The identification of "fuzzy profiles" through the *c*-means clustering

Silvestro Montrone and Paola Perchinunno*
Dept of business and law studies,
University of Bari, Italy
E-mail: silvestro.montrone@uniba.it
E-mail: paola.perchinunno@uniba.it
*corresponding author

Abstract: The numerous concepts of socio-economic hardship are, furthermore, attributable to a traditional distinction between absolute and relative conditions of hardship. The options of scientific research were therefore oriented towards the establishment of a multi-dimensional approach, sometimes abandoning dichotomous logic in order to arrive at fuzzy classifications in which each unit belongs and, at the same time, does not belong, to a category. A multidimensional index that considers hardship as the overall condition of being disadvantaged and deprived seems the most appropriate in view of the socio-economic differential analysis of demographic phenomena. The approach used in this work to synthesize and measure the conditions of the hardship of a population is based on a clustering procedure (Fuzzy *c*-means) aimed at outlining various *not defined a priori profiles*, which should be assigned to each family with different socio-economic behaviours. In comparison with conventional methods, this clustering method allows a set of data to belong not only to a main cluster but also to two or more clusters with "fuzzy profiles".

Keywords: fuzzy logic; hardship.

Biographical notes: Silvestro Montrone is Full Professor of Statistics at the University of Bari. The research topics are mainly oriented to solving problems of data mining. He is the coordinator of the PhD program in Statistics at the University of Bari and the University "Our Lady of Good Counsel" in Tirana and is author of several contributions both nationally and internationally.

Paola Perchinunno is a Researcher of Statistics at the University of Bari Her themes mainly focus on data integration techniques and fuzzy logic. She is Chair and member of the organizing committee of the session "Econometrics and Multidimensional evaluation in Urban Environment" (EMEUE) of International Conference on Computational Science and its Applications (ICCSA).

This paper is a revised and expanded version of a paper entitled 'The Lifestyles of Families through Fuzzy C-Means Clustering' presented at ICCSA 2014 (International Conference on Computational Science and Its Applications), Session: Econometrics and Multidimensional Evaluation in Urban Environment (EMEUE), Guimaraes, Portugal, 30 June – 3 July 2014.

The contribution is the result of joint reflections by the authors, with the following contributions attributed to S. Montrone (Sections 1, 2 and 5) and to P. Perchinunno (Sections 3 and 4).

1 Introduction

The objective of this report is the individuation of different profiles, not defined a priori, of each family behaviors with socio-economic specific.

In literature, poverty, concerning its economic nature, is usually defined as an insufficiency of the resources necessary to guarantee a high level of well-being with respect to certain predefined standards. There is a general agreement that evaluating poverty means measuring the economic resources of individual families with respect to the economic resources of other families. The use of monetary variability (in terms of consumption and income) is based on the implicit assumption of equivalence between available economic resources and well-being. Such minimum levels of well-being may be expressed in terms of being absolute or relative. A transversal approach is therefore proposed as an alternative to the above, considered as subjective, through which the poor are defined as those who identify themselves as such, even if this identification is revealed as a result of the comparison that they operate with the rest of society in terms of perceived wellbeing. It is, therefore, the "perception" an individual may have of their own condition that allows for the identification of the measure of poverty to a far greater degree than the assessment of external observers would allow.

The presence of a varied range of definitions on the theme of poverty underlines the necessity of no longer relying on a single indicator, but on a group of indicators which are useful in the definition of living conditions of various subjects (multidimensional approach). In this context the adoption of a fuzzy numbers theory, introduced by L. A. Zadeh [1,2], is considered as valuable and allows the intrinsic complexity of the phenomenon under investigation to be adequately taken into account. Indeed, Zadeh underlined that nature frequently does not present us with a set composed of clearly separate objects, to which it is possible to apply classical principals of set theory such as that of the principle of non-contradiction or the excluded middle; he thus introduced the concept of the degree of membership which, in opposition to classical theory (according to which a specific property may be proved as either true or false) also allows for possible intermediate values of veracity. The principal advantage of fuzzy logic lies precisely in its ability to align itself with human interpretation and, in the case in point, allows for the rejection of the "rich/poor" dichotomy in order to take into account the variety of levels which exist between the two extreme conditions (marked material hardship and high-level wellbeing).

The approach used in this work to synthesize and measure the conditions of the hardship of a population is based on a clustering procedure (Fuzzy c-means) aimed at outlining various not defined a priori profiles, which should be assigned to each family with different socio-economic behaviors. In comparison with conventional methods, this clustering method allows a set of data to belong not only to a main cluster but also to two or more clusters with "fuzzy" profiles.

2 The multidimensional approach for the construction of indicators

2.1 The data source

The data source used in order to construct indicators of socio-economic hardship is derived of the Family Lifestyles survey conducted by the University of Bari "A. Moro" (December 2012 - January 2013). The "Family Lifestyles survey" collected significant information on income, spending behavior, and on the use of financial loans by families with children, resident in the metropolitan city of Bari.

The objective of the survey, carried out by the University of Bari was that of analyzing issues associated with the measurement of socio-economic hardship created by the difficulty of attributing a single and generally agreed definition.

More specifically, the project aims to identify the lifestyles of *young families* in the urban area of Bari, as it is believed that this type of household can feel the economic crisis with greater attention.

The survey was addressed to the "young" couples with children in preschool and school age, by the distributing of questionnaires at the Educational Institutions that have been identified by sampling to be able to represent the urban area of the city of Bari. The technical instrument to collect information was a questionnaire to outline the social profile of a citizen through a variety of resources and information about the financial difficulties, the well-being arising from the possession of goods, the housing situation, level of education and occupation. 4420 completed questionnaires were returned.

A methodology based on objective variables (those resources actually available to families) was accompanied by subjective measurements based on the perception of the family in terms of its social and economic condition.

2.2 The construction of the indicators of hardship

The profound economic and social transformations witnessed in recent decades have underlined the necessity of analyzing the phenomenon of hardship in terms of its multiple facets. The identification of the poor as a subject living on the edge of society (as, for example, the homeless) appears to have already been superseded in favor of a growing academic focus placed on general context, including both economic hardship and social exclusion.

The numerous definitions found in the literature are almost all retraceable to the traditional distinction between absolute and relative poverty. The first understood as the incapacity to reach an objective level of wellbeing, independent of relevant social and temporal contexts; the second definition is, instead, based on the assumption that the social condition of an individual cannot be adequately defined without taking the environment in which they live as a starting point. In one case an individual is thus considered as poor if he is not be able to satisfy his primordial needs; whilst in the other case, if the individual live is in a worse state than the standard of the particular community in which he is located.

A transversal approach is therefore proposed as an alternative to the above, considered as subjective, through which the poor are defined as those who identify themselves as such, even if this identification is revealed as a result of the comparison that they operate with the rest of society in terms of perceived wellbeing. It is, therefore, the "perception" an individual may have of their own condition that allows for the identification of the measure of poverty to a far greater degree than the assessment of external observers would allow.

It should, therefore, be noted that such a line of enquiry is part of the wider trend which attempts to focus, in particular, on the multidimensional nature of poverty, i.e. on the necessity to take into consideration not only one single indicator but a group of indicators (considered useful in the definition of greater or lesser degrees of hardship in the individual observed). This approach recalls, in particular, the work of Renè Lenoir [3] on social exclusion, the human poverty index in the United Nations Report on Human Development as well as the work of A. Sen [4] on functioning and capability.

In order to obtain a measurement of the level of socio-economic hardship of the families interviewed, sets of indicators were constructed for the detection of the possession or absence of functional goods, the ability to bear certain costs, the perception of the evolution of the economic condition of the family etc.. Such sets of indicators were used in order to obtain a fuzzy value corresponding to the level of hardship of each family. The indices were chosen in order to identify levels of socio-economic hardship and were calculated so as to match the high values of the index with a high level of hardship and low values of the index with higher levels of well-being [5,6].

They were grouped into several sets characterized by different situations: *difficulty in paying debts/instalments or buying food staples* (mortgages, other debts and taxes, utility bills, food staples); *difficulty in paying for education, health or unforeseen expenses* (costs of school meals and other subsidies for children; voucher for medical treatment in public hospitals, private medical care or other unexpected expenses); *difficulty in purchasing other goods and services* (consumption of meat or fish at least once every two days, heating or air-conditioning in the home, purchase of clothing items when needed, going to the cinema/theatre at least once a month, going on holiday for one week a year) and *difficulty in participating in events* (social, religious, sporting, political, voluntary, or cultural).

3. Methodological approach

3.1 The fuzzy logic

The classical logic is based on the truth of propositions. In particular, a proposition can be true or false (or in the language computational can assume a value equal to 0 or equal to 1). This bi-valent logic (true-false; 0-1) goes back to Aristotle and even earlier to Parmenides, who in 400 BC introduced the dichotomy of "true-false". This logic has been the subject of many developments from the Middle Ages to the modern age, until the last century and the present day (antinomy Russell, Godel's theorem) and is based on correctness of deductive reasoning [7].

In the early 30s the first proposals of poly-valent logics are proposed by the mathematician Lukasiewicz, with the three-valued logic, and by physical Black, with vague sets [8,9].

In 1965 Lotfi A. Zadeh, engineer and professor at the University of Berkeley, California, well known for his contributions to systems theory, proposed a poly-valent logic to infinite values between 0 and 1, which called fuzzy logic. It is a term that means shaded, blurred, frayed. In general a concept is said fuzzy, when it corresponds to a class of elements that do not have well-defined boundaries and for which there are, therefore, partial truths.

According to the traditional conception of a given object can belong to a set or not: the membership is, therefore, bivalent and there are intermediate cases. According to the fuzzy logic, on the other hand, an object can belong partially to a set considered, that belong to a certain extent. Similarly, in classical logic a statement is semantically evaluated as true or false, while in fuzzy logic it is assigned a value of partial truth, which is best suited to so many situations where you cannot have absolute certainty about the characteristics of the phenomenon.

Starting from the concept of degree of membership Zadeh published his first articles [10,11], and gave rise to fuzzy logic.

The development of fuzzy theory initially stems from the work of Zadeh and subsequently draws upon Dubois and Prade [12] and their definition of a methodological basis. Fuzzy theory develops from the assumption that every unit is associated contemporarily to all categories identified and not univocally to only one, on the basis of ties of differing intensity expressed by the concept of degrees of association. Fuzzy methodology in the field of "poverty studies" in Italy has been recently employed in the work of Cheli and Lemmi [13] who define their method "total fuzzy and relative" (TFR) on the basis of the previous contribution from Cerioli and Zani [14].

The Total Fuzzy and Relative (TFR) model is used in order to summarize the values emerging from analysis in a single "blurred" fuzzy value which, as described above, measures the degree of membership of an individual in the range between 0 (condition of well-being) and 1 (hardship). Such a method consists in the construction of a function of membership to the fuzzy totality of the poor which is continuous in nature, and able to provide a measurement of the degree of poverty present within each unit.

3.2 The fuzzy C-means

Cluster analysis is highly advantageous as it provides "relatively distinct" (or heterogeneous) clusters, each consisting of units (families) with a high degree of "natural association". Different approaches to cluster analysis are characterized by the need to define a matrix of dissimilarity or distance between the n pairs of observations.

The cluster analysis allows to identify the profiles families who meet certain descriptive characteristics, not defined a priori. The cluster analysis is, in fact, a multivariate analysis technique through which you can group the statistical units in classes, so that the observations are as homogeneous as possible within the classes and the possible heterogeneous between the different classes [15,16,17].

This technique starts with the choice of an algorithm which defines the rules of how to group units into subgroups based on their similarity.

Depending on of data, you have different sizes. For quantitative data have distance measures, for qualitative data have association measures. Once the choice of measurement to be used, there is the choice of method or algorithm of classification and the criterion of aggregation / subdivision. The most common methods of classification are: aggregation or divisive hierarchical methods [18,19], non-hierarchical methods [20].

In our work, in order to identify the profile of poverty arising from the socio economic indicators, it was decided to choose a procedure of fuzzy cluster, in particular the *Fuzzy c-means (FCM)*. It is a clustering method that allows a set of data to belong not only to a main cluster but also to two or more clusters.

The c-means differs from the k-means objective function through the additions of the u_{ik} membership values and the fuzzifier m that determines the level of cluster fuzziness.

A fuzzy c-partition of **Y,** (subset of $\boldsymbol{R}^N$), is that which characterizes the membership of each sample point with all clusters through the identification of a membership function that assumes values of between zero and one. The sum of the memberships for each sample point must be equal to one.

This method was developed by Dunn [21] and later by Bezdek [22,23,24]. Several clustering criteria have been proposed for identifying optimal fuzzy *c*-partitions in Y. Of these, the most popular and well-studied method to date is associated with the generalized least-squared errors functional:

$$J_m(U,v)=\sum_{i=1}^{N}\sum_{k=1}^{c}(u_{ik})^m \left\|y_i-v_k\right\|^2 \qquad \textbf{(1)}$$

where [22]:

- $\mathbf{Y}=(y_1,y_2,\ldots,y_i,\ldots,y_N)\in R^d$ is a data set;
- c is the number of cluster with $2\le c\le N$,
- m is weighting exponent with $m\ge 1$,
- U, a $N\times c$ matrix, is a fuzzy *c*-partition of **Y**,
- $v=(v_1,v_2,....,v_c)$ is a vector of centres,
- $v_i=(v_{i1},v_{i2},....,v_{id})$ is the center of cluster *i*.

The FCM algorithm, via iterative optimization of J_m, produces a fuzzy *c* partition of the **Y** data set. The steps to be followed are:

1. determine the number of clusters $2\le c\le N$ e $m\ge 1$;
2. initialize the fuzzy *c*-partition $U^{(0)}$ with random numbers $u_{ik}\in[0,1]\ \forall i,j$, $\sum_{k=1}^{c}u_{ik}\ =\ 1\ \forall i$ and $0<\sum_{i=1}^{N}u_{ik}\ <\ N\ \forall N$;
3. calculate the *c* cluster centres with a general equation for the *k*-th cluster centre:

$$v_k=\frac{\sum_{i=1}^{N}(u_{ik})^m y_i}{\sum_{i=1}^{N}(u_{ik})^m}$$

4. Subsequently updating $U^{(b)}$ the membership matrix, at step *b*. $U^{(b+1)}$ is calculated with the equation:

$$u_{ik}=\frac{1}{\sum_{j=1}^{c}\left(\frac{\left\|y_i-v_k\right\|}{\left\|y_i-v_j\right\|}\right)^{2/(m-1)}} \qquad \textbf{(2)}$$

5. Finally, $U^{(b)}$and $U^{(b+1)}$ are compared through a matrix norm: the stopping rule is

$\left\| U^{(b)} - U^{(b+1)} \right\| \leq \varepsilon$, otherwise calculate the new *c* cluster centres at step 3.

4 The case study

4.1 Introduction

The cluster analysis allowed the identification of several family profiles derived from the fuzzy applications.

The cluster analysis on the 20 variables under observation, placing m = 1 and *c* = 4, produces only *four different main clusters*, each of which has *its own average profile*, thanks to which every family belongs exclusively to one cluster (*K-means*). In particular:

- *Cluster 1* refers to those families perceiving a situation of greater hardship in all types of expenses;
- *Cluster 2* refers to those families who do not perceive any hardship in each expense;
- *Cluster 3* presents a profile of medium-high hardship, albeit different, with peaks corresponding only to certain expenses;
- *Cluster 4* demonstrates a low hardship profile in terms of bearing material costs but a high level of hardship in participating in social and cultural events or recreational activities (Fig. 1).

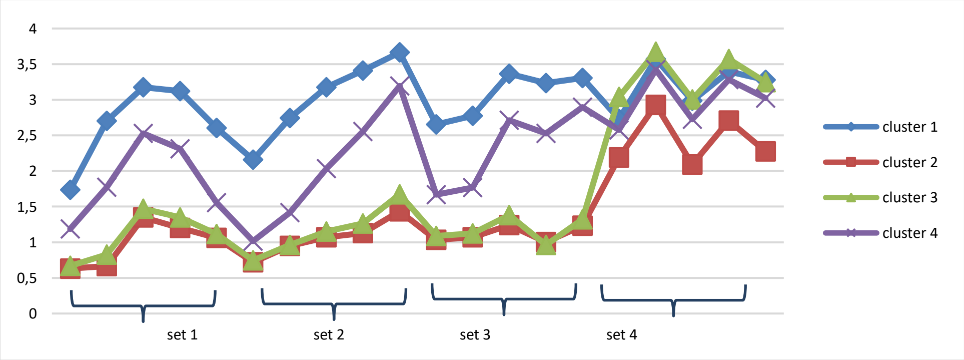

Figure 1 Classification of families per cluster membership and level of hardship.

The fuzzyfier *m* determines the level of cluster fuzziness. By varying *m* we will see how the cluster will change by creating "overlapping" between different profiles.

However, the profiles of belonging to a single cluster or a set of clusters, can be identified through Fuzzy *c*-means. Two simulations are carried out by placing *m=1.1* and *m=2.0* (Table 1). Increasing *m*, the main clusters disappear, leaving only the cluster “highly fuzzy" determined by the overlap of multiple clusters.

Table 1 Composition of clusters by placing m = 1.1 and m = 2.0.

Cluster m=1.1	**Absolute value**	**%**
Cluster 1	732	29.2
Cluster 2	557	22.2
Cluster 3	435	17.4
Cluster 4	454	18.1
Cluster 1,2	55	2.2
Clusters 1,4	93	3.7
Clusters 2,3	101	4.0
Clusters 2,4	64	2.6
Clusters 3,4	1	0.0
Clusters 1,2,4	11	0.4
Clusters 2,3,4	4	0.2
Total	**2,507**	**100**

Cluster m= 2.0	**Absolute value**	**%**
Cluster 1	4	0.2
Clusters 1,3	882	35.2
Clusters 1,4	2	0.1
Clusters 2,4	1232	49.1
Clusters 1,2,4	324	12.9
Clusters 1,2,3,4	63	2.5
Total	**2,507**	**100**

Another simulation is carried out by placing *m=1.5*. This simulation is the most appropriate since it is used to divide the different families in the main cluster and cluster "fuzzy" characterized by profiles derived from a mixture of two or more characteristics of the main cluster. Four different clusters are thus obtained, each of which with its own average profile and for which every family belongs to only one cluster and other 6 clusters of families that do not specifically belong to a well-defined cluster but belong to two or three clusters, as shown in Table 2.

Table 2 Composition of clusters by placing m=1.5 and the average value of hardship.

Cluster m=1.5	**Absolute value**	**%**	**Value of hardship**
Cluster 1	202	8.1	3.3
Cluster 2	368	14.7	1.2
Cluster 3	213	8.5	1.7
Cluster 4	201	8.0	2.2
Clusters 1,4	467	18.6	2.7
Clusters 2,3	643	25.6	1.5
Clusters 2,4	61	2.4	1.5
Clusters 3,4	156	6.2	2.2
Clusters 1,3,4	17	0.7	2.6
Clusters 2,3,4	179	7.1	1.8
Total	**2,507**	**100.0**	

Through a representation of the clusters on the basis of the average values of hardship, the following “fuzzy” relations between the different average profiles are obtained on the basis of membership to the different clusters.

By using the profile of the fuzzy cluster *2, 3, 4* together with the profiles of the single clusters 2, 3, 4 as an example, it should be highlighted whether there are some contributions of the single clusters to the fuzzy one. From the analysis of the profiles shown in Figure 2 it can be deduced that the cluster 4 and the “cluster 2,3,4” are strongly associated: until the variable 16 an attraction of the profiles 2 and 3 is felt on the profile 2,3,4 downwards. When the behaviour of the profiles 2 and 3 diverges, there is the strongest association between the “cluster 2, 3, 4” and 4.

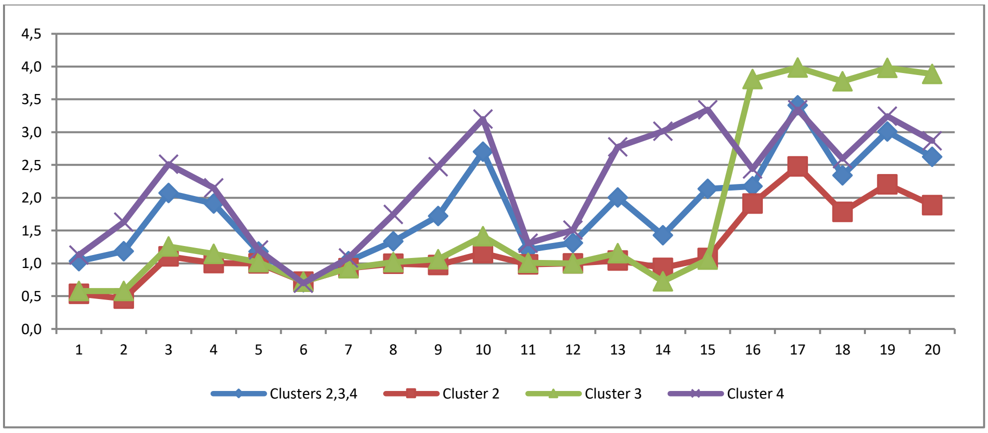

Figure 2 Classification of families per cluster membership and level of hardship (Cluster 2,3,4,).

5. Conclusion

The current analysis has attempted to quantify the influence of income and of family typology (number of members) in order to understand how family lifestyles may evolve. The risk of poverty estimated on the basis of “objective” indicators, such as income or levels of debt, is completely independent from the state of awareness of those directly involved. It is also useful, however, to observe the “subjective” perception of Italian people in relation to their standard of living and to the recurring causes of economic and social hardship.

The present study attempts to overcome the old classifications between poor and non-poor families by creating “fuzzy profiles” among those living in different circumstances. Through the different applications carried out in this work it is possible to create fuzzy profiles highlighting the specific peculiarities of small groups not strictly belonging to a defined profile but to a mix of different profiles (Fuzzy c-means).

The results obtained show that it is possible to identify the social stratification on the basis of the different components that influence the behavior of households at the same time. In particular, the results obtained from the simulation profiles allow detection of well-defined, symptomatic of situations of strong social discomfort and above all economic.

It is hoped that the changes in the profile of poverty that emerge from analysis, conducted with different criteria, provide important insights not only to better explain and understand the phenomenon of economic hardship, but also to obtain information on social policies to reduce poverty.

References

1. Kosko, Bart (1993). Fuzzy Thinking: The New Science of Fuzzy Logic. Hyperion. ISBN 0-7868-8021-X.
2. Zadeh, L.A., (1978), Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems, 1(1), 3-28.
3. Lenoir R. (1974) Les Exclus. Un francais surd ix. Seuil, Paris.
4. Sen A. (1994) Well-Being, Capability and Public Policy, Giornale degli Economisti e Analisi di Economia, vol.L III (N.S.).
5. Montrone, S. and Perchinunno, P. (2012) “Socioeconomic Zoning: Comparing Two Statistical Methods” In: Montrone, Perchinunno (eds.) Statistical Methods for Spatial Planning and Monitoring. Contributions to Statistics, Springer, Milan Heidelberg New York Dordrecht London: ISBN: 978-88-470-2750-3, ISSN: 1431-1968, pp. 93-118.
6. Montrone, S., Campobasso, F., Perchinunno, P., Fanizzi, A. (2010) “A Fuzzy Approach to the Small Area Estimation of Poverty in Italy” In: Gloria Phillips-Wren, Lakhmi C. Jain, Kazumi Nakamatsu, and Robert J. Howlett (Eds.), Advances in Intelligent Decision Technologies, Smart Innovation, Systems and Technologies n.4, ISSN 2190-3018, ISBN 978-3-642-14615-2, Springer, Heidelberg, pp. 309–318.
7. Veronesi M., Visioli A. (2003), Logica fuzzy. Fondamenti teorici e applicazioni pratiche. Franco Angeli, Milano.
8. Cammarata S. (1994), Sistemi fuzzy. Un’applicazione di successo dell’intelligenza artificiale. ETAS.
9. Cammarata S. (1997), Sistemi a logica fuzzy. Come rendere intelligenti le macchine, ETAS.
10. Zadeh L. A. (1965) Fuzzy sets. Information and Control., 8(3): pp. 338-353.
11. Zadeh L. A. (1968). Fuzzy algorithms. Information and Control (5): pp. 94-102.
12. Dubois, D., Prade, H. (1980) Fuzzy sets and systems. Academic Press, Boston, New York London.
13. Cheli, B., Lemmi, A. A (1995) Totally Fuzzy and Relative Approach to the Multidimensional Analysis of Poverty. Economic Notes vol. 24, n° 1, 115--134
14. Cerioli, A., Zani, S. (1980), A Fuzzy Approach to the Measurement of Poverty. In: Dugum, C., Zenga, M. (eds.) Income and Wealth Distribution, inequality and Poverty. Springer Verlag, Berlin (1990)
15. Fabbris L. (1990), Analisi esplorativa di dati multidimensionali, Cleup editore.
16. Green P.E., Frank R.E., Robinson P.J. (1967), Cluster Analysis in text market selection, Management science.
17. Jardine N., Sibson R. (1971), Mathematical taxonomy, Wiley, London.
18. Johnson S.C.(1967), Hierarchical clustering schemes, Psycometrika.
19. Everitt B.S. (1979), Unresolved problems in cluster analysis, Biometrics.
20. Andemberg M. (1973), Cluster analysis for applications, New York Academic Press.
21. Dunn (1973) “A fuzzy relative of the ISODATA process and its use in detecting compact, well-

separated clusters" Journal Cibern, vol 3, pp 32-57

22. Bezdek J. C., Ehrlich R., Full W. (1984) FCM: the fuzzy c-means clustering algorithm In: Computer e geosciences vol. 10, n. 2-3, pp. 191-205.
23. Bezdek J. C., Cannon R. L., Dave J. V. (1986) Efficient Implementation of the fuzzy c-means clustering algorithms IEEE Transactions on patters analysis and machine intelligence, vol 8 , n. 2, pp. 248-255.
24. Bezdek J. C. (1981): "Pattern Recognition with Fuzzy Objective Function Algorithms", Plenum Press, New York.