APPROXIMATIONS FOR THE DISTRIBUTION OF FUZZY SCAN STATISTICS

Laureano Rodríguez Corvea*[1], Gladys Casas Cardoso**[2], Ricardo Grau Abalo**[3], Onel Gomez Vegas*:[4]

*Department of Informatics, Medical Sciences Faculty, Sancti Spíritus, Cuba

** Bioinformatic Group, Computer Sciences Department, Mathematics, Physics and Computer Sciences Faculty, Central University of Las Villas, Cuba

ABSTRACT

The classical temporal scan statistic is used to identify disease clusters. In recent years, this method has become a very popular technique and its field of application has been notably increased. Many bioinformatics problems have been solved with this technique. In this paper a new scan method is proposed: fuzzy scan. Three approximations for the distribution of fuzzy scan statistic are presented and the behaviors of the classic and fuzzy scan techniques are studied with simulated data. Finally, ROC curves are calculated and it is demonstrated the superiority of the fuzzy scan technique compared with the classic scan method.

KEY WORDS: Scan statistic, fuzzy scan, simulating study, ROC curves

MSC: 62P10

RESUMEN

El método scan clásico se utiliza para identificar conglomerados de enfermos. Recientemente se ha convertido en una técnica muy popular y su campo de aplicación se ha incrementado notablemente. Numerosos problemas bioinformáticos se han resuelto con ayuda de esta técnica. En este trabajo se propone un nuevo método: el scan borroso. Se presentan tres aproximaciones para el cálculo de su distribución probabilística y se compara el comportamiento de ambas técnicas: clásica y borrosa, con datos simulados. Para finalizar se calculan las curvas ROC y se demuestra la superioridad de los métodos borrosos comparados con el clásico.

1. INTRODUCTION

The field of Statistics is constantly challenged by new exciting and complicated problems. Vast amounts of data are being generated in many fields, and the job of statistician is to understand "what the data say". This process is called by many authors: "*learning from data*". [5]

The learning problems can be roughly categorized as supervised or unsupervised. In supervised learning, the target is to predict the value of an outcome measure based on a number of input measures. In unsupervised learning, there is not outcome measure, therefore the goal is to describe the associations among a set of input measures. [5]

This paper is related with the mathematical details of an unsupervised statistical technique: the temporal scan statistic. We refer to this method as classic scan technique in the rest of the document. It has become a popular method for the detection of disease clusters, and also for the detection of clusters in other application fields. This can be consulted in several references, for instance: [3, 4, 6, 7, 9].

The paper is structured as follows: next section is dedicated to explain the Classic Scan Technique. Then, the Generalized Scan Method is introduced. Afterwards, another section explains the main ideas of the Fuzzy Scan Technique. Bases of the Simulation Study are also shown and the experiments prove the feasibility of our contribution. Finally, results and conclusions are provided.

[1] corvea@uclv.edu.cu
[2] gladita@uclv.edu.cu
[3] rgrau@uclv.edu.cu
[4] drissp@infomed.sld.cu

2. THE CLASSIC SCAN TECHNIQUE

Suppose we have a set of patients with a specific disease. For each patient we have the day of diagnostic or the day of onset of the disease. Scan statistics are commonly used to investigate the presence of at least one cluster of dates. The method is known in the literature as "moving window analysis", and its target is to scan a small fixed window of length t over the dates, calculating the number of cases for each window. The maximum value is known as the scan statistic, and it will be denoted by η_{max}.

Under some specified homogeneity null hypothesis H_0 on X (Poisson point process), the approach entails the specification of a critical value λ such that $P(\eta \geq \lambda) = \alpha$. If the observed maximum is larger or equal to λ, it is possible to infer that there is at least one significant cluster of dates [11].

The analysis of the scan process has been considered by many authors, including [9-11]. For a few simple models, exact values of the probability are available; but they can not be applied to large databases. Many applications require approximations to this value. In 1982 Naus published an important approximate formula to calculate p-value, based in exact results. [10]

$$p = P^*(\eta, \lambda L, 1/L) = 1 - Q^*(\eta, \lambda L, 1/L) \qquad (1)$$

where $L = T/t$, T represents the total time analyzed and t is the window length.

Q^* can be approximated for any $L>2$ starting from the values $L = 2$ and $L = 3$.

$$Q^*(\eta, \lambda L, 1/L) \approx Q^*(\eta, 2\lambda, 1/2)\left[Q^*(\eta, 3\lambda, 1/3) / Q^*(\eta, 2\lambda, 1/2)\right]^{L-2} \qquad (2)$$

The recipe (2) is easily calculable by using a personal microcomputer [9].

The exact calculation of $Q^*(\eta, 2\lambda, 1/2)$ and $Q^*(\eta, 3\lambda, 1/3)$ is based on a theorem also demonstrated in [8] and whose essence is summarized here:

For w>2, $p_i = e^{-\lambda} \lambda^i / i!$, $F_w = \sum_{i=0}^{w} p_i$, $\lambda > 0$, one has that:

$$Q^*(\eta, 2\lambda, 1/2) = F_{\eta-1}^2 - (\eta - 1) p_\eta \, p_{\eta-2} - (\eta - 1 - \lambda) p_\eta \, F_{\eta-3}$$
$$Q^*(\eta, 3\lambda, 1/3) = F_{\eta-1}^3 - A_1 + A_2 + A_3 - A_4$$

where:

$$A_1 = 2 p_\eta F_{\eta-1}\left((\eta - 1) F_{\eta-2} - \lambda F_{\eta-3}\right)$$
$$A_2 = 0.5 p_\eta^2 \left((\eta - 1)(\eta - 2) F_{\eta-3} - 2(\eta - 2) \lambda F_{\eta-4} + \lambda^2 F_{\eta-5}\right)$$
$$A_3 = \sum_{r=1}^{\eta-1} p_{2\eta-r} F_{r-1}^2$$
$$A_4 = \sum_{r=2}^{w-1} p_{2\eta-r} \, p_r \left((r - 1) F_{r-2} - \lambda F_{r-3}\right)$$

where $F_i = 0$ for all i <0.

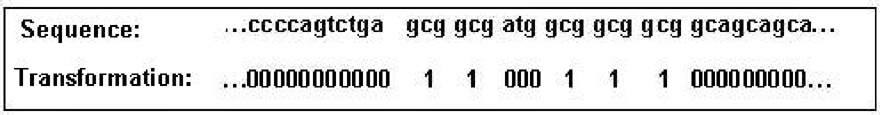
The recipe (2) can be calculated for non integer values of L. This makes the difference from other mathematical expressions that were used with these objectives. Besides, being the approach less restrictive, several authors demonstrated that (1) is more precise than others [2, 10, 15].

As can be seen in [1, 10], all these formulas use probability and cumulative Poisson distribution. We want to focus on the fact that Poisson distribution is discrete, that is, it is defined only for integer values.
In this section a mathematical description of the Classic Scan Technique is presented. The Scan Method emerges to solve an epidemiological problem: dates cluster detection, but it has been successfully applied in many different fields [3, 4, 6, 9].

3. THE GENERALIZED SCAN TECHNIQUE

In the previous section, we commented that the classic scan technique emerges to solve an epidemiologic problem: the dates' cluster detection of non-common disease. But, this method can be modified in order to be applied to other problems [3, 7]. To do this, it is often necessary to transform the input data, i. e. the dates, into a binary sequence. The number one will represent the interest category and the zero value will correspond to the other categories.

For example, the detection of some substring inside a DNA sequence is a classical problem in bioinformatic. Suppose that it is necessary to detect repeats of “gcg”. Then, the “gcg” substring will be replaced by 1, and the other letters will be replaced by 0, see figure 1 [13 y 14].



```
Sequence:          ...ccccagtctga   gcg gcg atg gcg gcg gcg gcagcagca...
Transformation:    ...00000000000    1   1  000  1   1   1  000000000...
```

Figure 1. Binary transformation of a DNA sequence

The objective of the method is the same: a window of fixed length is moved over the whole sequence and the maximum number of "ones" inside the window is calculated. Then, the signification formula of Naus (1) is applied and the p-value is obtained.

4. THE FUZZY SCAN TECHNIQUE

A modification to the classic scan method is proposed. The main idea of the new technique is to change the fixed length moving window by a fuzzy length window with a membership function in both extremes of the interval. [13]

The membership function is a graphical representation of the participation magnitude of each value in the binary sequence.

It associates a weighting with each of the inputs that are processed in the extremes of the moving window. The fuzzy scan technique uses the membership values as weighting factors to determine their influence on the fuzzy output number of cases detected in the interval.

$$\text{Fuzzy Window}_{\mathrm{k}} = \begin{cases} (\mathrm{i} - \mathrm{k} + \mathrm{g} + 1) * \left(\dfrac{\mathrm{s}_{\mathrm{i}}}{g+1} \right) & \mathrm{i} = \mathrm{k} - \mathrm{g}, \ldots, \mathrm{g} \\ \mathrm{s}_{\mathrm{i}} & \mathrm{i} = \mathrm{k}, \ldots, \mathrm{k} + t - 1 \\ (\mathrm{k} + \mathrm{t} + \mathrm{g} - \mathrm{i}) * \left(\dfrac{\mathrm{s}_{\mathrm{i}}}{g+1} \right) & \mathrm{i} = \mathrm{k} + t, \ldots, \mathrm{k} + t + \mathrm{g} - 1 \end{cases}$$

where: $s_1, s_2, \ldots, s_n$: binary sequence,

t : fixed window length,

g : length of the fuzzy part of the new window. We shall call it smoothed.

Figure 2 is included in order to graphically illustrate these ideas.

	Scan Method (k = 4 and t = 5)	
	Classic	**Fuzzy (g =1)**
Binary sequence:	0 1 0 **1 0 1 0 1** 1 0 0 0 1 0 1	0 0 1 **0 1 0 1 0 1 1** 0 0 0 1 0 1 0
Window:	**3**	**0 + 3 + .5**
Scan statistic:	$\eta_{max} = 3$	$\eta^*_{max} = 3.5$

Figure 2. Graphical representation of classic and fuzzy windows

The mathematical formulation of the test is essentially the same: the method scans the data by using a fuzzy moving window. The window is now fuzzy, thus the maximum number of cases reported in a window, (that is, the fuzzy scan statistic η^*_{max}) will be a real number, not an integer as in the classic method. Then, we have a problem with the computation of the p-value, because Poisson distribution is only defined for integer values.

We propose three different variants to calculate the p-value:

1. To approximate the real value to its integer part. Probability and cumulative Poisson functions are frequently used in the p-value calculus. Therefore, error propagation may be not so low. We refer to this method as fuzzy approximation 1.

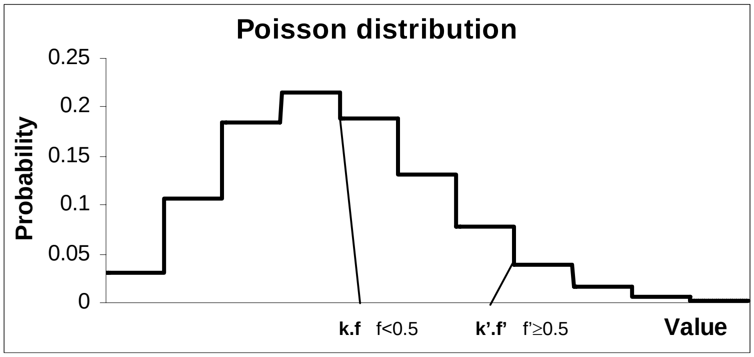


Figure 3: Adjusted Poisson distribution probability, using approximation 1

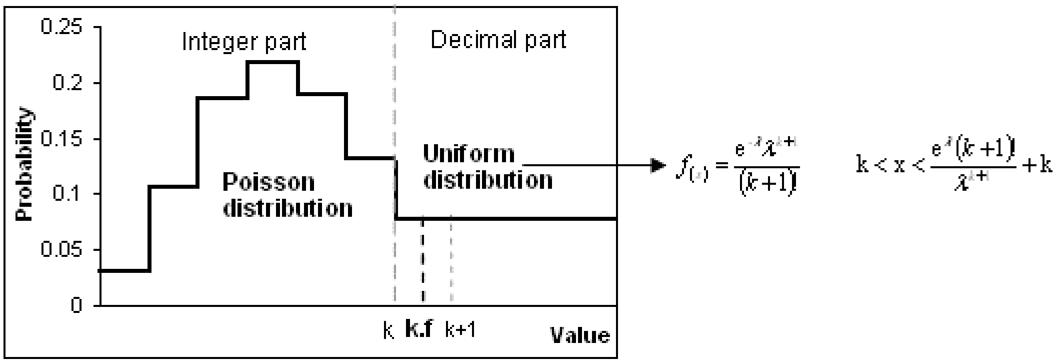
2. To approximate the real value by using the combination of two distributions: Poisson until the lowest integer value, and uniform to estimate the decimal part. We will denote the real number as k.f, where k is the integer part and f the factionary one.

The original formula presented by Naus in [10], need to be updated according to the next expressions because the statistic is now a real number.

- $$P\left[x \le k.f\right] = \sum_{n=0}^{k} \frac{\lambda^n}{n!} e^{-\lambda} + f \left(\frac{e^{-\lambda} \lambda^{k+1}}{(k+1)!} \right)$$
- $$P\left[x = k.f\right] = \frac{e^{-\lambda} \lambda^{k}}{k!} + f \left(\frac{e^{-\lambda} \lambda^{k}}{k!} - \frac{e^{-\lambda} \lambda^{k+1}}{(k+1)!} \right)$$
- $$A3 = \sum_{r=1+f}^{k.f} P\left[x = 2\,k.f - r\right] * P\left[x \le r - 1\right]^2$$

- $$A4 = \sum_{r=2+f}^{k.f} P\left[\,x = 2\,k.f - r\,\right] * P\left[\,x = r\,\right]\left((r-1)P\left[\,x \le r-2\,\right] - \lambda\,P\left[\,x \le r-3\,\right]\right)$$

We refer to this method as fuzzy approximation 2.



$$f_{(x)} = \frac{e^{-\lambda}\lambda^{k+1}}{(k+1)!} \qquad k < x < \frac{e^{\lambda}(k+1)!}{\lambda^{k+1}} + k$$

Figure 4: Adjusted Poisson distribution probability, using approximation 2

3. To approximate the real value by using two interpolating functions. An interpolating function was used to fit the data. In general terms, interpolation is a method of constructing a function from a discrete set of known data points.

In our case, the set of known data is the Poisson probability distribution data or the Poisson cumulative distribution data. With these datasets we build two other functions: an interpolating Poisson probability distribution and an interpolating Poisson cumulative distribution, see Figure 5. We used an interpolating function of degree four. Naus formula should be also updated as in approximation 2.

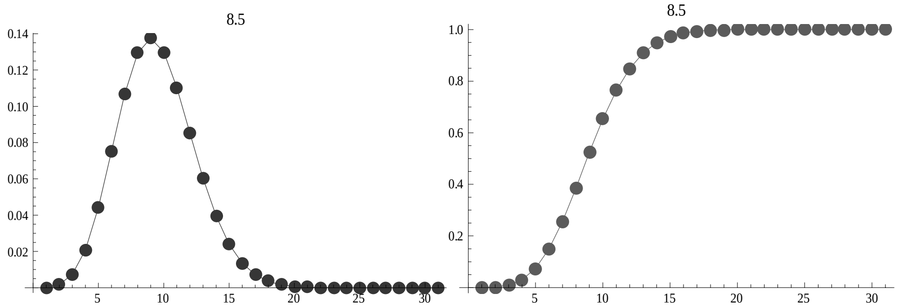


Figure 5: Interpolating function for probability and cumulative Poisson datasets

We refer to this method as fuzzy approximation 3.

Finally, the response of the technique is partitioned into two fuzzy sets with the labels: significative and no significative. There are several membership functions, S-shape was selected because it is continuous and smoothed. Then, each fuzzy set has the next membership function:

No significative:

$$S_{(u,\,0.05,\,0.0625,\,0.075)} = \begin{cases} 0 & u \le 0.05 \\ 2\left(\dfrac{u - 0.05}{0.025}\right)^2 & 0.05 < u < 0.0625 \\ 1 - 2\left(\dfrac{u - 0.075}{0.025}\right)^2 & 0.0625 \le u < 0.075 \\ 1 & u \ge 0.075 \end{cases}$$

Significative:

$$S_{(u,\,0.05,\,0.0625,\,0.075)} = \begin{cases} 1 & u \le 0.05 \\ 1 - 2\left(\dfrac{\text{u - 0.05}}{0.025}\right)^2 & 0.05 < u < 0.0625 \\ 2\left(\dfrac{\text{u - 0.075}}{0.025}\right)^2 & 0.0625 \le u < 0.075 \\ 0 & u \ge 0.075 \end{cases}$$

Figure 6 shows a graphical representation of the two membership functions.

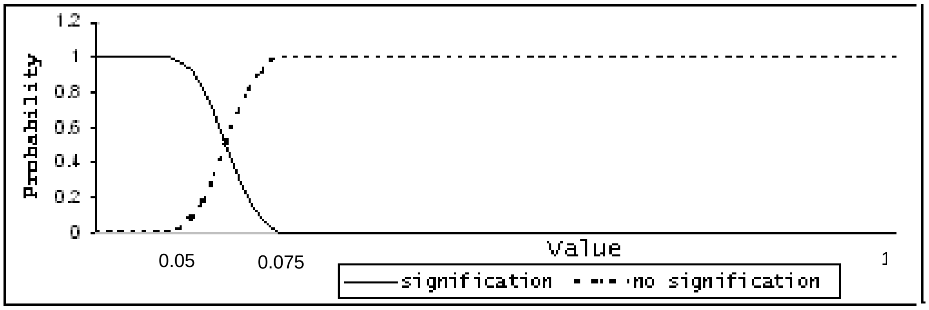


Figure 6: Fuzzy sets membership functions: significant and no significant.

A maximum desfuzzification method is applied in order to obtain a final crisp signification value [8].

6. BASES OF THE SIMULATION STUDY

The purpose of this section is to determine the capacity of the Fuzzy Scan techniques to detect actual clusters and to compare its results with the classic scan technique results.

Scan method capacity for detecting clusters depends on the ability of the user to select the correct value for the moving window length. A bad choice of this parameter can hide true clusters or show putative ones.

Two different kinds of datasets were generated: one for true clusters and other for the wrong ones. 1000 different sequences were generated for each kind of dataset, in order to be used as input data for both scan techniques. The capacity of computing correct responses was calculated using the formula described in the previous section.

6.1 Generating True Clusters Dataset

Binary sequences were generated according to the following principles:

- The first third part of the sequence was generated according to a Bernoulli distribution with parameter 0.2. The number of ones inside the sequence will be small.
- The second part was generated in accordance with another Bernoulli distribution. This time the parameter was higher: 0.8 in order to obtain a significant increment of the number of ones. Doing this, we assure that there is, at least, one cluster.
- The third part of the sequence was generated as the first one.

Besides, in order to characterize the behavior of the test, we generate sequences of different lengths, from 50 to 200 cases.

6.2 Generating False Clusters Dataset

We generate a binary sequence according to a Bernoulli distribution with 0.3 as parameter value. The number of ones will be dispersed in the whole string.

We also generate sequences of different lengths: small (50 cases), median (100 cases) and large (200 cases) in order to determine if the behavior of the methods is related with the sequence length.

6.3 Results and Discussions

Figure 7 graphically shows the results of the classic and the fuzzy scan techniques for true cluster datasets.

In all cases, the continuous line corresponds to the classic scan technique, while the other curves correspond to the three different variants of the fuzzy scan method, see figure 7. Visually there is not difference among the three fuzzy scan curves and there is difference between any fuzzy curve and the continuous curve representing the classic method.

Particularly, the fuzzy techniques improve the results of the classic scan method, for small values of the window's length. The range of true cluster detection capacity is superior in the fuzzy technique, in comparison with the classic one.

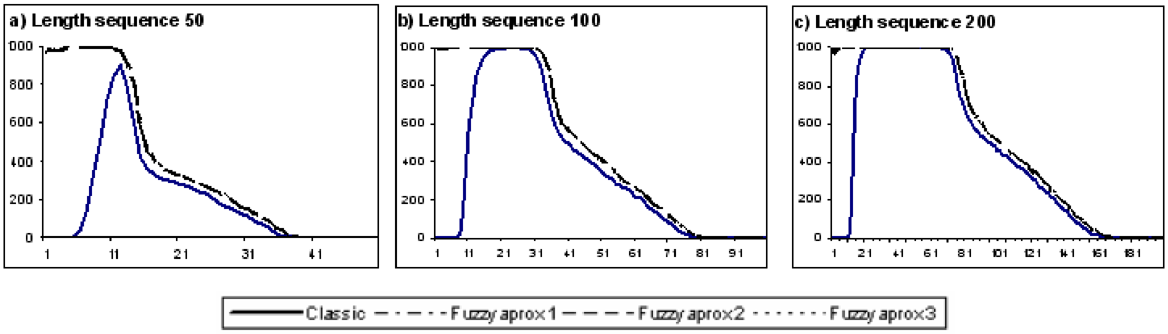


Figure 7. Results of classic and Fuzzy Scan methods for the signification set, using true clusters datasets and length sequences equal to a) 50, b) 100 and c) 200

False databases clusters were also presented as input sequences to the classic and fuzzy techniques. In all cases, the wrong classification percent was approximately zero.

The detection of clusters by applying a scan technique can be seen as a classification problem. Given a sequence of length n, the problem is to detect if there is a cluster or not. Four different classifiers are examined in this paper: the classic scan technique and the three fuzzy variants.

Receiver Operating Characteristics (ROC) graphs is a useful technique for visualizing, organizing and selecting classifiers and visualizing their performance [12]. The area under the ROC curve was calculated for all classifiers. Results are shown in Table 1.

Table 1. Area under the ROC curve

Length sequence	Scan technique			
	Classic	Fuzzy approx 1	Fuzzy approx 2	Fuzzy approx 3
50	0.840	0.856	0.874	0.862
100	0.870	0.908	0.914	0.911
200	0.890	0.911	0.912	0.911

As can be seen, the area under ROC curves is larger in fuzzy scan techniques compared with the equivalent area in the classic scan technique. Besides, if it fixed the sequence length, it is possible to observe that the ROC area is approximately the same for the three fuzzy techniques and smaller in the classic scan method. This fact demonstrates the superiority of the fuzzy scan methods compared with the classic one, when solving the cluster detection problem.

Besides, notice that in all techniques there is a relationship between the sequence length and the window length for obtaining correct results for true clusters. The range of the windows length values detecting true clusters is wider in the fuzzy scan techniques in comparison with the range of the classic method. This fact also demonstrates the superiority of the fuzzy methods.

In particular, fuzzy techniques solve the problem of the small values for the moving window, and increase the detection capacity of larger ones.

However, in general terms, the values of the window length must be smaller than the half of the used sequence length, in order to achieve correct results. But this is not a real constraint. The original method arises in order to scan a small window over a long period of time (or over a large sequence). Generally, bioinformatic applications are based in the analysis of large DNA sequences, then, this constraint will not be a real problem.

RECEIVED MAY 2008
REVISED JANUARY 2009

REFERENCES

[1] CRESSIE, N. (1977): On Some Properties of the Scan Statistic on the Circle and the Line. **Journal o**f Applied Probability, 14: 272-283.

[2] GLAZ, J. (1993): **Approximations for the tail probabilities and moments of the scan statistic.** Statistics **in Medicine**, 12**,**. 1845-1852.

[3] GLAZ, J. and BALAKRISHNAN, N. (1999): **Scan Statistics and Applications**. Statistics for Industry and Technology, ed. Birkauser. 324. Boston.

[4] GLAZ, J., NAUS, J. and WALLENSTEIN, S. (2001): **Scan Statistics**. Springer Verlag. N. York.

[5] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2001): **The elements of statistical learning/** Springer, Montreal..

[6] KULLDORFF, M. (1997): A spatial scan statistic. **Communications in Statistics. Theory and Methods**, 26: 1481–1496.

[7] LANGRAND, C. (2005): Scan Statistics: definición y ejemplos, **Seminario ANY 2005. Universidad Politécnica de Cataluya. España.**

[8] MARTIN DEL BRIO, B. AND SANZ MOLINA, A. (2005): **Redes Neuronales y Sistemas Difusos**. Segunda edición/. Alfaomega, México.

[9] NAUS, J. (1965): Clustering of random points in two *dimensions.* **Biometrika**, 52: 263–267.

[10] NAUS, J. (1982): Approximations for distributions of Scan statistics. **Journal of the American Statistical Association**, 77,. 177-183.

[11] PRIEBE, C., CONROY, J., MARCHETTE, D. and PARK, Y (2005): Scan Statistics on Enron Graphs. **Computational & Mathematical Organization Theory,** 11: . 229-247.

[12] PROVOST, F. and FAWCETT. R. (1998): Analysis and visualizacion of classifier performance: compararison under imprecise class and cost distributions. **IIIrd International Conference on Knowledge Discovery and Data Mining (KDD'97)**. Newport Beach (USA, 1997),. 43-48.

[13] RODRÍGUEZ, L., CASAS, G., GRAU, R. and MARTÍNEZ, Y. (2008): Fuzzy Scan Method to detect Clusters. **International Journal of Biomedical Sciences, © www.waset.org Spring 2008**, 3: 111 -115.

[14] RODRÍGUEZ, L., CASAS, G., GRAU, R. and PUPO, M. (2008): Generalización de dos métodos de detección de conglomerados. Aplicaciones en Bioinformática. **Revista de Matemática: Teoría y Aplicaciones,** 15: . 27 - 40.

[15] SAHU, S., BENDEL, R. and SISON, C. (1993): Effect of relative risk and cluster configuration on the power of the one-dimensional Scan statistics. **Statistics in Medicine**, 12, 1853-1865.