Statistica & Applicazioni - Vol. VII, n. 2, 2009, pp. 149-171

A FUZZY CLUSTERING APPROACH TO IMPROVE THE ACCURACY OF ITALIAN STUDENT DATA. AN EXPERIMENTAL PROCEDURE TO CORRECT THE IMPACT OF OUTLIERS ON ASSESSMENT TEST SCORES

Claudio Quintano*
Rosalia Castellano**
Sergio Longobardi***

SUMMARY

The aim of this paper is to introduce a new approach to outlier analysis in which the detection is carried out on data with a hierarchical structure and a complex pattern of variability, e.g. pupils in classes, employees in firms, etc. In particular, we analyze the data collected by the Italian National Evaluation Institute of the Ministry of Education (INVALSI) in which the micro units –students- are nested within classes and schools, with a strong presence of outliers at the second level -class- of hierarchy. By the analysis of within class variability, we have developed a procedure to detect outlier units at class level combining the factorial analysis with a fuzzy clustering approach. The purpose of this method is to go over the dichotomous logic which classifies each unit as outlier or not outlier (hard clustering), computing an ''outlier level'' measure for each unit and in such a way calibrating the correction of overstimation of children ability due to the outlier presence.

Keywords: *outlier correction, data accuracy, assessment test scores.*

1. INTRODUCTION

Outliers are generally identified as observations which appear to be inconsistent with the remaining of the data (Barnett and Lewis, 1994). Many studies focus on detection of outlier units (Hodge and Austin, 2004; Hawkins, 1980) and propose several methods to deal with this problem (Iglewicz and Hoaglin, 1993). In this paper, we introduce a new approach to outlier analysis in which the detection is carried out on students' data with a hierarchical structure.

* Dipartimento di Statistica e Matematica per la Ricerca Economica - Università degli Studi di Napoli Parthenope - via Medina, 40 - 80133 NAPOLI
(e-mail: claudio.quintano@uniparthenope.it).

** Dipartimento di Statistica e Matematica per la Ricerca Economica - Università degli Studi di Napoli Parthenope - via Medina, 40 - 80133 NAPOLI
(e-mail: lia.castellano@uniparthenope.it).

*** Dipartimento di Statistica e Matematica per la Ricerca Economica - Università degli Studi di Napoli Parthenope - via Medina, 40 - 80133 NAPOLI
(e-mail: sergio.longobardi@uniparthenope.it).

This paper results from the common work of the Authors under the coordination of Claudio Quintano. Rosalia Castellano has written Sections 5 and 6, while Sergio Longobardi has written Sections 1, 2, 3 and 4.

The paper considers data on student performance assessments collected by the Italian National Evaluation Institute of the Ministry of Education (INVALSI) in the school years 2004/05 and 2005/06, focusing on the results of the primary classes.

The INVALSI survey is conducted every year and it evaluates, through a closed items test, the students' knowledge in three areas: *reading, mathematics and science*. The survey investigates the whole population of the second and fourth year of primary school students and a sample of secondary level students (beginners at lower secondary, first and third class of the upper secondary).

The tests are made up of a different number of items on the basis of the school level and the assessment area. Each dataset, at student level, is created for each school level and assessment area (totally 15 dataset). Every dataset contains the following variables: *gender, region, school, class, item answers and student final score.*

Some descriptive analyses have shown the presence of outlier units at class level, which leads to biased distributions of the average scores by class.

This anomaly leads us to suppose that many primary school teachers have provided an excessive support to the pupils during the performance test. Consequently, the computed score for each student of some classes may be subject to some bias due to teacher intervention.

In this context, the teacher support might be considered similar to an interviewer effect (Biemer, Groves, Lyberg, Mathiowetz and Sudman, 1991) and we might suppose that the student's score is affected by an error component which inflates the measurement errors (Braverman, 1996; Wentland and Smith, 1993).

Under these conditions, we propose a two-stage method for evaluating and correcting the overestimation of childrens' ability found in primary classes.

In the first stage, classes of students with both very high average score and the within variability close to zero have been detected through a factorial analysis.

The second stage consists in implementing a weighting system that assigns a weight to every class based on the probability of belonging to the set of outlier units which is calculated by a fuzzy clustering algorithm (Driankov, Hellendoor, and Reinfark, 1994; Klir and Folger, 1988). The final output of this procedure is a modified distribution that shows a decrease in the mean, median and mode with respect to the original one. Moreover, the correction factor is able to improve the skewness and to smooth the data distribution.

Finally, the main features of units with high probability to be classified as outliers are analyzed in order to evaluate a relationship between the geographical distribution of classes and the presence of outliers.

The paper is structured as follows: in Section 2 we analyze the features of class mean score distributions and highlight the presence of outliers. In Section 3 the developed procedure is described. Section 4 illustrates the effects of the correction procedure on the original distributions. In Section 5, we consider the relationship between the geographical location of the students and the presence of outliers. Finally, in Section 6 conclusions are made.

2. THE AVERAGE CLASS SCORE DISTRIBUTIONS

The results of explorative data analysis[1] show an upward bias in the distributions of the average scores by class in each assessment area (reading, mathematics and science) for the primary students.

Primary school data highlight the presence of many classes where a large percentage of students (close to 100%) gave the same answer to each question and consequently, received the same score. Furthermore, all answers are often correct and the whole class achieved the top score (100 points).

Figure 1 shows the distributions of mean performance score by class for the second class of primary school in the school years 2004/05 and 2005/06.

The graphical comparison highlights that the primary school distributions show an upward bias and an anomalous presence of high frequencies in correspondence of the maximum values of distribution, then the considerable presence of outlier classes has produced a unimodal distribution where the mode is equal to the top score.

In order to confirm the presence of outlier classes and their impact on average score distribution, the correlation coefficient between the class average score and its standard deviation has been computed.

The correlation (Table 1) is significantly negative for the primary level classes (-0.7 for the second year and -0.6 for the fourth year) and it's close to zero for the secondary ones.

TABLE 1. - *Correlation between class mean score and its standard deviation*

School level	*2004/2005*			*2005/2006*		
	Reading	Mathematics	Science	Reading	Mathematics	Science
Second class of primary school	−0.725	−0.731	−0.740	−0.729	−0.758	−0.688
Fourth class of primary school	−0.643	−0.442	−0.688	−0.580	−0.329	−0.598
First class of lower secondary school	−0.224	−0.154	−0.406	−0.266	0.153	−0.180
First class of upper secondary school	−0.212	0.201	−0.024	−0.158	0.280	0.156
Third class of upper secondary school	0.126	0.217	0.094	0.208	0.403	0.236

[1] For a comparison among the distributions of all school levels see: Quintano, Castellano and Longobardi (2007).

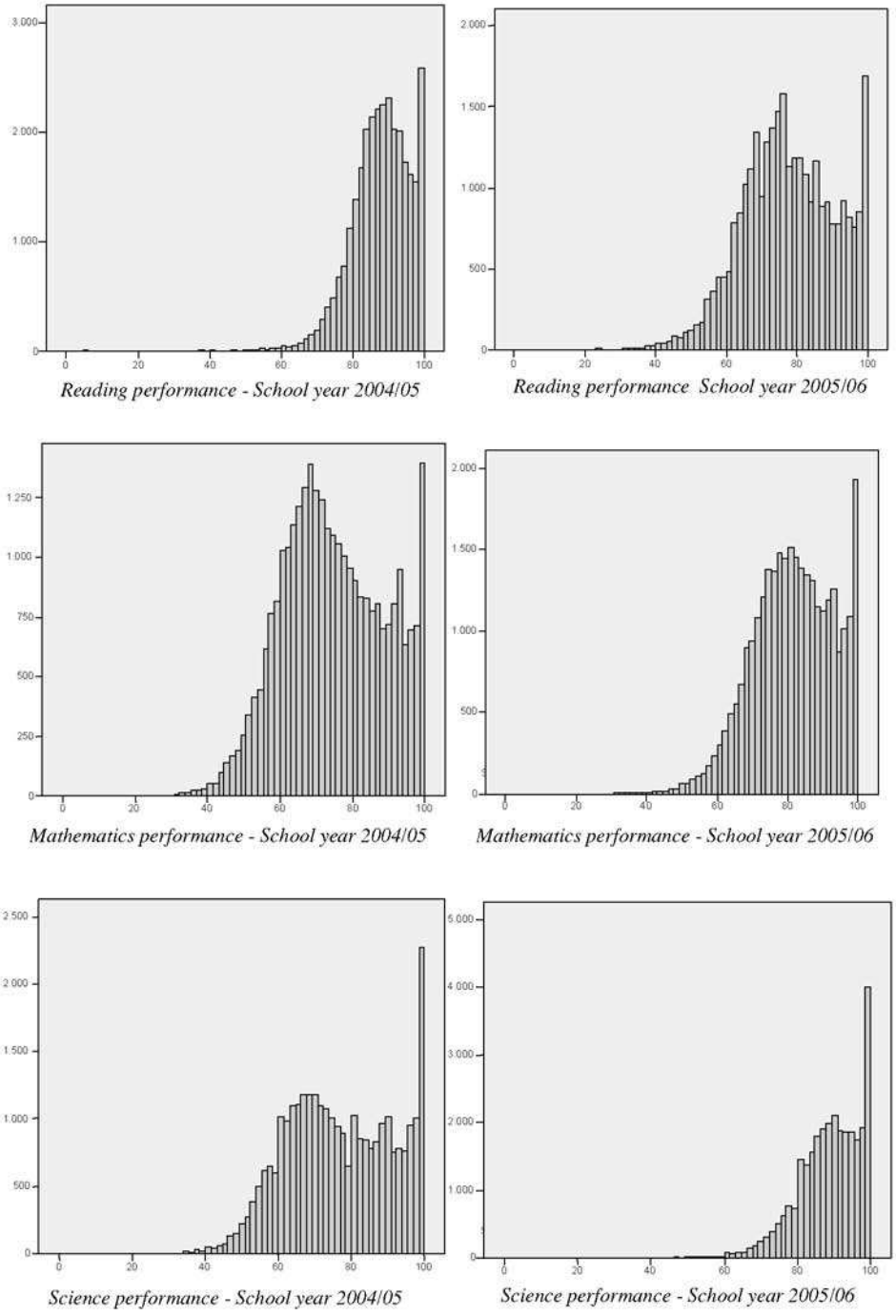

FIGURE 1. - *Distributions of mean score, at class level, of student performance in second-class of primary school, school years 2004/05 and 2005/06*

The values of this coefficient stand out in that the increase of the average class score is related to the reduction of within class variability but these findings are consistent only for the primary class students.

Excessive teacher support to pupils, due to their young age, could be a plausible determinant of this anomaly. Indeed, only the classes of very young students – especially the second class of primary school – are affected by this irregularity. For this reason, the detection and correction method has been limited to the primary level data. In the Table 2 some information about the structure of the primary school dataset are reported.

TABLE 2. - *Structure of the student dataset for second and fourth classes of primary school*

School level	*School year*	*Schools*	*Classes*	*Students*	*Average class size*	*Standard deviation of class size*
Second class of primary school	2004/05	7,497	30,031	556,231	18.39	5.27
	2005/06	7,517	29,996	544,704	18.16	5.36
Fourth class of primary school	2004/05	7,481	29,712	534,936	18.00	5.28
	2005/06	7,524	30,091	546,460	18.16	5.35

3. IDENTIFICATION AND CORRECTION OF OUTLIER UNITS

The proposed procedure is aimed at managing the presence of outlier classes and consequently, at improving the quality of the survey. The methodology classifies a class as outlier if the within variability of the final score is close to zero and there is a low percentage of missing data.

The detection and correction procedure consists of two steps:

- At the first step the units, at students level, with too many missing or invalid answers have been erased. Then, some homogeneity indexes at class level, have been computed.
- At the second step, an index has been computed which expresses, for each class, the degree of belonging to an outlier cluster. Then, on the basis of this membership index, a correction factor has been elaborated to adjust the average class score distribution.

The procedure has been applied to all data collected from primary students (second class and fourth class) who had participated to INVALSI survey in the school year 2004/05 and 2005/06. Furthermore, by comparing the results of the correction procedure, the distribution patterns look very similar in terms of both school year and assessment area. Consequently, we will limit the comment to the mathematics performance of second primary student in the school year 2004/2005.

3.1 *Data cleaning procedure and computation of class level indicators*

Primarily, the micro units – students – who have not given the minimum number of answers to compute a performance score are considered as ''pseudo-non respondents'' and consequently 47,884 units (8.6%) have been dropped from dataset (listwise deletion).

After this data cleaning procedure, the following indexes, at class level, are computed:

Class mean score $\bar{p}_j$:

$$\bar{p}_j = \frac{\sum_{k=1}^{N_j} p_{kj}}{N_j} \tag{1}$$

where:
p_{kj} denotes the score of k^{th} student of j^{th} class
N_j denotes number of respondent students of j^{th} class

Class standard deviation score σ_j:

$$\sigma_j = \sqrt{\frac{\sum_{k=1}^{N_j} (p_{kj} - \bar{p}_j)^2}{N_j}} \tag{2}$$

where:
p_{kj} denotes the score of k^{th} student of j^{th} class
$\bar{p}_j$ denotes the class j^{th} mean score
N_j denotes number of respondent students of j^{th} class

Class non-response rate MC_j:
The class non-response rate MC_j expresses the collaboration of each class to respond to all the test questions, equal to

$$MC_j = \frac{\sum_{k=1}^{N_j} M_{kj}}{N_j Q} \tag{3}$$

where:
M_{kj} denotes the number both of item non-responses and of invalid responses for the k^{th} student of the j^{th} class
Q denotes the number of administered item to j^{th} class. It is a constant for each assessment area (reading, mathematics and science) and for each school level
N_j denotes the number of respondent students of j^{th} class

The values of this class non-response rate vary in the range $0-1$. It is equal to 0 when there are no missing or invalid responses for the j^{th} class, while it reaches its maximum (equal to 1) when all students of j^{th} class have given only missing or invalid answers.

Class index of answer homogeneity $\bar{E}_j$:
The index of answer homogeneity $\bar{E}_j$ is developed based on Gini's measure of heterogeneity. It is the following:

$$\bar{E}_j = \frac{\sum_{s=1}^{Q} E_{sj}}{Q} \tag{4}$$

It is the mean of the Q Gini indexes (E_{sj}) computed for each s^{th} test question.

The numerator of $\bar{E}_j$ is a Gini measure of homogeneity E_{sj} and it is computed for each s^{th} test question administered to each student of j^{th} class:

$$E_{sj} = 1 - \sum_{t=1}^{h} \left(\frac{n_t}{N_j}\right)^2 \tag{5}$$

Where:

$\frac{n_t}{N_j}$ denotes the relative frequency of students of j^{th} class that has given the t^{th} answer to s^{th} question.

The Gini measure is equal to zero when all students of j^{th} class have given the same answer to the s^{th} question, while it reaches the maximum value: $(h-1)/h$ (h is the number of alternative answers to question s^{th}) when there is perfect heterogeneity of answers to s^{th} question in the j^{th} class.

Thus, $\bar{E}_j$ is the mean for each j^{th} class of the Q Gini indexes and it is between 0 and $(h-1)/h$, inclusive. It is equal to zero when all students of j^{th} class have given the same answers to all test questions, while it reaches the value $(h-1)/h$ when in the j^{th} class the answer heterogeneity is maximum.

Summarising, the first stage of editing procedure consists in deleting the non response units, at student level, and then in computing the following indexes of class answer behaviour:

- Class mean score $\bar{p}_j$
- Class standard deviation score σ_j
- Class non-response rate MC_j
- Class index of answer homogeneity $\bar{E}_j$

3.2 *Dimensionality reduction by Principal Component Analysis (PCA)*

At the second step, the size of the data matrix, composed of the four indexes at class level, is reduced to two components by using Factor Analysis with a principal component extraction (Jolliffe, 2002).

The first two principal components account for 92% of the total variance (Table 3).

TABLE 3. - *Eigenvalues of correlation matrix R, simple and cumulative percentage of explained variability by the principal component analysis*

COMPONENT	*Initial Eigenvalues*		
	TOTAL	*% of Variance*	*Cumulative %*
1	2.956	73.911	73.911
2	0.723	18.086	91.997
3	0.288	7.211	99.208
4	0.032	0.792	100.000

The component matrix, Table 4, shows the correlations between the four observed variables and the first two principal components.

TABLE 4. - *Correlation between the indexes of class answer behaviour and the first two components correlations*

Observed variable	*Component*	
	1	*2*
Class mean score $\bar{p}_j$	–0.946	0.117
Class standard deviation score σ_j	0.880	–0.134
Class non-response rate MC_j	0.670	0.742
Class index of answer homogeneity $\bar{E}_j$	0.940	–0.286

The first component is highly correlated with class mean score and with the two variability indicators.

The correlation between the first factor and the class mean score is significantly negative (–0.946), while the relationship between the same factor and the two indexes of variability (standard deviation and index of answer homogeneity) is positive (0.880 and 0.940 respectively). These values suggest that the first component might be interpreted as an ‘‘outlier identification axis’’.

The second component is most highly correlated with class non-response rate, thus this factor might be defined as the ‘‘index of class collaboration to survey’’.

A plot of the variables (Figure 2) where each factor loading is a coordinate along the corresponding factor axis is useful in order to illustrate the interpretation of the axes graphically.

As reported in Figure 2, the class mean score has a high negative loading on the first factor, while the within variability indexes show a positive loading and then these variables are projected on the first axis in opposite position in respect to the mean score. The position of observed variables on the factorial plane is also explained by the negative correlation between the class average score and its’ standard deviation (Table 1).

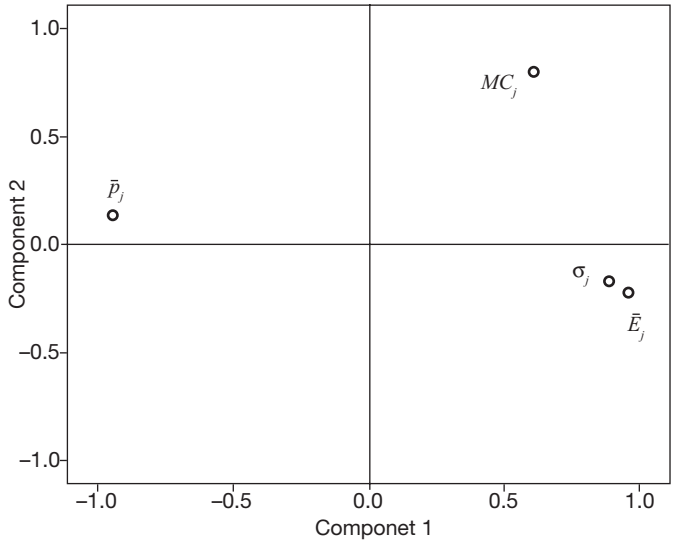

FIGURE 2. - *Loading plot corresponding to the first two components*

The axis interpretation suggests that the points-classes on factorial plane with negative first factor scores will be considered as outlier classes since they are distinguished by high average scores and minimum, close to zero, within variability; while the points-classes on the first quadrant and positive first factorial scores might be considered as not outlier classes since they have within variability more than zero and the class average score lower than the maximum.

The second principal component is considered as an ''index of class collaboration to survey''. Since the class non-response rate has a positive loading on the second factor (0.742), the student classes with a low number of missing data will be distinguished by high second factor scores.

3.3 *Outlier detection by the Fuzzy k-Means approach*

The classification of the outlier classes is based on a fuzzy classification approach – the *Fuzzy k-Means* (FKM) – developed by Bezdek (1981) and Dunn (1974).

The *Fuzzy k-means* is a fuzzy version of the non-overlapping partition model *hard k-means* and it is based on the generalized fuzzy variance criterion:

$$J_{FKM} = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^{m} \left\| x_j - v_i \right\|^2 \tag{6}$$

Where $u_{ij} \in [0,1], \sum_{i=1}^{c} u_{ij} = 1$ represents the membership degree of object

$j(1 \leq j \leq n)$ in group $i\,(1 \leq i \leq c)$, $V = [v_1, v_2, ..., v_c]$, $v_i \in [R^n]$ is a vector of cluster centers, and $\left\|x_j - v_i\right\|^2$ is the Euclidean norm between x_j *and* v_i.

The extension is made by introducing a weight $m(1 \leq m \leq \infty)$, named '*fuzziness factor*', which characterizes the family.

If $m = 1$, the obtained solution is a non-overlapping partition. If m tends to infinity then the membership degree values for each class become close to $1/c$. The fuzzy partition degree grows with m, and Pal and Bedzek (1995) underline that the Fuzzy k-means provides better performance for m in the range 1.5-2.5.

The cluster centroids and the respective membership functions that solve the minimization problem of the J_{FKM} function are:

$$v_i = \frac{\sum_{j=1}^{n} u_{ij}^m x_j}{\sum_{j=1}^{n} u_{ij}^m}, \qquad 1 \leq i \leq c \tag{7}$$

$$u_{ij} = \left[\sum_{k=1}^{c} \left(\frac{\left\|x_j - v_i\right\|^2}{\left\|x_j - v_k\right\|^2} \right)^{1/(m-1)} \right]^{-1}, \; 1 \leq i \leq c, \; 1 \leq j \leq n \tag{8}$$

The optimization of the classification strategy is subdivided into five stages: Initialization (I and II), Iteration (III and IV) and Stop Criterion (V).

I) Determining the cluster number c and fixing the fuzziness parameter m (see Section 3.4).
II) Calculation of the group centroids v_i using the expression (7).
III) Construction of a new fuzzy partition matrix (determination of the new membership values) using equation (8). If an object j keeps a distance *0* from the centre of cluster i, the value of u_{ij} is equal to 1 and the membership values of j towards the remaining classes is equal to 0.
IV) Calculation of the group centroids associated to the partition determined in III.
V) If the improvement in J_{FKM} is less than a certain threshold (e) the steps III and IV are iterated.

The final output of the fuzzy k-means is a matrix where for each class we report the membership degree to every cluster.

By the projection of the centroids on the factorial axes, it is possible to detect the cluster centroid with an outlier profile on the basis of the principal component interpretation.

Then the steps of the correction procedure are:

- clustering the classes by fuzzy k-means algorithm
- the projection on the factorial axes of the cluster centroids.
- the detection of outlier cluster centroid

- the computation, for each class, of a correction factor on the basis of degree of belonging to outlier cluster centroid
- the correction of the class average score in proportion of this membership.

A key stage of the whole procedure is the fuzzy clustering calibration i.e. the choice of the fuzzy clustering parameters: number of clusters (c) and fuzziness level (m).

3.4 *Calibration strategy of fuzzy clustering approach*

The fuzzy clustering is an unsupervised learning technique and then the results of this technique are affected from the choice of two parameters: number of the clusters (c) and fuzziness index (m).

The calibration strategy followed in this paper consists of two steps: firstly, the optimal number of the clusters is established by computing some validity measures, secondly m is determined by analyzing the sensitivity of the final results from the FKM algorithm (correction weight assigned to each student class) to variation in the level of the fuzziness.

The validity measures computed to assess the goodness of the fuzzy partitions and to obtain the optimum number of c are Fuzziness Performance Index (FPI), Normalized Classification Entropy (NCE) and Separation index (S). The values of these indices are calculated for m equals to 1.5, 2.0 and 2.5 for checking if any difference exists in the general structure of the indices for different fuzziness parameters. The upper boundary of the number of the clusters is determined as 20.

The fuzzy performance index (FPI) is defined as (Roubens, 1982):

$$FPI = 1 - \frac{cPC}{c-1} \tag{9}$$

Where PC is the partition coefficient proposed by Bedzek (1974) to measure the amount of overlap between clusters:

$$PC = \frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \tag{10}$$

The Normalized Classification Entropy (Roubens, 1982) is denoted by:

$$NCE = \frac{PE}{\log c} \tag{11}$$

Where PE is the partition entropy (Bedzek, 1981):

$$PE = -\frac{1}{n}\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\log u_{ij} \tag{12}$$

The FPI and MPE are used for evaluating the fuzziness of the solutions. The low-

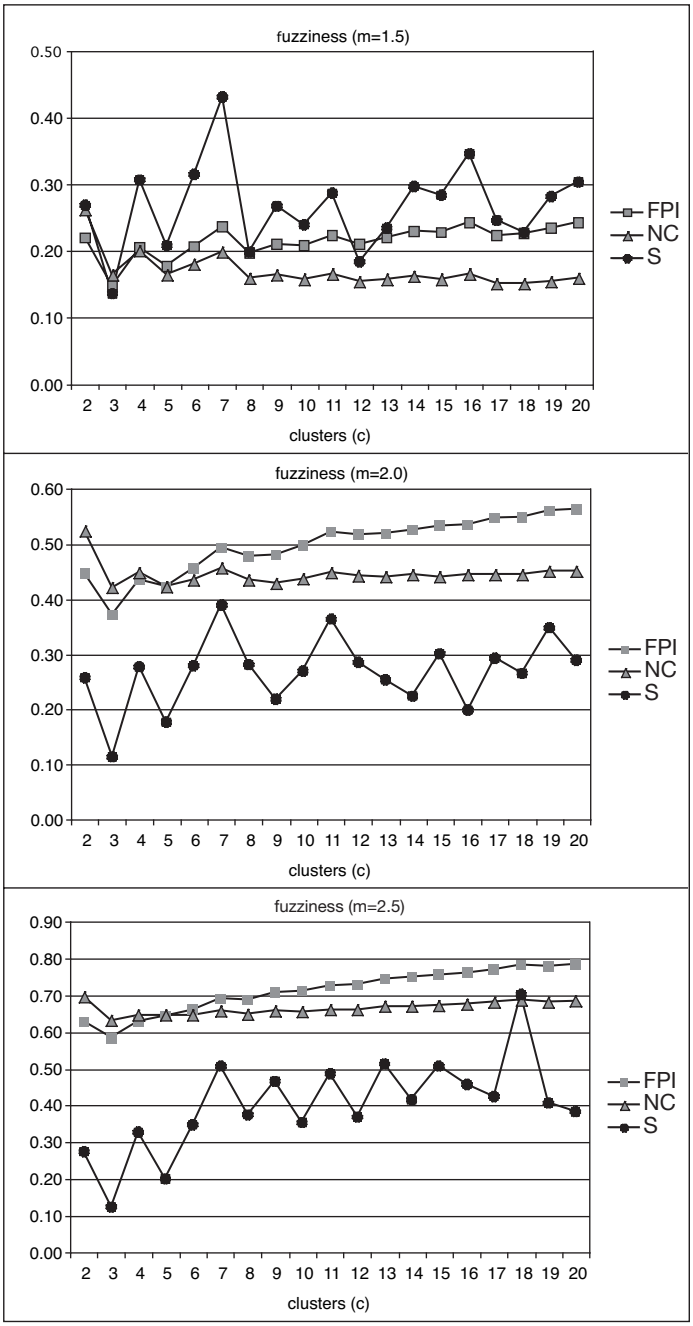

FIGURE 3. - *Fuzzy performance Index (FPI), Normal Partition Entropy (NPE) and Separation index (S) in correspondence of three level of fuzziness (m=1.5, 2.0, 2.5) and for c∈[2,20]*

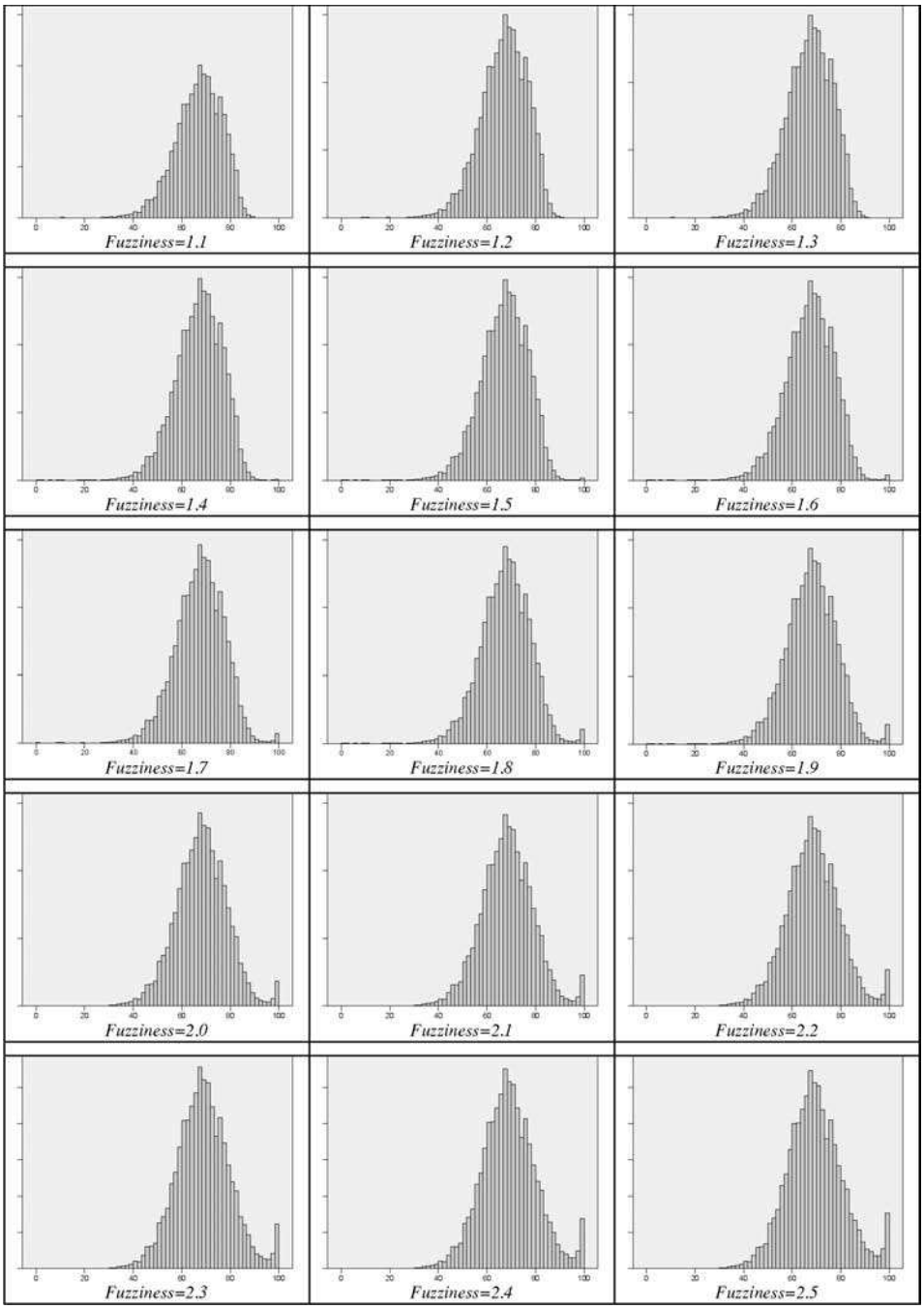

FIGURE 4. - *Distributions of adjusted class mean score for different levels of fuzziness: $1.1 \leq m \leq 2.5$ and for $c = 3$*

er the FPI and MPE values are, the more suitable is the corresponding solution (McBratney & Moore, 1985).

The third measure is the separation index S (Xie and Beni, 1991):

$$S = 1 - \frac{\sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}^2 \left\| x_j - v_i \right\|^2}{n \min_{i,j} \left\| v_i - v_j \right\|^2} \tag{13}$$

Where the numerator denotes the compactness by the sum of square distances within clusters, while the denominator denotes separation by the minimal distance between clusters. The smaller the value of S, the better the compactness and separation between the clusters.

The optimal number of clusters could be established on the basis of minimizing these three measures.

According to Figure 3, the 3-clusters solution shows the lowest values of each validity index. Moreover this choice is also the best solution for each level of fuzziness parameter m (1.5, 2.0, 2.5).

Given $c = 3$, the second step of the calibration is the choice of the fuzziness index (m). In this context, m is determined by analyzing the sensitivity of the results from the FKM algorithm (correction weight assigned to each student class) to variation in the level of the fuzziness. The final results are examined by varying m from 1.1 to 2.5 with an increment of 0.1.

Considering the distribution of mean score per class after the outlier correction (Figure 4), it's clear that the shape of weighted distribution tends to be closer to a normal distribution with low values of m, and the upward bias is minimized with $m \leq 2$.

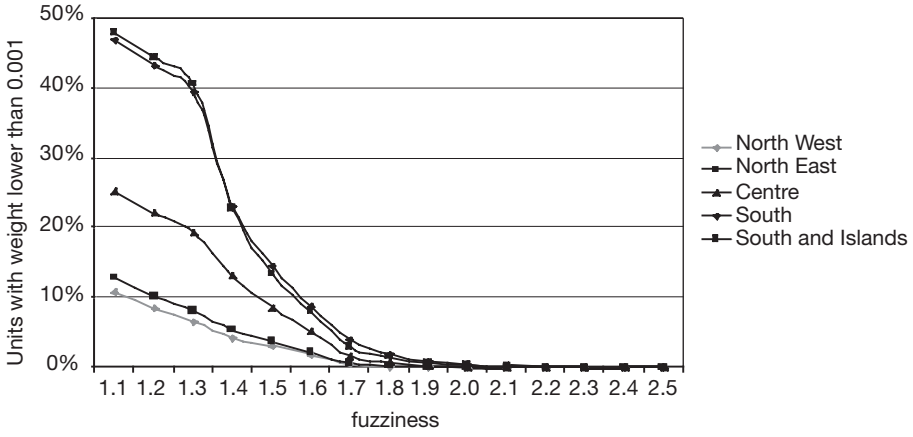

FIGURE 5. - *Variation of number of classes with a weight close to zero (lower than 0.001) in correspondence of increasing values of fuzziness level: $1.1 \leq m \leq 2.5$ and for $c = 3$*

In addition, for values of $m < 1.6$ the correction procedure excludes from the dataset an high rate (higher than 15%) of Southern student classes.

The Figure 5 shows the percentage of classes, for each area, with a weight close to zero (lower than 0.001) in correspondence of different values of m. This evidence leads to consider a value of $m \geq 1.6$.

Focusing on the distribution of mean score of the Southern Italy (including Sicily and Sardinia) classes (more affected by the outliers presence), the correction procedure tends to equalize the central tendency measures (mean, mode and median) for m in the range 1.7-1.9, while the mode reaches the maximum value (100) for m higher than 2.1 (Figure 6).

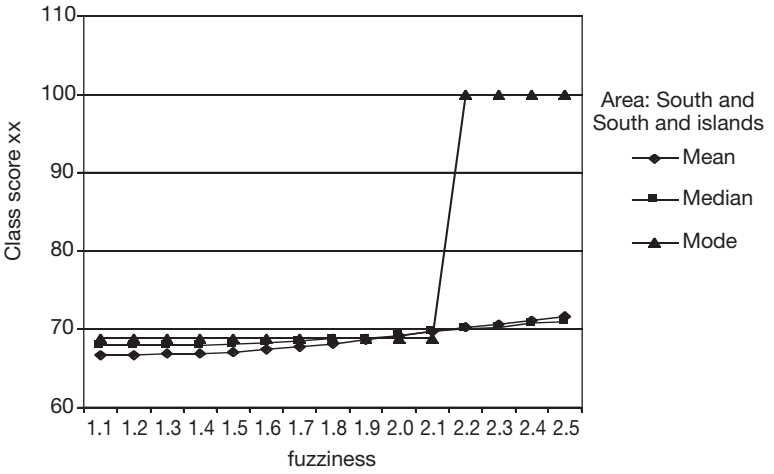

FIGURE 6. - *Variation of some weighted central tendency measures of class mean score with increase in the value of fuzziness level: $1.1 \leq m \leq 2.5$ and for $c = 3$*

Finally, according to these findings, an optimal choice of the fuzziness level could be m=1.7, this solution seems to be a good trade-off between the ''strength'' of the correction and the normalization of data distribution.

3.5 *A measure of the outlier level*

After the calibration stage, on the basis of the two factorial dimensions the student classes are classified in c=3 clusters with the fuzziness parameter m equal to 1.7. The Figure 7 shows the projection on the factorial plane of the cluster centroids.

Based on the principal component interpretation, Cluster 2 gathers the classes that present an outlier profile. This cluster is distinguished by:

- high negative scores on the ''outliers identification axis'' (x-axis) that characterizes high class average scores and minimum within variability with respect to scores and test answers;

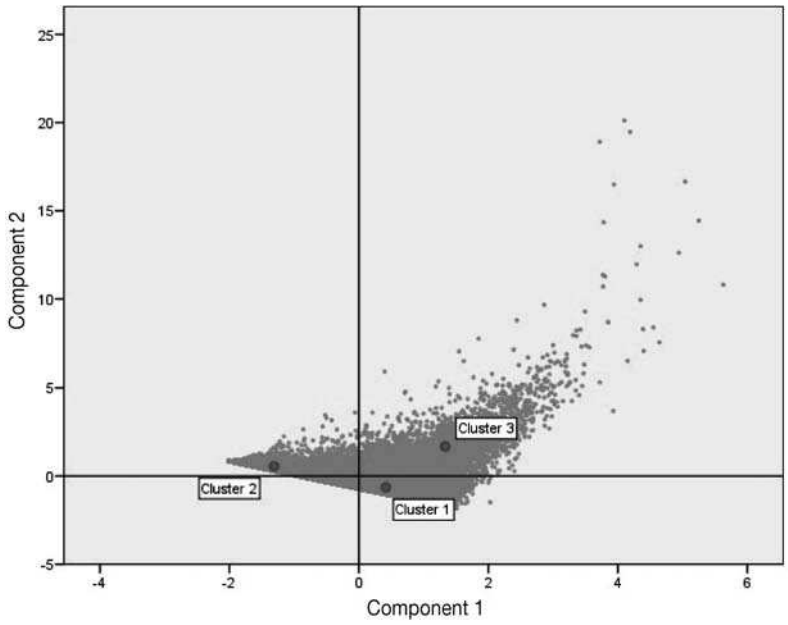

FIGURE 7. - *Projection on factorial plane of centroids computed by fuzzy k-means algorithm, c = 3, m = 1.7*

- factorial scores close to zero with respect to the ''index of class collaboration to survey'' (y-axis) that indicates a low presence of missing items in the class data and a full compilation of the performance test.

Denoting the outlier cluster with the term ''a'', for the j^{th} class the degree of belonging to this cluster is equal to:

$$u_{aj}$$

this measure varies from 0 to 1 and it is the membership to the outlier cluster or otherwise it can be interpreted as a measure of ''outlier level'' of the j^{th} class. Then the correction factor of average score of class j^{th} can be expressed as the complement to one of u_{aj}:

$$w_j = 1 - u_{aj} \tag{14}$$

This coefficient shall be used to weight the average score of each class in function of the outlier level of the class; then each class score will be weighted by this coefficient and the students' class with high degree of belonging to cluster 2 (outlier cluster) will have a low weight while the class very far from this cluster (low value of u_{aj}) will have a weight close to 1.

4. THE EFFECTS OF THE CORRECTION PROCEDURE

The basic inspiration principle of the whole procedure is to go over the dichotomous logic which classifies each unit as outlier or not outlier (hard clustering), in order to develop a fuzzy approach that allows us to compute an ''outlier level'' measure for each unit and consequently, to calibrate the correction in optimal way.

The impact of the correction procedure was analysed using a graphical comparison between the two distributions before and after the weight application (Figure 8).

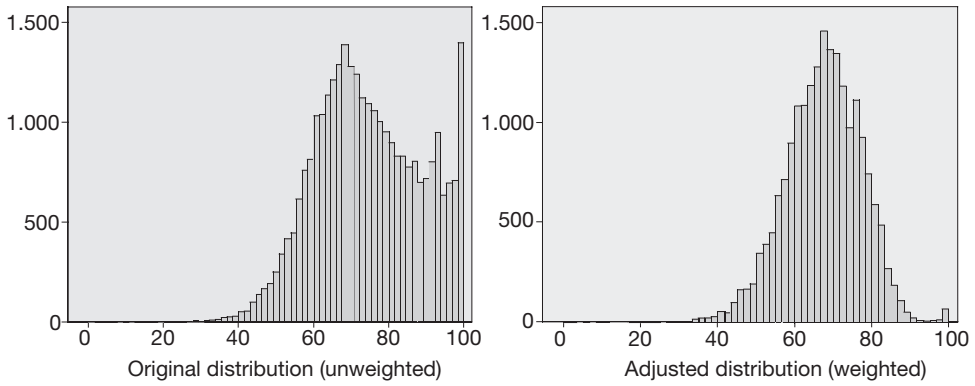

FIGURE 8. - *Comparison between the original class mean distribution and the adjusted one*

The comparison between the two distributions (original mean scores and weighted mean scores) shows that the shape of weighted distribution is closer to a normal distribution, although it is leptocurtik (kurtosis equal to 0.771) and it shows a light negative skewness (skewness index equal to –0.287).

Again, it is distinguished from the prior distribution by the lack of two modes and by the reduction of the high frequencies peak in correspondence of the higher values of the variable.

Focusing on the descriptive statistics of Table 5, the values of the second and third quartile are decreased and after the correction, the mean, the median and the mode of distribution are quite close to one another.

TABLE 5. - *Comparison between unweighted average score per class and the weighted score per class according to the factor* $w_j = 1 - u_{aj}$

	Original distribution	*Adjusted distribution*
MEAN	74.71	67.39
MODE	100.00	68.75
I QUARTILE	64.42	60.93
MEDIAN	73.61	67.78
III QUARTILE	85.94	74.25

5. THE GEOGRAPHICAL LOCALIZATION OF STUDENTS' CLASS AS DETERMINANT OF THE PRESENCE OF OUTLIERS

An analysis of ''outlier level'' coefficient (u_{aj}) distributions by geographical regions is performed in order to evaluate the relationship between the school location and the presence of outlier classes. The analysis is carried out by the comparison of box plots of ''outlier level'' by regions (Figure 9).

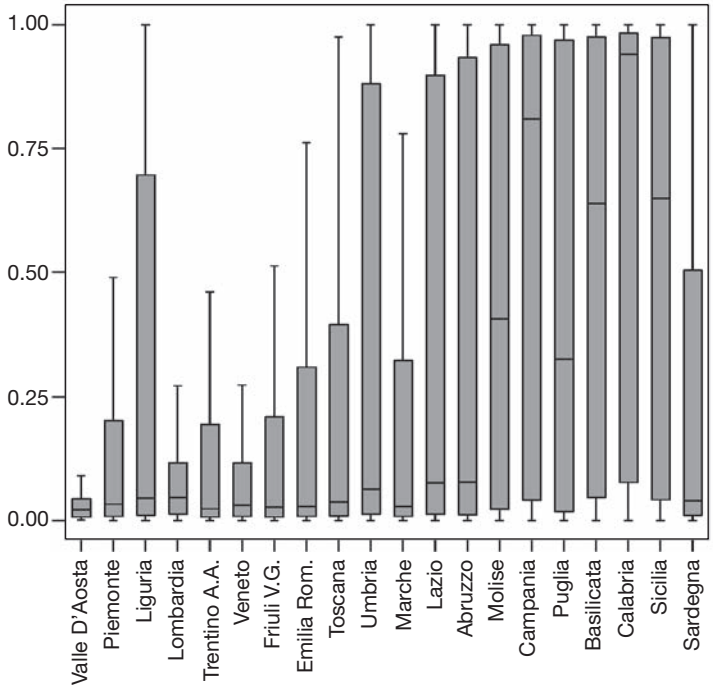

FIGURE 9. - *Graphics by box plot of the index u_{aj} distributions (j-th unit degree of belonging to outlier points group)*

The box plots allows us to classify the regions into two groups:

- The first group includes the regions of Central and Northern Italy, the classes of these regions show a low probability of being considered as outlier units. Indeed, the median of u_{aj} for the region including in this group vary between 0 and 0.08.
- The second group encompasses all regions of Southern Italy that are distinguished by higher values of u_{aj}. Particularly, the third quartile of outlier coefficient is higher than 0.9 for these regions. This means that 25% of students' classes of Southern Italy might be considered outliers with a probability superior to 0.9.

Then the considerable difference between the Southern regions distributions of u_{aj} and those of the Northern and the Central regions leads us to suppose that the

anomalies of the average score distribution at national level are generated by a suspicious answer behaviour limited to the Southern classes.

To confirm this hypothesis, it is interesting to observe the class average score distributions only for the Northern and Central Region students' classes (Figure 10).

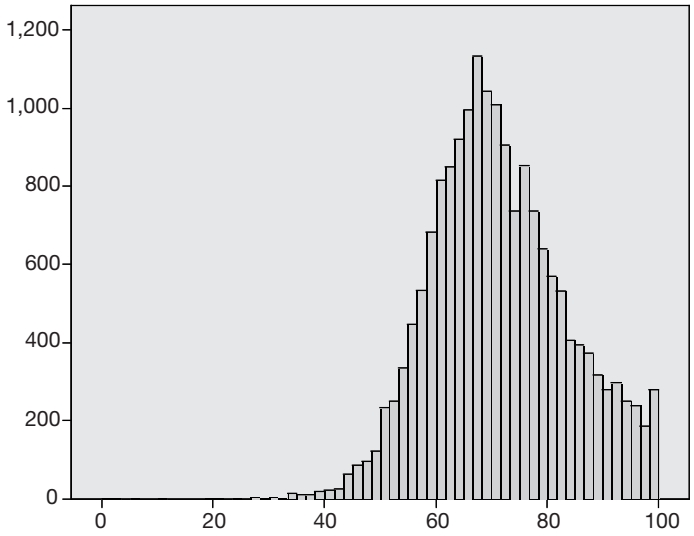

FIGURE 10. - *Unweighted average score per class computed on the data collected from second class primary students participating to mathematics assessment in the school year 2004/2005 from Northern and Central Regions*

This distribution does not show the anomalies of the national score distribution, in fact, it is unimodal and it does not show high peaks of frequencies in correspondence of the highest values of the variable.

Consequently, the upward bias on the distribution of the average scores by class would be ascribed to the presence of outlier classes concentrated in the Southern Italy. To explain these regional disparities it is supposed that the primary teachers have provided excessive support to the pupils during the performance test.

This motivation might explain the anomaly homogeneity of within class answer and the high average score. After the weighting procedure, the score difference in favour of Southern regions is decreased and the regional adjusted scores show lower differences in comparison of the original -unweighted- ones (Figure 11).

6. CONCLUSIONS

In this paper, an outlier detection and correction procedure is developed in order to improve the accuracy of data collected by the Italian National Evaluation Institute of the Ministry of Education (INVALSI).

The INVALSI survey aims to evaluate, every year, the student's knowledge of

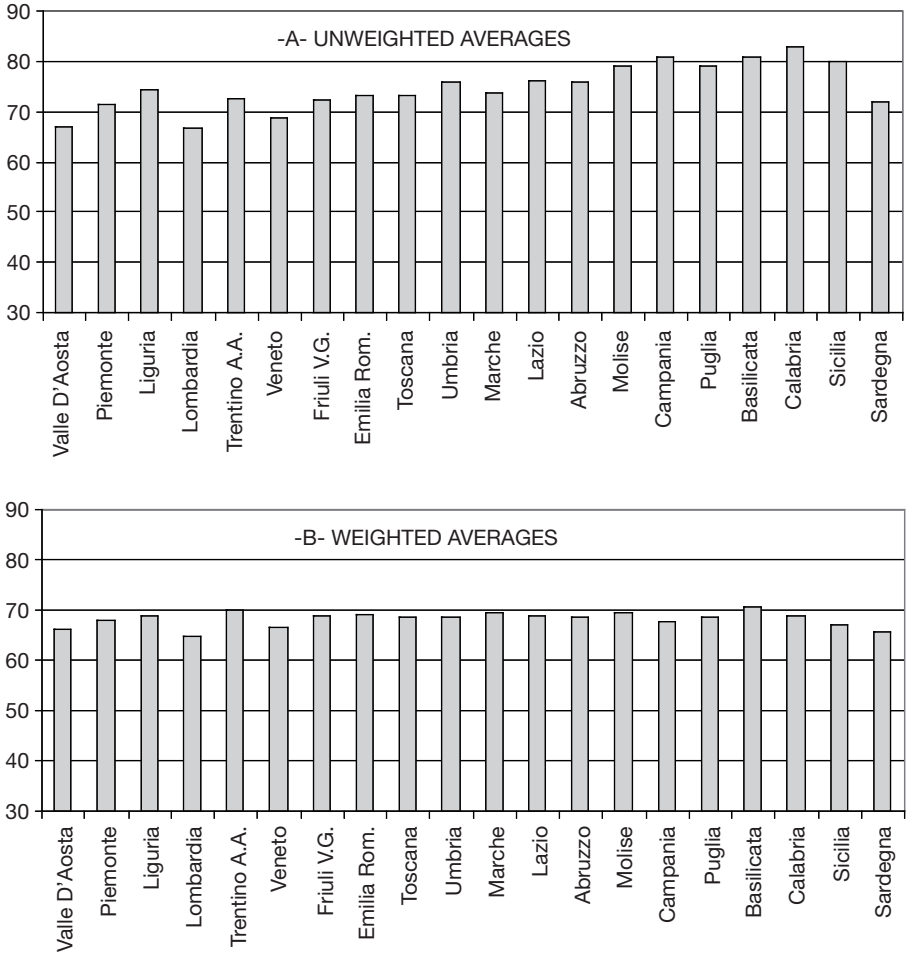

FIGURE 11. - *Unweighted and weighted average score per class*

reading, mathematics and science at primary and secondary level. The questionnaires are administered by the teacher of each class.

The tests are made up of a different number of items on the basis of the school level and the assessment area.

Every dataset, at student level, is created for each school level and assessment area (total 15 datasets) and contains the following variables: gender, region, school, class, item answers and student final score.

Looking at these data we noted too many classes of students distinguished by a mean score corresponding to the maximum value (100 points) and this effect is emphasized for students at primary school.

This anomaly leads us to suppose that many primary school teachers have provided excessive support to the pupils during the performance test. Consequently, the

computed score for each student of some classes may be subject to some bias due to teacher intervention.

In this context, teacher support might be considered similar to an interviewer effect (Biemer, Groves, Lyberg, Mathiowetz and Sudman, 1991) and then we might suppose that the student's score is affected by an error component which inflates the measurement errors.

Under these conditions, we have considered as outliers the classes distinguished by the average score close to the top score (100 points) and a within variability of the answers close to zero.

A specific approach is developed to detect the outlier units in such hierarchical structure where the schools are the primary units, the classes the secondary units and the pupils the tertiary ones.

The proposed procedure consists of two steps:

- At the first step, the units, at student level, with too many missing or invalid answers have been erased. Then, an homogeneity index, at class level, has been computed.
- At the second stage, an index has been computed which expresses, for each class, the degree of belonging to an outlier cluster. Then, on the basis of this membership index, a correction factor has been elaborated to adjust the average class score distribution.

On the basis of this approach we derived a set of modified distributions of primary class scores. The effect of the adopted procedure seems to show a shape closer to a normal distribution, but with a slight skewness.

Furthermore, the analysis of correction factor distribution by Italian regions has allowed to study the geographical distribution of outlier units and to highlight the strong relationship between outlier presence and the localization of students' classes. This evidence brought us to hypothesize the presence of some problems in the assessment system of INVALSI.

Finally, these findings suggest revision of the data collection procedures, especially the administration of the questionnaire, in order to avoid the presence of outliers and to improve the data quality.

ACKNOWLEDGEMENT

This paper was supported by the 2008 Endowment Funds of the Department of Statistics and Mathematics for Economic Research of University of Naples ''Parthenope'' in the framework of the current studies on ''The Italian scholastic system: evaluation and intervention proposals by the analysis of the international and national student assessment surveys''.

RIASSUNTO

Il lavoro propone un nuovo approccio di analisi degli outlier che si focalizza su dati con struttura gerarchica ed un pattern di variabilità complesso (studenti e scuole, dipendenti e società, pazienti e reparti, etc.). In particolare, si cerca di ''mitigare'' l'influenza di unità anomale che caratterizzano

i dati degli studenti della scuola primaria rilevati dall'Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e Formazione (INVALSI). Sulla base della variabilità intraclasse è stata sviluppata una procedura che combinando l'analisi fattoriale con le tecniche di fuzzy clustering permette di identificare le classi di studenti anomale. Il criterio ispiratore dell'intera procedura di correzione è quello di attribuire ad ogni unità un peso opposto alla probabilità di appartenere al cluster di unità anomale, in tal modo si supera il limite della logica dicotomica di classificare in modo ''drastico'' un'osservazione come outlier o meno (hard clustering), a favore di un approccio sfumato (fuzzy) che permette di quantificare, rispetto ad ogni classe, il livello di anomalia e conseguentemente di tarare adeguatamente l'intervento correttivo.

REFERENCES

Barnett V., Lewis T. (1994). *Outliers in Statistical Data*. Wiley, New York.

Bezdek J.C., (1974). Cluster validity with fuzzy sets. *Journal of Cybernetics,* **3**, 58-73.

Bezdek J.C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms.* Plenum Press, New York.

Biemer P.P., Groves R.M., Lyberg L.E., Mathiowetz N.A., Sudman S. (1991). *Measurement Errors in Surveys*. Wiley, New York.

Braverman M. (1996). Sources of Survey Error: Implications for Evaluation Studies. *New Directions for Evaluation: Advances in Survey Research*, **70**, 17-28.

Driankov D., Hellendoor H., Reinfark M. (1994). *An Introduction to Fuzzy Control*. Springer, New York.

Dunn J.C. (1974). A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well Separated Clusters. *Journal of Cybernetics*, **3**, 32-57.

Hawkins D. (1980). *Identification of Outliers.* Chapman and Hall, London.

Hodge V., Austin J. (2004). A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, **22**(2), 85-126.

Iglewicz B., Hoaglin D.C. (1993). *How to detect and handle Outliers*. ASQC Quality Press, Milwaukee.

Jolliffe I.T. (2002). *Principal Component Analysis*. Springer, New York.

Klir G., Folger T. (1988). *Fuzzy Sets, Uncertainty and Information*. Prentice Hall, New York.

McBratney A.B., Moore A.W. (1985). Application of fuzzy sets to climatic classification. *Agricultural and Forest Meteorology,* **35**, 165-185.

Pal N.R., Bezdek J.C. (1995). On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy systems*, **3**(3), 370–379.

Quintano C., Castellano R., Longobardi S. (2007). *Una procedura di qualità basata sulla fuzzy clustering per l'individuazione e la correzione dei dati anomali nell'ambito del Servizio Nazionale di Valutazione Scolastica degli apprendimenti (SNV)*. Progetto 032 Finvali 2005, available at: www.na.iac.cnr.it/finvali/reports.htm

Roubens M. (1982). Fuzzy clustering algorithms and their cluster validity. *European Journal of Operational Research*, **10**(3), 294-301.

Wentland E.J., Smith K.W. (1993). *Survey responses: An evaluation of their validity*. Academic Press, San Diego.

Xie X.L., Beni G. (1991). A validity measure for fuzzy clustering. *IEEE Transactions of Pattern Analysis and Machine Intelligence,* **13**, 841-847.