

## Dependence Analysis of the Market Index Using Fuzzy c- Means Algorithm

Renato Aparecido Aguiar  
Engenharia Elétrica  
Centro Universitário da FEI  
São Bernardo do Campo, Brazil  
preraguiar@fei.edu.br

Roberto Moura Sales  
Engenharia Elétrica  
Escola Politécnica da USP  
São Paulo, Brazil  
roberto@lac.usp.br

**Abstract**— In this paper an investigation of the dependence of the return associated to the market index, with respect to some attributes of the financial market is presented. The methodology of analysis is based on the application of a fuzzy clustering means algorithm to some companies of the oil and gas sector and textile sector. The results show that the return of the market index is a variable which depends on some indexes of the financial market. The market indexes of companies in the same sector have the property of describing the behavior of the return of market index. The results are based on quarterly data ranging from 1994 to 2000.

**Keywords**- Pattern Recognition; Attribute Dependence; Fuzzy Clustering Means; Stock Classification.

### I. INTRODUCTION

The prediction of some financial market variables and, more specifically, the prediction of the return of the market index, which may be a large profit or loss for the investors, has been a challenge for statistics. The purpose is to use such prediction to assist an investor in decision making, altering its behavior biased by optimism or pessimism, with regard to the market. Recent studies have suggested that trading strategies guided by forecasts on the direction of the change in the price level are more effective and may generate higher profits [11], [12].

However, considering the existing models for prediction of the market indexes, there arise the questions: on what variables does the market index depend? What economic variables may influence the market index? In this sense, the methodology proposed here aims to investigate, through a clustering algorithm, the dependence of the market index with regard to some financial market variables.

The proposed methodology is based on the technique of pattern recognition using fuzzy c- means algorithm and, the market index used is the stock exchange index of the São Paulo state (Ibovespa). The results suggest that the Ibovespa depends on six attributes of the financial market, consisting of financial ratios for certain sectors of the economy.

This paper is organized as follows: in section II are present the basic concepts relating to pattern recognition using the fuzzy c- means algorithm; section III presents the methodology and main results and section IV presents the conclusions and some comments.

### II. FUZZY C- MEANS ALGORITHM

Among pattern recognition methodologies, whose main objective is to recognize objects with similar features, defining thus a pattern class, the fuzzy c- means (FCM) algorithm is a very useful tool. This pattern class, also known as cluster, is a set of like objects, defined on the basis of some attributes or features relating to the objects [1], [7].

There are fundamentally two techniques of clustering: hierarchical and partitional. The hierarchical clustering technique is mostly applied in biological sciences and represented through a tree, known as dendrogram, that enables the visualization of the clusters and its relationships [7]. On the other hand, the partitional clustering technique is based on the description of the objects through its features, represented by a  $n \times p$  pattern matrix [10]. Each row of this matrix defines an object and each column denotes a feature of the objects. Therefore, given a  $n \times p$  pattern matrix where  $n$  corresponds the number of elements and  $p$  the number of attributes of each element, the aim is to determine a certain number of clusters, such that the objects in a cluster are more likely to each other than to objects in different clusters. This likeliness is defined by distance among the objects, such that the objects more similar are closer and the less similar farer off. For instance, assume two hypothetical clusters obtained through a  $n \times p$  pattern matrix, as shown the Fig. 1.

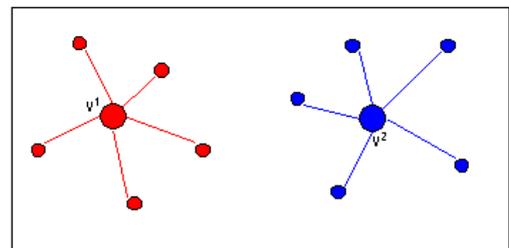


Figure 1. Hypothetical fuzzy clustering

In Fig. 1, each cluster is represented by a center  $v$ , denominated cluster center, surrounded by objects. This center  $v$  is viewed as a reference of likeliness among the

objects; in other words, the objects of a cluster may be represented by center  $v$  of this cluster.

The distance that measures the likeliness between an object and a cluster may be defined by Euclidian distance, or square distance, between the object and the cluster centers, as described in (1).

$$d_{ik} = d(x_k, v_i) = \|x_k - v_i\| = \left[ \sum_{j=1}^p (x_{kj} - v_{ij})^2 \right]^{1/2} \quad (1)$$

The identification of a cluster occurs through determination of a center, such that the sum of the square distances of this center to objects that constitutes the cluster be minimized. The fuzzy c- means (FCM) algorithm, as different from the traditional pattern recognition techniques in which an object belongs or not to a particular cluster, permits an object to belong to more than a cluster simultaneously, however with different membership grades. The FCM as a variation of the partitional technique, also uses as a membership criteria the square distance between the object and the center of the cluster [7].

The FCM technique is derived from fuzzy set theory, originally introduced by Lotfi Zadeh as a mean of handling uncertainty [9]. Fuzzy sets are considered with respect to a base set  $X$  of objects of interest. The essential idea is that each object  $x \in X$  is assigned a membership grade  $\mu(x)$  taking values in  $[0,1]$ , with  $\mu(x)=0$  for non-membership,  $\mu(x)=1$  for total membership and  $0 < \mu(x) < 1$  for partial membership. A fuzzy subset of  $X$  is a subset  $\{(x, \mu(x)) : x \in X\}$  of  $X \times [0,1]$  for some function  $\mu : X \rightarrow [0,1]$  [2].

The algorithm that implements the FCM generates a fuzzy partition by minimizing the objective function as shown in (2) [3], [8].

$$J = \min \sum_{k=1}^n \sum_{i=1}^c (\mu_{ik})^m \|x_k - v_i\|^2 \quad (2)$$

In (2),  $n$  represents the number of objects to be classified (in this paper, companies),  $x_k$  are the vectors that contain the attributes of the objects (financial indexes of the companies),  $c$  indicates the number of clusters,  $v_i$  are the vectors of the cluster centers and  $\mu_{ik}$  the membership grades of each object  $k$  with regard to each cluster  $i$ . The aim is unveil the minimal local of (2) through of the FCM algorithm.

So therefore, differentiating the objective function represented in (2) with respect to  $v_i$  and with respect to  $\mu_{ik}$

and applying the condition  $\sum_{i=1}^c \mu_{ik} = 1$ , are obtained, respectively, (3) and (4) [3], [10].

$$v_i = \frac{1}{\sum_{k=1}^n (\mu_{ik})^m} \sum_{k=1}^n (\mu_{ik})^m x_k \quad i=1, \dots, c \quad (3)$$

$$\mu_{ik} = \frac{\left( \frac{1}{\|x_k - v_i\|^2} \right)^{1/(m-1)}}{\sum_{j=1}^c \left( \frac{1}{\|x_k - v_j\|^2} \right)^{1/(m-1)}} \quad k=1, \dots, n \quad (4)$$

The FCM algorithm is executed through the following steps [10]:

**Step 1:** initialize a membership matrix, such that

$$\sum_{i=1}^c \mu_{ik} = 1.$$

**Step 2:** calculate the cluster centers through (3).

**Step 3:** calculate, through (4), the new membership matrix using the center vectors obtained in the step 2.

**Step 4:** repeats steps 2 and 3 until the value of objective function stops decreasing, according to precision adopted.

### III. METHODOLOGY AND RESULTS

In this section, the methodology employed for the analysis of dependence of the market index with regard to some market variables is introduced. This methodology is based in the pattern recognition technique using FCM algorithm and the data or features utilized for dependence analysis are financial indexes of public companies, including some return indexes related to stock evaluation, profitability and debt. These indexes have been collected every trimester from the Economatica data base [4], between the 4<sup>th</sup> trimester/1994 and the 3<sup>rd</sup> trimester/2000.

For the development of this study various indexes related to liquidity, profitability, debt and stock evaluation were tested. The group of indexes that produced most significant results were : Net Margin, Return On Operating Assets-ROOA, Return On Equity-ROE, Debt/Equity, Price/Earnings per share, Price/Book value per share [5], [6].

The FCM algorithm acts as a "filter", separating the stocks in two clusters. The analysis is based on quarterly information published by companies, collected from the 4<sup>th</sup> trimester of 1994 through the 3<sup>rd</sup> trimester of 2000. In each quarter  $t$  was applied the FCM algorithm in a  $n \times p$  pattern matrix, in which each row corresponds a company and each column to financial indexes of the companies. Two clusters have been obtained and the average financial return that each cluster produces at the end of the trimester  $t + 1$  is calculated according to (3), where  $V_{t-1}$  e  $V_t$  are the values of the stocks or of the Ibovespa in  $t - 1$  e  $t$ , respectively.

$$R_t = \frac{V_t - V_{t-1}}{V_{t-1}} \times 100[\%] \quad (5)$$

The cluster with larger average financial return is called winner and the one with smaller average financial return is called lose. In this case, the classification of the groups as winner or lose has been possible only at the end of the trimester  $t + 1$ .

As results obtained two clusters (or patterns) and, thereon, was calculated the mean financial return produced by each cluster in the trimester  $t + 1$ , the cluster that yielded a higher mean financial return was denominated winner cluster and the cluster with least mean financial return was denominated lose cluster. Thereafter, the mean financial return produced by the winner cluster was compared with the return generate by Ibovespa in the same period. As an example, was describe below what was done for the 1<sup>st</sup> quarter of 1997 for companies of the oil and gas sector. Applying the FCM algorithm in a  $15 \times 7$  pattern matrix ( $n = 15$  companies and  $p = 7$  financial indexes) corresponding to 1<sup>st</sup> quarter of 1997, two clusters represented by the membership matrix showed in table I were obtained. Table II shows the financial returns of each company and the financial return of Ibovespa, in the 2<sup>nd</sup> quarter of 1997.

In this paper, a validity measure for fuzzy clustering was applied for validate the choice of two cluster for the data set used, according to (4). In (4),  $F$  is inversely proportional to the overall average overlap between pairs of fuzzy subsets and maximum value of  $F$  is assumed to produce valid clustering of the data set [13], [14].

$$F = \frac{1}{n} \sum_{i=1}^c \sum_{j=1}^n (\mu_{ij})^2 \quad (6)$$

The higher value of  $F$  was obtained for partition with two clusters, resulting  $0,65 \leq F \leq 0,78$  in the period ranging from the 4<sup>th</sup> trimester/1994 until the 4<sup>th</sup> trimester/1997.

It is emphasized that, in particular, there is no membership sharing between any pairs of fuzzy clusters if  $F = 1$ . Thus, is valid to separate the data set used here in two clusters, using fuzzy c-means algorithm.

TABLE I. MEMBERSHIP MATRIX-1<sup>o</sup> QUARTER/1997

Companies		Membership Grades	
		$\mu_1$	$\mu_2$
Ciquine	a	0,8205	0,1795
Copesul	b	0,1085	0,8915
Ipiranga Dist	c	0,0442	0,9558
Ipiranga Pet	d	0,0177	0,9823
Ipiranga Ref	e	0,0684	0,9316
Oxiteno	f	0,0731	0,9269
Petrobras BR	g	0,2398	0,7602
Petrobras	h	0,6283	0,3717
Petroflex	i	0,9909	0,0091
Petroq.Uniao	j	0,9343	0,0657

Polialden	k	0,0148	0,9852
Politeno	l	0,2940	0,7060
Supergasbras	m	0,5904	0,4096
Trikem	n	0,8070	0,1930
Unipar	o	0,5361	0,4639

TABLE II. FINANCE RETURN-2<sup>o</sup> QUARTER/1997

COMPANIES	FINANCIAL Return
Ibovespa	36,77
Ciquine	2,08
Copesul	-6,01
Ipiranga Dist	4,22
Ipiranga Pet	2,59
Ipiranga Ref	29,05
Oxiteno PN	66,48
Petrobras BR	54,34
Petrobras	39,81
Petroflex	32,32
Petroq.Uniao	21,20
Polialden	49,27
Politeno	64,26
Supergasbras	5,10
Trikem	-26,18
Unipar	31,98

The table I indicates eight companies with higher membership grades in cluster 2 (b, c, d, e, f, g, k, l) and seven companies in cluster 1. Table II shows the financial return of each asset in the 2<sup>nd</sup> quarter of 1997. The mean financial return of the each cluster in this period was calculated: for cluster 2 (33,03%) and for cluster 1 (15,19%). The t-statistics test of the difference between the means for small samples is meaningful to level of 5%. So, cluster 2 obtained in the 1<sup>st</sup> quarter, yielding in the 2<sup>nd</sup> quarter of 1997 a higher mean financial return was denominated the winner cluster. Moreover, according to table II, it may be viewed that this mean financial return is closer of the return of 36,77%, generated by Ibovespa in the 2<sup>nd</sup> quarter of 1997.

The same procedure was done for all quarters, from the 4<sup>th</sup> quarter of 1994 through the 3<sup>rd</sup> quarter of 2000. Fig. 2 presents a comparison of the returns generate by winner clusters and returns generated by Ibovespa in all quarters from the 4<sup>th</sup> quarter of 1994 until the 3<sup>rd</sup> quarter of 2000 for companies of the oil/petrochemical sector.

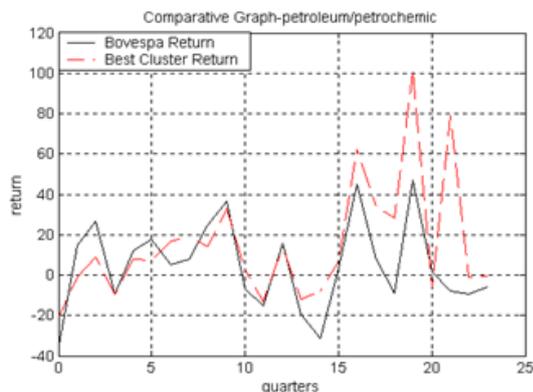


Figure 2. Comparative Graphs of Quarterly Financial Returns

The same methodology was applied for companies in the textile sector. The results confirm the same trend as compared to the petroleum/petrochemical sector. Fig. 3 shows the comparison between the average return of the winner cluster (textile sector) and the return of the Bovespa index in every quarter since 4<sup>th</sup> quarter of 1994 until the 3<sup>rd</sup> quarter of 2000.

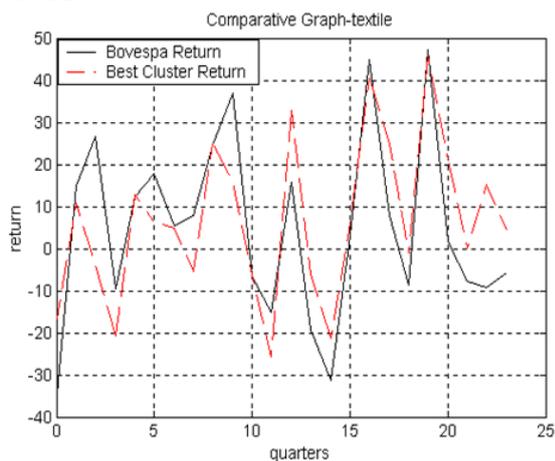


Figure 3. Comparative Graphs of Quarterly Financial returns

Note that in the Fig. 2 and Fig. 3 the average quarterly return of the winner cluster follows closely the Ibovespa. An increase in the return of the winner cluster represents an increase in the return of the Ibovespa. Thus, the results suggest that there is a dependency of the Ibovespa with respect to the indexes used here. In other words, the set of financial ratios effectively used has the property of describing the behavior of the stock exchange index of São Paulo (Ibovespa).

#### IV. CONCLUSION

This study investigates the feasibility of patterns identification for companies through the fuzzy clustering technique, using the FCM algorithm. These patterns were determined from the quarterly financial indexes of the companies' financial statements.

The results obtained from the petroleum/petrochemical and textile sectors show that the cluster denominated winner, that is, the cluster that contains the more efficient companies, owns a mean financial return superior to the lose cluster and, moreover, the profitability of the stocks belonging to the winner cluster follows the same tendencies and sometimes overcomes the profitability of the Ibovespa in the following quarter.

The results obtained show that the financial return of the cluster called winner, has the property of modelling the behavior of the return of the market index (Ibovespa). In other words, has the key feature to indicate the tendency of the Ibovespa. Thus, the financial return of Ibovespa is a dependent variable of the indexes used in this paper, indicating that these indexes can be used as predictors for the Ibovespa.

#### REFERENCES

- [1] J. Bezdek and S.K. Pal. *Fuzzy Models for Pattern Recognition*. IEEE, pp. 88-94, 1992.
- [2] P. Diamond and P. Kloeden. *Metric spaces of fuzzy sets theory and applications*, 1994.
- [3] J.C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybernetics*, V. 3, N. 3, 1973, pp. 32-57.
- [4] Economática Ltda (2002). *Support Software for Investors*, [www.economatica.com.br](http://www.economatica.com.br), accessed on 2002
- [5] F. Fabozzi, F. Modigliani, and M. Ferri. *Foundations of financial markets and institutions*, New Jersey: Prentice Hall, 1994.
- [6] L.J. Gitman. *Princípios de Administração Financeira*. Harper & Row do Brasil, 1994.
- [7] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. New Jersey: Prentice Hall, 1988.
- [8] J. C. Bezdek. "Pattern Recognition with Fuzzy Objective Function Algorithm". *Plenum Press, New York and London*, 1981
- [9] L. A. Zadeh. "Fuzzy sets". *Information and Control*, vol. 8, p.338-353, 1965.
- [10] H. J. Zimmermann. "Fuzzy Set Theory and its Application". *Kluwer Academic, Boston*, 1996
- [11] M.T. Leug, H. Daouk and A. Chen. Forecasting Stock Indices: a comparison of classification and level estimation models. *International Journal of Forecasting*, v. 16, p. 173-190, 2000.
- [12] E. Fama and K. French. The Cross-Section of Expected Stock Returns. *Journal of Finance*, v. 47, p. 427-465, 1992.
- [13] J. C. Bezdek. "Numerical Taxonomy with Fuzzy Sets". *J. Math. Biol.*, vol. 1, p. 57-71, 1974
- [14] X. L. Xie and G. Beni. "A Validity Measure for Fuzzy Clustering". *IEEE Trans. Pattern Anal. Machine Intell.*, v. PAMI-13, p. 841-847, 1991.