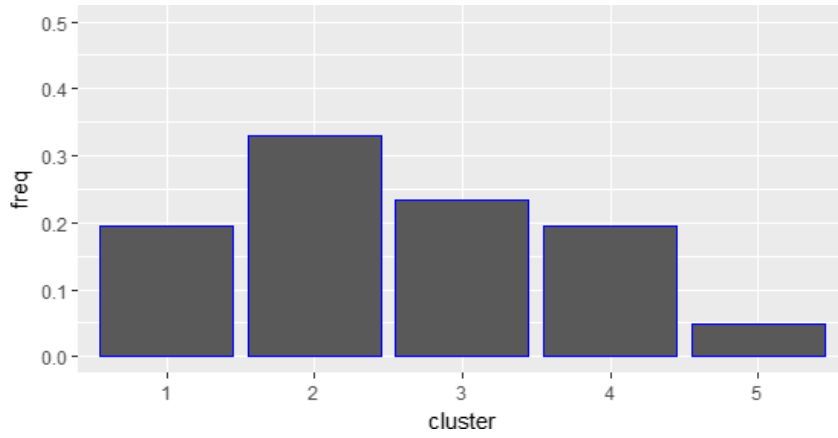
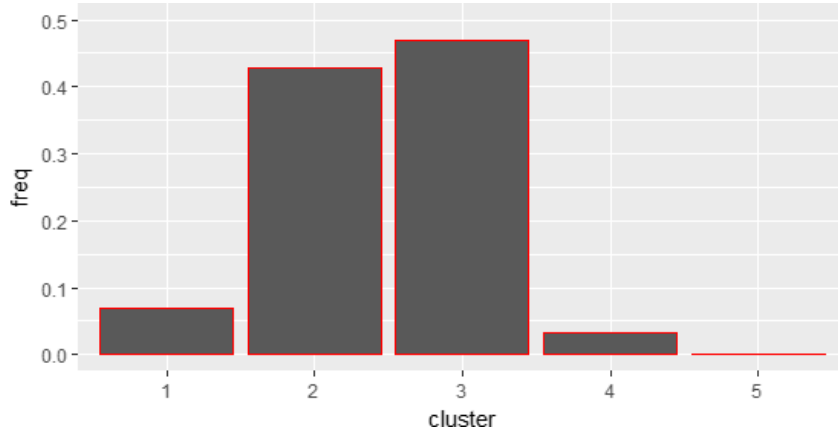
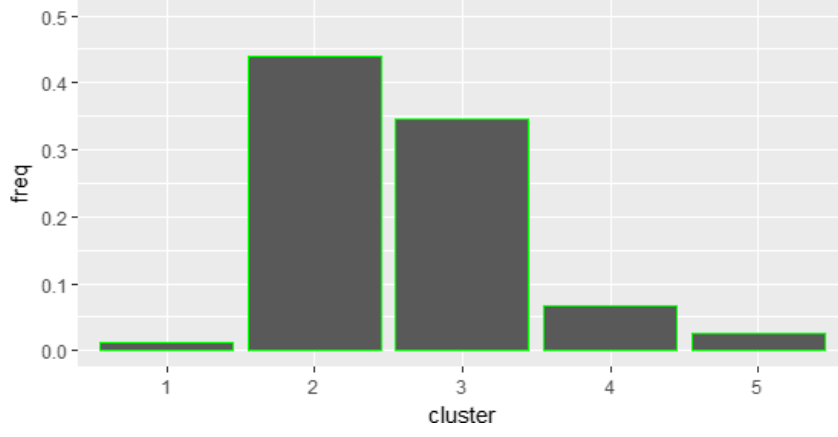


(a)



(b)



(c)

Fig. 7. Proportion of schools belonging to the five subpopulations, within the three geographical Italian macroareas (a) northern, (b) central and (c) southern Italy

algorithm to INVALSI data for 2013–2014 as a tool for clustering Italian schools. The semiparametric EM algorithm places itself in the branch of literature concerning the algorithms proposed in Aitkin (1996) and Azzimonti *et al*. (2013). In particular, our algorithm is inspired by that proposed in Azzimonti *et al*. (2013), but it introduces the major improvement, among others, that the covariates are group specific, meaning that they can vary both in number of observations and in range of assumed values across groups. Moreover, with respect to the algorithm proposed in Aitkin (1996) and the literature about GMMs and latent class analysis, the advantage of the semiparametric EM algorithm is that it does not need to fix *a priori* the number of discrete masses (subpopulations), but, conditionally on certain parameter values, the algorithm itself identifies the number of discrete support points. This has great value in applications where the number of subpopulations is not known *a priori* and the aim is therefore to find out how many and which different trends exist within the data. This concept is particularly relevant in the era of big data, where there is the need to identify latent structures within big and complex databases.

The semiparametric EM algorithm, when applied to the INVALSI data, can identify subpopulations of schools, within which student achievement trends differ. Among the identification of the number of subpopulations, which reveals how many different trends exist within the sample of Italian schools, the weights that are associated with the subpopulations give further information about the clustering. In a context in which we do not know *a priori* which is the expected trend, the subpopulations that are associated with higher weights represent the most common behaviour, whereas the less numerous subpopulations (those associated with lower weights) represent those schools whose impact differs from the majority. This draws attention to what determines whether schools belong to the minority subpopulations. In particular, the algorithm identifies five school subpopulations that represent different school associations with their student achievements trends, seen as the ability of junior secondary schools to train students to obtain certain skills at the end of the 3 years, given their skills at the beginning of schooling, adjusting for their socio-economic index ESCS. In the INVALSI framework, schools are associated with a *positive or negative effect*, based on the final performances of their students and given their students' initial skills. Among these five subpopulations, a subpopulation containing schools with a negative effects is immediately evident. This subpopulation contains schools that have students who tend to underperform, with respect to their performance 2 years before, since they have on average very low scores, even if 2 years before, when they started to attend these schools, they obtained higher scores. Regarding positive effects, we interpret the subpopulation with the highest intercept and positive slope (subpopulation 1) as the best, in terms of school effect, since it contains schools that can train students to reach high performances, even if they had low performances at the beginning of schooling. It is worth saying that, from a policy perspective, the definition of the *best school effect* is currently in debate. Indeed, it is reasonable to consider a school in which all students obtain very high scores, without heterogeneity, as a school with a good effect, but, in contrast, a different point of view emphasizes the advantages of having heterogeneity within the school. In this perspective, the role of the school is continuously to raise the students goals to urge pupils to perform even better, using competition and variation to motivate them.

After the identification of school subpopulations, the paper focuses on another actual and interesting topic, i.e. their interpretation *a posteriori*. In particular, we explore the associations between school subpopulations and school level characteristics, showing that only geographical areas, the percentage of immigrants, a dummy for private or public school and school principal's education turn out to be significantly associated. This evidence suggests that the school level variables at our disposal do not explain the differences in school effects. On the basis of the

fact that the school subpopulations are clearly different in their effect on student attainments, the lack of stratification of school level variables across subpopulations might mean that the observed school level variables do not reflect the real school characteristics (i.e. they are not measured in the right way) or there are other latent aspects, that we cannot measure, which might explain the different effects of schools on their students.

In the future, our aim is to deepen the analysis on the characterization of the estimated school subpopulations, considering other information about the school environment, which we have not been able to measure until now. Moreover, from a methodological point of view, our aim is to develop a multivariate version of the EM algorithm for semiparametric mixed effects models, to consider two (or more) response variables and to relax the linearity assumptions, considering also the case of other functional forms. In the INVALSI framework, since the data set contains both student scores in reading and in mathematics, it would be possible to apply the multivariate version, in which the response variable would be the bivariate vector of reading and mathematics scores, and, consequently, to cluster schools or classes on the basis of both their effects on reading and mathematics student attainments, analysing the interactions between these two fields.

References

Agasisti, T., Ieva, F. and Paganoni, A. M. (2017) Heterogeneity, school-effects and the north/south achievement gap in Italian secondary education: evidence from a three-level mixed model. *Statist. Meth. Appl.*, **26**, 157–180.

Agasisti, T. and Vittadini, G. (2012) Regional economic disparities as determinants of student's achievement in Italy. *Res. Appl. Econ.*, **4**, no. 2.

Aitkin, M. (1996) A general maximum likelihood analysis of overdispersion in generalized linear models. *Statist. Comput.*, **6**, 251–262.

Azzimonti, L., Ieva, F. and Paganoni, A. M. (2013) Nonlinear nonparametric mixed-effects models for unsupervised classification. *Computnl Statist.*, **28**, 1549–1570.

Bock, R. D. (2014) *Multilevel Analysis of Educational Data*. Amsterdam: Elsevier.

Bock, R. D. and Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika*, **46**, 443–459.

Bryk, A. S. and Raudenbush, S. W. (1988) Toward a more appropriate conceptualization of research on school effects: a three-level hierarchical linear model. *Am. J. Educ.*, **97**, 65–108.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F. and York, R. (1966) The Coleman report. *Equality of Educational Opportunity*. Washington DC: US Government Printing Office.

Goldstein, H. and Spiegelhalter, D. J. (1996) League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion). *J. R. Statist. Soc.* A, **159**, 385–443.

Hanushek, E. A., Rivkin, S. G. and Taylor, L. L. (1996) Aggregation and the estimated effects of school resources. *Technical Report*. National Bureau of Economic Research, Cambridge.

Heinen, T. (1996) *Latent Class and Discrete Latent Trait Models: Similarities and Differences*. New York: Sage.

Lin, L. I. (2000) Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statist. Med.*, **19**, 255–270.

Lindsay, B. G. (1983a) The geometry of mixture likelihoods: a general theory. *Ann. Statist.*, **11**, 86–94.

Lindsay, B. G. (1983b) The geometry of mixture likelihoods, part ii: the exponential family. *Ann. Statist.*, **11**, 783–792.

Masci, C., De Witte, K. and Agasisti, T. (2018) The influence of school size, principal characteristics and school management practices on educational performance: an efficiency analysis of Italian students attending middle schools. *Socio-Econ. Planng Sci.*, **61**, 52–69.

Masci, C., Ieva, F., Agasisti, T. and Paganoni, A. M. (2016) Does class matter more than school?: Evidence from a multilevel statistical analysis on Italian junior secondary school students. *Socio-Econ. Planng Sci.*, **54**, 47–57.

Masci, C., Ieva, F., Agasisti, T. and Paganoni, A. M. (2017) Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements. *J. Appl. Statist.*, **44**, 1296–1317.

Masci, C., Johnes, G. and Agasisti, T. (2019) Student and school performance across countries: a machine learning approach. *Eur. J. Oper. Res.*, to be published.

McCulloch, C., Lin, H., Slate, E. and Turnbull, B. (2002) Discovering subpopulation structure with latent class mixed models. *Statist. Med.*, **21**, 417–429.

Muthén, B. (2004) Latent variable analysis. In *The Sage Handbook of Quantitative Methodology for the Social Sciences*, pp. 345–368. Thousand Oaks: Sage.

Muthén, B. and Shedden, K. (1999) Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, **55**, 463–469.
Nagin, D. S. (1999) Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychol. Meth.*, **4**, no. 2, 139–157.
Pinheiro, J. C. and Bates, D. M. (2000) Linear mixed-effects models: basic concepts and examples. In *Mixed-effects Models in S and S-Plus*, pp. 3–56. New York: Springer.
Proust-Lima, C., Letenneur, L. and Jacqmin-Gadda, H. (2007) A nonlinear latent class model for joint analysis of multivariate longitudinal data and a binary outcome. *Statist. Med.*, **26**, 2229–2245.
Raudenbush, S. and Bryk, A. S. (1986) A hierarchical model for studying school effects. *Sociol. Educ.*, **59**, 1–17.
Raudenbush, S. W. and Willms, J. (1995) The estimation of school effects. *J. Educ. Behav. Statist.*, **20**, 307–335.
R Development Core Team. (2014) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Sani, C. and Grilli, L. (2011) Differential variability of test scores among schools: a multilevel analysis of the fifth-grade INVALSI test using heteroscedastic random effects. *J. Appl. Quant. Meth.,* **6**, 88–99.
Sarrico, C. S., Rosa, M. J. and Manatos, M. J. (2012) School performance management practices and school achievement. *Int. J. Product. Perform. Mangmnt*, **61**, 272–289.
Sirin, S. R. (2005) Socioeconomic status and academic achievement: a meta-analytic review of research. *Rev. Educ. Res.*, **75**, 417–453.
Vanthienen, J. and De Witte, K. (2017) *Data Analytics Applications in Education*. New York: Taylor and Francis.
Vermunt, J. K. and Magidson, J. (2002) Latent class cluster analysis. *Appl. Latnt Class Anal.*, **11**, 89–106.