# A MEASURE OF QUALITY FOR EVALUATING METHODS OF SEGMENTATION AND EDGE DETECTION[1]

*Barranco-López, Vicente*
Departamento de Electrotecnia y Electrónica. Universidad de Córdoba
Avd. Menéndez Pidal s/n. 14004 Córdoba. Spain
Email: el1balov@uco.es

*Román-Roldán, Ramón and Gómez-Lopera, Juan Francisco*
Departamento de Física Aplicada. Universidad de Granada
Campus Fuente Nueva. 18071 Granada. Spain
Email: {rroman, jfgomez}@ugr.es

*Martínez-Aroza, José*
Departamento de Matemática Aplicada. Universidad de Granada.
Campus Fuente Nueva. 18071. Granada. Spain
Email: jmaroza@ugr.es

## ABSTRACT

A new measure of quality is proposed for evaluating the performance of available methods of image segmentation and edge detection. The technique is intended for the evaluation of low-error results and features objective assessment of discrepancy with the theoretical edge, in tandem with subjective visual evaluation using the neighbourhood and error-interaction criteria. The proposed mathematical model is extremely simple, even from the perspective of computational execution. Encouraging results were obtained for a selection of test images, especially in relation to other recently-proposed and/or currently employed quality measures.

## 1. INTRODUCTION

### 1.1 Quality measures for edge detection and segmentation methods

Authors currently working in the field of low-level image segmentation frequently point to the need for a quality measure that would allow the evaluation and comparison of the variety of procedures available ([1], [5]). Some researchers ([2], [3], [7], [8]) have even stepped outside their usual sphere of interest in an effort to satisfy this demand by proposing new quality measures, as indeed is true of the present authors.

Although the problem is often referred to as one of determining the quality of a segmentation method, most of the algorithms for computing the proposed measures work within a window sliding across the image, thus whether or not the image has been effectively segmented is not taken into account. Only a few methods actually explore the obtained segments [6], most measures being best suited to edge detection. Yet in the design of segmentation quality measures, effective edge detection can prove highly useful. The present proposal is based on this concept.

### 1.2 On existing quality measures

Recent surveys ([2], [7]) have revealed the shortcomings of currently-available measures. For instance, the *merit factor f*, proposed by Pratt [5], sometimes lacks additivity with respect to mistakes. The *expansion factor* $f_e$ is similarly defective. The *discrepancy D* (called $P_e$ in [7]) is insensitive to the location of the mistakes relative to the edge, while the *error probability* $P_e$ [5] ignores errors due to missing edge pixels.

### 1.3 The nature of the proposed measure

The proposed measure is a hybrid of the so-called "*empirical discrepancy*" and "*empirical goodness*" by Zhang [7]. It combines the pixel-by-pixel objective discrepancy between the obtained edges and the theoretical ones with an evaluation of each mistake depending on the assessment by human observers.

The measure belongs to a class that may be called the *low-error model* of quality measures, not intended for excessive or aberrant (not usually found in edge images) errors. In other words, the method should be applied to

[1] This work was supported in part by grants TIC94-535 and MAR-97-0464-C04-02 from the Spanish government.

edge images with few mistakes, and most of these should be close to the theoretical edge.

## 1.4 Definitions

**Mistake:** the discrepancy between the detected edge and the real edge is due to individual discrepancies arising from pixel to pixel, which are here referred to as mistakes. These may be of two kinds:

**Bit:** a mistake due to excess, when a pixel is erroneously defined as an edge pixel.

**Hole:** a mistake due to failure of the method to recognise an edge pixel as such.

# 2. CHARACTERIZING THE MEASURE

In order to implement the initial statement (discrepancy, empirical goodness and low-error) in a mathematical expression for a quality measure *R***,** the following bases are established.

## 2.1 Discrepancy

The theoretical edge must be known so that the mistakes produced can be detected and evaluated one by one. However, the edge must be defined by convention. When two homogeneous regions are touching and the boundary has to be defined as a one-pixel-thick dividing line, the newly-defined edge pixels clearly must be placed in one of the two areas. Two different methods might perfectly locate the desired boundary within an image, but while one method will draw the edge on the darker side, the other will always place it in the lowermost region. The binary images of the edges obtained by the two methods will obviously be different, but both should be classed as error-free since the actual detection was perfect. To prevent this ambiguity in edge definition from influencing the number of mistakes, the theoretical edge may be assumed to lie on the side with the fewest mistakes.

The measure *R* should be expressed as the sum of the values assigned to all mistakes in the edge image. These values must take into account the presence (or not) of other mistakes produced close to the one being evaluated. Therefore, a sum expression for *R* does not imply that it be a merely additive function of single mistakes, nor even an increasing function with respect to the number of mistakes.

## 2.2 Empirical goodness

This feature is based on two criteria for the appearance of a mistake in the edge image.

### *2.2.1 Neighbourhood*

The way a human observer appreciates an edge image depends on the distance of each mistake from what is perceived to be the true edge. However, this is true only for very short distances. In fact, a mistake (bit) produced just next to the edge is assumed to be linked to it, while bits located only two or three pixels away from the edge are no longer perceived thus. Hence in the proposed measure, by contrast with some quality measures which include the distance to the edge as a variable, a mistake value will not depend on distance if this is greater than one or two pixels. The fact that only strict closeness to the edge will taken into account is then an essential and novel criterion that introduces a notable simplification in the mathematical expression for *R*.

### *2.2.2 Mistake interaction*

Also, human appreciation of closely-clustered mistakes is more acute than when the same mistakes are scattered throughout the image. An observer assessing a group of mistakes lying within a small neighbourhood (*V*) will allot them different values (interaction), and these may be extenuating or aggravating.

For *V* centred on a given mistake, if there is one or more mistake of the same type then visual appreciation is generally more severe. Conversely, if the accompanying mistakes are not of the same type and are placed horizontally or vertically, not obliquely, it will appear that the true edge pixels have been only slightly shifted, and the connectivity of the edge is preserved. Hence, the evaluation may be extenuating for different mistakes in *V*, but aggravating for the same type.

## 2.3 The low-error model

The low-error nature of the proposed method means that the evaluation of empirical goodness will be performed within a small window centred on each mistake. Here, the simplest solution was chosen: a 3x3 square evaluation window. In practice, the results justified this decision and any improvement gained using larger window sizes would not appear to merit the extra effort required.

# 3. THE MEASURE OF QUALITY

Following these premises, the mixed measure of quality *R* is defined as the sum of all the (positive) values assigned to mistakes. These very simple values fulfil the above-mentioned empirical goodness criteria.

$$R = \sum_{\text{bits}} \frac{a \cdot (1 + b \cdot n_b)}{1 + p \cdot n_e + i_{bh} \cdot n'_h} + \sum_{\text{holes}} \frac{c \cdot (1 + h \cdot n_h)}{1 + i_{hb} \cdot n'_b} \qquad (1)$$

where the sums extend to bits and holes, each with a different contribution to the measure. Each term corresponds to a mistake, and is determined according to the content of a window *V* centred on the mistake, where the variables are:

$n_b$ = # of bits in *V*, minus the central mistake,

$n_h$ = # of holes in *V*, minus the central mistake,

$n'_b$ = # of bits in direct contact with the central mistake,

$n'_h$ = # of holes in direct contact with the central mistake,

$n_e$ = # of real edge pixels in *V*,

and the coefficients represent:

$a, c$ = the balance between bits and holes, and the scale of the measure,

$b, h$ = the aggravating interaction between same type mistakes,

$p$ = the extenuating proximity to the edge, only for bits,

$i_{bh}, i_{hb}$ = the extenuating interaction between different type mistakes.

The proposed mathematical expression for *R* contains a set of coefficients that have been determined by applying *R* to a training battery of typical error patterns. The training process itself is omitted here for shortness, the final values being

$$a = c = 1, \quad p = 1/3, \quad b = h = 1/2, \quad i_{bh}, i_{hb} = 2 \qquad (2)$$

The possibility remains open of adapting the behaviour of *R* by means of another training process, either because a different set of observers is available, or because a specific application is desired.

## 4. RESULTS AND COMPARISON WITH SEVERAL QUALITY MEASURES

The proposed measure and other available measures were applied to a selected set of edge images provided by segmentation procedures and results were compared in order to assess the good behaviour of *R*.

### 4.1 Test images

Four series of test edge images were selected from a segmentation for comparison of the different methods. By human comparison, each series was ordered according to quality. The images (see Figure 1) were selected in such a way that, in the opinion of the observers, there was no doubt about the correct arrangement. That is, the difference in quality between each two consecutive images in a series was great enough to expect any other qualified observer to agree.

These were the criteria used in selecting the test images:

- They were synthetic images, this being a requirement for the application of the quality measures proposed for this comparison, and most were of the “discrepancy”, or mixed, type.
- They were of the same type as those utilised by other published studies on this subject [4] (new proposed measures, or surveys of measures), in order that the reader may more easily assess the results.
- Bearing in mind that *R* is designed as a hybrid measure, the prior assessment of test images must also be performed using the two criteria, empirical goodness and discrepancy. Since combining both in a subjective appreciation is difficult, they were applied separately. Thus, the first three series contain scattered and grouped mistakes, but no systematic error, whereas part d) in Table 1 refers to images with only continuous errors —displacements, duplications, etc.

The fourth series has not been shown in Figure 1 owing to obviousness. The real edge is a straight vertical line, and the test edges are as follows: 1) a two-pixel-thick vertical line over the edge; 2) a one-pixel-thick vertical line displaced one pixel to the right with respect to the real edge; 3) a three-pixel-thick vertical line centred on the edge; 4) same as 3) but shifted one pixel to the right; 5) same as 1) but shifted one pixel to the right, thus not covering the real edge; and 6) same as 5) but shifted again to the right.

### 4.2 Quality measures applied. Results

Besides *R*, the quality measures $P_e$, *f*, *D*, $f_e$ were applied to the test images. The unnormalized numerical results are in Table 1. The comparison of these results was performed not directly from the numerical values, but from the resulting ordered sequence in each series instead. The following criteria were used:

a) Very close values must be considered as equal; for each measure, an inequality tolerance of 1% of the range was established.

b) Given an ordered sequence *S*, its *deviation degree* ***G(S)*** with respect to the correct order (1,2,3,... by construction) was defined as the minimum number of transpositions between adjacent elements needed to correctly arrange *S*. This choice is justified because it represents the number of individual disorders produced in *S* with respect to the expected order, and this is what we wish to determine, instead of the mere deviation of each element from its correct location in the list.

If, according to a), an ordered sequence *S* contains equalities, then its deviation degree *G(S)* is obtained as the average of $G(S_p)$ for all the sequences $S_p$ differing from S only in a permutation of equal-valued elements. The sequences produced by all the measures are also shown in Table 1, as well as the corresponding deviations *G(S)*. The correct sequence $S_c$ is given in the second column as a reference.

### 4.3 Comments

The results reported in the previous paragraph show excellent behaviour for the proposed measure *R*. In particular it can be said that:

1. *R* gives the minimum deviation in series d. This is particularly important since systematic errors produced in segmentation processes are of this type.
2. Deviation degrees were markedly smaller for *R* than for the remaining measures in series a and b. In series c *R* is the second, very near from the best.
3. It can be seen that measures *D*, $P_e$, *f* and $f_e$ had relatively good behaviour in certain series, but not in all. This corroborates the argument given in the critical introduction at the beginning of this paper.

## 5. CONCLUSIONS

A new measure of quality that allows comparison of edge detection and segmentation procedures has been designed and tested. It is intended to provide an answer to a need felt and expressed by the scientific community working in this field and contributes new procedures without having appropriate quantitative comparison criteria.

The proposed measure has the following characteristics:

- It is a hybrid of empirical discrepancy and empirical goodness, which appears to be the type of measure best suited to the general problem.
- Its mathematical expression is simple and intuitive, also having a high computational efficiency.
- It has shown to have good behaviour, in the sense that it correctly reflects the expected quality order in representative series of test images.
- It is a very versatile measure, since it may be adapted to more specific criteria of visual appreciation. In particular, it contains seven coefficients that could be adapted to the desired application by means of a training process.

## 6. REFERENCES

[1] I.E. Abdou, and W.K. Pratt, “Quantitative Design and Evaluation of Enhancement / Thresholding Edge detectors”, Proceedings of the IEEE, Vol. 67, No 5, May, 1979.

[2] Q. Huang, and B. Dom, “Quantitative Methods of Evaluating Image Segmentation”, IEEE International Conference on Image Processing, 1995.

[3] P.L. Palmer, H. Dabis, and J. Kittler “A Performance Measure for Boundary Detection Algorithms”, Computer Vision and Image Understanding Vol. 63, No 3, May, pp 476-494, 1996.

[4] D.J. Park, K.M. Nam, and R-H. Park, “Edge Detection in Noisy Images Based on the Co-occurrence Matrix”, Pattern Recognition, Vol. 27, No 6, pp 765-775, 1994.

[5] W. Pratt, “Digital Image Processing”, Wiley-Intrescience, 1991.

[6] Y.J. Zhang, J.J., and Gerbrands, “Objective and quantitative segmentation evaluation and comparison”, Signal Processing, No 39, pp 43-54, 1994.

[7] Y.J. Zhang, “A Survey on Evaluation Methods for Image Segmentation”, Pattern Recognition, Vol. 29, No 8, pp 1335-1346, 1996.

[8] Q. Zhu, “Efficient evaluations of edge connectivity and width uniformity”, Image and Vision Computing, No 14, pp 21-34, 1996.

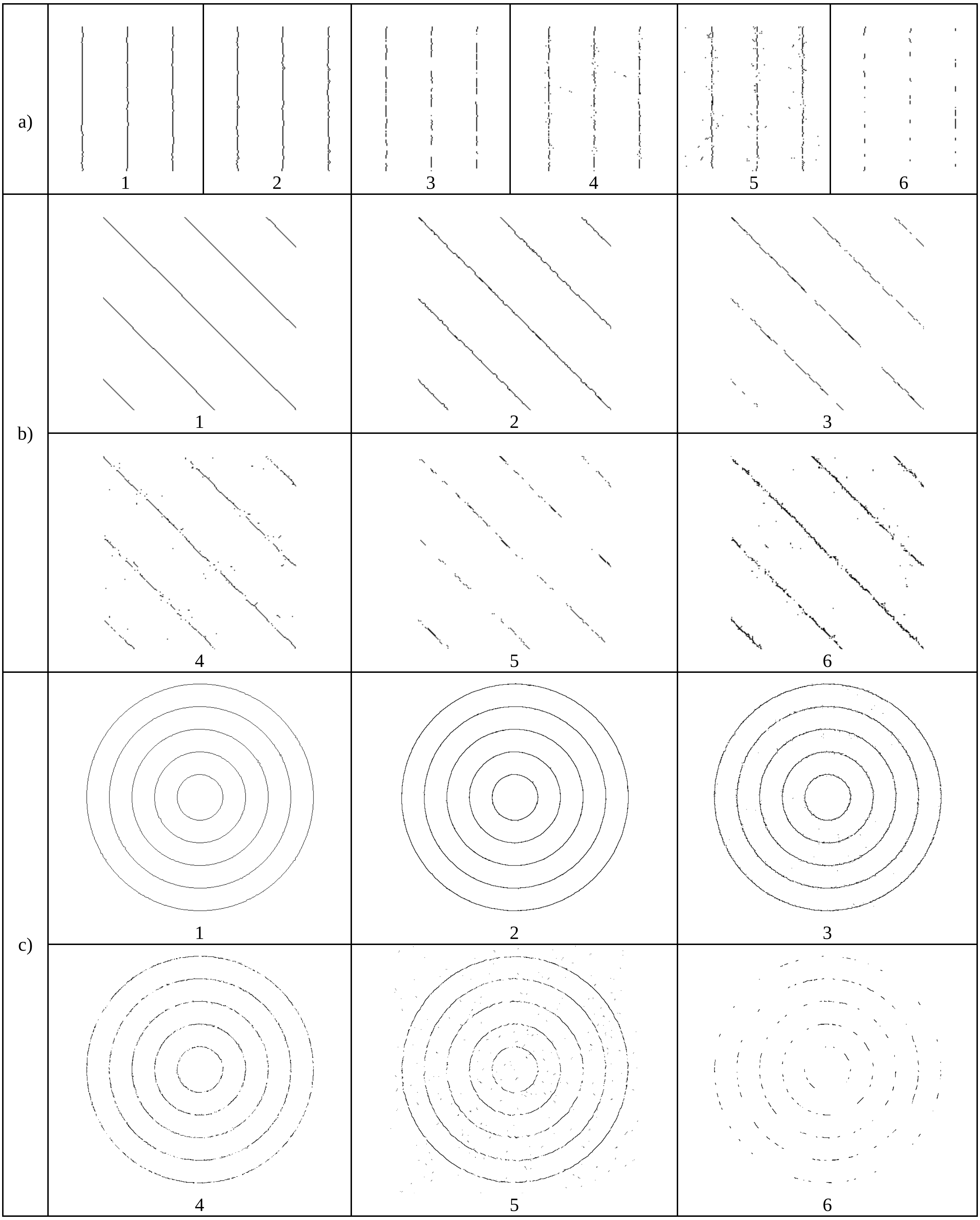

Figure 1 : series of segmented images.

| Serial | $S_c$ | $10*P_e$ | $S_{Pe}$ | $10^2*D$ | $S_D$ | $10*f$ | $S_f$ | $10*f_e$ | $S_{fe}$ | $10^{-2}*R$ | $S_R$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| a) | 1 | 1,54 | 3 | 0,58 | 1 | 0,923 | 1 | 0,500 | $1^*$ | 0,73 | 1 |
| | 2 | 2,54 | $4^*$ | 0,91 | 2 | 0,869 | 2 | 0,473 | 4 | 1,28 | 2 |
| | 3 | 1,21 | 2 | 1,04 | 3 | 0,629 | 5 | 0,500 | $2^*$ | 3,24 | 3 |
| | 4 | 2,54 | $5^*$ | 1,27 | 4 | 0,674 | 4 | 0,389 | 5 | 3,47 | 4 |
| | 5 | 4,65 | 6 | 1,66 | 6 | 0,686 | 3 | 0,296 | 6 | 4,11 | 5 |
| | 6 | 0,31 | 1 | 1,51 | 5 | 0,243 | 6 | 0,500 | $3^*$ | 7,01 | 6 |
| | ***G*** | | **7,5** | | **1** | | **3** | | **5,5** | | **0** |
| b) | 1 | 3,51 | 1 | 0,80 | 1 | 8,26 | 1 | 0,500 | 1 | 1,37 | 1 |
| | 2 | 7,34 | 5 | 1,44 | 3 | 6,89 | 2 | 0,483 | 2 | 3,39 | 2 |
| | 3 | 5,53 | 3 | 1,48 | 4 | 5,43 | 5 | 0,481 | 3 | 5,16 | 3 |
| | 4 | 6,87 | 4 | 1,56 | 5 | 5,83 | 4 | 0,376 | 5 | 5,71 | 4 |
| | 5 | 3,79 | 2 | 1,37 | 2 | 3,72 | 6 | 0,468 | 4 | 7,08 | 5 |
| | 6 | 12,42 | 6 | 1,86 | 6 | 5,91 | 3 | 0,381 | 6 | 9,74 | 6 |
| | ***G*** | | **5** | | **3** | | **4** | | **1** | | **0** |
| c) | 1 | 1,61 | 2 | 0,42 | 1 | 9,19 | 1 | 4,91 | 3 | 5,14 | 1 |
| | 2 | 5,44 | 4 | 0,81 | 2 | 8,16 | 2 | 4,96 | 1 | 19,05 | 2 |
| | 3 | 8,69 | $5^*$ | 1,43 | 5 | 7,15 | 4 | 4,47 | 5 | 39,23 | 4 |
| | 4 | 4,45 | 3 | 1,17 | 3 | 7,74 | 3 | 4,78 | 4 | 35,95 | 3 |
| | 5 | 8,71 | $6^*$ | 2,02 | 6 | 5,70 | 5 | 3,81 | 6 | 72,36 | 6 |
| | 6 | 1,42 | 1 | 1,29 | 4 | 2,58 | 6 | 4,93 | 2 | 68,82 | 5 |
| | ***G*** | | **7,5** | | **3** | | **1** | | **6** | | **2** |
| d) | 1 | 10,0 | $1^*$ | 1,88 | 1 | 7,50 | 1 | 5,00 | $1^*$ | 4,80 | 1 |
| | 2 | 10,0 | $2^*$ | 3,75 | $2^*$ | 5,00 | 4 | 5,00 | $2^*$ | 5,60 | 2 |
| | 3 | 20,0 | $3^{\blacklozenge}$ | 3,75 | $3^*$ | 6,67 | 2 | 5,00 | $3^*$ | 9,60 | 3 |
| | 4 | 20,0 | $4^{\blacklozenge}$ | 3,75 | $4^*$ | 5,67 | 3 | 3,50 | $4^{\blacklozenge}$ | 25,20 | 5 |
| | 5 | 20,0 | $5^{\blacklozenge}$ | 5,63 | $5^{\blacklozenge}$ | 3,50 | 5 | 3,50 | $5^{\blacklozenge}$ | 24,20 | 4 |
| | 6 | 20,0 | $6^{\blacklozenge}$ | 5,63 | $6^{\blacklozenge}$ | 1,50 | 6 | 1,50 | 6 | 43,20 | 6 |
| | ***G*** | | **3,67** | | **2** | | **2** | | **2** | | **1** |

Table 1: numerical results and ordered secuencies

$^*$,$^{\blacklozenge}$: the values in the serie are different less than the 1% of the serie's range.