Clustering of Oscillating Dynamical Systems from Time Series Data Bases

A. Tsimpiris and D. Kugiumtzis
Department of Mathematical, Physical and Computational Sciences of Engineering,
Aristotle University of Thessaloniki, Greece
e-mail: alkisser@gen.auth.gr

Abstract: The clustering of time series from oscillating dynamical systems requires appropriately selected features. We employed features stemming from methods of linear and nonlinear analysis of time series, such as autocorrelation and Lyapunov exponents, as well as features estimating oscillation characteristics. Optimal feature forward selection and standardization method, under the standard k-means clustering algorithm, were assessed using Monte Carlo simulations on known oscillating deterministic and stochastic systems. The clustering efficiency was measured by the corrected Rand index and the results showed the prevalence of oscillating-related features. The same setting was applied to real-world oscillating time series, i.e. epileptic electroencephalograms and optokinetic signals, giving good discrimination of pathological states.

Keywords: clustering, data base, time series, oscillating system, nonlinear measures.

1. Introduction

Numerous records of oscillating signals are registered in diverse applications ranging from physiology to acoustics and meteorology, where there is often need for classification or grouping of similar records. Thus there is a growing interest in time series clustering, where the data objects are formed from time series measurements [1]. There are three basic classes of time series clustering approaches in terms of the data object types, i.e. the data object is a raw time series (in the time or frequency domain, e.g. see [2, 3]), a set of features extracted from the time series (e.g. the cross-correlation in [4]), or a set of parameters of a model estimated on the time series (e.g. autoregressive moving average models in [5, 6]). Here, we follow the feature-based approach as suggested also in [7]. To the best of our knowledge there has not yet been a systematic investigation and evaluation of time series features for clustering and it seems that in every work the selected features are related to a specific approach, e.g. features extracted from wavelet transform of the time series [8].

Advances in nonlinear dynamics and synchronization offer new alternatives for determining the nature of the underlying system to the time series, as well as methods and measures to identify them from observed time series [9, 10, 11]. Nonlinear measures in conjunction with other statistical measures have evidently given rise to different clustering approaches [12, 13]. For problems involving oscillating time series, characteristic features of the oscillations in the time series (or of a single oscillation segment) have been considered [14, 15].

The present work aims at assessing the discriminating power of features that measure different characteristics of the time series, i.e. simple statistical and linear features, such as skewness and autocorrelation, nonlinear features, such as mutual information and largest Lyapunov exponent, and oscillation-related features, such as oscillation peak and period. These features may have different scale and distribution, so that the clustering efficiency is heavily dependent on the standardization of the features. Therefore, we evaluated the clustering also for a range of well-known standardization techniques and for a new standardization technique that makes use of the arbitrary-to-Gaussian transform. The motivating examples for the presented work are the analysis of electroencephalograms (EEG) in epileptic patients [16, 17] and optokinetic signals (OKN) from healthy people and others suffering from vertigo [18, 19].

In Section 2 the main parts of the clustering approach are presented with emphasis on the extracted features and the standardization techniques. Then the sequential feature selection that finds the optimal feature subset and the comparison of clusterings are presented in Section 3. In Section 4 the feature selection and standardization techniques are assessed using Monte Carlo simulations on time series from known oscillating dynamical systems . Then in Section 5 the clustering setting is applied to EEG and OKN time series and finally in Section 6 the results are discussed and conclusions are given.

2. Clustering Design

The clustering approach followed here is based on extracting features from time series and applying the EM algorithm and k-means partition algorithm on the standardized feature vectors.

2.1. Time Series Features

The main focus of this work is on the investigation of appropriate features from oscillating time series for clustering purposes. Ten features are studied here and they can be classified to standard statistical measures (skewness, kurtosis, autocorrelation), nonlinear measures (three-point autocorrelation, mutual information, largest Lyapunov exponent) and average and standard deviation (SD) of oscillation characteristics (local maxima and oscillation period). They are denoted $q_1, q_2, \ldots, q_{10}$ and are briefly described in Table 1. The set of these 10 features is expected to reflect the attributes of the underlying mechanism of the time series, including harmonic and nonlinear dynamics. This selection is by no means optimized, as there may be other features that capture additional information in the time series. However, the selected features are representative for each of the three types.

Note that while the autocorrelation for a delay τ, $r(\tau)$, in q_3 is a linear measure the q_4 feature of the higher-order autocorrelation, $r3(\tau)$, is a nonlinear correlation measure [20]. The mutual information $I(\tau)$ is considered a more advanced nonlinear correlation measure, as opposed to $r3(\tau)$, because it involves distribution functions rather than averages. Specifically, the joint mass probability $\text{P}(i,j) = \text{P}(x_t, x_{t-\tau})$, and the marginal mass probabilities $\text{P}(i) = \text{P}(x_t)$ and $\text{P}(j) = \text{P}(x_{t-\tau})$ are computed for all bins of a partition of $\{x_t\}_{t=1}^{n}$. One can consider the mutual information sum, $SI(\tau_{\text{max}})$, as the nonlinear analogue of the Box–Pierce sum of squared autocorrelations, $Q(\tau_{\text{max}})$, for the same maximum delay time τ_{max}.

The largest Lyapunov exponent L_1 measures a different type of nonlinearity, that is the rate of divergence of nearby trajectories in a reconstructed state space. The points reconstructed from the time series with embedding dimension m and delay time τ are

Table 1: *Features extracted from a time series $\{x_t\}_{t=1}^n = \{x_1, x_2..., x_n\}$ with mean $\bar{x}$ and SD s organized in three categories.*

Notation	*Name*	*Expression*
Standard Statistical Features		
q_1	Skewness	$\lambda = \left(\sum_{t=1}^n (x_t - \bar{x})^3\right) / (ns^3)$
q_2	Kurtosis	$\kappa = \left(\sum_{t=1}^n (x_t - \bar{x})^4\right) / (ns^4) - 3$
q_3	Autocorrelation Sum (Box-Pierce)	$Q(\tau_{\text{max}}) = n \sum_{\tau=1}^{\tau_{\text{max}}} r(\tau)^2$, where $r(\tau) = \left(\frac{1}{n-\tau} \sum_{t=\tau+1}^n (x_t x_{t-\tau} - \bar{x}^2)\right) / s^2$
Nonlinear Features		
q_4	Higher-Order Autocorrelation	$r3(\tau) = \frac{\sum_{t=2\tau+1}^n (x_t-\bar{x})(x_{t-\tau}-\bar{x})(x_{t-2\tau}-\bar{x})}{\sum_{t=1}^n (x_t-\bar{x})^3}$
q_5	Mutual Information Sum	$SI(\tau_{\text{max}}) = \sum_{\tau=1}^{\tau_{\text{max}}} I(\tau)$, where $I(\tau) = \sum_{i,j} P(i,j) \ln \frac{P(i,j)}{P(i)P(j)}$
q_6	largest Lyapunov Exponent	$L_1 = \frac{1}{n'h} \sum_{t=1}^{n'} \ln \frac{\delta_{h,t}}{\delta_{0,t}}$,
Oscillation–related Features		
q_7	Average Period	$\overline{T} = \frac{1}{M} \sum_{j=1}^M T_j$, where T_j is the period of oscillation j, $j = 1, \dots, M$.
q_8	Average Peak	$\overline{z} = \frac{1}{M} \sum_{j=1}^M z_j$, where z_j is the peak of oscillation j
q_9	SD of Period	$s_T = \sqrt{\frac{1}{M-1} \sum_{j=1}^M (T_j^2 - \overline{T}^2)}$
q_{10}	SD of Peak	$s_z = \sqrt{\frac{1}{M-1} \sum_{j=1}^M (z_j^2 - \overline{z}^2)}$

$\mathbf{x}_t = \left[x_t, x_{t+\tau}, \dots, x_{t+(m-1)\tau}\right]^T$, for $t = 1, \dots, n'$, where $n' = n - (m-1)\tau - h$. When the underlying to the time series system is chaotic, it is expected that on average the initial distance of two nearby trajectories starting at $\mathbf{x}_t$ and $\mathbf{x}'_t$, defined as $\delta_{0,t} = ||\mathbf{x}_t - \mathbf{x}'_t||$, grows exponentially, i.e. $\delta_{h,t} = ||\mathbf{x}_{t+h} - \mathbf{x}'_{t+h}||$ increases exponentially with time step h [9].

The last four features are supposed to capture the main trend and variability of the oscillation peaks and periods. When the oscillations are not distinct there is an apparent problem of estimating the peaks. Any algorithm for estimating the peak has to involve a free parameter, such as the filter order when smoothing the time series in order to make the peaks detectable. We choose to work on the raw data and detect a peak when the local maximum of a sliding data window of fixed length τ_w is the center point, where we set $\tau_w = \tau_{\text{max}}/4$ (note that $\tau_{\text{max}}/4$ is the maximum delay for q_3 and q_5).

We determine the feature specific parameters uniformly for all time series involved in the clustering problem. We compute the autocorrelation function for some randomly selected time series and compute the average for the delay of zero-autocorrelation and the delay for the first autocorrelation maximum. The former is the estimate of τ used in the computation of q_4 and q_6, and the latter is the estimate of τ_{max} used in the computation of q_3 and q_5. The embedding dimension m used in q_6 is estimated from the method of false nearest neighbors applied to the time series selected above and for the estimated τ [21]. In the computation of m, as well as q_6, the box-assisted method is used to facilitate state space reconstruction and computationally efficient search for nearest neighbors, as implemented in the TISEAN package [22].

2.2. Standardization Methods

Due to the wide scale range and variety of distributions of the involved features, the effect of each feature on the partition is expected to be related to the standardization scheme. Therefore the clustering is optimized over the most well-known standardization methods, i.e. linear, variation (z-score), logistic and logarithmic. We introduce also a new standardization scheme (to the best of our knowledge), which transforms the sample cumulative density function (cdf) of each feature, as this is estimated on all time series in the data base of size, say K, call it $\hat{F}$, to the standard Gaussian cdf, Φ. The transform is actually done by applying the inverse Φ on $((i)+1)/K$, where (i) is the rank of the feature value for time series i. The Gaussian standardization scheme may perform well in the presence of outliers within clusters because outlying feature values are transformed closer to the mass of values, as implied by the bell-shape Gaussian distribution. The expressions for all standardization methods are given in Table 2.

Table 2: *Standardization methods for a feature q_j ($j=1,\dots,10$), estimated on K time series, giving $q_{j,1},\dots,q_{j,K}$ with mean $\bar{q}_j$, SD s_{q_j}, maximum $q_{j,max}$ and minimum $q_{j,min}$.*

Method	*Expression*
linear	$y_{j,i} = \frac{q_{j,i}-q_{j,\text{min}}}{q_{j,\text{max}}-q_{j,\text{min}}}$
logistic	$y_{j,i} = \frac{1}{1+e^{q_{j,i}}}$
logarithmic	$y_{j,i} = \ln(q_{j,i} - q_{j,\text{min}} + 1)$
variation	$y_{j,i} = \frac{q_{j,i}-\bar{q}_j}{s_{q_j}}$
Gauss	$y_{j,i} = \Phi^{-1}(\hat{F}(q_{j,i}))$

2.3. Clustering Method

To cluster the time series we chose the popular, fast and efficient k-means partitioning algorithm [23]. This algorithm starts with a random set of cluster centers and by making use of the Expectation-Maximization (EM) algorithm converges iteratively to the final set of clusters. We used the Euclidean metric to compute distances and we let EM algorithm run for 1000 iterations.

3. Clustering Efficiency

In our setting, we assume we know the original partition of the data base of time series. For each given set of features and standardization technique, the clustering algorithm finds a partition solution. We want to assess the agreement of the computed partition to the original partition for the whole range of feature subsets combined with all standardization techniques in order to find the optimal feature subset and standardization technique. In the following, we present a measure for the agreement of the partitions, which is further used to search efficiently across all feature subsets.

3.1. Cluster Comparison

There are different measures for cluster comparison in the literature that all boil down in measuring the correlation of the two partitions [24, 25]. We use the Corrected Rand Index (CRI), which is a standard measure for partition comparison [24, 26, 27]. Suppose two partitions of K data objects (feature vectors estimated on K time series), the first with R clusters and the second with C clusters. Then CRI is given as

$$\text{CRI} = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{k_{ij}}{2} - \binom{K}{2}^{-1}\sum_{i=1}^{R}\binom{k_{i.}}{2}\sum_{j=1}^{C}\binom{k_{.j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{R}\binom{k_{i.}}{2}+\sum_{j=1}^{C}\binom{k_{.j}}{2}\right] - \binom{K}{2}^{-1}\sum_{i=1}^{R}\binom{k_{i.}}{2}\sum_{j=1}^{C}\binom{k_{.j}}{2}} \qquad (1)$$

where k_{ij} is the number of objects in the i-cluster of the first partition and the j-cluster of the second, $k_{i.}$ is the number of objects in the i-cluster of the first partition and $k_{.j}$ is the number of objects in the j-cluster of the second partition. CRI ranges from -1 to 1, where 1 indicates exact agreement of the two partitions, values near zero indicate random agreement and negative values indicate disagreement.

We simplify the clustering search by matching the number of clusters from the k-means algorithm to the number of original groups of time series in the data base, so that in the computation of CRI the first partition is always the original one and $R = C$.

3.2. Feature Subset Selection

An exhaustive search for optimal clustering among all possible combinations of 10 features and 5 standardizations requires the computation of 5×2^{10} clusterings. Instead we adopt the algorithm of Forward Sequential Selection (FSS) of features for each standardization technique, which appears to be the most frequently used algorithm that drops the time complexity to the order of d^2, where d is the number of features [28, 29].

The evaluation function in FSS is the CRI, and with the regard to a feature subset S is denoted CRI(S). The FSS algorithm reads as follows, where l denotes the cardinality of the feature subset:

1. Set $l = 1$ and find the optimal single feature clustering, i.e. $S_l = \{q_{(l)}\}$, where $q_{(l)} = \arg\max_{q_i} \text{CRI}(\{q_i\})$ and $i = 1, \dots, d$ ($d = 10$ here).
2. For $l > 1$, compute clusterings for the feature subsets $S_{l-1}^{i} = S_{l-1} \cup q_i$, where $q_i \notin S_{l-1}$ and find the one with the largest CRI, i.e. $q_{(l)} = \arg\max_{q_i \notin S_{l-1}} \text{CRI}(S_{l-1}^{i})$.
3. If CRI($S_{l-1}^{(l)}$) $> \theta$CRI(S_{l-1}), then $S_l = S_{l-1}^{(l)}$, $l = l + 1$, and go to step 2, otherwise stop.

Starting with the best single feature clustering, the feature subset is augmented by adding a single feature at a time, but only if the partition accuracy is significantly improved. The improvement is controlled by the threshold parameter θ in step 3. In our computations we set $\theta = 1.05$ that corresponds to 5% improvement in CRI. Thus we retain small cardinality of the optimal feature subset by punishing the inclusion of features that give very marginal improvement in CRI.

4. Monte Carlo Simulations and Results

We evaluated the estimation of optimal features subset and standardization method with Monte Carlo simulations on data bases of time series from well-known oscillating systems. Each data base is comprised of three groups that correspond to different regimes or systems and each group has 50 time series of length N=1000 points.

4.1. The simulated data bases

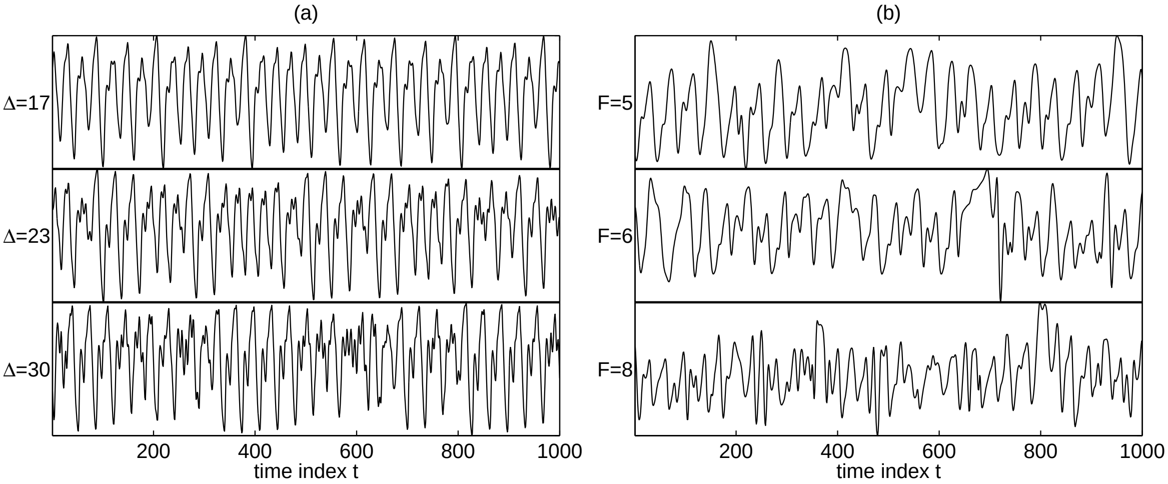
We formed three time series data bases described below.

Data Base 1 The data base is comprised of oscillating time series generated from the Mackey-Glass delay differential equation system [30]

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{0.2x(t-\Delta)}{1+[x(t-\Delta)]^{10}+0.1x(t)} \tag{2}$$

We considered three chaotic regimes of the system for $\Delta = 17, 23, 30$. The complexity of the regime increases with Δ and the fractal dimension changes from slightly larger than 2 for $\Delta = 17$ to slightly larger than 3 for $\Delta = 30$ [31]. Representative time series from each regime are given in Fig. 1a. Note that the three time series cannot easily be distinguished by eyeball judgement (maybe one can see a difference between the time series in the upper panel for $\Delta = 17$ and the time series in the lower panel for $\Delta = 30$).

Figure 1: *(a) Time series from the three regimes of the Mackey-Glass system as indicated at the right of each panel. (b) Time series of the $x(1)$ variable of the Lorenz-95 system at three regimes as indicated at the right of each panel.*



To simulate real-world-like conditions where data are often noisy, we considered also the case of adding white noise to the time series. The noise is Gaussian and has a standard deviation (SD) of 20% of the SD of the original data.

Data Base 2 This data base is comprised of more complex oscillating time series generated from the Lorenz–95 system of 10 differential equations

$$\frac{\mathrm{d}x}{\mathrm{d}t} = (x_{j+1} - x_{j-2})x_{j-1} - x_j + F, \qquad j = 1, 2 \ldots, 10, \tag{3}$$

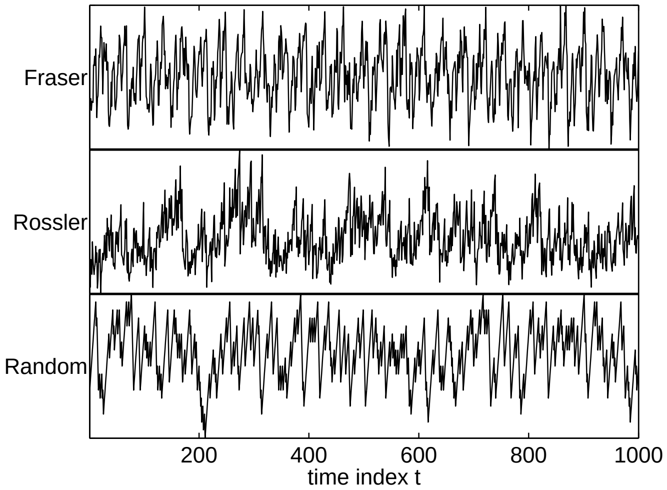
where F stands for the control parameter. We considered again three chaotic regimes of the system for $F = 5, 6, 8$ [32, 33] and we obtain the scalar time series for each regime from the first variable, $x(1)$. Note that the complexity of these regimes is higher than for those of the Mackey–Glass system, as shown in Fig. 1b. Again the three time series exhibit strong similarities with the one at the lower panel ($F = 8$) being the most complex. As before we considered also the case where 20% Gaussian white noise is added to the data.

Data Base 3 All time series in this data base consist of patterns of a rather linear upward trend followed by a somehow faster downward trend as opposed to the smooth oscillations of the time series in the two first data bases. Three systems are represented in the data base, each of a different type. The first type of time series is from the quasi-periodic (torus) system of Fraser [34]

$$\begin{bmatrix} \dot{x}(1) \\ \dot{x}(2) \\ \dot{x}(3) \\ \dot{x}(4) \end{bmatrix} = \begin{bmatrix} x(2) + a + x(1)(1 - x(1)^4) \\ 1 - (x(1) + 1)^3 \\ (x(4) + a + x(3)(1 - x(3))^4)/b \\ (1 - (x(3) + 1)^3)/b \end{bmatrix} \tag{4}$$

with parameters $a = -0.3811735$ and $b = \sqrt{5}+1$ and the observed variable is $x(2)+x(4)$. The second time series type is from the so-called Rössler hyper-chaos system

Figure 2: *Time series of the upward-downward trend pattern type from the three systems as indicated at the right of each panel.*



$$\begin{bmatrix} \dot{x}(1) \\ \dot{x}(2) \\ \dot{x}(3) \\ \dot{x}(4) \end{bmatrix} = \begin{bmatrix} -(x(2)+x(3)) \\ x(1)+ax(2)+x(4) \\ b+x(3)x(1) \\ cx(4)-gx(3) \end{bmatrix} \tag{5}$$

with parameters $a = 0.25$, $b = 3$, $c = 0.05$ and $g = 0.5$ and the observed variable is $x(4)$. Note that the apparently innocent multiplicative variable term in the third equation of (5) gives rise to chaos and the system has a strange attractor of fractal dimension slightly higher than 3 [35]. Opposite to the deterministic nature (quasi-periodic and chaotic) of the two first systems the third system is stochastic. The time series of successive patterns of random upward and downward trend is generated as follows. A series of alternating random turning points bounded to a predefined range is first generated (for this we use the BRWAD model in [36]). Then samples are filled at a fixed sampling time between the turning points according to a fixed upward and downward velocity. The downward velocity (going from a local maximum to the next local minimum) is higher to assimilate a faster downward trend. In order to make the distinction among the three different systems harder, we added 40% Gaussian white noise to the time series of the first two systems. Representative time series for the three systems are given in Fig. 2.

4.2. Results from Monte Carlo simulations

We generated 1000 Monte Carlo realizations of each data base. For each Monte Carlo realization, the 10 features in Table 1 are extracted from each of the 150 time series and standardized with the 5 techniques in Table 2. Then the following procedure is repeated for each standardization technique. The optimal feature subset and the corresponding CRI are found using the FSS algorithm (see Section 3.2). Note that the feature subset may vary, and it varies indeed, across the realizations of the same data base. Therefore the ”best” feature subset for the given standardization and data base is the one that was found by FSS most of the times over the 1000 realizations.

The summary results for all data bases are given in Table 3. First, we note that CRI values are very high and have very small variability for all data bases indicating that the selected features retrieve the initial groups of time series. To comprehend the level of accuracy it should be noted that a CRI value of about 0.98 indicates that only one time series out of 150 is misclassified. An impressive result of the simulations is that only very few features (sometimes even only a single one) could attain very high level of clustering accuracy. The best feature subset contained consistently an oscillation-related feature, coupled with a linear or nonlinear correlation measure and this result holds for all standardization techniques. Generally, there was little difference of the most frequently selected feature subsets across the standardization techniques.

Specifically, the feature couple q_7, q_3 performed best at the data base 1 (Mackey-Glass system) in the absence of noise, as shown in Fig. 3a for a single realization. Note that the group for $\Delta = 23$ could be clearly discriminated from the group for $\Delta = 30$ if only q_3 was used and from group for $\Delta = 17$ if only q_7 was used. When noise was added the clustering accuracy dropped and no feature could improve the CRI obtained only with q_3. However, at 1/4 of the realizations and when using Gauss standardization the combination of q_3, q_7, q_5 rise the CRI to the level of the noise-free case.

Table 3: *Monte Carlo results on feature subset selection for all data bases. The order of appearance of features indicate the order they are selected by FSS.*

Standardization	*Frequency*	*Features*	*Average CRI*	*SD of CRI*
Mackey-Glass system $\Delta = 17, 23, 30$, *noise-free*				
Linear	939	q_3, q_7	1.00	0.01
Logistic	908	q_3, q_7	1.00	0.00
Logarithmic	187	q_9, q_7	0.98	0.02
Variation	873	q_8, q_5	0.95	0.03
Gauss	249	q_3, q_8, q_5	0.99	0.01
Mackey-Glass system $\Delta = 17, 23, 30$, *20% noise*				
Linear	602	q_3	0.88	0.04
Logistic	607	q_3	0.87	0.04
Logarithmic	392	q_3	0.88	0.04
Variation	659	q_3	0.87	0.04
Gauss	232	q_3, q_7, q_5	0.96	0.02
Lorenz-95 system, $F = 5, 6, 8$, *noise-free*				
Linear	833	q_8, q_5	0.96	0.02
Logistic	879	q_8, q_5	0.95	0.03
Logarithmic	715	q_8, q_5	0.94	0.03
Variation	873	q_8, q_5	0.95	0.03
Gauss	649	q_8, q_5	0.91	0.03
Lorenz-95 system, $F = 5, 6, 8$, *20% noise*				
Linear	734	q_8, q_5	0.93	0.03
Logistic	757	q_8, q_5	0.92	0.04
Logarithmic	717	q_8, q_5	0.93	0.04
Variation	740	q_8, q_5	0.92	0.04
Gauss	614	q_8, q_5	0.87	0.04
Fraser torus, Rössler hyper-chaos and stochastic trends				
Linear	591	q_8, q_3	0.99	0.01
Logistic	711	q_8, q_3	1.00	0.01
Logarithmic	982	q_5	0.99	0.01
Variation	766	q_8, q_3	1.00	0.01
Gauss	248	q_5, q_8	0.94	0.02

For the data base 2, the couple q_8, q_5 was always the best for all standardizations and regardless of the presence of noise. As shown in Fig. 3b the two features together form distinct clusters. For data base 3, the couple q_8, q_3 was often the best giving very high clustering accuracy as shown in Fig. 3c for a single realization. It is notable that the same high level of accuracy could be attained only with q_5 when the logarithmic standardization was used.

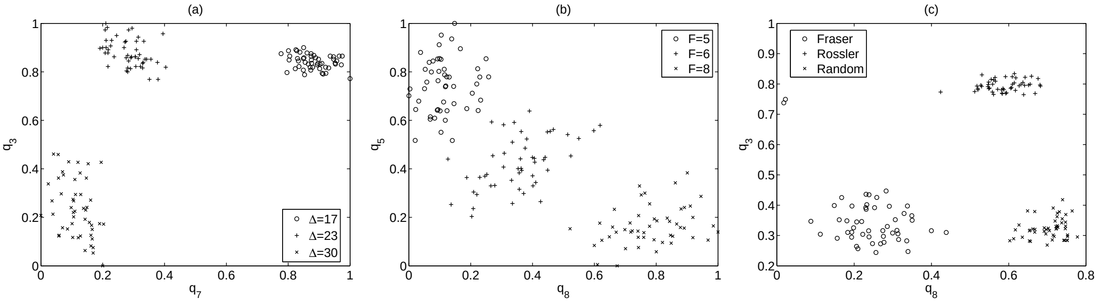
The overall results show that the features considered here could identify correctly the groups of time series for all the systems encountered in this study. This shows that these features, and in particular the oscillation-related features, are good candidates for clustering problems with oscillating time series. On the other hand, it is possible that for a different selection of more complex or more similar to each other systems, this approach

Table 4: *The EEG recordings at the early and late preictal state of 4 patients.*

patient	*duration*	*late preictal*	*early preictal*	*time series*
1	75sec	2 recordings	1 recording (5hrs before)	75
2	120sec	1 recording	1 recording (5hrs before)	50
3	110sec	1 recording	1 recording (1hr before)	50
4	85sec	1 recording	2 recordings (1 hr and 5hrs before)	75

may not give the same high clustering accuracy. This is actually the case with the real data presented in the next Section.

Figure 3: *(a) Scatter diagram of (q_7, q_3) standardized using the "linear" transform for 150 time series from the three regimes of the Mackey–Glass system as shown in the legend. (b) The same as (a) for (q_8, q_5) and the Lorenz–95 system with noise. (c) The same as (a) for (q_8, q_3) standardized using the variation transform and for the systems of data base 3.*



5. Real Data Clustering

We apply the same setting of feature extraction, standardization techniques and feature selection on two data bases of real-world oscillating time series.

5.1. Epileptic EEG

Extra-cranial EEG were recorded by a system of 25 channels with sampling frequency 200Hz from 4 epileptic patients and we analyzed separately the EEG data base for each patient. The recordings were taken under two different states of the brain: one at the late pre-ictal state that lasted from 80sec to 2min before seizure onset and the other at the early preictal state that regards times of 1hr or/and 5hrs before seizure onset.

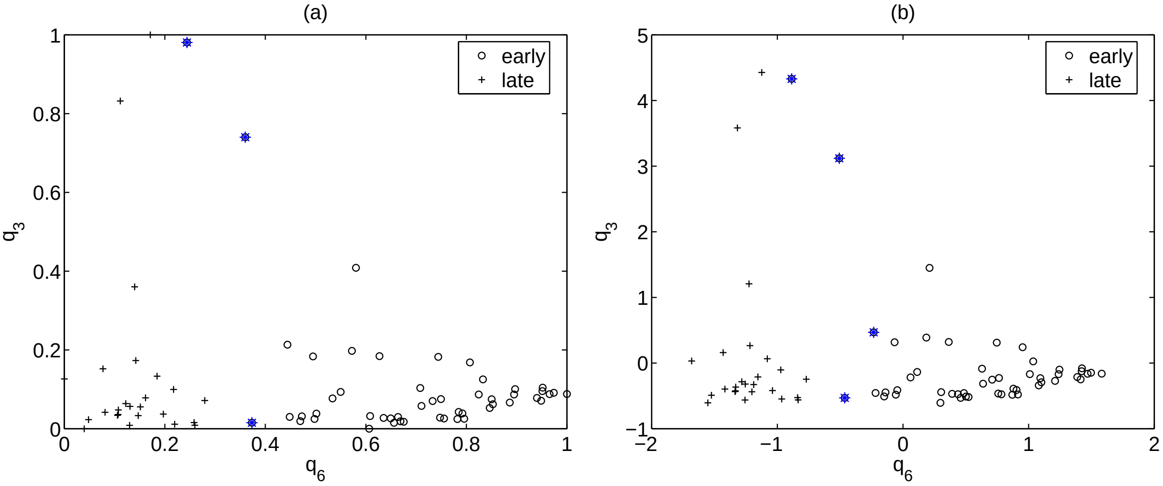
The time series length was fixed for each patient in order to avoid bias in the feature extraction due to data size. The description of the early and late preictal groups in the data base for each patient is given in Table 4. The preictal EEG time series are oscillating time series, similar in appearance to the simulated time series of data bases 2. However, it should be noted that irregular patterns, such as spikes or trends, may occur in the EEG time series that can be assigned to physiological activity or artifacts. We did not preprocessed the EEG data, so such patterns are not removed, making the clustering problem more adverse.

The objective here is to investigate the combination of features and standardization that classify best the EEG time series to the respective groups (early and late preictal states). The results of the FSS algorithm are given in Table 5. High level of clustering accuracy could be achieved for all patients with just a couple of features. The best clustering accuracy was found when three features were selected for patient 1 with only 3 time series misclassified (CRI=0.89) and for patient 3 with only 1 time series misclassified (CRI=0.92). The most frequently selected feature appears to be the largest Lyapunov exponent (q_6 in patient 1 and 4), which is in agreement with many research works in the literature (e.g. see [37]). As seen in Fig.4 the next best feature to q_6 contributed little (but significantly as the FSS algorithm gave) for the correct discrimination of the early and late preictal groups. Also we can see that the linear standardization attained misclassified 3 time series of the early preictal state group while the variation standardization misclassified one more. For patient 2, q_7 gives best results and attains alone high CRI when combined with Gauss standardization.

Table 5: *Optimal feature subset, standardization technique and clustering performance (CRI) for the EEG data.*

	Patient 1		*Patient 2*		*Patient 3*		*Patient 4*	
Standardization	*Features*	*CRI*	*Features*	*CRI*	*Features*	*CRI*	*Features*	*CRI*
Linear	q_6, q_3, q_2	0.89	q_9	0.35	q_5, q_8	0.40	q_6	0.61
Logistic	q_6, q_9	0.75	q_7, q_4	0.57	q_3	0.40	q_6, q_1	0.74
Logarithmic	q_6	0.66	q_7, q_4	0.84	q_2, q_5, q_{10}	0.92	q_6	0.57
Variation	q_6, q_3	0.79	q_9, q_6	0.70	q_5, q_8	0.40	q_6, q_1	0.70
Gauss	q_6, q_7	0.79	q_7	0.84	q_5, q_8	0.84	q_6, q_{10}	0.53

Figure 4: *(a) Scatter diagram of (q_6, q_3) standardized using the "linear" transform for 75 EEG time series of patient 1. The two groups are denoted as given in the legend and the misclassified EEG channels are denoted with an asterisk. (b) The same as in (a) but for the variation standardization.*



The variance in the CRI across standardizations is very large indicating the caution that should be taken in using the standardization technique. For example, the linear standardization gives the highest CRI for patient 1 and the lowest for patient 2 and 3, which is actually much worse than the CRI obtained with less favorable standardizations (logarithmic and Gauss).

5.2. Optokinetic Signals

Optokinetic (OKN) signals were measured in routine examination of people suffering from vertigo and healthy people as well (e.g. professional divers). For the description of the OKN signals that we used here see [18] (we used the data for rotating velocity of 60^0 and left direction). The OKN signals are characterized by successive patterns of upward and downward trends, very much like these in the time series of data base 3.

The same setting as for the EEG was applied to the data base of OKN signals for 9 healthy people and 10 patients. The results are given in Table 6. All but the Gaussian standardization techniques perform poorly and there is no indication that the inclusion of any features on the single feature subset of q_8 or q_1 can produce any better clustering results. The Gaussian standardization involves three misclassifications as shown in Fig. 5a (2 objects of ”patient” type were placed to the cluster of ”healthy” and one object of the ”healthy” type to the cluster of ”patient”). These two features have not symmetric unimodal distribution and this makes the clustering more difficult and inaccurate when the standardization does not change the form of the distribution, as shown in Fig. 5b for the linear standardization. Note that due to the small size of this data base each misclassification decreases significantly the CRI measure. The clustering results with Gaussian are comparable or even better than results from other works using a single nonlinear measure [18, 19].

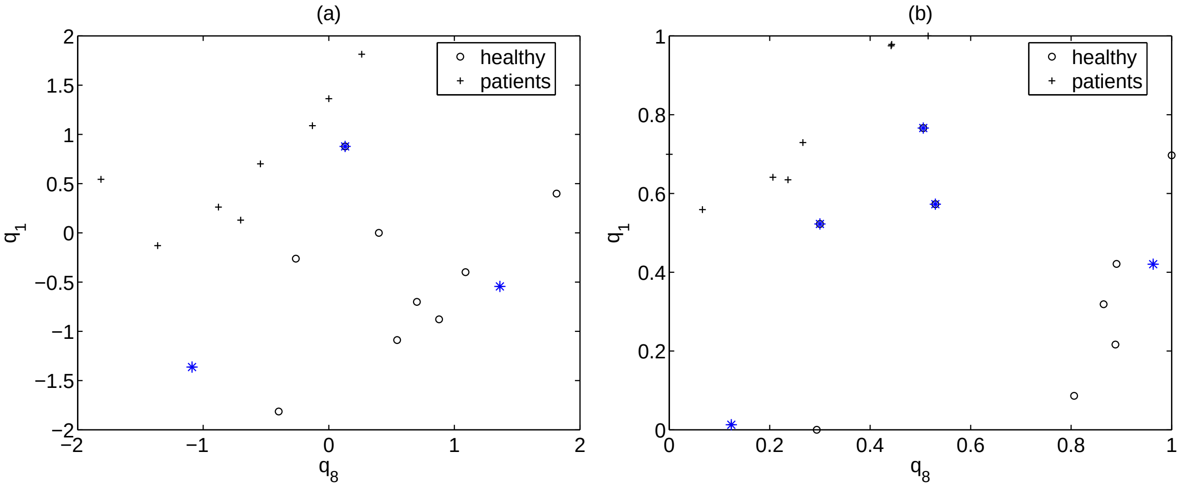
Table 6: *Optimal feature subset, standardization technique and clustering performance (CRI) for the OKN data.*

Standardization	Features	CRI
Linear	q_8	0.19
Logistic	q_8	0.19
Logarithmic	q_1	0.10
Variation	q_8	0.19
Gauss	q_8, q_1	0.44

6. Discussion

The work was focused on identifying the most relevant features for clustering data bases of oscillating time series. So, the approach is different from other standard clustering approaches in that we use the original partition of the data base in the search for the best feature subset. The evaluation measure CRI compares each time the estimated clustering to the original partition. This approach is similar to the training of candidate models in a time series or regression problem setting, where the candidate models are the feature subsets. In the same sense, CRI is the analogue to the correlation coefficient between real and fitted values estimated on the training set. The next important step is the clustering of oscillating time series of unidentified type, which is analogue to the out-of-sample

Figure 5: *(a) Scatter diagram of* (q_8, q_1) *standardized using the "Gauss" type of transform for 19 OKN time series. The groups of healthy and patients are denoted as given in the legend and the misclassified OKN time series are denoted with an asterisk. (b) The same as in (a) but for the "linear" standardization.*



predictions. This comprises the topic of future work, particularly relevant for the two real applications presented here.

We modified the Forward Sequential Selection (FSS) algorithm using the CRI as evaluation function and a threshold of improvement in CRI when stepping to a larger feature subset. This method could trace the most relevant features of the time series. We considered 10 features measuring statistical properties of the data, linear and nonlinear dynamics and oscillation characteristics. The Monte Carlo simulations on known oscillating systems showed that only a couple of features, one consistently being an oscillation-related feature, could achieve the highest clustering accuracy. The clustering of EEG and OKN data bases showed again the preference to small feature subset. Oscillation-related features were picked up here as well. In particular, the average peak magnitude was selected for the clustering of OKN, which was found optimal also for the clustering of simulated time series possessing similar oscillating patterns to those in OKN (upward and downward linear trends). On the other hand, the most important feature for the EEG turned out to be the largest Lyapunov exponent, having smaller values in the late preictal state. This result is in complete agreement with works dedicated in the use of the largest Lyapunov exponent to trace changes in the preictal state of the EEG [37].

The search of optimal features was combined with the standardization and we included the four most well-known techniques. To these we added a new one that transforms the feature values in order to have Gaussian (bell-shape) distribution. The Monte Carlo simulations did not show any dramatic differences across the standardization techniques. However, the results on the real data were different. For the clustering of EEG records, the discrimination of late preictal from early preictal states was good with the Gaussian and logarithmic standardizations in the cases that it was poor with other standardizations, and mainly the linear standardization. As for the discrimination of patients and healthy from the OKN signals, the Gaussian standardization outperformed all the others giving satisfactory clustering performance, in view of the discrimination success reported so far in the literature [38, 39].

For the real data, the analysis is by no means exhaustive and small data bases were chosen to illustrate the variability of performance under different feature subsets and standardization techniques, but also the strength of the feature-based clustering approach. For EEG, instead of using all channels from one short epoch, more appropriate would be to form the data base from EEG recordings of selected channels over many different epochs. We plan to extend the clustering to such larger data bases and the same holds for the OKN data base of patient and healthy patients.

Clustering of time series data bases becomes an emerging research area. Among different clustering approaches, the feature-based clustering can be effective only when the selected features extracted from the time series are actually the most relevant. Very often the time series are oscillatory and this work have showed the high relevance of oscillation-related features in clustering problems. Note that the calculation of oscillation-related features is much shorter than of other more sophisticated measures, such as the mutual information and the largest Lyapunov exponent, which require long computation time. The work also have showed that the selection of standardization techniques is also very important and relates strongly to the selected features.

Acknowledgments

We thank Pål Larsson from the National Center of Epilepsy of Norway for providing the EEG data and Torbjørn Aasen from the Haukeland hospital, Bergen, Norway, for providing the OKN data.

References

[1] T. W. Liao. Clustering of time series data - a survey. *Pattern Recognition*, 38(11):1857–1874, 2005.

[2] A. Wismüller, O. Lange, D. R. Dersch, G. L. Leinsinger, K. Hahn, B. Pütz, and D. Auer. Cluster analysis of biomedical image time series. *International Journal of Computer Vision*, 46(2):103 – 128, 2002.

[3] M. Vlachos, P. S. Yu, V. Castelli, and C. Meek. Structural prediodic measures for time-series data. *Data Mining and Knowledge Discovery*, 12:1 – 28, 2006.

[4] C. Goutte, P. Toft, E. Rostrup, F. Å. Nielsen, and L. K. Hansen. On clustering fMRI time series. *NeuroImage*, 9:298 – 310, 1999.

[5] Y. Xiong and D-Y. Yeung. Time series clustering with ARMA mixtures. *Pattern Recognition*, 37:1675 – 1689, 2004.

[6] A. Bagnall and G. Janacek. Clustering time series with clipped data. *Machine Learning*, 58:151 – 178, 2005.

[7] E. Keogh and J. Lin. Clustering of time-series subsequences is meaningless: Implications for previous and future research. *Knowledge and Information Systems*, 8:154 – 177, 2005.

[8] H. Zhang, T. Ho, and W. Huang. Blind feature extraction for time-series classification using Haar wavelet transform. *Lecture Notes in Computer Science*, 3497:605 – 610, 2005.

[9] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge, 1997.

[10] A. Pikovsky, M. Rosenblum, and J. Kurths. *Synchronization: A Universal Concept in Nonlinear Science*. Vol 12 of Cambridge Nonlinear Science Series.

[11] D. Kugiumtzis. Statically transformed autoregressive process and surrogate data test for nonlinearity. *Physical Review E*, 66:025201, 2002.

[12] N. Gershenfeld, B. Schoner, and E. Metois. Cluster-weighted modelling for time series analysis. *Nature*, 397:329 – 332, 1999.

[13] R. Bellotti, M. Castellano, and F. De Carlo. A chaotic map algorithm for knowledge discorvery in time series: A case study on biomedical signals. *IEEE Transactions on Nuclear Science*, 51(3):553 – 557, 2004.

[14] S. Hirano and S. Tsumoto. Empirical comparison of clustering methods for long time-series databases. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 3430:268 – 286, 2005.

[15] I. Mierswa and K. Morik. Automatic feature extraction for classifying audio data. *Machine Learning*, 58:127 – 149, 2005.

[16] F. Mormann, T. Kreuz, R. G. Rieke, C. Andrzejak, A. Kraskov, P. David, C. E. Elger, and K. Lehnertz. On the predictability of epileptic seizures. *Clinical Neurophysiology*, 116(3):569–587, 2005.

[17] E. Hirsch, F. Andermann, P. Chauvel, J. Engel, F. Lopes da Silva, and H. Luders. *Generalized Seizures: from Clinical Phenomenology to Underlying Systems and Networks*. Elsevier, Paris, 2006.

[18] T. Aasen, D. Kugiumtzis, and S. H. G. Nordahl. Procedure for estimating the correlation dimension of optokinetic nystagmus signal. *Computers and Biomedical Research*, 30:95 – 116, 1997.

[19] M. Shelhamer. Nonlinear dynamic systems evaluation of ’rhythmic’ eye movements (optokinetic nystagmus). *Journal of Neuroscience Methods*, 83:45 – 56, 1998.

[20] T. Schreiber and A. Schmitz. Discrimination power of measures for nonlinearity in a time series. *Physical Review E*, 55(5):5443 – 5447, 1997.

[21] M. B. Kennel, R. Brown, and H. D. I. Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A*, 45:3403 – 3411, 1992.

[22] R. Hegger, H. Kantz, and T. Schreiber. Practical implementation of nonlinear time series methods: The TISEAN package. *Chaos*, 9:413, 1999.

[23] J. A. Hartigan and M. A. Wong. A k–means clustering algorithm. *Applied Statistics*, 28:100 – 108, 1979.

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

[25] M. Meilă. Comparing clusterings by the variation of information. In B. Schölkopf and M. K. Warmuth, editors, *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003*, volume 2777 of *Lecture Notes in Computer Science*, pages 173 – 187. Springer, 2003.

[26] B. S. Everitt, S. Landau, and M. Leese. *Cluster Analysis*. Arnold, 2001.

[27] F. A. T. de Carvalho, R. M. C. R. de Souza, M. Chavent, and Y. Lechevallier. Adaptive hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recognition Letters*, 27:167 – 179, 2006.

[28] K. Fu. *Sequential Methods in Pattern Recognition and Machine Learning*. Academic Press, 1968.

[29] D. W. Aha and R. L. Bankert. A comparative evaluation of sequential feature selection algorithms. In D. Fisher and H. Lenz, editors, *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 1 – 7, 1995.

[30] M. Mackey and L. Glass. Oscillation and chaos in physiological control systems. *Science*, 197:287, 1977.

[31] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. *Physica D*, 9:189 – 208, 1983.

[32] E. N. Lorenz. Predictability - a problem partly solved. In *Predictability. ECMWF, Seminar Proceedings*, Shinfield Park, Reading, UK, 1995.

[33] M. S. Roulston and L. A. Smith. Combining dynamical and statistical ensembles. *Tellus A*, 55:16 – 30, 2003.

[34] A. M. Fraser. Reconstructing attractors from scalar time series: a comparison of singular system and redundancy criteria. *Physica D*, 34:391 – 404, 1989.

[35] O. E. Rössler. An equation for hyperchaos. *Physics Letters A*, 71(2 – 3):155 – 157, 1979.

[36] A. Kugiumtzis, D. and. Kehagias, E. C. Aifantis, and H. Neuhaüser. Statistical analysis of the extreme values of stress time series from the Portevin-Le Châtelier effect. *Physical Review E*, 70(3):036110, 2004.

[37] L.D. Iasemidis, P.M. Pardalos, D-S. Shiau, W. Chaovalitwongse, K. Narayanan, S. Kumar, P.R. Carney, and Sackellares J.C. Prediction of human epileptic seizures based on optimization and phase changes of brain electrical activity. *Journal of Optimization Methods and Software*, 18(1):81 – 104, 2003.

[38] M. Shelhamer. On the correlation dimension of optokinetic nystagmus eye movements: Computational parameters, filtering, nonstationarity, and surrogate data. *Biological Cybernetics*, 76:237 – 250, 1997.

[39] S. W. Kuo, T. H. Yang, and Y. H. Young. Changes in vestibular evoked myogenic potentials after Meniere attacks. *The Annals of Otology, Rhinology, and Laryngology*, 114(9):717 – 721, 2005.