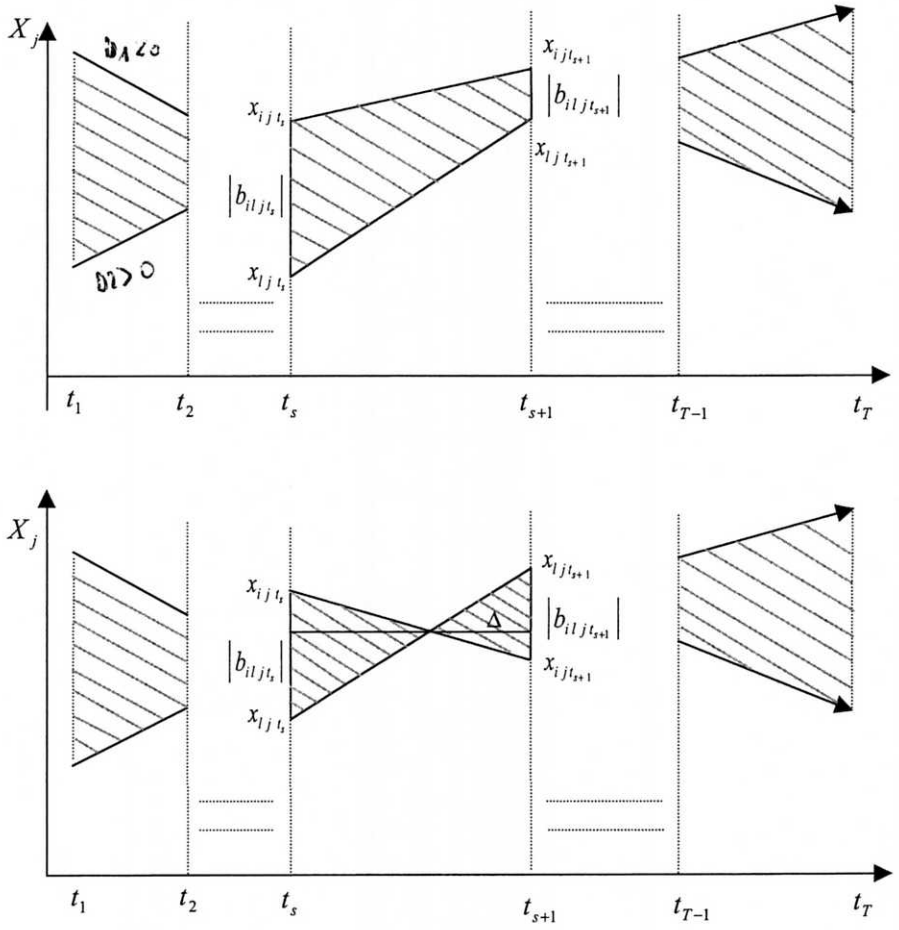


Fig. 8 – Possible configuration of a pair of time trajectories in the time interval $[t_s, t_{s+1}]$.

heights $(t_{s+1} - t_s)$ are similar, we have, by indicating with Δ the height of one of the two triangles (see Figure 8): $\dfrac{(t_{s+1} - t_s) - \Delta}{\Delta} = \dfrac{|b_{iljt_s}|}{|b_{iljt_{s+1}}|}$ and then $\Delta = \dfrac{(t_{s+1} - t_s)|b_{iljt_{s+1}}|}{|b_{iljt_s}| + |b_{iljt_{s+1}}|}$.

Therefore, the sum of the areas of the two similar triangles is:

$$\frac{\left|b_{il\,j\,t_s}\right|[(t_{s+1}-t_s)-\Delta]}{2}+\frac{\left|b_{il\,j\,t_{s+1}}\right|\Delta}{2}=\frac{1}{2}\left[\left|b_{il\,j\,t_s}\right|-\frac{\left|b_{il\,j\,t_s}\,b_{il\,j\,t_{s+1}}\right|}{\left|b_{il\,j\,t_s}\right|+\left|b_{il\,j\,t_{s+1}}\right|}+\frac{b^2_{il\,j\,t_{s+1}}}{\left|b_{il\,j\,t_s}\right|+\left|b_{il\,j\,t_{s+1}}\right|}\right].$$

$$(t_{s+1}-t_s)=\frac{(b^2_{il\,j\,t_s}+b^2_{il\,j\,t_{s+1}})(t_{s+1}-t_s)}{2\left(\left|b_{il\,j\,t_s}\right|+\left|b_{il\,j\,t_{s+1}}\right|\right)}.$$

Then, an expression of the area between two trajectories in $[t_s, t_{s+1}]$, that summarizes the two previous cases, is:

$$A_{il\,j\,t_s t_{s+1}}=\frac{(b^2_{il\,j\,t_s}+b^2_{il\,j\,t_{s+1}}+b_{il\,j\,t_s}b_{il\,j\,t_{s+1}}+\left|b_{il\,j\,t_s}\,b_{il\,j\,t_{s+1}}\right|)(t_{s+1}-t_s)}{2\left(\left|b_{il\,j\,t_s}\right|+\left|b_{il\,j\,t_{s+1}}\right|\right)}.$$

By taking into account all the sub-intervals $[t_s, t_{s+1}]$ (s = 1,T-1) of the analyzed period $[t_1, t_T]$, an 'areal' dissimilarity measure between multivariate time trajectories is:

$$_A d_{il}=\sum_{s=1}^{T-1}\sum_{j=1}^{J}A_{il\,j\,t_s t_{s+1}}=$$

$$=\sum_{s=1}^{T-1}\sum_{j=1}^{J}\frac{(b^2_{il\,j\,t_s}+b^2_{il\,j\,t_{s+1}}+b_{il\,j\,t_s}b_{il\,j\,t_{s+1}}+\left|b_{il\,j\,t_s}\,b_{il\,j\,t_{s+1}}\right|)(t_{s+1}-t_s)}{2\left(\left|b_{il\,j\,t_s}\right|+\left|b_{il\,j\,t_{s+1}}\right|\right)} \qquad (9)$$

This dissimilarity presents analogous characteristics to the cross sectional (instantaneous) dissimilarity. In fact, this dissimilarity is suitable for trajectories with few time points and can supply appreciable performances with smoothly varying trajectories. Moreover, like the cross sectional dissimilarity, it does not take into account the evolution (geometrically, the shape) of the trajectories (see Figure 9).

In particular, in Figure 9, we show two pairs of trajectories. In the first case the trajectories present the same shape, in the second one they do not. However, the portion of area between each pair of trajectories and then the areal dissimilarities are the same.

Finally, we observe that other kinds of 'areal' dissimilarities can be obtained by interpolating the univariate components of each multivariate trajectory with suitable *spline functions* (De Boor, 1978) or by considering the *Andrews' plot* (Andrews, 1972).

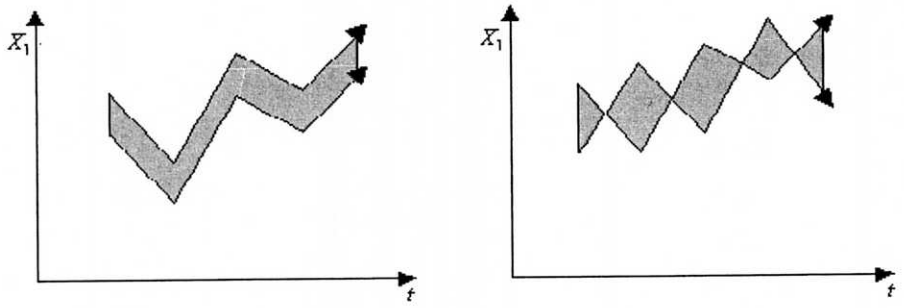


Fig. 9 – Example of (univariate) time trajectories with same shape (1st case) and different shape (2nd case) but same area.

6. The 'correlative' approach

The 'correlative' dissimilarity measures compare the trajectories by considering their time-dependence. We can distinguish two types of dissimilarities: the first type considers the dependence between different observations of different trajectories, according to the cross correlation coefficient of the univariate component of each trajectory, the second type compares each pair of trajectories by taking into account the autocorrelation coefficients which describe the dependence between different observations of the same trajectories.

A first dissimilarity belonging to the correlative approach is the so-called *lagged cross time dissimilarity* that compares each pair of trajectories, by considering the different time lags k, $k = 0, \pm 1, \pm 2, \ldots$. In particular, the squared Euclidean distance between the trajectories $\mathbf{X}_i$ and $\mathbf{X}_l$ at lag k is (Zani, 1983):

$$
{}_{1C}d^2_{il\,k} = \begin{cases} \displaystyle\sum_{s=1}^{T-k}\sum_{j=1}^{J}(x_{i\,j\,t_s} - x_{l\,jt_{s+k}})^2 = \sum_{s=1}^{T-k}\left\|\mathbf{x}_{it_s} - \mathbf{x}_{lt_{s+k}}\right\|^2, & k = 0,1,\ldots,K \\ \displaystyle\sum_{s=1-k}^{T}\sum_{j=1}^{J}(x_{i\,j\,t_s} - x_{l\,jt_{s+k}})^2 = \sum_{s=1-k}^{T}\left\|\mathbf{x}_{it_s} - \mathbf{x}_{lt_{s+k}}\right\|^2, & k = -1,\ldots,K. \end{cases} \tag{10}
$$

The (squared) distance (10) can be expressed easily in terms of the cross correlation.

Bothe *et al.* (1980) have proposed the following dissimilarities, which compare the autocorrelation and cross-correlation functions of each time trajectory:

$$ {}_{2C}d_{il} = \sum_{k=1}^{K}\sum_{j=1}^{J} w_k \frac{\left|\hat{r}_{ijk} - \hat{r}_{ljk}\right|}{1+\left|\hat{r}_{ijk} - \hat{r}_{ljk}\right|} \tag{11} $$

$$ {}_{3C}d_{il} = \sum_{k=1}^{K}\sum_{j=1}^{J} w_k \frac{\left|\hat{r}_{iljk} - \hat{r}_{lijk}\right|}{1+\left|\hat{r}_{iljk} - \hat{r}_{lijk}\right|} \tag{12} $$

where $\hat{r}_{ijk}$ is the global autocorrelation at *lag k* of the *i*-th time trajectory for the *j*-th variable, $\hat{r}_{iljk}$ is the cross correlation at *lag k* between the *i*-th and *l*-th time trajectories for the *j*-th variable, K is the maximum lag considered and w_k are subjective weights.

Other dissimilarities have been suggested by Košmelj (1986a), Košmelj and Lachet (1983) and Piccolo (1987). For instance, the latter author has proposed:

$$ {}_{4C}d_{il} = \left[\sum_{k=1}^{K}\sum_{j=1}^{J}(\hat{r}_{ijk} - \hat{r}_{ljk})^2\right]^{\frac{1}{2}}. \tag{13} $$

We observe that the correlative dissimilarities perform well with time trajectories with many time points.

So far, we have compared the trajectories by taking into account their global autocorrelation functions. However, as we know, the dynamic information of a time trajectory expressed by the global autocorrelation function is statistically equivalent to the information provided by the *spectrum*. Then, dissimilarity measures for time trajectories can be built comparing the respective spectra.

7. The 'structural' approach

In the 'structural' approach, we suppose that each time trajectory is decomposed in the trend, cycle, seasonal and accidental components. In this way, we compare the trajectories by considering these fundamental components. In particular, we define two kinds of structural dissimilarities: the first one based on the *trend* (and *multi-trend*); the latter one based on comparing the aspects linked to the *seasonality* of the analyzed trajectories.

1) The '*trend*' *dissimilarity*

In order to study the mutual relationships of time trajectories based on their average evolution in the long period, we can compare each pair of trajectories by considering the trend component.

For the j-th variable and i-th time trajectory, we define the following g–order polynomial trend:

$$\mathbf{x}_{ij} = \mathbf{T}\beta_{ij} + \mathbf{e}_{ij}$$

where $\mathbf{x}_{ij}$ is the $(T \times 1)$-vector of the observations for the i-th unit and the j-th variable; β_{ij} is the $((g+1) \times 1)$-vector of the regression parameters of the trend model; $\mathbf{e}_{ij}$ is the $(T \times 1)$-vector of residuals; $\mathbf{T}$ is the $(T \times (g+1))$-matrix, defined as follows:

$$\mathbf{T} = \begin{pmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^g \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & t_T & t_T^2 & \cdots & t_T^g \end{pmatrix}.$$

By assuming that, for each variable, the I trajectories have a trend of the same order, we propose the following dissimilarity:

$$_T d_{il}^2 = \sum_{j=1}^{J} \left\| \hat{\beta}_{ij} - \hat{\beta}_{lj} \right\|^2. \qquad (14)$$

where $\hat{\beta}_{ij} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{x}_{ij}$ is the least-squares estimate (analogously $\hat{\beta}_{lj}$).

2) The '*multi-trend*' *dissimilarity*

In several fields (Johnston, 1999) each time trajectory can be characterized, in different sub-periods of the considered time interval, by different types of linear trends. In these cases, for each trend, we consider spline functions (De Boor, 1978) and then, in order to compare the trajectories, we propose a dissimilarity that takes into account the coefficients of the trend spline, in each sub-period. In particular, by subdividing, for example, the period $[t_1,t_T]$ in three 3 sub-periods: $[t_1, t_{s^*}]$, $(t_{s^*}, t_{s^{**}}]$ and $(t_{s^{**}}, t_T]$, for the j-th variable and i-th time trajectory, we have the following linear trends:

$$x_{ij} = {}_1\beta_{0ij} + {}_1\beta_{1ij}\, t + u_{ij} \qquad t_1 \quad t \quad t_{s^*} \qquad 1^{st} \text{ period} \qquad (15a)$$
$$x_{ij} = {}_2\beta_{0ij} + {}_2\beta_{1ij}\, t + u_{ij} \qquad t_{s^*} \quad t \quad t_{s^{**}} \qquad 2^{nd} \text{ period} \qquad (15b)$$
$$x_{ij} = {}_3\beta_{0ij} + {}_3\beta_{1ij}\, t + u_{ij} \qquad t_{s^{**}} \quad t \quad t_T \qquad 3^{th} \text{ period.} \qquad (15c)$$

Then, we consider a linear spline function and estimate this function, by defining the following variables:

$$z_1 = t, z_2 = \begin{cases} 0 & 1 \le t \le t_{s^*} \\ t - t_{s^*} & t_{s^*} < t \end{cases} \text{ and } z_3 = \begin{cases} 0 & t \le t_{s^{**}} \\ t - t_{s^{**}} & t_{s^{**}} < t \le T \end{cases}$$

and re-parameterising the function as follows:

$$x_{ij} = \delta_{0ij} + \delta_{1ij} z_1 + \delta_{2ij} z_2 + \delta_{3ij} z_3 + u_{ij}. \tag{16}$$

By comparing equations (15a), (15b) and (15c) with (16), we have

$$\begin{array}{ll} {}_1\beta_{1ij} = \delta_{1ij} & {}_1\beta_{0ij} = \delta_{0ij} \\ {}_2\beta_{1ij} = \delta_{1ij} + \delta_{2ij} & {}_2\beta_{0ij} = {}_1\beta_{0ij} - \delta_{2ij} t_{s^*} \\ {}_3\beta_{1ij} = \delta_{1ij} + \delta_{2ij} + \delta_{3ij} & {}_3\beta_{0ij} = {}_2\beta_{0ij} - \delta_{3ij} t_{s^{**}}. \end{array} \tag{17}$$

By estimating, with ordinary least-squares, equation (16), we have continuous functions in nodes t_{s^*} and $t_{s^{**}}$ and substituting the estimated parameters $\hat{\delta}_{rij}$, $r = 1,2,3$, in (17), we obtain the estimates ${}_p\hat{\beta}_{qij}$, $p = 1,2,3$; $q = 1,2$ (Johnston, 1999). Alternatively, we can consider the constrained least-squares estimation (Johnston, 1999).

Then, a 'multi-trend' dissimilarity (in particular 'three-linear trend' dissimilarity) between trajectories is:

$${}_{MT}d_{il}^2 = \sum_{j=1}^{J} \left\| \hat{\delta}_{ij} - \hat{\delta}_{lj} \right\|^2 \tag{18}$$

where $\hat{\delta}_{ij} = (\hat{\delta}_{0ij}, \hat{\delta}_{1ij}, \hat{\delta}_{2ij}, \hat{\delta}_{3ij})'$ and $\hat{\delta}_{lj} = (\hat{\delta}_{0lj}, \hat{\delta}_{1lj}, \hat{\delta}_{2lj}, \hat{\delta}_{3lj})'$.

We can extend the previous results for τ sub-periods of the time interval and then for τ trends.

Notice that, it is preferable to utilize these dissimilarities for long time trajectories.

Finally, we observe that it is possible to consider also *quadratic splines, cubic splines*, and so on, (Buse and Lim, 1977; De Boor, 1978) and define, analogously to the linear case, dissimilarity measures between trajectories, by comparing their respective coefficients.

3) The *'seasonal' dissimilarity*

The seasonal component captures the systematic oscillations of each trajectory with period equal to a year. For instance, Hylleberg (1994) suggests the following definition: «Seasonality is the systematic, although not necessarily regular, intrayear movement caused by the change of the weather, the calendar, and

timing of decisions, directly or indirectly through the production and consumption decisions made by agents of economy. These decisions are influenced by endowments, the expectations and preferences of the agents, and the production techniques available in the economy».

A dissimilarity which compares univariate time trajectories, by taking into account the seasonal component has been proposed, for the first time, by Raveh and Tapiero (1980) and successively, by Raveh (1981). A dissimilarity which compares each pair of multivariate time trajectories (i-th and l-th) and considers the respective *seasonal indices* (or *seasonal patterns*) $\hat{S}_{ij\omega}$ and $\hat{S}_{lj\omega}$ (for the ω-th month) is (Raveh, 1981):

$$ {}_{s}d_{il} = \frac{1}{\Omega J} \sum_{\omega=1}^{\Omega} \sum_{j=1}^{J} \left| \hat{S}_{ij\omega} - \hat{S}_{lj\omega} \right| \tag{19} $$

where, for example, $\omega = 12$ for annual time trajectories.

In order to compute this dissimilarity measure it is necessary to consider trajectories with many time points.

8. An empirical comparison

In this Section, we illustrate an empirical comparison of the dissimilarity measures between multivariate time trajectories listed in Table 1.

The time array has been drawn from the «The Penn World Tables» (Mark 5.6), denoted PWT 5.6 (Heston and Summers, 1991). This time array can be drawn

Table 1
Typologies of dissimilarity measures for multivariate time trajectories

D_1 = cross sectional dissimilarity (D'Urso and Vichi)
D_2 = longitudinal (velocity) dissimilarity (D'Urso and Vichi)
D_3 = longitudinal (acceleration) dissimilarity (D'Urso and Vichi)
D_4 = compromise dissimilarity (D'Urso and Vichi)
D_5 = compromise dissimilarity (Saporta and Lavallard)
D_6 = compromise dissimilarity (Coppi and D'Urso)
D_7 = 'polygonal' dissimilarity
D_8 = 'areal' dissimilarity
D_9 = autocorrelative dissimilarity (Piccolo)
D_{10} = linear trend dissimilarity

also from the *web-site*: http://arcadia.chass.utoronto.ca/pwt/index.html). In particular, we have considered 15 European Countries (Austria, Belgium, Denmark, Finland, France, West Germany, Ireland, Italy, Luxemburg, Norway, Portugal, Spain, Sweden, The Netherlands, UK) and 3 macro-economic variables (GNP, Investment/GNP, Standard Living Index ((Consumptions + Government consumptions - Military expenses)/GNP)) observed in the Years 1974, 1979, 1984, 1989. We have calculated the different dissimilarities among the multivariate macroeconomic time trajectories of the 15 European Countries, obtaining ten (15×15) dissimilarity matrices indicated by D_1 - D_{10} (see Table 1).

Notice that, in order to compute the different dissimilarity measures, we have implemented an algorithm SAS/IML.

At this point, we can analyze the behaviour of the different types of dissimilarities, by comparing the corresponding configurations obtained by each dissimilarity matrix. In fact, a comparison of couples of such configurations is a comparison of the performance of the corresponding dissimilarity measures on the considered dataset.

Then, in order to compare the dissimilarity, we consider the following *strategy of analysis*:

1) Each dissimilarity matrix is analyzed via a *metric scaling* (see, for instance, Cox and Cox, 1994) to provide 10 configurations of 15 points, representing the European Countries.

2) Every possible pair of these 10 configurations is compared using *Procrustes analysis* (see, for instance, Cox and Cox, 1994). The (squared) Procrustes matrix is shown in Table 2.

3) Since the squared Procrustean values are squared distances[1] between corresponding configurations and hence, by implication, between the dissimilarities that generated them, we can apply the metric scaling to (squared) Procrustean matrix (Tab. 2). In this way, we obtain a configuration of 10 points in the bidimensional space (see Figure 10), which represent the 10 dissimilarity measures D_1-D_{10}. Then, the distance between any two such points represents the dissimilarity (Procrustean measure) between the configurations generated by the corresponding dissimilarities calculated among the multivariate time trajectories of the 15 European Countries. Furthermore, we applied to the (squared) Procrustean matrix a *hierarchical clustering method* (single linkage method), obtaining the same outcomes (see Dendrogram shown in Figure 11).

Notice that the metric scaling and the Procrustean analysis have been applied by utilizing, respectively, the software CLSCAL and PROCUST (see Cox and Cox, 1994).

1. Sibson (1978) has demonstrated that the Procrustean measure is a metric. Then, also the squared Procrustean matrix can itself be analyzed via multidimensional scaling.

Table 2
(Squared) Procrustean matrix

	D_1	D_2	D_3	D_4	D_5	D_6	D_7	D_8	D_9	D_{10}
D_1										
D_2	0.247									
D_3	0.478	0.249								
D_4	0.190	0.077	0.329							
D_5	0.102	0.146	0.380	0.088						
D_6	0.320	0.107	0.172	0.185	0.213					
D_7	0.431	0.194	0.097	0.264	0.335	0.167				
D_8	0.153	0.149	0.392	0.086	0.074	0.247	0.325			
D_9	0.512	0.346	0.431	0.334	0.416	0.422	0.339	0.352		
D_{10}	0.390	0.596	0.844	0.520	0.464	0.682	0.779	0.454	0.705	

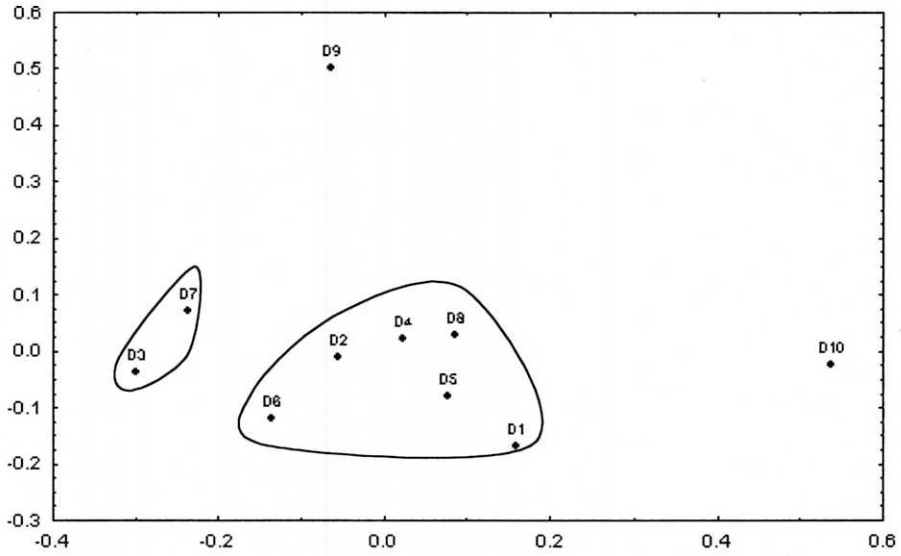


Fig. 10 – Final configuration.

Summing up, the comparative assessment based on the considered dataset provides the following results:

1) The dissimilarity measures D_9 and D_{10} are very distant from each other and with respect to other dissimilarities.

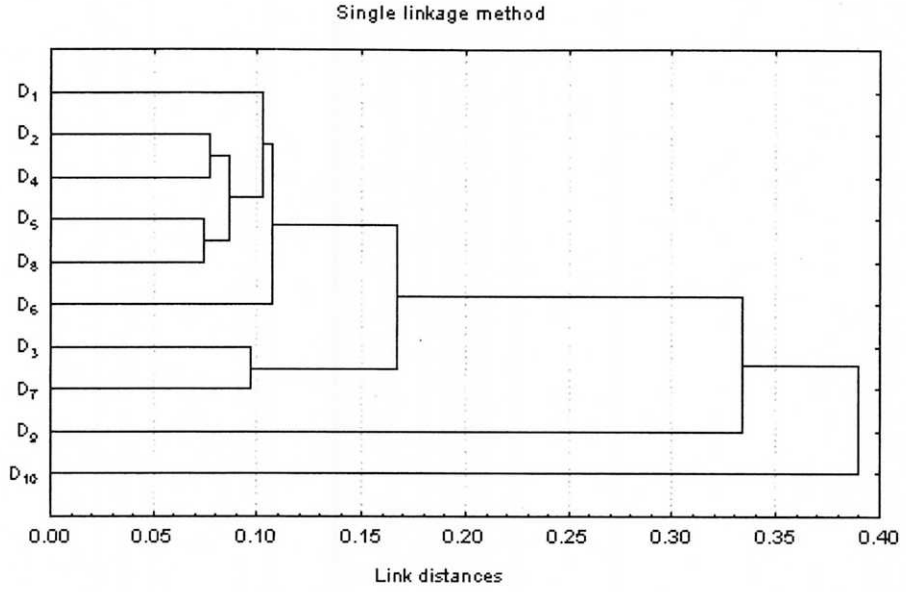


Fig. 11 – Dendrogram.

2) The vector of the weights of the mixed dissimilarity D_4 is $\alpha = (0.56; 0.63; 0.54)'$. Consequently, D_4 is more influenced, orderly, by the evolutive-velocity dissimilarity (D_2), cross sectional dissimilarity (D_1) and evolutive-acceleration dissimilarity (D_3). This conclusion is confirmed also by the final configuration (Fig. 10) and the dendrogram (Fig. 11).

3) In order to compute the compromise dissimilarity proposed by Saporta and Lavallard (D_5), the optimal *configuration* is $\{0,0,0,1\}$, *i.e.*, D_5 has been calculated by considering the cross sectional dissimilarity (D_1) in the three years 1974, 1979, 1984 and the longitudinal-velocity dissimilarity (D_2) in the time interval 1984-1989.

4) Concerning the Coppi-D'Urso's mixed dissimilarity, the instantaneous and longitudinal-acceleration components are relevant.

5) We notice that the closest point to D_6 is D_2, though, in the specific case, the velocity component is not relevant in the computation of D_6. Probably, the main justification of this situation is the attractive effect of D_1 on D_6.

6) The polygonal dissimilarity, indicated by the point D_7 in the final configuration, is next to D_3. From a theoretical point of view, this situation is explainable by taking into account the link between polygonal coefficients and acceleration (see Section 5.2).

7) From Figures 10 and 11, we can notice the proximity of the point D_8 with respect to D_4, D_5 and D_1. This indicates somehow the link between the 'areal' aspect and the instantaneous characteristics of each pair of time trajectories.

9. Final remarks

In this paper we have analyzed the algebraic and geometric characteristics of a three-way time array. In particular we have focused the attention on a specific geometric representation of the time array in the space of the units: the multivariate time trajectories. Then, we have classified different dissimilarity measures between multivariate time trajectories (some already present in literature, some others proposed here). A group of measures is based on a comparison of the trajectories from a geometrical point of view (cross sectional-longitudinal, polygonal and 'areal' dissimilarities). Another group of dissimilarities compares the trajectories by considering their correlative and structural features. Therefore, each dissimilarity measure takes into account a particular informational aspect of the trajectories. Summing up the characteristics of the different dissimilarities are shown in Table 3. Notice that, for all the examined dissimilarities, we have considered an exploratory approach. Moreover, for a critical appraisal, an empirical comparison has been carried out, among the various measures considered in the paper.

A suggestive perspective of research consists in considering other kinds of time trajectories, for instance asynchronous and imprecise (vague or fuzzy) trajectories (Coppi and D'Urso, 2000). Also the possibility to compare the trajectories by determining, for each trajectory, a sort of 'sufficient statistic' can be taken into account. Finally, we underline the interesting opportunity to adopt a probabilistic-inferential approach (autoregressive, spectral, etc.) in the definition of dissimilarities for classifying time trajectories as in Maharaj (1996; 1999; 2000); Martin (2000); Piccolo (1990). However, in these cases, we can observe that the application of the dissimilarities in the clustering framework is restricted to reasonably long stationary time trajectories (time series) that can be modeled as linear processes. Then «because in practice, we may encounter short or stationary series that cannot be fitted by linear models or nonstationary series that cannot be easily transformed to stationary series, further research is being undertaken to extend [the contributions suggested in this framework] so that [they] can be more generally applied» (Maharaj, 2000).

Acknowledgements

The author is grateful to the Editor and the Referees for their suggestions and comments.

Table 3
Features of the dissimilarities

Dissimilarity measure	*Approach*	*Comments*	*Preferable features of the trajectory*
Cross sectional (positional) (cf. (2))	Geometric	It compares each pair of trajectories by considering their *instantaneous position.*	Smoothly and fairly varying trajectories with few time points or not very long.
Longitudinal-velocity (cf. (3)) Longitudinal-acceleration (cf. (4))	Geometric	They consider, respectively, the *velocity* (*slope*) and *acceleration* (*concavity* and *convexity*) of the trajectories.	Wildly and smoothly varying trajectories with few time points or not very long.
Mixed (cf. (5), (6))	Geometric	They represent a *compromise* (suitable linear combination) of the positional, velocity and acceleration dissimilarities, where the weights of each component are objectively determined by means of optimization criteria (max variance and min inertia three-way distance).	Like cross sectional and longitudinal cases.
'Polygonal' (cf. (8))	Geometric	It takes into account the *shape* (*polygonal line*) of each trajectory by computing the so-called *polygonal coefficient.*	Trajectories not very long.
'Areal'(cf. (9))	Geometric	It compares the trajectories by measuring the area between each pair of trajectories. This dissimilarity is similar to the cross sectional (positional) one.	Smooth trajectories with few time points.
'Cross-correlative' (cf. (10)) 'Autocorrelative' (cf. (11)-(13))	Correlative	These dissimilarities compare each pair of trajectories, by considering, respectively, the dependence between different observations of the *same trajectories* and the dependence between different observations of *different trajectories*.	Trajectories with many time points.
'Trend' (cf. (14)) 'Multi-trend' (cf. (15))	Structural	They consider, respectively, the *polynomial trend* and the *linear spline* of each trajectory.	Trajectories with many time points and clear and evident tendency.
'Seasonal' (cf. (16))	Structural	It analyses the *seasonal coefficients* which summarize the seasonal behaviour of each trajectory.	Trajectories with many time points and seasonality.

REFERENCES

ANDREWS, D. F. (1972), Plots of High-Dimensional Data. *Biometrics*, 28, 125-136.

BENEDETTI, C. (1972), La rappresentazione delle poligonali e l'inversione dei momenti assoluti di ordine dispari rispetto ad un centro arbitrario. *Metron*, XXX, 1-4, 1-22.

BOHTE, Z., CEPAR, D. and KOŠMELJ, K. (1980), Clustering of time series. In *COMPSTAT'80*, 587-593. Heidelberg: Physica-Verlag.

BOX, G. E. P. and JENKINS, G.M. (1976), *Time Series Analysis. Forecasting and Control*. San Francisco: Holden Day.

BUSE, A. and LIM, L. (1977), Cubic Splines as a Special Case of Restricted Least Squares. *Journal of the American Statistical Association*, 72, 64-72.

CARLIER, A. (1986), Factor analysis of evolution and cluster methods on trajectories. In *COMPSTAT'86*, 140-145. Heidelberg Wien: Physica-Verlag.

CARROLL, J. D. and ARABIE, P. (1980), Multidimensional scaling. *Annual Review of Psychology*, 31, 438-457.

COPPI, R. and BOLASCO, S. (eds.) (1989), *Multiway Data Analysis*, Amsterdam: Elsevier Science Publisher, North Holland.

COPPI, R. and D'URSO, P. (2000), Fuzzy time arrays and dissimilarity measures for fuzzy time trajectories. In *Data Analysis, Classification, and Related Methods* (eds. H. A. L Kiers, J. P. Rasson, P. J. F. Groenen, M. Schader), 273-278. Berlin: Springer-Verlag.

COPPI, R. and D'URSO, P. (2001), The geometric approach to the comparison of multivariate time trajectories. In *Advances in Data Science and Classification* (eds. S. Borra, R. Rocci, M. Vichi, M. Schader), 93-100. Heidelberg: Springer-Verlag.

COPPI, R. and ZANNELLA, F. (1978), L'analisi fattoriale di una serie temporale multipla relativa allo stesso insieme di unità statistiche. In *Atti della XXIX Riunione Scientifica della SIS*. Bologna, 61-79.

COSTANZO, G. D. and D'URSO, P. (2000), A distance measure between structural models. In *Proceedings in Computational Statistics, COMPSTAT 2000* (eds. W. Jansen, J. G. Bethlehem), Utrecht 21-25 August 2000, 25-26. Netherlands, Statistics.

COX, T. F. and COX, M. A. A. (1994), *Multidimensional Scaling*. London: Chapman & Hall.

D'URSO, P. and VICHI, M. (1998), Dissimilarities between trajectories of a three-way longitudinal data set, in *Advances in Data Science and Classification* (eds. A. Rizzi, M. Vichi, H. H. Bock), 585-592. Springer-Verlag.

DE BOOR, C. (1978), *A Practical Guide to Splines*. New York: Springer-Verlag.

HYLLEBERG, S. (1994), *Modelling Seasonality*, *Advanced Texts in Econometrics*, 4, Oxford University Press.

JOLY, S. and LE CALVÉ, G. (1995), Three-way distances. *Journal of Classification*, 12, 191-205.

JOHNSTON, J. (1999), *Econometrica*, Milan: F. Angeli.

KAKIZAWA, Y, SHUMWAY, H. and TANIGUCHI, M. (1998), Discriminant and clustering for multivariate time series. *Journal of the American Statistical Association*, 93, 441, 328-340.

KIERS, H. A. L. (2000), Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics*, 14, 105-122.

KOŠMELJ, K. (1986), A two-step procedure for clustering time varying data. *Journal of Mathematical Sociology*, 12 (3), 315-326.

KOŠMELJ, K. and BATAGELJ, V. (1990), Cross-sectional approach for clustering time varying data. *Journal of Classification*, 7, 99-109.

KOŠMELJ, K. and LACHET, B. (1983), Aspect temporel des variables hydriques du haut rhône français. *Revue de Statistique Appliquée*, XXXI (2), 5-18.

MAHARAJ E. A. (1996), A significance test for classifying ARMA models. *Journal of Statist. Comput. Simul.*, 54, 305-331.

MAHARAJ E. A. (1999), Comparison and classification of stationary multivariate time series. *Pattern Recognition*, 32, 1129-1138.

MAHARAJ E. A. (2000), Cluster of time series. *Journal of Classification*, 17, 297-314.

MACGREGOR, J. F., JAECKLE, C., KIPARISSIDES C. and KOUNTOUDI, M. (1994), Process monitoring and diagnosis by multiblock PLS methods. *AIChE Journal*, 40, 826-838.

MARTIN, R. J. (2000), A metric for ARMA process. *IEEE Transactions on Signal Processing*, 48 (4), 1164-1170.

NOMIKOS, P. and MACGREGOR, J. F. (1995), Multivariate SPC charts for monitoring batch process. *Technometrics*, 37, 41-59.

HARSHMAN, R. A. and LUNDY, M. E. (1984), Data preprocessing and the extended PARAFAC model. In *Research Methods for Multimode Data Analysis* (eds. H. G. Law, C. W. Snyder, J. A. Hattie, R. P. McDonald), 216-284. New York: Praeger.

HESTON, A. and SUMMERS, R. (1991), The Penn World Table (Mark 5): an expanded set of international comparisons, 1950-1988. *Quarterly Journal of Economics*, 327-368.

PICCOLO, D. (1990), A distance measure for classifying ARIMA models. *Journal of Time Series Analysis*, 11 (2), 153-164.

RAVEH, A. (1981), A graphic display for characterization of seasonal pattern similarities of time series. *The Statistician*, 30 (3), 179-192.

RAVEH A. and TAPIERO, C. S. (1980), Finding common seasonal patterns among time series. *Journal of Econometrics*, 12, 353-363.

SAPORTA, G. and LAVALLARD, F. (1996), *L'Analyse des Données Évolutives. Méthodes et Applications*. Paris: Éditions Technip.

SIBSON, R. (1978), Studies in the robustness of multidimensional scaling: Procrustes statistics. *Journal of the Royal Statistical Society*, B, 40, 234-238.

ZANI, S. (1983), Osservazioni sulle serie storiche multiple e l'analisi dei gruppi. In *Analisi Moderna delle Serie Storiche* (ed. D. Piccolo), 263-274. F. Angeli.