Noname manuscript No.
(will be inserted by the editor)

Clustering of cross-sectional time series data

Agostino Tarsitano · Marianna Faclone

Received: date / Accepted: date

Abstract This paper deals with multivariate time-series describing each of a fixed set of units across a short sequence of time periods of unequal lengths. In particular, we present a method for clustering unbalanced time series cross-sectional (TSCS) data on the basis of a metric defined combining partial metrics computed on a selected set of social and economic indicators. We present results of some numerical experiments to study the behavior of the proposed method; in addition, a case study on quoted companies has been undertaken and the findings show that the new technique has acceptable results.

Keywords Consensus of classifications · short time series · PAM clustering · Panel data

1 Introduction

There is a continuous need for expansion of statistical methods to study data that combine observations both crossectionally and over time because they can exploit information more efficiently than cross-section alone or time series analysis alone. This is particularly true in the context of socio-economic studies. Here, the time series are usually characterized by a large sampling time, which cannot be shortened because the process of gathering data and delivering them to users takes long time and/or is expensive. Due to the limited number of observations, TSCS data are often pooled to increase data and to

A. Tarsitano. Corresponding author.
Dipartimento di Economia e Statistica - Università della Calabria
Via Pietro Bucci, cubo 1C, 87036 Rende (CS) - Italy
Tel.: +39-0984-492465
Fax: +123-45-492421
E-mail: agotar@unical.it

M. Falcone
E-mail: maryfalcon@libero.it

control for temporal and spatial bias in results based on a single unit or a single period. Such an approach, however, is not entirely satisfactory because it is unlikely in many cases that the units are statistically unrelated over time and space. The clustering of multivariate time series into homogeneous groups of cross-sectional units can be seen as a means rather than an end in itself; in fact, the pooling of time series can be carried out only within each group, and this has great potential to solve at least a part of the problems of spatial and temporal correlations.

TSCS data are usually organized into a three-way array of data referring to a fixed set divided into m units, indexed by $r = 1, 2, \ldots, m$, in which each unit is described by a given set of p different time series concerning various aspects of a particular problem.

$$\mathbf{X}_j = \begin{bmatrix} x_{1,j,t_{1,j}} & x_{1,j,t_{1,j}+1} & x_{1,j,t_{1,j}+2} \cdots & x_{1,j,t_{1,j}+n_{1,j}-1} \\ x_{2,j,t_{2,j}} & x_{2,j,t_{2,j}+1} & x_{2,j,t_{2,j}+2} \cdots & x_{2,j,t_{2,j}+n_{2,j}-1} \\ \cdots & \cdots & \cdots \quad \cdots & \cdots \\ x_{r,j,t_{r,j}} & x_{r,j,t_{r,j}+1} & x_{r,j,t_{r,j}+2} \cdots & x_{r,j,t_{r,j}+n_{r,j}-1} \\ \cdots & \cdots & \cdots \quad \cdots & \cdots \\ x_{m,j,t_{m,j}} & x_{m,j,t_{m,j}+1} & x_{m,j,t_{m,j}+2} \cdots & x_{m,j,t_{m,j}+n_{m,j}-1} \end{bmatrix} \quad j = 1, 2, \ldots, p \quad (1)$$

Notice that the time intervals of the p variables do not need to be of equal length: $n_j > 2, j = 1, 2, \ldots, p$; also, the time spans over which the units are observed may be different for the two starts and/or the two ends. In practice, each slide of (1) is a ragged array, which can be thought of as a rectangular array, certain of whose positions have been omitted. This type of data structure is extremely useful in regional analysis.

Preliminarily, the values $x_{r,j,t_{r,j}}$ are standardized to have a mean of zero and a variance of one through time. This type of invariance is useful for discarding heterogeneity of scales and/or baseline shifts regarding of the average level of the process, *i.e.* one process and a shifted or scaled version of it will have zero distance. Of course, centering changes to a common values the plot origin of the time series and scaling changes their orders of magnitude, but not the rankings, therefore all neighbor information is preserved.

The main objective of our research is to determine how to effectively utilize the combined strength of cross-sectional and temporal dimension of a common set of fixed indicators to classify short-run dynamics observed for a given set of units into two or more groups. We believe that a method for exploiting the richness of the data set (1) should be based on a distance function calculated at the univariate level, that is, for each single univariate time series. Inherent in this latter approach, as a necessary premise, is the consideration that the multi-sequences encountered in our type of studies are fairly short. In fact, when dealing with many time series of short duration, one wants more assigning a value to the dissimilarity between two observed sequences, rather than quantifying the strength of relationships between the stochastic processes that generate observations. According to [4], this implies a preference for the use of a distance function having a domain in the space of all time series, over a

distance function embedded in a lower dimensional space formed through, for example, fitting a parameterized model to each series.

The remainder of the paper is structured as follows. The second section introduces a distance measure for univariate short time series using geometric differences between their temporal profile. Initially, we calculated the distance between two univariate time series. The same computation can be performed far all possible pairs of units and for all variables in the data set. Subsequently, these indicator-specific distance matrices are merged into a global distance matrix, which is used for the classification of the units. Section 3 is devoted to a detailed description of our proposed methods to calculate a distance between multiple time series based on a weighted average of partial metrics. Cluster analysis of TSCS data is shown in Section 4 in terms of a consensus of classificazions. Our method aims at characterizing the unit behavior over time in order to be able to group units with similar dynamics, which, in turn, provide a mean to researchers and policy makers for better learning the action/reaction pattern that a certain type of unit shows in response to changing public policy or interventions to improve their conditions. Finally, the last section provides a summary and possible lines of future research.

2 Distance between short time series

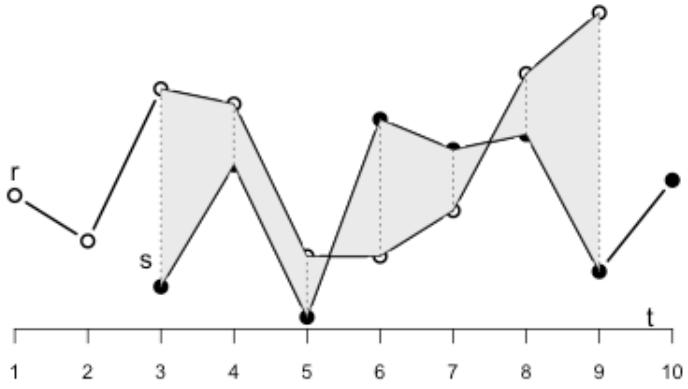
To compare and classify time series, various distance functions have been proposed in literature (see, for example, [7] and [15]), but all have shortcomings, especially when the observed sequences are short and/or have different lengths. In a recent review [18], discussed merits and drawbacks of several metrics and concluded that an "appropriate" distance measure for time series should be driven by the relative change of intensity as well as the temporal information because, only on this ground, distance-based clustering methods can be easily extended to time series.

To incorporate both factors, [19] defined a metric based on the observation that similar sequences have a small area enclosed between the polygonal curves representing them. Consider two univariate time series concerning the j-th variable observed in two different units $x_{r,j,t_{r,j}}, x_{s,j,t_{s,j}}$ of lengths, respectively, $n_{r,j}$ and $n_{s,j}$. The formula that we wish to evaluate has the form:

$$\delta\left(\mathbf{X}_{r,j}, \mathbf{X}_{s,j}\right) = \theta_{r,s,j}\left[\frac{\sum_{t=\alpha_{r,s,j}}^{\beta_{r,s,j}} \gamma_t A_t}{\sum_{t=\alpha_{r,s,j}}^{\beta_{r,s,j}} \gamma_t}\right],\ j = 1, 2, \ldots, p \tag{2}$$

The idea of the new metric is that the smaller the value of $\delta\left(\mathbf{X}_{r,j}, \mathbf{X}_{s,j}\right)$, the greater the ability of the two time series to reproduce each other. The symbol A_t in (2) denotes the area between the two polygons in the time interval $[t, t+1]$. More specifically, A_t is made of trapezoids or triangles. The latter are formed when polygons intersect and the intersection point belongs to $[t, t+1]$. In this case, the local contribution to the distance is obtained by summing the area of the two triangles. The graph in Figure 1 illustrates the idea.

Fig. 1: Distance between two time series.



The simple average of the A_t is commutative so that it ignores the temporal characteristics of the data being analyzed. In an attempt to circumvent this limitation, [19] applied a system of time-varying weights $\gamma_t, t = \alpha_{r,s,j}, \ldots, \beta_{r,s,j}$ to the polygons shown in Figure 1.

$$\gamma_t = 0.5 \left(\min\left\{|x_{r,j,t_{1,j}} - x_{s,j,t_{1,j}}|, |x_{r,j,t_{1,j}+1} - x_{s,j,t_{1,j}+1}|\right\} + \max\left\{|x_{r,j,t_{1,j}} - x_{s,j,t_{1,j}}|, |x_{r,j,t_{1,j}+1} - x_{s,j,t_{1,j}+1}|\right\}\right). \quad (3)$$

The larger is γ_t, the more important is the role of the corresponding local distance within the global distance function, while irrelevant differences have virtually zero weight.

The coefficients $\theta_{r,s,j} > 0$ are given by

$$\theta_{r,s,j} = \left[1 + \frac{\nu_j - (\beta_{r,s,j} - \alpha_{r,s,j})}{\nu_j}\right] \quad (4)$$

where

$$\alpha_{r,s,j} = \max\{t_{r,j}, t_{s,j}\}; \quad \beta_{r,s,j} = \min\{t_{r,j} + n_{r,j} - 1, t_{s,j} + n_{s,j} - 1\}; \quad \nu_j = \left(\max_{1 \le r \le m}\{n_{r,j}\} - 1\right) \quad (5)$$

The series r starts at t_1 and ends at t_9 whereas the series s starts at t_3 and ends at t_{10}; the two series overlap from t_3 through t_9 so that we have at least *seven* continuous observations during the period $t_1 - t_9$. Supposing that the largest time span for the j-th indicator is $\nu_j = 10$, the size-correction factor is $\theta_{r,s,j} = 13/9$ which implies that the area factor in the square brackets of (2) is increased by 44.4% with respect to the case in which both the sequences have a length equal to ν_j. The factor $\theta_{r,s,j}$ in expression (2) takes into account, within

certain limits, the possibility that units in the same panel may have a different time span for the same indicator. Although there is no specific reason to assume that the unobserved divergence between the units r and s is of a linear type, the choice (4) has the desirable effect of restraining formation of groups that share too few features. In fact, the rationale behind the introduction of this factor, not present in [19], is the generally accepted fact that the ability of any distance measure to distinguish among objects, decreases with the reduction of meaningful comparisons.

Usually, distances have the common thread of being scaled to vary in the unit interval:

$$0 \leq \delta_{r,s}^{(j)} \leq 1, \;\; j = 1, 2, \ldots, p \tag{6}$$

where zero is achieved only when the two time series are identical (once standardized) and one when the two time series maximally differ in all time points. Condition (6) is necessary, otherwise a combined representation of the distances would copy the structure of the indicator with the largest distances. In order to obtain the scaling relation in (6) we used

$$d_{r,s}^{(j)} = \frac{\delta_{r,s}^{(j)}}{\beta_j}, \quad \text{with} \quad \beta_j = \max_{r,s=1,2,\cdots m} \delta_{r,s}^{(j)}; \quad j = 1, 2, \cdots, p. \tag{7}$$

[14] shows that, since the lines forming the polygons in figure fig:Graf3 have their left and right end points on a vertical line, the area between them is a metric and this, in turn, implies that (2) satisfies the four conditions of non-negativity, reflexivity, symmetry, and triangle inequality.

A possible drawback of the metric (2) is its inability to discriminate between similar patterns starting at different points. This is not necessarily a problem. For instance, two time series with the same shape that do not occur concurrently on the time axis will have a high distance and, presumably, will be classified in two different groups. This is a correct decision if the initial conditions of the process are characterized by one or more exogenous variables.

The quantity $d\left(\mathbf{X}_{r,j}, \mathbf{X}_{s,j}\right)$ can be computed for all the pairs of the m units. These distances, once collected, form the $(m \times m)$ distance matrix $\mathbf{D}_j$ with elements $d_{r,s}^{(j)}$, relative to the j-th variable so that a comparison of two units becomes a comparison of two matrices. An application of the metric (2) that invokes each of the p variables provides a separate and potentially different basis for a multivariate analysis of the units. As a consequence, the comparison of two units can be considered a comparison of two sets of matrices. Of course, multivariate sequences have complex internal correlations that are neglected by the partial distance matrices, but the reduced amount of data in (1) usually cannot provide a reliable determination of all the parameters involved in a proper model of a real situation producing TSCS data.

3 Combining distances

The performance of a clustering algorithm critically depends on the measure used to form the distance matrix. This is also true for a classification of multi-

variate short time series. A promising approach consists in processing all units together and performing a single classification procedure based on a weighted average of all the partial distance matrices.

$$\mathbf{D} = \sum_{j=1}^{p} w_j \mathbf{D}_j \quad \text{with} \quad w_j \geq 0, \quad \text{and} \quad \sum_{j=1}^{m} w_j = 1. \tag{8}$$

Since the different variables could potentially generate a different solution, this is necessarily a complex account that emphasizes a great deal more relationship than any individual distance matrix. Although (8) is an attractive, easy, and reasonable method, caution should be taken as a linear combination might obscure profound structural changes which could have occurred in the pattern of some indicators. Units may rise and fall at different time points; some indicators follow largely predictable trajectories, but it is common to observe dynamics containing sine waves with varying amplitudes; sometimes trends are evident, sometimes are hidden by noise and complicated dependencies between space and time factors. Such differences might lead to an unequal or unfair contribution of indicators in the multivariate data and might thus yield suboptimal results by implicitly favoring a group of indicators over another.

To use the combined metric (8), a user must supply proper weights for the p variables. Some examples are listed below.

(a) **Subjective weighting**. The weights $w_j, j = 1, 2, \ldots, p$ could be determined on the basis of an *a priori* judgment of what is important and what should be prioritized with regard to the partial distance matrices. In other words, investigators give weights to types based on an intuitive understanding of the data, but if they do not know the context well, the assessment may be inadequate and introduce bias. For the moment we ignore this weighting scheme, although it should be considered seriously in future research.

(b) **Equal weighting**. If all indicators are equally effective when determining the global distance matrix, then the weights must be

$$w_j = 1/p, \ j = 1, 2, \ldots, p. \tag{9}$$

Equal weighting appears to be a valid practice when there are no theoretical or empirical grounds for choosing a different scheme.

(c) **Equalizing standard deviations of partial distances**. The significance of an indicator in determining the global distance depends, among other things, on the variability of $d_{r,s}^{(j)}, j = 1, 2, \ldots, p$ so that indicators leading to a high variance of pairwise distances will, thus, be more likely to have great influence into the global distance. To ensure the respect for this principle we can use the reciprocal of a measure of "variability" for the partial distances. For example

$$w_j = \sigma^{-1}\left[d_{r,s}^{(j)}\right] \left\{ \sum_{j=1}^{p} \sigma^{-1}\left[d_{r,s}^{(j)}\right] \right\}^{-1} \qquad j = 1, 2, \ldots, p \tag{10}$$

where $\sigma\left[d_{r,s}^{(t)}\right] > 0$ is the standard deviation of the distances in the strictly lower triangular part of $\mathbf{D}_j$. If the weights are based on (10) then the elements of the lower triangular part of the distance matrices are normalized to have a standard deviation of one.

(d) **Equalizing mean of partial distances**. The distances in the strictly lower triangular part of $\mathbf{D}_j$ can be normalized to have an average value of one in the data set.

$$w_j = \mu^{-1}\left[d_{r,s}^{(j)}\right]\left\{\sum_{j=1}^{p}\mu^{-1}\left[d_{r,s}^{(j)}\right]\right\}^{-1} \qquad j = 1, 2, \ldots, p \tag{11}$$

where $\mu\left[d_{r,s}^{(j)}\right] > 0$ is the average entry in the strictly lower triangular part of the j-th partial distance matrix.

(e) **Distatis**. An optimal system of weights, can be obtained using an expression for how much a certain variable affects the global distance. This can be derived from the total sum of squares of the elements of the partial distance matrices $\mathbf{D}_j, j = 1, 2, \ldots, p$ choosing the weights so as to maximize the variance of the elements in the global distance matrix $\mathbf{D}$ which, in turn, leads to the Distatis procedure developed by [1]. This method, requires that the $\mathbf{D}_j$'s are Euclidean distance matrices. The process of construction of the metric (2) gives no guarantee for the euclidity of the partial matrices, but in case $\mathbf{D}_j$ is not Euclidean, it is possible to determine a constant ϕ_1 such that a matrix with elements $d_{r,s}^{(j)} + \phi_1$ $r \neq s$ is Euclidean. See [10][theorem 7]. [16] show that if the $\mathbf{D}_j$'s are Euclidean, then $\mathbf{D}$ is also Euclidean. The weights $\mathbf{w} = (w_1, w_2, \ldots, w_p)$ are such that the patterns generated by variables more similar to the others get higher weights than those who disagree with most of the rest of the variables.

4 Clustering

Clustering of multiple time series of cross-sections of the type illustrated in (1), is an important issues that intersects many research fields. A usual task for such data is the discovery of units that move in a similar way. Also, representing clusters by typical units has theoretical and computational advantages when searching for anomaly patterns or aberrant behavior. Our point of departure is the hypothesis that there may be significant heterogeneities in the panel of time series from the different cross-sectional units and a natural clustering of such regions around an a priori unknown number of poles of attraction, in the sense that the patterns of the multisequences are more similar within each group than across groups.

Over the years, many methods have been used to find groups within patterns related to time periods. We choose the partitioning around medoids method (PAM) proposed by [13] for several reasons. First, it can operate directly on arbitrary distance matrices as a natural extension of distance-based

clustering on non-time series data. In fact, the computation of cluster centers does not require the presence of feature vectors allowing a flexible definition of what it means for two units (sequences or regions) to be "close". Moreover, there is no need for repeated calculation of distances since the PAM algorithm can simply look up distances from the distance matrix. [11] observe that distance-based clustering is a forced choice for applications requiring the classification of objects described exclusively by extrinsic properties in which only the relationships between individual data items are known (that is, their similarities or dissimilarities). Second, the typical representative of each cluster (the "medoid") is the most centrally located object (the multiple time series observed for a given region *i.e.* the region in the cluster whose average dissimilarity to all other regions in the same cluster is minimal; moreover, medoids are robust with respect to outliers. Third, PAM is a partitional algorithm that does not impose a hierarchical structure, which could not be necessarily present in the panel. Fourth, rather than selecting starting centers at random, PAM evaluates all possible starting centers and chooses the best centers to start cluster building and this gives consistent results when clustering is repeated.

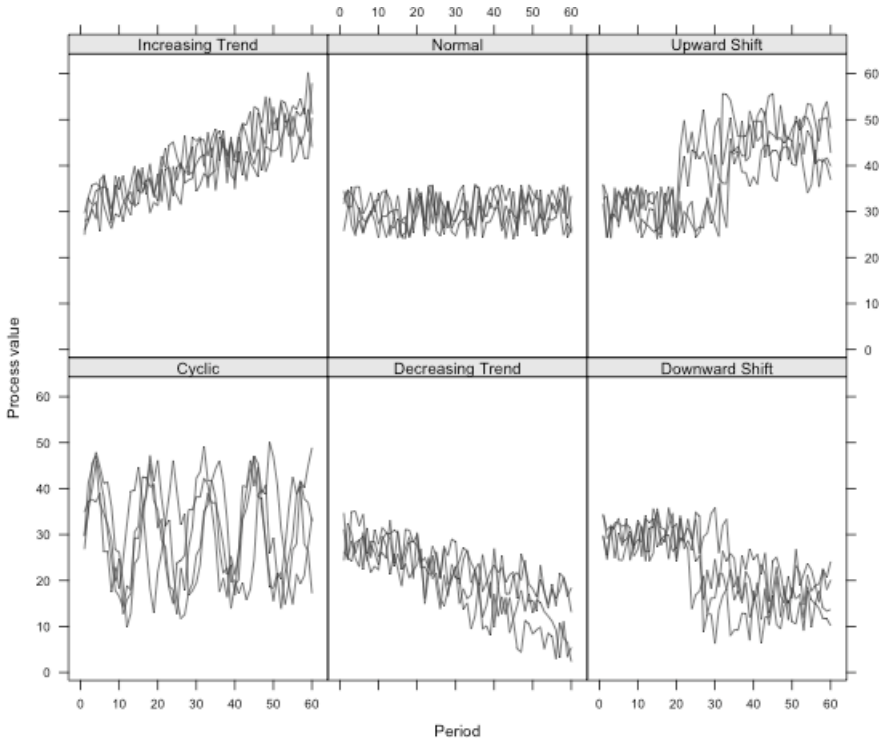
The PAM procedure searches k medoids that minimize the overall sum of distances of the sample units to their closest representative item [13] belonging to one of the k clusters. The number of cluster is assumed to be given. After an initial random selection of k tentative medoids, the PAM algorithm repeatedly tries to make a better choice of medoids. In this respect, we have used the *R* implementation of the PAM which assigns the initial clustering deterministically so that the results will always be identical and repeated runs to cope with random effects are not necessary. All of the possible pairs of items are analyzed, where one item in each pair is considered a representative item and the other is not. The quality of the resulting clustering is calculated for each such combination. Each unit is then assigned to the cluster corresponding to the nearest medoid until the medoids become fixed. The final result of PAM is a partition of the regions in k groups formed by multisequences with similar dynamics and guided by a leading unit that helps to explain how the group is shaped, how it evolves over time and what units we are to expect to be included in the group.

4.1 A modest experiment

A limited simulation was carried out to study the performance of the classification procedure discussed above. For this experiment we have chosen the "Synthetic Control Chart Time Series" (SCCTS) data set collected by [2] available from [8]. The data set contains 600 synthetically generated examples of control charts each having 60 time series values and there are six different classes with 100 representative examples from each class. We selected random samples without replacement of equal sizes from the group "normal" with code $1 - 100$; "cyclic": $101 - 200$; "increasing trend": $201 - 300$; "decreasing

trend": $301 - 400$; "upward shift": $401 - 500$; "downward shift": $501 - 600$. With the exception of "normal" patterns, all the other indicate abnormalities in the process that must be corrected. The data set is represented in Figure 2 reporting four examples for each of the six clusters. The following pairs of clusters will often be confused (normal/cyclic) (decreasing trend/downward shift) and (increasing trend/upward shift).

Fig. 2: Visual representation of example data.



We have divided the 60 observations of each series into ν consecutive equal subsequences of length $n_h = 60/\nu, h = 1, \ldots, \nu; \nu = 3, \ldots, 6$. For a number of cluster $k \in \{2, 3, 4, 5, 6\}$ we have selected $\{31, 21, 16, 13, 11)\}$ cases for each cluster from SCCTS of the type $1, 2, \ldots, k$. When $k \neq 6$, the classes were chosen at random (without replacement) form $1, 2, \ldots, 6$ in each replication of the experiment. To simulate time spans of different length, we have randomly omitted $0, 1, \ldots, t_1$ and $0, 1, \ldots, t_2$ contiguous entries at the beginning and end respectively of each subsequence, where $t_1 = t_2 \in \{0, 1, \ldots, \lfloor (n_h + 1)/4 \rfloor\}$. The kn_k ν-variate time series were classified using the global distance matrix $\mathbf{D}$ obtained using all the four weighting systems discussed in section 3 together with the PAM algorithm available in the R-package *cluster*.

To compare the stability of the above algorithms, the data generation has been repeated for 200 times for each system of weights, for each k and for each ν. In all the experiments, we set k to the to number of classes in the data set. Since we know the cluster to which each sequence belongs to, we can use the true clustering membership and assess the quality of the agreement between the true and the resulting cluster membership. The adjusted Rand index r has been chosen as cluster validation measure. See [12]. As it is well known, the r takes values on $[-1, 1]$ with one indicating a perfect concordance between the known and produced partition and values near zero or negative corresponding to cluster agreement found by chance. The numerical summaries are given in Table 1.
A high average of r means that clusters determined by the algorithm are in strict agreement with clusters from the original data. A low value of the standard deviation r indicates stability across the factors: number of groups, sequences and weightings of our experiment.

The optimal segmentation of the original time series is for $p = 4$ for which r is generally greater than for other alternative subdivisions. The mean value of r increases until it reaches its peak point at $p = 4$ at which point, precipitation occurs. This "evolution" does not always occur for $k = 2$, probably because of the major role played by the pairs of classes that are most often confused with each other (normal/cyclic and decreasing trend/downward shift).

The impact of missing values at the start and/or at the end of the sequences, apparantly, do not compromise the capable of capturing the dynamics of the time series, at lest not for the amount of missingness adopted in our experiment. In fact, the averge value of r is always around 0.5 for $k = 6$.

Weights based on the equalization of the mean of partial distances present the highest standard deviation. This is confirmed by the entries in the column Ar and in the row Ac in which is reported, respectively, the average across the values of k and the average across the number p. The Distatis weights tend to achieve slightly better results in terms of mean and standard deviation of r than the other systems. However, the Distatis method is a highly demanding technique and, where necessary, can be substituted by the weighting system based on the equalization of the standard deviation of the partial distances which has achieved comparable results.

4.2 An application to an unbalanced data set

The availability of fairly comparable data for time series of cross-sections has induced researchers to investigate and define the best way of utilizing the large quantity of information contained in the data. Pooled data are important in this sense because they contain the information necessary to deal with both the intertemporal dynamics and the individuality of the units being investigated [6]. For instance, it is usual to pool all time series data and estimate a common regression model or a global time series model instead of performing the usual identification, estimation and validation steps for each one ([9]).

Table 1: Average and standard deviation of r from 200 simulations of SCCTS

Weighting	p		$k=2$	$k=3$	$k=4$	$k=5$	$k=6$	Ar
Distatis	3	$E(R)$	0.604	0.618	0.548	0.513	0.476	0.552
		$Std(R)$	0.119	0.032	0.013	0.005	0.004	0.035
	4		0.659	0.636	0.592	0.545	0.521	0.591
			0.111	0.024	0.014	0.007	0.005	0.032
	5		0.655	0.627	0.582	0.532	0.506	0.580
			0.122	0.031	0.017	0.007	0.005	0.036
	6		0.656	0.584	0.572	0.532	0.498	0.568
			0.110	0.040	0.013	0.006	0.004	0.035
	Ac		0.644	0.616	0.574	0.531	0.500	0.573
			0.116	0.032	0.014	0.006	0.005	0.034
Uniform	3	$E(R)$	0.627	0.613	0.528	0.505	0.464	0.547
		$Std(R)$	0.130	0.036	0.014	0.005	0.004	0.038
	4		0.670	0.621	0.576	0.528	0.493	0.578
			0.127	0.031	0.014	0.006	0.004	0.036
	5		0.682	0.614	0.563	0.506	0.482	0.569
			0.115	0.038	0.017	0.007	0.004	0.036
	6		0.698	0.585	0.563	0.518	0.480	0.569
			0.118	0.049	0.016	0.006	0.003	0.038
	Ac		0.669	0.608	0.558	0.514	0.480	0.566
			0.123	0.039	0.015	0.006	0.004	0.037
Equal μ	3	$E(R)$	0.628	0.608	0.533	0.508	0.466	0.549
		$Std(R)$	0.123	0.035	0.014	0.123	0.123	0.084
	4		0.670	0.639	0.592	0.541	0.516	0.592
			0.119	0.026	0.013	0.123	0.123	0.081
	5		0.672	0.624	0.575	0.529	0.501	0.580
			0.112	0.035	0.123	0.123	0.123	0.103
	6		0.686	0.595	0.577	0.521	0.498	0.575
			0.112	0.045	0.123	0.123	0.123	0.105
	Ac		0.664	0.617	0.569	0.525	0.495	0.574
			0.117	0.035	0.068	0.123	0.123	0.093
Equal σ	3	$E(R)$	0.630	0.600	0.534	0.504	0.458	0.545
		$Std(R)$	0.111	0.028	0.012	0.006	0.005	0.032
	4		0.691	0.604	0.591	0.535	0.497	0.584
			0.106	0.038	0.015	0.009	0.004	0.034
	5		0.684	0.616	0.579	0.533	0.493	0.581
			0.104	0.033	0.017	0.007	0.005	0.033
	6		0.628	0.603	0.550	0.522	0.486	0.558
			0.112	0.028	0.013	0.007	0.005	0.033
	Ac		0.658	0.606	0.564	0.524	0.484	0.567
			0.108	0.032	0.014	0.007	0.005	0.033

All the calculations are executed by the statistical language R, [17] and the programs are available by the authors.

To illustrate our procedure, we use an example drawn from [3] reported as *EmplUK* data set in the package *plm* in *R*. The principal data source used is the published accounts of $p = 140$ quoted companies whose main activity is manufacturing and for which we have seven or more continuous observa-

tions during the period 1976-1984. The panel is unbalanced both in the sense that we have more observations on some firms than on others, and because these observations correspond to different points in historical time. We have $(\beta_{r,s,j} - \alpha_{r,s,j}) \geq 7$ and $(\nu_j = 9)$. The panel comprises $m = 4$ indicators: employment, real wage, gross capital stock and industry output. More information about the sample and the construction of these variables is provided in [3]. Table 2 reports the weights determined by the four systems considered in our paper.

Table 2: Weights for employment panel data

Weights	Employment	Wage	Capital stock	Output	ARI
Distatis	0.3136	0.1524	0.2888	0.2452	0.147
Uniform	0.2500	0.2500	0.2500	0.2500	0.149
Equal μ	0.2487	0.2243	0.2324	0.2946	0.233
Equal σ	0.2244	0.2991	0.2641	0.2124	0.160

It is immediately evident that, with the exception of the system of equal weights, each other system privileges a different variable giving rise to a different global or consensus distance matrix.

[3] allocated the companies to one of nine broad sub-sectors of manufacturing industry according to their main product by sales and value-added in that sector has been used as a measure of industry output. We have considered $k = 9$ the number of clusters to be identified. In this specific example, we have chosen the weights of the partial distance matrices according to the adjusted Rand index reported in the last column of Table 2. Weights based on the equalization of the mean of the partial distances adhere more strictly to the classification proposed by [3]. Table 3 shows the confusion matrix indicating the correct/incorrect predictions for each cluster and each sector.
There is an interesting overlap between sectors and clusters, but the number of firms misclassified is very high. Nonetheless the error rate could be reduced by aggregating very similar sub-sectors and eliminating several multivariate outliers (or abnormal situations) such as units $\{8, 89, 139\}$ which have been identified by our analysis.

5 Conclusions and suggestions for future research

Time series clustering is primarily based on the identification of a measure of distance between time series that takes into account the value and ordering of the points along the sequences; then the matrix of interdistances between units could be used as a raw input into well-known clustering procedures such as PAM methods. In this paper, a simple approach for performing similarity

Table 3: Confusion matrix for the *EmplUK* data set

	Cluster								
Sector	1	2	3	4	5	6	7	8	9
1	0	7	0	0	1	3	5	0	1
2	0	1	0	0	1	4	1	5	0
3	0	0	0	10	2	0	0	0	0
4	0	2	0	0	1	17	0	9	0
5	0	5	1	1	1	3	0	1	1
6	0	0	0	0	1	2	0	0	2
7	6	0	9	0	0	1	0	0	0
8	0	3	0	0	0	0	5	0	7
9	0	9	0	0	3	2	1	2	4

search in time series cross-sectional data has been proposed. TSCS data are important because they can exploit information more efficiently than cross-section alone or time series analysis alone. In this type of data, a fixed set of units is described by many time-dependent variables, each of which can provide a separate and potentially different basis of classification. One important feature of the method exposed in the previous sections is that a global or consensus distance matrix can be obtained as a weighted average of the partial distance matrices computed for the various indicators included in the data set. The method is appealing because of its easy of interpretation and good performance with multivariate short time series, particularly if the differential weights for each partial distance matrix are chosen to ensure that the variables with a greater classification ability are the most influential. Our main contribution in this work is to provide a distance function that is able to quantify the proximity between short-run dynamics even in the case the sequences have a different time span both between time-varying indicators and between cross-sectional units.

The main drawback of the method is the course of dimensionality associated with our methodology: when a large number of indicators are needed to assess properly the multidimensional nature of a problem, the differences between units are expected to be significantly reduced so that only a few basic stylized facts can emerge from the huge amount of information that is available. There are some doubts about the possibility that problems with high-dimensional data sets can be solved with a weighted average of distances. Future research will focus on system to compress the information contained in a TSCS data set. For instance, grouping the indicators into a restricted set of distinct constructs with separate effects on theoretically relevant patterns can be useful to prevent the decline in performance when large data sets are merged. Also, using statistical procedures that exploit the spatial nature of the TSCS data can reduce the classification error rate

References

1. Abdi, H. and O'Toole, A. J. and Valentin, D. and Edelman, B. DISTATIS: the analysis of multiple distance matrices. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, (USA), 42–47, (2005)
2. Alcock, R. J. and Manolopoulos, Y. Time series similarity queries employing a feature-based approach. In 7th Hellenic conference on informatics, Ioannina, Greece, (1999).
3. Arellano, M. and Bond, S. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. The Review of Economic Studies, 58, 277–297, (1991).
4. Bagnall, A.J. and Janacek, G.J. and de la Iglesia B. and Zhang, M. Clustering time series from mixture polynomial models with discretised data. Proceedings of the second Australasian Data Mining Workshop, 105–120, (2003).
5. Beck, N. Time-series-cross-section data: what have we learned in the past few years? Annual Reviews of Political Science, 4, 271–293, (2001).
6. Dielman, T.E. Pooled cross-sectional and time series data: A survey of current statistical methodology. The American Statistician, 37, 111–122, (1983).
7. D'Urso, P. Dissimilarity measures for time trajectories. Journal of the Italian Statistical Society, 9, 53–83, (2000).
8. Frank, A. and Asuncion, A. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences, (2010).
9. Fruhwirth-Schnetter, S. and Kaufmann, S. Model-based clustering of multiple time series. Journal of Business and Economic Statistics, 26, 78–89, (2008).
10. Gower, J. C. and Legendre, P. Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification, 3, 5–48, (1986).
11. Handl, J. and Knowles J. Multiobjective clustering around medoids. Proceedings of the Congress on Evolutionary Computation (CEC 2005), Vol. 1, 632–639, (2005).
12. Hubert, L. and Arabie, P. Comparing partitions. Journal of Classification, 2, 193–218, (2005).
13. Kaufman, L. and Rousseeuw, P. J. Finding groups in data: an introduction to Cluster Analysis, John Wiley & Sons, New York, (1990).
14. Kundu, S. Conflating two polygonal lines. Pattern Recognition, 39, 363–372, (2006).
15. Liao, T. W. Clustering of time series data - a survey. Pattern Recognition, 38, 1857–1874, (2005).
16. Pavoine, S. and Vallet, J. and Dufour, A-B and Gachet, S. and Herve, D. On the challenge of treating various types of variables: application for improving the measurement of functional diversity. Oikos, 18, 391–402, (2009).
17. R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, (2009).
18. Tarsitano, A. Classification of multiple short time series. In: Atti del convegno ISI2009. Pescara, september 23-25, Cleup, Padova (Italy), 1, 539–542, (2009).
19. Tarsitano, A. Classification of short time series. Advances and Applications in Statistical Sciences, 1, 215–231, (2010)