Computación y Sistemas. Vol. 7 No. 1 pp. 1-16


# Recognition of system behaviours based on temporal series similarity

**D. Llanos**[a], **F.J. Cuberos**[b], **J. Meléndez**[a], **Fco. I. Gamero**[a], **J. Colomer**[a] **and J.A. Ortega**[c]
[a]Grupo eXiT, University of Girona. Av. Lluis Santalo s/n E-17071-Girona (Spain)
[b]Dpto. de Planificación. Radio Televisión de Andalucía.
Crta. San Juan-Tomare km. 1,3. S.J. Aznalfarache-Sevilla (Spain)
[c]Dpto. de Lenguajes y Sistemas Informáticos. University of Seville.
Avda Reina Mercedes S/n. Sevilla (Spain)
fcuberos@rtva.es, {dllanosr, quimmel, gamero, colomer}@eia.udg.es, ortega@lsi.us.es

## Resumen

*Existen multitud de aproximaciones al estudio de los sistemas que evolucionan en el tiempo. Este artículo revisa trabajos previos relacionados con series temporales y evalúa tres aproximaciones enfocadas a la comparación de dicho tipo de series. Dos aproximaciones están basadas en los principios del algoritmo Dynamic Time Warping (DTW) y una de ellas usa representación cualitativa basada en episodios. Ambas estrategias son discutidas y aplicadas en la diagnosis de un sistema de tanques y en la recuperación de registros de perturbaciones obtenidos en una subestación de distribución eléctrica. La tercera aproximación usa un índice de similitud definido por etiquetas cualitativas. Cada etiqueta representa un rango de valores que, desde una perspectiva cualitativa, podemos considerar similares. Esta aproximación se prueba con dos conjuntos de datos. Este estudio se completa con un estudio del ruido y de otros posibles etiquetados.*

**Palabras clave:** *Series Temporales, Análisis de series temporales, Análisis Cualitativo, Programación Dinámica, Formas, Conocimiento Cualitativo, Ruido.*

## Abstract

*There are different approaches to the temporal study of time evolving systems. This paper reviews previous works related to time series and it evaluates three approaches focused to the comparison of these type of series. Two approaches are based on the principles of Dynamic Time Warping algorithm (DTW) and one of them uses qualitative representation based on episodes. Both strategies are discussed and applied to a tank system diagnosis and retrieval of registers of perturbations gathered in a electric distribution substation. The third approach uses a similarity index defined by qualitative labels. Each label represents a range of values that, from a qualitative perspective, we may consider similar. This approach is tested with two datasets. This study is completed with a evaluation of noise and other possible labellings.*

**Keywords:** *Temporal series, Time-series Analysis, Qualitative Analysis, Dynamic Programming, Shapes, Qualitative Knowledge, Noise.*

# 1 Introduction

There are different approaches to study time evolving systems (dynamic systems theory, temporal series modelling, statistical analysis, etc.). The aim is to compare sequences representing system behaviour in order to identify similar situations. Although, the comparison of time series has been extensively treated in different fields as signal processing, statistical analysis or dynamical programming, it is still an open issue.

The identification of qualitative sequences for process diagnosis offers the possibility to monitor complex systems using reasoning mechanisms based on knowledge extracted from previous sequences. For instance process diagnosis based on symptoms (described by sequences) could be defined by reusing past experiences; (association between sequences and its diagnosis). Typical tools used with this purpose are expert systems or learning schemas built on the CBR (Case Based Reasoning) concept. A more extended explanation of CBR methodology and foundations can be consulted in (Aamodt and Plaza, 1994), (Lenz M. *et al.*, 1998). The application of this approach to industrial process suffers from the drawback that heuristic knowledge is easily represented by symbols whereas process acquisition systems provide monitoring systems with numerical data. Consequently, knowledge based decision systems are usually forced to work in a higher level of abstraction using symbolic variables instead of raw data coming from sensors. Thus, the existence of qualitative representation strategies to interface systems with decision systems is being treated in the literature (Colomer *et al.*, 1997), (Colomer J., 1998 ), (Struss P., 2002). Assuming the existence of those qualitative descriptions, next step is to define similarity metrics to identify similar symptoms in order to fire rules or to

retrieve cases representing similar behaviours in order the assess the process behaviour.

This study considers both numeric and qualitative sequences and evaluates different criteria to measure the similarity between sequences. Although, emphasis is done in the evaluation of similarity algorithms, the conversion of numeric times series into sequences of qualitative episodes is also treated because both concepts are strongly tied. This is the case of the representation language described in (Agrawal *et al.*, 1995b) based on the definition of a language (SDL) represent sequences of data by means of shapes allowing a fuzzy retrieval.

Three main approaches are evaluated in this work in order to deal with similarity problem in time sequences. Two of them are based on the principles of Dynamic Time Warping algorithm. Both strategies are discussed and applied to recognize behaviours in specific domains: $DTW$ is applied to a tank system diagnosis and retrieval of registers of perturbations gathered in a electric distribution substation. The third approach defines a new index to deal with the problem of the Longest Common Subsequence ($LCS$). The analysis of a semiqualitative model of logistics growth with delay is used to test this approach.

This paper is organized as follows: Related work are described in section 2, special attention to Shape Definition Language ($SDL$) and the problem of the Longest Common Subsequence ($LCS$) is given. In section 3, Dynamic Time Warping ($DTW$) algorithm and variations of it are summarized. In section 4, applications of $DTW$ are presented. Section 5 introduces the Qualitative Similarity Index ($QSI$), section 6 present a theorical study of noise sensitivity and section 7 computes this sensitivity in a practical sample. The last section presents the conclusions and future works.

## 2 Related Work

There has been many works on comparison of time series.

A time sequence $\vec{x}$ of length $n$ can be considered a point in a $n$-dimensional space. The natural approach to the similarity problem is to apply existing multidimensional indexing. But this techniques suffer the *dimensionality curse*: only works when the number of dimensions is low (usually 15). The normal length of time series makes impossible to index the complete sequence.

To overcome this problem a popular solution is dimensionality reduction, the sequence is replaced by a subset of values. The distance between the new series representations and the original series must be preserved. An index with the subset of values extracted from the original data is built. This index provide an efficient comparison of time series.

One of the older techniques of dimensionality reduction is transform the series from the time domain to frequency domain by means of a transform function, based on the *Euclidean distance* preservation stated in the Parseval's theorem and the results of (Oppenheim *et al.*, 1975).

The indexation of the firsts coefficients of the *Discrete Fourier Transform*, $DFT$, was the method, called *F-index*, presented in (Agrawal *et al.*, 1993) and (Rafiei and Mendelzon, 1998). The index was constructed with a $R^*$-tree, (Beckmann *et al.*, 1990). Some works extend this technique to subsequence matching as (Faloutsos *et al.*, 1994).

From other perspective, there are papers that let the user to define the concept of similarity. A set of geometric transformations, as moving average, time warping or time scaling, are used in (Goldin and Kanellakis, 1995),(Rafiei and Mendelzon, 1997) and (Rafiei 1999).

(Chan and Wai-chee, 1999) propose using Haar transform, from the Discrete Wavelet Trasform ($DWT$) family, instead of $DFT$. There is no advantage of this approach over $DFT$ as was established in (Wu *et al.*, 2000).

Other great group of works propose the selection of a set of the original values of the time series as representation of the series. (Keogh and Smyth, 1997) and (Keogh and Pazzani, 1998) select a piecewise linear segmentation. (Keogh and Pazzani, 2000) and (Yi and Faloutsos, 2000) use an approximation to the original series using constant segments, and finally (Keogh *et al.*, 2001) continues this work with an adaptative method to compute the length of segments.

Applying the concepts from human perception the landmark model identify the important points in a time series. The method define a $n$-th order landmark as a point where the $n$-th derivative is zero. Some landmarks are removed from the set of representing values if they are to close to other landmarks. This model was introduced in the paper (Perng *et al.*, 2000).

There is an important number of works based on *Dynamic Time Warping*. A deep review of these papers is made in following sections.

In the paper (Cheung and Stephanopoulos, 1990), the study of series with different time scales from a qualitative perspective is proposed.

(Jagadish *et al.*, 1995) presents a domain independent framework to manage similarity. The framework is composed by a pattern language, a transform rules language and a query language.

From a new perspective, (Shatkay and Zdonik 1996) proposes to represent pieces of the time series by

Bézier, polinomial and lineal functions.

(Kahveci *et al.*, 2001) and (Kahveci *et al.*, 2002) focus their works on similarity of multi-attributes sequences.

There are other papers covering specialized versions of the similarity problem from continuos queries to parallel algorithms. Here we have presented a short review of papers related with time series similarity and a deeper analysis can be found in (Cuberos *et al.* 2002).

In the next subsections well see the $SDL$ language and $LCS$ algorithm due its key paper in the definition of the $QSI$ approach.

### 2.1 Shape Definition Language (*SDL*)

This language proposed in (Agrawal *et al.*, 1995b) is very suitable to create queries about the evolution of values or magnitudes along the time. The method consists of the conversion of the series into a string of symbols.


Figure 1: Sample of range division.

The fundamental idea in $SDL$ is to divide the range of the possible variations between adjacent values in a collection of disjoint ranges, and to assign a label for each one of them. Figure 1 represents a sample division into three regions of the positive axis.

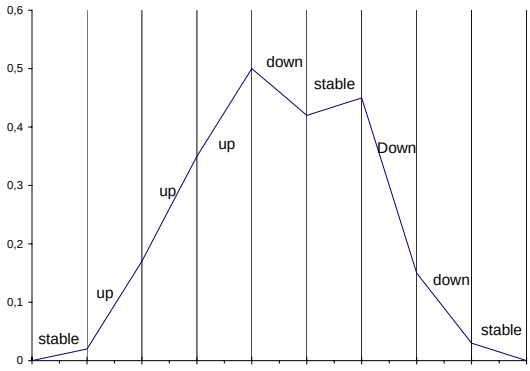


Figure 2: Time series and assigned labels.

The behaviour of a series may be described taking into account the transitions between consecutive values. A derivative series is obtained by means of the difference of amplitude among the consecutive values of the time series. The value of this difference matches in one of the disjoint ranges, and therefore this definition of the value produces a label of the alphabet. Figure 2 shows an example of a translation using the set of symbols $(Down, down, stable, zero, up, Up)$.

Every string of symbols may describe an infinite number of curves.

### 2.2 Longest Common Subsequence ($LCS$)

Working with different kinds of sequences, one of the most used similarity measures is the *Longest Common Subsequence* ($LCS$) of two or more given sequences. $LCS$ is a longest collection of elements which appears in both sequences and in the same order.

The algorithms to compute $LCS$ are well known and a deeper analysis of them is detailed in (Paterson and Dancík, 1994).

Our interest in $LCS$ come from:

- The $SDL$ language generates a string of symbols from the original time series, so it is possible to apply the $LCS$ algorithm to find a "distance" between two time series, abstracting the shapes of the curves.
- The $LCS$ is a special case of the Dynamic Time Warping ($DTW$) algorithm reducing the distance increment of each comparison to 0 or 1 depending on the presence, or absence of the same symbol. So $LCS$ inherits all the $DTW$ features.

The first work applying $LCS$ and transformations functions to time series is (Das *et al.*,1997). Later, it was extended to multidimensional trajectories in (Vlachos *et al.*, 2002).

## 3 Dynamic Time Warping Algorithm and Variations

Most of algorithms that try to measure similarity between time use the Euclidean distance or some variation in order to provide a distance between sequences. However, Euclidean distance could produce an incorrect measure of similarity because it is very sensitive to small distortions in the time axis. A method that tries

to solve this inconvenience is Dynamic Time Warping ($DTW$), this technique uses dynamic programming (Sakoe and Chiba, 1978),(Silverman and Morgan, 1990) to align time series with a given template so that the total distance measure in minimized (Figure 3). $DTW$ has been widely used in word recognition to compensate the temporal distortions related to different speeds of speech. Next, a brief notion of $DTW$ is described. Given two time series $X$ and $Y$, of length $m$ and $n$ respectively

$$X = x_1, x_2, ..., x_i, ..., x_m \; ; Y = y_1, y_2, ..., y_j, ..., y_n \quad (1)$$

To align the two sequences, $DTW$ will find a sequence $W$ of $k$ points on a $m$-by-$n$ matrix where every element $(i, j)$ of the matrix contains the local distance $d(x_i, y_j)$ between the points $x_i$ and $y_j$. This is illustrated in (Figure 4). The path $W$ is a contiguous set of matrix elements that minimize the distance between the two sequences.

$$W = w_1, w_2, ..., w_k \quad \max(m, n) \leq k \leq m + n \quad (2)$$

$$w_k = [i_k, j_k] \quad (3)$$

where $i_k$ and $j_k$ denote the time index of trajectories $X$ and $Y$ respectively. In order to find the best path $W$, some constraints on the matching process are considered:

- Constraints at the endpoints of the path, $w_1 = [1, 1]$ and $w_k = [m, n]$.
- Continuity constraints, matching paths cannot go backwards in time, this is achieved forcing $i_{k+1} \geq i_k$ and $j_{k+1} \geq j_k$.

The path is extracted by evaluating the cumulative distance $D(i, j)$ as the sum of the local distance $d(x_i, y_j)$ in the current cell and the minimum of the cumulative distances in the previous cells. This can be expressed as:

$$D(i,j) = d(x_i, y_j) + min[D(i-1, j-1), D(i-1, j), D(i, j-1)] \quad (4)$$

Several modifications of this technique have been introduced in order to apply the method in several situations. In (Keogh and Pazzani, 1999) a modification of $DTW$ is introduced to operate on a higher level of data abstraction through a piecewise linear representation. (Keogh and Pazzani, 2001) consider a higher level feature of shape considering the first derivative of the sequences. (Caiani it et al., 1998) adapt the $DTW$ approach to the analysis of the left ventricular volume signal for an optimal temporal alignment between pairs of cardiac cycles. (Vullings *et al.*, 1998) implement a piecewise linear approximation and segment the signal into separate heartbeats. $DTW$ also is used in (Kassidas *et al.*, 1998) to synchronise batch process trajectories in order to reconcile timing differences among them.

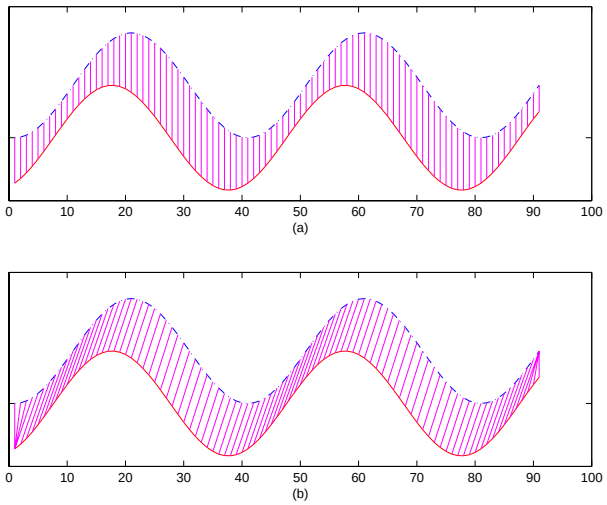


Figure 3: Two signals with similar shape. a) Euclidean distance b)DTW

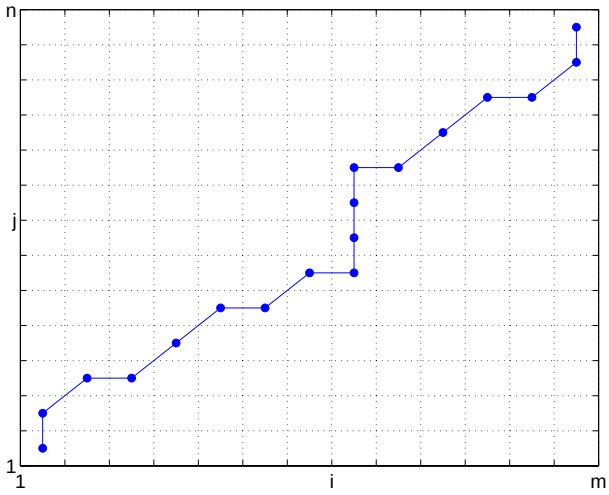


Figure 4: An example warping path.

### 3.1 Derivative Dynamic Time Warping-DDTW

The weakness of $DTW$ is in the features it considers. It only considers a data points $Y$-axis value. For example if we consider two data points ($x_i$ and $y_j$) which have identical values, but $x_i$ is part of a rising trend and $y_j$ is part of a falling trend. $DTW$ considers a mapping between these two points ideal, although intuitively we would prefer not to map a rising trend to a falling trend.

To prevent this problem, in (Keogh and Pazzani, 2001), a modification of $DTW$ was proposed. It does not consider the $Y$-values of the data points, but rather considers the higher level feature of "shape". Information about shape consists in the first derivative of the sequences; this algorithm was called Derivative Dynamic Time Warping ($DDTW$). As before we construct an $n$-by-$m$ matrix where the $(i_{th}, j_{th})$ element of the matrix contains the distance $d(x_i, y_j)$ between the two points

$x_i$ and $y_j$. With $DDTW$ the local distance measure is the square of the difference of the estimated derivatives of $x_i$ and $y_j$. While there exist sophisticated methods for estimating derivatives, particularly if one knows something about the underlying model generating the data, we use the following estimate for simplicity and generality:

$$D_x[x] = \frac{(x_i - x_{i-1}) + ((x_{i+1} - x_{i-1})/2)}{2} \qquad 1 < i < m \qquad (5)$$

This estimation is simply the average of the slope of the line through the point in question and its left neighbor, and the slope of the line through the left neighbor and the right neighbor. Empirically this estimation is more robust to outliers than any estimation considering only two data points. Note the estimation is not defined for the first and last elements of the sequence. Instead we use the estimates of the second and penultimate elements respectively.

### 3.2 Combining *DTW* and Episodes based Representations-EpDTW

Representations by means of episodes provide a good tool for situation assessment. On the one hand, uncertainty, incompleteness and heterogeneity of process data make the qualitative reasoning a good tool. On the other hand, reasoning not only with instantaneous information, but with historic behaviour of processes is necessary. Moreover, since a great deal of process data is available for the supervisory systems, to abstract and use only the most significant information is required. The representation of signals by means of episodes provides an adequate response to these necessities.

The general concept of episode was introduced in the field of qualitative reasoning by (Williams, 1986), who defined an episode as a set of two elements: a time interval, named temporal extent and a qualitative context, providing the temporal extension with significance. This definition allows defining an episode as explicitly as the qualitative context.

The formalism described in (Meléndez and Colomer, 2001) extend previous formalism to both qualitative and numerical context in order to be more general. It means that allows building episodes according to any feature extracted from variables. According to this formalism, a new representation allows to describe signal trends depending on the second derivative, that can be computed by means of a band-limited FIR differentiator (Colomer and Meléndez, 2001) in order to avoid noise amplification. The qualified first derivative at the beginning and end of each episode is used in order to obtain a more significant representation. Then, a set of 13 types of episodes is obtained (Figure 5).

Other proposed modification of the *DTW* algorithm consists on apply *DTW* not in original time series but in its episodes based representations. The representation of a sequence as episodes reduces the calculation time by decreasing the amount of manipulated data. Likewise, the qualitative character that defines an episode avoids the problem of the variability in the $Y$-axis. Therefore *DTW* can be used to align episodes to obtain a global distance. The problem is to define a local distance between episodes. In this sense, a chart of distances has been defined where the 13 types of episodes described above are related. Distances are based on the qualitative state and auxiliary characteristics that define the different types of episodes (Figure 6). However, these local distances could be subject to the criterion of the user, so one could give more importance to some episodes concerning another obtaining a different global distance and preserving the essential features of the process signal. This way, a new approach (*EpDTW*) of the *DTW* algorithm is created using episodes as a higher level representation of the signal.


Figure 5: Useful set of episodes

It is necessary to keep in mind that compared sequences could have different duration. This fact complicates the generalisation of the proposed technique. In the next example the length of the analysed sequences is different although not too dissimilar.

# 4 DTW/DDTW/EpDTW Applications

As application examples *DTW* and *DDTW* have been used in order to compare electric perturbations known as voltage sags (see example 1). In a second exam-

| | A | B | C | D | E | F | G | H | I | J | K | L | M |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | .72 | .85 | .7 | .62 | .67 | .75 | .9 | .8 | .87 | .95 | 1 | .67 |
| B | .72 | 0 | .7 | .62 | .77 | .82 | .75 | .75 | .87 | .95 | .87 | .67 | 1 |
| C | .85 | .7 | 0 | .52 | .8 | .85 | .6 | .27 | .9 | .82 | .65 | .8 | .87 |
| D | .7 | .62 | .52 | 0 | .45 | .6 | .6 | .6 | .82 | .9 | .82 | .87 | .95 |
| E | .62 | .77 | .8 | .45 | 0 | .27 | .6 | .85 | .65 | .82 | .9 | .95 | .87 |
| F | .67 | .82 | .85 | .6 | .27 | 0 | .55 | .8 | .27 | .6 | .85 | .9 | .75 |
| G | .75 | .75 | .6 | .6 | .6 | .55 | 0 | .55 | .6 | .6 | .6 | .75 | .75 |
| H | .9 | .75 | .27 | .6 | .85 | .8 | .55 | 0 | .85 | .6 | .27 | .67 | .82 |
| I | .8 | .87 | .9 | .82 | .65 | .27 | .6 | .85 | 0 | .4 | .8 | .85 | .7 |
| J | .87 | .95 | .82 | .9 | .82 | .6 | .6 | .6 | .4 | 0 | .45 | .7 | .62 |
| K | .95 | .87 | .65 | .82 | .9 | .85 | .6 | .27 | .8 | .45 | 0 | .57 | .77 |
| L | 1 | .67 | .8 | .87 | .95 | .9 | .75 | .67 | .85 | .7 | .57 | 0 | .72 |
| M | .67 | 1 | .87 | .95 | .87 | .75 | .75 | .82 | .7 | .62 | .77 | .72 | 0 |

Figure 6: Local distances between episodes

ple *EpDTW* has been used in a laboratory plant for diagnosis purposes (see example 2).

## 4.1 Example 1

Standard definition of sags is based on the minimum rms value obtained during the event and its duration is the time interval between the instant when the rms voltage crosses the voltage sag threshold (usually 90% of nominal voltage) and the instant when it returns to normal level (Bollen, 2000). A three-phase voltage sag is shown in Figure 7.

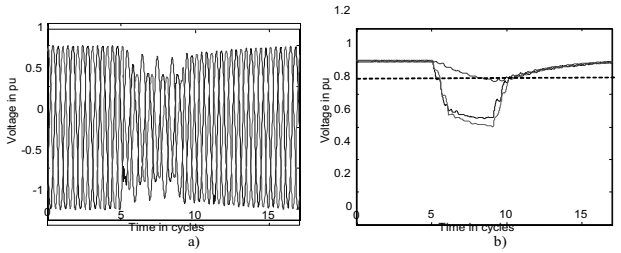


Figure 7: a)Example of a three-phase voltage sag b) rms voltage

An important sag feature known as *characteristic voltage* (Arrillaga *et al.*, 2000) can be obtained from the voltage magnitude and the voltage phase angle of the three phases. The characteristic voltage is the main indicator of the severity of the event. The absolute value (magnitude) of the characteristic voltage is comparable to the r.m.s. voltage for single-phase measurements and should be used to determine duration and retained voltage from three-phase measurements. In this work, characteristic voltage has been used to apply the *DDTW* algorithm in order to find similarity criteria among of a set of sag registers, see Figure 10.

The example shows the comparison between a new sag ($SALT18$) and the stored sags in order to retrieve the most similar one taking also into account the diagnostic (location and origin) of previous sag. Figure 8 and Figure 9 show the results obtained after applying *DTW* and *DDTW* respectively. Comparing both methods it was concluded that *DDTW* finds more similarity between the compared sag ($SALT18$) and the stored ones.

| Sag name | Global distance DTW | Location | Origin |
|---|---|---|---|
| SALT 18 | - | Distribution | Damaged conductor |
| SALT 2 | 2.46020E-05 | Distribution | Damaged conductor |
| SALT 4 | 2.78100E-05 | Distribution | Damaged conductor |
| SALT 5 | 3.01510E-05 | Distribution | Damaged conductor |
| SALT 1 | 4.09950E-05 | Distribution | Damaged conductor |
| SALT 3 | 5.29630E-05 | Distribution | Damaged conductor |
| SALT 9 | 5.23240E-04 | Transmission | Single phase trip. Successful reclose in one end line. |
| SALT 7 | 5.80530E-04 | Transmission | Single phase trip. Successful reclose in both end line. |
| SALT 10 | 9.65140E-04 | Transmission | Single phase trip. Successful reclose in both end line. |
| SALT 12 | 1.06890E-03 | Transmission | Single phase trip. Successful reclose in one end line. |
| SALT 17 | 1.30030E-03 | Transmission | Single phase trip |
| SALT 11 | 2.11770E-03 | Transmission | Single phase trip. Successful reclose in both end line. |
| SALT 6 | 3.85910E-03 | Distribution | Damaged conductor |
| SALT 13 | 3.94830E-03 | Distribution | Damaged conductor |
| SALT 14 | 6.13650E-03 | Distribution | Damaged conductor |
| SALT 15 | 2.17460E-02 | Distribution | Single phase trip |
| SALT 16 | 3.04110E-02 | Transmission | Damaged conductor |

Figure 8: Similarity results using DTW

| Sag name | Global distance DDTW | Location | Origin |
|---|---|---|---|
| SALT 18 | - | Distribution | Damaged conductor |
| SALT 2 | 3.92550E-06 | Distribution | Damaged conductor |
| SALT 4 | 4.52160E-06 | Distribution | Damaged conductor |
| SALT 1 | 6.40600E-06 | Distribution | Damaged conductor |
| SALT 3 | 7.48330E-06 | Distribution | Damaged conductor |
| SALT 9 | 7.98560E-06 | Transmission | Single phase trip. Successful reclose in one end line. |
| SALT 5 | 8.14680E-06 | Distribution | Damaged conductor |
| SALT 6 | 1.01270E-05 | Distribution | Damaged conductor |
| SALT 13 | 1.02760E-05 | Distribution | Damaged conductor |
| SALT 10 | 1.23390E-05 | Transmission | Single phase trip. Successful reclose in both end line. |
| SALT 12 | 1.27940E-05 | Transmission | Single phase trip. Successful reclose in one end line. |
| SALT 7 | 1.28730E-05 | Transmission | Single phase trip. Successful reclose in both end line. |
| SALT 11 | 1.71310E-05 | Transmission | Single phase trip. Successful reclose in both end line. |
| SALT 17 | 2.33520E-05 | Transmission | Single phase trip |
| SALT 16 | 2.78990E-05 | Transmission | Damaged conductor |
| SALT 14 | 6.57140E-05 | Distribution | Damaged conductor |
| SALT 15 | 7.98960E-05 | Distribution | Single phase trip |

Figure 9: Similarity results using DDTW

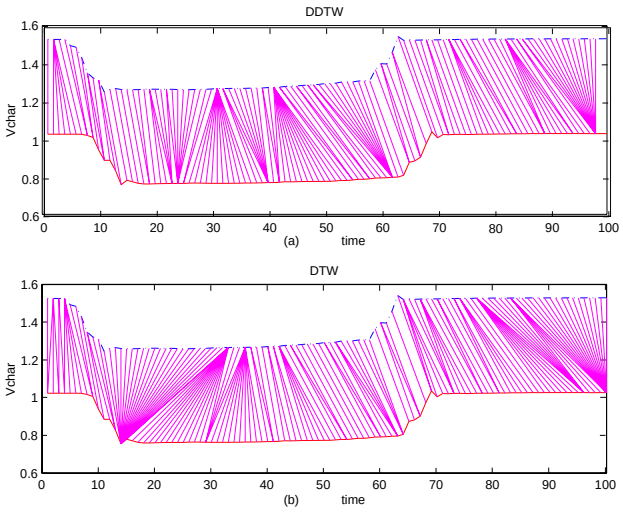


Figure 10: Voltage sag comparison a)DDTW b)DTW

Comparison between $SALT18$ and $SALT4$ has been made; Figure 10 shows the *DTW* and *DDTW* applied to both signals. As was explained before, each matrix element $(i, j)$ corresponds to the alignment between the points $x_i$ and $y_j$. This is illustrated in Figure 11, where path $W$ is a contiguous set matrix elements that defines a mapping between $X$ and $Y$. Look that path taken by the *DTW* is longer that the *DDTW* path.

## 4.2 Example 2

As application example, the *EpDTW* approach has been used in a laboratory plant for situation assessment purposes. In this plant (See Figure 12), level in tank A is controlled by means of a PID controller by pumping water from a reservoir (tank $B$). Monitored process variables are the level in tank A and the control signal (pump). Three valves ($V1$,$V2$ and $V3$) can be handled in order to simulate obstructions and leakages. Then several situations are possible by appropriate combination of opening and closing valves. Additionally, system dynamics can by slightly modified by filling or emptying the reservoir with external water. Then, input and output of external water in tank $B$ are also interesting situations to be detected.

The experiments have been developed under the assumption that two situations can not be overlapped.

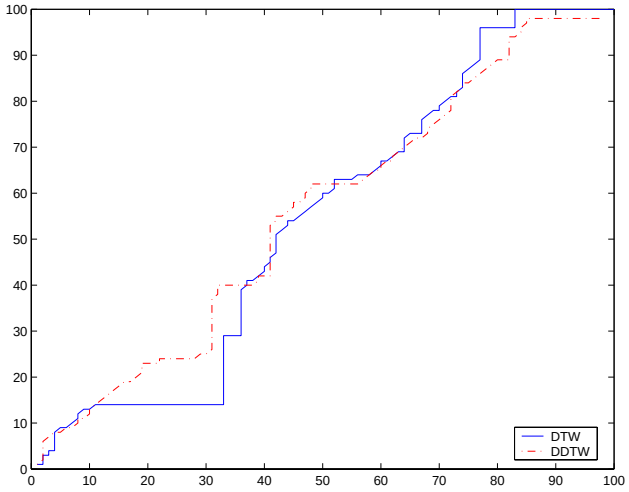


Figure 11: DTW and DDTW warping path compari-

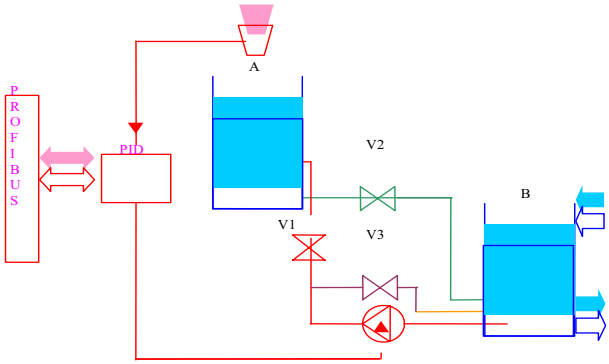


Figure 12: Laboratory plant

Thus, changes in the configuration of valves are only performed when process is in steady state. The monitoring system will be able to assess such situations and diagnose about the origin of misbehaviours according to the behaviour of measured signals described by sequences of episodes (Figure 5).

As example, a reduced set of registers has been built by obtaining the sequence of episodes for the two monitored variables in a time window of 70 seconds and the corresponding description of the situation. An example of an obstruction and its restoration for the level signal is showed in Figure 13. After testing the most

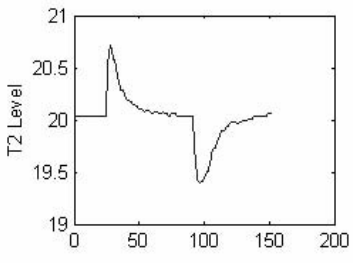


Figure 13: Example of level signal and its representations

common situations this set is composed by 29 registers (Figure 14). The assessed situation is divided in three parts: part of plant, component and diagnosis. The first field corresponds to the part of plant or operation that is being affected, in this case input of water or output of water. The second field points out the affected component of the plant, and in the last field, the corresponding diagnosis is indicated. In order to check the

| Case | Level | Control | Diagnosis | | |
|---|---|---|---|---|---|
| | | | Oper. | Comp. | Fault |
| 1 | –\| | –/– | In. A | pipe or pump | Obstruction |
| 2 | –\|\– | –/– | In. A | pipe or pump | Obstruction |
| 3 | –\|\– | –/– | In. A | pipe or pump | Obstruction |
| 4 | ∪– | –\|– | In. A | pipe or pump | End Obstr. |
| 5 | –/– | –\|– | In. A | pipe or pump | End Obstr. |
| 6 | –/|– | –\|\– | In. A | pipe or pump | End Obstr. |
| 7 | ∪|– | –\|\– | In. A | pipe or pump | End Obstr. |
| 8 | –/∩\– | –\|– | Out. A | pipe | Obstruction |
| 9 | –/∩\– | –\|\– | Out. A | pipe | Obstruction |
| 10 | –/|∩\|\– | –\|\– | Out. A | pipe | Obstruction |
| 11 | –\∪/– | –/– | Out. A | pipe | End Obstr. |
| 12 | –\|\∪/– | –/|– | Out. A | pipe | End Obstr. |
| 13 | –\∪/– | –/|– | Out. A | pipe | End Obstr. |
| 14 | –\∪/– | –/|– | In. A | pipe or pump | Leakage |
| 15 | –\∪– | –/|– | In. A | pipe or pump | Leakage |
| 16 | –\∪/ | –/|– | In. A | pipe or pump | Leakage |
| 17 | –\∪/ | –/– | In. A | pipe or pump | Leakage |
| 18 | –/∩\– | –\|\– | In. A | pipe or pump | Leak. Restored |
| 19 | –/∩\– | –\|– | In. A | pipe or pump | Leak. Restored |
| 20 | /∩\| | \|\– | In. A | pipe or pump | Leak. Restored |
| 21 | –/∩\| | –\|\– | In. A | pipe or pump | Leak. Restored |
| 22 | – | –\ | In. B | Ext. water | In. |
| 23 | – | –\ | In. B | Ext. water | In. |
| 24 | – | \– | In. B | Ext. water | End of In. |
| 25 | – | \– | In. B | Ext. water | End of In. |
| 26 | – | –/ | Out. B | Ext. water | Out. |
| 27 | – | –/ | Out. B | Ext. water | Out. |
| 28 | – | /– | Out. B | Ext. water | End of Out. |
| 29 | – | /– | Out. B | Ext. water | End of Out. |

Figure 14: The set of registers

methodology, each one of the 29 registers is compared with the other ones. Then, 841 similarity measures are carried out considering the pattern composed by the level and control signals. Similarity between symptoms (obtained by means of $EpDTW$) gives a normalized value where zero corresponds to identity. Then, similarity between registers is obtained with the average of the similarity for the two symptoms since for this process the two signals are considered with the same weight. From a general point of view, if the 29 registers are analyzed by ordering the value of similarity obtained concerning the rest of registers, it can be deduced that a threshold of 0.1 allows to obtain enough cases to do a correct situation assessment.

In a detailed example (Figure 15) the retrieved cases after comparing the register 15 with the rest are shown. The first line shows the register number while the similarity (less than 0.1) with respect to register 15 can be observed below. It can be considered that the tested register is a new register or that it already exists in the set of registers.

In this example, the initial supposition is that it

doesn't exist in the set of registers. So, the register 15 yields 6 cases with an inferior distance to 0.1 as result. The two more similar registers (13 and 14) offer different symptoms, nevertheless, the register 13 corresponds to a situation obtained by the restoration of a previous obstruction. Considering the monitory procedure, this obstruction hasn't existed previously and therefore the registers (13 and 11) are discarded. Analyzing the frequency of the remaining cases, 3/4 corresponds to pump leakage or input pipe leakage on tank A, therefore this is the offered assessment. If previous states are not kept in mind, the cases corresponding to restoration of previous obstructions they must be considered. Now, evaluating the frequency, 4/6 indicates problems in input of tank A, while the remaining 2/6 points out problems in the output. From the cases related to input, all cases indicate that the failure is located in the pump or pipe. So, this would be the proportioned diagnosis, with probability that the failure is caused for leakage.

| **Reg.** | 13 | 14 | 11 | 16 | 17 | 3 |
|---|---|---|---|---|---|---|
| **Dist.** | 0.0333 | 0.0333 | 0.0639 | 0.0639 | 0.0681 | 0.0806 |

Figure 15: Retrieved registers and similarity for register 15

# 5 Qualitative Similarity Index ($QSI$)

The idea of this index is the inclusion of qualitative knowledge in the comparison of time series. It is proposed a measure based in the matching of qualitative labels that represent the evolution of the series values. Each label represents a range of values that may be assumed as similar from a qualitative perspective . Different series with a qualitatively similar evolution produce the same sequence of labels.

The proposed approximation performs better comparisons than previously proposed methods. This improvement is mainly due to two characteristics of the index: it maximizes the exactness because it is defined using all the information of the time series, although there is always information loose in the process; and on the other hand, it focuses the comparison on the shape and not on the original values because it considers the evolution of groups as similar. It is interesting to note that we suppose that the time series are noise free and with a linear and monotonic evolution between samples.

Let $X = \langle x_0, ..., x_f \rangle$ be a time series. Our proposed approach is applied in three steps. First, a normalization of the values of $X$ is performed, yielding $\tilde{X} = \langle \tilde{x_0}, ..., \tilde{x_f} \rangle$. Using this series we obtain the differences series $X_D = \langle d_0, ..., d_{f-1} \rangle$, that it is translated to a string $S_X = \langle c_1, ..., c_{f-1} \rangle$. The similarity between two time series is calculated by means of the comparison of the two strings obtained from them, applying the previous transformation process, and then using the $LCS$ algorithm. The result is used as a similarity measure between the original time series.

## 5.1 Normalization

Keeping in mind the qualitative comparison of the series, it is made a normalization of the original numerical values in the interval [0,1]. This normalization is carried out to allow the comparison of time series with different quantitative scales.

Let $X = \langle x_0, ..., x_f \rangle$ be a time series, and let $\tilde{X} = \langle \tilde{x_0}, ..., \tilde{x_f} \rangle$ be the normalized temporal series obtained from $X$, as follows:

$$\tilde{x_i} = \frac{x_i - min(x_0, ..., x_f)}{max(x_0, ..., x_f) - min(x_0, ..., x_f)} \quad (6)$$

where $min$ and $max$ are operations that return the maximum a minimum values of a numerical sequence, respectively.

Let $X_D = \langle d_0, ..., d_{f-1} \rangle$ be the series of differences obtained from $\tilde{X}$ as follows:

$$d_i = \tilde{x_i} - \tilde{x}_{i-1} \quad (7)$$

This difference series will be used in the labelling step to produce the string of characters corresponding to $X$. It is interesting to note that every $d_i \in X_D$ is a value in the [-1,1] interval, as a consequence of the normalization process.

## 5.2 Labelling process

The proposed normalization in the previous section is focused in the slope evolution and not in the original values. A label may be assigned to every different slope, so the range of all the possible slopes is divided into groups and a qualitative label is assigned to every group.

The range division is defined depending on the parameter $\delta$ which is supplied by the experts according to their knowledge about the system. The value of this parameter has a direct influence in the quality of the results, therefore this is an open research area of this paper that we will detail in future work.

| Label | Range | Symbol |
|---|---|---|
| High increase | $[1/\delta, +\infty]$ | $H$ |
| Medium increase | $[1/\delta^2, 1/\delta]$ | $M$ |
| Low increase | $[0, 1/\delta^2]$ | $L$ |
| No variation | 0 | 0 |
| Low decrease | $[-1/\delta^2, 0]$ | $l$ |
| Medium decrease | $[-1/\delta, -1/\delta^2]$ | $m$ |
| High decrease | $[-\infty, -1/\delta]$ | $h$ |

Where the first column represents the qualitative label for every range of derivatives, which is shown in the second row. Last column contains the character assigned to each label. The proposed alphabet contains three characters for increases and three for decreases ranges, and one additional character for constant range. It is important to note that in our approach there is no application of the constraints presented in $SDL$ (Agrawal *et al.*, 1995b).

This alphabet is used to obtain the string of characters $S_X = \langle c_1, ..., c_{f-1} \rangle$ corresponding to the time series $X$, where every $c_i$ represents the evolution of the curve between two adjacent time points in $X$ and it is obtained from $X_D = \langle d_0, ..., d_{f-1} \rangle$ assigning to every $d_i$ its character in accordance with the above table.

This translation of the time series to a sequence of symbols lets us abstract from the real values and focus our attention on the shape of the curve. Every sequence of symbols describes a complete family of curves with a similar evolution.

Figure 16 shows a normalized curve with their derivative values and the assigned label to each transition between adjacent values. This example has been obtained selecting $\delta = 5$.

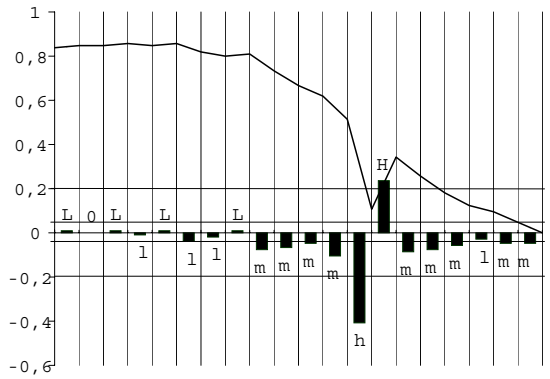


Figure 16: Sample of translation

## 5.3 Definition of $QSI$ Similarity

Let $X, Y$ be time series where $X = \langle x_0, ..., x_f \rangle$ and $Y = \langle y_0, ..., y_f \rangle$. Let $S_X, S_Y$ be the strings obtained when $X, Y$ are normalized and labelled.

The $QSI$ similarity between the strings $S_X, S_Y$ is defined as follows

$$QSI(S_X, S_Y) = \frac{\nabla(LCS(S_X, S_Y))}{m} \qquad (8)$$

where $\nabla S$ is the counter quantifier applied to string $S$. Counter quantifier yields the number of characters of $S$. On the other hand, $m$ is defined as $m = max(\nabla S_X, \nabla S_Y)$. Therefore, the $QSI$ similarity may be understood like the number of ordered symbols that we may find in the same order in both sequences simultaneously, and this value divided by the length of the longest sequence.

## 5.4 Comparison with other approach

The $QSI$ method has been compared with the algorithm introduced in (Keogh and Pazzani, 1999), called Segmented Dynamic Time Warping ($SDTW$). (Keogh and Pazzani, 1999) carries out a clustering process with a set of time series.

The $SDTW$ algorithm was tested with the Australian Sign Language Dataset from the UCI KDD (Bay, 1999) choosing 5 samples for each word. The data in the database are the 3-D position of the hand of five signers, records by means of a data glove.

The result was 22 correct clustering from 45 for $DWT$ and $SDTW$. Next, we used the similarity $QSI$ index, proposed in this paper, over the string obtained from the translation of the original values of the series. This time, the result was 44 correct clustering of 45.

For a detailed description of this comparison and the application of $QSI$ to a logistics growth model with a delay see (Ortega *et al.*, 2001).

Open questions on $QSI$ are the influence of noise in the index and the importance of the labelling schema in the results. We will try to answer these questions in the next section.

The three basic ways to divide the range of the possible slopes are:

- the values in an interval must be "similar",
- all the intervals have the same amplitude and
- every interval have the same number of elements.

The next three methods have been selected following these basic ideas.

- $CUM$ method. This method was developed and implemented in (González and Gavilán, 2000). This method makes a clustering of the initial values minimizing the average of the deviations, with the constraint that all the class marks be equally representative. This process is defined based on the statistical sampling techniques and a complete study can be found in(Cochran) and (González and Gavilán, 2000).
- *Amplitude*. The experience shows that the division of a group of values into ranges, or intervals, with

the same amplitude is the least noise sensitivity division. Selecting this method we want to verify this hypothesis in labelling.

- *Percentile*. We look for the intervals that present an approximate number of values. So every symbol has the same representation power in the set of series. The ends of the intervals are selected as the corresponding percentiles.

As the labelling methods have been presented, now we will analyze the influence of noise in $QSI$.

# 6 Noise and alternative labelling

Clearly, the noise sensitivity of $QSI$ depends on the labelling process, but we can analyze the sensitivity characteristic to any division.

Let $x_1, x_2, \cdots, x_T$ be a normalized time series and a set of values determining the ends of class intervals $L_0 < L_1, \cdots < L_k$ ($k$ class intervals, the ends can be non finite values). First, the differences between two consecutive values of the time series:

$$p(t) = \Delta x(t+1) = x_{t+1} - x_t, \qquad t = 1, \cdots, T-1$$

From this new series and the ends of the class intervals, a new series is computed:

$$\epsilon(t) = \min_{L_i} \{|p(t) - L_i|\,,\ i = 0 \cdots, k\}\,, \qquad t = 1, \cdots, T-1$$

This series verifies:

1. $\epsilon(t)$ is well defined and exists for all $t$.
2. $\epsilon(t) \geq 0$ for all $t$.
3. It comes true that:

$$\epsilon(t) \leq L = \frac{1}{2} \max\{L_i - L_{i-1},\ i = 1, \cdots, k\}\,. \quad (9)$$

This temporal series can be treated as a series of atemporal values. If a value $0 \leq \alpha < 1$ is chosen and the percentile of $\alpha$ order of the set $\{\epsilon(t)\}_{t=1}^{T-1}$ is calculated, and indicated $\epsilon_\alpha$ (see figure 17).

Associated with the differences series $p(t)$ the number:

$$p_\alpha = \frac{1}{2}\,\epsilon_\alpha$$

is considered, which does not depend on $t$.

With the study of the noise sensitivity of the temporal series $x_t$ being our target, a new normalized series is considered, in the form:

$$\widehat{x}_t = x_t + u(t) \cdot p_\alpha$$

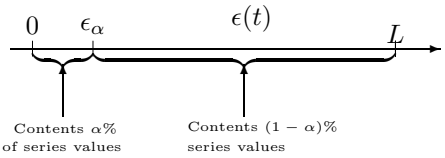


Figure 17: Percentile of series $\epsilon(t)$

where

$$-1 \leq u(t) \leq 1, \qquad \text{for alll } t$$

and the corresponding labelling to this series is computed. The differences are:

$$\begin{aligned} \widehat{p}(t) &= \Delta\widehat{x}(t+1) = \widehat{x}(t+1) - \widehat{x}(t) \\ &= x(t+1) - x(t) + (u(t+1) - u(t)) \cdot p_\alpha \\ &= p(t) + v(t) \cdot \epsilon_\alpha \end{aligned}$$

where $-1 \leq v(t) \leq 1$ and $t = 1, \cdots, T-1$. So we have:

$$-\epsilon_\alpha \leq \widehat{p}(t) - p(t) \leq \epsilon_\alpha$$

If we suppose $p(t) \in [L_i, L_{i+1}]$ then $p(t) - L_i \geq \epsilon(t)$ and $L_{i+1} - p(t) \geq \epsilon(t)$

$$\begin{aligned} \widehat{p}(t) - L_i &= \widehat{p}(t) - p(t) + p(t) - L_i \geq -\epsilon_\alpha + \epsilon(t) \\ L_{i+1} - \widehat{p}(t) &= L_{i+1} - p(t) + p(t) - \widehat{p}(t) \geq \epsilon(t) - \epsilon_\alpha \end{aligned}$$

and by the definition of $\epsilon_\alpha$ al least $(1-\alpha)\%$ is true that

$$\left.\begin{aligned} \widehat{p}(t) - L_i \geq 0 \\ L_{i+1} - \widehat{p}(t) \geq 0 \end{aligned}\right\} \Rightarrow \widehat{p}(t) \in [L_i, L_{i+1}]$$

Therefore, the labels assigned to the series $\widehat{p}(t)$ match in the same order with the the labels of $p(t)$ series.

From this reasoning we can conclude:

- Let $K$ be the normalization constant scale factor used in the normalization of the original series $x(t)$, then instead of $p_\alpha$ is defined

$$p_{1\alpha} = \frac{\epsilon_\alpha}{2\,K}.$$

- This way, we can assign to each labelling a value $p_\alpha$ of the **noise level** endured with a confidence level $1-\alpha$. If $p_\alpha$ value is relatively high, then we will have a great confidence in $QSI$ labelling provided for the series.
- As in statistics, we can determine, in a computer program, the confidence level of $\alpha$ in 5%, and therefore the labelling would have a confidence level for a level error $p_{0.05}$ of 95%.
- If for $\alpha$ high values, the $p_\alpha$ value took the value zero, then the labelling $QSI$ of the studied series would be very sensitive to the noise level.

This is valid to study the noise level supported a time series for a labelling scheme.

Now we will see the application of those different labelling to $QSI$.

| Name | Intervals | Clust. Success |
|---|---|---|
| Original2 | "-1,-.04,0,0,.04,1" | 44 |
| CUM | "-1,-0.083,-0.026,0.026,0.081,1" | 40 |
| Amplitude | "-1,-.6,-.2,.2,.6,1" | 25 |
| Percentile | "-1,-.05,-.01,.01,.05,1" | 39 |
| DTW | - | 22 |

Figure 18: Different labelling

| Name | Percentage of symbols | | | | |
|---|---|---|---|---|---|
| | S1 | S2 | S3 | S4 | S5 |
| Original2 | 26,40% | 16,75% | 13,87% | 15,58% | 27,39% |
| CUM | 12,04% | 19,84% | 35,45% | 20,75% | 11,92% |
| Amplitude | 0,00% | 1,51% | 96,88% | 1,57% | 0,04% |
| Percentile | 21,04% | 20,74% | 16,42% | 19,41% | 22,39% |

Figure 19: Labels distribution

| | Error level | % Labels 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| Original2 | 1% | 83,75 | 15,91 | 0,338 | 0 | 0 |
| | 2% | 78,84 | 19,53 | 1,523 | 0,11 | 0 |
| | 3% | 75,22 | 21,53 | 2,562 | 0,656 | 0,023 |
| | 4% | 71,64 | 23,42 | 3,443 | 1,373 | 0,123 |
| | 5% | 69,55 | 24,08 | 3,915 | 2,154 | 0,302 |
| | 6% | 67,22 | 25,11 | 4,241 | 2,94 | 0,492 |
| | 7% | 65,09 | 25,74 | 4,668 | 3,541 | 0,966 |
| | 8% | 63,55 | 25,92 | 4,924 | 4,269 | 1,338 |
| | 9% | 62,54 | 25,8 | 5,424 | 4,717 | 1,518 |
| | 10% | 61,05 | 25,97 | 5,581 | 5,371 | 2,036 |
| CUM | 1% | 94,57 | 5,426 | 0 | 0 | 0 |
| | 2% | 89,26 | 10,71 | 0,026 | 0 | 0 |
| | 3% | 84,46 | 15,12 | 0,415 | 0,004 | 0 |
| | 4% | 80,14 | 18,7 | 1,138 | 0,016 | 0 |
| | 5% | 76,58 | 21,32 | 1,971 | 0,13 | 0 |
| | 6% | 73,23 | 23,26 | 3,131 | 0,365 | 0,016 |
| | 7% | 70,98 | 24,29 | 4,124 | 0,574 | 0,026 |
| | 8% | 68,09 | 25,87 | 5,069 | 0,924 | 0,048 |
| | 9% | 66,3 | 26,65 | 5,716 | 1,246 | 0,088 |
| | 10% | 64,08 | 27,55 | 6,645 | 1,553 | 0,169 |

Figure 20: Number of label hops in presence of noise

# 7 An example

The noise sensitivity depends on the original series and on the intervals that define the labels too.

We will work with the Australian Sign Language Dataset. From this dataset 10 words from the 95 words in the database were selected. The noise is generated randomly for every value by means of a normal distribution.

The series are of different length and all shorter than 100 measures.

To see the influence of the labelling we will use the division techniques presented.

Applying the techniques to the ASL subset the next interval ends for the labelling definition are obtained.

From the original definition of $QSI$ with a $\delta = 5$ we get the first set of interval ends. As in the series included in the selected subset there are no values in the outer intervals, we reduce the number of labels to 5 and the ends of the intervals are $(-1, -.04, 0, 0, .04, 1)$. In the rest of this paper we will identify this set of intervals as $Original2$.

As the $Original2$ includes 5 symbols, the rest of the methods will be applied to obtain the same number of symbols.

The $CUM$ applied to the 50 series in the dataset with a number of 5 classes computes the set $(-1, -.083, -.026, .026, .081, 1)$.

With the selection of intervals of equal amplitude we have two options: to divide all the range $(-1, 1)$ or to divide only the zone in which values appear. As the results obtained with the two possibilities are very similar we will include only one of them, $(-1, -.6, -.2, .2, .6, 1)$.

The $Percentile$ method is applied for 5 regions, so there is 20% of each symbol in the set of series.

For comparison purpose the data obtained with $DTW$ algorithm is included.

First we will present the average $p_\alpha$, explained in the previous section, of the set of series in the ASL subset for each labelling scheme in the table below.

| $\alpha$ | Original2 | CUM | Amplitude | Percentile |
|---|---|---|---|---|
| 0.5 | 0.0163 | 0.0163 | 0.1570 | 0.0132 |
| 0.45 | 0.0130 | 0.0150 | 0.1495 | 0.0110 |
| 0.4 | 0.0116 | 0.0133 | 0.1446 | 0.0102 |
| 0.35 | 0.0111 | 0.0123 | 0.1378 | 0.0093 |
| 0.3 | 0.0092 | 0.0113 | 0.1298 | 0.0075 |
| 0.25 | 0.0071 | 0.0091 | 0.1206 | 0.0062 |
| 0.2 | 0.0043 | 0.0075 | 0.1123 | 0.0054 |
| 0.15 | 0.0022 | 0.0063 | 0.0998 | 0.0043 |
| 0.1 | 0.0010 | 0.0052 | 0.0878 | 0.0038 |
| 0.05 | 0.0002 | 0.0045 | 0.0685 | 0.0034 |

Now that we have a set of labelling processes we can check the quality of each one. As stated in previous works, the quality is defined, for us, as the number of correct clustering processes obtained with all the possible pairings of series representing two different words. As we have 10 words, the total of pairs is 45. The identification will be correct if the clustering process ends with two groups of five elements and each group contains series from the same word.

In figure 18 we present the number of correct clusterings for each labelling technique.

An important information is the distribution of symbols produced by every labelling method. This is shown in figure 19.

| | Error level | % Labels 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|---|
| Amplitude | 1% | 99,7 | 0,297 | 0 | 0 | 0 |
| | 2% | 99,49 | 0,51 | 0 | 0 | 0 |
| | 3% | 99,35 | 0,651 | 0 | 0 | 0 |
| | 4% | 99,01 | 0,995 | 0 | 0 | 0 |
| | 5% | 98,85 | 1,148 | 0 | 0 | 0 |
| | 6% | 98,47 | 1,527 | 0 | 0 | 0 |
| | 7% | 97,95 | 2,05 | 0 | 0 | 0 |
| | 8% | 97,52 | 2,478 | 0 | 0 | 0 |
| | 9% | 96,9 | 3,097 | 0 | 0 | 0 |
| | 10% | 96,37 | 3,631 | 0 | 0 | 0 |
| Percentile | 1% | 93,18 | 6,74 | 0,09 | 0,00 | 0,00 |
| | 2% | 87,06 | 12,17 | 0,754 | 0,015 | 0 |
| | 3% | 82,61 | 15,37 | 1,814 | 0,191 | 0,015 |
| | 4% | 78,55 | 18,09 | 2,748 | 0,556 | 0,051 |
| | 5% | 75,31 | 19,63 | 3,688 | 1,256 | 0,11 |
| | 6% | 72,77 | 20,69 | 4,551 | 1,799 | 0,197 |
| | 7% | 70,37 | 21,86 | 4,918 | 2,43 | 0,421 |
| | 8% | 68,06 | 22,99 | 5,236 | 2,994 | 0,72 |
| | 9% | 65,84 | 23,83 | 5,735 | 3,709 | 0,888 |
| | 10% | 64,63 | 24,07 | 6,098 | 4,126 | 1,08 |

Figure 21: Number of label hops in presence of noise cont.

The first evaluation of the noise influence in the labels can be achieved calculating the number of labels that are different between the original and the noisy series. But the magnitude of this change is important. So we define several levels of hop for a label, depending on the numbers of positions that differ the original and the noisy label. So all the labels that remain unchanged will have no hop, or a hop of level 0.

In the figures 20 and 21 the percentage of labels for every level of hop for the noise levels are presented. We use noise levels in the range from 1% to 10%.

We have to consider that the number of labels that remain unchanged is about the 60% at a noise level of 10%.

As the experience dictates, the least influence of noise is observed in the *Amplitude* labelling process.

But the number of the correct clusterings is more important for us that the change of symbols in the translated time series. So we will repeat the clustering for every labelling scheme and every level of noise.

The figures 22 a) to e) show the number of correct clusterings. As the noise is introduced in an aleatory way, we present the maximum and minimum values obtained for every labelling.

# 8 Conclusions and Further Work

This work shows some approaches in order to measure the similarity of time series. Since different patterns belonging to the same class of situations could have different time duration or magnitudes, two modifications of *DTW* algorithm are presented to compare and clas-

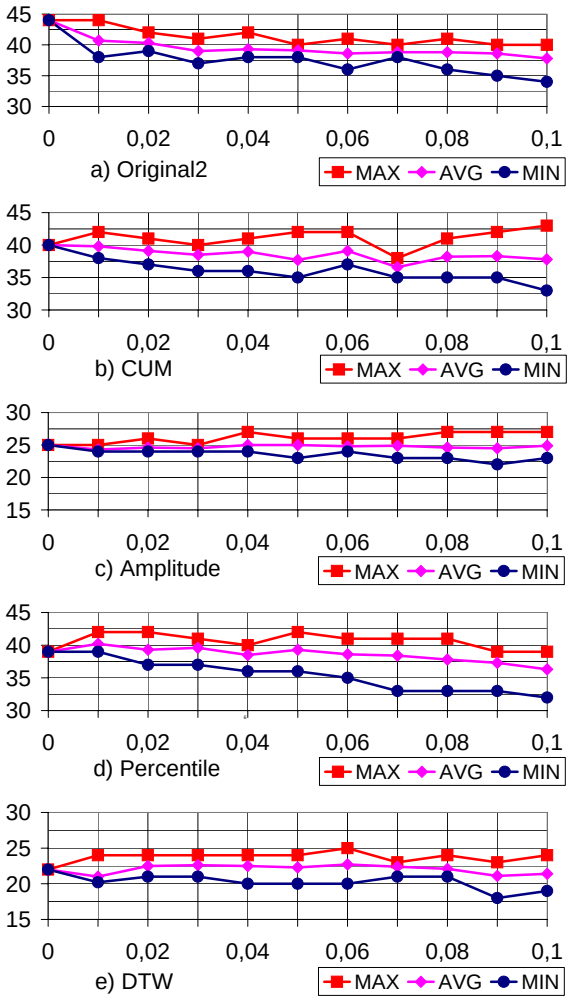


Figure 22: Clusterings success in the presence of noise

sify similar patterns. A first modification of *DTW* does not consider the $Y$-values of data point. The second is based on the integration of qualitative representation based on episodes and Dynamic programming.

Furthermore, we have reviewed the *QSI* index to measure the similarity of time series depending on its qualitative features. Also the proposed method achieves better results than previous algorithms with a similar computational cost.

We have studied the noise sensibility and others possible labelling schemas.

As it was expected, the labelling scheme that concentrates a high number of labels on few intervals is not very influenced by the presence of noise. This is shown by the *Amplitude* method. The *Original*2,*CUM* and *Percentile* methods are affected by noise in a higher level. All these methods have similar behaviours with noise.

We must conclude saying that the presence of noise in the clustering process has an influence similar to the level of the noise, the reduction of the number of correct clusterings is near lineal with the noise level.

We can remark this as a low influence of noise, as there is no level above which the results drop firmly. We have repeated the experiment with noise over 30% and the lineal relation is verified.

The idea for future works is the automation and the optimization of the division in ranges of the possible slopes to guarantee high quality clustering. When there are no information about the system which originated the time series, the *CUM* method can be used as a first approximation. Also, comparison between methods using qualitative information (*EpDTW* and *QSI*) should be done. Finally, it is necessary to extend these approaches to multivariate systems taking into account relevance of variables over the others. The inclusion of weights in this new approach must be studied to characterize this relevance.

# 9 Acknowledgement

This work is partially supported by the projects "Desarrollo de un sistema de control y supervisión aplicado a un reactor secuencial por cargas para la eliminación de materia orgánica, nitrógeno y Fósforo" (DPI2002-04579-C02-01) and DPI SECSE - "Supervisión Experta de la Calidad de Servicio Eléctrico" (DPI2001-2198) within the CICYT program from the Spanish government and FEDER funds.

# Bibliography

**Aamodt A.** and **Plaza E.**, Case-based Reasoning:Foundational Issues, Methodological Variations, and system approaches. *Artif. Intell. Communications, IOS Press*, Vol. 1, 1994, pp. 39-59.

**Agrawal R., Faloutsos C.** and **Swami A.**, Efficient similarity search in sequence databases. In *Proc. of the Fourth Intl. Conf. on Foundations of Data Organization and Algorithms (FODO '93)* , Chicago ,1993.

**Agrawal R., Lin K.I., Sawhney H.S.** and **Shim K.**, Fast similarity search in the presence of noise, scaling, and translation in time series databases. *The* $21^{st}$ *VLDB Conference* Switzerland, 1995.

**Agrawal R., Psaila G., Wimmers E.L.** and **Zaït M.**, Querying shapes of Histories. *The* $21^{st}$ *VLDB Conference* Switzerland, 1995b, pp. 502-514.

**Arrillaga J., Bollen M.H.J.** and **Watson N.R.**, Power quality following deregulation *IEEE Trans.* Vol. 88, No. 2, 2000, pp. 246 -261.

**Bay S.**, UCI Repository of KDD databases (http://kdd.ics.uci.edu/). Irvine, CA: University of California, Departamet of Information and Computer Science, 1999.

**Beckmann N., Kriegel H.-P., Schneider R.** and **Seeger B.**, "The $r^*$-tree: an eficient and robust access method for points and rectangles", *ACM SIGMOD*, 1990, pp. 322-331.

**Bollen M. H. J.**, "Understanding Power Quality problems" *IEEE Press.*, New York, 2000, pp. 139-251.

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

**Chan K.** and **Wai-chee F.A.**, Efficient time series matching by wavelets *Proc.* $15^{th}$ *International Conference on Data Engineering*, 1999.

**Cheung J.T.** and **Stephanopoulos G.**, Representation of process trend - Part II. The problem of scale and qualitative scaling, *Computers and Chemical Engineering* 14(4/5), 1990, pp. 511-539.

*Cochran W.G.* Técnicas de muestreo. *Edit. Continental*, Mexico, 6 edition.

**Colomer J.** and **Meléndez J.**, A family of FIR differentiators based on a polynomial least squares estimation. *Proc. Of the European Control Conference*, 2001, pp. 2802-2807.

**Colomer J.**, Representació qualitative assíncrona de senyals per la supervisó de sistemes dinàmics, Thesis Universitat de Girona, 1998.

**Colomer J., Meléndez J., De la Rosa J.L.** and **Aguilar J.**, A qualitative/quantitative representation of signals for supervision of continuous systems., *Proc. European Control Conference-ECC97*, Brussels, 1997.

**Cuberos F.J., Ortega J.A., Gasca R.M.** and **Toro M.** QSI - Qualitative Similarity Index. *QR-2002*. Sitges. Barcelona (Spain), (2002), pp. 45-51.

**Das G., Gunopulos D.** and **Mannila H.**, Finding similar Time Series, In *Proceedings of Principles of Data Mining and Knowledge Discovery*, $1^{st}$ European Symposium. Trondheim, Norway, 1997 , pp. 88-100.

**Faloutsos C., Ranganathan M.,** and **Manolopoulos Y.**, Fast subsequence matching in time-series databases. *The ACM SIGMOD Conference on Management of Data*, 1994, pp. 419-429

**Goldin D.Q.** and **Kanellakis P.C.**, On similarity queries for time-series data: constraint specification and implementation. In *1st Intl. Conf. on the Principles and Practive of Constraint Prog.*, Minneapolis, 1994, pages 419-429.

**González L.** and **Gavilan J.M.** Una metodología para la construcción de histogramas. Aplicación a los ingresos de los hogares andaluces, *XIV Reunión ASEPELT España* Oviedo (Spain) , 2000.

**Jagadish H., Mendelzon A.O.** and **Milo T.**, Similarity-based queries. In *Proc. of the 14th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS'95)*, San Jose, California, USA, 1995,pp. 36-45.

**Kahveci T., Singh A.** and **Gurel A.**, Shift and scale invariant search of multi-attribute time sequence. 2001.

**Kahveci T., Singh A.** and **Gurel A.**, Similarity Searching for Multi-Attribute Sequences, *SSDBM 2002*, Edinburgh, Scotland, 2002.

**Kassidas A., MacGregor J.F.** and **Taylor P.A.**, Synchronization of Batch Trajectories Using Dynamic Time Warping. *Proc. AIChE Journal* vol. 44, No. 4, 1998, pp. 864-875.

**Keogh E.J.** and **Smyth P.**, A probabilistic approach to fast pattern matching in time series databases, *Proceedings of the* $9^{th}$ *International Conference on Tools with Artifial Inteligence*, IEEE Press, 1998, pp. 578-584.

**Keogh E.J.** and **Pazzani M.J.**, An enhaced representation of time series wich allows fast and accurate classification, clustering and relevance feedback, *Proc.* $4^{th}$ *International Conference of Knowledge Discovery and Data Mining* , AAAI Press, 1998,pp. 239-241,.

**Keogh E.J.** and **Pazzani M.J.**, Scaling up Dynamic Time Warping to massive datasets, *Proc. Principles and Practice of Knowledge Discovery in Databases*, 1999.

**Keogh E.J.** and **Pazzani M.J.**, A simple dimensionality reduction technique for fast similarity search in large time series databases, In *Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2000.

**Keogh E.J., Chakrabarti K., Mehrotra S.** and **Pazzani M.J.**, Locally adaptative dimensionality reduction for indexing large time series databases, *SIGMOD*, 2001.

**Keogh E.J.** and **Pazzani M.J.**, Derivative Dynamic Time Warping, In *Proc. First SIAM International Conference on Data Mining (SDM'2001)*, Chicago, USA, 2001.

**Lenz M., Bartsch-Spörl B., Burkhard H.-D.** and **Wess S.**, Case-Based Reasoning Technology from Foundations to Applications. *Lecture Notes in Artificial Intelligence, State-of-the-Art-Survey, LNAI 1400*, Springer-Verlag, 1998.

**Meléndez J.** and **Colomer J.**, Episodes representation for supervision. Application to diagnosis of a level control system. *Proc. Workshop on Principles of Diagnosis DX'01*, Sansicario, Italy, 2001.

**Oppenheim V.** and **Schafer R.W.**, *Digital Signal Processing* , Prentice Hall, Englewood Cliffs, N.J., 1975.

**Ortega J.A., Cuberos F.J., Gasca R.M.,** and **Toro M.** , Comparación Cualitativa de Series Temporales. Indice Cualitativo de Similitud - QSI. *Computación y Sistemas, Vol. 5 No. 2*, 2001, pp. 96-108

**Paterson M.** and **Dancík V.** Longest Common Subsequences. Mathematical Foundations of Computer Science vol. 841 de LNCS, 1994, pp.127-142.

**Perng C.S., Wang H., Zhang S.R.** and **Parker D.S.**, Landmarks: a New Model for Similarity-based Pattern Querying in Time Series Databases. Proc. IEEE ICDE, 2000, pp. 33-42.

**Raifei D.** and **Mendelzon A.**, Similarity-based queries for time data series. In *Proc. of the ACM SIGMOD Intl. Conf. of Management of Data(SIGMOD '97)*, Tucson, 1998, pp. 13-24.

**Raifei D.** and **Mendelzon A.**, Efficient Retreival of similar time sequences using DFT. In *Proc. of the* $5^{th}$ *Intl. Conf. on Foundations of Data Organization and Algorithms (FODO '98)*. Kobe, 1998.

**Raifei D.**, On Similarity-based queries for time series data. In *Proc. of the* $15^{th}$ *International Conference on Data Engineering.* , Sydney, 1999.

**Sakoe H.** and **Chiba S.**, Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoustics, Speech, and Signal.*, vol. ASSP-26(1), 1978, pp. 43-49.

**Shatkay H.** and **Zdonic S.**, Approximate queries and representation for large data sequences. In *Proc. of the* $12^{th}$ *International Conference on Data Engineering.*, 1996, pp. 546-553.

**Silverman H.F.**, **Morgan D.P.**, The application of dynamic programming to connected speech recognition. *IEEE ASSP Magazine.* vol. 7, 1990, pp. 6-25.

**Struss P.**, Automated Abstraction of Numerical Simulation Models-Theory and Practical Experience, *Proc. of the* $16^{th}$ *International workshop on Qualitative Reasoning*, Barcelona-Spain, 2002, pp. 161-168.

**Vlachos M., Gunopoulos D.** and **Kollios G.**, Discovering Similar Multidimensional Trajectories. $18^{th}$ *International Conference on Data Engineering (ICDE'02).* San Jose, California (USA), 2002, pp. 673-684.

**Vullings H.J.L.M., Verhaegen M.H.G.** and **Verbruggen H.B.**, Automated ECG segmentation with dynamic time warping. *IEEE Intl. Conf. on Engineering in Medicine and Biology Society*, vol.20, No.1, 1998, pp.163-166.

**Williams B.C.**, Doing Time: Putting qualitative reasoning on firmer ground. *Proc. National Conf. on Artif. Intel. AAAI-86*, 1986, pp. 105-112.

**Wu D.,Agrawal D.** and **Abbadi A.**, A comparison of DFT and DWT based Similarity Search in Time-Series Databases. *Proc. of the* $9^{th}$ *International Conference on Information and Knowledge Management*, 2000.

**Yi B.K.** and **Faloutsos C.**, Fast time sequence indexing for arbitrary $L_p$ norms. *Proceedings of the* $26^{th}$ *Intl. Conf. on Very Large Databases*,Cairo, 2000.

**David Llanos.** Received his BS in Electronics Engineering from Pontificia Bolivariana University, Colombia in 2001. He is a PhD student in Information Technologies and member of the Department of Electronics, Computer Science and Automatic Control in the University of Girona. Currently he works in the project 'Expert Supervision of Power Quality-SECSE', DPI2001-2198. His area of interest is centered in fault diagnosis of dynamic systems reusing cases.

**Francisco J. Cuberos.** Received his BS in Computer Sciences Engineering from University of Seville, in 1993. He works as Analyst in Radio Televisión de Andalucía (R.T.V.A) since 1991. His investigation is focused in time series analysis from dynamic systems.

**Joaquin Meléndez.** Obtained a degree in Telecommunication Engineering at the Universitat Politècnica de Catalunya (UPC, Spain) in 1991 and the Ph.D. degree in Engineering by the Universitat de Girona in 1998. His a permanent professor at the Departament d'Electrònica i Informàtica (UdG) and his research activity is performed in the IIiA (Institute of Computer Science and Applications) of the same University in the field of knowledge-based techniques for fault detection, diagnosis and supervision of industrial processes and its application to real systems. Recent interest in this area is focused on the application of (Case Based Reasoning) CBR methodology for supervising dynamic processes and power quality assessment in power distribution plants.

**Fco. I. Gamero.** Received his BS in Electronics Engineering from Universitat Autònoma de Barcelona (U.A.B.) in 1999. He is a PhD student and member of the Department of Electronics, Computer Science and Automatic Control in the Universitat de Girona. Currently he works in the GROWTH project 'Advanced Decision Support Systems for Chemical/Petrochemical manufacturing processes' - CHEM. IST-2000-61200. His area of interest is centered in the application of qualitative information for the evaluation of dynamic systems, pattern recognition in the identification of process conditions and CBR (Case Based reasoning).

**Joan Colomer.** He obtained a degree in Sciences at the Autonomous University of Barcelona (Spain) in 1990 and the Ph.D. degree in Engineering by the University of Girona in 1998. Since 1992 he has been working at the University of Girona (Spain). He is member of the IIiA (Institute of Computer Science and Applications) of the UdG since its foundation. His main research interests are on knowledge based techniques for fault detection, diagnosis and supervision of industrial processes, especially on obtaining qualitative repres entation of signals for these purposes. He has been working in several national and international research projects. He has several publications in international journals and conferences. He is now leading a workpackage on process trend analysis in the GROWTH CHEM project (Advanced decision support system for chemical/petrochemical manufacturing processes) and a national project in supervision of waste water treatment plants.

**Juan Antonio Ortega.** Was born in 1968 and he obtained the Ph.D. degree in Computer Science in 2000 at the Seville University in Spain. He is professor since 1992 in the Department of Languages and Computer Systems at the Seville University. His research interests are in the temporal series and the global information systems and specifically in the domotic systems.