A fuzzy clustering approach to improve the accuracy of Italian students'data. An experimental procedure to correct the impact of the outliers on assessment test scores

Claudio Quintano[1], Rosalia Castellano[2], Sergio Longobardi[3]
[1]Univerisity of Naples "Parthenope", e-mail: claudio.quintano@uniparthenope.it
[2]Univerisity of Naples "Parthenope", e-mail : lia.castellano@uniparthenope.it
[3]Univerisity of Naples "Parthenope", e-mail: sergio.longobardi@uniparthenope.it

Abstract

The paper describes an experimental procedure for improving the accuracy of data collected by the Italian National Evaluation Institute of the Ministry of Education (INVALSI).
The INVALSI's survey is a national standardised assessment that aims to evaluate, every year, the student's knowledge of reading, mathematics and science at primary and secondary level. The paper focuses on the presence of outlier units, at class level, that may introduce an upward bias in the distribution of the average scores by class.
Then we propose a two-stage method for evaluating and correcting the overestimation of children ability that has been found at the primary classes.
At 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. 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.

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

1. Introduction

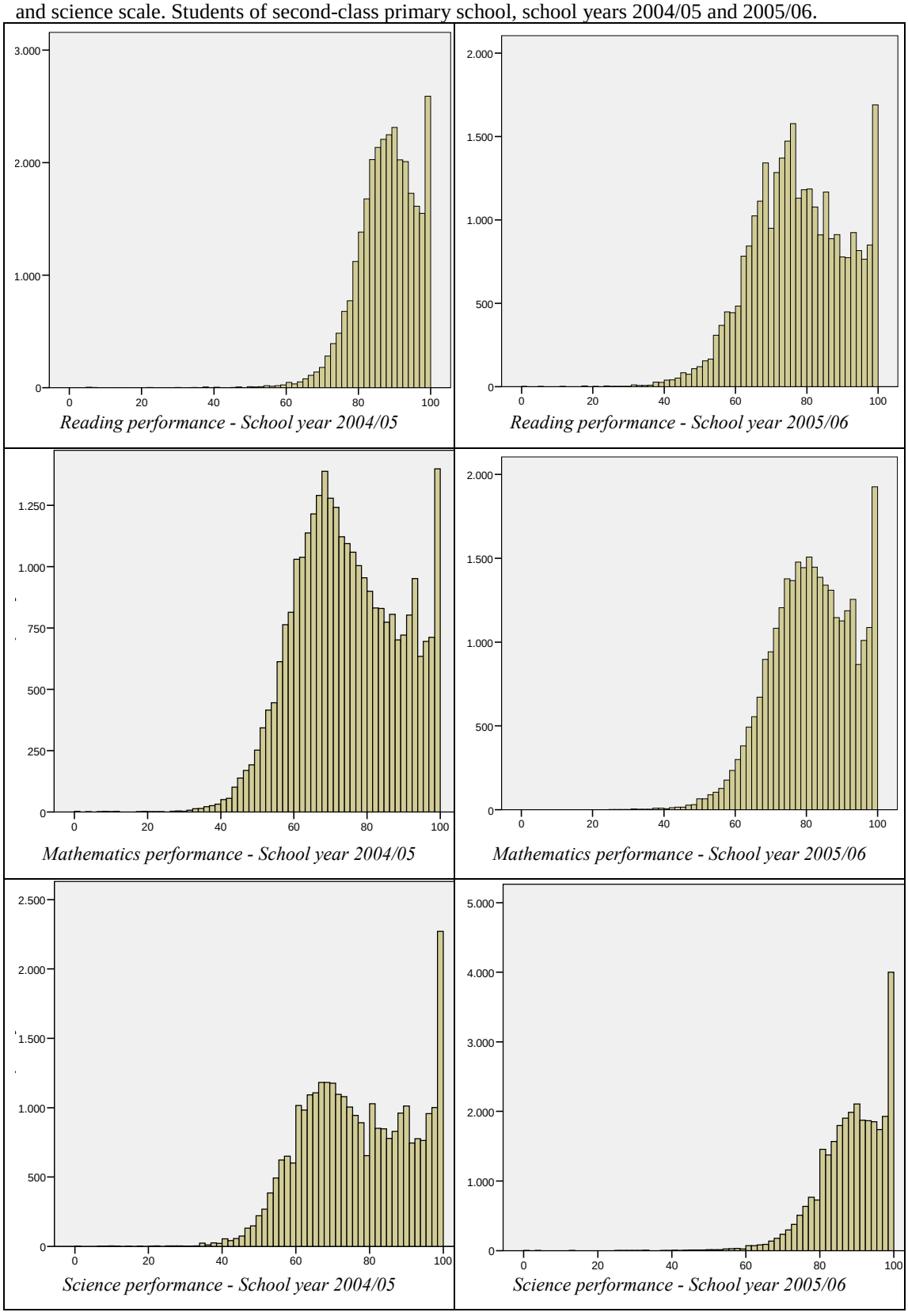
Outliers are generally identified as observations which appear to be inconsistent with the remaining of the data (Barnett and Lewis, 1994). Many studies focus on detection of outlier units (Hawkins, 1980; Hodge and Austin, 2004) and propose several methods to deal with this problem (Iglewicz and Hoaglin,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

micro units –students- are nested within classes and schools and take into account the 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 level class. Furthermore, we have adopted an innovative fuzzy approach. It allows 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 way calibrating the correction.
The paper considers data on students' 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. Every 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 analysis have showed the presence of outlier units, at class level, which brings to biased distributions of the average scores by class.
Consequently, the aim of this work is to improve the data quality of INVALSI survey by developing and implementing an editing procedure in order to find out and handle outlier classes of students.
This paper is structured as follows: in section 2 we analyze the features of class mean score distributions and highlight the outliers presence. In section 3 the developed procedure is described. Section 4 illustrates the effects of correction procedure on the original distributions. In section 5, we consider the relationship between the geographical localization of students' class and the outliers presence. Finally, in section 6 conclusions are made.

2. The average class score distributions

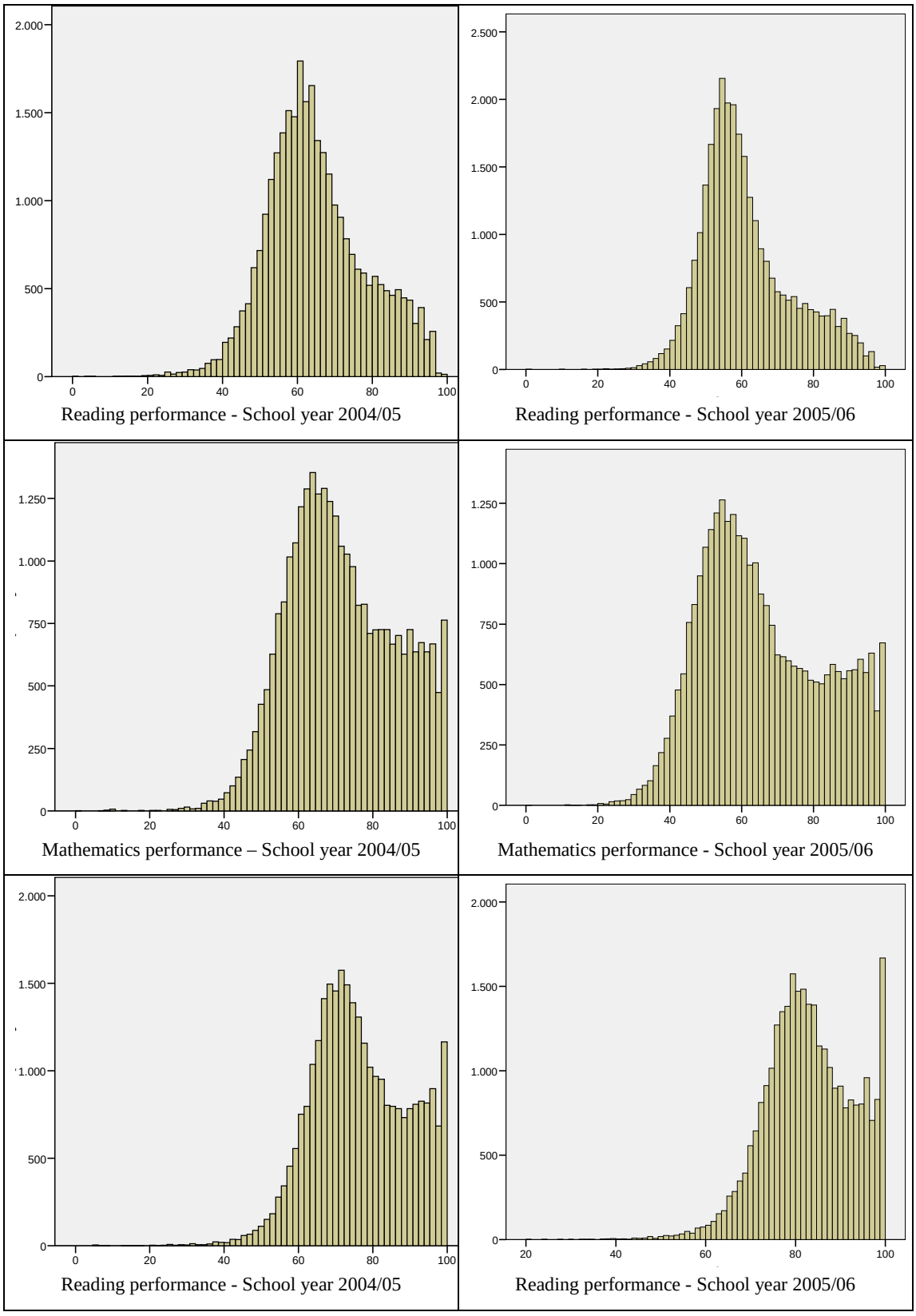
The results of explorative data analysis show an upward bias in the distributions of the average scores by class in each assessment area (reading, mathematics and science) only for the primary students.
Primary school data highlight the presence of many classes where a large percentage of students (close to 100%) has given the same answer to each question and, consequently, has received the same score. Furthermore, all answers are often correct and the whole class has achieved the top score (100 points).
The following figures show the distributions of performance mean score by class for each assessment area and all school levels in the school years 2004/05 and 2005/06.

Fig.1 - Distributions of mean scores, at class level, of student performance on the reading, mathematics and science scale. Students of second-class primary school, school years 2004/05 and 2005/06.



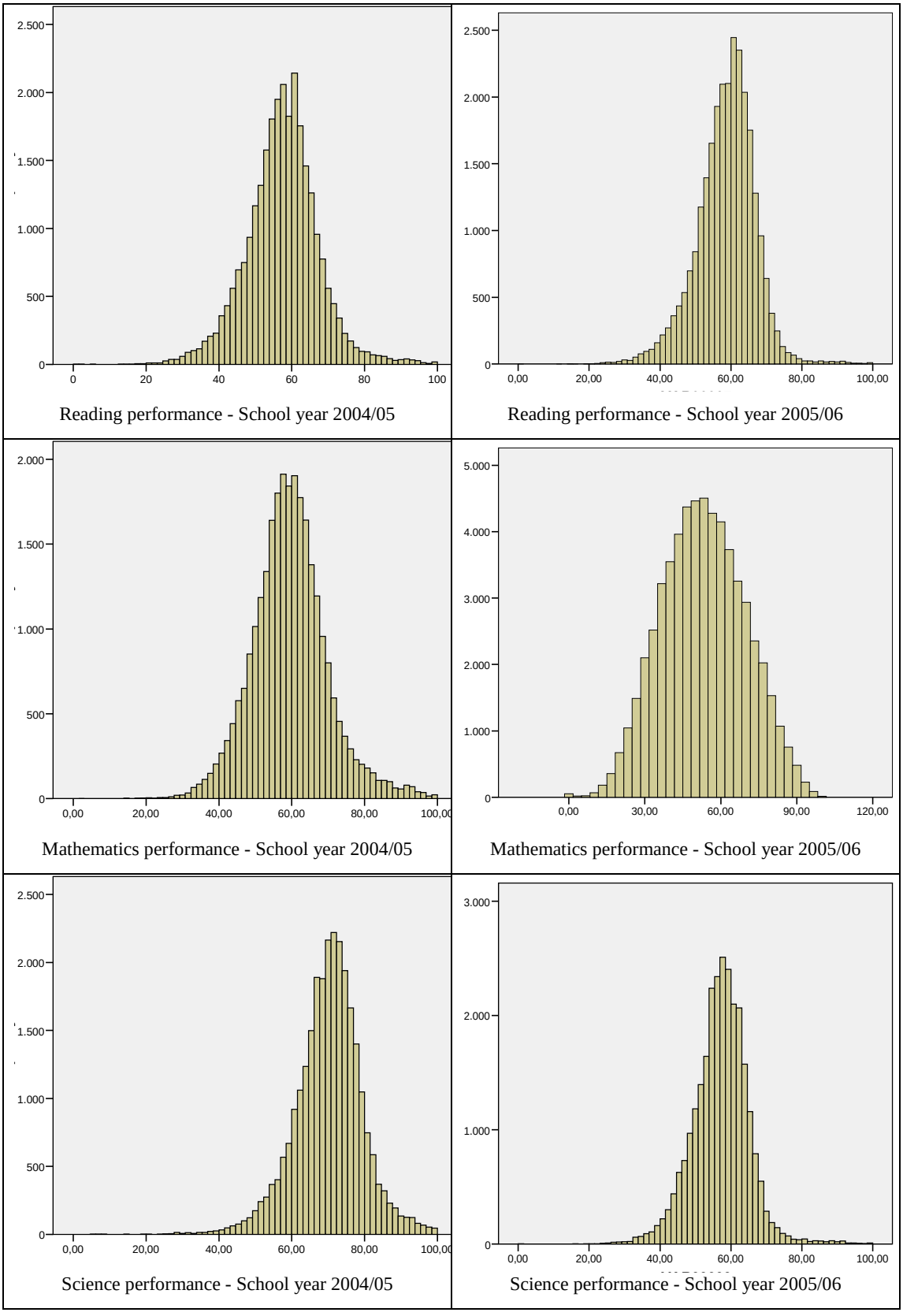
Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

Fig.2 - Distributions of mean scores, at class level, of student performance on the reading, mathematics and science scale. Students of fourth-class primary school, school years 2004/05 and 2005/06.



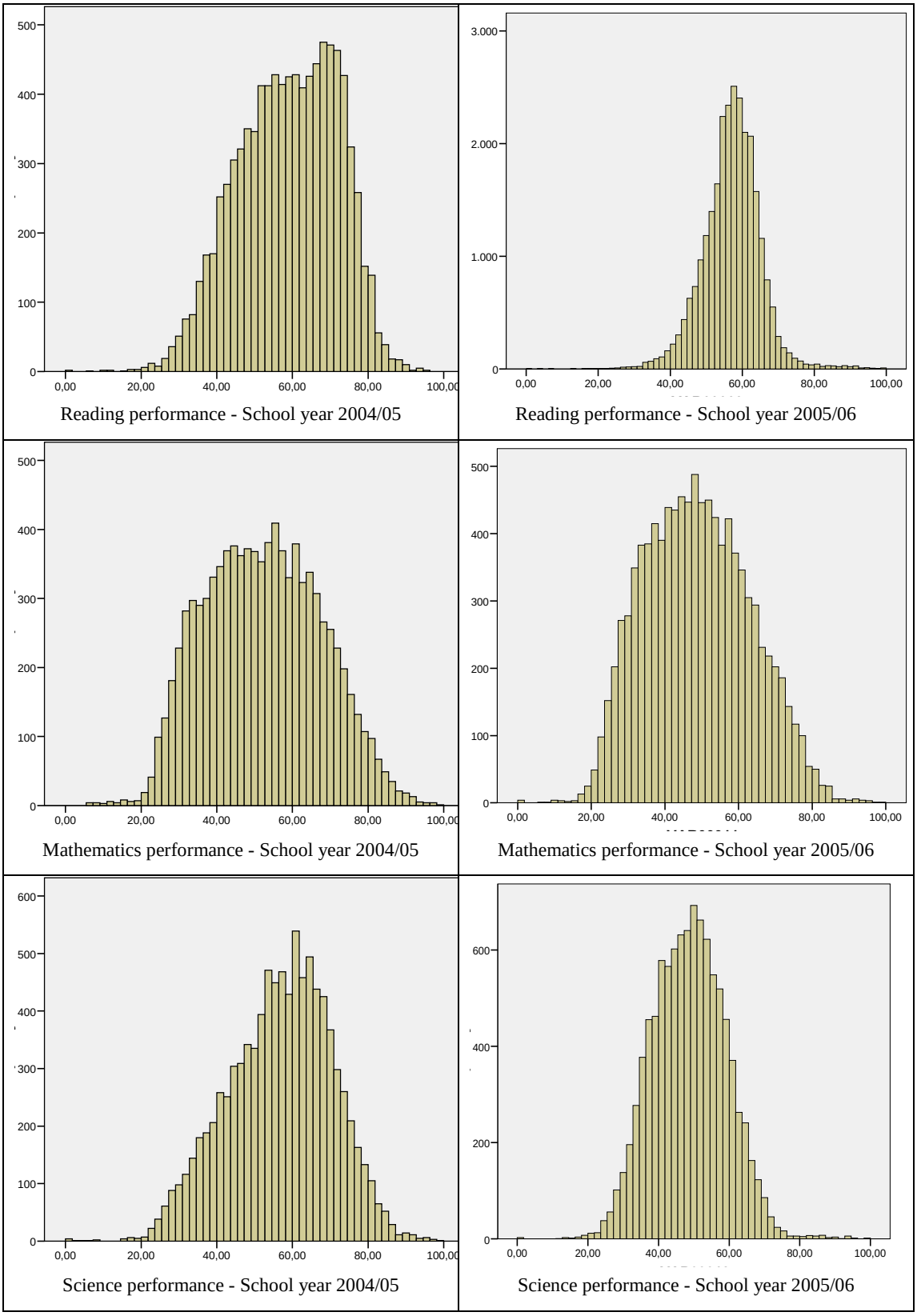
Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

Fig.3 - Distributions of mean scores, at class level, of student performance on the reading, mathematics and science scale. Students of first-class lower secondary school, school years 2004/05 and 2005/06.



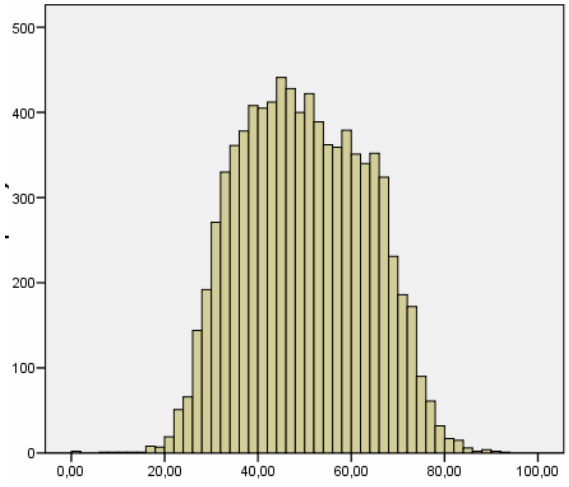
Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

Fig.4 - Distributions of mean scores, at class level, of student performance on the reading, mathematics and science scale. Students of first-class upper secondary school, school years 2004/05 and 2005/06.

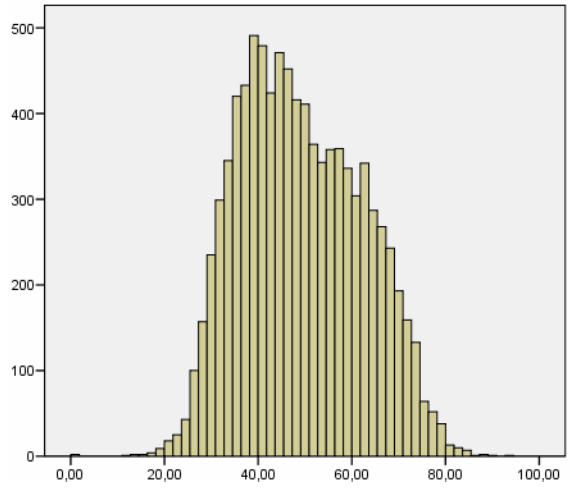


Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

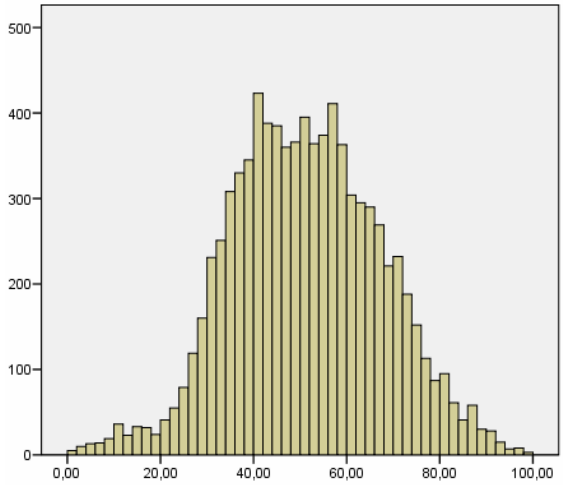
Fig.6 - Distributions of mean scores, at class level, of student performance on the reading, mathematics and science scale. Students of third-class upper secondary school, school years 2004/05 and 2005/06.



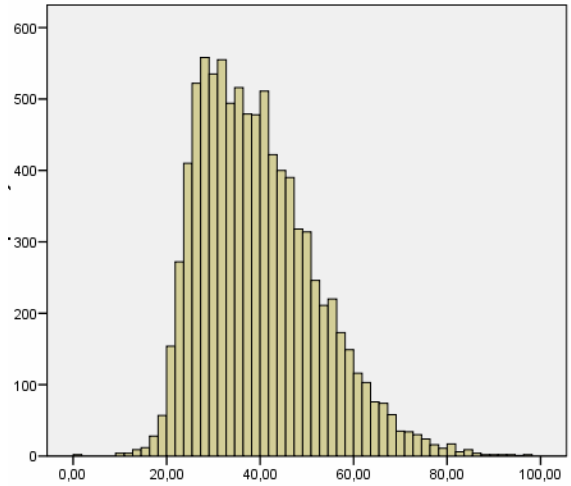
Reading performance - School year 2004/05



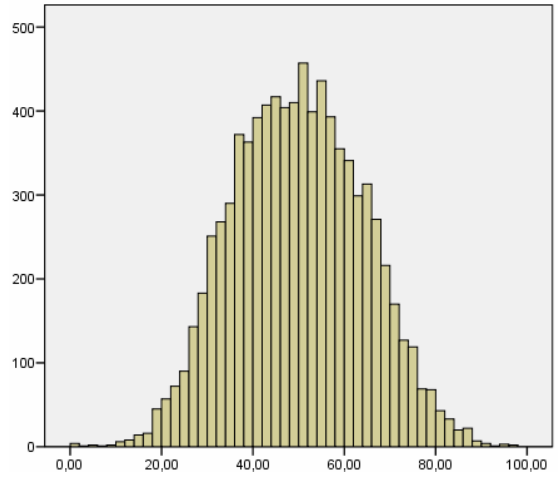
Reading performance - School year 2005/06



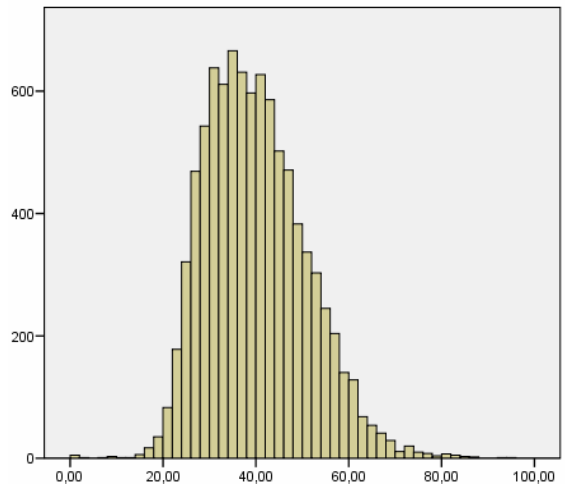
Mathematics performance – School year 2004/05



Mathematics performance - School year 2005/06



Science performance - School year 2004/05



Science performance - School year 2005/06

Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

The graphical comparison of the average score distributions by class has highlighted that the distributions of the secondary school, both at upper level and lower one, show a positive skewness and unimodality, while the primary school distributions show an upward bias and an anomalous presence of high frequencies in correspondence of the maximum values of distribution.
Looking only at the second year of primary classes, the considerable presence of outlier classes has produced an 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.

Tab.1 – Correlation between class mean score and its standard deviation.

	2004/2005			***2005/2006***		
	Reading	Mathematics	Science	Reading	Mathematics	Science
Second year primary school	-0,725	-0,731	-0,740	-0,729	-0,758	-0,688
Fourth year primary school	-0,643	-0,442	-0,688	-0,580	-0,329	-0,598
First year lower secondary school	-0,224	-0,154	-0,406	-0,266	0,153	-0,180
First year upper secondary school	-0,212	0,201	-0,024	-0,158	0,280	0,156
Third year upper secondary school	0,126	0,217	0,094	0,208	0,403	0,236

Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

The values of this coefficient stand out that the increase of the average class score is related with the reduction of within class variability but these findings are consistent only for the primary class student.
An excessive support from the teachers to the pupils, due to their young age, could be a plausible determinant of this anomaly, indeed, only the classes of very young students -especially 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.

3. How find out and correct 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 in presence of 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 stage, it has been computed an index 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 distributions 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 haven't given the minimum number of answers to compute a performance score are considered as "pseudo-non respondents" and consequently they have been dropped from dataset (listwise deletion).

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

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

$$\overline{p}_j = \frac{\sum_{i=1}^{N_j} p_{ij}}{N_j}$$

where:
p_{ij} denotes the score of ith student of jth class
N_j denotes number of respondent students of jth class

Standard deviation of mean score for jth class σ_j :

$$\sigma_j = \sqrt{\frac{\sum_{i=1}^{N_J} (p_{ij} - \overline{p}_j)^2}{N_j}}$$

where:
p_{ij} denotes the score of ith student of jth class
$\overline{p}_j$ denotes the class jth mean score
N_j denotes number of respondent students of jth class

The class non-response rate MC_j expresses the collaboration of each class to respond to all the test questions, it's equal to

$$MC_j = \frac{\sum_{i=1}^{N_j} M_{ij}}{N_j Q}$$

where:
M_{ij} denotes the number both of item non responses and of invalid responses for the ith student of the jth class

Q denotes the number of administered item to jth class. It's a constant for each assessment area (reading, mathematics and science) and for each school level
N_j denotes the number of respondent students of jth class

The values of this class non-response rate vary in the range 0-1. It's equal to 0 when isn't any missing or invalid response for the jth class, while it reaches its maximum (equal to 1) when all students of jth class have given only missing or invalid answers.

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

$$\overline{E}_j = \frac{\sum_{s=1}^{Q} E_{sj}}{Q}$$

It is the mean of the Q Gini indexes (E_{sj}) computed for each sth test question.
The numerator of $\overline{E}_j$ is a Gini measure of homogeneity E_{sj} and it is computed for each sth test question administered to each student of jth class:

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

Where:
$\frac{n_t}{N_j}$ denotes the ratio of students of jth class that has given the tth answer to sth question.
The Gini measure is equal to zero when all students of jth class have given the same answer to the sth question, while it reaches the maximum value: $\frac{h-1}{h}$ (h is the number of alternative answers to question sth) when there is perfect heterogeneity of answers to sth question in the jth class.

Thus, $\overline{E}_j$ is the mean for each j[th] class of the Q Gini indexes and it is between 0 and $\frac{h-1}{h}$, inclusive. It's equal to zero when all students of j[th] class have given the same answers to all test questions, while it reaches the value $\frac{h-1}{h}$ 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 $\overline{p}_j$
- Standard deviation of mean score σ_j
- Class non-response rate MC_j
- Index of answers' homogeneity $\overline{E}_j$

3.2 Dimensionality reduction by Principal component Analysis (PCA)

At the second step, the size of the data matrix, composed by the four indexes at class level, is reduced to two components by using Factor Analysis with a principal components extraction (Jolliffe, 2002).
The first two principal components account for 90% of the total variance (table 2).

Tab.2 - Eigenvalues of correlation matrix R, simple and cumulative percentage of explained variability by the principal component analysis applied to four indicators of class answer behaviour. Data collected from second class primary students participating to mathematics assessment in the school year 2004/2005.

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

Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

The component matrix, table 3, shows the correlations between the four observed variables and the first two principal components.

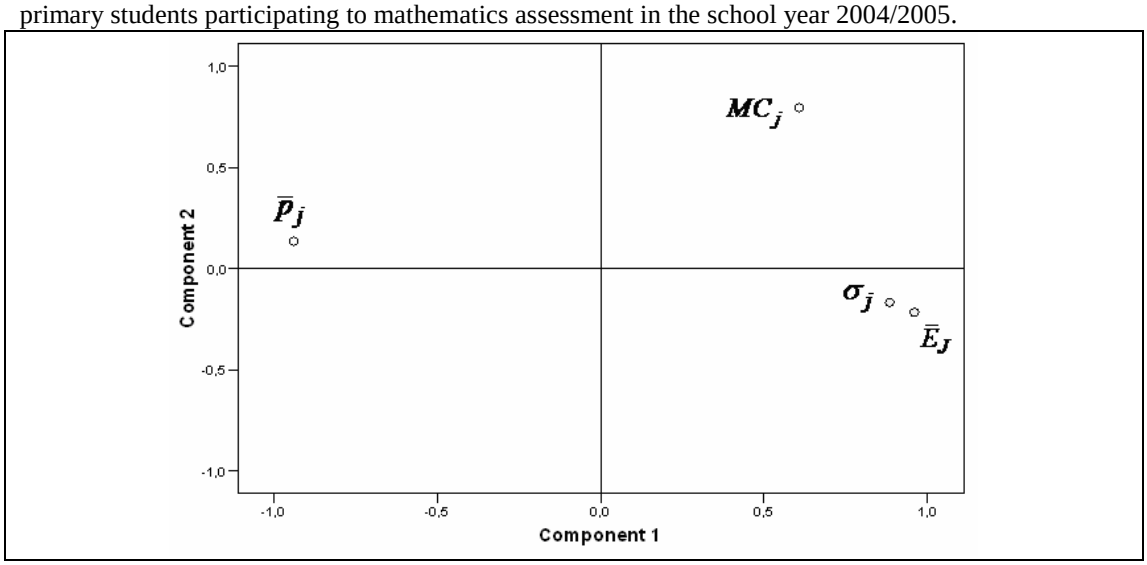
Tab.3 – Correlation between the indexes of class answer behaviour and the first two components correlations. Data collected from second class primary students participating to mathematics assessment in the school year 2004/2005.

Observed variable	Component	
	1	2
Class mean score $\overline{p}_j$	-0,946	0,117
Standard deviation of mean score σ_j	0,880	-0,134
Class non-response rate MC_j	0,670	0,742
Index of answers' homogeneity $\overline{E}_J$	0,940	-0,286

Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

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 both the standard deviation and the index of answers' homogeneity is very positive (0,880 and 0,940 respectively). These values suggest that the first component might be interpreted as "outliers identification axis".
The second component is most highly correlated with class non-response rate, then this factor might be defined as the "index of class collaboration to survey".
A plot of the variables (figure 6) where each factor loading is a coordinate along the corresponding factor's axis is useful in order to illustrate the axes interpretation graphically.

Fig.6 - Loading plot corresponding to the first two components. Data collected from second class primary students participating to mathematics assessment in the school year 2004/2005.

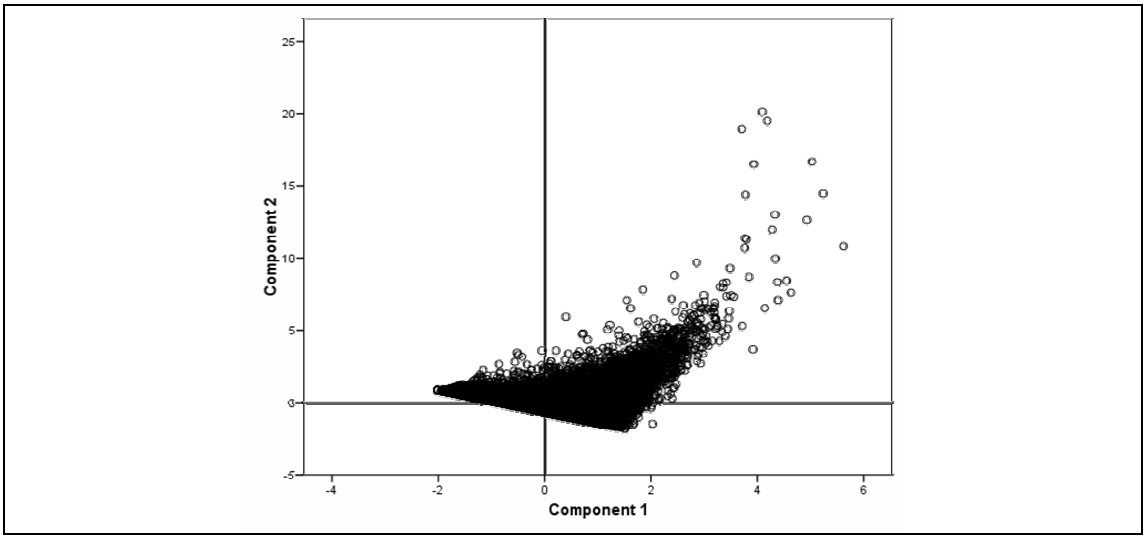


Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

As reported in figure 6, the class mean score has a high negative loading on the first factor, while the within variability indexes show a positive loadings 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 explained also by the negative correlation between the class average score and its standard deviation (table 1).
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 "index of class collaboration to survey". Since the class non response rate has a positive loading on the second factor (0,742), the students classes with a low number of missing data will be distinguished by high second factor scores.

Fig.7 – Projection on the first two factorial axis plane of second class primary students participating to INVALSI mathematics assessment in the school year 2004/2005.



Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

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

$$J_{FKM} = \sum_{n=1}^{N}\sum_{s=1}^{S} p_{ns}^{r} d_{ns}^{2}$$

Where $p_{ns} \in [0,1]$, $\sum_{s=1}^{s} p_{ns} = 1$ represents the membership degree of object n ($1 \le n \le N$) in group s ($1 \le s \le S$). The extension is made by introducing a weight *r* ($1 \le r \le \infty$), named '*fuzziness factor*', which characterizes the family.
If $r = 1$, the obtained solution would be a non-overlapping partition. If *r* tends to the infiniteness then the membership degree values to each class become close to $1/S$. The fuzzy partition degree grows with *r*, and 2 is the most used value.
The optimal strategy to minimize the J_{FKM} function, subdivided into classical stages: Initialization (I and II), Iteration (III and IV) and Stop Criterion (V):

I) Determining the cluster number s and fixing the parameter r.
II) Calculation of the group centroids using the expression:

$$v_{sk} = \frac{\sum_{n=1}^{N} p_{ns}^{r} x_{nk}}{\sum_{n=1}^{N} p_{ns}^{r}}$$

where x_{nk} represents the value of variable k ($1 \leq k \leq K$) for object n ($1 \leq n \leq N$).
III) Construction of a new fuzzy partition matrix (determination of the new membership values):
if an object *n* keeps a distance *0* from the centre of class *s*, the value of p_{ns} is
equal to 1 and the membership values of *n* towards the remaining classes is equal to 0;
if all the distances from an object to the centroids of the S groups are above 0, the membership values are determined by:

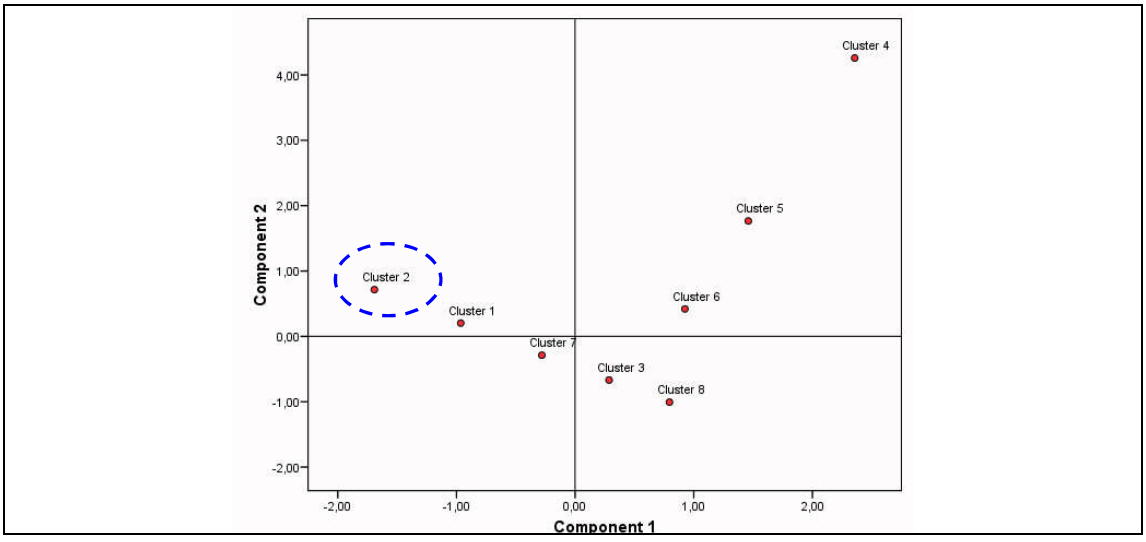
$$p_{ns} = \left[\sum_{t=1}^{s} (d_{ns} / d_{nt})^{2/r-1} \right]^{-1}$$

IV) Calculation of the group centroids associated to the partition determined in 3
V) Repetition of steps III and IV until the stop criterion is reached.
On the basis of the two factorial dimensions the students' classes are classified in s=8 clusters by a fuzzy k-means algorithm with the parameter r equal to 2.
In fuzzy clustering, each point has a degree of belonging to clusters rather than belonging completely to just one cluster.
This property is useful to assign at each students' class the degree of belonging to outlier cluster and then to correct the class average score in proportion of this membership.
The steps of the procedure are:

- clustering the classes by fuzzy k-means algorithm (s=8, r=2)
- the projection on the factorial axes of the cluster centroids.
- the detection of outlier cluster centroid
- for each class, the computation of a correction factor on the basis of degree of belonging to outlier cluster centroid

The output of the fuzzy k-means is a matrix where for each class is reported the membership degree to every cluster (s=8).
By the projection of the centroids on the factorial axes (figure 8), it's possible to detect the cluster centroid with an outlier profile.

Fig.8 – Projection on factorial plane of centroids computed by fuzzy k-means algorithm. Data collected from second class primary students participating to mathematics assessment in the school year 2004/2005.



Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

On the basis of principal components interpretation, the Cluster 2 gathers the classes that present an outlier profile: This cluster is distinguished by:

- High negative scores on "outliers identification axis" (x-axis) that indicates a high class average scores and minimum within variability respect to scores and test answers
- Factorial scores close to zero respect to the "index of class collaboration to survey" (y-axis) that indicate a low presence of missing items in the class data and a full compilation of performance test

Indicating the outlier cluster with the term "a", for the i^{th} class the degree of belonging to this cluster is equal to:

$$\mu_{ia}$$

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

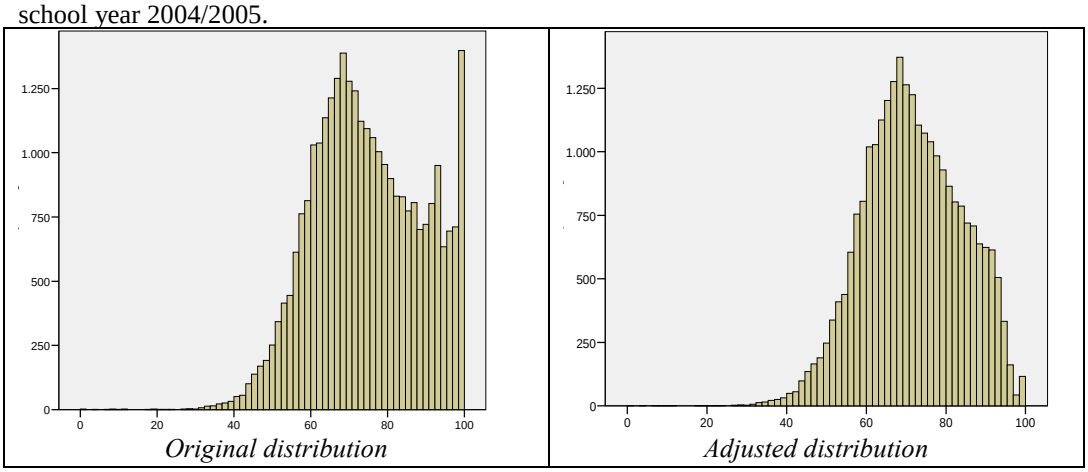
$$w_i = 1 - \mu_{ia}$$

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 μ_{ia}) 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 to compute an "outlier level" measure for each unit and, consequently, to calibrate the correction in optimal way.
The impact of correction procedure was analysed using a graphical comparison between the two distributions before and after the weight application (figure 9).

Fig.9 Comparison between the original class mean distribution and the adjusted one. Data collected from second class primary students participating to mathematics assessment in the school year 2004/2005.



Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

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's platycurtik (kurtosis equal to -0,16) and it shows a light negative skewness (skewness index equal to -0.114).
Again, it's distinguished from the prior distribution by the lack of two modes and by the reduction of the high frequencies peaks in correspondence of the higher values of the variable.
Focusing on the descriptive statistics of table 4, 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.

Tab. 4 – Comparison between average no weighted score per class and the weighted score per class according to the factor $w_i=1-\mu_{ia}$. Data collected from second year class primary students participating to mathematics assessment in the school year 2004/2005.

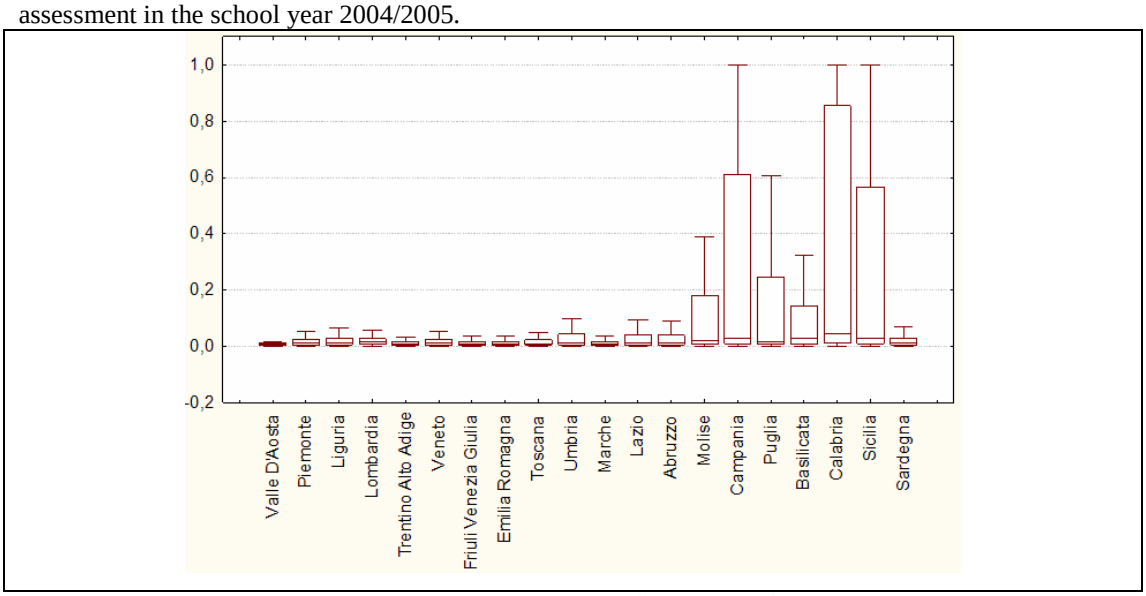
	Original distribution	**Adjusted distribution**
MEAN	74,71	71,67
MODE	100,00	68,75
I QUARTILE	64,42	63,12
MEDIAN	73,61	71,09
III QUARTILE	85,94	80,69

Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

5. The geographical localization of students' class as determinant of the outliers presence

An analysis of "outlier level" coefficient (μ_{ia}) distributions by geographical regions is performed in order to evaluate the relationship between the school localization and the presence of outlier classes. The analysis is carried out by the comparison of box plots of "outlier level" by regions (figure 10).

Fig.10 – Graphics by box plot of the index μ_{ia} distributions (i-th unit degree of belonging to outlier points group). Data collected from second year primary students participating to mathematics assessment in the school year 2004/2005.



Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

The empirical examination of the box plots allows to classify the regions in two groups:

- The first group includes Sardinia and the regions of Central and Northern Italy. The classes of these regions show low probability to be considered as outlier units. Indeed, the mean and the median of μ_{ia} for the region including in this group vary between 0 and 0,16.
- The second group encompasses all Southern Italy regions (Sardinia excluded) that are distinguished by higher values of μ_{ia}. Particularly, the third quartile of outlier coefficient is equal to 0,8 for Campania distribution and 0.6 for Calabria one. This means that 25% of students' classes in Campania and in Calabria might be considered outliers with a probability superior to 0,6.

The descriptive statistics in table 5 confirm the dependence between the presence of outlier classes and their geographical localization. Then the considerable difference between the southern regions distributions of μ_{ia} and the Northern and the Central ones leads 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.

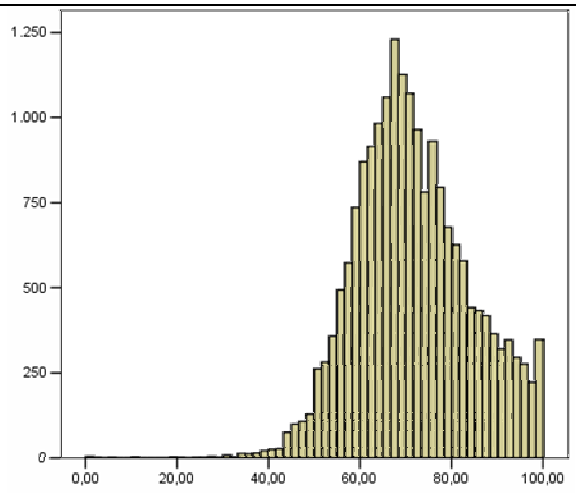
Tab.5 – Index μ_{ia} distributions (i-th unit degree of belonging to outlier points group for the geographical regions). Data collected from second year primary students participating to mathematics assessment in the school year 2004/2005.

REGIONS	MEAN	I QUARTILE	MEDIAN	III QUARTILE
Valle D'Aosta	0,010	0,005	0,010	0,010
Piemonte	0,060	0,010	0,010	0,020
Liguria	0,073	0,010	0,010	0,030
Lombardia	0,035	0,010	0,020	0,030
Trentino A.A.	0,032	0,010	0,010	0,020
Veneto	0,040	0,010	0,010	0,020
Friuli V.G.	0,041	0,010	0,010	0,020
Emilia R.	0,056	0,010	0,010	0,020
Toscana	0,067	0,010	0,010	0,020
Umbria	0,095	0,010	0,010	0,040
Marche	0,056	0,010	0,010	0,020
Lazio	0,133	0,010	0,010	0,040
Abruzzo	0,121	0,010	0,010	0,040
Molise	0,198	0,010	0,020	0,180
Campania	0,272	0,010	0,030	0,610
Puglia	0,214	0,010	0,020	0,250
Basilicata	0,195	0,010	0,030	0,140
Calabria	0,331	0,010	0,050	0,850
Sicilia	0,267	0,010	0,030	0,570
Sardegna	0,105	0,010	0,010	0,030
ITALIA	0,140	0,010	0,010	0,040

Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

To confirm this hypothesis, it's interesting to observe the class average score distributions only for the Northern and Central Regions students' classes (figure 11).

Fig.11 - Average no weighted 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.



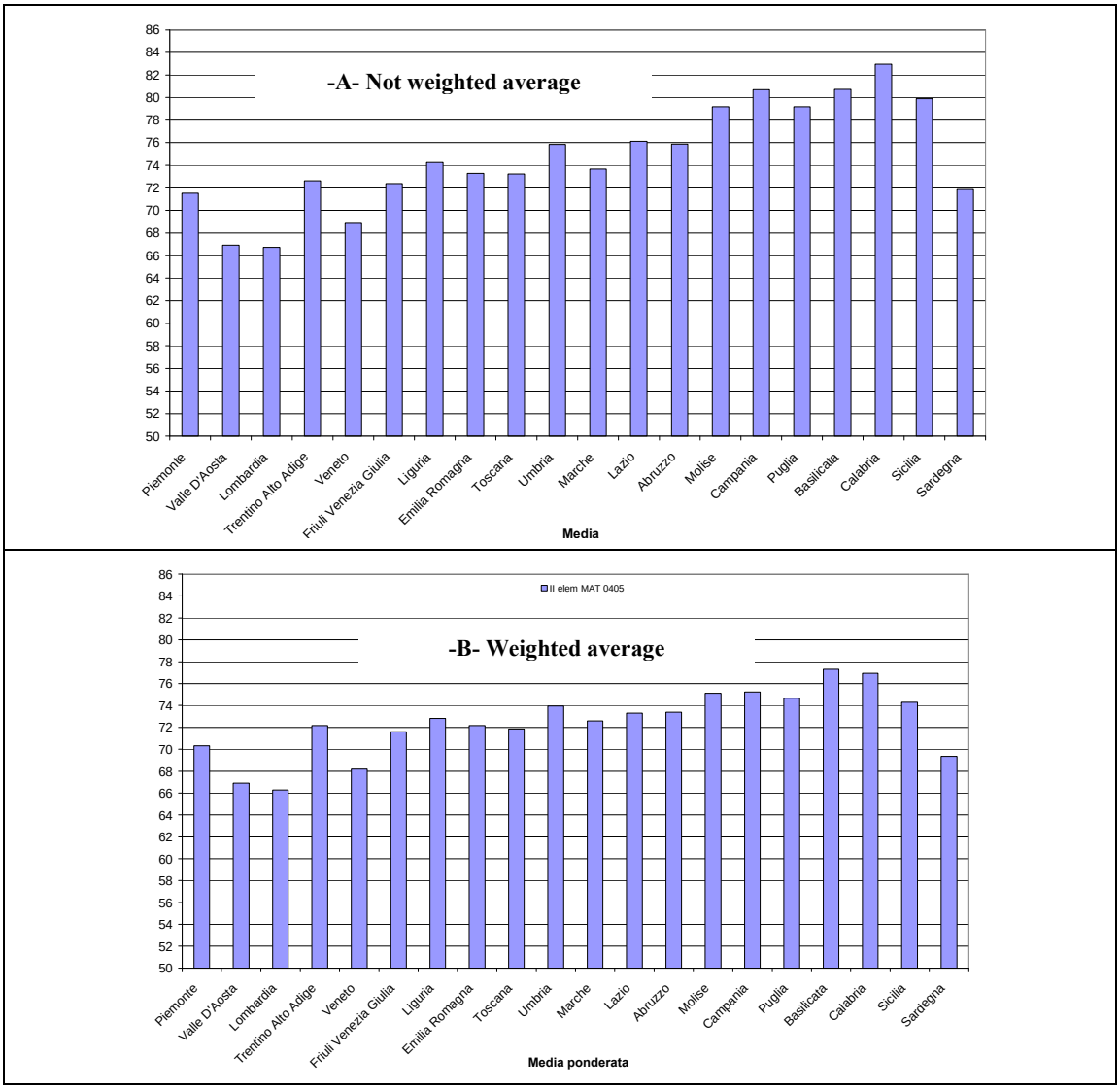
Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

This distribution doesn't show the anomalies of the national score distribution, in fact, it's unimodal and it doesn't 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's supposed that the primary teachers have provided an 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 -not weighted- ones (figure 12).

Fig.12 - Average not weighted -A- and weighted -B- score per class computed on the data collected from second class primary students participating to mathematics assessment in the school year 2004/2005.



Source: Authors'computations on Italian National Evaluation Institute of the Ministry of Education (INVALSI) data.

6. Conclusions

In this paper, an outliers' 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 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 (totally 15 dataset) 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 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 teachers' intervention.
In this context, the teacher support might be considered like a 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 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 stage, it has been computed an index 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 outliers' presence and the localization of students' classes. These evidences brought us to hypothesize the presence of some problems in the assessment system of INVALSI.
Finally, these findings suggest to revise the data collection procedures, especially the administration of the questionnaire, in order to avoid the outliers' presence and to improve the data quality.

References

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

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. e Sudman S. (1991), *Measurement Errors in Surveys*, Wiley, New York.

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

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

Hodge V., Austin J. (2004) A Survey of Outlier Detection Methodologies in *Artificial Intelligence Review*, Volume 22 , Issue 2, Kluwer Academic Publishers

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

Jolliffe I.T. (2002) *Principal Component Analysis*, Series: Springer Series in Statistics, 2nd ed., Springer, New York.