# Qualitative Comparison of Temporal Series. *QSI*

J.A. Ortega, F.J. Cuberos, R.M. Gasca, M. Toro, and J. Torres

**Abstract.** In this paper, the study of systems that evolve in time by means of the comparison of time series is proposed. An improvement in the form to compare temporal series with the incorporation of qualitative knowledge by means of qualitative labels is carried out. Each label represents a rank of values that, from a qualitative perspective, may be considered similar. The selection of labels of a single character allows the application of algorithms of string comparison. Finally, an index of similarity of time series based on the similarity of the obtained strings is defined.

## 1 Introduction

The study of the temporal evolution of systems is an incipient research area. The development of new methodologies to analyze and to process the time series obtained from the evolution of these systems is necessary. A time series is a sequence of real values, each one representing the value of a magnitude at a point of time. These time series are usually stored in databases.

We are interested in databases obtained from the evolution of dynamic systems. A methodology to simulate semiqualitative dynamic systems and to store the data into a database is proposed in [13]. This database may also be obtained by means of the data acquired from sensors installed in the real system. Anyway, there is a variety of applications that produce and store time series.

One of the biggest problems of working with time-series databases is to calculate the similarity between two given time series. The interest of a similarity measure is multiple: finding the different behaviour patterns of the system stored in a database, looking for a particular pattern, reducing the amount of relevance series previously to the application of analysis algorithms, etc.

Many approaches have been proposed to solve the problem of an efficient comparison. In this paper, we propose to carry out this comparison from a qualitative perspective, taking into account the variations of the time series values. The idea of our proposal is to abstract the numerical values of the time series and to concentrate on the comparison in the shape of the time series.

In this paper, time series with noise are not taken into account, and this is postponed for future works.

The remain of this paper is structured as follows: first, some related works that have been used to define our index will be analyzed. Next the *Shape Definition Language* will be introduced, which is appropriated to carry out the translation of the original values, and the problem of the *Longest Common Subsequence* ($LCS$) will also be explained. Next section will introduce our approach, the *Qualitative Similarity Index*. Finally, this index is applied to a semiqualitative logistics growth model with a delay.

## 2 Related Work

In the literature, different approximations have been developed to study time series. The shape definition language ($SDL$) is defined in [2]. This language $SDL$ is suitable for retrieving objects based on shapes contained in the histories associated with these objects. An important feature of the language is its ability to perform blurry matching where the user cares only about the overall shape. This work is the key to translate the original data into a qualitative description of its evolution.

On the other hand, the study of the problem of the Longest Common Subsequence ($LCS$) is also related to this paper, because we use ($LCS$) algorithms as the baseline to define $QSI$. A complete review of the most known solutions to this problem is collected in [15].

There have been many works on comparison of time series [7]. Most of them propose the definition of indexes, which are applied to a subset of values obtained from the original data. These indexes provide an efficient comparison of time series. They are defined taking into account only some of the original values. This improvement of speed produces a decrease in the accuracy of the comparison. These indexes are obtained applying a transformation from the time series values to a lower dimensionality space.

Other approaches differ in the way to carry out this mapping or in the selected target space. One option is to select only a few coefficients of a transformation process to represent all the information of the original series. In this approach, the change from the time domain to the frequency domain is carried out. The Discrete Fourier Transform ($DFT$) is used in [1] to reduce the series to the first Fourier Coefficients. A solution based in the Discrete Wavelet Transform ($DWT$) in a similar way is proposed in [5].

Other approaches reduce the original data in the time series, selecting a subset of the original values. A piece-wise linear segmentation of the original curve is used in [9]. In [10], the Dynamic Time Warping algorithm is applied over the segmented data, and finally in work [11] a straight dimensionality reduction with Piece-wise Constant Approximation is made, selecting a fixed number of values of the original data. This is known as *PCA-indexing*.

The last option is to generate a 4-tuple-feature vector extracted from every sequence [12]. A new distance function is defined as the similarity index.

In paper [6], the study of series with different time scales from a qualitative perspective is proposed.

## 3 Shape Definition Language (SDL)

This language, proposed in [2], is very suitable to create queries about the evolution of values or magnitudes along the time.

For any set of values stored for a period of time, 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.

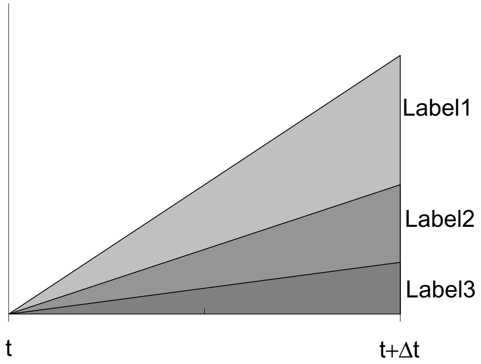

**Fig. 1.** Possible assignment of labels

Figure 1 represents a sample division into three regions of the positive axis. The behaviour of a time 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.

This translation generates a sequence of transitions based on an alphabet. The symbols of this alphabet describe the magnitude of the increments of the time series values. Every symbol is defined by means of four descriptors. The firsts two are the lower and upper bounds of the allowed variation from the initial value to the final value of the transition. The last two specify the constraints on the initial and final value of the transition respectively.

The alphabet proposed in [2] has only 8 symbols. This translation gives priority to the shape over the original values of the time series. This affirmation will be later explained. 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. All of them verify the constraints imposed for the symbols to the represented transitions. Figure 3 shows three different curves with the same sequence of symbols, even though the curves have different initial points. The *SDL* language is used to translate a time series described

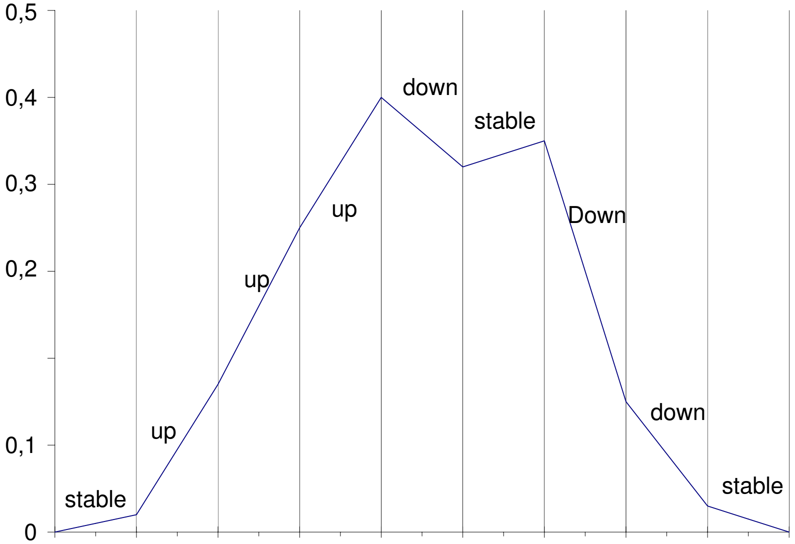

**Fig. 2.** Example of translation

by means of their numerical values into a string of symbols which represent the variations between adjacent values.

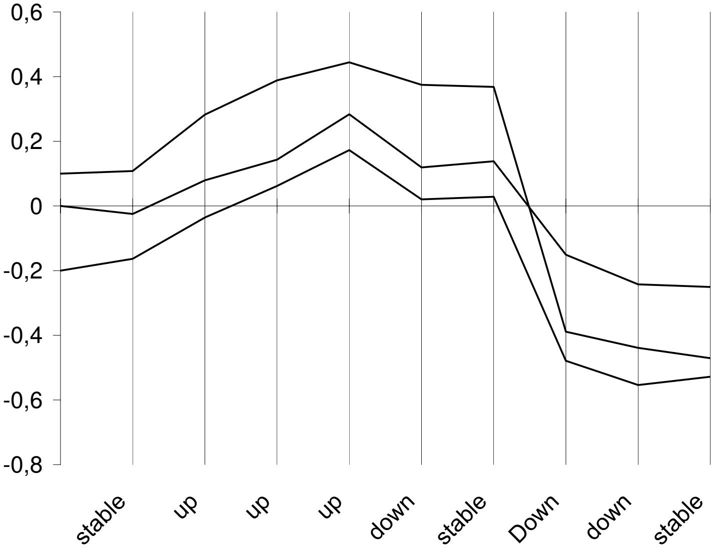

**Fig. 3.** Translation with identical sequence

## 4 Longest Common Subsequence (*LCS*)

Working with different kinds of sequences, from strings of characters to DNA chains, one of the most used similarity measure is the *Longest Common Subsequence* ($LCS$) of two or more given sequences. $LCS$ is the 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 [15].

Our interest in $LCS$ comes from a double point of view:

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

$DTW$ is an algorithm intensively used in the speech recognition area because it is appropriated to detect similar shapes that are non-aligned in the time axis. This lack of alignment induces catastrophic errors in the comparison of shapes which use the Euclidean distance.

The idea of $DTW$ is to find a set of ordered mappings between the values of two series, so the global distance warping cost is minimized.

## 5 Qualitative Similarity Index (*QSI*)

The idea of this index is the inclusion of qualitative knowledge in the comparison of time series. A measure based in the matching of qualitative labels that represent the evolution of the series values is proposed. 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: on the one hand, it maximizes the exactness because it is defined using all the information of the time series; 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 the time series are supposed to be noise-free between two samples and with a linear and monotonic evolution.

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, the difference series $X_D = \langle d_0, ..., d_{f-1} \rangle$ is obtained, which is translated into 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 with the original time series.

### 5.1 Normalization

Keeping in mind the qualitative comparison of the series, a normalization of the original numerical values in the interval [0,1] is performed. 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:

$$\tilde{x_i} = \frac{x_i - min(x_0, ..., x_f)}{max(x_0, ..., x_f) - min(x_0, ..., x_f)} \tag{1}$$

where $min$ and $max$ are operations that return the maximum and 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:

$$d_i = \tilde{x_i} - \tilde{x}_{i-1} \tag{2}$$

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 on the slope evolution and not on 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 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 will be detailed in future works.

| 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 | $[-\infty, -1/\delta]$ | $h$ |

In this table, the first column represents the qualitative label for every range of derivatives, which is shown in the second row. The last column contains the character assigned to each label. The proposed alphabet contains three characters for increase and three for decrease 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* [2].

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 points of time in $X$. It is obtained from $X_D = \langle d_0, ..., d_{f-1} \rangle$ assigning to every $d_i$ its character in accordance with the table above.

This translation of the time series into a sequence of symbols 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.

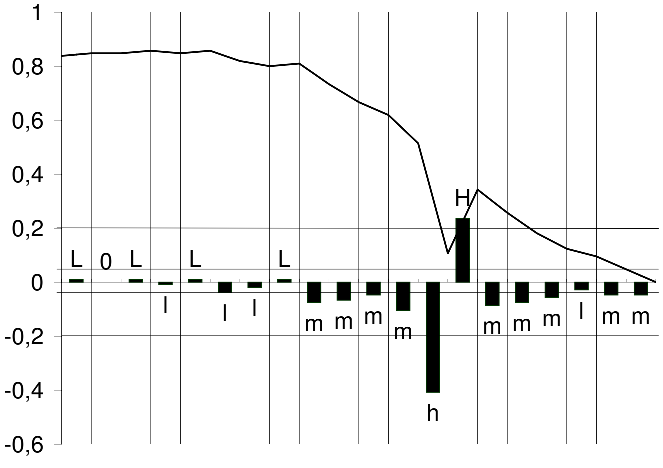

**Fig. 4.** Sample of translation

Figure 4 shows a normalized curve with their derivative values and the assigned label to each transition between adjacent values. This example has been obtained with $\delta = 5$.

### 5.3 Definition of *QSI* Similarity

Let $X, Y$ be two time series, $X = \langle x_0, ..., x_f \rangle$ 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

$$QSI(S_X, S_Y) = \frac{\nabla(LCS(S_X, S_Y))}{m} \tag{3}$$

where $\nabla S$ is the counter quantifier applied to string $S$. The 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 as the number of ordered symbols that may be found in the same order in both sequences simultaneously divided by the length of the longest sequence.

**Properties of *QSI*.** We are going to describe two properties of $QSI$.

Let $S_X, S_Y, S_Z$ be three strings of characters obtained from the transformation of three temporal series $X, Y, Z$ respectively. The definition of $QSI$ similarity verifies that:

*Property 1.* $QSI(S_X, S_Y)$ is a number in the interval $[0, 1]$. If $X$ is absolutely different of $Y$ it is 0.

$$\begin{gathered} \text{If } LCS(S_X, S_Y)) = \emptyset \\ \Downarrow \\ \nabla LCS(S_X, S_Y)) = 0 \\ \Downarrow \\ QSI(S_X, S_Y) = 0. \end{gathered} \tag{4}$$

The $QSI(S_X, S_Y)$ value increases according to the number of coincident characters. This number is 1 if $S_X = S_Y$.

$$\begin{gathered} \text{If } LCS(S_X, S_Y)) = S_X \\ \Downarrow \\ \nabla LCS(S_X, S_Y)) = S_X \\ \Downarrow \\ QSI(S_X, S_Y) = 1. \end{gathered} \tag{5}$$

*Property 2.* The length of the strings to compare has also an important influence. In this sense, two strings with approximated lengths and with a number of coincident symbols are more similar than two strings with the same number of coincident symbols but with different lengths:

$$\begin{gathered} \nabla S_X \approx \nabla S_Y, \nabla S_X \approx \nabla S_Z, \\ \nabla LCS(S_X, S_Y) \approx \nabla LCS(S_X, S_Z) \\ \Downarrow \\ QSI(S_X, S_Y) > QSI(S_X, S_Z) \end{gathered} \tag{6}$$

### 5.4 Comparison with Other Approaches

When a new approach is introduced, it is interesting to test its validity and the improvements with respect to the other approaches appeared in the literature. In this paper, our approach is going to be compared with the algorithm introduced in [10], called Segmented Dynamic Time Warping ($SDTW$). This algorithm carries out a clustering process with a set of time series. Every clustering process joins sets of data in subsets trying to minimize the similarity among the elements of every subset and the similarity between different subsets.

The $SDTW$ algorithm was tested with the Australian Sign Language Dataset from the UCI KDD [4] choosing 5 samples for each word. The data in the database are the 3-D position of the hand of five signers, recorded by means of a data glove.

In order to carry out the comparison between both approaches, the same 10 words used in [10] from the 95 words included in the database have been chosen. Next, for every possible pairing of different words (45), the 10 sequences (5 of each word) have been clustered, using a hierarchical average clustering with two different distance measures. First, the distance defined in the classic $DTW$ algorithm applied to the original numerical values of the series was used. The result was 22 correct clustering from 45. Then, the similarity $QSI$ index, proposed in this paper, over the string obtained from the translation of the original values of the series was used. This time, the result was 44 correct clustering from 45. Therefore, our index obtains a 97.7% of accuracy. The $QSI$ error is clearly lower than the obtained with $DWT$.

The total success obtained with $DWT$ is exactly the same reported by [10], but the success obtained with $QSI$ similarity is better.

## 6 Application of $QSI$ to a Logistics Growth Model with a Delay

The following generic names: logistic, sigmoidal, and s-shaped processes are given to those systems in which an initial phase of exponential growth is followed by another phase of approaching to a saturation value asymptotically (figure 5). This growth is exhibited by those systems for which exponential expansion is truncated by the limitation of the resources required for this growth. In literature, these models have been profusely studied. They abound both in natural processes, and in social and socio-technical systems. These models appear in the

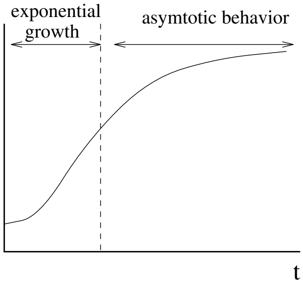

**Fig. 5.** Logistics growth curve

evolution of bacteria, in mineral extraction, in world population growth, economic development, learning curves, some diffusion phenomena within a given population such as epidemics or rumors, etc. In all these cases, their common behaviours are shown in figure 6. There is a bimodal behaviour pattern attractor: $A$ stands for normal growth, and $O$ for decay. It can be observed how it combines exponential with asymptotic growth. This phenomenon was first modelled by the Belgian Sociologist P.F. Verhulst in relation with human population growth. Nowadays, it has a wide variety of applications, some of which have just been mentioned.

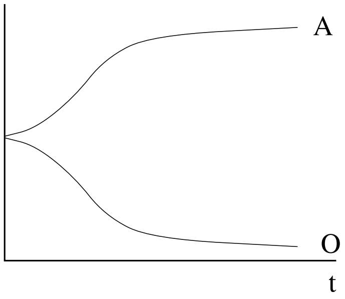

**Fig. 6.** Logistics growth model

Let $S$ be the qualitative model. If a delay in the feedback paths of $S$ is added, then its differential equations are

$$\Phi \equiv \begin{cases} \dot{x} = x(n\,r - m), \\ y = delay_\tau(x), \quad x > 0, \quad r = h_1(y), \\ h_1 \equiv \{(-\infty,-\infty),+,(d_0,0),+,(0,1), \\ \qquad +,(d_1,e_0),-,(1,0),-(+\infty,-\infty)\} \end{cases}$$

being $n$ the increasing factor, $m$ the decreasing factor, and $h_1$ a qualitative continuous function defined by means of points and the derivative sign among two consecutive points. These functions are explained in detail in [14]. This function has a maximum point at $(x_1, y_0)$. The initial conditions are

$$\Phi_0 \equiv \begin{cases} x_0 \in [LP_x, MP_x], \\ LP_x(m), \\ LP_x(n), \\ \tau \in [MP_\tau, VP_\tau] \end{cases}$$

where $LP, MP, VP$ are the qualitative unary operators *slightly positive, moderately positive and very positive* for the $x, \tau$ variables.

The methodology described in [13] is applied to this model in order to obtain the database of time series. This methodology transforms this semiqualitative model into a family of quantitative models. Stochastic techniques are applied to choose a quantitative model of the family. The simulation of one selected quantitative model generates a time series that is stored into the database. We would like to classify the different behaviours of the system applying the $QSI$ similarity to the obtained database. Figure 7 contains the table obtained when this index is applied between every two time series of the database. Figure 7 shows some of these time series. According to the obtained value $QSI$, three different behaviours appear in figure 7. These results to obtain the behaviour patterns of the system are in accordance to others appeared in the bibliography [3] and [8], where the results are concluded by means of a mathematical reasoning.

| 54X | 55X | 1X | 18X | 77X | 17X | 73X | Series |
|---|---|---|---|---|---|---|---|
| 0,87 | 0,872 | 0,41 | 0,44 | 0,494 | 0,43 | 0,376 | 50X |
| | 0,994 | 0,292 | 0,314 | 0,388 | 0,384 | 0,35 | 54X |
| | | 0,294 | 0,316 | 0,39 | 0,384 | 0,35 | 55X |
| | | | 0,758 | 0,792 | 0,598 | 0,586 | 1X |
| | | | | 0,754 | 0,58 | 0,552 | 18X |
| | | | | | 0,632 | 0,62 | 77X |
| | | | | | | 0,93 | 17X |

**Fig. 7.** $QSI$ similarity of the model

Figure 8 shows the time series grouped by these behaviours. The 50x, 55x and 54x time series have a behaviour which is labelled as *decay and extinction.* In a similar way, the 18x, 1x and 77x time series whose behaviour is classified as *recovered equilibrium* and, finally, the 17x and, 73x time series whose behaviour is labelled as *retarded catastrophe* are shown.

## 7 Conclusions and Further Work

In this paper, the $QSI$ index to measure the similarity of the time series depending on its qualitative features has been introduced. Furthermore, the proposed similarity index achieves better results than previous works with a similar computational cost.

In order to apply the $QSI$ similarity index between two time series, first of all, a normalization process is necessary. Next, the sequence of differences is obtained from this normalized series. Finally, this sequence is translated into a string of characters using a qualitatively defined alphabet. The $LCS$ algorithms are used to calculate the index $QSI$. The results obtained are in accordance with other previous works, although our approach produces an improved classification.

In the future, the idea is the automation and the optimization of the division into ranges of the possible slopes by studying the number of regions and their limits.

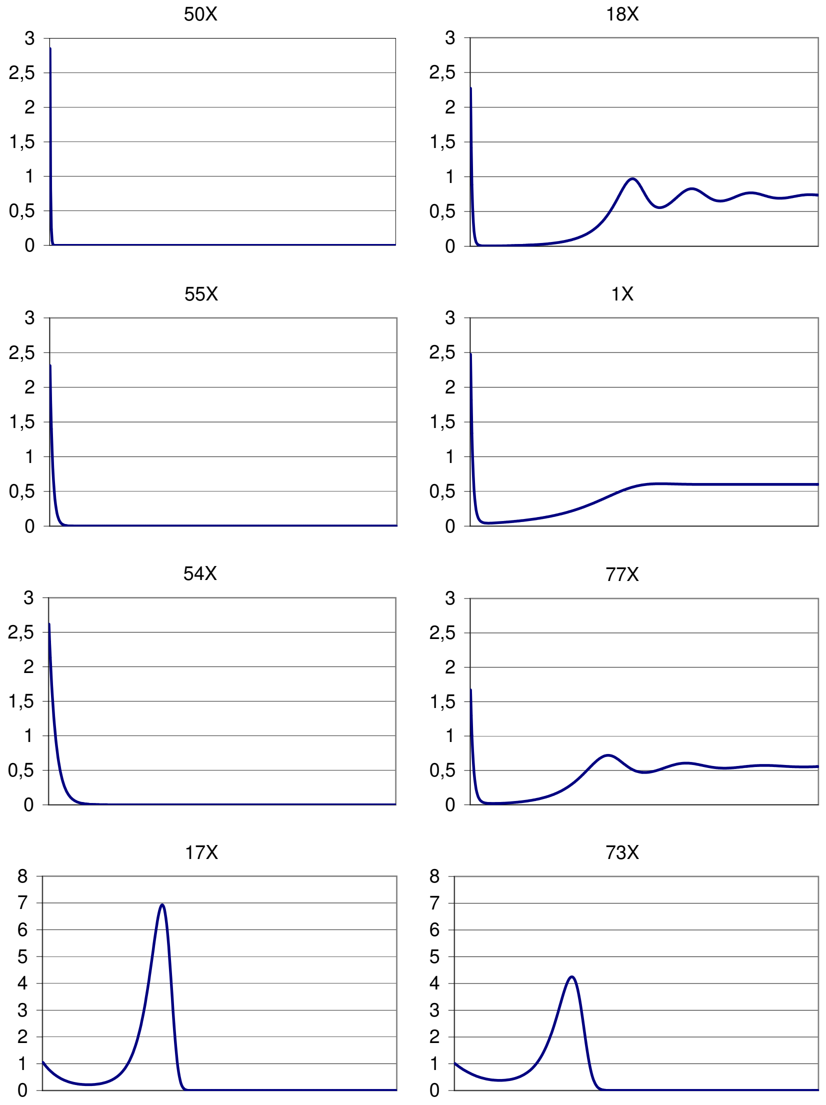

**Fig. 8.** Representations

**Acknowledgments.** This work was partially supported by the Spanish Interministerial Committee of Science and Technology by means of DPI2001-4404-E and DPI2000-0666-C02-02 programs.

## References

1. *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).
2. *Agrawal R., Psaila G., Wimmers E.L. and Zaït M.* Querying shapes of Histories. *The* $21^{st}$ *VLDB Conference* Switzerland, pp. 502-514 (1995).
3. *Aracil J., Ponce E. and Pizarro L.* Behavior patterns of logistic models with a delay *Mathematics and computer in simulation* 44: 123–141, (1997).
4. *Bay S.* UCI Repository of KDD databases (http://kdd.ics.uci.edu/). Irvine, CA: University of California, Departamet of Information and Computer Science. (1999).
5. *Chan K. and Wai-chee F.A.* Efficient time series matching by wavelets *Proc.* $15^{th}$ *International Conference on Data Engineering*, (1999).
6. *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), pp. 511-539, (1990).
7. *Faloutsos C., Ranganathan M., and Manolopoulos Y.* Fast subsequence matching in time-series databases. *The ACM SIGMOD Conference on Management of Data*, pp. 419-429 (1994).
8. *Karsky M. Dore J.-C. and Gueneau P.* Da la possibilité d'apparition de catastrophes différès. *Ecodecision No 6*, (1992).
9. *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*, pp. 239-241, AAAI Press (1998).
10. *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).
11. *Keogh E.J. and Pazzani M.J.* A simple dimensionality reduction technique for fast similarity search in large time series databases, (2000).
12. *Kim S-W, Park S. and Chu W.W.* An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases. *Proc. 17th IEEE Int'l Conf. on Data Engineering*, Heidelberg, Germany, (2001).
13. *Ortega J.A., Gasca R.M. and Toro M.* A semiqualitative methodology for reasoning about dynamic systems. $13^{th}$ *International Workshop on Qualitative Reasoning.* Loch Awe (Scotland), 169–177, 1999.
14. *Ortega J.A.*Patrones de comportamiento temporal en modelos semicualitativos con restricciones. Ph.D. diss., Dept. of Computer Science, Seville Univ, (2000).
15. *Paterson M. and Dancík V.* Longest Common Subsequences. *Mathematical Foundations of Computer Science* vol. 841 de LNCS, pp.127-142, (1994).