Task 7: Use of classification techniques for time series

Technical report on classification techniques

GRETA

1. Introduction

Clustering and classifications are two topics extensively discussed in the statistical literature with applications in many areas, including Physics, Biometrics, Information Technology, Economics, Finance, and in general in the data mining literature. As pointed out by Everitt (2001), Xu and Wunsch (2009), among others, classifying an object in a homogenous category may help in inferring its properties and features by looking at the elements included in that category.

Both clustering and classification methods have a common purpose: to split a large number of objects into smaller groups which are more homogeneous that the entire set, in such a way that the within group similarity is maximized and the between group similarity is minimized. The grouping of objects could follow different approaches, which are all based on a subjectively chosen measure of similarity or dissimilarity (in general terms a proximity measure). Note that the groups obtained from different clustering or classification approaches could vary in a sensible way, making the choice among alternative implementations very difficult. Unfortunately, the statistical literature does not yet contain a general criterion allowing a robust comparison among the outcomes of different clustering/classification algorithms. Despite this limitation, these grouping techniques have been proven to be useful in many disciplines, and using both cross sectional and time series data.

Within this work, following Xu and Wunsch (2009), we make a clear distinction between clustering and classification. In fact, despite the two words could be considered as synonyms, we label as classification a procedure which is directly controlled by the user (a supervised classification system). Differently, clustering is a procedure which is governed by an algorithm (an unsupervised classification system). To make an example, grouping time series on the basis of their integration properties (trend-stationary against difference-stationary) is a classification because, for a given series and on the basis of a test, the user decides the group to which it belongs; moreover, the groups are defined a-priori and clearly identified. Differently, the grouping of time series with respect to a similarity measure based, for instance, on the Pearson correlation coefficient, could be made by a numerical algorithm that produce a hierarchical structure across the objects, without the intervention of the user, and thus in an unsupervised case, giving a

clustering exercise. Notably, the groups created by an unsupervised classification could not be labelled as easily as in the case of the supervised classification. In other words, the supervised classification is based on a set of groups defined a-priori, while the unsupervised classification creates a set of groups data-driven which could be labelled only ex-post. We stress that labelling groups created by a clustering algorithm using the features of the objects included in each group may not be always possible, even if some help could come from other statistical approaches such as discriminant analysis.

We also highlight that, in the data mining literature, the word classification is often matched with the use of algorithms that try to predict the class of an object using available information on other objects. This could be done by a number of algorithms and approaches including neural networks. Within this paper, we are not interested in the prediction aspects but rather in the creation of the groups. Nevertheless, neural networks and genetic algorithms are included in the following sections as tools within clustering approaches.

The scientific literature already includes some surveys and books on clustering methods, Everitt (2001), Berkhin (2006), Xu and Wunsch (2009) among others, and, to our best knowledge, only one survey in the time series context, Liao (2005). This last contribution focuses on papers published in several scientific areas ranging from physics to finance, but the weight of economic and financial applications was really minor and many interesting contributions, most of them appeared in the last few years, were not included. Therefore, our work will fill the gap providing an extensive survey of time series clustering techniques which have been used or could be used for the grouping of economics and financial time series. As the reader may note, the references that will be cited are from journals belonging in the largest part to non-economic disciplines but tackling economic applications. This is a consequence of the development of clustering techniques as tools for data mining in areas related to non social sciences.

However, within economics and financial applications, clustering and classification approaches could have very relevant applications as we will show in a following sections. To summarize the possible uses of clustering (some of them not yet appeared in the literature) we mention the following. In the economic context, time series clustering could be used to detect comovements within time series by means of

clustering methods that compare the evolution of time series levels with approaches derived from dynamic time warping (which is a proximity measure), or their dynamic. Series that commove could then be used as predictors one for the other or to create composite indices by extracting the principal components. Similarly, we could group time series defining a proximity with respect to a given indicator obtaining groups which could be identified as leading, lagging and coincident, giving a further tool for the construction of composite indices. The search for similarities could be performed both on the series levels as well as on their structural components, seasonal, trend-cycle and irregular, to create groups homogeneous in term of trend-cycle (useful for business cycle dating), or similar with respect to the seasonal filter. The last case would allow a simpler seasonal adjustment procedure that would apply the same filter over series belonging to the same cluster (as suggested by Focardi and Fabozzi, 2004). Combining series with similar patterns or components, with or without a reference to a given index, could be useful also in term of point forecast of economic series. Furthermore, clustering methods may allow the identification of economic sectors or geographical areas with similar character and this additional piece of information could be used to improve the effectiveness of policy decisions (Crone, 1999).

In financial applications, time series clustering methods could be applied in the creation of classes of objects with similar credit rating, or market risk features. This would provide useful information for risk management purposes but also for investment management. In addition, clustering outcomes could supply a data driven segmentation of a number of financial instruments with relevant applications in asset management. Finally, data driven groups could be compared to a-priori groups associated to a classification for descriptive purposes that include the check for the degree of agreement and the possible impact on portfolio diversification (Pattarin et al., 2004).

Within this survey we will not consider applications of time series subsequences clustering. In fact, despite the possible relevant approach in finance (subsequences matching may be associated to the identification of technical analysis “figures”), they cannot have a direct interpretation apart in very specific circumstances, Inniss (2006). Some discussion on these approaches can be found in Keogh et al. (2003) ad Chen (2007).

The report proceeds as follow. Section 2 provides a short description of the time series classification, or supervised classification. The subsequent sections are devoted to the clustering of time series, or unsupervised classification. Section 3 summarizes the main aspects of time series clustering and Section 4 surveys some empirical studies of time series classifications in economic and finance. Section 5 concludes. The report contains also a more technical appendix on time series clustering.

2. Time Series Classification

The classification of a time series could be done in many ways, for example by referring to its descriptive statistics or to the random process which constitutes its data generating process (DGP). In this section we refer to the second characterisation that gives the advantage to consider the time series as an episode of a more complete process (with respect to its stochastic features). As we previously argument, the following characterisations could be used to perform a supervised grouping of time series.

For that reason it is convenient, first of all, to distinguish the time series which are originate by deterministic function of the time from that originated by a random process.

2.1 Deterministic vs. Stochastic

Most time series deal strictly with deterministic functions of time: $x(t) = f(t)$. Consequently, the characterisation of a time series is determined by the choice of the function *f(t)*. In a wider meaning, the function *f(t)* represents any mathematical expression, rule, or table. Because of this, future values of any deterministic time series can be calculated from past values. These time series are relatively easy to analyse as they do not change randomly, and we can make accurate assumptions about their past and future behaviour.

The literature on random process also consider deterministic processes determined by the Wold decomposition theorem. Here we stress that this “deterministic” processes are,

in effect, random processes and coincide with the definition of *singular* processes originally given by Wold (1938).

2.1.1 Deterministic Time Series

In the class of deterministic time series we can distinguish some sub-classes as:

2.1.1.1. Continuous-Time vs. Discrete-Time

As the names suggest, this classification is determined by whether or not the time axis is discrete (countable) or continuous. A continuous-time series will contain a value for all real numbers along the time axis. In contrast to this, a discrete-time series comes from sampling at non-continuous time, sometime created by using the *sampling theorem* to sample a continuous time series, so it will only have values at equally spaced intervals along the time axis.

2.1.1.2. Analog vs. Digital

The difference between analog and digital is similar to the difference between continuous-time and discrete-time. In this case, however, the difference is with respect to the value of the function (y-axis). Analog corresponds to a continuous y-axis, while digital corresponds to a discrete y-axis. An easy example of a digital time series is a binary sequence, where the values of the function can only be one or zero.

2.1.1.3. Periodic vs. Aperiodic

Periodic time series repeat with some period *T*, while aperiodic, or nonperiodic, time series do not. We can define a periodic function through the following mathematical expression, where *t* can be any number and *T* is a positive constant:

$$f(t) = f(T + t) \qquad (1)$$

The *fundamental period* of function, *f(t)*, is the smallest value of *T* that the still allows (1) to be true.

2.1.1.4. Causal vs. Anticausal vs. Noncausal

Causal time series are time series with values zero for all negative time, while anticausal are time series with values zero for all positive time. Noncausal time series are time series that have nonzero values in both positive and negative time.

2.1.1.5. Even vs. Odd

An even time series is any time series f such that $f(t) = f(-t)$. Even time series can be easily spotted as they are symmetric around the vertical axis. An odd time series, on the other hand, is a time series f such that $f(t) = -f(-t)$.

Using the definitions of even and odd time series, it can be shown that any time series can be written as a combination of an even and odd time series. That is, every time series has an odd-even decomposition. To demonstrate this, consider the following relation:

$$f(t) = \frac{1}{2}(f(t) + f(-t)) + \frac{1}{2}(f(t) - f(-t)) \qquad (2)$$

It can be seen that $f(t) + f(-t)$ fulfils the requirement of an even function, while $f(t) - f(-t)$ fulfils the requirement of an odd function.

2.1.1.6. Finite vs. Infinite Length

As the name applies, time series can be characterized as to whether they have a finite or infinite length set of values. Most finite length time series are used when dealing with discrete-time time series or a given sequence of values. Mathematically speaking, f(t) is a finite-length time series if it is nonzero over a finite interval $t_1 < f(t) < t_2$ where $t_1 > -\infty$ and $t_2 < \infty$.

The same classification can applied to the stochastic time series. In the following only the stochastic time series will be considered.

2.1.2 Stochastic time series: Stationary vs. non stationary

Stochastic time series is stationary if all its statistical properties do not vary with time. Processes whose statistical properties do change are referred to as non-stationary.

Statistical properties of time series are completely described by its family of finite dimensional distributions or its family of moment-generating functions if it exists. Thus, stationarity can be defined in terms of invariance of this family with respect the shifts of the stochastic time series along the time axis (*strict stationarity*). Similarly, the family of moment-generating functions must be invariant with respect the shifts. As a consequence of this property all the moments are time invariant.

A less strict definition of stationarity disregards the behaviour on the time of all the moments of the family of distributions and considers only those of the first and second moments (if they exist).

The stochastic time series for which the first and second moments are time invariant are said to be *stationary in wide sense* or *covariance stationary*. In this case the correlation function $\rho_{t,s}$, which generally depends on the choice of time *t* and time *s*, become dependent only on time lag $(t-s)$.

Examples of stationary time series in the discrete domain are given by iid sequences of random variables, white noise time series, and all the elements belonging to the class of general linear sequences (*regular sequences* in the sense of the Wold decomposition). This last includes the class of Autoregressive Moving Average sequences (ARMA).

In the case of continuous time domain, we can have the example of time series constructed by superpositions of n periodic oscillations of different frequencies (time series with a discrete spectrum). The Bochner-Khinchin theorem provides the general (spectral) representation of any covariance stationary time series in continuous domain.

One of the main advantages of the stationary property is given by the possibility to apply the ergodic theorem. According to the ergodic theorem, the mathematical expectation of both the time series x_t and the product $x_t x_{t+k}$, obtained by averaging the corresponding quantities over the whole space of experimental outcomes, can be replaced by the time averages of the same quantities. In other words, the ergodic theorem shows that it is possible, with probability one, to confine oneself (under specific conditions) to a single realisation of the time series and this constitutes a more favourable condition for applications to non experimental data.

2.1.2.1. Trend stationary vs. difference stationary

Non stationary time series can be composed by a stochastic covariance stationary component and a non stationary deterministic component. On the other hand, a non stationary time series can be originated by violations of the stationary conditions.

For example, let $u_t \sim ARMA(p,q)$ and the linear trend $f(t) = a + bt$, then the time series:

$$x_t = f(t) + u_t \tag{3}$$

is non stationary, because the first and the second moment depend on time. Also the process defined by:

$$x_t = x_{t-1} + \varepsilon_t \tag{4}$$

where ε_t is white noise, is non stationary because it is a random walk, with variance and correlation function which depend on time.

In the recent literature a distinction between trend stationary (TS) and difference stationary (DS) is preferred. This distinction is justified by the potential cost of misspecification of the data generating process mainly in terms of forecasting.

The two classes of non-stationary processes "have radically different implications for forecastability when the parameters of the processes are known: forecast-error variances grow linearly in the forecast horizon for the DS process, but are bounded for the TS process. This is unsurprising given that the unit root indefinitely accumulates previous disturbances, whereas in the TS process with known parameters, the conditional h-step ahead forecast error is simply that period's disturbance term. Uncertainty plays an add-on role in the TS process, but is integral to DS".

DS process is also named integrated process, indicated by *I(d)*, where *d* is the power of difference operator Δ applied to the process for obtaining stationarity. A stationary process is often named *I(0)* process, even though is more correct to say "process without unit root". In fact, following the definition of Johansen (1995):

A stochastic process Y_t *which satisfies that* $Y_t - E(Y_t) = \sum_{i=0}^{\infty} C_i \varepsilon_{t-i}$ *is called I(0) if* $C = \sum_{i=0}^{\infty} C_i \neq 0$.

then we can have stationary process that are not *I(0)* as in the case of the simple example:

$$Y_t = \varepsilon_t - \theta \varepsilon_{t-1} \tag{5}$$

which is *I(0)* only for $\theta \neq 1$.

There are substantial differences in appearance between a time series that is *I(0)* and another that is *I(1)*. In this respect a detailed discussion is given, for example, in Feller (1968) or in Granger and Newbold (1977).

Here we summarise the most evident different behaviours of the two class components.

(a) If $x_t \sim I(0)$ with zero mean then (i) the variance of x_t, is finite; (ii) an innovation has only a temporary effect on the value of x_t; (iii) the spectrum of x_t, $f(\omega)$, has the property $0 < f(\omega) < \infty$; (iv) the expected length of times between crossings of $x = 0$ is finite; (v) the autocorrelations, ρ_k, decrease steadily in magnitude for large enough *k*, so that their sum is finite.

(b) If $x_t \sim I(1)$ with $x_0 = 0$, then (i) variance x_t, goes to infinity as t goes to infinity; (ii) an innovation has a permanent effect on the value of x_t, as x_t is the sum of all previous changes; (iii) the spectrum of x, has the approximate shape $f(\omega) \sim A\omega^{-2d}$, where *d* is the order of integration, for small ω so that in particular $f(0) = \infty$; (iv) the expected time between crossings of $x = 0$ is infinite; (v) the theoretical autocorrelations, $\rho_k \to 1$ for all *k* as $t \to \infty$.

Other effects of misspecification are linked to the spuriously detrending integrated time series. Nelson and Kang (1981, 1983) argue that the regression of a driftless random walk against a time trend will result in the inappropriate inference that the trend is significant. Further, detrended random walks will exhibit spurious correlation. Integrated processes also pose problems for the empirical worker because of the probabilistic properties of the time series. In particular, conventional strong laws and central limit theory do not apply to standardized sums of the realizations of an

integrated process. These probabilistic properties and their statistical implica- tions have been extensively analyzed in work by Phillips (1986, 1987a, 1987b) and Phillips and Durlauf (1986).

2.1.2.2. With/without seasonal components

Decomposition of a series into a set of non-observable or latent components may be useful in time series analysis. Following the pioneer work of Persons (1919) we can think a time series composed of four types of fluctuations:

(1) A long-term tendency or trend.
(2) Cyclical movements super-imposed upon the long-term trend. These cycles appear to reach their peaks during periods of prosperity and their troughs during periods of depressions, their rise and fall constituting the business cycle.
(3) A seasonal movement within each year, the shape of which depends on the nature of the series.
(4) Residual variations due to changes impacting individual variables or other major events such as wars and national catastrophes affecting a number of variables.

Traditionally, the four variations *have been assumed to be mutually independent from one another* and specified by means of an additive or multiplicative decomposition model.
In the econometric literature there is a perennial question about the modelling the seasonality or disregarding this component using seasonality adjust data: Why do we seasonality adjust economic time series? Arguments are abundant on both sides, namely those who oppose it and those who agree (see, e.g., Ghysels (1996), Ghysels and Osborn (2001) for further discussion and references).
Ghysels (1988), among others, has argued that “economic theory” does not yield the decomposition used to seasonally adjust economic time series. Ghysels, in fact, showed that standard economic models do not yield orthogonal decompositions.

Econometricians have an enormous interest in forecasting and modelling seasonal time series. They understood that the inclusion of explicit descriptions of a trend and of seasonality in an econometric time series model is appropriate from a modelling and

forecasting point of view. Excluding such a description would lead to senseless out of-sample forecasts. Furthermore, often they are interested in common patterns across economic variables, including common trends and common seasonality.

This is the reason why they claim from data providers to make available the original data against the quite recently common practice to provide only so-called seasonally-adjusted data, obtained by . applying automatic filter routine to original data.

Following Franses and Paap (2004) we can say that the mechanical seasonal adjustment "is rather harmful, at least if one intends to use the estimated adjusted data for subsequent modelling. This holds in particular for cases in which one is interested in (i) examining the trend and the business cycle, (ii) when one wants to see how innovations get propagated through the model, and (iii) when one wants to forecast time series…".

"If, in any case, one still is interested in separating seasonality from the time series, one seems better off using the so-called model-based methods, instead of the mechanical filtering methods of the Census Bureau. These methods also allow one to provide confidence bounds around the seasonally-adjusted data, thereby making it explicit that such data are estimates and should not be confused with the original data…".

The relevance to cluster time series through various synthetic characteristics as, for example, their variances or autocorrelation functions, arises from the fact that the recent proposal of so-called periodic time series models consider data which seem to have different time series properties across different seasons. These properties concern just the autocorrelations and the variances, as well as the links between two or more time series in a regression context.

2.1.2.3. With/without cyclical (trend/cycle) components seasonal components

Another time series decomposition often used for univariate time series modelling and forecasting:

$$X_t = u_t + \varepsilon_t \qquad (6)$$

where u_t and are referred to as the "signal" and ε_t the "noise". The signal comprises all the systematic components of models, i.e. trend-cycle and seasonal components, and ε_t is the noise component. This model is used in "signal extraction" where the problem is

to find the “best” estimates of the signal given the observations corrupted by noise . The “best” estimates are usually defined as minimizing the mean square error.

Economists and Econometricians put a lot of attention to the trend component, which represents long-term smooth variations. The identification and estimation of long-term trend have encountered serious difficulties caused by the fact that the trend is a latent (non-observable) component.

Dagum and Cholette (2006) consider “long-period” a relative concept “since a trend estimated for a given series may turn out to be just a long business cycle as more years of data become available. To avoid this problem statisticians have used two simple solutions. One is to estimate the trend and the business cycles, calling it the trend-cycle. The other solution is to estimate the trend over the whole series, and to refer to it as the longest nonperiodic variation”.

A distinction between deterministic and stochastic trend is already discussed in the previous section. The same distinction can be done between deterministic and stochastic business cycle.

Deterministic models may consist of sine and cosine functions of different amplitude and periodicities. Stochastic models, usually of the ARIMA type involving autoregressive models of order *p* with complex roots, have also been used to model the trend-cycle.

The decomposition of economic time series into trend and cyclical components are very common in the applied works. The main reason is the attempt to distinguish between permanent and transitory behaviours for their important implications in monetary and fiscal policy. Examples are given by the measurement of potential output (permanent), output gaps (transitory) or the short-term or transitory link between inflation and real activity to smooth business cycle fluctuations pursued by the Central Bank. Nelson and Plosser (1982) are the precursors in investigating whether macroeconomic time series are better characterized as stationary fluctuations around a deterministic trend or as non-stationary processes that have no tendency to return to a deterministic path.

The way to formulate a dynamic model for such components is somewhat controversial issue.

A good discussion of the different issues which deal with the specification of a time series model for the trend-cycle component is exposed in Chapter 7 of Harvey and Proietti (2005).

2.1.2.4. Lead/lag with respect to a given indicators (ex. business cycle phases)

The use of lead-lag relations to predict business-cycle turning points dates back to the years immediately preceding World War I and the 1920s. In those periods, the US forecasting services considered mainly the tendency of stock prices to lead and short-term interest rates to lag business activity. Later NBER adopted a more complete indicator system of coincident, leading, and lagging indicators proposed by the researchers Arthur Burns and Wesley Mitchell (Mitchell and Burns, [1938], 1961). This original designation of leading indicators was further investigated and refined until a composite index based on the 12 most promising leading indicators was first systematically released in 1968. Since that time, the composite leading index (CLI) has undergone a number of significant revisions.

Hamilton and Quiros (1996) summarised in the best way how the researchers tried to find what the leading indicators actually lead, from the use of spectral analysis to identify the phase shift relating the CLI, to the evaluation of the usefulness of the CLI for identifying turning points, until the discussion about their alternative definitions. They also test whether the CLI is most useful as a linear predictor or for identifying turning points.

A classification of business cycle indicators may be useful. Conference Board lists the "most reliable" leading indicators that have been produced and revised from time to time. Cyclical indicators are classified into three categories—leading, coincident and lagging— based on the *timing* of their movements.

Following the Conference Board descriptions, the *coincident* indicators, such as employment, production, personal income, and manufacturing and trade sales, are broad series that measure aggregate economic activity; thus, they define the business cycle.

Leading indicators, such as average weekly hours, new orders, consumer expectations, housing permits, stock prices, and the interest rate spread, are series that tend to shift direction in advance of the business cycle.

The *lagging* indicators, in contrast to the leaders, tend to change direction after the coincident series. Among the three types of indicators the lagging series would seem to have little practical value for business cycles. On the contrary, they are helpful in warning us of structural imbalances that may be developing within the economy.

"These indicators represent costs of doing business, such as inventory-sales ratios, change in unit labor costs, average prime rate charged by banks, and commercial and industrial loans outstanding. Consumer and social costs are also represented by lagging indicators, such as the ratio of installment credit outstanding to personal income, the change in consumer prices for services, and average duration of unemployment. Thus, an accelerated rise in the lagging indicators, which often occurs late in an expansion, provides a warning that an imbalance in rising costs may be developing.

Moreover, the lagging indicators help confirm recent movements in the leading and coincident indicators, and thus enable us to distinguish turning points in these series from idiosyncratic movements".

Similarly, we can list the cyclical indicators looking at their *direction* relative to the business cycle. In this respect, they may be classified into other three categories— pro-cyclical, counter-cyclical and acyclical.

Pro-cyclical indicator is one that moves in the same direction as the economy. So if the economy is increasing, this indicator is also increasing, whereas if we are in a recession this indicator is decreasing. Gross Domestic Product (GDP), Consumption, Price Deflators of GDP, Personal Savings Rate, Employment - Payroll Jobs, Average Hourly Earnings, Employment Cost Index, Producer Price Index (PPI), Consumer Price is an example of a pro-cyclic economic indicator, are some examples of pro-cyclical indicators that are usually also coincident.

(Fixed) Investment, Change in Inventories, National Association of Purchasing Managers, Index of Leading Economic Indicators, Consumer Confidence, are also pro-cyclical indicators but usually of leading type.

Countercyclical indicator is one that moves in the opposite direction as the economy. The unemployment rate gets larger as the economy gets worse so it is a counter-cyclic economic indicator. It is also of lagging type. Government Consumption, Net Exports, International Trade (Exports, Imports, Trade Balance) are counter-cyclical and usually coincident indicators.

An *acyclic* indicator is one that has no relation to the health of the economy and is generally of little use.

2.2 Discriminant Analysis, Vector Support Machine and Cluster Analysis

Following the interesting paper on Discriminant Analysis and Clustering by Gnanadesikan et al. (1989), we can distinguish two broad categories of classification problems.

In the first, one has data from known or pre-specifiable groups as well as observations from entities whose group membership, in terms of the known groups, is unknown initially and has to be determined through the analysis of the data. In statistical terminology it falls under the heading of *discriminant analysis*. Discriminant analysis, also known as *supervised classification*, therefore uses known classifications of some observations (the *training set*) to classify others. The number of classes is assumed to be known.

Recent advances in statistics, generalization theory, computational learning theory, machine learning and complexity have provided new methods for supervised classification. Among these new methods, Support Vector Machines have attracted most interest in the last few years. Support Vector Machine (SVM) is a novel learning machine introduced first by Vapnik (1995), an algorithm based on the maximization of the margin of confidence of the classifier (where the margin is related to the minimal distance between the points and the classifier in some Euclidean space). Some studies reported that SVM was competitive and outperformed other classifiers including neural networks and linear discriminant analysis in terms of generalisation performance.

On the other hand there are classification problems where the groups are themselves unknown a priori and the primary purpose of the data analysis is to determine the groupings from the data themselves so that the entities within the same group are in some sense more similar or homogeneous than those that belong to different groups. This topic will be described in the following sections.

3. Time series clustering

Following the introduction, the classification of time series using clustering approaches belongs to the "unsupervised classification schemes". In this section we review the purposes and the approaches of time series clustering following the contributions of Liao (2005) and, Xu and Wunsch (2005, 2009) and with reference to a set of works appeared in the literature of different disciplines including economic and finance but also computer sciences and physic.

In origin, clustering methods have been applied for the identification of latent structures on static data sets. More recently, the approaches focusing on organizing data using concepts of similarity and aiming at maximizing the within groups similarity and, at the same time, minimizing the between group similarity, have been applied to time series data sets for the purposes listed in the introduction.

Following Han and Kamber (2001), and Liao (2005), clustering methods applied to time series data could be associated to three main categories: partitioning methods, hierarchical methods and model-based methods. Two other groups of clustering approaches cannot be easily extended to the time series domain: density-based methods and grid-based methods. For a description of these approaches see Han and Kamber (2001).

This contribution will present a review of the most known and used clustering approaches but with a subtle distinction from the statistical literature. In fact, despite we will group clustering algorithms into the classical three sets previously mentioned (partitioning, hierarchical and model based), our definition of model based clustering methods is different from the standard one. We define as model based (or statistical) a clustering method that uses an algorithm possibly based on a statistical model, but which cannot be considered as a special case of partitioning or hierarchical methods. The distinction will become clear with an example: a hierarchical clustering based on the estimated coefficients of an ARMA model is a model based clustering in the traditional interpretation, while in this paper it is a hierarchical clustering. We think this alternative approach avoids possible confusion between the method to be used for creating the groups and the additional elements that need to be specified when considering a clustering exercise: the grouping quantity, or distance between objects,

and the inputs of the grouping method. Within the inputs, and again following Liao (2005), we may identify three possible groups of approaches: the one passing to the grouping method the entire time series (the raw data based case), the approaches passing some features of the time series (moments, correlations, periodograms etc.), and the cases where the inputs are coming from an estimated model or statistical procedure which are not labelled as features (such as the coefficients of ARMA models, a wavelet transformation or the forecast densities) which are called model based. Within our distinction of clustering methods (the grouping rules), model based features could be used as input for partitioning or hierarchical clustering while the traditional statistical literature classify all approaches using model based input as model based approaches.
In this section we will review the clustering methods, the distances proposed by the literature, and then describe the possible different inputs that could be passes to a clustering algorithm. Furthermore, in the next subsection, we will also provide a general discussion on which are the problems one may face when dealing with a time series clustering case. Note that, unless differently and explicitly specified, clustering and classification will be used as synonyms within this section.

3.1 Distance measures

The purpose of cluster analyses is to determine a set of groups of objects which are similar or close with respect to a number of features. A central element is thus given by the criterion defining the similarity or dissimilarity between objects. Using a more general formulation, we could reason in term of proximity measures, as in Xu and Wunsch (2009), where proximity could be either similarity or dissimilarity. Within clustering analysis, groups are created in order to:

maximize similarity within groups and/or minimize similarity between groups,

or

minimize dissimilarity within groups and/or maximize dissimilarity between groups.

Similarity and dissimilarity measures (or in one term proximity measures) have thus a central role. The literature on clustering and classification includes a large number of distance measures. Some of them are collected in Appendix A.2. At this point we introduce the most common distance, the Euclidean, which will be used in the empirical applications of the companion Technical Report on Classification Techniques for Time Series Data (GRETA 2009).

Given two vectors y_i and y_i of dimension l that may contain either a time series or one or more quantities derived from a time series, the Euclidean distance is defines as follows:

$$d\left(y_j, y_i\right) = \sqrt{\sum_{l=1}^{H}\left(y_{i,l} - y_{j,l}\right)^2} \qquad (1)$$

In the following, unless differently stated, we assume the Euclidean measure is used. However, the reader should note that other measures could be considered, taking into accounts the comments included in Appendix A.2.

3.2 Clustering methods

As we previously stated, from a general point of view, the approaches or methods which could be followed for clustering time series could be grouped into three main sets: partitioning methods, hierarchical methods and model-based methods.

3.2.1. Partitioning methods

Partitioning methods points at grouping a set of M time series into K groups containing at least one element. The most known methods are the k-means and k-medoids algorithms described in McQueen (1967) and Kaufman and Rousseeuw (1990), respectively. These approaches provide a crisp partition of the M time series (objects belong to just one single cluster). These approaches may be generalized

allowing objects to belong with a given probability to more than one cluster (or, in other words, the objects have a different degree of similarity with respect to more than one cluster). These extensions are given by the fuzzy c-means and fuzzy c-medoids algorithms, see Bezdek (1987) and Krishnapuram et al. (2001), respectively.

The idea behind partitioning methods is also at the base of genetic algorithms implementing clustering methods. For examples on this literature see Estivill-Castro and Murray (1998), Hall et al. (1999), Krishna and Murty (1999) and, Meng et al. (2002).

In Appendix A.3 we provide some details on two approaches: the K-means and fuzzy c-means algorithm, and the Genetic Algorithm for Medoid Evolution (GAME). The K-means method could find relevant applications in time series clustering given its purpose of grouping subjects by minimizing the distance with respect to the group centre. For instance, when the purpose of the clustering is the identification of series characterised by similar seasonal patterns, the input of the clustering algorithm could be either by the periodogram of the series at given frequencies, or the seasonal patterns extracted by appropriate models (for instance by TRAMO). The K-mean approach will provide groups which can be interpreted as series characterized by close seasonal behaviours. Note that the groups will not define a linkage structure across the series, this additional feature is given by the methods included in the following section.

3.2.2. Hierarchical methods

These methods create a tree-based structure which represents a sequence of nested partitions of the M series to be grouped. Hierarchical methods could be further distinguished into agglomerative and divisive depending on the starting point of the tree structure. Agglomerative methods start from singleton clusters and end to a cluster including all series while divisive methods do exactly the opposite (from the entire set of objects to singleton clusters).

Hierarchical approaches are generally graphically represented by dendograms or binary trees and the final clustering results are obtained by cutting at some level the dendrogram.

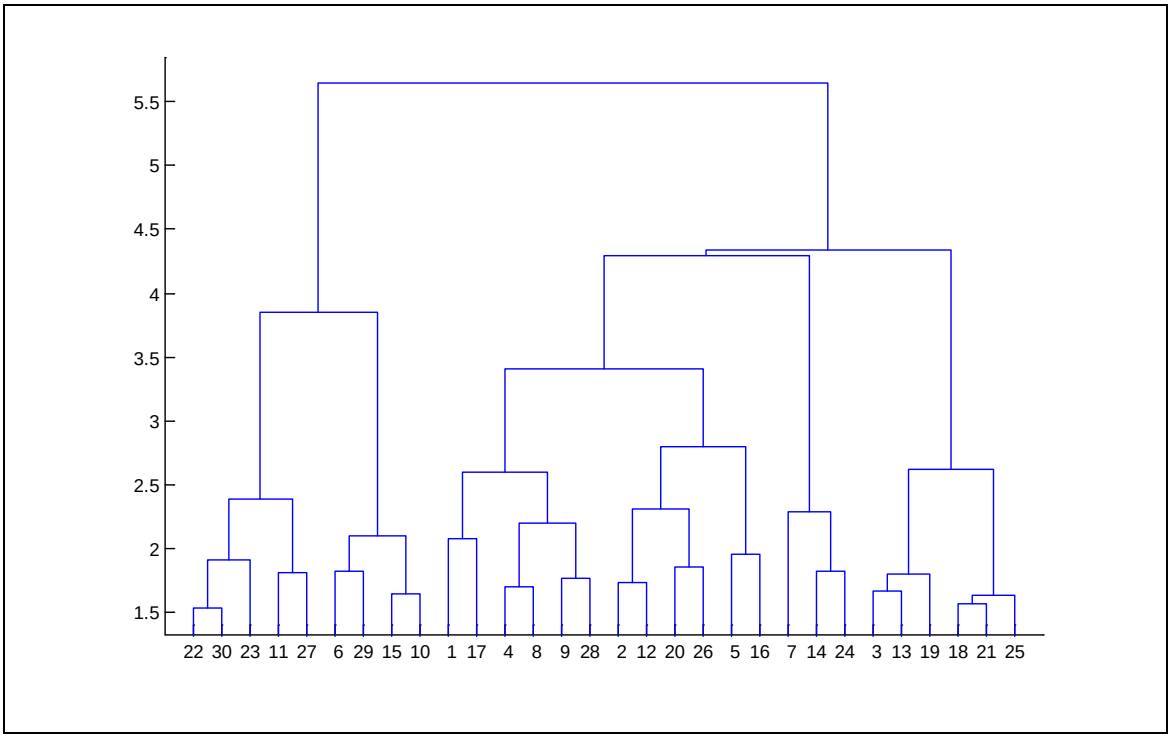

Figure 1: a Dendrogram

From the computation point of view, divisive methods are more intensive given that for a collection of M objects they need to consider 2^{M-1}-1 possible subsets. Agglomerative methods are thus generally preferred, even if they are still computational intensive methods requiring a number of objective function evaluations which is of order $O(M^2)$.

Agglomerative methods are implemented through a sequence of merge operations that build up clusters of objects. The approach starts with a set of M singleton clusters which are then subsequently grouped on the basis of a distance measure. Then the approach ends when all objects are included in a single group. Appendix A.4 reports some technical details of the agglomerative approach.

Hierarchical methods have been often used in economic and financial applications on time series clustering, see the following section for some examples. Inputs provided to hierarchical methods could be of different nature, starting from components capturing the dynamic of series (such as the correlation functions, the periodogram), or the series themselves. The resulting groups could be identified as set of series sharing common patterns or behaviours, with interpretations similar to those provided by partitioning methods. The additional element is given by the hierarchical structure, which contains information about the closeness of group of series. This could be of interest when the

classification based on the seasonal patterns include, for instance, industrial production indices extracted from the NACE classification up to some precision. The NACE classification has itself a hierarchical structure, and the clustering approach could evidence which component of a lower level precision in the NACE classification is closer to the upper level component. One question we could answer is, for instance: which industrial production index of the one digit NACE classification is closer in term of seasonal component to the total industrial production index?

[TO BE COMPLETED with contributions from Everitt for a deeper comparison across methods and from Xu and Wunsch, section 3.4, for more advanced approaches]

3.2.3. Model-based methods

In this group we include clustering approaches based on neural networks and with specific algorithms generally based on estimated models and statistical approaches which are not included in the partitioning and hierarchical methods. The literature often includes in this group partitioning or hierarchical clustering exercises for which the inputs are given by the outcome of estimated statistical models, such as the coefficients. We do not follow this standard practice and we consider the previous as special hierarchical and partitioning approaches.

3.2.3.1 Neural networks and Self-organizing maps

Neural networks are one example of model-based approaches for clustering and classification of time series. A specific class which has proven to be useful in this framework is that of self organizing maps, Kohonen (1990), and more recently Kiang (2001). In these peculiar networks, the inputs are represented by the feature matrix and the outputs are given by a map of neighbourhood relations across objects. By construction, self organizing maps perform clustering with the same spirit of k-means algorithms, with the advantage of not requiring the a-priori definition of the number of clusters.

Given the peculiar features of self organizing maps and, in general, of neural networks, these methods are appropriate for clustering exercises defined over feature vectors with the same dimension across objects. Consequently, raw data approaches with neural networks are discouraged. Wang et al. (2004) includes an example on the use of self-organizing maps where inputs are given by time series features.

3.2.3.2 Model based approaches

As we previously stated, we label as model based the clustering approaches which are using a specifically designed algorithm which is using model outcomes, or clustering approaches which are nested in a more general model estimation method. Due to the specificity of each algorithm, we will not present them in details but simply mention the works which are, in our opinion, the most interesting for economics and financial applications.

An example of the first group of approaches, specific algorithms not included in the partitioning or hierarchical classes, is in Otranto (2008). His algorithm is based on a subsequence of tests and comparison of time series on which variance models have been fitted. The method could be considered as an variation of the approach in Corduas and Piccolo (2008) and is related to that in Maharaj (2000).

Xiong and Yeung (2004), present an approach for clustering time series on the basis of mixture models following McLachlan and Basford (1988). They proposed an ARMA mixture approach with Bayesian elements to cluster a set of K time series into M groups whose components share a common data generating process. The main advantages of this class of methods are: the number of classes is identified by the model; they can handle series with different sample length given that the approach is based on a likelihood evaluation. A similar model based approach is in Oates et al. (2001) that combine Dynamic Time Warping and Hidden Markov Models for clustering of time series.

Finally, Beran and Mazzola (1999) proposed an interesting model based approach based on the application of smoothing filters of different amplitude to a set of time series. We believe this method has some potential for applications in economics and finance.

3.3 Defining clustering inputs

The grouping rules (the clustering method) and the distance measures are the two technical elements characterising a clustering exercise. A the third element is needed to complete from a methodological point of view the clustering application: the input of the clustering method, which, from a general point of view could be interpreted as a function of the empirical data.

Following Liao (2005) we group the possible inputs into tree sets: i) Raw data based: these methods use the time series as an input; ii) Feature based: they convert the time series into a set of features such as moments, correlations or cross-correlations, periodograms, among others; iii) Model based: these approaches consider as inputs elements coming from an estimate of a statistical model.

3.3.1 Raw data based methods

Raw data based methods are generally subject to a curse of dimensionality in the time series context. In fact, each observation over time of all time series is considered as a distinctive and relevant element by the clustering algorithm. As a result, on the one side, raw data based methods will create groups of series with similar patters (which could be useful in some financial and economic applications, such as the extrapolation of seasonal patterns or of pattern regularities); on the other side, the computation complexity increases with the sample length.

Raw data based methods are relevant if the purpose of the clustering is the identification of subsequences with multiple occurrences over time on a single time series. However, this has a limited applications in economics, but a far more relevant interpretation in finance where subsequences may be associated to the “figures” popularized by technical analysis.

We note that, generally speaking, raw data based consider as inputs the entire series. However, the time series could be provided to the clustering algorithms in their levels, period growth rates or annual growth rates. As a consequence, clustering exercises could present group of time series characterised by similar patterns on the levels (with possible scale effects), or over the growth rates. In this last situation, the groups could contain some relevant information with respect to the evolution of the business cycle.

In addition, the inputs of raw data based method could be represented by the components of macroeconomic time series, that is, the trend-cycle, seasonal and irregular component. In this additional circumstance, raw data based approaches allows the creation of groups characterised by similar trend-cycle evolution (or trend-cycle growth rates) with relevant application in business cycle (growth cycle) dating and detection. Alternatively, we may group time series of business indicators characterised by similar seasonal patterns allowing the application of similar seasonal filters.

Finally, raw data based methods represent the only case where the dynamic time warping measure could be used. In fact, dynamic time warping detects similarities between time series patterns irrespectively of scale effects. We also stress that raw data based approaches can be applied over a set of time series available over the same sample period and with the same sampling frequency.

The literature has proposed some approaches trying to reduce the curse of dimensionality in raw data methods. An interesting example in finance is given by Pattarin et al. (2004) that overcome the curse by extracting the principal components from the observation matrix (of time series observations), or Gavrilov et al. (2000) that compress the time series into fewer points.

We also include within the raw data based approaches all the studies suggesting the use of wavelet transformations of given time series, see Povinelli and Feng (1998), Chan and Fu (1999), Lin et al. (2004), and Zhang et al. (2006). A related approach we include within raw data methods is that of Beran and Mazzola (1999), that propose a hierarchical smoothing modelisation combining nonparametric kernel smoothing and parametric nonlinear regressions. They applied the proposed approach on musical time series but we believe interesting applications could be done also in economics and finance. In fact, within their framework, series may be represented as a combination of elements active at different time resolutions. This approach reminds the existence of global and local trends popularized in the technical analysis literature.

3.3.2 Feature based methods

An alternative and more effective approach to tackle the curse of dimensionality is to transform the time series into a smaller set of features. The possible choice of feature is quite extensive. We list here a set of the most appropriate for economic and financial

applications: i) Moments of the series (preferably over growth rates to avoid scale effects) up to orders 4 (higher orders could be used but with an increase in the uncertainty of their population values); ii) The Auto Correlation Functions or Partial Auto Correlation Functions that could be used to create groups of series characterised by similar dynamics. In this case a maximum lag should be fixed as a function of the sample length; iii) The periodogram evaluated at a common set of frequencies across all the time series considered; iv) The turning point dates associated to the time series.

The elements in points ii) and iii) can be evaluated over time series levels and growth rates if the purpose of the analysis is an economic application involving macroeconomic non financial data. Differently, in finance or in economic applications involving financial data, elements in ii) and iii) should be determined over growth rates and squared growth rates.

In addition, we note that the reliability (asymptotic properties) of features in i)-iii) is associated to the sample dimension, therefore calling for minimal sample dimension for robust and reliable clustering exercises.

By construction, feature based methods could be applied to time series of different length, the only impact being the reliability of some of the features included in the feature matrix, but without strong impacts on the overall clustering procedure.

Several papers proposed feature based clustering approaches with applications in economic and finance. Within the possible choices of moments and features characterizing the time series evolution, the variance plays a prominent role, in particular when dealing with financial variables. Among the studies dealing with the clustering of financial time series, we mention the approach of Micciché et al. (2003), one of the few classifying time series using a proxy of the variances, defined as

$$\sigma_{i,t} = 2\frac{\max\left(P_{i,t}\right) - \min\left(P_{i,t}\right)}{\max\left(P_{i,t}\right) + \min\left(P_{i,t}\right)} \qquad (1)$$

where the day *t* standard deviation is a function of the minimum and maximum price observed within the day. An interesting example combining different features is in Wang et al. (2004).

From a more general point of view, many contributions focused on the clustering of time series in the time domain, but some studies also considered the classification using frequency domain information. Those studies, starting from the contribution of Agrawal et al. (1993), are mainly based on the estimate of the power spectrum of a time series, which is made by the periodogram

$$P_i(\omega_h) = \frac{1}{T_i}\left|\sum_{t=1}^{T_i} x_{i,t} e^{-it\omega_h}\right| \qquad (2)$$

where T_i is the sample dimension of series i and ω_h is the frequency at which the periodogram is evaluated. Generally, the periodogram is computed for values of ω_h between 0 and π (or 0 and ½ in a normalized representation) which are generally fixed using the relation

$$\omega_h = \frac{h\pi}{\lceil T_i/2 \rceil} \quad h = 0,1,2,...\lceil T_i/2 \rceil \qquad (3)$$

where $\lceil a \rceil$ denotes the integer part of a. In this setup, features of the time series are represented by the periodogram values at a given set of frequencies.

As we already argument, within the clustering framework, a time series database could include series of different length. The approaches based on the periodogram are flexible enough to allow the classification irrespectively of the sample length. In fact, the periodograms of two series i and j, characterised by different lengths could be evaluated at a common set of frequencies. For instance, as discussed in Caiado et al. (2006, 2007, 2009), we could use the frequencies defined as in the previous equation using the smallest sample length, or add zeros to the shorter series to increase the sample length up to the desired value (the zero-padding procedure), or we could interpolate the periodograms at missing frequencies. In all cases, the feature vector would then include a set of periodogram values computed over a common set of m frequencies, and the distances could be determined, for instance, with the Euclidean metric as in Caiado et al. (2006, 2007, 2009), such as

$$d\left(y_i, y_j\right) = \sqrt{\frac{1}{m}\sum_{h=1}^{m}\left(P_i\left(\omega_h\right) - P_j\left(\omega_h\right)\right)^2} \qquad (4)$$

An additional feature of the use of periodogram ordinates in the feature vector is the existence of an asymptotic distribution for the ordinates. This allows the construction of a test for the equivalence of two periodograms evaluated on a common set of frequencies and thus the construction of a clustering exercise similar to that in Corduas and Piccolo (2008).

Caiado et al. (2006) contains an interesting comparison of clustering algorithms using different inputs, including raw data, features such as Auto Correlation Functions, Partial Auto Correlation Functions, Periodograms (standard and normalized), and estimated models (ARIMA models). They also consider a variety of distance measures from the Euclidean, the Mahalanobis, the AR metric of Piccolo (1990), and the Kullback-Liebler distance. Differently, Alonso et al. (2008) propose the use of the integrated periodogram to classify series

$$F_i\left(\omega_k\right) = \frac{\sum_{h=1}^{k} P_i\left(\omega_h\right)}{\sum_{h=1}^{m} P_i\left(\omega_h\right)} \qquad (5)$$

where m is again the total number of Fourier frequencies used for a series. The normalization of the integrated periodogram is not fundamental and may be excluded. The normalized version puts some emphasis on the shape while the un-normalized periofograms focus on the scale. The introduction of the integrated periodogram has many advantages: it provides smoother patters, it has good asymptotic properties and, its distribution always exists. In addition, it completely determines the underlying stochastic process. However, it also has a drawback: it does not work for non-stationary time series. Notably, Alonso et al. (2008) also introduce a distance measure based on integrals and two clustering algorithms for this periodogram based distance which are derivations of the K-medoid approach.

As noted in Caiado et al. (2009), the use of autocorrelations as components of the feature vector would produce similar results to the use of periodogram ordinates over a set of common frequencies.

Despite the appeal of periodogram ordinates as clustering inputs, Wang and Wang (2000) point out this approach could not properly identify the dissimilarities across the time series. The suggest to overcome this limitation by smoothing the periodogram ordinates by moving averages or weighting functions.

3.3.3 Model based methods

This third set of clustering methods aims at passing to the clustering algorithm the outcome of an estimated model such as: the estimated coefficients, with or without their covariance, in a raw format or after a transformation, or the forecast densities of a model.

Within a time series framework, the fitted models may belong to two large classes, the ARIMA models popularized by Box and Jenkins (1979) and the structural models by Harvey (1991). Both approaches provide as outputs of the estimation process a vector of estimated coefficients and a covariance matrix across estimated values. Passing these quantities to a hierarchical clustering algorithm may be combined with appropriately designed distance matrices, see section 3.2.7.

A different, but related, approach is that in Piccolo (1990), and Corduas and Piccolo (1999, 2008). The authors propose to use as features of each time series the autoregressive coefficients of the AR(∞) expansion of the best fitted ARIMA model . They then consider the following Euclidean distance

$$d\left(y_i = \left\{\pi_{i,k}\right\}_{k=1}^{K}, y_j = \left\{\pi_{j,k}\right\}_{k=1}^{K}\right) = \sqrt{\sum_{j=1}^{K}\left(\pi_{i,k} - \pi_{j,k}\right)^2} \qquad (6)$$

where $\pi_{i,k}$ is the lag *k* coefficient of the AR(∞) representation of the ARIMA model fitted on series *i*. Note that only *K* lags are used, allowing thus, by construction, to compare series of different lengths. Note that this distance is also called AR metric. The approach of Piccolo (1990) has a further advantage: the AR metric has an asymptotic

distribution allowing the construction of inferential procedures for the clustering of time series. Corduas and Piccolo (2008) apply such an approach, using a dichotomous variable derived from the AR metric and identifying pairs of time series that were accepting the null hypothesis of zero AR metric. In such a case, permutation algorithms have been used to create distance matrices which are block-diagonal.

Finally, Corduas and Piccolo (2008) also note that the AR metric is similar to a clustering based on the forecasting functions of ARIMA models given the relevance of AR(∞) coefficients in the prediction from that model class.

3.4. The optimal selection of group numbers and the comparison of alternative partitions

In the previous sections we presented alternative approaches for clustering data, in particular dealing with hierarchical and partitioning methods, and allowing for different inputs. However, on the one side, partitioning methods requires a-priori the definition of the number of clusters in which the collection of objects has to be divided, while, on the other side, both methods call for approaches allowing a comparison of the groups created using different algorithms or different data input.

The literature provides a number of methods for the comparison of clustering outcomes, distinguishing the case where the true classification is known, from the one where the true classification is unknown. Given the purposes of this report, we present in the following some criteria for the comparison of alternative classifications when the true groups are not known. Some details and references for criteria requiring the knowledge of the true classification could be found in Xu and Wunsch (2009).

If we consider a hierarchical clustering exercise, the standard tool for its validation is the Cophenetic Correlation Coefficient, see Rohlf and Fisher (1968) and Jain and Dubes (1988), among others. This measure assumes values between -1 and 1, and values close to 1 suggest a good agreement between the hierarchical structure and the data. As a result, given two alternative hierarchical clustering outcomes, the one with higher Cophenetic Correlation should be preferred.

Differently, when partitioning methods are compared, rules for the optimal choice of the number of clusters should be considered. In this case the literature contains several

methods than can be used. Without reporting them in details, we just mention that the basic idea is to construct a statistics that has to be maximized over the possible range of K, the number of groups. A survey of the methods, indicators and quantities could be found in Xu and Wunsch (2009), Milligan and Cooper (1985), and Gan et al. (2007).

3.5 Time series clustering: problems and solutions

The clustering of time series is affected by a number of problems that may limit its usefulness. Some elements have already been mentioned in the previous sections. Here we recall them providing some additional comments.

In principle, clustering techniques requires that the objects of study have the same dimension. Within a time series framework, however, this may not be the case. In fact, many economic or financial time series may have either different lengths, or different sampling frequencies. The reasons for this different size of the objects could be of different nature: revisions of economic series, newly added economic indicators, or, simply, unavailability of old data (as an example, economic data are not available for eastern European countries for years before mid 80's).

However, despite the different sample length of frequency, a time series clustering exercise could be in any case of interest. In such a situation, some approaches are feasible others are not. In fact, raw data based time series clustering approaches cannot be applied, while feature or model based approaches are feasible since they convert series of different length into vectors of features or parameters which are comparable across the objects.

Feature based and model based approaches should be also preferred when dealing with large datasets since they allow a relevant reduction of the computational burden behind any time series clustering method.

A second aspect that needs to be mentioned is the clustering of multivariate data. This may apply when the objective is the clustering of subjects characterised by a collection of time series. As an example, this may realize when we are interested in clustering countries by using the time series data of many economic indicators. Unfortunately, the

literature on these approaches is limited and computationally demanding. Some result are included in Maharaj (1999 and 2000).

The statistical literature considers also the clustering of entire time series paths for two main purposes: whole matching, that is grouping series with similar patterns, or subsequences matching, that is, searching for peculiar behaviours that repeat over time in the same series or in different series. However, subsequent matching may present some problem, see Keog et al. (2003), and Chen (2007) for a discussion on the interpretation problems of time series subsequences clustering. Further problems associated with whole range matching is given by the curse of dimensionality, time series may be quite long and create computational problems. These aspects favours the use of feature based approaches for time series clustering.

4. Applications in the Literature

We conclude this report with a section devoted to the economic and financial applications of time series clustering methods. Despite the possible applications are numerous, the literature includes a limited number of papers considering time series clustering of economic data. Many more contributions are related to financial data. References to applications in other areas of statistics are available in Liao (2005).

4.1 In economics

In Economics, most of the time series clustering applications available in the literature use Industrial Production data.

Caiado et al. (2006) present a clustering example based on a set of US seasonally adjusted industrial production indices showing that the clusters may be associated to different average growth rates over the sample period used. Differently, Corduas and Piccolo (2008) report an empirical application of the AR metric of Picoolo (1990) over a set of Industrial Production Indices for Italy. Their classification is based on a distance matrix defined over the rejection of the null hypothesis of equal ARIMA structure of pairs of series. Then, to perform the clustering they used a permutation algorithm.

Caiado et al. (2007, 2009) analyse the seasonally adjusted industrial production indices of many European and industrialised countries using a frequency domain distance. Their study reports that the countries are clustered in groups homogeneous with respect to the countries degree of development. Finally, Vilar et al. (2009) extend Alonso et al. (2006) on the use of L-norms based forecast densities and present an example based on the industrial production indices of 21 countries. They note that the clusters based on L^1 norm should be preferred. A related work is that of Galbraith and Lu (2001) which propose and discuss the use of clustering techniques in the evaluation of industrial performances.

Other authors consider general macroeconomic data for time series clustering. For instance Crone (1999) uses macroeconomic data to construct an index for the US continental contiguous states. The basic data could be considered as features characterizing the economy of each state. The index has been then used to cluster states into economic regions.

Similarly to Corduas and Piccolo (2008), Maharaj (2000) considers the clustering of time series using an agglomerative model based approach which includes a hypothesis testing procedure. The paper includes an empirical examples on Australian dwelling units financed by all lenders.

Few papers consider clustering exercises with a focus on seasonal patterns. Among these we cite Kumar et al. (2002) a study that propose a distance measures for series characterised by the presence of a seasonal pattern. They applied their methodology to the clustering of retail data classified by departments (shoes, shirts, jewellery...) and classes (men's winter shoes, formal shirts...). They also include a comparison of their distance with traditional ones showing its superiority.

Some authors consider different type of data. Di Matteo et al. (2004) perform a cluster analysis of US interest rates by using a correlation based distance and an ultra-metric distance for hierarchical clustering. Their results points at the presence of clusters characterised by interest rates with close maturities (a somewhat expected outcome). Alonso et al. (2006) consider an application to CO2 emissions using hierarchical approaches based on an L^2 norm between forecast densities. Finally, Xiong and Yeung (2004) propose and apply a model based clustering approach on Personal Income and Population data of US states. They include a comparison of their outcome with those of

Kalpakis et al. (2001), that used the same dataset with more traditional methods. Similar data are also used in Zhang et al. (2006).

4.2 In finance

Financial applications of time series clustering are much more diffused in the scientific literature than economic examples. The main reason is given by the possible uses of clusters in the exploratory analysis of financial market data. Despite these applications are not the primary concern for Eurostat, we believe they represent interesting examples on the possible uses of time series clustering. In addition, the methodology here employed could be applied and generalized for the use with macroeconomic data, structural indicators, and short term business statistics. Some possible extensions of the methods to areas of interest of Eurostat are mentioned in the following.

Panton et al. (1976), probably the first paper on clustering financial time series. They provide an application based on correlation distances across equity market returns. In the same field, the exploratory analysis of financial data, we include the Mantegna (1999) study the clustering of Standard and Poor's 500 and Dow Jones Industrial Average components using a distance based on the correlation between asset returns. The results included in this paper show evidence of discrepancies between the minimum spanning tree derived from the clustering procedure and the classification of companies using economic sectors and subsectors. Notably, in the hierarchical clustering procedure the author uses the subdominant ultrametric distance presented in Appendix A.2. Techniques similar to those of Mantegna could be employed for the comparison of an a-priori classification, for instance the NACE, and the hierarchical structure creating by a clustering exercise on economic time series. Approaches similar to that of Mantegna (1999) are those in Bonanno et al. (2000), that study the association between financial markets by means of a clustering of stock market indices. This paper considers minimum spanning trees based on a correlation-based distance and determine a hierarchical structure across a set of 51 world equity indices extracted from the Morgan Stanley Capital International database. Examples based on few time series could be easily replicated in the economic context by clustering the same time series (say GDP) across the EU member states.

An interesting example, for the methods used in comparing the clusters created by modifying the features provided to the grouping algorithm is the one in Gavrilov et al. (2000). The authors consider the clustering of S&P 500 components by mean of hierarchical approaches, they report the results for different features, and compare the outcomes.

Few contributions considered the clustering of interest rates and bond indices time series. Bernaschi et al. (2002) classify US bond using a correlation based metric and show that clusters are given by bond with close maturities. Similar, and expected, results are provided by Di Matteo and Aste (2002), that classify Eurodollar forward rates using correlation based distances and ultrametric distances for hierarchical linkage of clusters.

Financial applications also include model based clustering. Otranto (2004), Otranto (2008), and Otranto and Trudda (2008a,b) consider the classification of financial data using the metric proposed by Piccolo (1990) applied to GARCH parameters. Their results highlight that clustering approaches could be used for grouping time series characterised by similar variance dynamic patterns. The papers contain also some results associated with the comparison through statistical tests of the distances between assets and provide a specific agglomerative algorithm. The ideas included in the previous papers could be extended and applied to economic data with moderately high frequency (weekly) or with economic series strongly related to the financial markets (interest rates) that could present heteroskedasticity.

Some authors consider financial applications for testing new distances or clustering methods. For instance, Bonanno et al. (2003) consider a minimum spanning tree derived from a clustering algorithm using the correlation distance defined in equation (A.15). In their analysis they focus on a set of more than one thousand stocks traded in the New York Stock Exchange and compare the minimum spanning tree clustering with the classification of companies based on the economic sector. They show evidence of the presence of a hierarchical structure not completely associated to the economic sectors. Similar approaches could be followed, for instance, in the clustering of sections of the Eurostat database containing a large number of series. This could be the case of the Labour Market data, or of the External Trade domain.

Micciché et al. (2003, 2004) focus on the clustering of highly capitalised stocks traded at the New York Stock Exchange by using a distance metric based on the correlations. This study has some relevant differences with respect to similar contributions. In fact, it uses the Spearman rank correlation instead of the standard Pearson correlation, it verifies the stability over time of the clusters composition, and it also focus both on asset returns and on asset volatilities identified from a proxy based on the range. The reported results point out the lower stability of clusters based on the volatility compared to those derived from returns. This study could be of interest in economic applications where series included in a domain are characterized by extremely different levels of volatility.

Few examples consider clustering of managed financial products. Pattarin et al. (2004) present a GAME clustering for Italian mutual funds for the purposes of comparing an a-priori declared investment style with a data-driven grouping. The results reported show evidence of a relevant agreement between the two classifications with some mutual funds sensibly deviating. A similar approach could be applied in the clustering of National Accounts of different countries. In fact, National Accounts as a whole for a single country could be considered as a portfolio of assets, and the GDP may represent the total portfolio value.

In a framework similar to that of Pattarin et al. (2004), we mention Dose and Cincotti (2005), and Lisi and Corazza (2008). In particular, Dose and Cincotti (2005) introduce the clustering as a tool for passive portfolio management in the identification of index tracking stocks. Their purpose is to cluster assets in groups which are closely reproducing (or not reproducing) a reference index. In their analysis they consider both the correlation based distance as well as the distance based on the percentage difference across prices. They show, by mean of a complete linkage agglomerative approach, how the clustering allows a reduction in the tracking errors of passive management. This study could be of interest in business cycle applications where the business cycle itself could represent the reference index, and the purpose could be the identification of coincident, leading and lagging indicators.

Basalto et al. (2006) use Hausdorff clustering for the stocks included in the Dow Jones Industrial Average index. As proximity measure they adopt the distance based on the linear correlation coefficient described in (A.15) and computed over the log-returns of

the assets. Their analysis, despite based on few variables (the index contains 30 variables), contains some interesting results related to the cluster composition and their evolution over time. The most interesting aspect of this study is given by the attempt of checking the evolution over time of the clusters and of their components. Clearly, such a purpose could be of interest in economic applications where subject under study are member states and the interest could be in the identification of the stability of groups containing, for instance, developed markets and transition markets. The movement from transition to developed group may have relevant impacts on policy choices.

Caiado and Crato (2007) used a model based clustering with distance measures defined over the periodogram and considering a set of stock market indices. Their approach is based on the estimate of GARCH models on series of different lengths. Notably, the authors point out how the distances based on the correlation completely fail in capturing the dynamics of the time series and is not suitable for series of different lengths. This last result deserves particular attention when considering business cycle applications of clustering, where the dynamic of series could play a relevant role.

5. Conclusions

This document summarizes some of the elements characterizing the clustering of time series. Beside the theoretical construction of clustering exercises, we devoted the largest part of our report to the possible definition of clustering inputs and to the examples already included in the statistical literature. In addition, we make a clear distinction between supervised classification, which has been discussed in Section 2, and unsupervised classification, or clustering, which was presented in Section 3.

References

1. Agrawal, R., Faloutsos, C., and Swami, A., 1993, Efficient similarity search in sequence databases, Research report, IBM Almaden Research Center, California.
2. Alonso, A.M., Barrendero, J.R., Hernadez, A., and Justel, A., 2006, Time series clustering based on forecast densities, Computational Statistics and Data Analysis, 51, 762-776.
3. Alonso, A.M., Casado, D., Pintado, S.L., and Ramo, J., 2008, A functional data based method for time series classification, Working paper 08-74, Statistics and Econometrics Series, 27, Universidad Carlos III de Madrid.
4. Baragona, R., 2000, Genetic algorithms and cross-correlation clustering of time series, working paper University of Rome "La Sapienza".
5. Basalto, N., Bellotti, R., De Carlo, F., Facchi, P., Pantaleo, E., and Pascazio, S., Hausdorff clustering of financial time series, Physica A, 379, 635-644.
6. Bauwens, L., and Rombouts, J.V.K., 2003, Bayesian clustering of many GARCH models, preprint.
7. Beran, J., and Mazzola, G., 1999, Visualizing the relationship between two time series by hierarchical smoothing models, Journal of Computational and Graphical Statistics, 8 (2), 213-238.
8. Berkhin, P., 2006, A survey of clustering data mining techniques, in Kogan, J., Nicholas, C., and Teboulle, M., (Eds.), Grouping multidimensional data: recent advances in clustering, Springer Berlin Heidelberg New York.
9. Bernaschi, M., Grilli, L., and Vergni, D., 2002, Statistical analysis of fixed income market, Physica A, 308, 381-390
10. Bezdek, J.C., 1987, pattern recognition with fuzzy objective function algorithms, Plenum Press, New York and London.
11. Bonanno, G., Vandewalle, N., and Mantegna, R.N., 2000, Taxonomy of stock market indices, Physical Review E, 62, 6, R7615-8.
12. Bonanno, G., Caldarelli, G., Lillo, F., and Mantegna, R.N., 2003, Topology of correlation-based minimal spanning trees in real and model markets, Physical Review E, 68, 046130-1/4.

13. Box, G.E.P., and Jenkins, G.M, 1979, *Time Series Analysis: Forecasting and Control*, Holden-Day
14. Caiado, J. and Crato, N., 2007, A GARCH-based method for clustering of financial time series: international stock market evidence, http://mpra.ub.uni-muenchen.de/2074/.
15. Caiado, J., Crato, N., and Pena, D., 2006, A periodogram-based metric for time series classification, Computational Statistics and Data Analysis, 50, 2668-2684.
16. Caiado, J., Crato, N., and Pena, D., 2007, Comparison of time series with unequal length, http://mpra.ub.uni-muenchen.de/6605/.
17. Caiado, J., Crato, N., and Pena, D., 2009, Comparison of time series with unequal length in the frequency domain, http://mpra.ub.uni-muenchen.de/15310/.
18. Chan, K. And Fu, A.W., 1999, Efficient time series matching by wavelets, Proceedings of the 15th International Conference on Data Engineering, 126-133
19. Chen, J.R., 2007, Useful clustering outcomes from meaningful time series clustering, Proceedings of the sixth Australasian conference on Data mining and analytics, 70, 101-109
20. Clements, M. P.; Hendry, D. F. (2001), "Forecasting with difference-stationary and trend-stationary models", *Econometrics Journal*, Vol. 4, Issue 1, pp. 1-19.
21. Corduas, M., and Piccolo, D., 1999, An application of the AR metric to seasonal adjustment, Bulletin of the International Statistical Institute, LVIII, 217-218.
22. Corduas, M., and Piccolo, D., 2008, Time series clustering and classification by the autoregressive metric, Computational Statistics and Data Analysis, 52, 1860-1872.
23. Crone, T., 1999, Using State indexes to define economic regions in the US, Federal Reserve Bank of Philadelphia, Working Paper 99-19.
24. Dagum, E. B. and Cholette, P. A. (2006), *Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series*, New York: Springer Science

25. Di Matteo, T., and Aste, T., 2002, How does the eurodollar interest rate behave?, International Journal of Theoretical and Applied Finance, 5 (1), 107-122.
26. Di Matteo, T., Aste, T., and Mantegna, R.N., 2004, An interest rate cluster analysis, Physica A, 339, 181-188.
27. Dose, C., and Cincotti, S., 2005, Clustering of financial time series with application to index and enhances index tracking portfolio, Physica A, 355, 145-151.
28. Duda , R. , Hart , P. , and Stork , D., 2001,. Pattern classification, 2nd edition, John Wiley and Sons, New York, NY.
29. Dunn, J.C., 1974, A fuzzy relative ISODATA process and its use in detecting compact well-separated clusters, Journal of Cybernetics, 3, 32-57.
30. Estivill-Castro, V. and Murray, A.T., 1998, Spatial clustering for data mining with genetic algorithms, http://citeseer.nj.nec.com/estivill-castro97spatial.html.
31. Everitt, B., Landau, S. and Leese, M., 2001, Cluster analysis 4th edition, Arnold, London.
32. Feller, W., (1968): *An Introduction to Probability Theory and Its Applications*, Volume I. New York: John Wiley
33. Focardi, S.M. and Fabozzi, F.J., 2004, Clustering economic and financial time series: exploring the existence of stable correlation conditions, Finance Letters, 2 (3), 1-9.
34. Franses, P. H., and Paap, R. (2004), *Periodic time series models*, Oxford, U.K.: Oxford University Press
35. Galbraith, J.K. and Lu, J., 2001, Cluster and discriminant analysis on time series as a social science research tool, in Inequality and Industrial Change: A Global View, Glabraith, J.K. and Berner, R. (eds.), Cambridge University Press
36. Gan, G., Ma, C., and Wu, J., 2007, Data clustering: theory, algorithms and applications, ASA-SIAM Series on Statistics and Applied Probability
37. Gavrilov, M., Anguelov, D., Indyk, P., and Motwani, R., 2000, Mining the stock market: which measure is the best? In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Boston, MA, US, 487-496

38. Ghysels, E. (1996), “On the Economics and Econometrics of Seasonality”, in *Advances in Econometrics*, Sixth World Congress, ed. C. A. Sims, Cambridge, U.K.: Cambridge University Press, pp. 257–316.
39. Ghysels, E., and Osborn, E. (2001), *The Econometric Analysis of Seasonal Time Series*, Cambridge, U.K.: Cambridge University Press
40. Ghysels, E. (1988), “A Study Towards a Dynamic Theory of Seasonality for Economic Time Series”, *Journal of the American Statistical Association*, 85, pp. 168–172.
41. Gower. J., 1967, A comparison of some methods of cluster analysis, Biometrics, 23 (4), 623-628.
42. Gnanadesikan, R., et al., 1989, Discriminant Analysis and Clustering: Panel on Discriminant Analysis, Classification, and Clustering, *Statistical Science*, 4, 1, pp. 34-69.
43. Granger, C. W. J., and P. Newbold (1977): *Forecasting Economic Time Series*. New York: Academic Press
44. GRETA, 2009, Technical report on classification techniques for time series data.
45. Hall, L.O., Özyurt, B. and Bezdek, J.C., 1999, Clustering with a genetically optimized approach, IEEE Transactions on Evolutionary Computation, 3 (2), 103–112.
46. Han, J., and Kamber, 2001, M., Data mining: concepts and techniques, Morgan Kaufmann, San Francisco.
47. Harvey, A.C., 1991, Forecasting, structural time series models, and the Kalman filter, Cambridge University Press
48. Harvey, A. and Proietti, T. (2005), *Readings in Unobserved Components Models*, Oxford: Oxford University Press.
49. Inniss, T.R., 2006, Seasonal clustering technique for time series data, European Journal of Operational Research, 175, 376-384
50. Jain, A. And Dubes, R., 1988, Algorithms for clustering data, Prentice Hall, Englewood Cliffs, NJ.
51. Johansen, S. (1995), *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*, Oxford University Press Inc., New York

52. Johnson, S., 1967, Hierarchical clustering schemes, Psychometrika, 32 (3), 241-254.
53. Kakizawa, Y., Shumway, R.H., and Taniguchi, M., 1998, Discriminant and clustering for multivariate time series, Journal of the American Statistical Association, 93 (441), 328-340
54. Kalpakis, K., Gada, D., and Puttagunta,V., 2001, Distance measures for effective clustering of ARIMA time-series, Proceedings of the 2001 IEEE International Conference on Data Mining, 273–280.
55. Kaufmann, L., and Rousseeuw, 1990, Finding groups in data: an introduction to cluster analysis, Wiley, New York.
56. Keogh, E., Lin, J., and Truppel, W., 2003, Clustering of time series subsequences is meaningless: implications for previous and future research, Proceedings of the Third IEEE International Conference on Data Mining, 115-124
57. Kiang, M.Y., 2001, Extending the Kohonen self-organizing map networks for clustering analysis, Computational Statistics and Data Analysis, 38, 161-180
58. Kohonen, T., 1990, The self organizing maps, Proceedings IEEE, 78 (9), 1464-1480
59. Krishna, K., Murty, M.N., 1999, Genetic *k*-means algorithms, IEEE Transactions on Systems, Man and Cybernetics, Part B: Cybernetics, 29 (3), 433–439.
60. Krishnapuram, R., Joshi, A., Nasraoui, O., and Yi, L., Low-complexity fuzzy relational clustering algorithms for web mining, IEEE Transactions of Fuzzy Systems, 9 (4), 565-607.
61. Kumar, M., Patel, N.R., and Woo, J., 2002, Clustering seasonality patterns in the presence of errors, KDD 2002, ACM.
62. Lance, G. And Williams, W., 1967, A general theory of classification sorting strategies: 1. Hierarchical systems, Computer Journal, 9, 373-380.
63. Liao, T.W., 2005, Clustering of time series data – a survey, Pattern Recognition, 38, 1857-1874.
64. Lin, J., Vlachos, M., Keogh, E., and Gunopulos, 2004, Iterative incremental clustering of time series, Lecture Notes in Computer Science, Advances in

Database Technology - EDBT 2004, 9th International Conference on Extending Database Technology, Heraklion, Crete, Greece, 521-530

65. Lisi, F., and Corazza, M., 2008, Clustering financial data for mutual fund management, in Perna, C., and Sibillo, M. (Eds), Mathematical and statistical methods for insurance and finance, Springer, 157-164.

66. Lisi, F., and Otranto, E., 2008, Clustering Mutual Funds by Risk Levels, CRENoS , Working Paper n. 2008/13

67. Maharaj, E.A., 1999, Comparison and classification of stationary multivariate time series, Pattern Recognition, 32, 1129-1138

68. Maharaj, E.A., 2000, Clusters of time series, Journal of Classification, 17, 297-314

69. McLachlan, G.J. and Basford, K.E., 1988, Mixture models: inference and applications to clustering, Marcel Dekker, New York.

70. Mantegna, R.N., 1999, Hierarchical structure in financial markets, The European Physical Journal B, 11, 193-197.

71. McQueen, J., 1967, Some methods for classification and analysis of multivariate observations, in LeCam, L.M., and Neyman, J., (Eds.), Proceedings of the fifth Berkley symposium on mathematical statistics and probability, 1, 281-297.

72. Meng, L., Wu, Q.H., and Yong, Z.Z., 2002, A genetic hard *c*-means clustering algorithm, Dynamics of Continuous, Discrete and Impulsive Systems, Series B: Applications and Algorithms, 9, 421–438.

73. Micciché, S., Bonanno, G., Lillo, F., and Mantegna, R.N., 2003, Degree stability of a minimum spanning tree of price return and volatility, Physica A, 324, 66-73.

74. Micchiché, S., Lillo, F., and Mantegna, R.N., 2004, Correlation based hierarchical clustering in financial time series, proceedings of 31st Workshop of the International School of Solid State Physics, Erice, Sicily, Italy, 20 - 26 July 2004.

75. Milligan, G., and Cooper, M., 1985, An examination of procedures for determining the number of clusters in a data set, Psychometrika, 50, 159-179

76. Mitchell, W. C., and A. F. Burns. [1938], 1961. Statistical indicators of cyclical revivals. Reprinted in *Business cycle indicators*, ed. G. H. Moore. Princeton, N.J.: Princeton University Press.
77. Murtagh, F., 1983, A survey of recent advances in hierarchical clustering algorithms, Computer Journal, 26 (4), 354-359.
78. Nelson C. R., and H. Kang (1981): "Spurious Periodicity in Inappropriately Detrended Time Series", *Econometrica,* 49, 741-751
79. Nelson C. R., and H. Kang (1983): "Pitfalls in the Use of Time as an Explanatory Variable in Regression", *Journal of Business and Economic Statistics*, 2, 73-82.
80. Nelson, C. R. and Plosser, C. R. (1982), Trends and random walks in macroeconomic time series: Some evidence and implications, *Journal of Monetary Economics*, 10, 2, pp. 139-162
81. Oates, T., Firoiu, L., and Cohen, P.R., Using dynamic time warping to bootstrap HMM-based clustering of time series, Lecture Notes in Computer Science, 1828, 35–52.
82. Otranto, E., 2004, Classifying the market volatility with ARMA distance measures, Quaderni di Statistica, 6, 1-19.
83. Otranto, E., 2008, Clustering heteroskedastic time series by model-based procedures, Computational Statistics and Data Analysis, 52, 4685-4698.
84. Otranto, E., and Trudda, A., 2008a, Evaluating the Risk of Pension Funds by Statistical Procedures, in Lakatos, G.M. (Ed.), Transition Economies: 21st Century Issues and Challenges, 189-204, Nova Science Publishers, Hauppauge
85. Otranto, E., and Trudda, A., 2008b, Classifying the Italian pension funds via GARCH distance, in Perna, C., and Sibillo, M., (Eds.), Mathematical and Statistical Methods for Insurance and Finance, Springer, 189-197.
86. Panton, D.B., Lessig, V.P., and Joy, O.M., 1976, Comovement of international equity markets: a taxonomic approach, Journal of Financial and Quantitative Analysis, September, 415-432.
87. Pattarin, F., Paterlini, S., and Minerva, T., 2004, Clustering financial time series: an application to mutual fund style analysis, Computational Statistics and Data Analysis, 47, 353-372.

88. Persons, W.M. (1919): “Indices of business conditions”, *Review of Economic Statistics*, 1, 5-107.
89. Phillips, P. C. B. (1986): "Understanding Spurious Regressions in Econometrics," *Journal of Econometrics*, 33, 311-340
90. Phillips, P. C. B. (1987a): "Time Series Regression with a Unit Root," *Econometrica*, 55, 277-301.
91. Phillips, P. C. B. (1987b): "Asymptotic Expansions in Nonstationary Vector Autoregressions," *Econometric Theory*, 3, 45-68.
92. Phillips, P. C. B., and S. N. Durlauf (1986): "Multiple Time Series Regression with Integrated Processes," *Review of Economic Studies*, 53, 473-495
93. Piccolo, D., 1990, A distance measure for classifying ARMA models, Journal of Time Series Analysis, 11, 153-164.
94. Povinelli, R.J. and Feng, X., 1998, Temporal pattern identification of time series data using pattern wavelets and genetic algorithms, Artificial Neural Networks in Engineering, Proceedings, 691-696.
95. Rohlf, F., and Fisher, D., 1968, Tests for the hierarchical structure in random data sets, Systematic Zoology, 17, 407-412
96. Sneath, P., 1957, The application of computers to taxonomy, Journal of General Microbiology, 17, 201-226.
97. Sokal, R. And Michener, C., 1958, A statistical method for evaluating systematic relationships, University of Kansas Science Bulletin, 38, 1409-1438.
98. Sorensen, T., 1948, A method of establishing groups of equal amplitude in plan sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons, Biologiske Skrifter, 5, 1-34.
99. Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, New York: Springer-Verlag.
100. Vilar, J.A., Alonso, A.M., and Vilar, J.M., 2009, Non-linear time series clustering based on non-parametric forecast densities, Computational Statistics and Data Analysis, forthcoming.
101. Vilar, J.M., Vilar, J.A., and Pértega, S., 2009, Classifying time series data: a nonparametric approach, Journal of Classification, 26, 3-28

102. Ward, J., 1963, Hierarchical groupings to optimize an objective function, Journal of the American Statistical Association, 58, 236-244.

103. Wang, C., and Wang, X.S., 2000, Supporting content based searches on time series via approximation, preprint.

104. Wang, X., Smith, K.A., Hyndman, R., and Alahakoon, D., 2004, A scalable method for time series clustering, Technical Report, Department of Econometrics and Business Statistics, Monash University, Victoria, Australia

105. *Wold, H. O. A. (1938), A study in the analysis of stationary time series*, Almqvist & Wiksells (Uppsala)

106. Xiong, Y. and Yeung, D., 2004. Time series clustering with ARMA mixtures. Pattern Recognition 37, 1675–1689.

107. Xu, R., and Wunsch, D., 2005, Survey of clustering algorithms, IEEE Transactions on neural networks, 16 (3), 645-678.

108. Xu, R., and Wunsch, D., 2009, Clustering, IEEE Press Series on Computational Intelligence, Wiley.

109. Zhang, H., Ho, T.B., Zhang, Y., and Lin, M., 2006, Unsupervised feature extraction for time series clustering using orthogonal wavelet transform, Informatica, 30, 305-319

Table 1: Lance and William's parameter combination for agglomerative hierarchical clustering algorithms

Clustering algorithm	α_i	α_j	β	δ	Distance $d\left(y_{(ij)}, y_l\right)$
Single linkage	½	½	0	-½	$\min\left(d\left(y_i, y_l\right), d\left(y_j, y_l\right)\right)$
Complete linkage	½	½	0	½	$\max\left(d\left(y_i, y_l\right), d\left(y_j, y_l\right)\right)$
Group average linkage	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	0	0	$\frac{1}{2}\left(d\left(y_i, y_l\right) + d\left(y_j, y_l\right)\right)$
Weighted average linkage	½	½	0	0	$\frac{n_i}{n_i + n_j} d\left(y_i, y_l\right) + \frac{n_j}{n_i + n_j} d\left(y_j, y_l\right)$
Median linkage	½	½	-¼	0	$\frac{1}{2} d\left(y_i, y_l\right) + \frac{1}{2} d\left(y_j, y_l\right) - \frac{1}{4} d\left(y_j, y_i\right)$
Centroid linkage	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	$-\frac{n_i n_j}{\left(n_i + n_j\right)^2}$	0	$\frac{n_i}{n_i + n_j} d\left(y_i, y_l\right) + \frac{n_j}{n_i + n_j} d\left(y_j, y_l\right) - \frac{n_j n_j}{\left(n_i + n_j\right)^2} d\left(y_j, y_j\right)$
Ward's methods	$\frac{n_i + n_l}{n_i + n_j + n_l}$	$\frac{n_j + n_l}{n_i + n_j + n_l}$	$-\frac{n_l}{n_i + n_j + n_l}$	0	$\frac{n_i + n_l}{n_i + n_j + n_l} d\left(y_i, y_l\right) + \frac{n_j + n_l}{n_i + n_j + n_l} d\left(y_j, y_l\right) - \frac{n_l}{n_i + n_j + n_l} d\left(y_j, y_j\right)$

Appendices

A.1. Notation

The object of this study is given by a set of *M* time series $x_{i,t}$, $t = 1, 2, ...T$, $i = 1, 2, ...M$ which may not be always available over the *T* points of the sample. We presume that the time index is defined over months and that the time series could be observed with a monthly, quarterly or annual frequency. Series with a lower order frequency with respect to the month, will be associated to the monthly observation at the end of the measurement period and will have, by construction, a set of missing values. In addition, series could be characterised by further missing values not associated to the sampling/measurement frequency but rather to the complete unavailability of the data for part of the sample. We will make clear in the following when we refer to series measured at the monthly or at a different frequency.

For each time series we will define an *H*-dimensional vector y_i that contain the inputs of the clustering. This vector could include some relevant features of the series $x_{i,t}$. The collection of all the y_i vectors define the pattern matrix $\boldsymbol{y}$ which has dimension *HxM*.

The possible content of $\boldsymbol{y}$ will be described in an appropriate section.

In principle, supervised classification systems could be imagined as systems based on a $\boldsymbol{y}$ matrix composed by a single feature assuming values on a known and fixed scale which allows a direct creation of cluster. Differently, unsupervised classification systems use a pattern matrix containing at least one feature for which is sensible to compute ratios of observations between different series. Unsupervised classification systems derive a set of groups performing iterative steps or estimations on the $\boldsymbol{y}$ matrix content.

A.2. A collection of distance measures

This section contains a list of the most common distance measures used in time series classification, together with distance measures that could provide relevant insight for time series clustering. We assume that the main interest is the clustering or classification of M time series, each of them associated to a vector of features y_j, $j = 1, 2, ...M$ with dimension H. The matrix containing the proximity measure computed over all pairs of clusters (where clusters may be composed even by a single time series) is defined as the proximity matrix and we will denote it as $\boldsymbol{\mathcal{D}}$ if it is created using a dissimilarity measure, $\boldsymbol{\mathcal{S}}$ if a similarity index is used, or $\boldsymbol{\mathcal{P}}$ if we refer to a general proximity matrix ($\boldsymbol{\mathcal{P}}$ will be used for properties or results valid both for similarity and dissimilarity measures).

We also observe that the $\boldsymbol{\mathcal{P}}$ matrix has dimension *MxM* while the pattern matrix $\boldsymbol{\mathcal{Y}}$ (the matrix collecting all feature vectors) has dimension *HxM*, where *H* is the number of features included in the vectors y_j.

In general terms, distance measures are defined in order to satisfy a minimal number of properties. Let us denote by $d(y_i, y_j)$ a generic distance (dissimilarity) between the feature vectors y_j and y_i. The function $d(y_i, y_j)$ is a distance metric if it satisfy the following properties:

1 Symmetry: $d(y_i, y_j) = d(y_j, y_i)$;

2 Positivity: $d(y_i, y_j) \geq 0$ for $i, j = 1, 2, ...M$;

3 Triangle inequality: $d(y_i, y_l) + d(y_l, y_j) \geq d(y_i, y_j)$;

4 Reflexivity: $d(y_i, y_j) = 0$ if and only if $y_i = y_j$.

Conditions 1) and 2) are required to define $d(y_i, y_j)$ as a distance while 3) and 4) are needed for $d(y_i, y_j)$ to become also a metric.

Differently, if we denote by $s(y_i, y_j)$ the similarity between y_j and y_i, the function $s(y_i, y_j)$ is a similarity metric if the following properties are satisfied:

1 Symmetry: $s(y_i, y_j) = s(y_j, y_i)$;

2 Positivity: $s(y_i, y_j) \geq 0$ for $i, j = 1, 2, ...M$;

3 Triangle inequality: $s(y_i, y_l)s(y_l, y_j) \leq \left[s(y_i, y_l) + s(y_l, y_j) \right] s(y_i, y_j)$;

4 Reflexivity: $s(y_i, y_j) = 0$ if and only if $y_i = y_j$.

As for the dissimilarity or distance functions properties 1) and 2) are required for $s(y_i, y_j)$ to be a distance while 3) and 4) to label it as a similarity metric.
We now review the most common proximity measures used in the clustering literature.

A.2.1 The Euclidean, Minkowsky, Manhattan and Sup distances

Probably, these are the most commonly used and known distance metrics. Given two feature vectors y_j and y_i, the Euclidean distance (or L_2 norm) is defined as:

$$d(y_j, y_i) = \sqrt{\sum_{l=1}^{H} (y_{i,l} - y_{j,l})^2} \qquad \text{(A.1)}$$

As noted by Duda et al. (2001) and Xu and Wunsch (2009), the Euclidean distance has the relevant feature of providing clusters invariant to rotations or translations in the spaced spanned by the pattern matrix Y. However, it has also some drawbacks: it is highly influenced by features that are dominant in absolute values (relatively to other features); linear or other transformations of the pattern matrix may manipulate the ordering relations between objects. To solve these aspects, features are generally standardised with respect to the rows moments of the pattern matrix Y: the elements of

the H-dimensional feature vectors y_j, $j = 1,2,...M$ are replaced by standardised values as follow

$$\tilde{y}_{l,j} = \frac{y_{l,j} - \mu_l}{\sigma_l},\ i = 1,2,...H,\ j = 1,2,...M \tag{A.2}$$

where

$$\mu_l = \frac{1}{M}\sum_{j=1}^{M} y_{l,j} \tag{A.3}$$

and

$$\sigma_l = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}\left(y_{l,j} - \mu_l\right)^2} \tag{A.4}$$

Differently, the features may be standardised with respect to their range as

$$\tilde{y}_{l,j} = \frac{y_{l,j} - \min\left(y_{l,j}\right)}{\max\left(y_{l,j}\right) - \min\left(y_{l,j}\right)},\ l = 1,2,...H,\ j = 1,2,...M \tag{A.5}$$

Note that the minimum and maximum quantities are determined over the rows of the pattern matrix $\boldsymbol{y}$. These quantities, as well as the mean μ_l and standard deviations σ_l are feature-specific.

The Euclidean distance may be generalised unrestricting the powers in equation (1) obtaining the Minkowsky distance (or L_p norm).

$$d\left(y_j, y_i\right) = \left[\sum_{l=1}^{H}\left|y_{i,l} - y_{j,l}\right|^p\right]^{\frac{1}{p}} \tag{A.6}$$

Clearly, the Euclidean distance is obtained by setting p=2, whether two other popular distances are associated to the limiting cases p=1 and p=∞. When the parameter is equal to 1, the Minkowsky distance collapses onto the Manhattan distance or L_1 norm

$$d\left(y_j, y_i\right) = \sum_{l=1}^{H} \left| y_{i,l} - y_{j,l} \right| \tag{A.7}$$

while when p goes to infinity, we obtain the Sup distance of L_∞ norm

$$d\left(y_j, y_i\right) = \max_{1 \le l \le H} \left| y_{i,l} - y_{j,l} \right| \tag{A.8}$$

A.2.2 The Mahalanobis distance

This distance is defined as

$$d\left(y_j, y_i\right) = \left(y_j - y_i\right) S^{-1} \left(y_j - y_i\right) \tag{A.9}$$

where S is the covariance matrix across the features

$$S = \frac{1}{M} \sum_{i=1}^{M} \left(y_i - \mu\right)\left(y_i - \mu\right)' \qquad \mu = \frac{1}{M} \sum_{i=1}^{M} y_i \tag{A.10}$$

The Mahalanobis distance includes as special case the squared Euclidean distance, which is obtained when the features are not correlated. One drawback of this measure is its possible computational complexity for large values of *H*.

A.2.3 The Point Symmetry distance

Under the assumption of symmetry for the clusters, a different distance could be used.

$$d\left(y_i, y_l\right) = \min_{j=1,2...M,\ \text{and}\ j\neq i} \frac{\sqrt{\sum_{k=1}^{H}\left(\left(y_{k,i} - y_{k,l}\right) - \left(y_{k,j} - y_{k,l}\right)\right)^2}}{\sqrt{\sum_{k=1}^{H}\left(y_{k,i} - y_{k,l}\right)^2} + \sqrt{\sum_{k=1}^{H}\left(y_{k,j} - y_{k,l}\right)^2}} \qquad \text{(A.11)}$$

Where y_l is a reference point such as a cluster centroid. This distance measure evaluates the distance of objects with respect to a given point given the other *M-1* objects.

A.2.4 Measures based on correlations

An alternative approach points at defining distance measures based on the Pearson correlation coefficient. Let us define the correlation coefficient as

$$\rho\left(y_i, y_j\right) = \frac{\sum_{l=1}^{H}\left(y_{i,l} - \overline{y}_i\right)\left(y_{j,l} - \overline{y}_j\right)}{\sqrt{\sum_{l=1}^{H}\left(y_{i,l} - \overline{y}_i\right)^2 \sum_{l=1}^{H}\left(y_{j,l} - \overline{y}_j\right)^2}} \qquad \text{(A.12)}$$

where

$$\overline{y}_i = \frac{1}{H}\sum_{l=1}^{H} y_{i,l} \ \text{ and } \ \overline{y}_j = \frac{1}{H}\sum_{l=1}^{H} y_{j,l}\ . \qquad \text{(A.13)}$$

Then a distance measure could be defined as

$$d\left(y_i, y_j\right) = \frac{1 - \rho\left(y_i, y_j\right)}{2} \qquad \text{(A.14)}$$

or

$$d\left(y_i, y_j\right) = \sqrt{2\left(1-\rho\left(y_i, y_j\right)\right)}$$

(A.15)

or

$$d\left(y_i, y_j\right) = \left(\frac{1-\rho\left(y_i, y_j\right)}{1+\rho\left(y_i, y_j\right)}\right)^{\beta} \tag{A.16}$$

where β>0.

An alternative approach uses the cross-correlations between two time series. Such an approach assumes that the features vectors y_i and y_j contain the entire time series (which we assume to be of the same length). If we define the cross correlation at lag k as

$$\rho_k\left(y_i, y_j\right) = \frac{\sum_{t=k}^{T}\left(y_{i,t}-\overline{y}_i\right)\left(y_{j,t-k}-\overline{y}_j\right)}{\sqrt{\sum_{t=k}^{T}\left(y_{i,t}-\overline{y}_i\right)^2\sum_{l=1}^{T}\left(y_{j,t-k}-\overline{y}_j\right)^2}} \tag{A.17}$$

where $\overline{y}_j$ and $\overline{y}_i$ are again the sample means, a distance based on cross-correlations may be

$$d\left(y_i, y_j\right) = \sqrt{\frac{1-\rho_0^2\left(y_i, y_j\right)}{\sum_{k=1}^{K}\rho_k^2\left(y_i, y_j\right)}} \tag{A.18}$$

see Liao (2005). The statistical literature contains other dissimilarity measures based on the cross-correlations, such as:

$$d\left(y_i, y_j\right) = \min_{-m\le k\le m}\left(1-\left|\rho_k\left(y_i, y_j\right)\right|\right), \tag{A.19}$$

where *m* is the maximum lead/lag;

$$d\left(y_i, y_j\right) = \min_{-m \le k \le m} w(k)\left(1 - \left|\rho_k\left(y_i, y_j\right)\right|\right), \tag{A.20}$$

where $w(k)$ is an appropriate weighting function;

$$d\left(y_i, y_j\right) = \sum_{k=0}^{m} w(k) \frac{\left|\rho_k\left(y_i, y_j\right) - \rho_{-k}\left(y_i, y_j\right)\right|}{1 + \left|\rho_k\left(y_i, y_j\right) - \rho_{-k}\left(y_i, y_j\right)\right|}; \tag{A.21}$$

$$d\left(y_i, y_j\right) = \sum_{k=0}^{m} w(k) \frac{\left|\rho_k\left(y_i, y_i\right)\rho_k\left(y_j, y_j\right) - \rho_k\left(y_i, y_j\right)\rho_{-k}\left(y_i, y_j\right)\right|}{1 + \left|\rho_k\left(y_i, y_i\right)\rho_k\left(y_j, y_j\right) - \rho_k\left(y_i, y_j\right)\rho_{-k}\left(y_i, y_j\right)\right|}, \tag{A.22}$$

where $\rho_k\left(y_i, y_i\right)$ denotes the autocorrelation function, see Baragona (2000) and therein cited references for additional details.

A.2.5 Distance measures for discrete variables

The previous distances are designed for continuous random variables. However, the statistical literature also includes distances which are more appropriate for variables or features assuming values in a discrete range such as binary variables or limited dependent variables.

Given the purposes of this survey, the review of clustering methods and approaches for economics and financial time series, we do not present here the distance measures for discrete variables. The readers interested in this topic could find some information in Xu and Wunsch (2009), sections 2.5 and 2.6.

A.2.6 Dynamic time warping

The Dynamic Time Warping distance is based on a generalization of the traditional algorithms for comparing two sequences one characterized by discrete observations and the other by continuous values.

In general terms, given two feature vectors y_i and y_j formed by the time series observations (the two vectors in this case may have different length, call them T_1 and T_2), the Dynamic Time Warping search for an optimal alignment of the series such that the distance between observations is minimized.

To determine the optimal Dynamic Time Warping path (the set of paired observations of the series), at first, we must determine the proximity matrix P whose entries are given by the Euclidean distance for each pair of observations in the feature vectors y_i and y_j

$$d\left(y_{j,t}, y_{i,m}\right) = \sqrt{\left(y_{i,t} - y_{j,m}\right)^2} \text{ for } t = 1, 2, ...T_1 \text{ and } m = 1, 2, ...T_2 \quad \text{(A.23)}$$

Given this matrix, Dynamic Time Warping finds a path satisfying the following restrictions:

1. Boundary condition: the path starts at the top left corner of the matrix (where the time index of both series is equal to 1, $d\left(y_{j,1}, y_{i,1}\right)$) and ends and the bottom right corner of the matrix (where the time index of both series is at the final point in the corresponding samples $d\left(y_{j,T_1}, y_{i,T_2}\right)$);
2. Continuity condition: the path considers moves over adjacent cells of the matrix P (from $d\left(y_{j,t}, y_{i,m}\right)$ we could move to one of the following cells $d\left(y_{j,t+1}, y_{i,m}\right)$, $d\left(y_{j,t}, y_{i,m+1}\right)$ or $d\left(y_{j,t+1}, y_{i,m+1}\right)$);
3. Monotonicity condition: the path should move over cells in the matrix P which are monotonically spaced over time (that is the row index, or the column index or both indices of two cells representing two consecutive points in the path must always increase by 1).

The Dynamic Time Warping path is the one that minimize the distance between series and the Dynamic Time Warping distance is defined as

$$d\left(y_j, y_i\right) = \min_A \frac{\sum_{(t,m)\in A} d\left(y_{j,t}, y_{i,m}\right)}{card\left(A\right)} \tag{A.24}$$

where A is a set of row and column indices (and by construction of time indices) satisfying conditions 1 to 3 above, card(A) is the cardinality of the set A (the number of elements). The Dynamic Time Warping distance thus search for the path associated to the minimum distance between the series observations.
Note that the optimal solution could be found by following this recursion:

a. Initialize the distance as $d_{1,1}\left(y_j, y_i\right) = d\left(y_{j,1}, y_{i,1}\right)$
b. Update the distance as
$$d_{t,m}\left(y_j, y_i\right) = d\left(y_{j,t}, y_{i,m}\right) + \min\left\{d_{t-1,m}\left(y_j, y_i\right), d_{t,m-1}\left(y_j, y_i\right), d_{t-1,m-1}\left(y_j, y_i\right)\right\}$$
c. Stop at $d_{T_1,T_2}\left(y_j, y_i\right) = d\left(y_j, y_i\right)$

A.2.7 Distances based on estimated features

In some cases the feature vector could be composed by estimated quantities, such as sample moments or estimated coefficients. In such a case, the quantities included in the feature vectors could be associated to a multivariate density function with mean equal to the unknown population value of the estimated features and variance equal to the population variance of the features (P denotes population values):

$$y_j \sim D\left(y_j^P, \Sigma\left(y_j^P\right)\right). \tag{A.25}$$

In a generalized approach, we may assume that the feature vector is associated to an estimate of the covariance matrix between the components of the vector (an estimate of $\Sigma\left(y_j^p\right)$). In this situation, we could design a distance measure taking into account both the feature vector and the covariance between its constituents, defined as

$$d\left(y_i, y_j\right) = \left(y_i - y_j\right)' \left(\Sigma\left(y_i\right) + \Sigma\left(y_j\right)\right)^{-1} \left(y_i - y_j\right) \tag{A.26}$$

Such a measure have been proposed by Caiado, Crato and Pena (2007) and used in Caiado and Crato (2007). Note that this distance is similar to the Mahalanobis distance but is using a different weighting matrix. It is also similar to the test statistics for Hausman-type tests (reference to be included).

A.2.8. Distances based on transformations of the time series

Alonso et al. (2008) introduced a distance for the evaluation of the discrepance between the periodograms of two different time series. We generalize their distance to the case where the feature vector includes the estimates of a monotonic function based on the underlying time series. The distance could be defined as

$$d\left(y_i, y_j\right) = \int_a^b \left(F_i\left(w\right) - F_j\left(w\right)\right) dw \tag{A.27}$$

where $F_i\left(w\right)$ is an monotonic function in $\left[a,b\right]$. Following Alonso et al. (2008), this function could be the integrated periodogram, or, generalising their approach, the cumulative density function.

A similar approach has been presented in Alonso et al. (2006) and Vilar et al. (2009). Given two time series *i* and *j* available up to time *T*, and their forecast density for time *T+h*, they suggest the use of an L-norm distance across densities

$$d\left(y_i, y_j\right) = \int \left| f_{i,T+h}\left(w\right) - f_{j,T+h}\left(w\right)\right|^a dw \tag{A.28}$$

where $f_{i,T+h}(w)$ is the forecast density for series *i*, *a* is the order of the norm which was set to 2 by Alonso et al. (2006) and to 1 in Vilar et al. (2009). The authors provide also a proof of the consistency of the sample estimator of the distance and note that the L-norms offer computational advantages with respect to alternative measures that could be used to compute the distance between densities. They also highlight that their approach requires stationarity of the underlying time series. We note that the forecast density could be derived using different methods, including the bootstrap approach proposed by Alonso et al. (2006).

A.2.9. Distances based on time series paths and quantities

Kumar et al. (2002) propose a variation of the Euclidean distance to cluster the seasonality patterns of time series taking also into account the different dispersion of the time series around the seasonal component. They propose the following distance

$$d\left(y_i=\left\{x_{i,t},s_{i,t}\right\}_{t=1}^{T},y_j=\left\{x_{j,t},s_{j,t}\right\}_{t=1}^{T}\right)=\sum_{t=1}^{T}\frac{\left(s_{i,t}-s_{j,t}\right)^2}{\left(x_{j,t}-s_{j,t}\right)^2+\left(x_{i,t}-s_{i,t}\right)^2} \quad \text{(A.29)}$$

where the feature vectors contain the original time series ($x_{j,t}$ and $x_{i,t}$) and an estimate of the corresponding seasonal patterns ($s_{j,t}$ and $s_{i,t}$). The distance is then composed as a ratio between the Euclidean distance among the seasonal patterns standardised by the squared deviation between the original series and the seasonal pattern. As a result, the denominator could be considered as a pooled local estimate of the error variance. Under the assumption of normality of errors, the distance is distributed as a Chi-square density with T-1 degrees of freedom and thus it could also be used to test for the equivalence across series. Kumar et al. (2002) also notes that this distance is scale invariant.

We note that the distance proposed by Kumar et al. (2002) could be further generalized to take into account the possible contemporaneous presence of a trend-cycle and seasonal component. In fact, the seasonal pattern in the previous equation could be

replaced by either the trend-cycle component or by the combination of the seasonal pattern and trend-cycle. Note that the classification based over trend-cycle could be used to extract subsets of series characterized by similar business cycle phases which could be later labeled as coincident, leading or lagging with respect to the true underlying and dated business cycle.

A different distance was proposed by Dose and Cincotti (2005) and based on the percentage difference across series. Their distance is defined as

$$
\begin{aligned}
&d\left(y_i = \left\{x_{i,t}\right\}_{t=1}^{T}, y_j = \left\{x_{j,t}\right\}_{t=1}^{T}\right) = \min\left\{d_1, d_2\right\} \\
&d_1 = \min_{a\in\mathbb{R}}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_{i,t} - a x_{j,t}}{x_{i,t}}\right)^2\right\} \\
&d_2 = \min_{a\in\mathbb{R}}\left\{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{x_{j,t} - a x_{i,t}}{x_{j,t}}\right)^2\right\}
\end{aligned}
\tag{A.30}
$$

Note that the Dose and Cincotti's distance was designed to create clusters of time series with close patterns over their levels, with an emphasis of financial asset prices. We note that the approach could be applied over macroeconomic series levels or over their components such as trend-cycle or seasonal patterns. In this alternative interpretation, the distance could be used to cluster macroeconomic time series characterized by similar seasonal component or similar business cycle phases as in the case of the Kumar et al. (2002) distance.

A.2.10. Probabilistic distance measures

In this set we include a number of distances that share a common feature: they depend on the density function of the feature vector or on a transformation of this density function. Two examples are given in Liao (2005) that mention the Kullback-Liebler divergence and the Chernoff information divergence. A related approach is in Vilar et al. (2009) that consider a divergence based on the periodograms of two time series. A detailed discussion of probability based distances is included in Kakizawa et al. (1998).

A.3. Technical details on partitioning clustering methods

A.3.1 K-means and fuzzy c-means

This methods perform a clustering of M time series by iterating on two main steps: the distribution of the M objects into the K clusters and the update of the cluster centres (or cluster membership).

Given M time series $x_{i,t}$, $t = 1,2,...T$, $i = 1,2,...M$ and the M H-dimensional vectors y_i containing the inputs for the clustering (the possible form of the inputs will be described in a following section), the K-means method starts by fixing the number of clusters, K, to be determined (sections 5 will consider the issue of the appropriate choice of K), by an arbitrary selection of K cluster centres c_j, $j = 1,2,...K$ (the K-dimensional vector of cluster centres is called C). Note that cluster centres have the same dimension of y_i. The second relevant element is the objective function of the method which could be represented as follow:

$$J(U,C) = \sum_{j=1}^{K}\sum_{i=1}^{M} u_{ji} d\left(y_i, c_j\right)^2$$

(A.31)

where $u_{ij} \in \{0,1\}$, $\sum_{j=1}^{K} u_{ji} = 1$ for $i = 1,2,...M$, $U = \left\{u_{ji}, j = 1,2,..K, i = 1,2,...M\right\}$ and $d(a,b)$ is one of the distance measures defined in section 3.4. Note that U is a selection matrix assigning each series to a given cluster. Finally, a small number ε should be chosen in order to define a stopping criteria.

Then, the method proceeds following this iterative scheme:

1) Assign the M series to the K clusters and evaluate the function $J(U,C)$;
2) Minimize the function $J(U,C)$ with respect to U conditionally to a choice for the vector C;

3) Update cluster centres by minimizing $J(U,C)$ with respect to C conditionally to the U vector define at step 2) or by determining cluster centres as

$$c_j = \frac{\sum_{i=1}^{M} u_{ji} y_i}{\sum_{i=1}^{M} u_{ji}}, \quad j = 1,2,...K \tag{A.32}$$

4) Stop if the change in C is smaller than ε otherwise repeat steps 2) and 3).

Step 1) basically initialize the whole procedure while step 2) reassign the time series to clusters in order to minimize the within groups distances. Step 3) recomputed cluster centres once series reallocation has been performed.

Fuzzy c-means are generalizations of the K-means algorithms where the matrix U is allowed to assume values between 0 and 1, see Dunn (1974). In this case U (we define it as membership matrix) must satisfy $0 \le u_{ij} \le 1, \sum_{j=1}^{K} u_{ji} = 1$ for $i = 1,2,...M$, and $0 < \sum_{i=1}^{M} u_{ji} < M$ for $j = 1,2,...K$ while the objective function becomes

$$J(U,C) = \sum_{j=1}^{K} \sum_{i=1}^{M} \left(u_{ji}\right)^{\delta} d\left(y_i, c_j\right)^2 \tag{A.33}$$

with $\delta \ge 1$. In this case the clusters may be identified using the following iterative procedure for given values of m, ε and K.

1) Assign the M series to the K clusters and initialize the matrix U;
2) Evaluate cluster centres using

$$c_j = \frac{\sum_{i=1}^{M} u_{ji} y_i}{\sum_{i=1}^{M} u_{ji}}, \quad j = 1,2,...K \tag{A.34}$$

3) Update the membership matrix as

$$u_{ji} = \left(\frac{1}{d\left(y_i, c_j\right)^2} \right)^{\frac{1}{m-1}} \left[\sum_{j=1}^{K} \left(\frac{1}{d\left(y_i, c_j\right)^2} \right)^{\frac{1}{m-1}} \right]^{-1}, \quad j = 1,2,...K, \quad i = 1,2,...M \quad \text{(A.35)}$$

if $y_i \neq c_j$ otherwise set $u_{ji} = 1$ for $j = i$, and $u_{ji} = 0$ for $j \neq i$;

4) Stop if the Euclidean distance between two subsequent evaluations of the membership matrix differ for less than ε, otherwise repeat steps 2) and 3).

The k-means method could be generalised in many ways by supplying different kind of inputs extrapolated from time series. Some examples are the series components (Trend-Cycle, Seasonal and Irregular), the Fourier transformation, or the Wavelet decomposition. An interesting approach based on the Wavelets is given in Lin et al. (2004).

A.3. 2 Genetic Algorithm for Medoid Evolution (GAME)

In this approach, Genetic Algorithms are used in combination with the definition of cluster centers (the medoids). The GAME approach aims at determining both the number of cluster, their composition and the number of features (within the H supplied) which are relevant in the classification.

Within the GAME clustering, the objective function to be optimized is a mathematical transposition of the classification objective: minimizing the within group variance and maximize the between group variance.

In general, this is equivalent to minimize the within group covariance of the features and maximize the between groups covariance of the features. Recalling that the feature vectors y_i contain each H features, and assuming the existence of G groups, the within groups covariance is

$$W = \sum_{g=1}^{G} \left(\sum_{i=1}^{m_g} \left(y_i - c_g \right) \left(y_i - c_g \right)' \right) \tag{A.36}$$

where c_g is the cluster centroid computed as

$$c_g = \frac{1}{m_g} \sum_{i=1}^{m_g} y_i \tag{A.37}$$

m_g is the dimension of group g and $\sum_{g=1}^{G} m_g = M$.

In a similar way, we define the between group covariance as

$$B = \sum_{g=1}^{G} m_g \left(c - c_g \right) \left(c - c_g \right)' \tag{A.38}$$

where c is the total centroid.

Given that the total covariance matrix is given by the sum of W and B, two ratios could be used as objective functions:

i) the Variance ratio criterion $\frac{tr\left(B\right)/\left(G-1\right)}{tr\left(W\right)/\left(M-G\right)}$, which, by construction, does not take into account the covariance between features. the optimum is associated to a maximized variance ratio;

ii) the Marriott's criterion $G^2 \frac{|B|}{|W|}$ which associates optimal partitions to its minimum value.

A.4. Technical details on agglomerative methods

The method may be represented by the following set of steps

1) Start with M clusters and compute the proximity matrix P whose entries are the distances for all pairs of objects (is an MxM matrix at the initialization);
2) Identify the minimum distance $\min_{1\le i,j\le M,\, i\neq j} d\left(y_i, y_j\right)$ and combine the two objects y_j and y_i into a new cluster $y_{(ij)}$;
3) Update the proximity matrix P computing the distances between y_{ij} and all other clusters;
4) Iterate 2) and 3) until a single cluster remains.

The critical element of agglomerative algorithms is given by step 3), where the new cluster is used to determine distance measures with respect to the other existing clusters. There exist a number of methods for the computation of these distances, and most of them could be represented with a recursive representation due to Lance and Williams (1967). Given two objects y_i and y_j which are merged in the cluster $y_{(ij)}$ and a third object y_l the distance $d\left(y_{(ij)}, y_l\right)$ can be represented as follow:

$$d\left(y_{(ij)}, y_l\right) = \alpha_i d\left(y_i, y_l\right) + \alpha_j d\left(y_j, y_l\right) + \beta d\left(y_j, y_i\right) + \delta \left| d\left(y_j, y_l\right) - d\left(y_i, y_l\right) \right| \qquad \text{(A.39)}$$

where the values of α_i, α_j, β and δ depend on the weighting scheme adopted. The most used parameter designs are reported in Table 1 which also appeared in Murtagh (1983), Jain and Dubes (1988), Everitt et al. (2001), and Xu and Wunsch (2009).

[INSERT HERE TABLE 1]

Table 1 assumes that the objects can be replaced by clusters of dimension n_i, n_j and n_l, respectively. The last column report the form of the composed distance.

The single linkage algorithm proposed by Sneath (1957) and then analysed by Jain and Dubes (1988), Johnson (1967) and Everitt et al. (2001) defines the distance as the smallest one across the elements included in the two combined clusters. For this reason, the method is also called nearest neighbours. As stated by Everitt et al. (2001), the single linkage algorithm works well when clusters are clearly separated one from the other. Single linkage nearest neighbour clustering could be alternatively represented by dendograms or by minimum spanning trees.

The complete linkage approach (Sorensen, 1948, Jain and Dubes, 1988, Everitt et al., 2001, Xu and Wunsch, 2009) use the farthest distance in combining clusters. As a result, it is most suited for the identification of small compact clusters.

The group average linkage (Sokal and Michener, 1958, Jain and Dubes, 1988, Everitt et al., 2001, Xu and Wunsch, 2009), as its name suggests, take the average of distances, while the weighted average linkage (McQuitty, 1966, Jain and Dubes, 1988, Xu and Wunsch, 2009) weights distances with cluster dimensions.

The centroid linkage (Sokal and Michener, 1958, Jain and Dubes, 1988, Everitt et al., 2001, Xu and Wunsch, 2009) provides a combination based on the cluster centres and can be considered a generalized version of the median linkage algorithm (Gower, 1967, Jain and Dubes, 1988, Everitt et al., 2001, Xu and Wunsch, 2009). The last assign equal weight to the combined clusters while the former weight clusters using their dimension.

Finally, the minimum variance method or Ward's method (Ward, 1963, Jain and Dubes, 1988, Everitt et al., 2001, Xu and Wunsch, 2009) implements a minimization of the within-class sum of squared deviations from cluster centres.

A further case not included in Table 1 is given by the Hausdorff clustering (Basalto et al., 2007) which can be considered an intermediate solution between the single linkage and the complete linkage approaches. In fact, given two clusters composed by I and J elements, respectively, and using a notation similar to that in Table 1, single and complete linkage correspond, respectively, to the following rules for determining the distance between the two clusters

$$d\left(y_{(I)}, y_{(J)}\right) = \min_{i,j} d\left(y_i, y_j\right) \quad i = 1,2,...I, \quad j = 1,2,,...J \tag{A.40}$$

$$d\left(y_{(I)}, y_{(J)}\right) = \max_{i,j} d\left(y_i, y_j\right) \quad i = 1, 2, ...I, \quad j = 1, 2, , ...J \,. \tag{A.41}$$

Differently, the Hausdorff clustering uses

$$d\left(y_{(I)}, y_{(J)}\right) = \max\left\{\max_i \min_j d\left(y_i, y_j\right), \max_j \min_i d\left(y_i, y_j\right)\right\} \quad i = 1, 2, ...I, \quad j = 1, 2, , ...J \,. \tag{A.42}$$

In section 3.2.9 we introduced a distance function proposed by Kumar et al. (2002) which was designed for series characterized by seasonal patterns. Kumar et al. (2002) suggest the use of such a distance within hierarchical clustering and also proposed an alternative way of combining series into a cluster. They introduced the following distance

$$\begin{aligned} & d\left(y_{ij} = \left\{s_{ij,t}, e_{ij,t}\right\}_{t=1}^{T}, y_l = \left\{x_{l,t}, e_{l,t}\right\}_{t=1}^{T}\right) \\ & s_{ij,t} = \left(\frac{1}{e_{i,t}^2} + \frac{1}{e_{j,t}^2}\right)^{-1}\left[\frac{s_{i,t}}{e_{i,t}^2} + \frac{s_{j,t}}{e_{j,t}^2}\right] \\ & e_{ij,t} = \left(\frac{1}{e_{i,t}^2} + \frac{1}{e_{j,t}^2}\right)^{-\frac{1}{2}} \end{aligned} \tag{A.43}$$

where $e_{l,t} = x_{l,t} - s_{l,t}$ is the deviation between the observed series and the estimated seasonal pattern, and the cluster composed by at least two elements is characterised by the combination of the seasonal patterns and the combination of the discrepancies of the component series and the corresponding estimated seasonal patterns. The combined seasonality patterns could then be interpreted as average seasonality within each cluster.

Within hierarchical clustering a further rule for determining the distances between objects belonging to different clusters has been defined in Di Matteo and Aste (2002) and called ultrametric distance. The distance is used within an iterative linkage procedure and for each pairs of objects *i* and *j* belonging to clusters *I* and *J*, respectively, is defined as

$$d\left(y_i, y_j\right) = \max\left\{d\left(y_l, y_m\right), l \in I, m \in J\right\} \quad \text{(A.44)}$$

which is the maximum distance between all couples of elements in the two clusters.

A related approach was used by Mantegna (1999), the subdominant ultrametric distance. In this case, the dendogram (or the minimum spanning tree) is used to identify the distance between two objects *i* and *j*. The subdominant ultrametric distance is the maximum distance detected between two objects in the path between *i* and *j*. As an example, assume that moving from *i* to *j* we cross objects *l* and *m* (from *i* we move to *l*, then to *m* and finally to *j*) than, the distance between *i* and *j* is the maximum between the following distances $d\left(y_i, y_l\right)$, $d\left(y_l, y_m\right)$ and $d\left(y_m, y_j\right)$.