FUZZY DATA ANALYSIS: A METHODOLOGICAL APPROACH

Alain Couturier [1)], Bernard Fioleau[2)]

[1)] CRGNA Nantes France, e-mail: *alaine.couturier@laposte.net*
[2)] Institut d'Administration des Entreprises Nantes France, e-mail: *Bernard.Fioleau@iae.univ-nantes.fr*

Abstract*:* This study aims at solving the problems usually created by the heterogeneity, the changing nature and the inaccuracy of data. In order to achieve this, we will process the latter inside the expertons theory, and therefore, unify them.

1. Introduction

The classical methods to analyze PCA type data (or of the type of component analysis,…) are based on the hypothesis of homogeneity and criteria independency as well as on hypothesis of lack of non linear link. Such hypothesis are usually non validated. From a theorical point of view, some treatments (centering, reduction, re-coding,…) are barely acceptable, even though they provide satisfying results in practice. The inaccuracy of measurements is not really treated.

In this study, we propose to process the heterogeneity, the changing pattern in time of accuracy of data in the context of expertons theory. For a criterion C, an experton is defined as a couple $(1\text{-}G_{inf}(x), 1\text{-}G_{sup}(x))$, where $G_{inf}(x)$ and $G_{sup}(x)$ are the cumulative functions of the lower and upper limits of the random variable X which vary in the interval of [0,1], identifying experts' notations.

Data are tri dimensional (individuals, criteria, time or place). At first, we present the model which enables to analyze the evolution of criteria along the time. Afterwards, we implement such model with a sample of 154 companies which have been observed between 1997-2001.

2. The basic theorical model

To each period k, where k varies from 1 to q, we associate a suite $[x_1, ..,x_i ,..,x_n]$, where x_i the value that we measure on the individual i. We ask the expert of domain to note every individual. To make the operation easier, he must also define three specifics break even points x_a, x_b and x_m according to the following rules:

1) each value inferior or equal to x_a is given a mark inferior or equal to 0.1

2) each value superior or equal to x_b is given a mark superior or equal to 0.9

3) each value equal to x_m is given a mark which is comprised between 0.5 - u = 0.5 + v .

The values 0.1 and 0.9 correspond to almost wrong and almost right on a ten points scale wrong to right. In this study, u = v. Among the potential shapes for the function, we choose a sigmoïd shape:

$$\frac{1}{1+e^{-(ax+b-u)}}$$

For each period, the process follows three steps:

1) Construction of the coding function and of the lower and upper limits functions. Such step consists in identifying the coefficients and determining the coefficients a and u.

2) Construction for each individual x of the interval $] F_{inf}(x) , F_{sup}(x) [$.

3) Construction of two random variables Xinf et Xsup with a support {0.05, 0.15, ... , 0.85, 0.95}, centers pf the classes associated into ten groups of steps of 0.10.

We therefore deduct the experton.

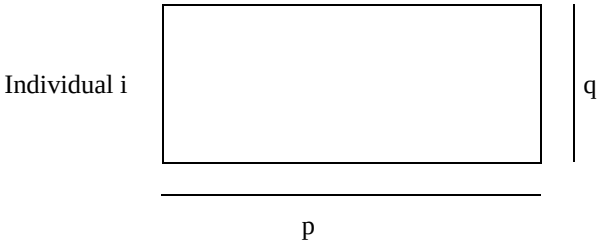
The treatment process can be summarize as following:

$$\text{Let D} = \underset{n}{T}^{i} * \underset{p}{C}^{j} * \underset{q}{T}^{k}$$

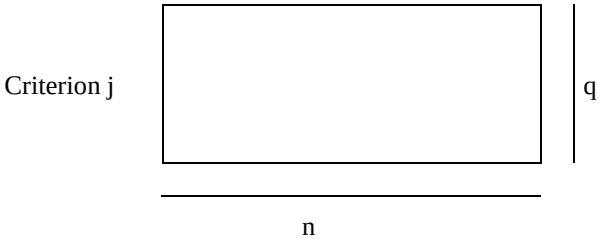
D_k is a table of n individuals and p criteria at the time k.

We must evaluate the coding function p*q of sigmoïd type. If the time has no influence on the evolution of the criteria, we must only evaluate p functions.

1) Analysis of the individuals. At each individual is associated a probability function:



2) Criteria analysis. At each criterion is associated a probability function:



Note: For the probability functions, we use a scalar product.

3. Implementation of the model, results and comments

3.1. Sampling

The approach described at point 2 has been implemented with a sectorial sample of companies. This sample comprises 154 companies specialized in retailing food.

The financial data concern the period from 1997 to 2001. Fifteen ratios describe various aspects of each company's situation, and their evolution over the last five years. The table 1 indicates the fifteen chosen ratios.

1 - $\frac{\text{Net Worth + Long Term Liabilities}}{\text{Fixed Assets}}$	2 - $\frac{\text{Long Term Liabilities}}{\text{Net Worth + Long Term Liabilities}}$
3 - $\frac{\text{Net Worth}}{\text{Total Assets}}$	4 - $\frac{\text{Quick Assets}}{\text{Current Liabilities}}$
5 - $\frac{\text{Inventory}}{\text{Cost of Sales}}$ *360	6 - $\frac{\text{Accounts Receivable}}{\text{Sales inc VAT}}$ *360
7 - $\frac{\text{Suppliers}}{\text{Cost of Goods inc VAT}}$ *360	8 - $\frac{\text{Interests}}{\text{Sales}}$
9 - $\frac{\text{Long-Term Debt}}{\text{Long-Term Debt + Owned Capital}}$	10 - $\frac{\text{Cash Flow}}{\text{Sales}}$
11 - $\frac{\text{Added Value}}{\text{Manpower}}$	12 - $\frac{\text{Added Value}}{\text{Fixed Assets}}$
13 - $\frac{\text{Added Value}}{\text{Sales}}$	14 - $\frac{\text{Current Earnings}}{\text{Total Assets}}$

15 - $\frac{\text{Net Profit}}{\text{Net Worth}}$	

Table 1: Ratios

One can notice that some ratios measure the economic profitability and gross profit of the companies. The company profitability is calculated excluding fiscal effects and nonrecurring events. Others ratios measure the productivity of labor and capital respectivement, the structural characteristics of the activity of the firms and the nature, level and structure of the debt of the companies in the sample.

The fact that all the companies in the sample come from the same sector should guarantee a broad homogeneity in economic activity. But for most of the ratios a very high deviation can be shown between the highest and lowest values observed. It should be noted however that these extremes are reported for only a very small number of companies, usually one or two, and that these companies are different from one ratio to another. Careful analysis shows that, whatever the year, around twenty companies show one ratio with a value different by more than three standard deviations from the average value.

3.2. Results and comments

Two types of analysis have been carried out about these tri dimensional data: one dealing with the criteria, the other one dealing with all the companies of the sample.

Regarding the criteria, the analysis enables to provide a classification of the fifteen criteria over the all period of observation. It emphasizes in particular, an opposition among three groups of three criteria each:

- group 1 includes the ratios 2, 5, 6
- group 2 includes the ratios 7, 9, 10
- group 3 includes the ratios 8, 11, 12

The information provided by each of those ratios can be considered as similar and allows a possible reduction of the number of ratios to be taken in consideration for a classification. Thus, the ratios 11 and 12 measuring respectively the economic efficiency and the productivity can be used individually without reducing the relevance of the classification. The same remark can be applied to the ratios 5 and 6 both defining the operating conditions of the companies.

The second analysis, which has been carried out on the 154 companies of the sample, enables to obtain a classification of all these companies based on the evolution of their ratios over a five year period.

The results first show, on each end of the distribution , that a few companies cannot fit in any group (outliers). On a scale rating from –0,11 to +0,38, the values associated to 149 companies are comprise between –0,09 et + 0,19 ; only four companies go far beyond these limits.

Several typologies could be proposed. We chose a three parts classification. In each part we have selected ten companies which exhibit a similar behavior. See attached in the appendix, the averages calculated for each of the fifteen ratios used. Moreover, each group has a very strong homogeneity. There is obviously a strong opposition between groups 1 and 3 on almost all criteria. Group 2 is somewhere between grou 1 and 3. The main features of the companies of the group one (the most powerful) are as follows:

- High productivity an profitability,
- Moderate financial debt but high global debt,
- Satisfactory investment financing,
- High auto financing.

If we take into account all the above characteristics, we can observe that companies of the group three and of the group one stand on opposite sides of the spectrum, and that companies of the last group stand in between.

4 Conclusion

The results of the experiments enable to validate the approach proposed by the theorical model, which has been proposed in the first part of this article.

The initial data having been homogenized prior to their processing, other classifications fuzzy or not fuzzy could be equally carried out. Besides, among possible extensions, this approach could also enable to meas-

ure the degree of implication of some criteria: on of the main difficulties for using those ratios in terms of classification is their high degree of correlation.

References

[1]. Volle .: "*Analyse des données*" (1993) Economica Paris.
[2]. Dazy F., Le Barzic J.F: "*L'analyse des données évolutives. Méthodes et Applications*" (1996) Editions Tecnip Paris.
[3]. Gil Aluja J: "*Elementos par una teoria de la decision en incertidumbre*" (1999) Milladoiro Santiago de Compostella.
[4]. Kaufman A: "*Les expertons*" (1987) Hermes Paris.
[5]. Lefebvre J: "*Introduction aux analyses statistiques multidimensionnelles*" (1976) Masson Paris
[6]. Couturier A, Fioleau B: "A fuzzy classification method for time-variable data" *AMSE Special issue SSIT 98* June 1999.

Average of the three groups

Ratios	Group 1	Group 2	Group 3
1	1,72	1,19	0,89
2	12,77	13,01	17,64
3	53,42	30,56	10,32
4	1,85	0,93	0,71
5	127,87	29,44	23,44
6	67,69	60,26	36,36
7	78,28	59,61	60,62
8	1,68	0,96	0,83
9	143,15	125,96	123,26
10	12,2	5,02	-0,31
11	96,62	79,32	42,06
12	1,01	0,74	1,00
13	38,54	22,53	13,85
14	0,11	0,07	0,01
15	0,16	0,13	0,46