SUMMARY OF THE MODIFICATIONS:

1) The three sentences from line 9 to line 15 have been put in the footnotes
2) References indicated have been introduced in addition to the five taken from the IJBIDM
3) English have been reviewed

Fuzzy cluster and validity indices in a socio-economic context

Silvestro Montrone, Paola Perchinunno* and Samuela L'Abbate
Department of Business and Law Studies,
University of Bari, Italy
E-mail: silvestro.montrone@uniba.it
E-mail: paola.perchinunno@uniba.it*
E-mail: samuela.labbate@uniba.it
*corresponding author

Abstract: The multidimensional nature of socio-economic hardship requires a multidimensional research approaches, oriented toward advanced solutions, able to capture the changing dimensions of the problem at hand. One of such approaches consists in abandoning traditional dichotomous logic in favor of a semantically richer fuzzy classification, in which each unit belongs and, at the same time, does not belong to a given category. Cluster analysis allows to identify the profiles families who meet certain descriptive characteristics not defined a priori. The approach used in this work to synthesize and measure hardship conditions is based on a clustering procedure known as Fuzzy clustering by Local Approximation of Membership (FLAME), and based on defining the neighborhood of each object and identifying cluster supporting objects. This clustering method not only allows for each instance of a data set to belong to a unique main cluster, but also that each instance can be shared by two or more clusters on the ground of suitably defined "fuzzy profiles".

Keywords: fuzzy clustering, hardship, flame, prototypes, cluster validity index.

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.

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 member of the organizing committee of the session "Econometrics and Multidimensional evaluation in Urban Environment" of International Conference on Computational Science and its Applications.

Samuela L'Abbate graduated in mathematics and will receive her PhD in statistics from the Statistics Department at the University of Bari, Italy, in May 2013. In the statistics field, she was involved in data mining, in particular, cluster analysis, a technique of unsupervised data mining.

The contribution is the result of joint reflections by the authors, with the following attributions: to S. Montrone (Sections 2 and 4.3), to P. Perchinunno (Sections 3 and 4.5) and to Samuela L'Abbate (Sections 4.1, 4.2 and 4.4). The introduction and conclusions are the result of the common considerations of the authors.

1 Introduction

Fuzzy clustering methods allow the objects to belong to several clusters simultaneously, with a different degrees of membership. Objects on the boundaries between several classes are not forced to fully belong to one of the classes, but rather are assigned membership degrees between 0 and 1 indicating their partial membership. The discrete nature of the hard partitioning also causes difficulties with algorithms based on analytic functional, since these are not differentiable.

Broadly speaking, clustering is the process of grouping a data in such a way that the similarity between data within a cluster is maximised, while the similarity between data of different cluster is minimised[1] [1, 2, 3, 4].

Fuzzy clustering is used for complex data sets and multi-dimensional, where the members have fuzzy relationships. Among the various techniques developed, algorithm-Fuzzy C-means (FCM) is the most popular, in which a piece of data has partial membership with each of the cluster centers predefined. In FCM, the cluster centres and membership values of the data points with them are updated through some iterations [5].

Fu and Medico [6] developed a clustering algorithm to capture dataset-specific structures at the beginning of DNA microarray analysis process, which is known as Fuzzy clustering by Local Approximation of Membership (FLAME). It worked by defining the neighbourhood of each object and identifying cluster supporting objects. Fuzzy membership vector for each object was assigned by approximating the memberships of its neighbouring objects through an iterative converging process. Ma and Chan [7] proposed an Incremental Fuzzy Mining (IFM) technique to tackle complexities due to higher dimensional noisy data, as encountered in genetic engineering

The objective of this report is the individuation of different profiles, not defined a priori, of each family behaviors with socio-economic specific. 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 [8, 9, 10, 11, 12, 13].

2 Fuzzy clustering by Local Approximation of Membership (FLAME)

The algorithm FLAME (Fuzzy clustering by Local Approximation of Membership) [6], worked by defining the neighbourhood of each object and identifying cluster supporting objects. Fuzzy membership vector for each object was assigned by approximating the memberships of its neighbouring objects through an iterative converging process.

[1] In other words, cluster analysis is a statistical technique used to generate a category structure which fits a set of observations, a key point being that there is no need for the classes to be identified prior to processing. The main input to clustering algorithms is the dissimilarity or distance matrix between the *n* pairs of observations. Distance measures are typically used with quantitative data, whereas association measure are often suitable to define dissimilarity measures mostly useful with categorical/attribute data.

The method involves three main steps: definition of the information structure on which to base the classification; application of a criterion of approximation to define the fuzzy membership of an object to one or more groups and final classification.

The information structure must necessarily be based on a criterion of accumulation points. In particular it is necessary to define the particular points that are "poles of attraction" of a certain neighborhood. This neighborhood, for a generic object *i*, is defined by *k* objects nearest or most similar to it. This k-nearest neighbors are denoted KNN *(i)* and points of attraction as *Cluster Supporting Object* (*CSO*).

Assigned the dataset X with *n* points and *m* variables, the density is calculated with the following expression:

$$d_i = \frac{s_{max}}{s_i} \qquad \textbf{(1)}$$

where:

- $S_i = \frac{1}{k}\sum_{j\in KNN(i)} dist\big(X_i, X_j\big) \quad \forall i = 1,2,\dots,n$
- $S_{max} = \max_i S_i$
- $dist(X_i, X_j)$ is the function of distance used between X_i and $X_j$profiles .

The density values can be divided into three classes:

1. Defined a threshold level (percentile), all values below this threshold are considered outlier. In the original formulation of the algorithm (1), this set of points is considered as a cluster; in this work, however, outliers were excluded for not interfering in the process of local approximation and therefore the size of the data set is reduced to $n' = n - \text{ number of outliers}$.
2. Identification of the *CSO*s consists of identifying those points with the maximum density. Number of *CSO*s, i.e. the number of clusters, is not defined a priori and depends on the fact that if a *CSO* belongs to neighborhood of another *CSO*, between the two will choose the one with the higher density. Furthermore another parameter that influences the number of *CSOs* is *k*: as *k* increases, the number of clusters decreases.
3. The points that are not outliers or are not *CSO*s are the points that need to be classified.

The algorithm then provides the assignment of fuzzy membership by a method of local approximation. The number of clusters is defined by the number of *CSO*s, representative points of the dataset.

At the beginning, each object has the same degree of fuzzy membership at all clusters, with the exception of *CSO*s. In fact, to each *CSO* is assigned a full membership to himself as if it were a single cluster.

Therefore defined with $C = \#CSO$ (the number of *CSO*), for each profile (point) is associated a vector which indicates the membership degree to the cluster *l*:

$$p(i) = (p_{i1}, p_{i2}, \dots, p_{il}, \dots, p_{iC}) \qquad \forall\ i = 1,2,\dots,n'$$

with $0 \le p_{il} \le 1$ and $\sum_{l=1}^{C} p_{il} = 1$

this vector is updated with an iterative process that minimizes an error function of local approximation.

$$E[p] = \sum_{i=1}^{n'} \sum_{l=1}^{C} \left\| p_{il} - \sum_{j \in KNN(i)} w_{ij}\, p_{jl} \right\|^2 \qquad \textbf{(2)}$$

where w_{ij} are the coefficients to calculate the relative distances to the nearest neighbors obtained by:

$$w_{i,j} = \frac{1/dist(X_i, X_j)}{\sum_{k \in KNN(i)} 1/dist(X_i, X_k)} \qquad \textbf{(3)}$$

with $\sum_{j \in KNN(i)} w_{ij} = 1$.

The function $E[p]$ can be minimized by solving the following linear equations:

$$p_{il} - \sum_{j \in KNN(i)} w_{ij}\, p_{jl} = 0, \quad i = 1,2,\dots,n' \quad \text{and} \quad l = 1,2,\dots,C \qquad \textbf{(4)}$$

that have a unique solution found by the iterative procedure defined as

$$p_{il}^{t+1} = \sum_{j \in KNN(i)} w_{ij}\, p_{jl}^{t} \qquad \textbf{(5)}$$

The last step is the construction of the clusters that can be done in two ways. By assigning each object to the cluster which has a greater degree of membership in order to have a partition crisp. Or, considering a threshold value for the degree of membership, assigning each object to one or more clusters according to its degree of membership if greater than the threshold value set. In this work, we opted for the second mode in order to have the intermediate profiles.

3. Cluster Validity Indices

3.1 Introduction

In literature there are several validity indices suitable for hard partition clustering. Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following three types: *External Index* used to measure the extent to which cluster labels match externally supplied class labels (Entropy); *Internal Index* used to measure the goodness of a clustering structure without respect to external information (Sum of Squared Error) and *Relative Index* used to compare two different clusters. Often an external or internal index is used for this function, e.g., SSE or entropy.

Sometimes these are referred to us as criteria instead of indices. However, sometimes criterion is the general strategy and index is the numerical measure that implements the criterion. There are two criteria proposed for clustering evaluation and selection of an optimal clustering scheme [14]:

- *Compactness*, measuring the internal cohesion between the cluster objects (as they are close to each other).
- *Separation*, measuring the separation between the clusters. There are three approaches measuring the distance between two clusters: Single linkage, complete linkage and comparison of centroids [15].

Using this approach of cluster validity our goal is to evaluate the clustering result of an algorithm using only quantities and features inherent to the dataset.

3.2 Validity indices for fuzzy cluster

The "cluster validity" identifies indices to measure the quality of the partition obtained by the clustering algorithm, and in particular are able to provide useful information for the choice of the optimum values of *c* and *m*. The indices of cluster validity can be classified into:

- *Indices of fuzziness,* which measure the degree of "fuzzy" of the partition
- *Indices of compactness and separation:* measuring the degree of compactness and heterogeneity between groups
- *Indices of fuzziness and compactness / separation* which measure both the degree of overlap of the groups and their degree of compactness and heterogeneity between groups.

The most common indices are the *Partition Coefficient* and the *Partition Entropy* proposed by Bezdek, define as performance measures based on minimizing the overall content of pairwise fuzzy intersection [16; 17].

The *Partition Coefficient* (PC) is defined as:

$$PC = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^{2} \qquad \textbf{(6)}$$

The *PC* index assumes the value of 1/c (where c is the number of clusters) in the case of fuzziness maximum partition (that is, when the degrees of membership are equally distributed among the cluster), it takes value of 1 if the partition obtained is a type crisp or each unit belongs to only one group, then the degrees of membership are all equal to one or zero.

The *Partition Entropy (PE)* is defined as:

$$PE = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}\, log_a u_{ik} \qquad \textbf{(7)}$$

The index ranges from *0* to $log_a c$ and increases with increasing fuzziness. In the case of maximum fuzziness *P*E is equal to $log_a c$, in the case of partition crisp PE is 0, that is to say that the degrees of membership are all either 1 or 0. The indices *PE* and *PC* are the two most common criteria for measuring the degree of overlap between the groups, however, suffer from a high sensitivity to the parameter *m*. When *m* is very close to 1 or very high the two indices lose their ability to discriminate between the various values of *c*. Bedzek has shown that when *m* is close to 1 we have that *PC = 1* and *PE = 0*, while when *m* is very high PC=1/ c and *PE=* $log_a c$.

To overcome this tendency of the two indices, Dave [18] proposed an index ranging from 0 to 1 which is 0 in case of partition maximally fuzzy and 1 in case of partition maximally crisp:

$$MPC(c) = 1 - \frac{c}{c-1}\left(1 - PC(c)\right) \qquad \textbf{(8)}$$

The indices of fuzziness proposed some limitations and are subject to some criticism. In scientific literature, it is criticized the exclusive dependence of these indices from the matrix of the degrees of membership without considering in any way the information contained in the starting data as well as relatively to the centroids of the clusters.

The result of a grouping is given by the centers of the clusters, obtained from the application of a particular clustering technique. The criterion of judgment obtained if a partition is acceptable may be based on a criterion of proximity of an observed point to the centroid, as occurs in crisp methods, or the ability of a centroid of reproducing the data by the use of probability of membership.

In a technique of fuzzy clustering, these probabilities, or degrees of membership, are the main result. An index of validity of the partition, it may be based, therefore, on the ability of reproduction of data, every point should be well represented using the degree of membership and the centroid (or prototypes) of the obtained clusters.

Practically, the index of validity should help you understand if the centers of the clusters obtained by a process of grouping are able to offer a global view of the data set, and this means a small error of quantification.

In the field of fuzzy classification this error is usually measured with the index Xie-Beni. The *Xie-Beni index* [19] also called the compactness and separation validity function, is representative of indices involving the membership values and the dataset. The Xie-Beni index (XB) defines the inter-cluster separation as the minimum square distance between cluster centers, and the intra-cluster compactness as the mean square distance between each data object and its cluster center. The optimal cluster number is reached when the minimum of XB is found.

The XB index, with *m = 2*, modified by Pal and Bezdek [20] is defined as:

$$XB = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^{2}\left\|x_j - v_i\right\|^2}{N min_{i,j}\left\|v_i - v_j\right\|^2} \qquad \textbf{(9)}$$

Where the numerator indicates the *compactness* of the fuzzy partition and the denominator indicates the strength of the *separation* between cluster. A smaller value of XB indicates a compact and well-separated clustering partition.

This index, although this can provide a more reliable on a wide range of choice, both for the number of clusters both for the weighting parameter blurred, it has two drawbacks:

1) the index decreases monotonically when the number of clusters becomes very large and up to the points that coincide with the centroids;

2) there is no interaction between the expected number of clusters and the weighting parameter fuzzy (numerical instability) due to his behavior when tends to infinity.

To attenuate these problems various authors proposed corrections to the list of Xie and Beni. For example, S.H. Kwon [21], Tang and Sun [22] have introduced a tool "punitive" to mitigate the downward trend of the index as that C → ∞.

In this paper we examine the index proposed by Saha and Bandyopadhyay [23] based on the compactness and the ability of reproduction of the original data. The index Fuzzy Vector Quantization, FVQ, is defined as follows:

$$FVQ = \frac{\sum_{i=1}^{N}\left\|x_i - x_i'\right\|^2}{N \min\limits_{i \neq j, i,j=1,2,\ldots C} d(v_i, v_j)} \qquad \textbf{(10)}$$

where the numerator $\sum_{i=1}^{N} \|x_i - x_i'\|^2$ is the error due to the partition, or the error committed in considering the reconstruction of x_i, original data. Dividing the total error for the total number of points, *N*, is obtained the average error of quantization.

With x_i' we mean the data reconstructed through the expression:

$$x_{ij}' = \frac{\sum_{k=1}^{C} u_{ik}^{m} v_{kj}}{\sum_{k=1}^{C} u_{ik}^{m}} \tag{11}$$

$d(v_i, v_j)$ is the Euclidean distance between the centroid, or prototypes, v_i and v_j. The denominator of the index represents the minimum distance between the centers of each possible pair of clusters. Therefore, an optimal value of *C*, the number of clusters, we obtain it considering the *FVQ* for every *C* = 2,3, ...

As you can see, FVQ is a composition of two factors: the average error (the numerator compared to N number of points) and the Euclidean distance:

$$D_C = \min_{i \neq j, i, j = 1,2,\ldots C} d(v_i, v_j). \tag{12}$$

The first factor indicates the average error being committed in the reconstruction of points x_i' using the cluster centers, probabilities of belonging to a cluster u_{ij} with j = 1, 2,...,C. The minimum value of this factor indicates that the centroids were built properly, that is, the entire data set is well represented by these centers and you can see that this value decreases as C.

The second factor, D_C, which measures the minimum distance between a pair of centroids, decreases with increasing value of C. But is our wish that in a logical grouping, the cluster centers must be well separated.

Because the objective is to determine a partitioning conceptually correct, this means having also groups well separated, and it is therefore necessary to find a fair compromise between the minimum and average error cluster number *C*.

4. The case study

4.1 Introduction

In this report the data source used in order to construct indicators of socio-economic hardship is that 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. 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.

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.

4.2 The Construction of Sets of Indicators with a Fuzzy Method

The development of fuzzy theory initially stems from the work of Zadeh [24] and subsequently draws upon Dubois and Prade [25] 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 [26] who define their method "total fuzzy and relative" (TFR) on the basis of the previous contribution from Cerioli and Zani [27].

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.

Given a set of **X** elements $x \in \mathbf{X}$, any fuzzy subset A of X is defined as follows:

$$A = \{X, f_A(x) \ \}$$

where $f_A(x): x \to [0,1]$ is defined as the membership function of the fuzzy subset A and indicates the degree of membership of *x* to A. Therefore $f_A(x) = 0$ indicates that *x* does not belong to A, while $f_A(x) = 1$ indicates that *x* belongs only to A. However, in the case of $0 < f_A(x) < 1$, *x* belongs partially to A, with a greater degree of membership the closer $f_A(x)$ is to 1.

Supposing an observation of k poverty indicators for every family, the function of membership of i*th* family to the fuzzy subset of the poor may be defined thus:

$$f(x_i) = \frac{\sum_{j=1}^{k} g(x_{ij}).w_j}{\sum_{j=1}^{k} w_j} \qquad i = 1,....,n \qquad \textbf{(13)}$$

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).

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. The indices were grouped into several sets characterized by different situations:

1. **Set 1: difficulty in paying debts/instalments or buying food staples** (mortgages, other debts and taxes, utility bills, food staples);
2. **Set 2: 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);
3. **Set 3: 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);
4. **Set 4: difficulty in participating in events** (social, religious, sporting, political, voluntary, or cultural).

4.3 Cluster Validity Indices

Proceeding in the cluster analysis, we have to define the optimal number of clusters. Considering various processing by varying the number of clusters, *C,* the *Min_Pts*, and the *percentage of outliers* we built the Table 1. This table shows how vary the values of the indices of *Xie-Beni* and of Fuzzy Vector Quantization.

Tab 1: *Cluster validity indices (Xie-Beni, Fuzzy Vector Quantization)*

Min_Pts	%_outlier	FVQ	XB	Min_Pts	%_outlier	FVQ	XB
30	1	0.81467	2.22617	33	1	1.75986	2.32031
30	2	0.5466	5.05592	33	2	0.88389	3.12193
30	3	0.75654	2.085	33	3	1.64406	2.17184
30	4	0.51739	4.7575	33	4	0.84723	2.94882
30	**5**	**0.72312**	**1.97923**	33	5	1.57727	2.07589
31	1	1.01709	3.15558	34	1	1.77382	2.24466
31	2	0.57355	5.17736	34	2	0.93287	4.33586
31	3	0.94755	2.95387	34	3	1.65706	2.10026
31	4	0.54353	4.87139	34	4	0.89393	4.09507
31	5	0.90827	2.81553	34	5	1.59026	2.05143
32	1	1.07744	1.89019	35	1	1.84242	1.39991
32	2	0.73057	10.75652	35	2	0.96349	18.25247
32	3	1.00693	1.76713	35	3	1.72112	1.48391
32	4	0.69746	10.13942	35	4	0.92293	17.24359
32	5	0.96650	1.68622	35	5	1.65161	1.25143

As has been said, we need to consider a compromise between sufficient number of clusters, and the index value of validity. Here we consider the combination of *Xie-Beni* index equal to 0.72312 and of *Fuzzy Vector Quantization* equal to 1.97923 which show us the best partition with 30 *Min_Pts* and 5% of outliers.

4.4 Identification of Cluster Supporting Object (CSO)

Primarily the FLAME identify *CSO* or those points with particular characteristics about which build the clusters. These points are the "prototype family" to which members of the group have similar profiles.

Tab 2: *CSO for different value of hardship*

Cluster	Value of hardship Set 1	Value of hardship Set 2	Value of hardship Set 3	Value of hardship Set 4
CSO 1	0.82	0.90	0.96	1.00
CSO 2	0.70	0.84	0.95	0.60

CSO 3	0.13	0.63	0.61	0.39
CSO 4	0.36	0.71	0.54	0.49
CSO 5	0.21	0.58	0.53	0.64

From the five *CSO* identified emerges as each of them has a well-defined profile. In particular, the *CSO* 1 presents high levels of hardship with respect to all 4 Set of indicators; then it will be a "prototype" of the family whit *hard economic and social conditions (high hardship)*. The *CSO* 2 also reflects a prototype of a family in conditions of *high hardship* as regards the first 3 Set of indicators; however they haven't difficult in Set 4 (difficulty in participating in events: social, religious, sporting, political, voluntary, or cultural). The *CSO* 3 concerns families who have low levels of hardship in all Set of indicators, thus showing a prototype of a *prosperous family*. The *CSO* 4, however, shows the average levels of hardship in all sets of indicators, except in Set 3 (difficulty in purchasing other goods and services) which doesn't show any difficulty. Finally, the *CSO* 5 also presents a *low profile of hardship* from the economic point of view (first 3 Sets of indicators) and slightly high regard on Set 4 (difficulty in participating in events: social, religious, sporting, political, voluntary, or cultural). It is therefore families in *good economic conditions* but with *some difficulty in social aspects*.

4.5 Identification of Fuzzy Clusters

After identifying the *CSO*, were generated clusters of two types: *Main clusters*, which have the characteristics similar to only one *CSO* and *Fuzzy Clusters*, with similar characteristics to one or more *CSO*, with a different level of membership (probability).

The different families are classified in five “Main Clusters” and in nine “Fuzzy Clusters", characterized by profiles derived from a mixture of two or more characteristics of the main five *CSO (tab. 3)*.

For example, the profiles of the *Fuzzy Cluster 4.5* presents an intermediate profile between the two main clusters, with further similarities to the cluster 4, compared to cluster 5. However the *Fuzzy Cluster* 5.4 assumes a profile more similar to the cluster 5, compared to cluster 4.

The Table 4 shows the average probabilities membership to different clusters. In particular, the *Fuzzy cluster 3,4,5*, has a higher probability membership to the *CSO* 3 (0.34), then to the *CSO* 4 (0.31) and finally to the *CSO* 5 (0.27).

***Tab 3:** Description of the clusters based on the number of households and the average value of hardship*

Cluster	**Number of families**	**%**	**Value of hardship Set 1**	**Value of hardship Set 2**	**Value of hardship Set 3**	**Value of hardship Set 4**
Cluster 1	46	2.6%	0.84	0.87	0.94	0.98

Cluster 2	94	5.4%	0.84	0.88	0.95	0.63
Cluster 3	1	0.1%	0.13	0.63	0.61	0.39
Cluster 4	87	5.0%	0.69	0.78	0.82	0.73
Clusters 5	1	0.1%	0.21	0.58	0.53	0.64
Clusters 2,4	48	2.8%	0.79	0.78	0.89	0.48
Clusters 3,4	17	1.0%	0.12	0.67	0.63	0.33
Clusters 4,2	26	1.5%	0.81	0.74	0.75	0.48
Clusters 4,5	861	49.3%	0.39	0.67	0.65	0.62
Clusters 5,4	386	22.1%	0.21	0.51	0.46	0.80
Clusters 3,4,5	23	1.3%	0.11	0.63	0.59	0.36
Clusters 4,3,5	32	1.8%	0.16	0.65	0.60	0.28
Clusters 4,5,3	82	4.7%	0.22	0.50	0.53	0.28
Clusters 5,4,3	7	0.4%	0.12	0.52	0.60	0.50
Outliers	34	1.9%				
Total	**1,745**	**100.0%**				

Tab 4: *Probability membership to the single clusters (values between 0 and 1)*

Cluster	Number of families	Prob 1	Prob 2	Prob 3	Prob 4	Prob 5
1	46	0.32	0.18	0.11	0.21	0.18
2	94	0.12	0.34	0.12	0.23	0.18
2,4	48	0.09	0.31	0.14	0.26	0.20
3	1	0.00	0.00	1.00	0.00	0.00
3,4	17	0.02	0.05	0.37	0.31	0.25
3,4,5	23	0.02	0.05	0.34	0.31	0.27
4	87	0.12	0.20	0.15	0.30	0.23
4,2	26	0.08	0.26	0.15	0.29	0.22
4,5	861	0.04	0.09	0.20	0.36	0.30
4,3,5	32	0.02	0.06	0.31	0.34	0.27
4,5,3	82	0.03	0.06	0.26	0.35	0.30
5	1	0.00	0.00	0.00	0.00	1.00
5,4	386	0.03	0.07	0.20	0.33	0.37
5,4,3	7	0.02	0.06	0.27	0.32	0.33

5. Conclusion

Numerous social marginalization phenomena are often studied without specific statistical data. Analyses of poverty are based on surveys of a general nature and often at a "macro" level. It would be of particular interest to perform in Italy, as has already been the case for several years in the United States, surveys of "micro" areas.

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

In this work the focus is to quantify the influence of income and of family typology (number of members) in order to understand how family lifestyles may evolve. The estimates' risk of poverty based on "objective" indicators, such as income or levels of debt are completely independent from the state of awareness of those directly involved. It is, however, also useful 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 analysis presented seeks to overcome old classifications between the poor and non-poor by creating "blurred" profiles between those living in different circumstances.

Through the application carried out in this work it is possible to:

- Analyze situations of family hardship through the synthesis of multi-dimensional sets of indicators (Total Fuzzy and Relative method);

- Define the optimal number of clusters through a Cluster Validity Indices;

- Identify Cluster Supporting Object (CSO), points with particular characteristics around which build Main clusters and Fuzzy Clusters.

- Create "fuzzy profiles" highlighting the specific peculiarities of small groups not strictly belonging to a defined profile but to a mix of different profiles, having different levels membership to the Main clusters [28, 29, 30, 31, 32, 33, 34].

It is hoped that the variations regarding the new family profiles emerging in general from analyses carried out with different criteria may provide a solid basis for not only a more accurate description and understanding of the phenomenon of economic hardship but also for developing new social policies that may contrast poverty.

References

1. Kwok T., Smith K. A. , Lozano S., Taniar D. (2002): "Parallel Fuzzy c-Means Clustering for Large Data Sets", Proceedings of the 8th International Euro-Par Conference, Lecture Notes in
2. Computer Science, Vol. 2400, Springer, pp: 365-374.
3. Fabbris L. (1990), Analisi esplorativa di dati multidimensionali, Cleup editore.
4. Green P.E., Frank R.E., Robinson P.J. (1967), Cluster Analysis in text market selection, Management science.
5. Jardine N., Sibson R. (1971), Mathematical taxonomy, Wiley, London.
6. Chattopadhyay S., Pratihar D.K., De Sarkar S.C., A Comparative Study of Fuzzy C-Means Algorithm and Entropy-Based Fuzzy Clustering Algorithms, Computing and Informatics, Vol. 30, 2011, 701–720.
7. Fu, L.—Medico, E.: FLAME: A Novel Fuzzy Clustering Method for the Analysis of DNA Microarray Data. BMC Bioinformatics. Vol. 8, 2007, No. 3, doi:10.1186/1471- 2105-8-3.
8. Montrone, S. and Perchinunno, P. (2015) 'The identification of 'fuzzy profiles' through the c-means clustering', Int. J. Business Intelligence and Data Mining, Vol. 10, No. 1, pp.62–72
9. Perchinunno, P. and Montrone, S. (2016) 'Different fuzzy cluster validity indices for the evaluation of the quality of the resulting partitioning', Int. J. Innovative Computing and Applications, Vol. 7, No. 2, pp.84–90.
10. Jian-Jiang Lu, Bao-Wen Xu, Yan-Hui Li, Da-Zhou Kang (2006) A family of Extended Fuzzy Description Logics Int. J. Business Intelligence and Data Mining, Vol. 1, No. 4, pp. 384-400.
11. Nikos Pelekis, Dimitris K. Iakovidis, Evangelos E. Kotsifakos, Ioannis Kopanakis (2008) "Fuzzy clustering of intuitionistic fuzzy data" Int. J. of Business Intelligence and Data Mining, 2008 Vol.3, No.1, pp.45 - 65
12. Lokesh Kumar Sharma, Simon Scheider, Willy Kloesgen, Om Prakash Vyas (2008) Efficient clustering technique for regionalisation of a spatial database, Int. J. of Business Intelligence and Data Mining, 2008 Vol.3, No.1, pp.66 - 81
13. Zahid Ahmed Ansari; Abdul Sattar Syed (2016) "Discovery of web usage patterns using fuzzy mountain clustering" International Journal of Business Intelligence and Data Mining (IJBIDM), Vol. 11, No. 1, 2016, pp 1-18.

14. Ma, P., Chan, K.: Incremental Fuzzy Mining of Gene Expression Data for Gene Function Prediction. IEEE Trans. on Biomedical Engineering, 2010.
15. Berry, M.J.A. and Linoff, G. (1996). Data Mining Techniques For Marketing, Sales and Customer Support. John Wiley & Sons, Inc., USA.
16. Halkidi M, Batistakis Y, Vazirgiannis M, (2001) On clustering validation techniques, *Journal of Intelligent Information Systems 17* (2), 107-145
17. Bezdek, J.C. (1974) Cluster validity with fuzzy sets, J. *Cybernet. 3* 58–74.
18. Bezdek, J.C. (1974) Numerical taxonomy with fuzzy sets, *J. Math. Biol.* 1 57–71.
19. Dave R.N. (1996), Validating fuzzy partition obtained through c-shells clustering, *Pattern Recognition Lett. 17* 613–623.
20. Xie X.L., Beni G. (1991), A validity measure for fuzzy clustering, *IEEE Trans. Pattern Anal. Mach. Intell.* 13 (8) 841–847.
21. Pal N.R., Bezdek J.C. (1995), On cluster validity for fuzzy c-means model, *IEEE Trans. Fuzzy Systems* 3 (3) 370–379.
22. Kwon, S.H. (1998), Cluster validity index for fuzzy clustering, *Electronics Letters,* Vol. 34, Issue: 22, pp.2176 - 2177.
23. Tang, Y.,Sun, F (2005), Improved Validation Index for Fuzzy Clustering, *American Control Conference*, June 8-10. Portland, OR, USA.
24. Saha, S.; Bandyopadhyay, S. (2007), A New Cluster Validity Index Based on Fuzzy Granulation-degranulation Criteria, *Advanced Computing and Communications, ADCOM*. International Conference on, pp. 353 - 358,
25. Zadeh, L.A: Fuzzy sets, Information and Control. 8(3), 338--353 (1965)
26. Dubois, D., Prade, H.: (1980) Fuzzy sets and systems. Academic Press, Boston, New York London
27. 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
28. 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 pp. 272-284.
29. Montrone S. Perchinunno P., L'Abbate S., Zitolo M.R. (2014) The Lifestyles of Families through Fuzzy C-Means Clustering, in B. Murgante et al. (Eds.) LNCS 8581, Springer International Publishing Switzerland, pp. 122–134.
30. Mahardhika Pratama; Jie Lu; Guangquan Zhang, , (2016) "Evolving Interval Type-2 Fuzzy Classifier", IEEE Transactions on Fuzzy Systems, Vol. 24(3), pp. 574-589
31. Mahardhika Pratama; Sreenatha. G. Anavatti; Plamen. P. Angelov; Edwin Lughofer , (2014) "PANFIS: A Novel Incremental Learning Machine", IEEE Transactions on Neural Networks and Learning Systems, Vol.25(1), pp.55-68
32. Mahardhika Pratama, Jie Lu, Sreenatha Anavatti, Edwin Lughofer, Chee-Peng Lim , (2016) "An Incremental Meta-Cognitive-based Scaffolding Fuzzy Neural Network", Neurocomputing, Vol. 171, pp.89-105
33. Mahardhika Pratamaa, Jie Lub, Edwin Lughoferc, Guangquan Zhangb, Sreenatha Anavattid , (2016) "Scaffolding Type-2 Classifier for Incremental Learning under Concept Drifts", Neurocomputing, Vol.191, pp. 304-329
34. Mahardhika Pratama; Sreenatha G. Anavatti; Edwin Lughofer (2013) "Evolving Fuzzy Rule-Based Classifier Based on GENEFIS", in Proceedings of the 2013 IEEE Conference on Fuzzy Systems (FUZZ-IEEE), pp.1-8,Hyderabad, India