QSI - Alternative Labelling and Noise Sensitivity

F.J. Cuberos[a], J.A. Ortega[b], F. Velasco[c], and L. González[c],
[a] Dept. Planificación-Radio Televisión de Andalucía, San Juan de Aznalfarache. Seville (Spain)
[b] Dept. de Lenguajes y Sistemas Informáticos. University of Seville (Spain).
[c] Dept. of Applied Economy I, University of Seville (Spain).
fjcuberos@rtva.es, ortega@lsi.us.es, {luisgon, velasco}@us.es

Abstract

There are different approaches to the temporal study of time evolving systems. In this paper, this study is carried out by means of the comparison of time series. This paper continues previous works on QSI and studies the noise sensitivity of this index. The noise sensitivity depends basically on the labelling process. This study is completed with a comparison with other possible labellings.
The alternative labelling techniques are selected keeping different goals in mind: A better representativeness of the class marks, a reduced noise sensitivity and a similar number of every symbol into which the series are translated. We have carried out a detailed study applying different levels of noise for all this labelling schemes and checking the quality of the obtained index.

Introduction

The study of the temporal evolution of systems is an incipient research area. It is necessary the development of new methodologies to analyze and to process the time series obtained from the evolution of these systems. These time series are usually stored in databases. It is necessary to develop new algorithms and techniques for its study.

A time series is a sequence of real values, each one representing the value of a magnitude at a point of time. A possible field of application is the comparison of time series in numeric databases. We are interested in databases obtained from the evolution of dynamic systems. A methodology to simulate semiqualitative dynamic systems it is proposed in (Ortega *et al.* 99). These simulations are stored into a database. This database may also be obtained by means of the data acquired from sensors installed in the real system. There is a variety of applications to produce and to store time series.

When we are working with time-series databases, one of the biggest problems is to calculate the similarity between two given time series. The interest of a similarity measure is multiple. In this paper, this interest is focused on: finding the different behaviour patterns of the system stored in a database, looking for a particular pattern, reducing the number of relevance series before applying analysis algorithms, etc, as was presented in (Cuberos *et al.* 02).

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 the comparison on the shape of the time series.

Assuming that similarity is a distance function of the time series, we catalogue the basic queries to manage a time series database in three groups:

- *Range query:* given a series, finding those series that are similar within a distance.
- *Nearest neighbor*: given a series, finding in the database the series which is the nearest neighbor in accordance with a defined distance.
- *All-pairs query*: finding all the pairs of series in the database that are within a distance of each other.

For the related work, a general review was presented in (Cuberos *et al.* 02).

Our main objective is to show the relation between noise and labelling. We will present several labelling techniques and we´ll study the influence of noise. So we will compare the noise sensitivity of QSI with all the labelling schemes.

The rest of this paper is structured as follows: first a review of QSI is presented, including a definition and basics related works. Next the different labelling methods are shown. Next the influence of labelling in noise sensitivity is studied. And finally the labelling schemes are applied to a real dataset.

Qualitative Similarity Index (QSI)

The idea of this index is the inclusion of qualitative knowledge in the comparison of time series. A measure based on 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.

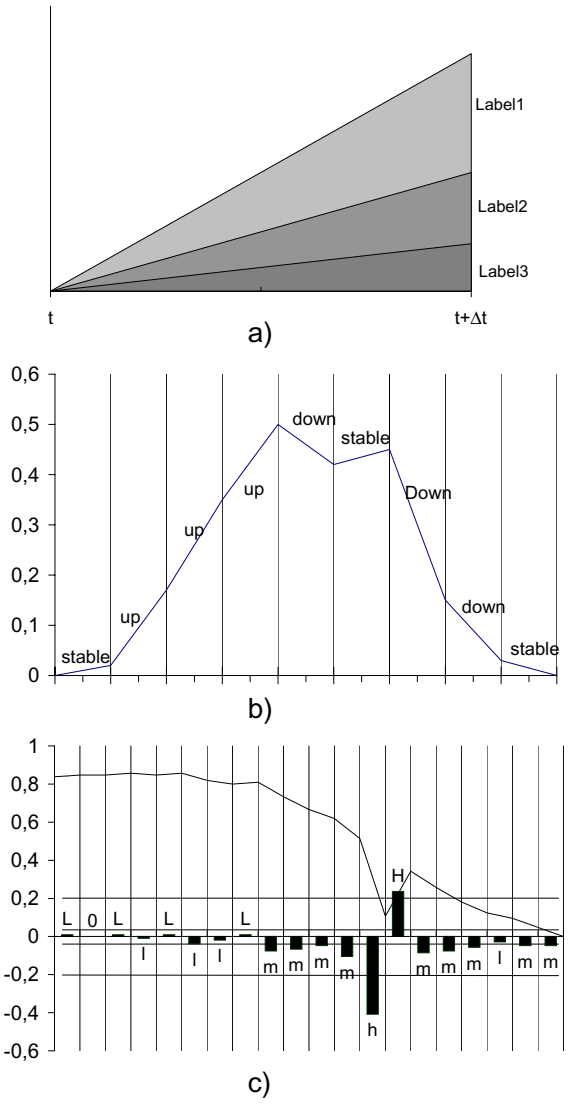

Figure 1: a) Sample of range division. b) Translation. c) Time series, differences and symbols assigned.

The QSI is based on the SDL language and the LCS algorithms. Now we will make a short review of both.

Shape Definition Language (SDL)

This language proposed in (Agrawal *et al.* 95b) 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.

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 1a represents a sample division into three regions of the positive axis.

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 1b shows an example of a translation using the set of symbols $(Down, down, stable, sero, up, Up)$.

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

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&Dancík94).

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.

DTW is an algorithm intensively used in the speech recognition area because it is appropriate to detect similar shapes that are non aligned in the time axis.

Definition of QSI

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.

Our proposed approach is applied in three steps. Let X be a time series, first a normalization of the values of X is performed. Using this normalized series, the difference series is obtained, which is translated into a string. 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.

Let $X = \{x_0, ..., x_T\}$ be a time series, and let $\tilde{X} = \{\tilde{x_0}, ..., \tilde{x_T}\}$ be the normalized temporal series obtained from X.

In (Cuberos *et al.* 02), as follows:

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

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

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

$$d_i = \tilde{x_{i+1}} - \tilde{x_i} \quad (2)$$

This difference series will be used in the labelling step to produce the string of characters corresponding to X.

The labelling process will be studied deeper in this paper in the next section.

Let X, Y be time series where $X = \langle x_0, ..., x_T \rangle$ and $Y = \langle y_0, ..., y_V \rangle$. 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

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

where ∇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 we may find in the same order in both sequences simultaneously, and this value divided by the length of the longest sequence.

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 2: 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 3: Labels distribution

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.

In (Cuberos *et al.* 02), the labelling process was defined based on a parameter δ which is supplied by the experts according to their knowledge about the system and following the table below.

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 column. The last column contains the character assigned to each label.

The proposed alphabet contains three characters for increases and three for decrease ranges, and one additional character for constant range.

This alphabet is used to obtain the string of characters $S_X = \langle c_0, ..., c_{T-1} \rangle$ 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 $X_D = \langle d_0, ..., d_{T-1} \rangle$ assigning to every d_i its character in accordance with the above table.

This translation of the time series into a sequence of symbols abstracts 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 1c shows a normalized curve with their derivative values and the assigned label to each transition between adjacent values. This example has been obtained with $\delta = 5$.

But in this paper the interest about noise sensitivity is joined to the study of other possible labelling strategies. 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.

Noise

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:

$$\epsilon(t) \leq L = \frac{1}{2} \max\{L_i - L_{i-1},\ i = 1, \cdots, k\}\,. \quad (4)$$

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 α order of the set $\{\epsilon(t)\}_{t=1}^{T-1}$ is calculated, and indicated ϵ_α (see figure 4).

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

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

is considered, which does not depend on t.

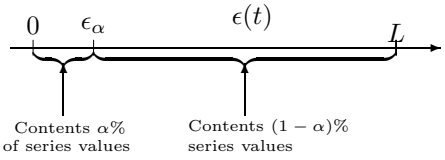

Figure 4: Percentile of series $\epsilon(t)$

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:

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

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 ϵ_α 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_α is defined

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

- This way, we can assign to each labelling a value p_α of the **noise level** endured with a confidence level $1-\alpha$. If p_α 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 α in 5%, and therefore the labelling would have a confidence level for a level error $p_{0.05}$ of 95%.
- If for α high values, the p_α 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.

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 the UCI KDD (Bay99) 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. From this dataset 10 words from the 95 words in the database were selected.

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.

	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 5: Number of label hops in presence of noise

	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 6: Number of label hops in presence of noise cont.

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_α, explained in the previous section, of the set of series in the ASL subset for each labelling scheme in the table below.

α	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 2 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 3.

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.

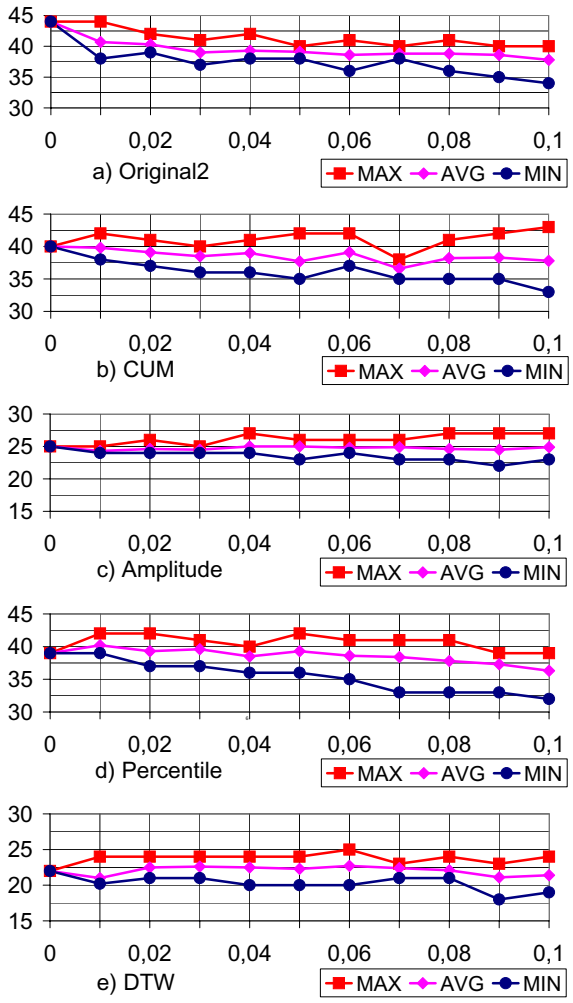

Figure 7: Clusterings success in the presence of noise

In the figures 5 and 6 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 7 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.

Conclusions and Further Work

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 the are no information about the system which originated the time series, the *CUM* method can be used as a first approximation.

Acknowledgments

This work was partially supported by the Spanish Interministerial Committee of Science and Technology by means of the programs DPI2001-4404-E. Moreover, it has been partly supported by the grant ACC-265-TIC-2001 given by the Junta de Andalucía.

References

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

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

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

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

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.

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.

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