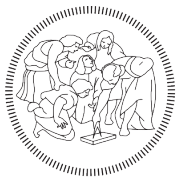


POLITECNICO
MILANO 1863

Department of Mathematics
Doctoral Program In Mathematical Models
and Methods in Engineering

Semi-parametric mixed-effects models for assessing public education systems

Doctoral Dissertation of:
Chiara Masci

Supervisor:
Prof. Anna Maria Paganoni

Co-supervisor:
Prof. Francesca Ieva

The Chair of the Doctoral Program:
Prof. Irene Maria Sabadini

Year 2019 – XXXI Cycle

Abstract

This thesis regards the development of semi-parametric mixed-effects models and their application to administrative educational databases for the analysis of student, class, school and university performances. The research aims go in two directions: the former is to develop novel statistical models and methods that represent a novelty and an improvement both in the statistical and in the educational literature; the latter is to investigate statistical methods that have the potential of being applied to educational data for addressing new and interesting research questions in the context of learning analytics. Being the hierarchical structure (e.g. students are nested within classes, that are in turn nested within schools...) the main characteristic of educational data, mixed-effects models, that are able to take into account the nested nature of the data, constitute the cross-sectional methodological core of the entire work. Our proposed approach consists in relaxing the parametric assumptions, both on fixed and random effects, of mixed-effects models in order to develop innovative and advanced statistical methods with the aim of improving the research about school or university effectiveness and of addressing new and unexplored issues in the educational research context.

Sommario

La tesi riguarda lo sviluppo di modelli a effetti misti semi-parametrici e la loro applicazione a dataset amministrativi educativi per l'analisi delle performance di studenti, classi, scuole e università. Gli obiettivi della ricerca sono due: il primo è di sviluppare modelli e metodi statistici innovativi che rappresentino un valore aggiunto sia nella letteratura statistica che in quella del campo dell'educazione; il secondo è di identificare metodi statistici che abbiano il potenziale, se applicati a dati educativi, di rispondere a nuove e interessanti domande di ricerca nel contesto del learning analytics. Dato che la principale caratteristica dei dati educativi è la loro struttura gerarchica (per esempio, gli studenti sono annidati nelle classi, che a loro volta sono annidate nelle scuole), il cuore metodologico di tutto il lavoro si basa sui modelli a effetti misti, che sono in grado di modellizzare la natura annidata dei dati. In particolare, ci proponiamo di rilassare le assunzioni parametriche di questi modelli, sia sugli effetti fissi, che su quelli casuali. I metodi non-parametrici così sviluppati, oltre ad essere metodologicamente innovativi, rappresentano un importante contributo alla ricerca sull'efficacia educativa di scuole e università e permettono di affrontare nuove tematiche nel contesto dell'apprendimento scolastico e universitario.

Acknowledgments

We are grateful to Tommaso Agasisti for his support, his collaboration and his inspiring contribution to the thesis, to Geraint Johnes for having hosted me as a visiting PhD student at Lancaster University and for his comments and contribution to the thesis and to Fritz Schiltz for his contribution. We thanks the H2020 EdEN project for its support, Patrizia Falzetti and INVALSI for having provided the original data and Umberto Spagnolini and Aldo Torrebruno for their comments and support during the work within the SPEET project.

Contents

Introduction **1**

1 Student and school performance across countries: a machine learning approach via random-effects regression trees and boosting **9**
1.1 Background and previous literature 11
1.2 The OECD-PISA dataset 15
1.3 Model and methods 19
1.3.1 An introduction to tree-based methods 19
1.3.2 Multilevel models and RE-EM trees 22
1.3.3 Regression trees and Boosting 23
1.4 Results 25
1.4.1 First stage: Estimating the determinants of students' test scores and school value-added by using RE-EM trees . . . 28
1.4.2 Second stage: Modelling the determinants of school value-added through regression trees and boosting 31
1.5 Discussion, concluding remarks and policy implications 37

2 Performing learning analytics via generalized mixed-effects trees **41**
2.1 Model and methods 43
2.1.1 Generalized mixed-effects tree model 43
2.1.2 Generalized mixed-effects tree estimation 45
2.2 Simulation study 47
2.3 Case study: application of mixed-effects tree algorithm to education PoliMi data 52
2.4 Conclusions 56

3 Semi-parametric mixed-effects models for the clustering of Italian schools **59**
3.1 Model, methods and simulation study 63
3.1.1 Semi-parametric mixed-effects model 63
3.1.2 The SPEM algorithm 65

3.1.3 Simulation study 70
3.2 Case study: application of SPEM algorithm to education INVALSI data 78
3.2.1 The INVALSI 2013/2014 dataset 78
3.2.2 SPEM algorithm applied to INVALSI data 80
3.2.3 Association between school characteristics and school subpopulations 85
3.3 Conclusions 88

4 Multivariate semi-parametric mixed-effects models for the joint clustering of Italian classes 93
4.1 Model, methods and simulation study 96
4.1.1 Bivariate semi-parametric mixed-effects model 97
4.1.2 The MSPEM algorithm 100
4.1.3 Simulation study 103
4.2 Case study: application of the MSPEM algorithm to INVALSI data 109
4.2.1 The INVALSI 2016/2017 dataset 109
4.2.2 MSPEM algorithm applied to INVALSI data 113
4.2.3 Characterization of the subpopulations of classes 120
4.3 Conclusions 122

Conclusions 125

A OECD-PISA results 129

Introduction

Student learning is a long and complex process that sees many different factors acting on it. During their careers, students receive inputs from their peers as well as from the school and class they are attending. Also personal motivation, family, friends and the geographical context play a fundamental role in student learning. All these aspects contribute to the education of each pupil, making the educational activity a complex process whereby inputs are converted into outputs. Moreover, the way in which various inputs affect the output is likely to vary substantially across the education systems that operate in different countries or contexts.

Learning Analytics has been defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (*https://tekri.athabascau.ca/analytics/*) and, in the last decades, it is receiving increasing attention. Dedicated programs test students in their scholastic skills and collect information about classes and schools they are attending, both at national and international levels, in many countries. These data are then stored in administrative databases that constitute rich ensembles of information. Many researchers analyze educational administrative data, as well as other type of educational data, both to identify the determinants of student learning and educational providers effectiveness, and to develop policy implications aimed at improving education systems across the world.

Within this context, the aim of the thesis is twofold: the former is to develop novel statistical models and methods that represent an improvement both in the technical and in the educational literature; the latter is to investigate statistical methods that have the potential of being applied to educational data for addressing new and interesting research questions in the context of learning analytics.

The education system can be characterized as a hierarchical system in which different levels of grouping are nested within each others. In primary or secondary education, students are nested within classes, that are in turn nested within schools, that are in turn nested within districts and so on so forth. In the same way, in higher education, students are nested within degree programs, that

are in turn nested within universities and so on so forth. Each one of these levels has a role in the learning process of students. Measuring how much of student education is due to each grouping level is not trivial, but is of extreme interest. Disentangling the effects given to the levels of grouping on the student learning process is important in the perspective of identifying the most influential level on which it is possible to act with the aim of improving education systems and, consequently, student education level. In this perspective, the hierarchical structure of educational data represents the main characteristic that drives our choice about the statistical models and methods to be developed and used. This is the reason why our modeling approach, cross-sectional to the entire work, is based on mixed-effects (or multilevel) models (Pinheiro and Bates, 2000), that, to the best of our knowledge, are one of the most appropriate tools to fit nested data (Bock, 2014; Agasisti et al., 2017a).

From a modeling point of view, the application of hierarchical models to educational data is straightforward. In Raudenbush (1988), the author explains the advantages of applying these models in an educational context. He states that two primary goals motivate the application of hierarchical models in education: first, the inclusion of data from many groups strengthens estimation of random effects for each group (a researcher, for instance, seeks to estimate a regression equation for a particular school), and second it improves inference about the fixed effects (a researcher, for instance, asks why some kinds of schools have smaller regression slopes than others). The application of hierarchical linear modeling enables researchers to go beyond the classical questions, such as why do students in some schools have higher achievements than others, to ask about why structural relationships vary across groups. These models also offer advantages in dealing with aggregation bias associated with nested data structure.

Being multilevel models able to quantify the part of variability in the response variable that is due to each level of grouping, when applied to educational data, they are useful to measure the "school effect", intended as the impact that the school the student is attending has on his/her achievements with respect to other schools (Bryk and Raudenbush, 1988; Coleman et al., 1966; Hanushek et al., 1996; Raudenbush and Bryk, 1986). In Bryk and Raudenbush (1988), the authors state the importance of considering the "unit-of-analysis" (students, classes, schools), when speaking about educational research, and they argue that hierarchical models should constitute the basic paradigm for quantitative research on student learning. Moreover, in Raudenbush and Bryk (1986), the authors, given the hierarchical structure of educational data, underline the importance of measuring school effects, as intended before, and present different approaches to analyze nested data. In Coleman et al. (1966), the authors view education as a process in which student performance (output) is produced from inputs including school resources, teacher quality, family attributes, and peer quality. In their perspective, policy attention should be focused on inputs that

are both directly controlled by policymakers (characteristics of schools, teachers, curricula, etc.) and those that are "uncontrolled" (family, friends, learning capacities of the student, etc.). Also in Hanushek et al. (1996), the authors show that schools' characteristics are important for determining student outcomes.

For these reasons, multilevel approaches have been broadly applied in the educational literature. Raudenbush himself applies hierarchical models in various educational studies (Bryk and Raudenbush, 1988; Willms and Raudenbush, 1989; Raudenbush and Bryk, 1986). Other examples are given by Goldstein (1987); Rumberger (1995); Grilli and Rampichini (2007); Plewis (2011); Sani and Grilli (2011); Shen et al. (2012); Sun et al. (2012); Martínez (2012); Masci et al. (2016b, 2017a); Agasisti et al. (2017b), that apply multilevel linear or logit models considering different levels of grouping, such as class, school, Local Education Authority (LEA), degree programs or geographical areas.

Even where these approaches do indeed model the hierarchical structure of data, however, their parametric modellistic assumptions are in general too restrictive due to the complexity, the interactions and the heterogeneity that characterize educational data. In this perspective, our innovative approach consists in relaxing some of the parametric assumptions of mixed-effects models, both on fixed and random effects respectively, in order to develop innovative and flexible statistical methods able to address new and interesting research questions and to extract major information from complex data. The literature about the development of non-parametric mixed-effects models for educational applications is very limited and we will enter in its merit in the following chapters of the thesis. Figure 1 summarizes the scenario about mixed-effects models in education, distinguishing our novel contribution from the state of the art, both from a methodological and an educational point of view.

We distinguish models with parametric fixed or random effects from models with non-parametric effects, as well as univariate from multivariate models (i.e. models in which the answer variable is univariate or multivariate). This last modeling distinction is worthwhile in educational applications. In a multivariate setting, student skills in different fields (e.g. reading, mathematics, science) can be modeled as multiple responses of the same model and, consequently, their structural correlation can be properly investigated. The interactions among different school subjects, indeed, are often stronger than what we expect, especially at class level, where the dynamics of learning processes in different school subjects are particularly intertwined. As shown in Figure 1 and as anticipated in previous paragraphs, univariate and multivariate parametric mixed-effects models have already been applied in the literature. Also our first approach to the topic was by means of parametric mixed-effects models (Masci et al., 2016a,b, 2017a). In Masci et al. (2016b) and Agasisti et al. (2017b), we apply univariate three-levels linear models to Italian administrative data, considering students nested within classes, in turn nested within schools. We estimate class and

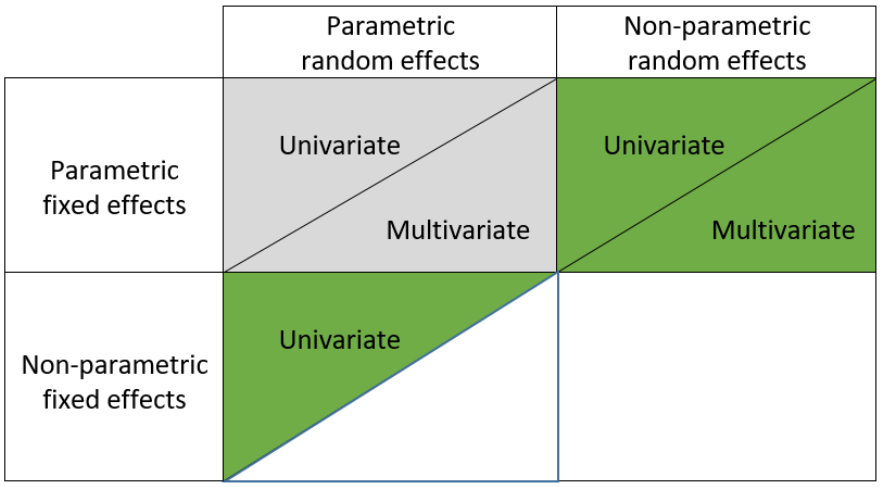


Figure 1: Scenario of mixed-effects models in education. Grey box represents existing statistical models already applied in the educational field while green boxes represent our novel contribution to the literature, both in the methodology and in the application.

school value-added, modeled as random intercepts of the multilevel model and intended as the contribution that the class and the school give to their student achievements, for mathematics and reading student achievements respectively. In Masci et al. (2017a), we apply the same model but extended to the case of a bivariate response variable, for jointly modeling the class and school value-added on mathematics and reading student achievements.

The new contributions of the thesis regard the development of mixed-effects models with non-parametric fixed effects (that fulfills the bottom left box in the table of Figure 1) and of mixed-effects models with non-parametric random effects (that fulfills the top right box in the table of Figure 1) to be applied to educational data to address new issues in the educational research. About mixed-effects models with non-parametric fixed effects, we propose two studies. In the former, our methodological approach is based on mixed-effects regression trees (Sela and Simonoff, 2012) and boosted regression trees (Friedman, 2001). This methodology is not new in the statistical literature, but what we will prove to be innovative is its effective and revealing application to the educational field. In the latter, we propose a new method that extends the classification tree model (Friedman et al., 2001) to handle clustered data structures. Regarding the case of mixed-effects models with non-parametric random effects, our contribution is again both methodological and novel in the educational research field, since we propose a new statistical method, a mixed-effects linear model, that can also handle a multivariate response, where the random effects follow a discrete distribution with an unknown number of support points, that, applied to educational

data, represents an improvement in the research about educational providers effectiveness.

In particular, all these concepts, that correspond to the contribution boxes in Figure 1, are developed as the cores of four different research lines, that are described in four chapters of the thesis in the following way:

- **Chapter 1**: The first research line regards the application of random-effects regression trees and boosted regression trees to international educational data. In a methodological perspective, the parametric assumption on the fixed effects of linear mixed-effects models are relaxed, being the linear functional form of the fixed effects replaced by a regression tree. The flexibility of this model results to be of great advantage in easily modelling both non-linearities and interactions among the variables. When applied to worldwide educational data, this methodology allows the identification of complex patterns across the variables and gives an improved description of the structurally different educational production functions across countries, leading to new and interesting insights in a policy implication perspective. Applying mixed-effects regression trees, we identify student level characteristics associated to student performances and we estimate school value-added, that, in a second step, can be characterized in terms of school level variables by means of boosted regression trees. The results explained in this chapter are gathered in Masci et al. (2018b). It is worth noting that the work that we present in this chapter is not our first attempt of using regression trees in educational application. We also applied tree-based methods to analyze Hungarian educational data in a work (Schiltz et al., 2018) that is part of the H2020 Education Economics Network (EdEN) project (www.edenproject.eu). We do not include this work in the thesis since the methodology is for a large part overlapped to the one that we propose in this chapter.

- **Chapter 2**: The second research line regards the proposal of an innovative statistical method, that is a generalization of mixed-effects trees for a response variable in the exponential family, Generalized Mixed-Effects Trees (GMET), and its application to higher education data for modeling student dropout. We perform a simulation study in order to validate the performance of our proposed method and to compare GMET to classical models. Given that the analysis of university careers and of student dropout prediction is one of the most studied topics in the area of learning analytics, in the case study, we apply GMET to model bachelor student dropout in different degree programs of Politecnico di Milano. The model is able to identify discriminating student characteristics and estimate the degree program effect on the probability of student dropout. The results explained in this chapter are gathered in Fontana et al. (2018).

- **Chapter 3**: The third research line regards the development of a novel semi-parametric mixed-effects linear model, together with an EM algorithm to estimate its parameters, and its application to Italian educational data as a tool to perform an unsupervised classification of Italian schools. We relax the parametric assumption on the distribution of the random effects of mixed-effects models and we assume them to follow a discrete distribution with an a priori unknown number of support points. This modelling induces an automatic clustering of the higher level of hierarchy (enabling the identification of subpopulations) and can be used in multiple classification problems. Among being an innovative method in the statistical scenario and representing a significant improvement in the context of mixed-effects models, this model contributes to the research about school effectiveness since, when applied to educational data considering students nested within schools, it identifies subpopulations of schools that differ in terms of distribution of student outcomes and that can be characterized a posteriori by school level variables. The results explained in this chapter are gathered in Masci et al. (2017b).

- **Chapter 4**: The last research line evolves as an extension of the model presented in Chapter 3, since it regards the development of a *multivariate* semi-parametric mixed-effects linear model, i.e. a model that handles a multivariate response variable. In this proposed model, the random effects are assumed to follow a multivariate discrete distribution where the numbers of support points are unknown and allowed to be different between multiple responses. This modelling enables to jointly model the presence of subpopulations in the higher level of hierarchy and it is totally new to the literature both from a technical/statistical point of view and for the potential that modelling the joint behaviors of the identified subpopulations has from an interpretative point of view. When applied to Italian educational data considering students nested within classes, this model allows to identify subpopulations of classes that differ in their joint effect on reading and mathematics student achievements. The results explained in this chapter are gathered in Masci et al. (2018a).

We are interested in analyzing three different databases: two of them regard lower education institutions (from primary to upper secondary schools), while one regards higher education institutions (universities). The two lower education databases of our interest are the Programme for International Student Assessment (PISA) database and the Italian Institute for the Educational Evaluation of Instruction and Training (INVALSI) database.

PISA (*www.oecd.org/pisa*) was initiated by the OECD and it is a triennial international survey (started in 2000) which aims to evaluate worldwide education systems by testing the skills and knowledge of 15-year-old students. At

each survey, over half a million students, representing 28 million 15-year-olds in 72 countries and economies, take the internationally agreed two-hour test. Students are assessed in science, mathematics, reading, collaborative problem solving and financial literacy. Moreover, a wide array of data concerning a set of student and school levels characteristics are available, thanks to questionnaires completed by students and school principals.

INVALSI (*www.invalsi.it*), following a procedure similar to OECD-PISA, tests Italian students all over the country since 2004, both in their mathematics and reading skills. These tests are administered at several grades, starting from primary schools up to the end of secondary schools, producing data that collect multiple observations for each student. Students are tested at grades II and V of primary school, at grade III of junior secondary school and at grade II of upper secondary school[1]. Moreover, INVALSI, by means of questionnaires, collects information about students themselves, teachers, classes, schools and school principals. By doing this, it creates a dataset that contains a rich picture of the personal and scholastic situation of each student.

Both these datasets allow to compare the performances of students coming from heterogeneous education systems and that attend different classes, in different schools, in diverse geographical areas, but with the same yardstick.

The higher education database is taken from the Student Profile for Enhancing Engineering Tutoring (SPEET) project, an Erasmus$^+$ project started in 2017, whose objective can be stated as determine and categorize the different profiles for engineering students across Europe. Partners of the project are Politecnico di Milano (Italy), Universidad Autonoma de Barcelona (Spain), Universidad de Leon (Spain), Instituto Politécnico de Bragança (Portugal), Opole University of Technology (Poland) and Dunarea de Jos University of Galati (Romania). The aim of the project is to collect engineering students performances and collateral data from partner organizations and to apply data mining techniques in order to get a characterization of students profile, identifiable by labels and features. The results can be used to generate an information technology (IT) tool as a support of tutoring activities.

These three datasets are characterized by very high dimensions: they contain information about a huge number of students (hundred thousand observations) and the features collected about each student are most of the time extracted by questionnaires composed by hundreds of singular questions. For this reason, a big effort in data preprocessing is required to obtain the final dataset to be used in the analysis. Moreover, the big size of data implies a series of computational issues that need to be faced, regarding, for example, the convergence timing of the algorithms.

[1] In Italy, students attend five years of primary school, three years of junior secondary school and five years of upper secondary school.

All analyses undertaken in this thesis are conducted using the statistical software R (R Core Team, 2014) and the main functions of the code will be soon released in dedicated R packages.

Chapter 1

Student and school performance across countries: a machine learning approach via random-effects regression trees and boosting

The educational activity involves a complex process whereby inputs (such as human and financial resources) are converted into outputs. By analogy with the type of production function that is typically used to analyse the technology of a firm, the labour and capital inputs used by a school are likely to influence its output. But, since students themselves form both an input and output, and since they themselves are transformed by the experience of education, such a simple framework fails adequately to capture some key salient features of the process. This is a very well-known challenge in the existent literature about Educational Production Function (EPF). Indeed, the learning process of students is influenced by students' own characteristics, those of their family, their peers, the neighbourhood in which they live, as well as by the characteristics of the school that they are attending. Moreover, the way in which various inputs (at different levels) affect output is likely to vary substantially across the educational systems that operate in different countries. A common characteristic of all educational systems is the hierarchical structure in which students are nested within classes, that are nested within schools, that are in turn nested within cities and so forth. Establishing the structure of such a hierarchy is a non-trivial exercise, not least because this structure may be different across countries. Exploring international datasets which contain information about students' performance in more countries can be a rational approach to understand how the differences among educational systems can have an impact on students' results, all else

equal (Hanushek and Woessmann, 2010).

Our aim in this chapter is to analyze the OECD-PISA 2015 database in order to identify which are the student and school level characteristics that are related to students' achievement, with the aim of investigating the impact of these characteristics on the outcome. We analyze the school systems of nine large developed countries: Australia, Canada, France, Germany, Italy, Japan, Spain, UK, USA. Specifically, our research questions are:

- Which student level characteristics are related to student achievement?
- How much of the total variability in student achievement can be explained by the difference between schools and how can we estimate the school value-added?
- Which school level characteristics are related to school value-added and in what way?
- How do co-factors interact with each other in determining outcomes simultaneously?
- How do these relationships between inputs/covariates and outputs/test scores vary across countries?

In order to address these issues, we run a two stage-analysis, that departs from traditional EPFs approach and embraces a Machine Learning strategy:

1. In the first stage, we apply multilevel regression trees (RE-EM tree, Sela and Simonoff (2012)) in which we consider students (level 1) nested within schools (level 2). By means of this model we can both analyse which are the student level variables that are related to student achievements and estimate the school value-added, as a random effect (grouping factor in the hierarchical model).
2. In the second stage, we apply regression trees and boosting to identify which are the school level characteristics related to school value-added (estimated at first stage), how they are related with the outcome and how they interact among each other.

The set of analytical tools that we use to examine these issues is new to the literature, but is quickly gaining in popularity. Tree-based methods can be classified as a *Machine Learning* (ML) approach. The main difference between statistical and ML approaches is that while the former starts by assuming an appropriate data model and then estimates the parameters from the data, the latter avoids starting with a data model and rather uses an algorithm to learn the relationships between the response and the predictors (in our setting, students'

test scores and their determinants, respectively). Furthermore, ML approach assumes that the data-generating process is *complex* and *unknown* and tries to identify the dominant patterns by observing inputs and the responses (Elith et al., 2008).

Tree-based methods (extended to accommodate the multilevel context) fit the problem in hand well for several reasons. First of all, this methodology takes into account the hierarchical structure of data. The two levels of analysis are students (level 1) that are nested within schools (level 2) and it is worth disentangling the portions of variability explained at each level. Multilevel models are well suited to this. Secondly, our tree-based methodology does not force any particular functional form on the input-output relationship, and it allows for interactions among the predictors. This point is essential because the functional form of the relationships between the covariates and the outcome is unknown a priori and forcing it to be linear can considerably bias the results and, critically, it does not allow discovery of the most likely relationships between the variables. Moreover, there are reasons to believe that the educational context is intrinsically characterised by interactions among variables, since inputs are various and coexist in the same environment. So, tree-based models, that are able to let the variables interact and that identify which interactions are relevant in influencing the outcome, are definitely attractive (Mullainathan et al., 2017). Thirdly, the method allows a clear graphical representation of the results that helps in communicating them to policy practitioners. Alongside the deep interrogation of interactive effects, we consider this to be a major benefit of this approach.

The remainder of the chapter is organised as follows: in Section 1.1 we review the existing literature and, in so doing, motivate our model choice; in Section 1.2 we present the PISA dataset and the countries that we analyse; Section 1.3 discusses the methodological approach (multilevel trees and boosting); in Section 1.4 we report the results and in Section 1.5 we derive conclusions and policy implications.

1.1 Background and previous literature

In recent decades, many researchers have studied the determinants of student achievement, in order to develop policy implications aimed at improving educational systems across the world. The statistical methods proposed by the literature in this perspective are various - including linear regression, multilevel linear models and stochastic frontier analysis - in each case aimed at parameterising the educational production function (EPF). While a complete literature review of previous studies that use a EPF approach is beyond the scope of this chapter, we report important points from existing contributions that can be considered as relevant for interpreting our approach. Specifically, we focus on

those studies which adopt a cross-national perspective in modelling the determinants of students' educational performance by means of economic models and statistical and econometric empirical tools. Indeed, our main contribution to the academic literature stems from the relevance of the innovations brought by the ML strategy to explore differences in educational production across countries.

The Programme for International Student Assessment (PISA) was initiated by the OECD, and has been running since 2000. It involves standardised testing of 15 year olds across a large number of countries. Over the 15 years for which data are now available, PISA results have revealed that there are big discrepancies across education systems. The data allow direct comparisons of student performance in science, reading and mathematics, leading to a ranking of the countries and identifying those that score the best results (OECD, 2016). PISA2015 data, for example, show that Singapore achieves the best results in the scientific area, followed by Japan, Estonia, Finland and Canada. For our purposes, the most interesting aspect of the PISA data is the possibility that they offer to compare the marginal effects of student and school levels variables on students' performance. Gender, immigrant status, socio-economic status (SES), proportion of disadvantaged students, school size and characteristics of the school principal are all variables that have been found to be very important in some countries but less so in others (Owens, 2013; Stacey, 2015). For example, in almost all countries boys perform on average better than girls in the scientific subjects, with the notable exception of Finland, where girls have on average higher results than boys. As another example, after accounting for socio-economical status, immigrant students have a double probability compared to their not immigrant counterparts to achieve low results in scientific subjects (Peña-López et al., 2016). Focusing on mathematics, four Asian countries outperform *all* other economies - Singapore, Hong Kong (China), Macao (China) and Chinese Taipei - and Japan is the strongest performer among all the OECD countries.

Policy responses to internationally reported PISA results have differed among participating countries. For example, in some country groups PISA deficits have been associated with a push towards more centralised control, while others have responded with much more focused reforms implemented with the specific aim of raising PISA (or similar) test scores over time (Wiseman et al., 2013).

What is clear to experts and analysts worldwide, therefore, is that the educational systems, in their structural, internal complexity and in their various aspects, vary within and across countries. Different variables play a role and sometimes with different impacts in influencing educational results in different contexts. Analysing international datasets like PISA therefore calls for the use of a flexible model, able to identify the significant variables within each system and to fit data with different patterns. Indeed, imposing the same coefficient on the correlation between covariates and educational results in all countries is

inappropriate and even the inclusion of country fixed-effects - shifting only the intercept - is not obviously an adequate solution. Therefore, it is necessary to employ more flexible instruments for the analysis of patterns that go beyond the simply "fixed-effects" which impose homogeneity of the interactions between key variables within countries.

The EPF literature builds upon the work of Coleman, Hanushek, and others by viewing education as a process in which students' performance or output (attainment or years of schooling completed) is produced from inputs including school resources, teacher quality, family attributes, and peer quality. Because outcomes cannot be changed by fiat, policy attention has focused on inputs. These include inputs that are both directly controlled by policymakers (characteristics of schools, teachers, curricula, etc.) and those that are not so controlled (family, friends, the learning capacities of the student, etc.) (Hanushek, 2008). While a large part of the effect on students' attainments is due to these "uncontrolled" characetristics of students (Coleman et al., 1966), many researchers have found that schools' and teachers' characteristics are also of importance in determining outcomes (Hanushek et al., 1996; Angrist and Lavy, 1999; Rivkin et al., 2005; Word et al., 1990).

In this chapter, we try to find out which are the inputs that are related with students' performances (output) and in our perspective, three main points need to be taken into account when modelling the educational production functions:

- *Data levels of grouping*: educational data have a hierarchical structure and it is important to distinguish and disentangle the portion of variability in student achievements due to different levels of grouping (between and within classes and schools).

- *Realistic assumptions*: since the educational system is a complex and unknown process, the model assumptions are a sensitive issue and are one of the main weak points of the parametric approaches to the problem. Most of the statistical approaches force the data to be explained through a functional form chosen *a priori*, but the imposition of such a functional form may be inappropriate - either because it does not reflect the underlying technology in some contexts (countries) or, even in none. Therefore, there is the need of a flexible approach that does not force any functional relationships among the variables, where the functional form is not known and that admits the eventuality that the relationship between a covariate (for instance, school resources) and educational results (for example, students' test scores) may be non linear.

- *Interactions*: interactions between cofactors (both within and between levels) are inevitable, as, for example, the relationship between average socioeconomic status of students and class/school size. In such a perspective,

modelling the educational production function would require the inclusion of interaction factors that better describe how covariates combine to influence educational performances.

Most of the classical statistical techniques used in the literature to model educational data do not fulfill these requirements. As we state in the Introduction, the application of hierarchical models to educational data is straightforward and multilevel approaches have been broadly applied in the literature. Even where all the approaches that we list in the Introduction do indeed model the hierarchical structure of data, however, they still force the covariates to have a linear (or a defined functional form) relationship with the outputs, without allowing possible heterogeneous interactions among the predictors.

The innovation of the present study involves the combination of the EPF approach with a multilevel approach to estimation using a machine learning (ML) method. This allows us to relax the parametric assumptions and to discover the data generating process that lies behind our data. The fundamental insight behind ML approaches is as much statistical as computational and its success is largely due to its ability to discover complex structure that does not need to be imposed by the researcher in advance. It manages to find complex and very flexible functional forms in the data without simply overfitting: it finds functions that work well out-of-sample (Mullainathan et al., 2017).

Spurred by the need to relax the parametric assumptions and to explain complex systems, some researchers have already adopted a ML approach for studying some key economic and social relevant issues. *Varian* (Varian, 2014) states that

"*conventional statistical and econometric techniques such as regression often work well, but there are issues unique to big datasets that may require different tools. First, the sheer size of the data involved may require more powerful data manipulation tools. Second, we may have more potential predictors than appropriate for estimation, so we need to do some kind of variable selection. Third, large datasets may allow for more flexible relationships than simple linear models. Machine learning techniques such as decision trees, support vector machines, neural nets, deep learning, and so on may allow for more effective ways to model complex relationships.*"

Various studies on the comparison of the performance of regression and classification trees and conventional statistical methods have already been done: *Fitzpatrick & Mues* (Fitzpatrick and Mues, 2016), for example, apply different modelling approaches for future mortgage default status and they show that boosted regression trees significantly outperform logistic regression. *Savona* (Savona, 2014) realizes an early warning system for hedge funds based on specific red flags that help detect the symptoms of impending extreme negative returns and the contagion effect. He uses regression tree analysis to identify a series of

splitting rules that act as risk signals and he compares these results with the ones obtained applying logistic regression, showing that they are consistent.

The work proposed in this chapter is not the first in which regression trees have been applied in an educational context. *Thomas & Galambos* (Thomas and Galambos, 2004) apply regression and decision trees to investigate how students' characteristics and experiences affect satisfaction. The data mining approach is able to identify the specific aspects of students' university experience that most influence students' satisfaction, in a survey of students in Iowa city (IA). *Ma* (Ma, 2005) analyses students' performances at middle and high schools employing a two-stage analysis, the first stage of which involves estimation of the rate of growth in mathematics achievements of each student, by means of a hierarchical linear model (HML), while the second stage applies classification and regression trees (CART) to students' characteristics. *Cortez & Silva* (Cortez and Silva, 2008) apply some Data Mining (DM) methods such as regression trees and random forests to relate Portuguese secondary school students' scores in mathematics and reading to students' characteristics. *Grayson* (Grayson, 1997) merges results of students at York University in Toronto that were surveyed at the end of the first year with information on grades from administrative records, by means of regression trees.

In this chapter, we relax the assumption of linear effects of student-level covariates on their performance, instead modelling this relationship by means of flexible regression trees. In the first stage of the analysis, we therefore combine multilevel models with regression trees. In the second stage, when exploring the factors associated to the school value-added, we again employ regression trees, combining this method with a boosting procedure, so gaining more precise estimates of determinants of school performance. This type of research is very much in its infancy. We are aware of only one other study, Gabriel et al. (2017), - conducted concurrently with and independently of the present research - that uses regression trees in an education context. That study also draws on PISA data, but focuses specifically on mathematics achievement in Australia.

1.2 The OECD-PISA dataset

The Programme for International Student Assessment (PISA) data assesses student performance, on a triennial basis, in science, mathematics, reading, collaborative problem solving and financial literacy. In our analysis, we use PISA data for 2015, focusing on 9 countries: Australia, Canada, France, Germany, Italy, Japan, Spain, UK and USA. The selection of countries is motivated by the attempt of representing different "types" of educational systems: Anglo-Saxon, Asian, Continental-Europe and Southern Europe. Future research will be realized to extend the analysis to other educational regimes, such as Nordic

countries, South America and Africa. We also need to keep the number of countries quite limited, for favoring easy interpretation of results and their comparison. PISA requires both students and school principals to compile a questionnaire. We therefore have information both at student and school levels. The school questionnaire contains around 30 multiple choice questions about (i) school background information, (ii) school management, (iii) teaching staff, (iv) assessment and evaluation, (v) targeted groups (eg how schools might organise instruction differently for students with different abilities) and (vi) school climate. Meanwhile the student questionnaire contains around 50 multiple choice questions about the (i) student, student's family and student's home (home resources, parents support), (ii) student's view about his/her life (anxiety, effort, collaboration, perception of school climate), (iii) student's school, (iv) student's school schedule and learning time and (v) student's view on science. In addition, students are required to undertake tests in several subjects, and, upon completion, is awarded ten scores for each subject, measuring different abilities within each subject. For example, in science, these scores measure students' ability to explain phenomena scientifically, to evaluate and design scientific enquiry, and to interpret data and evidence scientifically; in reading, they measure student's ability in retrieving information, forming a broad understanding, developing an interpretation, reflecting on and evaluating the content of a text, reflecting on and evaluating the form of a text, etc.; and in mathematics, they measure students' ability in identifying the mathematical aspects of a problem situated in a real-world context and identifying the significant variables, recognising mathematical structure (including regularities, relationships and patterns) in problems or situations, simplifying a situation or problem in order to make it amenable to mathematical analysis and so on. The ten scores are very highly correlated within each subject (coefficient of correlation $\simeq 0.8/0.9$). In each country, test scores have been standardised in order to have mean = 500 and standard deviation = 100. Some other variables, noted in the following tables, are indicators built by PISA and have been standardised so that the mean = 0 and standard deviation = 1. An example is ESCS, which is a weighted average of measures of parental education, wealth, home educational resources and cultural possessions. In our analysis, we focus on mathematics test scores, choosing just one of the ten scores (the same one for each country) as answer variable. We report in Tables 1.1 and 1.2 the variables used in our two-stage analysis, with full definitions[1].

[1]We report here the students' score in mathematics, since this will be our response variable in the model. We do not consider students' scores in other educational subjects in the analysis. In order to have a complete overview of the data collected by PISA, refer to the PISA 2015 technical report in *http://www.oecd.org/pisa/data/2015-technical-report/.*

Variable name	Type	Explanation
MATH SCORE	num	Mathematics PISA test score (mean = 0, sd = 1)
GENDER	0/1	0=male 1=female
ESCS	num	Socio-economical status (mean = 0, sd = 1)
IMMIGRANT	cat	0 = not immigrant student 1 = first generation immigrant 2 = second generation immigrant
TIME HOMEWORK	int	Number of hours of student homework per week
HISCED	cat	Highest level of education of parents (levels from 0 to 6)
VIDEO GAME	0/1	Whether the student plays video games or not
SPORT	0/1	Whether the student plays sport or not
DISCIPLIN CLIMATE	num	How is the disciplinary climate in class
TEACHER SUPPORT	num	Teacher support in class
MMINS	num	Hours of mathematics lessons/week
BELONG	num	Subjective well-being: sense of belonging to school
MOTIVAT	num	Student Attitudes, Preferences and Self-related beliefs: motivation
ANXTEST	num	Personality: test anxiety
COOPERATE	num	Collaboration and teamwork dispositions: Enjoy cooperation
PARENTS SUPPORT	num	Parents emotional support
CULTURAL POSSESSION	num	Cultural possession at home
HOME EDUCAT RESOURC	num	Home educational resources

Table 1.1: List of student level variables of PISA2015 survey used in the analysis, with the relative explanations. Note: we report here only the test score in mathematics that we use as answer variable in the first stage of the analysis. In each country, we standardize the test score in order to have mean = 0 and sd = 1. All variables from "DISCIPLIN CLIMATE" to the end are indicators built by PISA and have mean = 0 and sd = 1.

Variable name	Type	Explanation
# STUDENTS	num	Number of students in the school
RATIO-COMPUTER-STUD	num	Number of available computers per student
MANAGEMENT1	1/6	How much the school principal uses student performance results to develop school's educational goals
MANAGEMENT2	1/6	How much the school principal discusses schools' academic goals with teachers at faculty meetings
STUD-ADMIT-RECORD	0/1	Whether the students are admitted to the school depending on their previous scores or not
PRIVATE	0/1	0 = Public school 1 = Private school
% GOVERN FUNDS	num	Percentage of school funds given by the government
TEACHERS-INADEQ	1/4	How much the principal thinks that teachers are inadequate (on a 1 to 4 scale)
MATERIALS-INADEQ	1/4	How much the principal thinks that materials are inadequate (on a 1 to 4 scale)
INFRASTRUCT-INADEQ	1/4	How much the principal thinks that infrastructures are inadequate (on a 1 to 4 scale)
RATIO-STUDENTS-TEACHER	num	Student-teacher ratio
RATIO-STUDENTS-TEACHER5	num	Student-teacher with level 5 ratio
% STUD SPECIAL NEEDS	num	Proportion of students with special needs
% DISADVANT STUDENTS	num	Proportion of disadvantaged students in terms of socio-economical index
STUDENTS TRUANCY	1/4	Students truancy (on a 1 to 4 scale)
STUD-NO-RESPECT-TEACH	1/4	Students lack respect for teachers (on a 1 to 4 scale)
TEACHER ABSENTEEISM	1/4	Teacher absenteeism (on a 1 to 4 scale)
% PARENTS SPEAK TEACHERS	num	Proportion of students' parents speaking with teachers at the meeting
% PARENTS IN SCHOOL GOVERN	num	Proportion of students' parents participating at the school government

Table 1.2: List of school level variables of PISA2015 survey used in the analysis, with the relative explanations. Note: all variables of type n_1/n_2 assume integer values ranging from n_1 to n_2, with the maximum value corresponding to n_2.

Table 1.3 reports the sample size in the different countries, specifying the number of students and the number of schools that participated in the PISA

survey. The sample sizes vary somewhat across countries, but we have chosen the countries used in our analysis so as to ensure that there are sufficient observations in each to allow robust conclusions to be drawn.

Lastly, it is worth noting that the percentage of missing data at student level is very low (about 2 to 5 % among countries), while at school level it is slightly higher (about 10 to 25 % among countries). We note, however, that a major advantage of tree-based algorithms concerns their performance in the presence of missing data - see for example Breiman et al. (1984) and Loh et al. (2016).

Country	# Students	# Schools
Australia	14,530	758
Canada	20,058	759
France	6,108	252
Germany	6,504	256
Italy	11,583	474
Japan	6,647	198
Spain	6,736	201
UK	14,157	550
USA	5,712	177

Table 1.3: Sample size in the 9 selected countries.

1.3 Model and methods

We develop and employ a two-stage procedure. In the first stage, we apply a mixed-effects regression tree (RE-EM tree), with only random intercept, in which we consider two levels of grouping: students (level 1) nested within schools (level 2). The response variable of the mixed-effects model is the student PISA test score in maths, this being regressed against a set of student level characteristics (fixed coefficients), plus a random intercept that describes the school effect. By means of this model, we can both estimate the fixed coefficients of the student level predictors on the outcome and the school value-added (corresponding to the random intercept). In the second stage, we regress the estimated school value-added against a set of school level characteristics, by means of regression trees and boosting.

1.3.1 An introduction to tree-based methods

Given an outcome variable and a set of predictors, tree-based methods for regression (James et al., 2013) involve a segmentation or stratification of the predictors space into a number of regions. In order to make a prediction for a given obser-

vation, we typically use the mean of the observations in the region to which it belongs. Building a regression tree involves two steps:

1. We divide the predictor space - that is, the set of possible values for $X_1, X_2 \ldots, X_P$ - into M distinct and non-overlapping regions, $R_1, R_2 \ldots, R_M$. For simplicity, we consider these regions as high-dimensional rectangles (or boxes);

2. For every observation that falls into the region R_m, we make the same prediction, which is the mean of the response values for the observations in R_m.

The regions are chosen in order to minimize the Residual Sum of Squares (RSS):

$$\sum_{m=1}^{M} \sum_{i \in R_m} (y_{im} - \hat{y}_{R_m})^2 \tag{1.1}$$

where $\hat{y}_{R_m}$ is the mean of the observations within the m-th box and y_{im} is the i-th observation within the m-th box.

It is useful to contrast this approach with the more conventional methods typically used in the education economics literature - namely a linear functional form imposed on the education production function. In particular, a linear regression model assumes the following functional form:

$$f(X) = \beta_0 + \sum_{p=1}^{P} X_p \beta_p; \tag{1.2}$$

(where P is the total number of predictors) whereas regression trees assume a model of the form:

$$f(X) = \sum_{m=1}^{M} c_m I_{(X \in R_m)} \tag{1.3}$$

where M is the total number of distinct regions and $R_1, \ldots, R_M$ represent the partition of feature space.

Determining which model is more appropriate depends on the problem: if the relationship among the features and the response is well approximated by a linear model, then an approach such as linear regression will likely work well, and will outperform a method such as a regression tree that does not exploit this linear structure (Varian, 2014). If instead there is a highly non-linear and complex relationship between the features and the response, then decision trees may outperform classical approaches. The complex nature of educational production renders this an ideal candidate for exploring the ability of trees-based methods to interrogate non-linearities and interactions in the data.

In order to give an example of how to read the result of a regression tree, let us imagine that we want to regress stadardised student test scores (that is a continuous variable with mean = 0 and standard deviation = 1) against three covariates: Economic Social and Cultural Status (ESCS, an indicator of socio-economic status defined to be a continuous variable with mean = 0 and standard deviation = 1), number of siblings (variable assuming integer values) and time spent on homework (variable assuming integer values) and that Figure 1.1 reports the result of the regression.

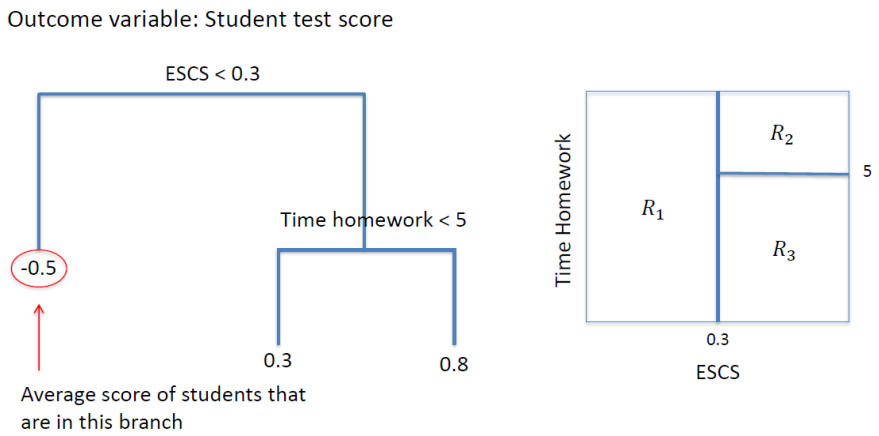


Figure 1.1: Example of the result of a regression tree. The answer variable is students' tests scores (continuous variable with mean = 0 and sd = 1) and the three covariates are: (i) socioeconomic index (ESCS, continuous variable with mean = 0 and sd = 1), (ii) number of siblings (integer variable) and (iii) time of homework (integer variable counting the hours of homework at home). The image on the left represents the partition of the covariate space into three regions, computed by the regression tree. The image on the right represents the regression tree. Variable "number of siblings" does not appear in either the two images, since it does not result to be statistically relevant.

First, we notice that the number of siblings does not appear in the tree. This means that this variable is not able to catch any variability in students' test scores and therefore, the tree excludes it from the splits. When reading the tree, every time the condition at the split point is satisfied, we follow the left branch, otherwise, we follow the one on the right. On the left side of the figure, we see the regression tree while on the right, we see the partition of the covariate space into three regions. The most important variable turns out to be ESCS: a student with an ESCS less than 0.3 follows the left branch yielding a predicted student test score of -0.3; instead, if the student's ESCS exceeds 0.3, he/she goes in the right branch and, at this point, if he/she studies less than 5 hours per week, his/her predicted score is 0.3, while if he/she studies more, it is 0.8. The algorithm itself identifies the threshold values in order to minimize the Residual Sum of Squares (RSS). Focusing on the interaction between the two

covariates, it is noteworthy that the variable "time of homework" matters if the ESCS is higher than 0.3, while it is irrelevant if the ESCS is lower than 0.3.

This brief and simplified explanation serves as a foundation for the methods that we discuss in the following two subsections: RE-EM trees and Boosting, which are the ones used in the empirical analysis of this chapter.

1.3.2 Multilevel models and RE-EM trees

RE-EM trees (Sela and Simonoff, 2012) work in a similar fashion to random effects (or multilevel) linear models (Snijders, 2011) but relax the linearity assumptions of the fixed covariates with the response. Given $J = \sum_{i=1}^{N} n_i$ individuals, nested within N groups, a two-level linear model takes the form:

$$y_{ij} = \beta_0 + \sum_{p=1}^{P} \beta_p x_{pij} + b_i + \epsilon_{ij} \tag{1.4}$$

where

$j = 1, \ldots, n_i$ is the index of the $j-$th individual within group i;

$i = 1, \ldots, N$ is the index of the $i-$th group;

y_{ij} is the answer variable of the individual j within group i;

$\boldsymbol{\beta}$ is the (P+1)-dimensional vector of fixed coefficients;

$x_{1ij}, \ldots, x_{Pij}$ are the P (fixed) predictors;

b_i is the (random) effect of the group i on the answer variable (value-added of group i)

and $\boldsymbol{\epsilon}$ is the vector of the residuals.

Both $\mathbf{b}$ and $\boldsymbol{\epsilon}$ are assumed to be normally distributed with mean 0 and variance σ_b^2 and σ_ϵ^2 respectively. The vector of fixed coefficients $\boldsymbol{\beta}$ is the same for all the N groups, while the random intercept b_i changes across groups (b_i is the value-added, positive or negative, of the $i-$th group). The larger is σ_b^2 the larger are the differences across groups.

RE-EM trees merge multilevel models with regression trees, substituting the linear regression of the fixed covariates with a regression tree. So, in place of a linear regression, a regression tree is built to model the relationship between the output (test scores) and the inputs (student characteristics). In our case, the individuals are the students and the groups are the schools. If we consider students (level 1) nested within schools (level 2), the two-levels model (with only random intercept), for pupil $j, j = 1, \ldots, n_i$, $N = \sum_{i=1}^{N} n_i$, in school $i, i = 1, \ldots, N$ takes the form:

$$y_{ij} = f(x_{ij1}, \ldots, x_{ijP}) + b_i + \epsilon_{ij} \tag{1.5}$$

with

$$b \sim N(0, \sigma_b^2), \tag{1.6}$$

$$\epsilon \sim N(0, \sigma_\epsilon^2) \tag{1.7}$$

where f(X) takes the form in (3) and

y_{ij} is the maths PISA test score of student j within school i;
$x_{ij1}, \ldots, x_{ijP}$ are the P predictors at student level;
b_i is the random effect of school i, which in this chapter is interpreted as a school-specific value-added (VA) to the educational performance of the student; and
ϵ_{ij} is the error.

It is generally assumed that the errors ϵ are independent across objects and are uncorrelated with the effects b. Note, however, that autocorrelation structure within the errors for a particular object is allowed; to do this, we allow the variance/covariance matrix of errors to be a non-diagonal matrix. The random effect b_i is still linear with the outcome, while the fixed covariates, that do not change across groups (schools) are related to the outcome by means of a regression tree.

Moreover, one of the advantages of multilevel models is that we can compute the Proportion of Variability explained by Random Effects (PVRE):

$$PVRE = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\epsilon^2}. \tag{1.8}$$

PVRE measures how much of the variability of test scores can be attributed to students' characteristics or to structural differences across schools - in other words, PVRE disentangles the variability of test scores between students from that between schools. Applying RE-EM trees to data of each of the 9 countries, we can both (i) analyse which are the student level variables that are related with students' achievements and in which way and (ii) estimate the school value-added (random effect b_i) to students' achievements and compute the proportion of student scores' variability given by differences across schools (PVRE). With the aim of adequatly considering the structural differences between countries, we estimate the educational production function as specified in the equation (5) separately for each country.

1.3.3 Regression trees and Boosting

Regression trees have a series of advantages: they do not force any functional relationship between the response variable and the covariates; they can be dis-

played graphically and are easily interpretable; they can handle qualitative predictors; they allow interactions among the variables and they can handle missing data. Nevertheless, they suffer from high variance in the estimation of the relationship between covariates and test scores and they are sensitive to outliers. For these reasons, methods have been developed that serve to reduce variance and increase predictive power; these include *bagging*, *random forests* and *boosting* (James et al., 2013).

Boosting (Elith et al., 2008) is a method for improving model accuracy, based on the idea that it is easier to find and average many rough rules of thumb, than to find a single, highly accurate prediction rule (Schapire, 2003). Related techniques - including bagging, stacking and model averaging - also build and merge results from multiple models, but boosting is unique amongst these in that it is sequential: it is a forward, stagewise procedure. In boosting, models (e.g. regression trees) are fitted iteratively to the data, using appropriate methods gradually to increase emphasis on observations that are modelled poorly by the existing collection of trees. Boosting algorithms vary in exactly how they quantify lack of fit and select settings for the next iteration. In the context of regression trees and for regression problems, boosting is a form of "functional gradient descent". Consider a loss function - in this case, a measure (such as deviance) that represents the loss in predictive performance of the educational production function due to a suboptimal model. Boosting is a numerical optimisation technique for minimising the loss function by adding, at each step, a new tree that is chosen from the available trees on the basis that it most reduces the loss function. In applying the Boosting Regression Tree (BRT) method, the first regression tree is the one that, for the selected tree size, maximally reduces the loss function. For each subsequent step, the focus is on the residuals: on variation in the response that is not so far explained by the model. For example, at the second step, a tree is fitted to the residuals of the first tree, and that second tree could contain quite different variables and split points compared with the first. The model is then updated to contain two trees (two terms), and the residuals from this two-term model are calculated, and so on. The process is stagewise (not stepwise), meaning that existing trees are left unchanged as the model is enlarged. The final BRT model is then a linear combination of many trees (usually hundreds or thousands) that can be thought of as a regression model where each term is a tree. A number of parameters control the model-building process: the *learning rate* (lr), that drives the velocity with which the tree is learning, that is, it shrinks the contribution of each tree; the maximum number of trees to be considered; the distribution of response variable; and the *tree complexity* (tc), that is the maximum level of interaction among variables (Elith et al., 2008).

The increase in predictive power obtained by adopting a BRT approach comes at a cost in terms of ease of interpretation. Indeed, with boosting it is no longer possible to display the tree graphically. But the results can nonetheless

be represented quite simply. BRT provides a ranking of the variables, based on their ability to reduce the *node purity* in the tree (Breiman, 2001), that is the significance of each variable. In order to measure the marginal impact of each predictor, Friedman (Friedman, 2001) has proposed the use of *partial dependence plots*. These plots are based on the following idea: consider an arbitrary model obtained by fitting a particular structure (e.g., random forest, support vector machine, or linear regression model) to a given dataset. This dataset includes N observations y_k of a response variable y, for $k = 1, 2, \ldots, N$, along with P covariates denoted x_{ik} for $i = 1, 2, \ldots, P$ and $k = 1, 2, \ldots, N$. The model generates predictions of the form:

$$\hat{y}_k = F(x_{1k}, x_{2k}, \ldots, x_{Pk}) \tag{1.9}$$

for some mathematical function $F(\ldots)$. In the case of a single covariate x_j, Friedman's partial dependence plots are obtained by computing the following average and plotting it over a useful range of x values:

$$\Phi_j(x) = \frac{1}{N} \sum_{k=1}^{N} F(x_{1,k}, \ldots, x_{j-1,k}, x, x_{j+1,k}, \ldots, x_{P,k}) \tag{1.10}$$

The idea is that the function $\Phi_j(x)$ tells us how the value of the variable x_j influences the model predictions $\hat{y}$ after we have "averaged out" the influence of all other variables.

It is possible to visualise also the *joint* effect of two predictors on the response variable. The multivariate extension of the partial dependence plots just described is straightforward: the bivariate partial dependence function $\Phi_{i,j}(x, y)$ for two covariates x_i and x_j is defined analogously to $\Phi_j(x)$ by averaging over all other covariates, and this function is still relatively easy to plot and visualise. In particular:

$$\Phi_{i,j}(x, y) = \frac{1}{N} \sum_{k=1}^{N} F(x_{1,k}, \ldots, x_{i-1,k}, x, x_{i+1,k}, \ldots, x_{j-1,k}, y, x_{j+1,k}, \ldots, x_{p,k}) \tag{1.11}$$

We therefore apply BRT in each country, in the second stage of our analysis, using the estimated school value-added (first stage) as response variable and a set of school-level characteristics as predictors.

1.4 Results

We begin by comparing the results of PISA test in mathematics across the 9 selected countries. Table 1.4 reports descriptive statistics and Figure 1.2 shows their distributions.

Country	Mean	Median	sd
Australia	481.587	480.903	94.443
Canada	505.021	504.813	85.757
France	496.997	503.998	94.647
Germany	509.170	511.604	87.814
Italy	500.235	501.275	89.483
Japan	532.66	536.96	89.256
Spain	491.361	493.681	83.519
UK	490.765	492.591	85.577
USA	467.383	467.286	88.089

Table 1.4: Descriptive statistics of students' PISA2015 test scores in mathematics in the 9 selected countries.

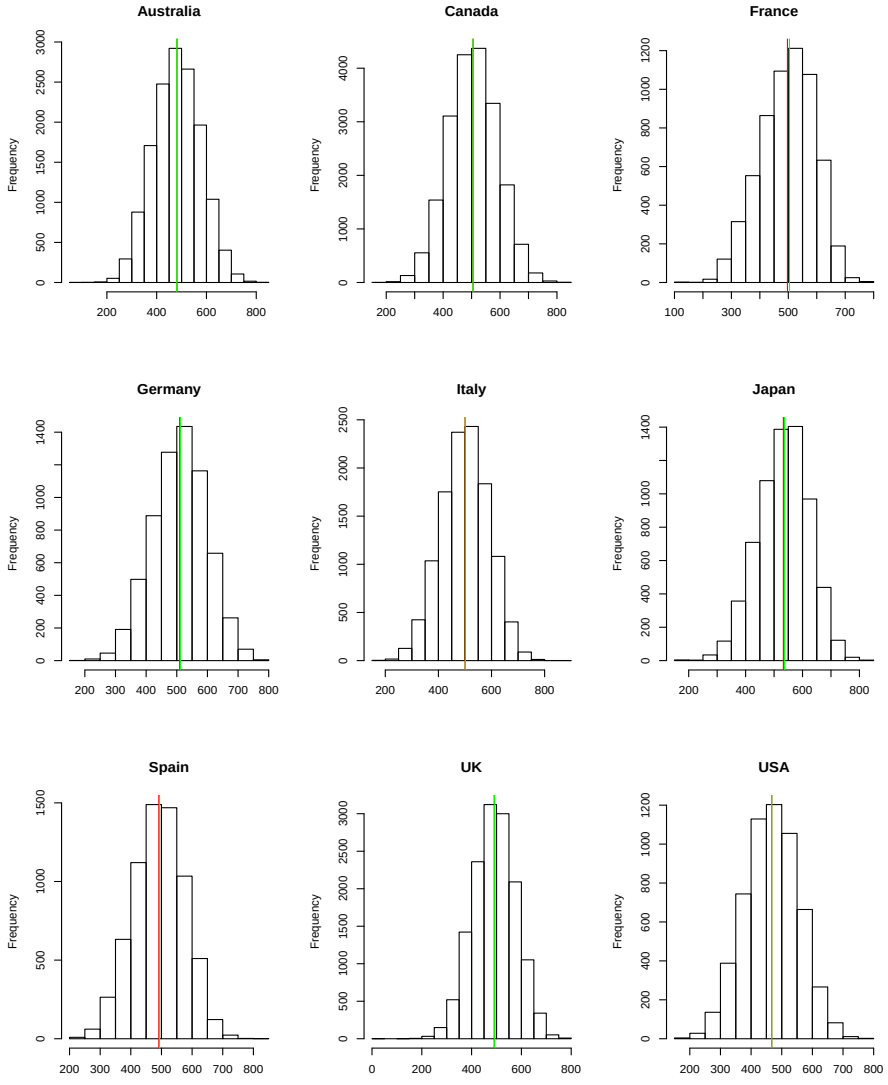


Figure 1.2: Histograms of PISA students test scores in mathematics in the 9 selected countries. Red line refers to the mean, green one to the median. Note: by construction, PISA test scores are standardized at the international level for having mean = 500 and standard deviation = 100.

Japan is the country where students, on average, perform higher test scores, followed by Germany, while USA is the country where students report the lowest scores. In almost all the countries, the mean and median are quite close, suggesting that the distributions are symmetric; France and Japan are exceptions, where in both cases the mean is somewhat smaller than the median, suggesting that there is a slightly higher proportion of students with relatively low test scores.

Country	σ^2_ϵ	σ^2_b	PVRE	PV
Australia	0.690	0.125	15.41%	33.59%
Canada	0.724	0.143	16.49%	29.93%
France	0.464	0.419	47.47%	55.28%
Germany	0.525	0.437	45.44%	50.17%
Italy	0.568	0.395	41.04%	45.57%
Japan	0.510	0.437	46.13%	50.32%
Spain	0.706	0.068	0.08%	30.11%
UK	0.695	0.162	18.97%	32.51%
USA	0.689	0.132	16.15%	33.45%

Table 1.5: RE-EM trees results in the nine selected countries.

1.4.1 First stage: Estimating the determinants of students' test scores and school value-added by using RE-EM trees

RE-EM trees are fitted, separately for each country, using the standardised students' PISA test score in maths as response (in each country students' scores have been standardized, having mean 0 and standard deviation 1) and the entire set of student level variables shown in Table 1.1 as predictors. A random intercept is given by the grouping factor of students within schools (identified by school ID). Results of this first stage comprise the regression tree with the coefficients for the inputs of individual students' characteristics, the proportion of explained variability by the multilevel model (PV) and the PVRE, within each country.

Figure 1.3 shows the trees of fixed student level covariates in each country[2], while Table 1.5 shows the estimated variance of errors, estimated variance of random effects, PV and PVRE of the RE-EM trees models.

The ability of student features to explain students' achievements varies markedly across countries. In some countries, a quite substantial proportion of the differences in students' achievements are explained by student level variables such as socio-economic index, immigrant status, anxiety in dealing with the scholastic life, self-motivation and so on. France, Japan and Germany, that have high PVs (55.28%,50.32% and 50.17% respectively), are examples of this kind. In other countries, such as Canada and Spain, it seems that these student characteristics are not sufficient to explain much of the variability in outcomes. Despite these differences, Figure A.1 in Appendix A shows that the impact of several types of student characteristics are coherent across countries. In almost all the countries, the grape of the most important variables includes (1) the indicator that

[2]We only report here the figure for Australia, while the figures for other countries are reported in Appendix A in Figure A.1.

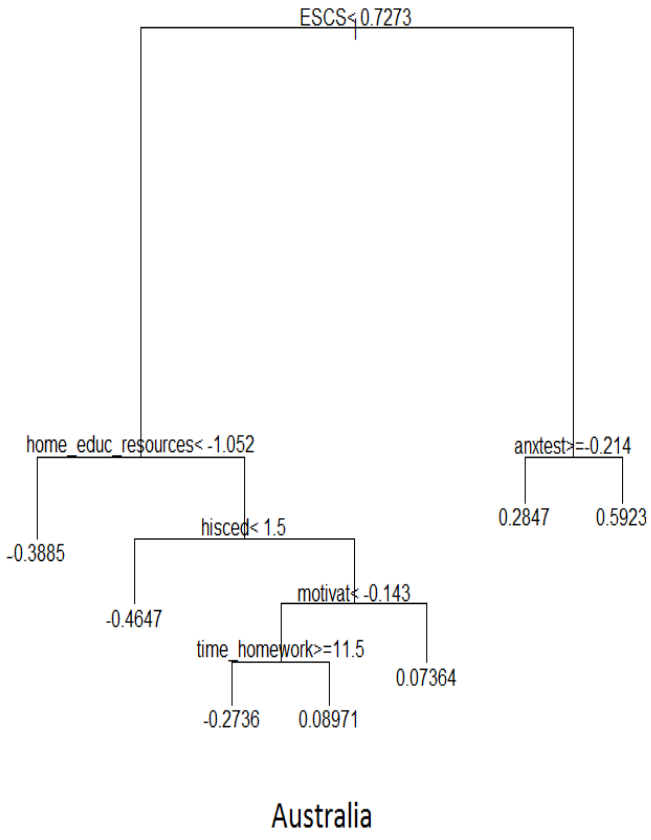


Figure 1.3: Fixed effect tree of first stage analysis (RE-EM tree in model 1.5) in Australia.

measures students' self-reported anxiety toward tests, (2) socio-economic index (ESCS) and (3) the indicator measuring the self-reported motivation. In particular, the ESCS turns out to be the most important variable within five countries out of the nine (Australia, France, Spain, UK and USA). In Canada, Germany and Italy, the most significant variable is ANXTEST: students that feel anxious in their studies have on average lower test scores than more confident students. Japan is the only country where students' self-motivation is the most important variable: if a student has an index of self-motivation less than a certain threshold (in this case, less than -0.9017), then no other variables matter in predicting achievement; otherwise, parents' education and anxiety matter. Other recurrent variables are the highest educational level of parents (HISCED), the educational resources at home, the disciplinary climate and the number of minutes in the maths lesson. Parental education is a particularly relevant variable in Australia, Italy and Japan. Higher levels of parental education are associated with better student achievement. While in Australia and Italy, the different impact of parental education is between parents with less or more than ISCED2 (lower secondary), in Japan the difference is between students with parents with less or more than ISCED4 (post-secondary). Disciplinary climate results to be an important factor in UK and USA: apparently, students that perceive a good disciplinary climate in the class, perform on average better than others.

When tuning to the estimation of school value-added, it differs across countries, with some countries showing a stronger role of schools in affecting test scores than others. In France, for example, almost the 50% (PVRE = 47.47%) of the unexplained variability among students is captured by the "school effect". This means that results of students attending different schools also differ, probably due to heterogeneity in schools' quality. By way of contrast, Spain is a country in which students' achievements are quite homogeneous across schools (PVRE = 0.08%). In general, schools have a clear role to play in explaining the variability of students' scores in France, Japan, Germany and Italy (about 40/45%); in Australia, Canada, UK and the USA, a smaller - but still non-negligible - portion of variability is explained at school level (about 15/20%). This is a finding with very clear policy implications - policies aimed at schools (rather than, say, families) are likely to have much more potency in the former group of countries than in the latter.

Different students' achievements across schools may be the consequence of different school policy and teaching programmes or of the socio-economic composition of the school body (Orfield et al., 2012). While the available data and the proposed methodology do not allow investigation of the channels that drive the causal relationships between schools' characteristics and test scores, the next section uses regression trees and boosting to show correlations between schools' features and their estimated "value-added".

1.4.2 Second stage: Modelling the determinants of school value-added through regression trees and boosting

In the second stage of the analysis, we run, within each country, a regression model based on trees and boosting. The response variable is the school value-added, as estimated at the first stage, while the predictors are the school level variables described in previous section and contained in the questionnaire filled by school principals. Figure 1.4 and Table 1.6 show the variable importance ranking within each country[3] and the proportion of total variability explained by the model, respectively.

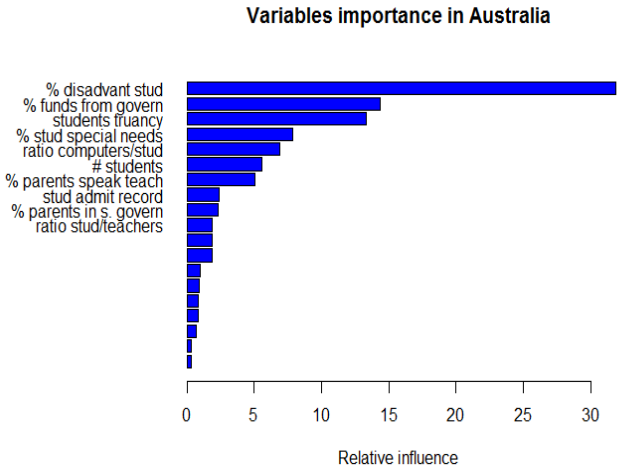


Figure 1.4: School level variables importance ranking in the second stage of the analysis in Australia. Boosting creates a ranking of the relative influences of the covariates on the outcome variable (school value-added). To lighten the reading, we report here only the first ten most important variables (where the most important variable is the one able to catch the biggest part of variability in the outcome).

	Australia	Canada	France	Germany	Italy
PV	40.36%	28.09%	59.13%	53.08%	28.09%
	Japan	Spain	UK	USA	
PV	30.87%	14.15%	39.12%	35.81%	

Table 1.6: Proportion of explained variability (PV) of the second stage boosting model, in the 9 selected countries.

[3]We only report here the figure for Australia, while the figures for other countries are reported in Appendix A in Figure A.2.

We report in the figures only the ten most important variables within each country, both because the remaining variables are statistically irrelevant and to lighten the reading. School size ("# students"), proportion of disadvantaged students, proportion of students with special needs, students' truancy and the ratio of computers to students are typically the most important variables in each country (see Figure A.2 in Appendix A). This means that the school value-added is mainly associated with students' socioeconomic composition and to school size, more so than with managerial characteristics or proxies for resources, as inadequacy of materials and infrastructure. Besides these four main variables, participation of parents, measured both as proportion of parents speaking with teachers and participating in school governance, and the percentage of funds given by the government are also important in some countries to qualify the estimated schools value-added.

Describing the patterns of the impact of school variables on schools value-added

After identifying the important variables, in order to detect the magnitude and the way in which these predictors are associated with the response, we visualise in Figure 1.5 the partial plots of the four most significant variables within each country[4], noting that these differ across countries.

The proportion of disadvantaged students is one of the four most important variables in all the countries except for Japan. Schools with higher proportions of disadvantaged students are those with lower estimated value-added. On average, schools with a high proportion of disadvantaged students suffer a negative impact on performances. In particular, in almost all countries, the impact of this variable on schools value-added is negative in its range from 0% to 30/40%. By way of contrast, in the USA, schools in which the proportion of disadvantaged students lies between 0 and 20 tend not to differ in terms of outcomes *ceteris paribus*, while there is a monotonic negative association between the covariate and the response in the covariate range between 20 and 100. Thus, there are countries in which the substantial difference is between schools composed by only advantaged students and schools with a minimum proportion of disadvantaged ones, while there are countries, such as the USA, in which the the proportion of disadvantaged students is influential only if it is quite high (more than 20%).

Another important determinant of outcomes in all countries, with the exception of Australia, is school size. In general, bigger schools are associated with higher school value-added. The impact of this variable is highly nonlinear and this can be an explanation about why some previous literature fails to find any statistical (linear) correlation between performances and size. In all countries,

[4]We only report here the figure for Australia, while the figures for other countries are reported in Appendix A in Figure A.3.

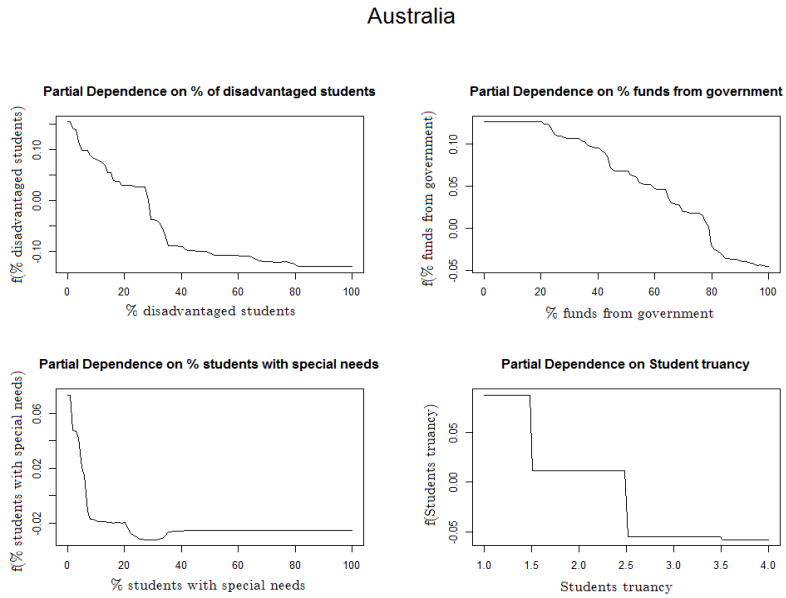


Figure 1.5: Partial plot of the four most important school level variables in the association with school value-added, in Australia. Note: the selection of the four most significant variables is taken from Figure 1.4 and the explanation of each school level covariate is given in Table 1.2.

except for Australia and USA, the school value-added rapidly increases when the school size ranges between about 500 and $1,000$ students. Schools smaller than 500 students perform in a quite similar way to schools larger than about $1,000$ students. The USA provides an interesting exception: very small schools (with fewer than 500 students) are associated with very high school value-added, while there is a negative peak corresponding to schools attended by about 500 students, that is the value associated with the lowest school value-added. Again, from 500 on, larger schools are estimated to have higher value-added.

The proportion of students with special needs is important as a determinant of outcomes in all countries, except Canada and Japan. Schools with a higher proportion of students with special needs are associated to lower school value-added. Again, there is a gap in the response value when the covariate ranges between 0% and 20%. The number of schools with more than 20% of students with special needs is small, but still we have observations in this range that do not differ in their impact on the response.

Another recurrent important variable is the one measuring the students truancy. Students truancy is an indicator about how much students take seriously their presence at school and therefore, their education. In Australia, Canada, Japan and USA it is one of the four most important variables. Schools with higher proportion of students that tend to skip school days are associated to lower school value-added, in a quite intuitive way, with strong effects after a threshold when the number of days skipped is > 2.5.

The percentage of funds given to the school from the government is a key determinant of schools' effectiveness in both Australia and Japan. In Australia, the trend is very well defined: when the percentage of funds given by the government increases, the school value-added decreases. From the literature (Marginson, 1993; Anderson, 1993), we know that in Australia, private schools, which receive less funds from the government respect to public schools, are more likely to perform better than public ones and therefore these two aspects are probably strongly connected. Even if a dummy variable for public/private schools is considered, the percentage of funds given by the government still reflects some of the public/private heterogeneities and it is actually able to catch more variability in the response than the dummy variable. Also in Japan the partial effect of the percentage of funds given by the government on the school value-added is related to the difference between private and public schools. In Japan, contrary to Australia, PISA2015 data indicate that private schools have, on average, lower performance when compared with public schools. Moreover, private schools usually receive about 40/50% of their funds from the government. The trend of the impact of the covariate on the response is less clear than the one in Australia.

Lastly, in Canada and in Italy the percentage of parents speaking with teachers or participating in school governance are important. An increase in cofactor values is positively associated with the school value-added: schools in which

parents are actively interested in their children's education experience more favourable outcomes than do others. Likewise, in Spain the percentage of parents participating in school governance, when in the range from 0 to 50%, has a positive effect on outcomes.

The last variable that appears in the four most important variables of France, Germany, Japan and UK is the number of computers per student ("ratio comp / stud"). This covariate has a counterintuitive association with school value-added. In Japan and UK (see Japan and UK panels in Figure A.3 in Appendix A), an increase of number of computers per student is associated with a decrease in school value-added. In Germany (see Germany panel in Figure A.3 in Appendix A), there is a peak around 0.4 and a trough around 0.6. Lastly, in France (see France panel in Figure A.3 in Appendix A), the highest value-added corresponds to zero computers, but there is a peak around 1, maybe suggesting that one computer per person is the right balance. A possible interpretation of these trends is that too many computers (more than one per person) may be sign of inefficient management of school funds. Alternatively it might be the case that national policies have concentrated the IT facilities in less advantaged schools with lower test scores - in this case, the statistical relationship would be biased.

Describing the impact of joint variables on schools value-added

Up to this point, we have investigated the partial effect of predictors one by one, on a *ceteris paribus* basis. But one of the main strengths of the regression tree approach is that it allows consideration of circumstances in which more than one cofactor changes simultaneously, so affecting simultaneously the dependent variable (in our case, school value-added). We now turn, therefore, to focus on the visualisation of the joint effect of two predictors on the response, and in so doing investigate the interaction effect of the most significant variables within each country (Figure 1.6)[5]. Again, the choice of the variables to be included in the graphical illustration is based on the variables that, in the different countries, turned out to be most important in affecting the estimated schools value-added.

In several countries, the impact on outcomes of the joint association between the proportion of disadvantaged students and school size is of interest. From Australia and USA panels, we know that in most countries larger schools perform better than smaller ones and schools with a high proportion of disadvantaged students perform less successfully than others. The extent to which differences in school size affect outcomes depends critically on how high is the proportion of disadvantaged students, however. In Italy and Spain, the proportion of disadvantaged students seems to have a clear negative impact even in the

[5]We only report here the figure for Australia, while the figures for other countries are reported in Appendix A in Figure A.4.

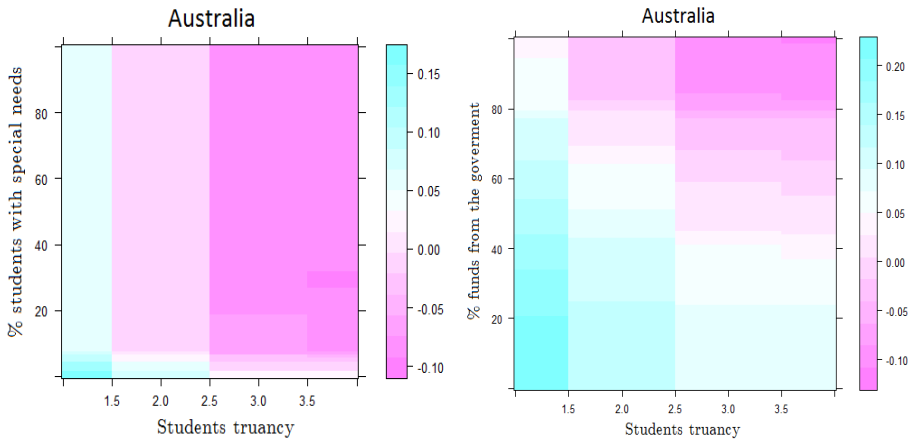


Figure 1.6: Joint partial plot of the most important school level variables in association with school value-added, in Australia. Notes: 1. Colors represent the scale of the values of the response (school value-added). 2. The selection of variables is based on the group of the variables that turn out to be significant in previous steps.

big schools, while small schools with a low proportion of disadvantaged students are not associated with negative effect on value-added. In UK and USA, the interaction is much weaker in the sense that the high proportion of disadvantaged students has a negative impact, almost independently from the school size. The difference between these two countries is that while in UK the threshold value of proportion of disadvantaged students to have a negative impact on the response is about 20/30%, in the USA is much higher, around 70/80%.

Interaction between two variables about the students' socioeconomic composition - namely the proportion of socioeconomically disadvantaged students and proportion of students with special needs - is also interesting and instructive. In France, schools in which both percentages are low perform better than the average while schools where both percentages are high perform worse. However, schools with a high proportion of disadvantaged students nevertheless manage average performance if they have a very small proportion of students with special needs (and *vice versa*). In Germany and Italy, schools with a low proportion of disadvantaged students perform better than the average and the increasing proportion of students with special needs does not affect this performance. On the contrary, schools with a high proportion of disadvantaged students perform worse than the average and the increasing proportion of students with special needs worsens the results even more. In UK, the increase in both proportions contributes to lower school value-added in an almost symmetric way.

Truancy is another variable whose interaction with school size and school body composition is worthy of investigation. "Truancy" is defined by OECD as the propensity for students to skip classes without justification. In Japan,

truancy is associated with very low school value-added only when considering small schools, while, even if it has again a negative impact, we still have positive school value-added in big schools with high students truancy. In USA, schools with low levels of truancy perform better than the average while schools with high truancy rates perform worse than the average, but there is an important interaction with school size - truancy has a more negative association to test scores in smaller rather than in larger schools. In Australia and in Canada, the interaction between students truancy and proportion of disadvantaged students is similar: schools with both high (low) truancy and high (low) proportion of disadvantaged students are associated with negative (positive) school value-added. But, schools with high truancy rates and a low proportion of disadvantaged students (and *vice versa*), are still able to achieve average performance.

In Australia and Japan, truancy and percentage of funds given by the government are very important variables but they interact in an heterogeneous way to affect schools' performance. In Australia, schools with both high (low) students truancy and high (low) percentage of funds given by the government are associated with negative (positive) effects on school value-added, but, in all the other cases, this relationship doen not hold. Instead in Japan, schools with low (high) students truancy perform worse (better) than the average, almost independently from the percentage of funds given by the government.

The last interaction that deserves attention is the one between school size and percentage of parents participating in school governance in Spain: the size of the school is associated with positive school value-added, but only if parents actively participate at the school government and are interested in their children's education.

The visualization of joint partial plots to characterise the determinants of schools value-added proves to be a powerful tool for analysts and decision makers. Indeed, these figures provide an immediate sense of which are the variables with more or less influence on schools value-added, while simultaneously providing information covering the whole distribution of the impacting variables, without forcing to concentrate on average correlations.

1.5 Discussion, concluding remarks and policy implications

The availability of large scale datasets allowing comparative analysis of educational performance has been a major boost to researchers interested in the educational production function. In this chapter, we have applied new methods of analysis, drawn from the machine learning literature, to examine the determinants of students' test scores and schools value-added. The results confirm many of the relationships we knew already from statistical analysis, but provide

a new and enriched understanding of how both nonlinearities amongst and interactions between cofactors determine educational performance. These insights come from a recognition that the education process is *complex*, *unknown* in its specific mechanisms and *heterogeneous across countries*. The tree-based methods that we use represent an *inductive* and non *deductive* way to explain the associations among variables, having two main advantages respect to the classical statistical methods: they do not force any functional relationships between the response (students' results) and the covariates (students' characteristics) and they allow for interactions among the variables.

The first stage of our analysis shows that student-level variables are able to explain part of the variability in their achievements: socio-economic index, anxiety, motivation, gender, and parental education are some of the most influential variables. Their association to test scores and their ability in explaining variability in students' achievements differ substantially across countries. The percentage of variability in students' achievements explained at school level (schools value-added in our terminology here) also varies across countries. Those countries in which the estimated variance of schools value-added is high are characterised by heterogeneity at school level. On the contrary, countries where the variance of schools value-added is limited in magnitude offer a more homogeneous experience across schools. There are clear policy implications in noting, for example, that the ratio of students to teachers has high relative influence in Canada, Japan and Spain, but not elsewhere. In many countries, the actions that can most effectively improve educational outcomes are not educational policies per se, but rather social policies.

After estimating the school value-added in the first stage, we correlate it to school level characteristics in the second stage. Again, we find different school level variables associated to school value-added across countries. The main focus in this stage is the effect of interactions between cofactors, which is modelled by means of joint partial plots. As we have seen, the impact on performance of changes in one variable often depends crucially on the value of other explanatory variables.

Tree-based methods complement linear regression models of educational performance by augmenting them with a richer interrogation of the data. The impact of student and school level variables are often not simply linearly associated with students' achievements; we have uncovered evidence in the data of considerably more complex (and intuitively plausible) patterns. The strength of the machine learning method, in this perspective, is that they literally "learn from the data", finding the dominant patterns without any assumption. Armed with the refined understanding of how different policies can impact differently on schools in various circumstances, policy-makers can better implement change aimed at improved performance.

Several policy implications can be drawn from our analysis. The results show

the relationship between test scores and both school and individual factors to be quite complex, and this presents a challenge to naïve interpretations of school performance tables. A particularly salient aspect of this complexity relates to differences across countries in the impact on educational performance of variables that are not usually thought to pertain to educational policy. Notably in several countries in this study (but not in others), the first branch of the regression tree is defined by ESCS - indicating that (in these countries, but not elsewhere) issues in the sphere of education might most effectively be addressed using social rather than educational policies. The machine learning tools used thus highlight in sharp relief some issues with high policy relevance.

The results obtained in the present chapter should be viewed alongside other research drawn from the literature on educational production functions. In common with much contemporary applied economic research, these studies place emphasis on causality. Further research is needed to introduce sophisticated analysis of causality in the machine learning context, specifically as it applies in the sphere of education.

Chapter 2

Performing learning analytics via generalized mixed-effects trees

The work presented in this chapter is part of the international SPEET project (Student Profile for Enhancing Engineering Tutoring), an ERASMUS$^+$ project aiming to open a new perspective to university tutoring systems. It intends to extract useful information from academic data provided by its partners[1] and to identify different Engineering students profiles across Europe (*www.speet-project.com/the-project*). Here, our goal is to find out which indicators may discriminate between two different student profiles: *dropout* students, who permanently finish their career for any reason other than the achievement of the Bachelor of Science (BSc) degree, and *graduate* students, who complete their career with the achievement of academic qualification. This choice is motivated by the fact that, across all SPEET partners, almost a student out of two leaves his/her Engineering studies before obtaining the BSc degree. If it was possible to know as soon as possible to which profile a student belongs, it would be of valuable help for tutors to improve counseling actions.

Data provided by universities usually includes indicators about the socio-economic background and both current and previous performance of the students. However, academic success depends on different factors, both internal and external (Barbu et al., 2017). The dataset we use in our analysis includes more than 18,000 BSc careers from Politecnico di Milano: it essentially consists of student record data, so it just partially covers these factors. Similar dataset structures have already been used in recent developments oriented to the prediction of performance and detection of dropouts or students at risk (Romero and

[1]Universitat Autonoma de Barcelona (UAB) - Spain; Instituto Politecnico de Braganca (IPB) - Portugal; Opole University of Technology - Poland; Politecnico di Milano (PoliMi) - Italy; Universidad de Leon - Spain; University of Galati *Dunarea de Jos* - Romania.

Ventura, 2010). The hypothesis is that both background and career indicators are enough to identify the students at risk and to draw the attention of tutors, who should complete the student profile with further information.

In our situation, students are naturally nested within the degree programme they are attending. In addition, further levels of hierarchy are possibile, such as programmes within faculties, faculties within universities and finally universities within countries. While investigating the learning process, it is necessary to disentangle the effects given by each level of hierarchy (Bock, 2014). Indeed, if the clustered aspect of the data is not inspected, it may result in a loss of likely valuable information. Multilevel models take into account the hierarchical nature of data and are able to quantify the portion of variability in the response variable that is attributable to each level of grouping (Goldstein, 2011). Generalized Linear Mixed Models (GLMM) fit a multilevel model on a binary response variable, but they impose a linear effect of covariates on a transformation of the response variable (Agresti, 2003). On the contrary, tree-based methods such as the CART model learn the relationship between the response and the predictors by identifying dominant patterns in the training data (Breiman et al., 1984). In addition, these methods allow a clear graphical representation of the results that is easy to communicate. The goal of our study is to propose a novel method able to preserve the flexibility of the CART model and to extend it to a clustered data structure, where multiple observations can be viewed as being sampled within groups.

In the literature this is not the first time in which tree-based methods are adopted to deal with longitudinal and clustered data. In Sela and Simonoff (2012) a regression tree method for longitudinal or clustered data is proposed. This method is called Random Effects Expectation-Maximization (RE-EM) tree. Independently, in Hajjem et al. (2011) a Mixed-Effect Regression Tree (MERT) model is proposed. If clustered observations are considered, these are extensions of a standard regression tree to the case of individuals nested within groups. These methods use observation-level covariates in the splitting process and can deal with the possible random effects associated to those covariates. However, they both deal with a Gaussian response variable and they are not suitable to a classification problem.

In Hajjem et al. (2017) the MERT approach is extended to non-gaussian data and a generalized mixed effects regression tree (GMERT) is proposed. This algorithm is basically the PQL algorithm used to fit GLMMs where the weighted linear mixed-effect pseudo-model is replaced by a weighted MERT pseudo-model. Lastly, the most recent work is proposed in Speiser et al. (2018), where the authors develop a decision tree method for modeling clustered and longitudinal binary outcomes. Even if the aim of the model is very similar to ours, their model only handles binary outcomes using a Bayesian GLMM.

Following a different strategy, our proposed method intends to generalize

the RE-EM tree approach. In particular, in this chapter we expand its use to different classes of response variables from the exponential family: this would allow to extend it to a classification setting. At the same time this method can deal with the grouped data structure, similarly to traditional multilevel models. As in RE-EM tree estimation, we develop an algorithm that disentangles the estimation of fixed and random effects. That is, an initial tree is built ignoring the grouped data structure, a mixed-effects model is fitted based on the resultant tree structure, and a final mixed-effects tree is reported.

In this chapter, we apply this model to the Politecnico di Milano dataset. In this specific case, we can identify which fixed-effects covariates discriminate between dropout and graduate students. Through a GMET model, we can relax the assumption of linear effects of student-level covariates on their performance and we can identify which interactions relevantly influence the career status. In addition, the choice of a multilevel model allows to estimate the degree programme effect on the predicted probability of obtaining the degree.

The chapter is organized as follows. In Section 2.1 we describe model and methods - generalized mixed tree algorithm (GMET) - and in Section 2.2 we show a simulation study. In Section 2.3 we describe the PoliMi dataset, we report the application of the proposed algorithm to the case study and outline the results. Finally, in Section 2.4 we draw our conclusions.

2.1 Model and methods

In this section, we present the proposed generalized mixed-effects tree model (Subsection 2.1.1) and the algorithm for the estimation of its parameters (Subsection 2.1.2).

2.1.1 Generalized mixed-effects tree model

We start considering a generic GLMM. This model is an extension of a generalized linear model that includes both fixed and random effects in the linear predictor (Agresti, 2003). Therefore, GLMMs handle a wide range of response distributions and a wide range of scenarios where observations are grouped in groups rather than completely independently. For a GLMM with a two-level hierarchy, each observation j, for $j = 1, \ldots, n_i$, is nested within a group i, for $i = 1, \ldots, I$. Let $\boldsymbol{y}_i = (y_{1i}, \ldots, y_{n_i i})$ be the n_i-dimensional response vector for observations in the i-th group. Conditionally on random effects denoted by $\boldsymbol{b}_i$, a GLMM assumes that the elements of $\boldsymbol{y}_i$ are independent, with density function from the exponential family, of the form

$$f_i(y_{ij}|\boldsymbol{b}_i) = \exp\left[\frac{y_{ij}\eta_{ij} - a(\eta_{ij})}{\phi} + c(y_{ij}, \phi)\right]$$

where $a(\cdot)$ and $c(\cdot)$ are specified functions, η_{ij} is the natural parameter and ϕ is the dispersion parameter. In addition, we have

$$\begin{aligned} E[y_{ij}|\boldsymbol{b}_i] &= a'(\eta_{ij}) = \mu_{ij} \\ Var[y_{ij}|\boldsymbol{b}_i] &= \phi\, a''(\eta_{ij}) \end{aligned}$$

A monotonic, differentiable link function $g(\cdot)$ specifies the function of the mean that the model equates to the systematic component. Usually, the canonical link function is used, i.e., $g = a'^{-1}$. From now on, without loss of generality the canonical link function is used. In this case, the model is the following (McCulloch and Searle, 2004):

$$\begin{aligned} \boldsymbol{\mu}_i &= E[\boldsymbol{Y}_i|\boldsymbol{b}_i] \qquad i = 1, \ldots, I \\ g(\boldsymbol{\mu}_i) &= \boldsymbol{\eta}_i \\ \boldsymbol{\eta}_i &= X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i \\ \boldsymbol{b}_i &\sim N_q(\mathbf{0}, \Psi) \qquad ind. \end{aligned} \tag{2.1}$$

where i is the group index, I is the total number of groups, n_i is the number of observations within the i-th group and $\sum_{i=1}^{I} n_i = J$, $\boldsymbol{\eta}_i$ is the n_i-dimensional linear predictor vector. In addition, X_i is the $n_i \times (p+1)$ matrix of fixed-effects regressors of observations in group i, $\boldsymbol{\beta}$ is the $(p+1)$-dimensional vector of their coefficients, Z_i is the $n_i \times q$ matrix of regressors for the random effects, $\boldsymbol{b}_i$ is the $(q+1)$-dimensional vector of their coefficients and Ψ is the $q \times q$ within-group covariance matrix of the random effects. Fixed effects are identified by parameters associated to the entire population, while random ones are identified by group-specific parameters.

Our proposed Generalized Mixed-Effects Tree (GMET) method expands the use of tree-based mixed models to different classes of response variables from the exponential family. At the same time the method can deal with the grouped data structure as GLMMs do. We now specify the GMET model. The random component of this model consists of a response variable Y from a distribution in the exponential family. The fixed part in the GMET is not linear as in (2.1) but it is replaced by the function $\boldsymbol{f}(X_i)$ that is estimated through a tree-based algorithm. Thus, the matrix formulation of the model is the following:

$$\begin{aligned} \boldsymbol{\mu}_i &= E[\boldsymbol{Y}_i|\boldsymbol{b}_i] \qquad i = 1, \ldots, I \\ g(\boldsymbol{\mu}_i) &= \boldsymbol{\eta}_i \\ \boldsymbol{\eta}_i &= \boldsymbol{f}(X_i) + Z_i\boldsymbol{b}_i \\ \boldsymbol{b}_i &\sim N_q(\mathbf{0}, \Psi) \qquad ind. \end{aligned} \tag{2.2}$$

where i is the group index, I is the total number of groups, n_i is the number of observations within the i-th group and $\sum_{i=1}^{I} n_i = J$. In addition, $\boldsymbol{\eta}_i$ is the

n_i-dimensional linear predictor vector and $g(\cdot)$ is the link function. Finally, X_i is the $n_i \times (p+1)$ matrix of fixed-effects regressors of observations in group i, Z_i is the $n_i \times q$ matrix of regressors for the random effects, $\boldsymbol{b}_i$ is the $(q+1)$-dimensional vector of their coefficients and Ψ is the $q \times q$ within-group covariance matrix of the random effects. As in a GLMM, $\boldsymbol{b}_i$ and $\boldsymbol{b}_{i'}$ are independent for $i \neq i'$. Fixed effects are identified by a non-parametric CART tree model associated to the entire population, while random ones are identified by group-specific parameters.

Without loss of generality, let us now specify model (2.2) for the case of a binary random variable and univariate random effect. The logit function is the canonical link function:

$$g(\mu_{ij}) = g(p_{ij}) = \log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \text{logit}(p_{ij}).$$

Here, the random-effects structure simplifies to a random intercept. The model formulation for observation y_{ij} may therefore be written as:

$$\begin{aligned} Y_{ij} &\sim \text{Bernoulli}(p_{ij}) \qquad i = 1, ..., I \qquad j = 1, ..., n_i \\ p_{ij} &= E[\, Y_{ij} | b_i] \\ \text{logit}(p_{ij}) &= f(\boldsymbol{x}_{ij}) + b_i \\ b_i &\sim N(0, \sigma^2) \qquad ind. \end{aligned} \tag{2.3}$$

where we observe $\boldsymbol{x}_{ij} = (x_{1ij}, .., x_{ijp})^T$, a $(p+1)$-dimensional vector of fixed-effects covariates for each observation j in group i.

2.1.2 Generalized mixed-effects tree estimation

In this subsection we show the algorithm for the estimation of the parameters of the GMET model (2.2). The basic idea behind the algorithm is to disentangle the estimation of fixed and random effects. The structure of the algorithm is the following:

1. Initialize the estimated random effects $\boldsymbol{b}_i$ to zero.
2. Estimate the target variable μ_{ij} through a generalized linear model (GLM), given fixed-effects covariates $\boldsymbol{x}_{ij} = (x_{ij1}, ..., x_{ijp})^T$ for $i = 1, ..., I$ and $j = 1, ..., n_i$. Get estimate $\hat{\mu}_{ij}$ of target variable μ_{ij}.
3. Build a regression tree approximating f using $\hat{\mu}_{ij}$ as dependent variable and $\boldsymbol{x}_{ij} = (x_{ij1}, ..., x_{ijp})^T$ as vector of covariates. Through this regression tree, define a set of indicator variables $I(\boldsymbol{x}_{ij} \in R_\ell)$ where the index ℓ ranges over all of the terminal nodes in the tree.
4. Fit the mixed effects model (2.2), using y_{ij} as response variable and the set of indicator variables $I(\boldsymbol{x}_{ij} \in R_\ell)$ as fixed-effects covariates. Specifically, for $i = 1, ..., I$ and $j = 1, ..., n_i$, we have $g(\mu_{ij}) = I(\boldsymbol{x}_{ij} \in R_\ell)\gamma_\ell + \boldsymbol{z}_{ij}^T \boldsymbol{b}_i$. Extract $\hat{\boldsymbol{b}}_i$ from the estimated model.

5. Replace the predicted response at each terminal node R_ℓ of the tree with the estimated predicted response $g(\hat{\gamma}_\ell)$ from the mixed-effects model fitted in Step 4.

The GLM in Step 2 is fitted through maximum likelihood. The maximum likelihood estimates can be found using an iteratively reweighted least squares algorithm or a Newton-Raphson method (McCullagh and Nelder, 1989).

The fitting of the tree in Step 3 can be achieved using any tree algorithm, based on any tree-growing rules that are desired. Here, tree building is based on the CART tree algorithm (Breiman et al., 1984). After building a large tree T_0, pruning is advised to avoid overfitting on training data. In principle, any tree-pruning rule could be used; here, we propose cost-complexity pruning (Friedman et al., 2001). It considers a sequence of nested trees indexed by a nonnegative tuning parameter α which controls the trade-off between the subtree's complexity and its fit to the training data. For each value of α exists a subtree $T \subset T_0$ to minimize

$$\sum_{\ell=1}^{|T|} \sum_{x_i \in R_\ell} (y_i - \hat{y}_{R_\ell})^2 + \alpha |T|. \tag{2.4}$$

Here, $|T|$ indicates the number of terminal nodes of tree T. When $\alpha = 0$, then the subtree T will simply be equal to T_0. However, as α increases, the quantity (2.4) will tend to be minimized for a smaller subtree. We can select a value of α using a validation set or using K-fold cross-validation: for example, we can pick $\tilde{\alpha}$ to minimize the average CV error. Tree building and pruning is implemented in `R` library `rpart` (Therneau et al., 2017), according to the CART tree-building algorithm and cost-complexity pruning. In order to ensure that initial trees are sufficiently large, we set the complexity parameter to zero. Thus, the largest tree is grown then pruned based on ten-fold cross-validation error. Instead of choosing the tree that achieves the lowest CV error, we use the so-called *1-SE rule*: any CV error within one standard error of the achieved minimum is marked as being equivalent to the minimum. Among all these equivalent models in terms of CV error, the simplest one is chosen as final tree model.

The generalized linear mixed model in Step 4 can be estimated using fitting techniques that were previosly described. Different statistical packages can estimate those type of models: the `glmer` function of the `R` library `lme4` (Bates et al., 2014) is used here. It fits a generalized linear mixed model via maximum likelihood. For a GLMM the integral must be approximated: the most reliable approximation is adaptive Gauss-Hermite quadrature, at present implemented only for models with a single scalar random effect, otherwise Gaussian quadrature is used.

Prediction for new observations

After estimating a GMET it is possible to make out-of-sample predictions for new observations. Suppose the tree is estimated on data from groups $i = 1, .., I$ for observations y_{ij}, $j = 1, ..., n_i$. Given a new observation $\boldsymbol{x}_{ij'}$ we are able to output its corresponding response since we know the estimation of the fixed-effects function $f(\cdot)$, of the random effects $\boldsymbol{b}_i$ and of the associated covariance matrix Ψ. We may look for two types of prediction:

- predict response $y_{ij'}$ given a new observation $\boldsymbol{x}_{ij'}$ for a group in the sample $i \in \{1, ..., I\}$. We define it a *group-level prediction*.
- predict response $y_{i'j'}$ given an observation $\boldsymbol{x}_{i'j'}$ for a group i' for which there are no observations in our current sample, or for which we do not know the group it belongs to. We define it a *population-level prediction*.

For the first type of prediction, we estimate $f(\boldsymbol{x}_{ij'})$ using the estimated tree and attributes $\boldsymbol{x}_{ij'}$ and then add $\boldsymbol{z}_{ij'}^T \boldsymbol{b}_i$ on the linear predictor scale, and get back to the response scale through the inverse link function $g^{-1}(\cdot)$. As we underlined before, random-effects coefficients $\boldsymbol{b}_i$ are known from the estimation process.

For the second type of prediction, we have no information to evaluate $\boldsymbol{b}_i$. A possible solution is to set it to its expected value of 0, yielding the value $\hat{f}(\boldsymbol{x}_{i'j'})$, and transform it back to the response scale through the inverse link function. As noted in Sela and Simonoff (2012), in this case we might expect that methods that do not incorporate random effects would have comparable performance to those that do, as long as the sample is large enough so that the fixed-effects function $f(\boldsymbol{x}_{ij'})$ is well-estimated by both types of methods.

2.2 Simulation study

In this section we compare the performance of the proposed GMET method to standard classification trees on different simulated binary outcomes datasets.

We first use a variation of a simulation design proposed in Hajjem et al. (2017). It has a two-level data structure of $I = 50$ groups with $n_i = 60$ observations each: 10 observations in each group are included in the training sample, and the other 50 observations constitute the test sample. Therefore, $N_{\text{train}} = 500$, while $N_{\text{test}} = 2500$. Setting $i = 1, ..., I$ and $j = 1, ..., n_i$, the response values y_{ij} are simulated according to a Bernoulli distribution with conditional probability of success μ_{ij}. Both fixed and random effects are used to generate μ_{ij}. Overall, we consider 10 different Data Generating Processes (DGP) outlined in Table 2.1 by combining different fixed- and random-effect specifications.

Let us define the fixed-effect structure. Eight random variables $X_1, ..., X_8$, independent and uniformly distributed in the interval $[0, 10]$, are generated. While

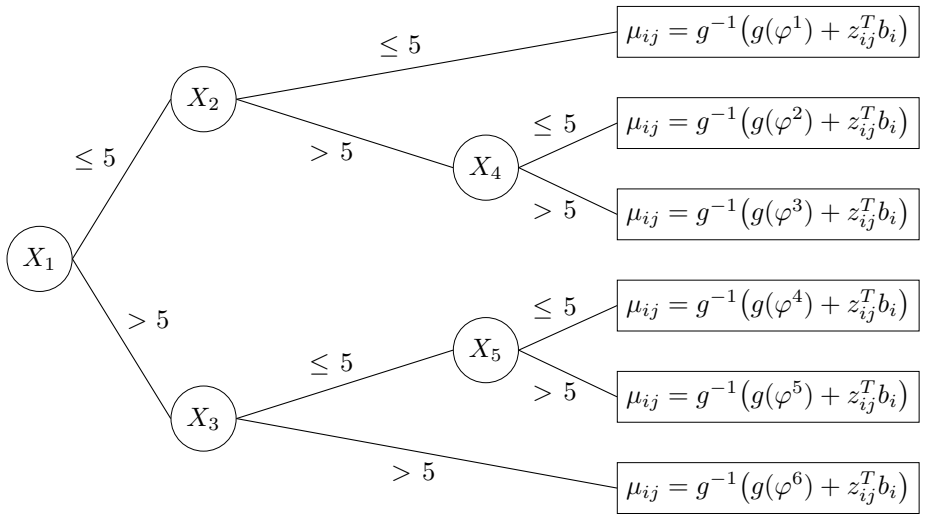


Figure 2.1: Mixed-effects tree structure used to generate the conditional probability of success μ_{ij} in the simulation study.

all of them are being used as predictors, only five of them are actually used to generate μ_{ij}, based on the tree rule summarized in Figure 2.1. Each observation is classified into one of the six terminal nodes according to the values $x_{ij1}, ..., x_{ij5}$. Within each leaf, values $\varphi^1, ..., \varphi^6$ denote the probabilities of success when the random effects b_i are equal to zero:

Leaf 1: if $x_{1ij} \leq 5 \wedge x_{2ij} \leq 5$ then $\mu_{ij} = g^{-1}\big(g(\varphi^1) + z_{ij}^T b_i\big)$;

Leaf 2: if $x_{1ij} \leq 5 \wedge x_{2ij} > 5 \wedge x_{4ij} \leq 5$ then $\mu_{ij} = g^{-1}\big(g(\varphi^2) + z_{ij}^T b_i\big)$;

Leaf 3: if $x_{1ij} \leq 5 \wedge x_{2ij} > 5 \wedge x_{4ij} > 5$ then $\mu_{ij} = g^{-1}\big(g(\varphi^3) + z_{ij}^T b_i\big)$;

Leaf 4: if $x_{1ij} > 5 \wedge x_{3ij} \leq 5 \wedge x_{5ij} \leq 5$ then $\mu_{ij} = g^{-1}\big(g(\varphi^4) + z_{ij}^T b_i\big)$;

Leaf 5: if $x_{1ij} > 5 \wedge x_{3ij} > 5 \wedge x_{5ij} > 5$ then $\mu_{ij} = g^{-1}\big(g(\varphi^5) + z_{ij}^T b_i\big)$;

Leaf 6: if $x_{1ij} > 5 \wedge x_{3ij} > 5$ then $\mu_{ij} = g^{-1}\big(g(\varphi^6) + z_{ij}^T b_i\big)$;

where $g(\cdot)$ is the logit link function. Two different possibilities are specified for the fixed effects: in the *large* fixed-effects specification, the standard deviation of the typical probabilities across the leaves is higher than in the *small* one (0.37 versus 0.24).

The random component $b_i \sim N(0, \Psi)$ is generated according to three different possibilities:

- No random effects: $\Psi = 0$;
- Random intercept: $z_{ij} = 1 \quad \forall i, \forall j$ and $\Psi = \psi_{11}$;

- Random intercept and slope, which add a linear random effect for the fixed-effect covariate X_1, uncorrelated from the random effect on the intercept. That is, $z_{ij} = [1 \quad x_{1ij}]^T \quad \forall i, \forall j$ and $\Psi = \begin{bmatrix} \psi_{11} & 0 \\ 0 & \psi_{22} \end{bmatrix}$.

Within each fixed effects scenario with random effects, we consider two specifications (*low* and *high*) for the covariance matrix Ψ to account for different levels of magnitude of the between-group variability.

DGP	Random component				Fixed component						
	Structure	**Effect**	ψ_{11}	ψ_{22}	**Effect**	φ^1	φ^2	φ^3	φ^4	φ^5	φ^6
1	No random	–	–	–	Large	0.10	0.20	0.80	0.20	0.80	0.90
2	effect	–			Small	0.20	0.40	0.70	0.30	0.60	0.80
3	Random Intercept	Low	4.00	–	Large	0.10	0.20	0.80	0.20	0.80	0.90
4		High	10.00	–							
5		Low	0.50	–	Small	0.20	0.40	0.70	0.30	0.60	0.80
6		High	4.00	–							
7	Random Intercept and Slope	Low	2.00	0.05	Large	0.10	0.20	0.80	0.20	0.80	0.90
8		High	5.00	0.25							
9		Low	0.25	0.01	Small	0.20	0.40	0.70	0.30	0.60	0.80
10		High	2.00	0.05							

Table 2.1: Data Generating Processes (DGP) for the simulation study.

Simulation results

We fit four different models for each one of the 10 DGPs: a standard binary classification tree model (*Std*), a random intercept GMET model (*RI*), a random intercept and slope GMET model (*RIS*), a parametric mixed-effects logistic regression model (*MElog*) that uses the true model leaves' indicators as fixed covariates. As noted in Hajjem et al. (2011) the MElog model could not be a real competitor of any other model. Indeed, it is not possible in practice to specify this parametric structure without knowing the underlying data generating process. This model only serves as a reference to compare the performance of the other models. In tree-based models, we fix to 10 the maximum depth parameter and to 20 the minimum number of observations necessary to attempt a split. After fitting each model on the training set, we can compute the corresponding predicted probability $\hat{\mu}_{ij}$ and the predicted class $\hat{y}_{ij}$ of observation j in group i in the test dataset. While the former is directly estimated by the algorithm, the latter depends on the threshold value μ_k^* used to classify subjects in the test set: $\hat{\mu}_{ij} \geq \mu_k^* \Rightarrow \hat{y}_{ij} = 1$ where $(i, j) \in$ test. There are at most K distinct fitted values μ_k, with $K \leq I|T|$. We use each of them to classify observations in the training

set and we fix the threshold μ_k^* as the one that yields the closest proportion of class 1 to the actual proportion of class 1 in the training set.

We measure the predictive performance by:

- the *predictive mean absolute deviation* (PMAD) of the estimated probability

$$\text{PMAD} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{I} \sum_{j=1}^{n_i^{\text{test}}} |\mu_{ij} - \hat{\mu}_{ij}|$$

- the *predictive misclassification rate* (PMCR)

$$\text{PMCR} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{I} \sum_{j=1}^{n_i^{\text{test}}} |y_{ij} - \hat{y}_{ij}|.$$

The mean, standard deviation, minimum and maximum of the PMAD and the PMCR over 50 runs were calculated and are reported in Table 2.2.

We observe that when there is no random effect (DGPs 1 and 2), the standard classification tree algorithm performs better, specifically when the fixed effect is large. However, when random effects are present (DGPs 3 to 10), the mixed effects classification tree performs better than the standard classification tree in terms of average PMAD. The highest improvement in PMAD using a mixed tree model is observed when both the fixed and the random effects are large (16.50% in DGP4 - *Std vs RI* and 16.78% in DGP8 - *Std vs RIS*). The lowest improvement is observed when both the fixed and the random effects are small (2.35% in DGP5 - *Std vs RI* and 2.34% in DGP9 - *Std vs RIS*). Analogous considerations can be made about PMCR. In addition, GMETs perform better than standard trees even when we fit a mixed tree whose random component is over-specified (like in DGPs 3-6, *Std vs RIS*) or under-specified (like in DGPs 7-10, *Std vs RI*) in relation to the true data generating process.

Next, we compare the performance of the GMET approach to the results of the MElog reference model. If the DGP does not include random effects, the difference in PMAD and PMCR is higher when the fixed effects are large (DGP1). When random effects are large and fixed effects are small (DGPs 6 and 10), the GMET model performs closer to the MElog model. In terms of PMAD, this difference equals to 4.75% and 4.41% in DGPs 6 and 10 respectively; in terms of PMCR it equals to 3.38% and 2.95% respectively. The difference in predictive accuracy between the two models reaches the maximum when random effects are small and fixed effects are large (DGPs 3 and 7). In terms of PMAD, this difference equals to 9.69% and 9.28% in DGPs 3 and 7 respectively; in terms of PMCR it equals to 7.46% and 7.92% respectively.

DGP	Random effect	Fixed effect	Fitted model	PMAD (%)				PMCR (%)			
				mean	sd	min	max	mean	sd	min	max
1	NO RANDOM EFFECT	Large	Std	5.35	1.53	2.71	8.79	17.20	1.40	14.64	20.52
			RI	20.22	2.31	15.15	24.72	31.09	2.63	26.04	37.68
			RIS	20.36	2.36	13.14	24.88	31.03	2.40	24.24	35.48
			MElog	3.11	0.88	1.42	4.95	17.79	3.18	14.52	24.24
2		Small	Std	12.93	2.78	7.01	19.28	33.16	2.18	28.92	38.60
			RI	13.99	1.78	9.84	17.19	37.57	1.88	32.72	41.64
			RIS	14.08	1.82	9.93	17.81	37.33	1.79	33.16	41.56
			MElog	4.16	1.30	1.02	6.45	29.32	1.63	26.96	33.16
3	Low	Large	Std	23.83	2.94	17.53	29.88	30.53	3.13	23.32	38.20
			RI	**18.28**	1.47	15.07	22.67	**26.80**	1.86	22.84	31.92
			RIS	18.43	1.31	15.28	21.89	26.84	1.73	22.72	30.76
			MElog	**8.59**	0.87	6.02	10.56	**19.34**	1.29	16.08	22.48
4	High		Std	32.05	2.37	26.90	37.59	37.80	2.65	32.08	44.96
			RI	15.55	1.28	12.49	18.71	21.62	1.88	16.32	26.56
			RIS	15.66	1.27	12.52	18.91	21.71	1.87	16.56	26.40
			MElog	8.09	0.76	6.04	10.06	16.32	1.53	13.32	19.80
5	Low	Small	Std	17.89	2.32	13.28	22.48	35.30	2.23	31.40	41.40
			RI	15.54	1.58	12.52	19.12	35.89	2.18	30.76	41.20
			RIS	15.76	1.56	12.76	19.63	36.12	2.14	31.20	41.32
			MElog	8.63	0.92	6.49	10.53	28.90	0.95	27.20	31.84
6	High		Std	29.47	2.22	24.56	35.08	41.42	2.36	36.36	45.48
			RI	**14.11**	1.46	10.17	17.38	**26.23**	2.35	21.40	30.96
			RIS	14.25	1.49	10.39	17.81	26.27	2.40	21.28	31.20
			MElog	**9.36**	0.98	7.07	11.25	**22.85**	1.70	19.12	26.08
7	Low	Large	Std	23.24	2.49	18.54	29.68	29.61	2.91	23.44	38.44
			RI	19.59	1.37	15.42	22.51	27.89	1.98	22.16	31.20
			RIS	**19.29**	1.40	15.15	22.22	**27.84**	1.82	22.08	31.08
			MElog	**10.01**	1.02	8.07	11.91	**19.92**	1.37	17.20	24.04
8	High		Std	32.89	2.61	27.47	38.04	38.69	3.67	31.64	46.32
			RI	17.52	1.57	14.29	20.85	22.03	2.04	17.48	26.08
			RIS	16.11	1.41	12.90	18.93	21.26	1.92	17.04	25.48
			MElog	9.86	1.02	7.82	13.16	16.59	1.48	13.20	20.36
9	INTERCEPT & SLOPE Low	Small	Std	18.15	2.25	13.36	24.73	35.34	2.56	31.36	42.64
			RI	15.84	1.17	12.37	18.61	35.83	1.92	30.84	40.48
			RIS	15.81	1.24	12.41	19.05	35.76	1.92	31.28	39.80
			MElog	9.31	0.86	7.95	11.06	29.11	0.94	26.76	30.96
10	High		Std	29.09	2.06	24.21	33.51	41.64	2.45	37.16	49.76
			RI	15.88	1.26	13.60	19.77	27.66	1.97	23.00	32.76
			RIS	**15.21**	1.15	13.20	18.32	**27.20**	1.93	21.96	31.64
			MElog	**10.80**	1.02	9.20	13.06	**24.25**	1.69	20.32	28.04

Table 2.2: Results of the 50 simulation runs in terms of predictive probability mean absolute deviation (PMAD) and predictive misclassification rate (PMCR). In bold, DGPs in which the performance gap between MElog and GMET is the largest or the smallest are marked.

2.3 Case study: application of mixed-effects tree algorithm to education PoliMi data

In this section, we describe the PoliMi dataset and we apply the generalized mixed-effects tree algorithm to these data. Using a GMET model, we can identify discriminating fixed-effects covariates and estimate the degree programme effect on the predicted success probability. In addition, we also analyse the accuracy of this model in predicting dropout careers.

The PoliMi dataset consists of 18,612 careers in Bachelor of Science (BSc) that began between A.Y. 2010/2011 and 2013/2014. Students are nested within $I = 19$ degree programmes.[2] A descriptive analysis shows that a high percentage of students leaves the Politecnico before obtaining the degree. Therefore, our goal is to find out which student-level indicators could discriminate between two different profiles: *dropout* and *graduate* students.

We assume the binary GMET model (2.3) where student j is nested within degree programme i. The response variable Y is the career `status`, a two-level factor we code as a binary variable:

- `status` $= 1$ for careers definitely completed with graduation;
- `status` $= 0$ for careers definitely concluded with a dropout.

We would like to make predictions at the very early stage of the academic career. So, we choose as predictors five variables available at the time of enrollment and three more variables collected just after the first semester of studies. The list and explanation of student-level variables to be included as covariates is reported in Table 2.3. In addition we choose as grouping variable the degree programme at the time of the enrollment (factor `DegreeProgramme`) which has 19 levels. The influence of the grouping factor on the predictor is modeled through a group-level intercept b_i. We randomly split the dataset into training and test subsets, with a ratio of 80% for training and 20% for evaluation. Thus, the training subset equals to 14,890 careers while the test subset amounts to 3,722 careers.

While growing the tree, we fix to 10 the maximum depth parameter and to 20 the minimum number of observations necessary to attempt a split. Figure 2.2 shows the estimated mixed-effects tree for the graduating probability. Every internal node has its corresponding condition that splits it into two sons: if the condition is true, observations are sent down the tree through the left son, while through the right son if the condition is false. In addition, all nodes report two values: the estimated graduating probability and the percentage of observations in the node over the total training set. We remind that variable

[2]We are considering the following Engineering programmes: Aerospace, Automation, Biomedical, Building, Chemical, Civil, Civil and Environmental, Electrical, Electronic, Energy, Computing Systems, Environmental and Land Planning, Industrial Production, Management, Materials and Nanotechnology, Mathematical Mechanical, Physics, Telecommunications.

Variable	Description	Type of variable
`Sex`	gender	factor (2 levels: M, F)
`Nationality`	nationality	factor (Italian, foreigner)
`PreviousStudies`	high school studies	factor (*Liceo Scientifico*, *Istituto Tecnico*, Other)
`AdmissionScore`	PoliMi admission test result	real number
`AccessToStudiesAge`	age at the beginning of the BSc studies at PoliMi	natural number
`WeightedAvgEval1.1`	weighted average of the evaluations during the first semester of the first year	real number
`AvgAttempts1.1`	average number of attempts to be evaluated on subjects during the first semester of the first year (passed and failed exams)	real number
`TotalCredits1.1`	number of ECTS credits obtained by the student during the first semester of the first year	natural number

Table 2.3: List and explanation of variables at student level to be included as covariates in the GMET model.

`PreviousStudies` has been coded as a three-level factor with levels S (*Liceo Scientifico*), T (*Istituto Tecnico*) and O (other high school studies). The number of ECTS obtained in the first semester of the first year is used as first split: students who obtained less than 13 ECTS are associated to lower success probability (0.16 versus 0.86). Then, students are further classified using other explainatory variables: we can notice that Italian students who obtained more than 24 ECTS have the highest predicted success probability (0.95). Other variables actually used to split smaller internal nodes are `Nationality` and `PreviousStudies`: in these nodes, students who attended *Istituto Tecnico* and foreign students have lower predicted success probability than the others. Through this model, it is possible to point out significant interactions among the covariates: for example, variable `Nationality` is used to split the group of students that obtained at least 13 ECTS, while this same variable does not appear in the complementary branch of the tree. Finally, covariates `Sex`, `AdmissionScore` and `AvgAttempts1.1` do not compare in the trees, so they do not appear to have strong influence on how a career ends.

Using the tree structure in Figure 2.2, we can get a population-level prediction for new observations that do not include the effect of the programme. However, if we also specify the level of the random effect covariate, our model is able to adjust this prediction to account for this effect and make a group-

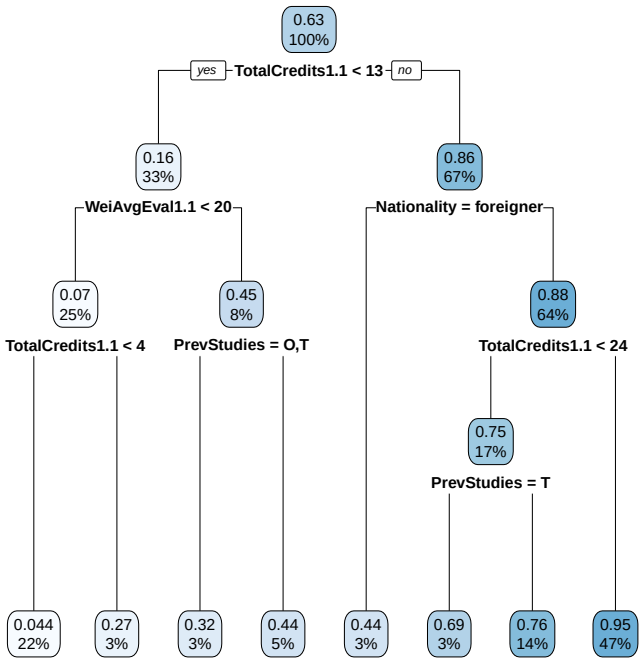


Figure 2.2: Estimated mixed-effects tree of model (2.3) for the graduating probability.

specific prediction. Indeed, we can extract coefficients $\hat{b}_i$ from the full estimated mixed model (2.3) and provide different predictions for different programmes within each leaf of the tree structure. Figure 2.3 shows the estimated random effects for all 19 groups in the dataset. The coefficients b_i are rearranged by their point estimate. In many groups, the 95% confidence interval does not overlap the vertical line at zero, underlining substantial differences between the groups. If we use this model to estimate the graduating probability, in many of the groups it is significantly different from the average one. After fixing all other covariates, levels *Environmental and Land Planning Engineering* and *Civil and Environmental Engineering* have higher positive effect on the intercept: being a student from one of these programmes improves the log odds by 1.051 and 0.705 respectively. On the contrary, studying either *Civil Engineering* or *Electrical Engineering* penalizes the log odds by 0.680 and 0.546 respectively.

Since we are using a multilevel model we can account for the interdependence of observations by partitioning the total variance into different components due to the clustered data structure in model (2.3). The *Variance Partition Coefficient* (VPC) is a possible measure of intraclass correlation: it is equal to the percentage of variation that is found at the higher level of hierarchy over the total variance (Goldstein et al., 2002). The idea of VPC was extended using the latent variable approach, to define a method to partition the total variance in the case of a binary response and group-specific intercept as random effects

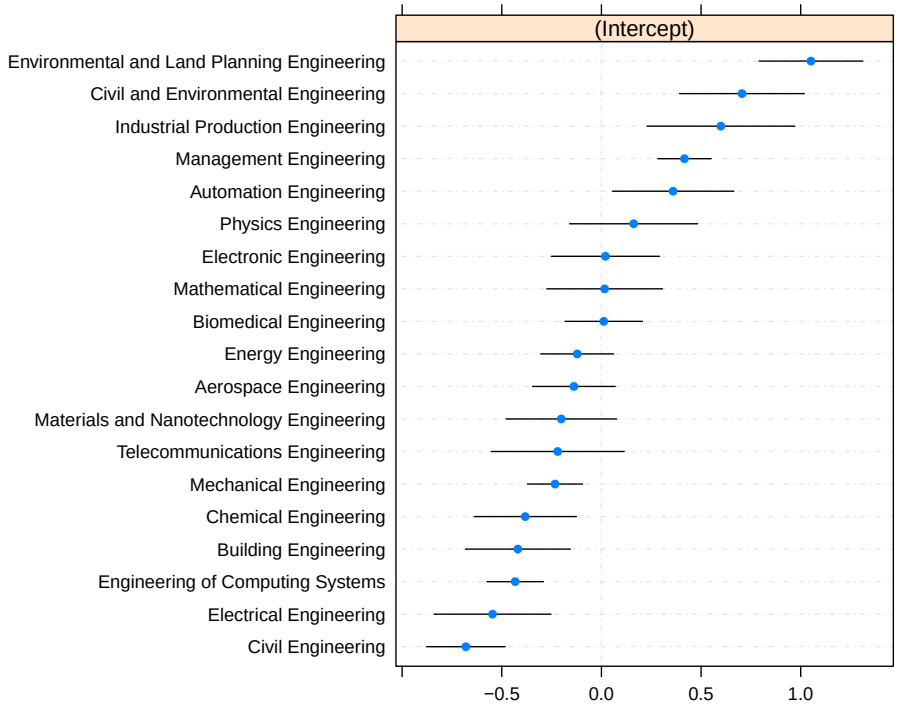


Figure 2.3: Estimated random intercept for each degree programme in model (2.3). For each Engineering programme, the blue dot and the horizontal line marks the estimate and the 95% confidence interval of the corresponding random intercept.

structure (Browne et al., 2005). In this case, the Variance Partition Coefficient is constant across all individuals and it can be estimated as:

$$\text{VPC} = \frac{\hat{\sigma}_b^2}{\hat{\sigma}_b^2 + \sigma_{lat}^2} = \frac{0.2988}{0.2988 + \pi^2/3} = 0.0612$$

where $\hat{\sigma}_b^2$ is the estimated variance of the random intercept and σ_{lat}^2 is the residual variability that can neither be explained by fixed effects, nor through the group features that are represented by the random intercept. In this case, it is equal to the variance of the standard logistic distribution. This VPC value means that 6.12% of variation in the response is attributed to the classification by degree type. This value underlines the need to use a mixed model.

We can now evaluate the performance of the model and its predictive quality using the test data. For each test observation, we are given a full set of covariates: therefore, we are able to compute an estimate $\hat{p}$ of the probability of successfully concluding the BSc and getting the degree. We use this estimate to define a binary classifier based on model (2.3): we choose $p_0 = 0.6$ as optimal cutoff value through ROC curve analysis. For 20 iterations, we randomly split the observations in training and test set, we fit a GMET model on the training set and we classify test observations using the optimal threshold value. At

the end, we compute the average accuracy, sensitivity and specificity and their standard deviation, reported in Table 2.4. High values of accuracy, sensitivity and specificity point to a good effectiveness of the model. In addition, the model performance is robust, as highlighted by the low standard deviation of mean performance indexes.

It is interesting to compare these average performance indexes against those obtained using different methods. This approach has similar accuracy to a standard classification tree (0.878 versus 0.879), but its accuracy shows less variability across the iterations. For example, its standard deviation of accuracy is 0.5% against 2.8% for a classification tree. Since we are interested in the detection of dropout careers, we should compare mean sensitivity using different models. Using mixed-effects trees, we get higher sensitivity than using standard classification trees (0.835 versus 0.800). Thus, the choice of a mixed-effects model seems appropriate: the degree programme is a meaningful covariate for the prediction of career `status`. A mixed-effects tree is slightly less sensitive than a classifier build through a GLMM (0.835 versus 0.850), suggesting that a tree-like structure for fixed effects might not be as suitable as the GLMM one. However, it has other advantages like offering an easily interpretable model that could be graphically displayed and understood.

Index	**Mean**	**Std deviation**
Accuracy	0.860	0.006
Sensitivity	0.816	0.012
Specificity	0.886	0.008

Table 2.4: Performance indexes of a classifier based on the mixed-effects tree of model (2.3).

2.4 Conclusions

This chapter proposes a multilevel tree-based model for a non-gaussian response (GMET algorithm), shows a simulation study and applies the GMET algorithm to the PoliMi careers dataset as a tool to find discriminating student-level variables between two different student profiles (graduate and dropout) and to estimate the degree programme effect on the predicted success probability.

The GMET model can deal with a grouped data structure, while providing easily interpretable models that can outline complex interactions among the input variables. In the simulation study, the performance of the proposed mixed-effects tree method is a marked improvement over the CART model when the data generating process (DGP) includes random effects, even if of small magni-

tude. In addition, the performance of the GMET model is closer to the one of the benchmark logistic model that is fitted assuming the whole specification of the DGP. Although our study focuses on the binary response case, the mixed-effects tree approach could be extended to other types of response variables. Using a suitable link function, we could study if the method is appropriate to model different outcomes such as counts data or a multinomial factor response. Moreover, ensemble methods which use a mixed-effects tree as base learner may be developped.

In our case study, the effectiveness of the GMET model in dropout prediction is comparable to the ones of more established classification methods. A GMET model with high accuracy and sensitivity has been obtained by considering information available at the time of the admission and the career of the first semester of studies. In addition, our study identifies a significant effect of the Engineering programme on dropout probability.

In the context of the SPEET project, a future development could be the extension of our analysis to the other project partners in order to compare the programme effect at country level. This would allow us to relate this effect to programme-level variables and we could establish if the same profiles of students with dropout risk arise at country level. Moreover, in accordance to the validity and the potential of GMET method when applied to model student dropout prediction, our future perspective goes in the direction of major applications in the Learning Analytics area. This method, when applied to educational data, can be a useful tool to support the definition of best practices and new tutoring programmes aimed at enhancing student performances and reducing student dropout. A worthwhile aspect regards also the approach that teachers and students have with respect to its results. Indeed, this method is also valuable in the perspective of recommendation systems, since, if its results are interpreted and communicated in the right way, they can be used to drive students in their career choices.

Chapter 3

Semi-parametric mixed-effects models for the clustering of Italian schools

The nature and the magnitude of the school impact on student attainments strongly depend on the type of school system and related regulations. There are countries where the education system is totally centralized and, therefore, school programs and practices are very homogeneous across the territory. On the other hand, in the last years the dynamics of education systems are changing and more and more countries are decentralizing the power on decision about education, giving more autonomy to schools (Sarrico et al., 2012). This phenomenum leads to differences across schools that are reflected on differences across student achievements. Studies on PISA data show that Italy is a country where the percentage of variability in student achievements due to the grouping factor (i.e. schools) is quite high with respect to other countries (Masci et al., 2018b). This means that in Italy the value-added, seen as the positive or negative impact, that schools give to their students is relevant: in other words, attending a certain school instead of another might lead to different results in student's skills. Schools differ under many aspects: size, location, school body composition, teachers, school principal management style and much more. All these aspects contribute to the student learning process, creating heterogeneity within their achievements.

Focusing on the Italian context, many studies confirm that the magnitude of the school effect, intended as the positive or negative value-added of the school, on student attainments is substantial. In Agasisti et al. (2017b); Masci et al. (2016b, 2017a) the authors observe that the percentage of variability in student attainments in INVALSI tests explained by the random effect depends on the geographical macro-area and differs between mathematics and reading performances. In particular, this percentage is higher in mathematics and especially in

Southern Italy, reaching peaks of 20%. Moreover, results of PISA data in Italy report that, in mathematics, the Percentage of Variability explained by Random Effects - PVRE- exceeds the 40% (Masci et al., 2018b).

The aim of this study is to identify latent subpopulations of Italian schools that differ in the evolution of their student attainments across different years. The goal is to reduce the set of numerous Italian schools into a series of subpopulations, each of which contains schools with a similar impact on student achievements and across which these impacts differ. To this aim, we need a model that takes into consideration the hierarchical structure of data, but that also identifies a latent structure among the higher level of hierarchy. Therefore, we apply a multilevel model in which we model the subpopulations by choosing as random effects a discrete distribution P^* with an unknown finite number of mass points, that is able to detect a latent structure among the Italian schools, the higher level of hierarchy. This model can be interpreted as an in-built unsupervised classification tool, since it identifies a clustering structure among groups, without knowing a priori nor the clusters of belonging neither their size. From a practical point of view, in Italy students must attend five years of primary school, three years of junior secondary school and five years of upper secondary school. We are aware of the challenges that estimating the pure school effect implies (Goldstein and Spiegelhalter, 1996; Raudenbush and Willms, 1995), indeed, we will not refer to "school effect" in the classic way. Rather, since we focus on junior secondary schools, our "school effect" can be interpreted as the ability of these schools in receiving students from the primary schools with certain skills and give them new and possibly increased skills at the end of the three years, aware of the fact that students might not be randomly assigned to schools. So our research mainly aims at identifying subpopulations of schools, standing on the relationship between their students test scores at the beginning and at the end of the three years (grades 6 and 8 respectively). Supposing that we can model the relationship between students test scores at different grades by means of linear models, which means that student scores at different grades are assumed to be linearly correlated, the regression line between the two grades test scores might be characterized by different parameters across schools. In other words, we try to identify subpopulations of Italian junior secondary schools, characterized by different trends in their student achievements, where the number of subpopulations is unknown a priori.

In the methodological literature, two lines of research about the identification of subpopulations are (a) Growth Mixture Models (GMM) (Muthén, 2004; Muthén and Shedden, 1999) and (b) Latent Class Mixture Models (McCulloch et al., 2002; Nagin, 1999; Vermunt and Magidson, 2002; Asparouhov and Muthen, 2008; Vermunt, 2011). Conventional growth modelling is applied to longitudinal data and it is used to estimate the average growth, the amount of variation across individuals in growth intercept and slopes and the influence of covariates

on this variation. It can be described as a random effect model where intercept and slope vary across individuals. However, conventional growth models assume that individuals come from a single population and that a single growth trajectory can approximate the entire population. Growth mixture models relax this assumption and assume that there are differences in growth parameters across unobserved subpopulations. They allow for the existence of latent trajectory classes where different groups of individual growth trajectories vary around different behaviors. In other words, the average association between covariates and the outcome varies across latent classes and also, within classes, individuals also vary randomly in their coefficients. The results are separate growth models for each latent class. Latent Class Growth Analysis (LCGA) is a special case of GMM where the variance and covariance estimates for the growth factors are assumed to be fixed at zero, assuming that all the individuals within a latent class are homogeneous. Individuals within a latent classes are assumed to have identical random effects. Conceptually, these methods are very similar to the one that we propose, especially the special case of LCGA, since we also assume that individuals within latent classes have identical random effects. Nonetheless, there are two main differences between our approach and the one of GMM and LCGA. The former is that GMM/LCGA are thought for modelling longitudinal changes and not regression[1]. The latter is that GMM/LCGA need to fix a priori the number of latent classes, while our approach estimates it together with the other unknown parameters. There are numerous extensions and applications of GMM (Lin et al., 2000; Proust-Lima et al., 2007), but none of them includes the estimation of the number of latent classes. They rather estimate the parameters fixing different number of masses and they choose the best one comparing goodness of fit indices. Latent class mixture models are even more related to our approach since they consider linear mixed models where the assumption of normality of random effects is relaxed. They also assume a discrete distribution for the random effect coefficients and they are used to uncover distinct subpopulations (latent classes) and classify individuals. But also this approach requires a fixed number of latent classes, chosen a priori. In the framework of latent structure analysis, an other branch of research related to ours is the one about Latent Trait Analysis (LTA) (Bock and Aitkin, 1981; Heinen, 1996). LTA, also called Item Response Theory (IRT), is used for the analysis of categorical data. It performs the reduction of a set of binary or ordered-category variables into a smaller set of factors and it is mainly used for data exploration or theory confirmation. The common aspect of this method with the ones described above and, at the same time, the main difference with our method is, again, the fact that they need to fix a priori the number of latent factors. The choice of the

[1] One of the characteristics of the models for longitudinal data is that the set of time instants in which the dependent variable is evaluated is the same within each group/individual, meaning that the covariate is fixed across the groups/individuals.

number of latent classes (mass points) is not trivial when the sample is very big or the knowledge about possible different trends across the individuals (groups) is limited. Our case study represents a clear example of a sample composed by hundreds of groups, within which we do not know how many different subpopulations exist. For this reason, in the perspective of performing dimensionality reduction without any assumption about the final dimension, we need to develop an approach that estimates, together with the other parameters, also the number of existing subpopulations. In this sense, our approach brings a significant improvement with respect to the existing literature.

In particular, we develop and apply an EM algorithm for semi-parametric mixed-effects models (Bock and Aitkin, 1981), for hierarchical data (students nested within schools), in order to perform an in-built classifier of the grouping factor (schools). The algorithm is inspired by the ones proposed in Aitkin (1996) and Azzimonti et al. (2013), but with substantial changes. The idea is that we perform a linear two-level model, in which we consider students nested within schools, where the random effect (school effect) is semi-parametric since it follows a discrete distribution with an unknown number of support points. The algorithm itself identifies the number of support points, that is the number of subpopulations in which schools are grouped, standing on the achievements trend of their students. In the educational literature, multilevel linear models have already been applied to INVALSI data, with a view to estimating school value-added, modeled by means of parametric distributions, after adjusting for student characteristics (Agasisti et al., 2017b; Masci et al., 2016b, 2017a; Sani and Grilli, 2011). Nonetheless, our method has a different scope since it does not seek to estimate individual value-added for each school, but it looks for subpopulations of schools with homogeneous value-added. Both the algorithm and its application to the educational context are new to the literature.

From an interpretative point of view, the consequence of the identification of subpopulations of schools is that we can recognize how many and which different behaviors characterize Italian schools and, therefore, identify a latent structure within them. In particular, the distribution of schools across subpopulations reveals which is the most common trend (the most numerous subpopulation) and identifies subpopulations of anomalous schools, that are those subpopulations containing less schools with different impact on student achievements. In a second stage, this enables the profiling of subpopulations by means of school level variables. The idea is that there could be variables at school level that influence the different student achievements trends across schools. Therefore, in the second part of the analysis we explore the presence of patterns of school characteristics among subpopulations of schools by means of multinomial regression models.

This chapter is organized as follows: in Section 3.1 we describe the model and methods - SPEM algorithm - and we present a simulation study; in Section 3.2 we

describe the INVALSI dataset and report the application of SPEM algorithm to INVALSI data, show the results and explore the relation between subpopulations and school characteristics; in Section 3.3 we draw our conclusions.

The code for SPEM algorithm is available upon request to the authors.

3.1 Model, methods and simulation study

In this section, we present the semi-parametric mixed-effects model (Subsection 3.1.1), the EM algorithm for the estimation of its parameters (Subsection 3.1.2) and a simulation study (Subsection 3.1.3). Since we learn from previous learning analytics on Italian data that there exist patterns of student achievements across different Italian schools (Agasisti et al., 2017b; Masci et al., 2016b, 2017a), we are interested in evaluating how the association between previous and current student test scores does change across different Italian schools and, in particular, in identifying subpopulations of schools within which this association is identical. Therefore, the model that we develop is a two-level linear model (in the application, students represent level 1 and schools represent level 2) with a discrete distribution with a finite number of support points on the random effects. This modelling allows to identify a latent structure of subpopulations in the higher level of grouping (in the application, schools).

3.1.1 Semi-parametric mixed-effects model

We start considering a general mixed-effects (two-level) linear model, where each observation j, for $j = 1, \ldots, n_i$, is nested within a group i, for $i = 1, \ldots, N$. The model takes the following form:

$$\begin{aligned} \mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \qquad & i = 1, \ldots, N \\ \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{1}_{n_i}) \qquad & ind. \end{aligned} \tag{3.1}$$

where i is the group index, N is the total number of groups, n_i is the number of observations within the i-th group and $\sum_{i=1}^{N} n_i = J$. $\mathbf{y}_i = (y_{1i}, \ldots, y_{n_i i})$ is the n_i-dimensional vector of response variable within the $i-$th group, $\mathbf{X}_i$ is the $n_i \times (p+1)$ matrix of covariates having fixed effects, $\boldsymbol{\beta}$ is the (p+1)-dimensional vector of fixed coefficients, $\mathbf{Z}_i$ is the $n_i \times (r+1)$ matrix of covariates having random effects, $\mathbf{b}$ is the (r+1)-dimensional vector of random coefficients and $\boldsymbol{\epsilon}_i$ is the vector of errors. Fixed effects are identified by parameters associated to the entire population, while random ones are identified by group-specific parameters.

In the parametric framework of mixed-effects linear models, random coefficients are assumed to be distributed according to a Normal distribution with unknown parameters that, together with the coefficients of fixed effects and σ^2, can be estimated through methods based on the maximization of the likelihood or the restricted likelihood functions (Pinheiro and Bates, 2000).

The main novelty introduced here is that we move to a semi-parametric framework, assuming the coefficients $\mathbf{b}_i$ to be distributed according to a discrete distribution P^*, assuming M sets of values $(c_{0l}, \ldots, c_{rl})$ for $l = 1, \ldots, M$, where $M \leq N$. This means that each group i, for $i = 1, \ldots, N$, is assigned to a subpopulation l, that is characterized by random parameters $(c_{0l}, \ldots, c_{rl})$. This semi-parametric modelling enables to identify a latent structure among the groups, that are clustered by the model into an unknown number of discrete masses. Therefore, the two main advantages are that, first of all, we can identify how many latent subpopulations exist within the groups of data and, second, we can estimate the parameters associated to each subpopulation, pointing out their differences.

Under these assumptions, the semi-parametric mixed-effects model takes the following form:

$$\begin{aligned} \mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{c}_l + \boldsymbol{\epsilon}_i \qquad i = 1, \ldots, N \qquad l = 1, \ldots, M \\ \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{1}_{n_i}) \qquad ind. \end{aligned} \tag{3.2}$$

In particular, from now on, without loss of generality, we consider the case with one random intercept, one random effect and one fixed effect[2]:

$$\begin{aligned} \mathbf{y}_i = \mathbf{x}_i\beta + \mathbf{1}c_{0l} + \mathbf{z}_i c_{1l} + \boldsymbol{\epsilon}_i \qquad i = 1, \ldots, N \qquad l = 1, \ldots, M \\ \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{1}_{n_i}) \qquad ind. \end{aligned} \tag{3.3}$$

where $\mathbf{1}$ is the n_i-dimensional vector of 1, $M \leq N$ is the number of subpopulations (mass points) unknown a priori. Coefficients $\mathbf{c}_l$, for $l = 1, \ldots, M$, are distributed according to a probability measure $\mathcal{P}^*$ that belongs to the class of all probability measures on $\mathbb{R}^2$. $\mathcal{P}^*$ is a discrete measure with M support points that can then be interpreted as the mixing distribution that generates the density of the stochastic model in (3.3). The ML estimator $\hat{\mathcal{P}}^*$ of $\mathcal{P}^*$ can be obtained following the theory of mixture likelihoods in Lindsay et al. (1983a,b), where the author proves the existence, discreteness and uniqueness of the semi-parametric maximum likelihood estimator of a mixing distribution, in the case of exponential family densities. In particular, the author faces statistical problems (existence, discreteness, support size characterization and uniqueness) transforming them in geometrical problems, concerning support hyperplanes of the convex hull of the likelihood curve. So, the ML estimator of the random effects distribution can be expressed as a set of points $(\mathbf{c}_1, \ldots, \mathbf{c}_M)$, where $M \leq N$ and $\mathbf{c}_l \in \mathbb{R}^2$ for $l = 1, \ldots, M$, and a set of weights $(w_1, \ldots, w_M)$, where $\sum_{l=1}^{M} w_l = 1$ and $w_l \geq 0$ for each $l = 1, \ldots, M$. Given this, we propose an algorithm for the joint estimation of σ^2, β, $(\mathbf{c}_1, \ldots, \mathbf{c}_M)$ and $(w_1, \ldots, w_M)$, that is performed through

[2]This choice is due to the case considered in the application to INVALSI dataset, in Section 3.

the maximization of the likelihood, mixture by the discrete distribution of the random effects,

$$L(\boldsymbol{\beta}, \sigma^2|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = \sum_{l=1}^{M} \frac{w_l}{(2\pi\sigma^2)^{\frac{J}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\sum_{j=1}^{n_i}(y_{ij} - \beta x_{ij} - c_{0l} - c_{1l}z_{ij})^2\right\}, \quad (3.4)$$

with respect to the fixed coefficient β, the error variance σ^2 and the random effects distribution $(\mathbf{c}_l, w_l)$, for $l = 1, \ldots, M$. For each $l = 1, \ldots, M$, $\mathbf{c}_l$ represents the group-specific parameters and w_l the corresponding weight in the mixture equation (3.3).

The algorithm that we propose is inspired by the one proposed in Azzimonti et al. (2013), but it considers the linear functional dependence between response and predictors and it makes three main improvements: (i) the optimization of the Maximization step is computed in closed form, (ii) the covariates can be group specific[3] and (iii) the initialization of the parameters is done in a more efficient and flexible way. The first point directly derives from the linearity assumption. The idea at the base of the algorithm is also similar to the one proposed in Aitkin (1996), but while in Aitkin (1996) the authors need to fix a priori the number of discrete points of the mixing distribution, our algorithm identifies itself the number of support points M, standing on given tolerance values that we fix depending on the problem.

3.1.2 The SPEM algorithm

The proposed EM algorithm is an iterative algorithm that alternates two steps: the expectation step (E step) in which we compute the conditional expectation of the likelihood function with respect to the random effects, given the observations and the parameters computed in the previous iteration; and the maximization step (M step) in which we maximize the conditional expectation of the likelihood function. The observations are the values of the answer variable y_{ij} and of the covariates z_{ij} and x_{ij}, for $j = 1, \ldots, n_i$ and i=1,...,N. The parameters to be estimated are the random coefficients $\mathbf{c}_l$ with their weights w_l, for l=1,...,M, the fixed coefficient β and the variance σ^2. The algorithm allows the number n_i, for $i = 1, \ldots, N$, of observations to be different across groups, but, within each group missing data are not handled, i.e. missing values of y, z and x for the n_i units are not allowed. At each iteration, the EM algorithm updates the parameters in order to increase the likelihood in (3.4) and it continues until the convergence or until a fixed number of iterations (`it`) is reached. In particular, the update is given by:

[3]With the term "group specific covariates" we mean individual level covariates that are allowed to vary in terms of number of observations and assumed values across the groups.

$$w_l^{(up)} = \frac{1}{N}\sum_{i=1}^{N} W_{il} \qquad \text{for } l = 1, \ldots, M \tag{3.5}$$

$$(\beta^{(up)}, \mathbf{c}_1^{(up)}, \ldots, \mathbf{c}_M^{(up)}, \sigma^{2(up)}) = \arg\max_{\beta, \mathbf{c}_l, \sigma^2} \sum_{l=1}^{M}\sum_{i=1}^{N} W_{il}\, \ln p(\mathbf{y}_i|\beta, \sigma^2, \mathbf{c}_l) \tag{3.6}$$

where

$$W_{il} = \frac{w_l\, p(\mathbf{y}_i|\beta, \sigma^2, \mathbf{c}_l)}{\sum_{k=1}^{M} w_k\, p(\mathbf{y}_i|\beta, \sigma^2, \mathbf{c}_k)} \tag{3.7}$$

and

$$p(\mathbf{y}_i|\beta, \sigma^2, \mathbf{c}) = \frac{1}{(2\pi\sigma^2)^{\frac{n_i}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}(y_{ij} - \beta x_{ij} - c_{0l} - c_{1l} z_{ij})^2\right\}. \tag{3.8}$$

The weight $w_l^{(up)}$ is the mean over the N groups of their weights related to the $l-$th subpopulation. Coefficients W_{il} represent the probability of $\mathbf{b}_i$ being equal to $\mathbf{c}_l$ conditionally to observations $\mathbf{y}_i$ and given the fixed coefficient β and the variance σ^2.

The maximization (M step) in equation (3.6) involves two steps and it is done iteratively. In the first step, we compute the *arg-max* with respect to the support points $\mathbf{c}_l$, keeping β and σ^2 fixed to the last computed values. In this way, we can maximize the expected log-likelihood (computed in the E step) with respect to all support points $\mathbf{c}_l$ separately, that means

$$\mathbf{c}_l^{(up)} = \arg\max_{\mathbf{c}} \sum_{i=1}^{N} W_{il}\, \ln p(\mathbf{y}_i|\beta, \sigma^2, \mathbf{c}) \qquad l = 1, \ldots, M. \tag{3.9}$$

Since we are considering the linear case, it is possible to perform this maximization step in closed-form. With regard to model (3.3), the estimates of the random effects are obtained by means of the weighted least squares method and are the following:

$$\hat{c}_{0l} = \frac{\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i}(y_{ij} - \hat{\beta} x_{ij} - \hat{c}_{1l} z_{ij})}{n_i \sum_{i=1}^{N} w_{il}} \tag{3.10}$$

and

$$\hat{c}_{1l} = \frac{\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} y_{ij} z_{ij} - \frac{(\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} y_{ij})(\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} z_{ij})}{n_i \sum_{i=1}^{N} w_{il}}}{\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} z_{ij}^2 - \frac{(\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} z_{ij})^2}{n_i \sum_{i=1}^{N} w_{il}}} + \frac{\frac{\hat{\beta}(\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} z_{ij})(\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} x_{ij})}{n_i \sum_{i=1}^{N} w_{il}} - \hat{\beta} \sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} x_{ij} z_{ij}}{\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} z_{ij}^2 - \frac{(\sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} z_{ij})^2}{n_i \sum_{i=1}^{N} w_{il}}}. \tag{3.11}$$

In the second step, we fix the support points of the random effects distribution computed in the previous step and we compute the *arg-max* in equation (3.6) with respect to β and σ^2. Again, this step can be done in closed-form and the estimates of the parameters, with regard to model (3.3), obtained by means of the weighted least squares method, are:

$$\hat{\beta} = \frac{\sum_{l=1}^{M} \sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} (y_{ij} x_{ij} - \hat{c}_{0l} x_{ij} - \hat{c}_{1l} z_{ij} x_{ij})}{\sum_{l=1}^{M} \sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} x_{ij}^2} \tag{3.12}$$

and

$$\hat{\sigma}^2 = \frac{\sum_{l=1}^{M} \sum_{i=1}^{N} w_{il} \sum_{j=1}^{n_i} (y_{ij} - \hat{\beta} x_{ij} - \hat{c}_{0l} - \hat{c}_{1l} z_{ij})^2}{n_i \sum_{l=1}^{M} \sum_{i=1}^{N} w_{il}}. \tag{3.13}$$

Notice that, since $w_l = p(\mathbf{b}_i = \mathbf{c}_l)$, then

$$\begin{aligned} W_{il} &= \frac{w_l \, p(\mathbf{y}_i | \beta, \sigma^2, \mathbf{c}_l)}{\sum_{k=1}^{M} w_k \, p(\mathbf{y}_i | \beta, \sigma^2, \mathbf{c}_k)} = \frac{p(\mathbf{b}_i = \mathbf{c}_l) \, p(\mathbf{y}_i | \beta, \sigma^2, \mathbf{c}_l)}{p(\mathbf{y}_i | \beta, \sigma^2)} = \\ &= \frac{p(\mathbf{y}_i, \mathbf{b}_i = \mathbf{c}_l | \beta, \sigma^2)}{p(\mathbf{y}_i | \beta, \sigma^2)} = p(\mathbf{b}_i = \mathbf{c}_l | \mathbf{y}_i, \beta, \sigma^2). \end{aligned} \tag{3.14}$$

Therefore, in order to compute the point $\mathbf{c}_l$ for each group i, for $i = 1, \ldots, N$, we maximize the conditional probability of $\mathbf{b}_i$ given the observations $\mathbf{y}_i$, the coefficient β and the error variance σ^2. So that, the estimation of the coefficients $\mathbf{b}_i$ of the random effects for each group is obtained maximizing W_{il} over l, that is

$$\hat{\mathbf{b}}_i = \mathbf{c}_{\tilde{l}} \quad \text{where} \quad \tilde{l} = \arg\max_{l} W_{il} \quad i = 1, \ldots, N. \tag{3.15}$$

As anticipated before, the initialization of the support points is done in a robust and generalizable way. The algorithm starts considering N support points for the coefficients of random effects and a starting estimate for the coefficients of fixed effects. In particular, the initialization of all these parameters is done in the following way:

- random effects: the starting N support points are obtained fitting a simple linear regression within each group and estimating the couple of parameters (both the intercept and the slope) for each one of the N groups. The weights are uniformly distributed on these N support points[4];
- fixed effects: the starting values of β and σ^2 are estimated by fitting a unique linear regression on the entire population (without distinction among the groups).

Nonetheless, if the number of starting support points N is extremely large, the algorithm is relatively slow and using N starting support points becomes not strictly necessary. In this case, the initialization of the support points of the random effect distribution is done in the following way:

- we choose a number N* < N of support points;
- we extract N* points from a uniform distribution with support on the entire range of possible values, that is estimated by fitting N distinct linear regressions for each one of the N groups, as before, and identifying the minimum and the maximum values;
- we uniformly distribute the weights on these N* support points.

During the iterations, the EM algorithm performs the support reduction of the discrete distribution, in order to identify $M < N$ mass points in which the N groups are clustered. The support reduction is made standing on two criteria. The former is that we fix a threshold D and if two points $\mathbf{c}_l$ and $\mathbf{c}_k$ are closer than D, in terms of euclidean distance, they collapse to a unique point $\mathbf{c}_{l,k}$, where $\mathbf{c}_{l,k} = \frac{\mathbf{c}_l + \mathbf{c}_k}{2}$ with weight $w_{l,k} = w_l + w_k$. The first two masses collapsing to a unique point are the two masses with the minimum euclidean distance, among the couples of masses with euclidean distance less than D, and so on so forth. The latter is that, starting from a given iteration up to the end, we fix a threshold $\tilde{w}$ and we remove mass points with weight $w_l \leq \tilde{w}$ or that are not associated to any subpopulation. D and $\tilde{w}$ are two tuning parameters that tune the estimates of the subpopulations. The choice of D depends on how much we want to be sensitive to the differences among subpopulations: the higher is D, the lower is the number of subpopulations and the less homogeneous are the groups within subpopulations. D depends also on the order of magnitude of the data. The choice of $\tilde{w}$ depends on the minimum number of groups that we allow within each subpopulation. When one or more mass points are deleted, the remaining weights are reparameterized in such a way that they sum up to 1:

[4]This is not the only possibility to estimate the starting support points. A valuable alternative is to fit a classical multilevel model, with N groups, where both the intercept and the slope are random coefficients.

$$
\begin{aligned}
S_w &= \sum_{l=1}^{M^{new}} w_l^{old} \\
w_l^{new} &= \frac{w_l^{old}}{S_w} \quad \forall l = 1, \ldots, M^{new}
\end{aligned}
\tag{3.16}
$$

where M^{new} is the total number of masses after deleting the ones associated to weight $w_l \leq \tilde{w}$ or not associated to any subpopulation, $\mathbf{w}^{old}$ are the old remaining weights and $\mathbf{w}^{new}$ are the new reparameterized weights.

The sketch of the algorithm is shown in Algorithm 1 in Appendix. At each iteration k, the algorithm, given the estimated number of mass points, estimates all the parameters in (3.3) in an iterative way, updating the coefficients of both fixed and random effects, until convergence or until it reaches the maximum number of sub-iterations fixed a priori for this stage (`itmax`). At the beginning of the iterative process, the algorithm performs the dimensional reduction of the mass points standing only on the distance between the mass points. When the estimates are stable, meaning that all the differences between the estimates of the parameters at two consecutive iterations are smaller than fixed tolerance values, or after a given number of iterations `it1`, the algorithm continues performing the dimensional reduction of the support points standing also on the criterion of the minimum weight $\tilde{w}$. The final convergence is reached when all the differences between the estimates of the parameters at two consecutive iterations are smaller than fixed tolerance values. In particular, we fix the tolerance values for the estimates of both the parameters of fixed and random effects to `tollF` and `tollR` respectively, which depend on the scale of the parameters.

The introduction of the maximum number of iterations `it`, `it1` and `itmax` (as just explained in this section) depends on the complexity of the data and on the consequent convergence rate and its use is merely to avoid an infinite loop.

It is worth noting that since the optimization steps are done in closed-form, the algorithm is not particularly time-consuming and, in both the simulation study and in the application, it converges in less than 20 iterations.

In the presentation of the algorithm, as well as in the simulation study that will be presented in the next subsection, we focus on the case of a linear model with two covariates, where both one slope and the intercept are considered as random effects. This is due to the upcoming application of the algorithm to the case study of INVALSI dataset, in which we make this choice of fixed and random parameters. Nonetheless, the SPEM algorithm allows to consider as random effects both the intercept and one slope, as well as only one of them. Moreover, its extension to the case with p covariates among the random effects, i.e. $\mathbf{c} \in \mathbb{R}^{p+1}$, is analytically straightforward and it implies only a computational issue.

3.1.3 Simulation study

In order to validate the proposed estimation algorithm, we perform two simulation studies: the former considers the case of a population containing three latent subpopulations and the latter considers the case of a population with no latent subpopulations. In this way, we can test the algorithm in the presence of clear subpopulations and also in the case in which there are no clear subpopulations. We apply the algorithm considering different values of D, in order to test how the results do change by changing the threshold parameter and we provide a measure of the uncertainty of classification by computing the entropy in the weights matrix W. We consider a linear model with two covariates.

For the first simulation study, we generate a dataset containing 100 groups of variables (100 level 2 units), where each group is composed by an answer variable and two covariates. We sample the variables in order to have 3 different latent subpopulations within the 100 groups, that is, in order to create 100 cohorts of data characterized by three different linear correlations. For this purpose, we generate 100 response variables as the result of 3 distinct linear combinations of 3 couples of covariates, plus some errors. The three subpopulations contain 40, 25 and 35 groups respectively. The data are simulated in the following way:

$$\begin{cases} \mathbf{y}_i = \beta \mathbf{x}_1 + c_{01} + c_{11}\mathbf{z}_1 + \boldsymbol{\epsilon}_i & i = 1, \dots, 40 \\ \mathbf{y}_i = \beta \mathbf{x}_2 + c_{02} + c_{12}\mathbf{z}_2 + \boldsymbol{\epsilon}_i & i = 41, \dots, 65 \\ \mathbf{y}_i = \beta \mathbf{x}_3 + c_{03} + c_{13}\mathbf{z}_3 + \boldsymbol{\epsilon}_i & i = 66, \dots, 100 \end{cases} \tag{3.17}$$

where coefficients β and $\mathbf{c}_l$, for $l = 1, \dots, 3$ are reported in Table 3.1, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, 3)$ and the covariates are sampled by Normal distributions with different parameters. In particular,

$$\begin{aligned} \mathbf{x}_1 &\sim \mathcal{N}(0.30, 0.16), & \mathbf{z}_1 &\sim \mathcal{N}(50, 100), \\ \mathbf{x}_2 &\sim \mathcal{N}(0.28, 0.16), & \mathbf{z}_2 &\sim \mathcal{N}(51, 100), \\ \mathbf{x}_3 &\sim \mathcal{N}(0.27, 0.16), & \mathbf{z}_3 &\sim \mathcal{N}(49, 100), \end{aligned} \tag{3.18}$$

where $\mathbf{z}_1$ and $\mathbf{x}_1$ have 100 observations, $\mathbf{z}_2$ and $\mathbf{x}_2$ have 90 observations and $\mathbf{z}_3$ and $\mathbf{x}_3$ have 95 observations ($9,575$ level 1 units in total). Therefore, the dimensional choices of the generated data are the following:

- Number of groups = 100
- Number of subjects within groups = $\begin{cases} 100 & \forall \quad \text{group i} \in \{1, \dots, 40\} \\ 90 & \forall \quad \text{group i} \in \{41, \dots, 65\} \\ 95 & \forall \quad \text{group i} \in \{66, \dots, 100\} \end{cases}$

The choice of the size, of the parameters and of the distribution is arbitrary. Our choice for the values of x and z is driven by the case study. We sample x and z in order to obtain values in the same range of the ones in the INVALSI application. Other choices are possible and do not affect the validity of results.

	c_0	c_1	β
l=1	20	1.00	1.50
l=2	30	0.05	1.50
l=3	40	0.50	1.50

Table 3.1: Coefficients used for data simulation in Eq. (3.17). Each row corresponds to a subpopulation l. The intercept and the coefficient of $\mathbf{z}$ differ across subpopulations (c_0 and c_1 respectively), while the coefficient of x (β) is fixed.

Also the choice of the coefficients in Table 3.1 is arbitrary. This choice of parameters is driven by the case study, since we choose values for $\mathbf{c}_l$, for $l = 1, \ldots, 3$ and β in the same range of the ones obtained in the INVALSI application. For coherence with the upcoming INVALSI case study, that considers both the slope and the intercept as random, we choose different values for both the intercept and the coefficient of variable $\mathbf{z}$ across the three subpopulations, while we maintain the coefficient of $\mathbf{x}$ fixed. Figure 3.1 shows the 3d image of one simulated dataset.

Looking at Figure 3.1, it is possible to recognize three different linear correlations among the data, identified by the three distinct "clouds" of points. Groups of points characterized by similar linear correlations are automatically associated to similar colors by the software R and this helps in the visual inspection of the 3 subpopulations.

The model that we fit takes the following form:

$$\mathbf{y}_i = \beta \mathbf{x}_i + c_{0l} + c_{1l}\mathbf{z}_i + \boldsymbol{\epsilon}_i, \tag{3.19}$$

where $i = 1, \ldots, 100$ and $l = 1, \ldots, M$ where M is unknown a priori to the algorithm. We apply the algorithm 100 times, to different simulated datasets for the same model, for each different value of $D = \{0.5, 0.8, 1, 2, 3\}$ and considering the following choice of the other parameters: $\tilde{w} = 0.05$, `it`=30, `it1`=20, `itmax` = 20 and `toolF`=`tollR`= 10^{-4}. The following box summarizes the simulation study.

$\forall$ D $\in$ {0.5, 0.8, 1, 2, 3} and for (k in 1:100)

- generate $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3$ according to Eq. (3.18) and $\mathbf{y}$ according to Eq. (3.17);
- apply the SPEM algorithm to the generated data.

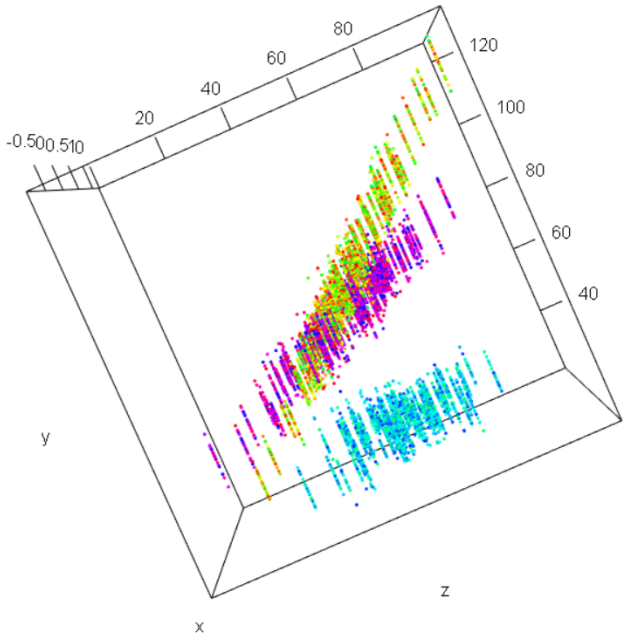


Figure 3.1: Plot of the simulated data obtained by Eq. (3.17) and (3.18). Each one of the 100 groups has a different color. Data with similar behaviors are automatically assigned to similar colors by software R.

The number of times, out of the 100 runs, in which the algorithm allocates the right subpopulation to each one of the 100 groups, for different values of D, is shown in Table 3.2 (in all the runs, the algorithm converges before the maximum number of iterations).

$D = 0.5$	$D = 0.8$	$D = 1$	$D = 2$	$D = 3$
34	84	92	98	68

Table 3.2: Number of times, out of the 100 runs, in which the algorithm allocates the right subpopulation to each one of the 100 groups for different values of D.

In the case in which D is equal to 0.5, the algorithm correctly assigns the belonging of the groups to the three subpopulations 34 times out of 100. In the remaining 66 cases, the algorithm identifies more than three subpopulations. This means that the threshold value $D = 0.5$ is too small and the algorithm is, consequently, too sensitive to the variations among the data. On the other side, in the case in which D is equal to 3, the algorithm correctly assigns the belonging of the groups to the three subpopulations 68 times out of 100 (identifying less than three subpopulations in the remaining 32 cases). This means

that for values of D higher than 3, the algorithm is not perfectly sensitive to the differences among the groups and it sometimes collapses groups presenting different trends into the same subpopulation. In the cases of D={0.8, 1, 2} the algorithm correctly assigns the subpopulations 84, 92 and 96 times out of 100, respectively, that represents a good proportion. The results of the estimates of the parameters for the two "best" choices of D are shown in Table 3.3.

		$\hat{c}_0$		$\hat{c}_1$		$\hat{\beta}$		$\hat{w}$
		Mean	sd	Mean	sd	Mean	sd	
D=1	l=1	20.034	0.170	0.999	0.003	1.477	0.005	0.40
	l=2	40.001	0.197	0.500	0.003			0.25
	l=3	30.032	0.292	0.049	0.005			0.35
D=2	l=1	20.011	0.154	1.000	0.003	1.505	0.004	0.40
	l=2	40.038	0.176	0.499	0.004			0.25
	l=3	29.987	0.236	0.050	0.004			0.35

Table 3.3: Distribution of the parameters of model in Eq. (3.19), estimated by the SPEM algorithm, obtained in the runs in which three populations are identified. Results are shown both for D = 1 and D = 2. Within each choice of D, each row corresponds to a subpopulation l. The intercept and the coefficient of $\mathbf{z}$ differ across subpopulations (c_0 and c_1 respectively), while the coefficient of $\mathbf{x}$ (β) is fixed. $\hat{w}$ represents the weight estimated for each subpopulation.

Starting from 100 distinct groups, the SPEM algorithm, in most of the cases, identifies three subpopulations (M = 3) that are represented by the estimates $(\hat{\mathbf{c}}_l, \hat{w}_l)$, for each $l = 1, \ldots, M$, and $\hat{\beta}$ shown in Table 3.3. The estimates obtained with D=1 and D=2 are coherent. The mean of each parameter distribution is centered very close to the real value of the parameter used to simulate the data and standard deviations are very small[5]. Moreover, masses' volumes are proportional to the percentage of data that belongs to each mass. In this case, the algorithm correctly assigns the 100 groups to the three subpopulations, so that, the three volumes are proportional to 0.40, 0.25 and 0.35, respectively. For one of the 100 simulated datasets in which the algorithm identifies the three clusters, data with the three identified regression planes are shown in Figure 3.2. In Figure 3.2, observations that belong to the same subpopulation are associated to the same color and, in this simulation, the algorithm associates each observation to the correct subpopulation. The three identified regression planes are able to fit the three distinct clouds of data in a precise way. In order to have a measure of the uncertainty of classification of the SPEM algorithm, we can observe the matrices of the weights W that we obtain in each run and evaluate

[5]In order to test the equality of the mean of each parameter distribution to the parameters shown in Table 3.1, we test the normality of each parameter distribution by means of Shapiro test, obtaining p-values > 0.1 for all of them, and we perform a t-test for each parameter (c_0, c_1 ad β), obtaining p-values > 0.2 for all the tests.

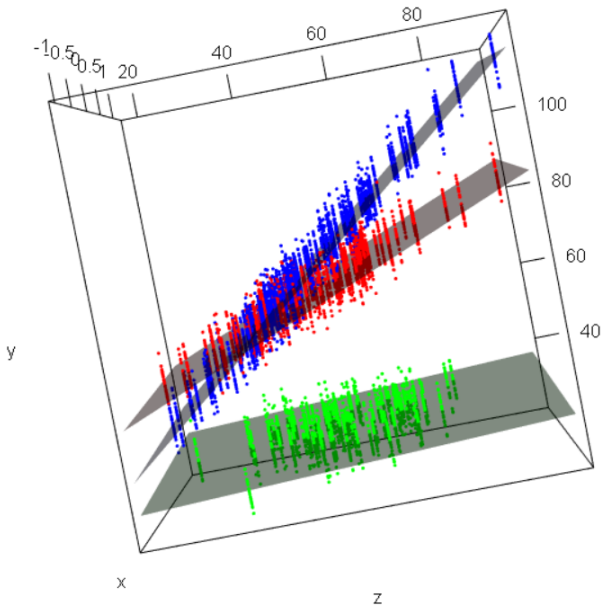


Figure 3.2: Result of the SPEM algorithm applied to a simulated dataset according to Eq. (3.17) and (3.18). Colors represent the three subpopulations that the algorithm identifies and planes are the estimated linear regression planes within each subpopulation. Each group is painted with the color of the subpopulation to which it belongs.

the level of uncertainty with which the algorithm assigns each group to a cluster. This uncertainty of classification can be evaluated by measuring the entropy of the rows of the matrix W. In the best case, that is when the algorithm assigns each group i to a cluster l with probability 1, each row of the matrix W would be composed by M-1 values equal to zero and a value equal to 1. In this scenario, the entropy $E_i = -\sum_{l=1}^{M} W_{il} ln(W_{il})$ of each row i of the matrix W would be equal to 0. The more the distribution of the weights is uniform on the M mass points, the higher is the entropy. The worst case when $M = 3$ is the one in which the distribution of the weights of a group i is uniform on the 3 clusters ($w_{il} = 1/3$ for $l = 1, 2, 3$), that corresponds to an entropy $E_i = -3 \times (1/3) ln(1/3) = 1.098$. We compute the entropy of each row of W for the 100 runs and we show here the distribution of the mean on the 100 runs of the entropy measured for each group i, in the cases of $D = 0.8$, $D = 1$ and $D = 2$.

The mean and the standard deviation of the entropy estimated when $D = 0.8$, $D = 1$ and $D = 2$ are shown in Table 3.4.

These very low values of the entropy (see Figure 3.3 and Table 3.4) suggests that the level of uncertainty of classification, for these three values of D, is very low, since the distribution of the weights w_{il}, for $i = 1, \ldots, N$ and $l = 1, \ldots, 3$

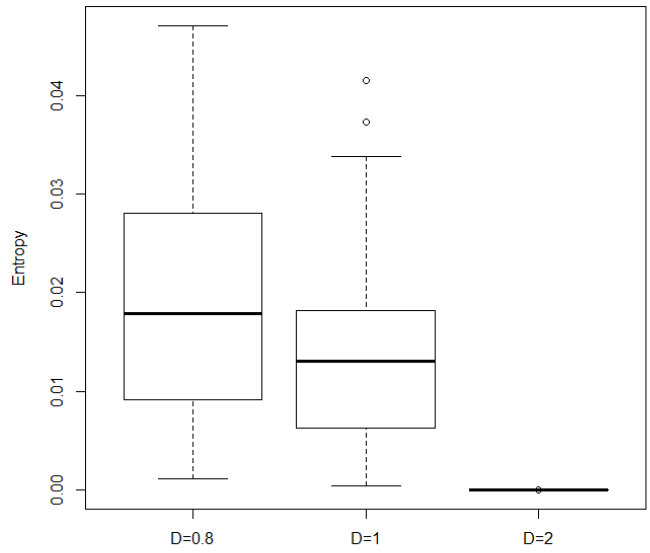


Figure 3.3: Boxplots of the entropy computed in the 100 runs, for D=0.8, D=1 and D=2 . Each boxplot represents the distribution of the entropy measured for each group, obtained by mediating the entropy in the 100 runs.

	mean	sd
D=0.8	0.019	0.012
D=1	0.013	0.008
D=2	9.7×10^{-13}	9.6×10^{-12}

Table 3.4: Mean and standard deviation of the entropy estimated when D=0.8, D=1 and D=2 on the 100 runs of the simulation, for the choice of data in Eq. (3.18) and coefficients in Table 3.1.

results to be very concentrated on single mass points. In particular, the case in which $D = 2$ has the lowest entropy and results to be the case with the lowest level of uncertainty of classification.

We can conclude that, in this simulation study, the SPEM algorithm is able to identify the latent structure that elapses within the 100 groups of data. In particular, it can identify which is the effective number of subpopulations in which the data are nested and it can characterize each one of these subpopulations by means of the estimates of the associated parameters.

In the second simulation study, we generate a population without latent subpopulations and we analyze the performance of the SPEM algorithm. We choose

one of the previous set of parameters and we generate 100 response variables in the following way:

$$\mathbf{y}_i = 20 + 1.5\mathbf{x}_i + 1\mathbf{z}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, 100, \tag{3.20}$$

where $\epsilon_i \sim \mathcal{N}(0, 3)$ and x_i and z_i are defined as in Eq. (3.18). Again, we apply the algorithm 100 times to 100 different simulated datasets and this process is repeated for values of $D = \{0.5, 0.8, 1, 2, 3\}$ and considering the following choice of the other parameters: $\tilde{w} = 0.05$, `it`=30, `it1`=20, `itmax` = 20 and `toolF`=`tollR`= 10^{-4}. The number of times, out of the 100 runs, in which the algorithm identifies only one subpopulation, for different values of D, is shown in Table 3.5.

$D = 0.5$	$D = 0.8$	$D = 1$	$D = 2$	$D = 3$
52	74	90	100	100

Table 3.5: Number of times, out of the 100 runs, in which the algorithm identifies only one subpopulation, for different values of D.

For $D = 2$ and $D = 3$, the algorithm always recognizes that there are no subpopulations. For smaller values of D, sometimes the algorithm catches heterogeneities among the 100 groups of data and identifies the presence of latent subpopulations. For $D = 2$, Table 3.6 shows the distribution of the estimated coefficients in the 100 runs.

	$\hat{c}_0$		$\hat{c}_1$		$\hat{\beta}$		$\hat{w}$
	Mean	sd	Mean	sd	Mean	sd	
l=1	20.012	0.099	0.999	0.002	1.493	0.081	1

Table 3.6: Distribution of the parameters of model in Eq. (3.20), estimated by the SPEM algorithm, obtained in the 100 runs. Results are shown for $D = 2$, but are coherent with any other choice of $D \geq 2$.

Regarding the uncertainty of classification, the entropy in the simulations done with $D = 2$ and $D = 3$ is zero, since 100 times out of 100 the algorithm identifies one population and each group has probability 1 to be assigned to it. In the cases of lower values of D, the algorithm sometimes identifies more than one population and the distribution of the entropy related to these cases are shown In Figure 3.4.

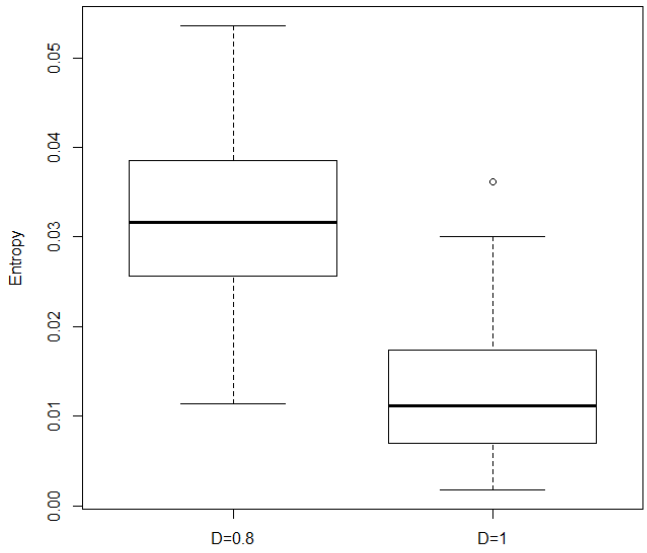


Figure 3.4: Boxplots of the entropy computed in the 100 runs, for $D = 0.8$ and $D = 1$. Each boxplot represents the distribution of the entropy measured for each group, obtained by mediating the entropy in the 100 runs.

The mean and the standard deviation of the entropy estimated when $D = 0.8$ and $D = 1$ are shown in Table 3.7.

	mean	sd
D=0.8	0.032	0.008
D=1	0.012	0.007

Table 3.7: Mean and standard deviation of the entropy estimated when D=0.8 and D=1 on the 100 runs of the simulation, for the choice of coefficients in Eq. (3.20).

Also in this case, the estimates of the parameters result to be significantly equal to the parameters used to generate the data[6].

In general, by changing the value of D, we make the algorithm more or less sensitive to the heterogeneity among the groups of data, that is given both by the clustering induced by construction and by the remaining randomness in the model (e.g. by the error term). From this perspective, a graphical visualization of the results can help in the choice of D.

[6]Again, we test the normality of each parameter distribution, obtaining p-values of the Shapiro test ¿ 0.1 for all of them, and the t-tests for the null hypotheses $c_0 = 5$, $c_1 = 3$ and $\beta = 10$ give p-values ¿ 0.2.

3.2 Case study: application of SPEM algorithm to education INVALSI data

In this section, we describe the INVALSI dataset (Subsection 3.2.1) and we apply the SPEM algorithm to these data, in order to identify subpopulations of Italian schools (Subsection 3.2.2). In a second step, we characterize the identified subpopulations by means of school level variables (Subsection 3.2.3).

3.2.1 The INVALSI 2013/2014 dataset

INVALSI is an Institute that tests Italian students at different grades and at different years. The data that we analyze in this chapter are taken from the INVALSI survey of 2013/2014. Among others, the survey provides several information both at student and at school level. Students, in addition to solve tests in different school subjects, have to fill out a questionnaire about theirselves, their family situations and their habits. Moreover, also school principals have to fill out a questionnaire about himself/herself, his/her school practices and management, school body composition and school size, school structures, infrastructures and school climate. The dataset collects information about $8,946$ students nested within 586 schools. The aim of applying the SPEM algorithm to INVALSI data is that we are interested in exploring the different relations between student performances at grade 6 and 8, across Italian junior secondary schools, adjusting for the student socio-economical index. For this reason, we select only three variables at student level to employ in the analysis:

- MATH8: student mathematics test score at grade 8 (students attending the last year of junior secondary school in the year 2013/2014);
- MATH6: student mathematics test score at grade 6 (students attending the first year of junior secondary school in the year 2011/2012);
- ESCS: student socio-economical index.

Student test scores range between 0 and 100, while the ESCS is an indicator built by INVALSI as a continuous variable with mean = 0 and variance = 1. This indicator considers (i) parents' occupation and educational qualifications, and (ii) whether the student owns certain items at home (for instance, the number of books). In general, pupils with an ESCS greater than or equal to 2 are socially and culturally highly advantaged. Figure 3.5 and Table 3.8 show variables distributions and descriptive statistics respectively.

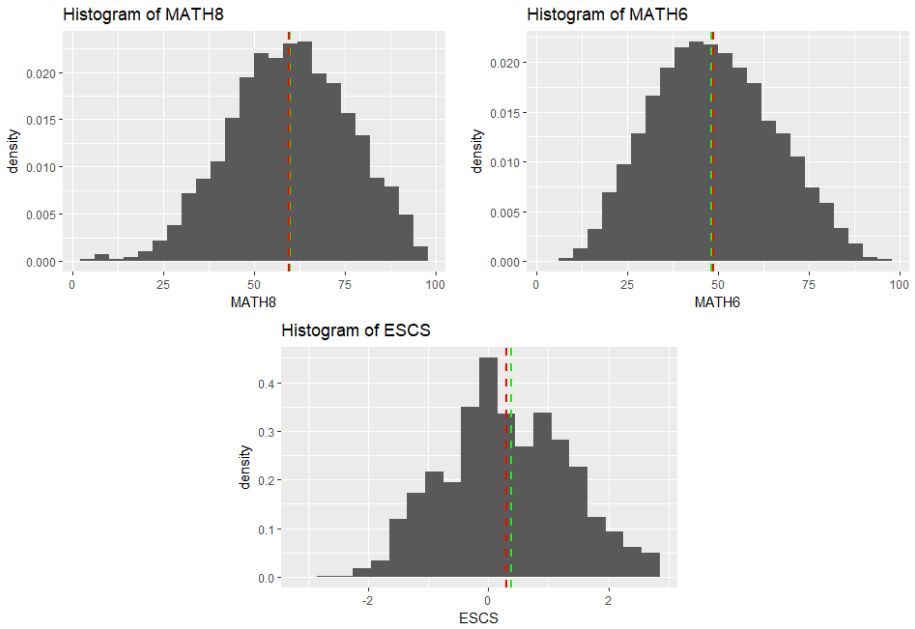


Figure 3.5: Histograms of students' INVALSI test scores at grade 8, at grade 6 and socio-economical index (ESCS). Red lines refer to the means, green ones to the medians.

	Mean	sd	Median	IQR
MATH8	59.73	16.49	60.98	23.29
MATH6	48.69	16.83	48.26	24.55
ESCS	0.30	1.02	0.38	1.40

Table 3.8: Descriptive statistics of student level variables employed in the analysis.

Moreover, we have information about the macro-area of localization of schools. About 59% of schools is in Northern Italy, 18% is in Central Italy and 23% is in Southern Italy. Geographical information is a very relevant aspect since many studies in Italy confirm that there are significant discrepancies between student and school performances across the three geographical macro-areas (Agasisti and Vittadini, 2012; Agasisti et al., 2017b; Masci et al., 2016b, 2017a).

Since, in a second stage of the analysis, we will look for a characterization of the identified school subpopulations, Table 3.9 reports the school level variables that we are interested in, with their descriptive statistics. In particular, variables concern three aspects of schools. The first one concerns the *school body composition*: school mean socio-economical index, percentage of females, immigrants, late/early-enrolled students[7], school size and the dummy for private/public school. The second one is about the *school principal's features*:

[7]Late/early-enrolled students are those students who started the school grade later or earlier respect to their peers.

gender, age, education and years of experience. Lastly, we have three composite indicators[8] about (i) school climate and human relations, (ii) managerial practices and principal's strategy and (iii) structures and resources of the school.

Variable Name	Mean	sd	Median	IQR
Mean ESCS	0.26	0.54	0.27	0.58
Female percentage	50.11	10.83	50.00	14.28
Immigrant percentage	10.52	11.15	8.01	16.66
Early-enrolled student percent	1.21	4.13	0.00	0.00
Late-enrolled student percent	8.52	8.02	6.66	13.04
Number of classes	20.15	3.77	21.00	5.01
Number of school complexes	5.37	2.81	6.01	5.00
Private	8.21%	–	–	–
Principal features:				
Gender(Female=1)	70.01%	–	–	–
Age	55.13	7.49	56.00	11.00
Master after degree(yes=1)	22%	–	–	–
Scientific education(yes=1)	14.62%	–	–	–
Year of experience	9.23	7.79	7.00	10.00
Year of experience in the actual school	5.08	5.18	3.00	5.00
Experience in an other district	25.37%	–	–	–
Experience with INVALSI	51.34%	–	–	–
Composite indicators:				
Ind 1: school climate and human relations	0.96	0.09	1	0
Ind 2: managerial practices and principal's strategy	0.86	0.11	0.83	0.12
Ind 3: structures and resources of the school	0.94	0.09	1	0.11

Table 3.9: School level variables of the database used in the analysis, with their descriptive statistics.

3.2.2 SPEM algorithm applied to INVALSI data

The aim of this subsection is to apply the EM algorithm for semi-parametric mixed-effects models to INVALSI database of 2013/2014 as a tool for clustering

[8]The computation of these three *composite indicators* is shown in Masci et al. (2016a).

Italian schools standing on their student attainments. The correlation between previous student scores (grade 6) and current student scores (grade 8) changes across schools, in the sense that the effects that schools give to student attainments are heterogeneous and depend on different school characteristics. From this perspective, student scores at grade 8 can be seen as the result of student scores two years before (grade 6) combined with the effect of having attended a particular school for two years. The idea is to find out how student test scores at grade 6 and grade 8 are related to each other in different schools and in which schools these relationships are similar. In other words, we look for how many and which different trends exist in the scores of students attending Italian schools and, standing on the results, we group schools into different subpopulations. In this perspective, the SPEM algorithm works as an in-built classifier, since it performs the grouping of schools into subpopulations, without knowing a priori the number of subpopulations.

Standing on previous literature, it is reasonable to think that there is a linear correlation between student scores at grade 6 and at grade 8 (Agasisti et al., 2017b; Masci et al., 2016b, 2017a). We therefore consider a semi-parametric two-level linear model (where students represent the first level and schools the second one), with student test scores at grade 6 and student socio-economical index as random and fixed effects respectively, allowing both the intercept and the coefficient of student test scores at grade 6 to be random/school-specific. For each student j, $j = 1, \dots, n_i$, and each school i, $i = 1, \dots, N$, given that N is the total number of schools, J is the total number of students and $\sum_{i=1}^{N} n_i = J$, the model takes the following form:

$$\begin{aligned} \mathbf{y}_i = \mathbf{x}_i\beta + \mathbf{1}b_{0i} + \mathbf{z}_i b_{1i} + \boldsymbol{\epsilon}_i \qquad & i = 1, \dots, N \\ \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{1}_{n_i}) \qquad & ind. \end{aligned} \tag{3.21}$$

where the answer variable $\mathbf{y}_i = (y_{1i}, \dots, y_{n_i i})$ is the mathematics test score at grade 8 (MATH8) of the n_i students within school i, while the covariate $\mathbf{z}_i = (z_{1i}, \dots, z_{n_i i})$ and the covariate $\mathbf{x}_i = (x_{1i}, \dots, x_{n_i i})$ are respectively the mathematics test score at grade 6 (MATH6) and the socio-economical index (ESCS) of the n_i students within the i-th school. The choice of considering ESCS as fixed effect and MATH6 as random one is due to the fact that we are interested in exploring how the correlation between MATH6 and MATH8, seen as the reflex of schools ability in training students to achieve certain results, given their students starting potential, varies among schools.

In order to have robust estimates, we select, from the dataset presented in Section 3.1, only the schools that have at least ten students. The resulting dataset consists of $6,188$ students nested within 363 schools.

The SPEM algorithm is applied, considering $\tilde{w} = 0.015$, $D = 0.8$, `it`=30, `itmax`=`it1`=20 and `tollR`=`tollF`=10^{-4}. Given these parameters, the algo-

rithm identifies $M = 5$ distinct subpopulations, whose estimates of parameters are shown in Table 3.10.

Subpopulation	$\hat{\beta}$	$\hat{c}_0$	$\hat{c}_1$	$\hat{w}$
Subpopulation 1	1.417	46.028	0.454	12.2%
Subpopulation 2	1.417	22.579	0.707	39.6%
Subpopulation 3	1.417	30.293	0.648	37.5%
Subpopulation 4	1.417	31.207	0.393	8.8%
Subpopulation 5	1.417	25.359	0.027	1.9%

Table 3.10: ML estimates of coefficients of model (3.21) obtained applying the SPEM algorithm to a selection of INVALSI data of 2013/2014.

The coefficient β in Table 3.10 is the coefficient related to ESCS (fixed effect). Its positive value (1.417) suggests that, on average, students with high socio-economical index are associated to high performances, in line with previous literature (Sirin, 2005). The estimated $\hat{w}_l$, for $l = 1, \ldots, M$, express the percentage of Italian schools belonging to each subpopulation l, for $l = 1, \ldots, M$. We identify two main subpopulations (subpopulation 2 and subpopulation 3 in Table 3.10), that contain about the 77% of the total population, while the remaining 23% is distributed across the three other subpopulations. Regarding the analysis of the coefficients of random effects, Figure 3.6 helps us in their visualization.

Looking at Figure 3.6, it is immediately evident that there is a quite anomalous subpopulation, identified by lilac color, characterized by a very low slope (subpopulation 5 in Table 3.10). From an interpretative point of view, this subpopulation contains the "worse" set of Italian schools. Indeed, it is characterized by both low intercept and slope and this means that students in these kind of schools have on average low results at grade 8, even if they had good results at grade 6. In other words, students have on average low scores, without variability depending on their previous performances: students that had good results at grade 6, after attending two years in a secondary school belonging to subpopulation 5, have on average low performances, similar to the ones of those students that performed worse than them two years before. On the other side, the best scenario is represented by the subpopulation on the top of Figure 3.6, identified by red color (subpopulation 1 in Table 3.10), that is characterized by a very high intercept (46.028) and a still high slope (0.454). These values suggest that even students that had very low scores at grade 6, obtain high scores at grade 8 with respect to their counterparts attending schools belonging to other subpopulations. Moreover, the value of the slope suggests that, even if students had on average an improvement on their performances, there is still heterogeneity across students that performed differently two years before, in the sense that

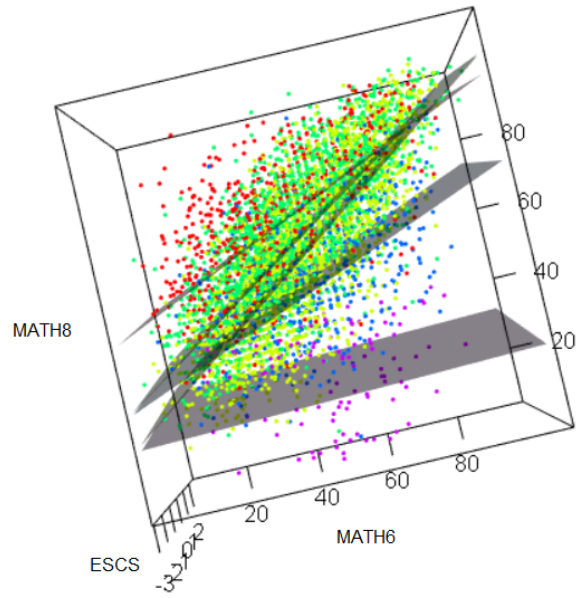


Figure 3.6: Plot of INVALSI data with the five regression planes identified by the SPEM algorithm , for model (3.3). Parameters are shown in Table 3.10. Colors represent the five subpopulations.

best students continue to perform the best with respect to the average.

Thanks to the multilevel structure, we can also compute the Percentage of Variability explained by Random Effects (PVRE), that, in our case, is the percentage of variability in student test scores explained at school level:

$$PVRE_{School} = \frac{\sigma^2_{School}}{\sigma^2_{School} + \sigma^2_{Residuals}}.$$

Given the two-level semi-parametric model:

$$\mathbf{y}_i = \beta \mathbf{x}_i + c_{0l} + c_{1l}\mathbf{z}_i + \boldsymbol{\epsilon}_i,$$

the variance of random effects is given by

$$\sigma^2_{School} = \sigma^2_{c_0} + 2Cov(c_0, c_1)\bar{z} + \sigma^2_{c_1}\bar{z}^2.$$

Computing the empirical values of $\sigma^2_{c_0}$, $Cov(c_0, c_1)$ and $\sigma^2_{c_1}$ from the estimated parameters, we obtain a PVRE equal to 70.48%. This quantity confirms the significance of the random effects in explaining the answer, since about the 70% of the explained variability at student level is explained by differences across schools.

In order to provide an index for the goodness of fit of the model, we provide a leave-one-out cross-validation, we compute the Mean Square Error (MSE) and

we compare it with the ones obtained considering (i) the same model but with all the parameters as fixed effects and (ii) the parametric mixed-effects models with the same choice of random and fixed effects. Table 3.11 reports the three MSE computed on the student test scores.

	Parametric FE model	Parametric RE model	SPEM random intercept/slope
MSE	155.91	111.55	118.69

Table 3.11: Mean Square Error (MSE) computed in three models: (i) parametric fixed-effects model (Parametric FE model); (ii) parametric mixed-effects models with both intercept and covariate as random effects (Parametric RE model); (iii) Semi-parametric mixed-effects model with both intercept and slope as random effects (SPEM random intercept/slope).

The MSE obtained with the fixed-effects model is the highest one (155.91) and it departs from the ones obtained by both the parametric and semi-parametric mixed-effects models (111.55 and 118.69 respectively). Standing on the nature of the problem, we expect the parametric mixed-effects model to perform the best, since it fits the trend of the data within each school. Nonetheless, the semi-parametric mixed-effects model produces a slightly bigger MSE, but it extrapolates a new kind of information from the data. Indeed, while the parametric approach is able to estimate the parameters of a model, that is based on an already known structure of the data, the semi-parametric approach makes a further step, since it is able to identify a new structure within the data, that is the existence of a new, latent level of grouping. From an interpretative point of view, the identification of subpopulations is highly informative in the perspective of identifying those groups that depart from the common behavior. Indeed, among the identification of subpopulations itself, what really matters is the identification of the minority subpopulations, that are those subpopulations containing a small percentage of the entire population, characterized by different properties with respect to the majority. In our application to INVALSI database, subpopulations 2 and 3, that are very close to each other and contain almost the 80% of the schools, represent the most common trend, but the subpopulations that deserve more attention are subpopulations 1,4 and 5, that are the ones containing a smaller percentage of schools that behave differently from the majority. Moreover, the relatively small difference between the MSEs of the two approaches suggests that the subpopulations structure identified by the SPEM algorithm catches almost all the heterogeneity across the impacts of Italian schools, meaning that the subpopulations are quite homogeneous.

The further consequence of the identification of a latent structure within the data is that subpopulations likely derive from some unknown characteristics

of schools, that lead to these differences. In a general perspective, the interpretation a posteriori of subpopulations of data is important per se, especially when speaking about Big Data, where the identification of patterns within a big amount of data, marked by a complex and unknown structure, is particularly relevant. For this reason, in the next subsection, we try to find out whether there are patterns of school level variables that characterize the estimated subpopulations.

3.2.3 Association between school characteristics and school subpopulations

Applying the SPEM algorithm to INVALSI data, we discover a structure of subpopulations that clearly reflects heterogeneities among the impacts of Italian schools. In particular, we identify five different subpopulations, that emerge from five different behaviors of schools in affecting the evolution of their student achievements. We are interested in exploring a posteriori these subpopulations, in order to investigate whether there are school characteristics that are associated to them. Actually, among these five subpopulations, subpopulation 2 and subpopulation 3 in Table 3.10, that are characterized by similar parameters and that contain almost the 80% of the entire set of schools, represent the majority of schools. Consequently, we consider the union of subpopulation 2 and 3 as the reference subpopulation S_{ref}, that represents the reference trend. Our interest is to see how the school characteristics of the other three subpopulations (subpopulation 1,4 and 5 in Table 3.10) differ from the reference subpopulation. To this end, we apply a multinomial logit model by treating all the school level characteristics shown in Table 3.9 as covariates and as outcome variable the belonging to the four subpopulations.

For each group (school) $i = 1, \dots, N$ and each subpopulation l = {1,4,5}, the model takes the following form:

$$ln\Big(\frac{P(Y_i = l)}{P(Y_i = S_{ref})}\Big) = \beta_{0l} + \sum_{q=1}^{Q} \beta_{lq} X_{iq}. \tag{3.22}$$

where X is the N×Q matrix of school level covariates shown in Table 3.9, where Q is the total number of school level covariates. The results of the model in Eq. (3.22) are shown in Table 3.12.

Variable Name	subpop 1	subpop 4	subpop 5
Intercept	−2.287	−0.560	−23.363
Mean ESCS	−0.335	−0.043	0.087
Female percentage	0.013	−0.016	−0.011
Immigrant percentage	−0.069.	−0.077.	−0.246
Early-enrolled student percent	−0.095	0.030	0.014
Late-enrolled student percent	0.013	0.034	−0.012
Number of classes	−0.035	−0.008	0.067
Number of school complexes	0.078	−0.126	0.086
Private	0.884	−9.187***	−6.147***
Principal features:			
Gender (Female=1)	−0.192	−0.043	0.211
Age	0.018	−0.048	0.020
Master after degree (yes=1)	0.478	−0.577	0.981
Scientific education (yes=1)	−0.135	0.171	−6.019***
Year of experience	0.013	0.035	0.046
Year of experience in the actual school	−0.096	−0.034	−0.048
Experience in an other district	0.004	0.583	−1.390
Experience with INVALSI	−0.155	0.384	1.525
Composite indicators:			
Ind 1: school climate and human relations	0.327	−0.083	1.864
Ind 2: managerial practices and principal's strategy	2.899	−0.626	−5.762
Ind 3: structures and resources of the school	−3.588	2.726	6.553
Geographical area:			
Center	0.744	0.648	15.691***
South	1.201.	1.200.	14.687***

Table 3.12: Results of the multinomial logit model in Eq. (3.22). Coefficients are computed considering S_{ref}, the union of subpopulations 2 and 3, as the reference. Asterisks denote different levels of significance: . 0.01 < p-val < 0.1; * 0.001 < p-val < 0.01; ** 0.0001 < p-val < 0.001; *** p-val < 0.0001.

Among the big amount of school level variables, only four variables result to be associated to the belonging of schools to the four subpopulations: the percent-

age of immigrants, the dummy for public/private school, the kind of education of the school principal (humanistic/scientific) and the geographical area (Northern, Central and Southern Italy). With respect to the reference subpopulation, subpopulation 1 and 4 are more likely to contain schools with low percentages of immigrant students; Cluster 4 and 5 are less likely to contain private schools; subpopulation 5 is more likely to contain schools managed by school principals with a humanistic education rather than a scientific one; subpopulation 1 and 4 are more likely to contain schools in Southern Italy and subpopulation 5 is most likely to contain schools both in Central and Southern Italy. The fact that subpopulations 1 and 4 are more likely to contain schools with low percentages of immigrant students and are also more likely to contain schools in Southern Italy is actually an expected result since the majority of immigrant students in Italy live in Northern Italy. Subpopulations 1 and 4 are also the subpopulations with the highest intercepts and high positive slopes (see Table 3.10), being the best scenario of schools standing on our interpretation, and those schools result to be associated to Southern Italy and to low percentages of immigrant students. The fact that both subpopulations 4 and 5 are less likely to contain private schools reveals that private schools tend to be associated neither to the worst set of schools (subpopulation 5 of Table 3.10) nor to a very good set of schools (subpopulation 4 of Table 3.10).

Geographical differences represent an interesting aspect in the Italian educational context. Figure 3.7 reports the proportion of schools belonging to the five subpopulations, in the three geographical Italian macro-areas: Northern, Central and Southern Italy. Comparing Northern and Southern Italy, we can notice that the distribution of schools among subpopulations is different. In Northern Italy, we do not have any school belonging to subpopulation 5 and we have very few schools belonging to subpopulations 1 and 4: almost all schools belong to subpopulations 2 and 3. In Southern Italy, the distribution of schools among subpopulations is more uniform and it is possible to count a good quantity of schools belonging to each subpopulation.

The fact that, among the entire set of school level variables at our disposal, only four variables result to be significantly associated to the presence of subpopulations does not imply that there is no explanation for the presence of subpopulations of schools, but, most likely, these subpopulations derive from other dynamics, that we are not able to observe or to measure.

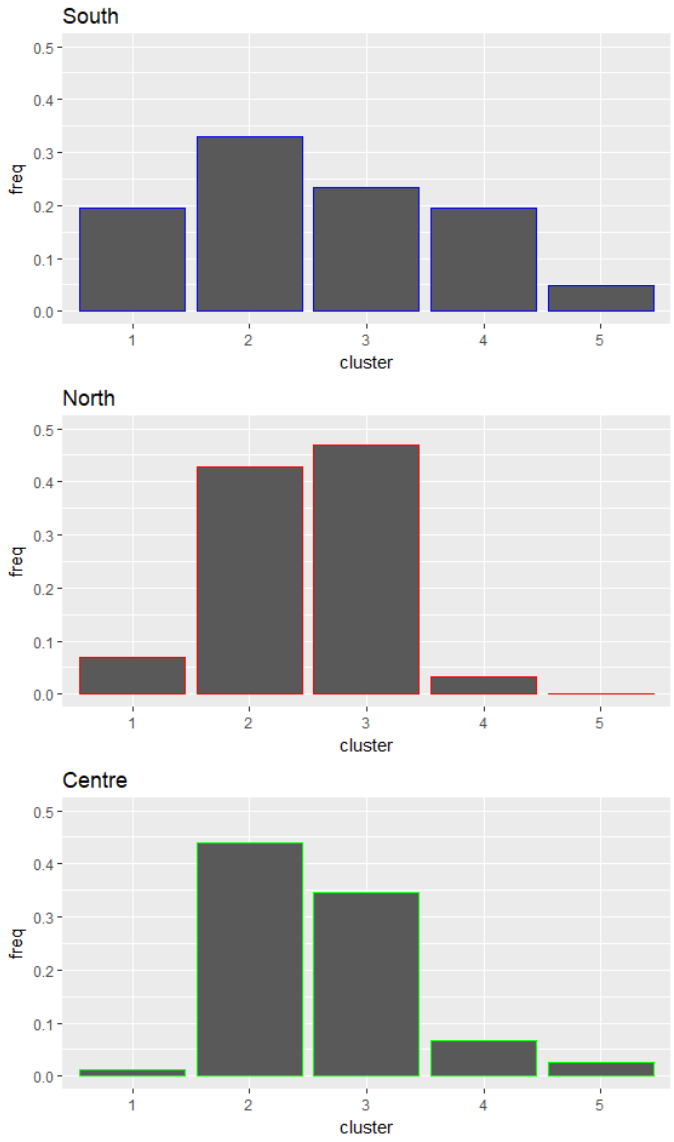


Figure 3.7: Proportion of schools belonging to the five subpopulations, within the three geographical Italian macro-areas: Northern, Central and Southern Italy.

3.3 Conclusions

This chapter proposes an EM algorithm for semi-parametric mixed-effects models (SPEM algorithm), shows a simulation study and applies the SPEM algorithm to INVALSI data of 2013/2014 as a tool for clustering Italian schools. The SPEM algorithm places itself in the literature branch concerning the algorithms proposed in Aitkin (1996); Azzimonti et al. (2013). In particular, our algorithm is inspired by the one proposed in Azzimonti et al. (2013) but it introduces the

major improvement, among the others, that the covariates are group specific, meaning that they can vary both in number of observations and range of assumed values across groups. Moreover, with respect to the algorithm proposed in Aitkin (1996) and the literature about Growth Mixture Models and Latent Class Analysis, the advantage of SPEM algorithm is that it does not need to fix a priori the number of discrete masses (subpopulations), but, standing on certain parameters, the algorithm itself identifies the number of discrete support points. This aspect has a great value in the 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 hera of Big Data, where there is the need of identifying latent structures within big and complex databases.

The SPEM algorithm, when applied to INVALSI data, is able to identify subpopulations of schools, within which student achievements trends differ. Among the identification of the number of subpopulations, that reveals how many different trends exist within the sample of Italian schools, the weights associated to the subpopulations, give a further information of the clustering. In a context in which we do not know a priori which is the expected trend, the subpopulations associated to higher weights represent the most common behavior, while the less numerous subpopulations (the ones associated to lower weights) represent those schools whose impact differs from the majority. This draws the attention on the determinants that bring schools to belong to the minority subpopulations. In particular, the algorithm identifies five school subpopulations that represent different school associations to their student achievements trends, seen as the ability of junior secondary schools in training students to obtain certain skills at the end of the three years, given their skills at the beginning of the school, adjusting for their socio-economical index (ESCS). In the INVALSI framework, schools are associated to a *positive or negative impact*, standing on the final performances of their students and given their students initial skills. Among these five subpopulations, the presence of a subpopulation containing schools with a negative impact is immediately evident. This subpopulation contains schools that have students which tend to underperform, with respect to their performance two years before, since they have on average very low scores, even if two years before, when they started to attend these schools, they obtained higher scores. Regarding positive impacts, we interpret the subpopulation with the highest intercept and positive slope (subpopulation 1) as the best one, in terms of school effect, since it contains schools able to train students to obtain high performances, even if they had low performances at the beginning of the school. It is worth to say that, from a policy perspective, the definition of the *best school effect* is currently debated. 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, on the other hand, a different point of view emphasizes the

advantages of having heterogeneity within the school. In this perspective, the role of the school is to continuously increase the student goals in order to stress the pupils to perform even better, using competition and variation to motivate them.

After the identification of school subpopulations, the chapter focuses on an other actual and interesting topic, that is their interpretation a posteriori. In particular, we explore the associations between school subpopulations and school level characteristics, showing that only geographical areas, percentage of immigrants, dummy for private/public school and school principal education result to be significantly associated. This evidence suggests that the school level variables at our disposal do not explain the differences in school impacts. Standing on the fact that the school subpopulations are clearly different in their effect on student attainments, the lack of a stratification of school level variables across subpopulations might means 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 are not able to measure, that might explain the different effects of schools on their students.

In a future perspective, our aim is to deepen the analysis on the characterization of the estimated school subpopulations, considering other information about the school environment, that we have not been able to measure until now. Moreover, from a methodological point of view, we aim at relaxing the linearity assumptions, to consider also the case of other functional forms. Lastly, in the next chapter, we develop the multivariate version of the SPEM algorithm, to consider two (or more) response variables. In the framework of INVALSI, since the dataset contains both the student scores in reading and 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 standing on both their effects on reading and mathematics student attainments, analyzing the interactions between these two fields.

Appendix

Algorithm 1: EM algorithm for semi-parametric mixed-effects models

input : Initial estimates for $(\mathbf{c}_1^{(0)}, \ldots, \mathbf{c}_M^{(0)})$ and $(w_1^{(0)}, \ldots, w_M^{(0)})$, with $M = N$;
Initial estimates for $\beta^{(0)}$ and $\sigma^{2(0)}$;
Tolerance parameters D, $\tilde{w}$, tollR, tollF, it, it1, itmax.

output: Final estimates of $\mathbf{c}_l^{(it)}$, $w_l^{(it)}$, for $l = 1, \ldots, M$, $\beta^{(it)}$ and $\sigma^{2(it)}$.

k=1; conv1=0; conv2=0;

while *(conv*1 == 0 *or conv*2 == 0 & $k < it$*)* **do**

compute the distance matrix DIST (where $DIST_{st} = \sqrt{(c_{0s} - c_{0t})^2 + (c_{1s} - c_{1t})^2}$ is the euclidean distance between each couple of mass points $s, t \; \forall s, t = 1, \ldots, M, s \neq t$);

if *(*$DIST_{st} < D$ & $DIST_{st} = min(DIST) \quad (\forall s, t = 1, \ldots, M, s \neq t)$*)* **then**

collapse masses s and t to a unique mass point;

compute the new distance matrix DIST;

if *conv*1 == 1 *or* $k \geq it1$ **then**

if $w_l^{(k)} \leq \tilde{w} \quad (\forall l = 1, \ldots, M)$ **then**

delete mass point l;

reparameterize the weights according to Eq. (3.16);

if *no changes are done* **then**

conv2=1;

given $\mathbf{c}_l^{(k-1)}, w_l^{(k-1)}$ for $l = 1, \ldots, M$, $\beta^{(k-1)}$ and $\sigma^{2(k-1)}$, compute the matrix W according to Eq. (3.7);

update the weights $w_1^{(k)}, \ldots, w_M^{(k)}$ according to Eq. (3.5);

$\beta^{(k,0)} = \beta^{(k-1)}$;

$\sigma^{2(k,0)} = \sigma^{2(k-1)}$;

$\mathbf{c}_l^{(k,0)} = \mathbf{c}_l^{(k-1)}$;

$w_l^{(k,0)} = w_l^{(k-1)}$;

keeping $\beta^{(k,0)}$ and $\sigma^{2(k,0)}$ fixed, update the M support points $\mathbf{c}_1^{(k,1)}, \ldots, \mathbf{c}_M^{(k,1)}$ according to Eq. (3.10) and (3.11);

keeping $\mathbf{c}_l^{(k,1)}, w_l^{(k,0)}$ for $l = 1, \ldots, M$ fixed, update $\beta^{(k,1)}$ and $\sigma^{2(k,1)}$ according to Eq. (3.12) and (3.13);

j=1;

while $(|\beta^{(k,j-1)} - \beta^{(k,j)}| \geq tollF \quad or \quad |\sigma^{2(k,j-1)} - \sigma^{2(k,j)}| \geq tollF \quad or \quad |\mathbf{c}_l^{(k,j-1)} - \mathbf{c}_l^{(k,j)}| \geq tollR) \quad \& \quad j \leq itmax$ **do**

j=j+1;

keeping $\beta^{(k,j-1)}$ and $\sigma^{2(k,j-1)}$ fixed, update the M support points $\mathbf{c}_1^{(k,j)}, \ldots, \mathbf{c}_M^{(k,j)}$ according to Eq. (3.10) and (3.11);

keeping $\mathbf{c}_l^{(k,j)}, w_l^{(k,j-1)}$ for $l = 1, \ldots, M$ fixed, update $\beta^{(k,j)}$ and $\sigma^{2(k,j)}$ according to Eq. (3.12) and (3.13);

set $\mathbf{c}_l^{(k)} = \mathbf{c}_l^{(k,j)}$ for $l = 1, \ldots, M$, $\beta^{(k)} = \beta^{(k,j)}$, $\sigma^{2(k)} = \sigma^{2(k,j)}$;

estimate subpopulation l for each group i according to Eq. (3.15);

if $(\beta^{(k)} - \beta^{(k-1)} < tollF)$ & $(\sigma^{2(k)} - \sigma^{2(k-1)} < tollF)$ & $(\mathbf{c}_l^{(k)} - \mathbf{c}_l^{(k-1)} < tollR)$ **then**

conv1=1;

k= k+1;

Chapter 4

Multivariate semi-parametric mixed-effects models for the joint clustering of Italian classes

In the previous chapter, we develop a semi-parametric two-level model able to identify a latent structure among the higher level of grouping, that in the educational application - in which we consider students nested within schools - is the school. With respect to similar methods present in the literature, the main advantage of our method is that it does not need to fix a priori the number of latent subpopulations, that is estimated by the algorithm together with the other parameters of the model. This aspect is of great value when analyzing big dataset where we do not have any prior about the number of existent subpopulations. Applying this method to INVALSI data, we classify schools standing on the evolution of their student achievements across years. In this sense, our concept of "school effect" is the effect that a school has on the evolution of its student achievements at different grades. In particular, we identify subpopulations of schools within which student mathematics test scores trends (measured by the linear relation of INVALSI test scores at different grades) are similar and, in a second step, we characterize a posteriori the identified subpopulations of schools by means of school level characteristics.

As stated in the Introduction, the INVALSI dataset has been previously studied by researchers interested in investigating the determinants of student, class and school performances. In some of our previous works (Agasisti et al., 2017b; Masci et al., 2016b, 2017a), considering the hierarchical nature of educational data, we apply mixed-effects linear models (Pinheiro and Bates, 2000) to INVALSI data in order to identify which are the student characteristics associated to student performances and to estimate how much of the variability

in student performance is due to their grouping in different classes and schools. These are the first attempts that aim at separating and estimating the effects of different levels of grouping on Italian student achievements. In Masci et al. (2016b, 2017a), we apply a three-level hierarchical linear model in which students are nested within classes that are in turn nested within schools and we measure the contribute of each of these levels on student INVALSI achievements variability. Results show that, after adjusting for student characteristics, the variability among student achievements explained at class level is much higher that the one explained at school level. Moreover, we find that this proportion changes across geographical areas and across educational subjects, i.e. reading and mathematics. By means of parametric mixed-effects linear models, we estimate the school or class effect, that, in these cases, is the value-added that each school or class gives to the performances of its students. Among the fact that the class effect results to be stronger than the school one, an other relevant result presented in Masci et al. (2017a) is that the correlation between the school effects on reading and mathematics student achievements is positive and statistically significant, while the correlation between some effects at class level is null. This result suggests that the effect of the school is usually coherent on the students performances in the two school subjects, driven by certain school characteristics, school principal practices, school body composition and school peers that result to have a similar impact on both the reading and mathematics learning processes. On the other way, the fact that the correlation among class effects in reading and mathematics is null suggests that there is not a strong common effect of the class environment on student performances in different school subjects, but the effect of the class on the school subject is uncorrelated across school subjects. One of the most likely interpretation of this result is that the class effect, rather than on class body composition or peers, mainly depends on teacher practices, that might be strongly different between reading and mathematics.

In this chapter, exploiting the results in the educational context shown in Masci et al. (2016b, 2017a) and in the light of the potential of the SPEM algorithm presented in the previous chapter, we propose an extension of the SPEM algorithm that is innovative both from a methodological and an interpretative point of view. We develop the *multivariate* SPEM algorithm, i.e. a semi-parametric linear mixed-effects model with a multivariate response, and we apply it in a case study that faces the new issue of the identification of subpopulations of Italian classes, whose effect results to be stronger than the school one. Inspired by the fact that in Masci et al. (2016b, 2017a) emerges that the biggest part of unexplained variability among student performances is explained at class level, we precisely focus our attention on the class level. We are interested in estimating the impact that attending different classes has on student performance trends, i.e. student performance evolution over time, and, in particular,

in jointly modeling these effects in reading and mathematics. Driven by this purpose, we modify the semi-parametric mixed-effects linear model presented in the previous chapter to allow a multivariate response variable and we apply it to INVALSI data, considering students nested within classes, for a joint analysis of the Italian class effect on student achievement trends in reading and mathematics. The model that we propose is a multivariate two-level linear model where the coefficients of random effects, under non-parametric assumptions, follow a multivariate discrete distribution with an unknown number of mass points. Each group (observation of the second level of hierarchy) is assigned to a subpopulation of groups, that is represented by specific values of the parameters of the multivariate mixed-effects linear model. The distribution of the coefficients of random effects is a multivariate discrete distribution where each dimension is allowed to have a different finite number, unknown a priori, of mass points. This formulation permits to estimate the marginal distribution of the random effects related to each one of the multiple response variables and, moreover, to estimate the joint distribution of random effects related to the multiple response variables, investigating the correlation among them. Again, the great advantage of this formulation is that the model also estimates the number of latent subpopualations and, moreover, the flexibility of the multivariate model that allows the number of mass points of the marginal distributions of random effects to be different across the multiple response variables constitutes a further important plus. Together with the model, we propose the Expectation-Maximization (EM) algorithm to estimate its parameters.

This methodology is totally new to the literature. In Chapter 3, we state that the semi-parametric mixed-effects linear model, on which we base our multivariate model, enters in the research line about the identification of subpopulations of the Growth Mixture Models and of Latent Class Mixture Models, but with the novelty and the advantage that, contrarily to these existing methods, it does not need to fix a priori the number of latent subpopulations to be identified. Numerous extensions and applications of GMM and LCMM has been done (Lin et al., 2000; Muthén and Asparouhov, 2015), but none of them include the modeling of a multivariate answer variable, where the latent subpopulations structure of groups (higher level of hierarchy) are allowed to differ across the responses, i.e. are response-specific. This means that our method represents the unique extension to the multivariate case of a method that is already innovative by itself.

Our case study consists in the application of the multivariate semi-parametric two-level linear model to the INVALSI data, considering students as first level and classes as second one. The model that we propose aims at identifying a latent clustering structure of classes where, within each subpopulation, the effect of the classes on their student achievement trends across years are similar. The model estimates a bivariate effect for each class, i.e. the effect of the class on

mathematics student achievement trends and the one on the reading ones. The aim is to identify how many different trends exist in student performances across classes, for both mathematics and reading, i.e. to identify how many and which are the mass points of the distribution of random effects (class effects) for both the first and the second response. Moreover, by looking at the joint distribution of the random effects, we are able to investigate the correlation between the class effects on reading and mathematics.

The presence of subpopulations among Italian classes, related to each response variable, is the consequence of student performance trends that differ across classes. These differences might be due to different class body compositions, peers or class climates. Moreover, the fact that the subpopulations of classes differ between reading and mathematics might be due to something that is not class specific, but that is school subject specific, like different teacher practices. Since the year 2012/2013, INVALSI submits questionnaires to teachers about their personal information, their education, their teaching practices and the environment of the class and school in which they work, creating an informative and new dataset that, until now and in this context, has been poorly explored. In this perspective, in order to investigate whether the different student achievement trends across classes are related to these teacher characteristics, in a second stage of the analysis, we look for associations between class and teacher level characteristics and the identified subpopulations.

The work presented in this chapter is innovative for two aspects. First, it proposes a novel statistical method to perform in-built, unsupervised clustering of the higher level of grouping of a multivariate multilevel model, without knowing a priori the number of subpopulations. Second, the case study that we propose, that consists in identifying subpopulations of Italian classes standing on their effects on INVALSI student achievements, is able to face an interesting research question, never addressed before.

The chapter is organized as follows: in Section 4.1 we present the multivariate semi-parametric two-level linear model and the EM algorithm to estimate its parameters (MSPEM algorithm) and we show a simulation study; in Section 4.2 we present the INVALSI 2016/2017 dataset, we apply the MSPEM algorithm to it, showing the results, and, in a second step, we characterize the identified subpopulations by means of class and teacher levels characteristics; in Section 4.3 we draw our conclusions.

4.1 Model, methods and simulation study

In this section, we present the multivariate semi-parametric mixed-effects linear model and the EM algorithm for the estimation of its parameters. For the sake of simplicity, we consider the case of a bivariate semi-parametric mixed-effects

linear model, i.e. the case of a response variable in $\mathcal{R}^2$, but the generalization to the case of a response variable in $\mathcal{R}^p$, for $p > 2$, is straightforward.

4.1.1 Bivariate semi-parametric mixed-effects model

Consider a bivariate two-level linear model, where each bivariate observation j, for $j = 1, \ldots, n_i$, is nested within a group i, for $i = 1, \ldots, N$. The model takes the following form:

$$\begin{aligned} \mathbf{Y}_i = \begin{pmatrix} \mathbf{y}_{1,i} \\ \mathbf{y}_{2,i} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \mathbf{X}_i + \begin{pmatrix} \mathbf{b}_{1,i} \\ \mathbf{b}_{2,i} \end{pmatrix} \mathbf{Z}_i + \boldsymbol{\epsilon}_i \qquad i = 1, \ldots, N \\ \boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i} \end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind. \end{aligned} \tag{4.1}$$

where i is the group index, N is the total number of groups, n_i is the number of bivariate observations within the i-th group and $J = \sum_{i=1}^{N} n_i$ is the total number of bivariate observations[1]. The components of model (4.1) are the following:

- $\mathbf{Y}_i = \begin{pmatrix} y_{1,1i}, \ldots, y_{1,n_i i} \\ y_{2,1i}, \ldots, y_{2,n_i i} \end{pmatrix}$ is the $2 \times n_i$-dimensional matrix of response variable within the i-th group[2],
- $\mathbf{X}_i$ is the $(p+1) \times n_i$-dimensional matrix of covariates of fixed effects,
- $\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}$ is the $2 \times (p+1)$-dimensional matrix of coefficients of $\mathbf{X}$,
- $\mathbf{Z}_i$ is the $(r+1) \times n_i$-dimensional matrix of covariates of random effects,
- $\mathbf{b}_i = \begin{pmatrix} \mathbf{b}_{1,i} \\ \mathbf{b}_{2,i} \end{pmatrix}$ is the $2 \times (r+1)$-dimensional matrix of coefficients of $\mathbf{Z}_i$,
- $\boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i} \end{pmatrix}$ is the $2 \times n_i$-dimensional matrix of errors and $\boldsymbol{\Sigma}$ is its variance/covariance matrix.

Fixed effects are identified by parameters associated to the entire population, while random ones are identified by group-specific parameters. In the parametric framework of bivariate linear mixed-effects models, the coefficients of random effects are assumed to be distributed according to a Normal distribution with mean vector equal to $\mathbf{0}$ and a variance/covariance matrix that is estimated,

[1] In subscript of each variable, we indicate by the number before the coma whether the variable is referred to the first or the second response variable (for example, $y_{1,ij}$ and $y_{2,ij}$ are the j-th first and second response variables within group i, respectively).

[2] We consider the case in which the number of observations of the two response variables is the same within each group, but is allowed to be different across the groups.

together with the other parameters of the model, through methods based on the maximization of the likelihood or the restricted likelihood functions (Pinheiro and Bates, 2000). This parametric distribution implies that, for each group i, the model estimates the coefficients $\mathbf{b}_i = (b_{i1}, \ldots, b_{i(r+1)})$ for the $(r+1)$ covariates of the random effects, meaning that the covariates of random effects are allowed to have N different associations to the response variables across the N groups.

Following the idea presented in the previous chapter, we relax the parametric assumptions about the coefficients of the random effects and we assume the bivariate coefficients $\mathbf{b}_i$ to follow a bivariate discrete distribution P^*, assuming $M \times K$ mass points $(\mathbf{c}_{11}, \ldots, \mathbf{c}_{MK})$, where each $\mathbf{c}_{mk}$ is the $2 \times (r+1)$-dimensional matrix of coefficients of random effects for the bivariate mass point related to the index (m, k), for each $m = 1, \ldots, M$ and $k = 1, \ldots, K$, where both M and K are smaller than N. The total number of mass points, that is $M \times K$, is unknown a priori and it is estimated together with the other parameters of the model. This modeling allows the identification of a bivariate clustering distribution among the N groups, where each group i is associated to a subpopulation, standing on the linear relationships between the two response variables and their covariates. In other words, the model identifies a bivariate latent structure among the groups, that also reveals the dependence among the two response variables. Under these assumptions, the semi-parametric bivariate mixed-effects model takes the following form[3]:

$$\begin{aligned} \mathbf{Y}_i = \begin{pmatrix} \mathbf{y}_{1,i} \\ \mathbf{y}_{2,i} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \mathbf{X}_i + \begin{pmatrix} \mathbf{c}_{1,m} \\ \mathbf{c}_{2,k} \end{pmatrix} \mathbf{Z}_i + \boldsymbol{\epsilon}_i \qquad & \begin{matrix} i = 1, \ldots, N \\ m = 1, \ldots, M \\ k = 1, \ldots, K \end{matrix} \\ \boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i} \end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind. \end{aligned} \tag{4.2}$$

Without loss of generality, we consider the case of a semi-parametric bivariate two-level linear model, with one random intercept, one random covariate and one fixed covariate[4]. Model (4.2) reduces to:

[3] Also for the parameters, we indicate by $\mathbf{c}_{\mathbf{1},*}$ and $\mathbf{c}_{\mathbf{2},*}$ the coefficients related to the first and second response variables, respectively.

[4] This choice is driven by the application in the case study shown in Section 4.2.

$$\mathbf{Y}_i = \begin{pmatrix} \mathbf{y}_{1,i} \\ \mathbf{y}_{2,i} \end{pmatrix} = \begin{pmatrix} c_{1,1m} \\ c_{2,1k} \end{pmatrix} \mathbf{1} + \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \mathbf{x}_i + \begin{pmatrix} c_{1,2m} \\ c_{2,2k} \end{pmatrix} \mathbf{z}_i + \boldsymbol{\epsilon}_i \qquad \begin{matrix} i = 1, \dots, N \\ m = 1, \dots, M \\ k = 1, \dots, K \end{matrix}$$

$$\boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i} \end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind. \tag{4.3}$$

where $\mathbf{1}$ is the n_i-dimensional vector of 1, M is the total number of mass points for the first response and K is the total number of mass points for the second response and both of them are not known a priori. Coefficients $\mathbf{c}_{mk}$, for $m = 1, \dots, M$ and $k = 1, \dots, K$ are distributed according to a discrete probability measure P^* that belongs to the class of all probability measures on $\mathcal{R}^4$. P^* can then be interpreted as the mixing distribution that generates the density of the stochastic model in (4.3). The ML estimator $\hat{P}^*$ of P^* can be obtained following the theory of mixture likelihoods in Lindsay et al. (1983a,b), as explained in the previous chapter. The ML estimator of the random effects distribution can be expressed as a set of points $(\mathbf{c}_{11}, \dots, \mathbf{c}_{MK})$ and a set of wights $(w_{11}, \dots, w_{MK})$, where $\sum_{m=1}^{M} \sum_{k=1}^{K} w_{mk} = 1$ and $w_{mk} \geq 0$, for $m = 1, \dots, M$ and $k = 1, \dots, K$. Each group i, for $i = 1, \dots, N$, is assigned to a subpopulation (m, k), standing on the fact that the first response belongs to subpopulation m and the second one to subpopulation k. Indeed, the marginal distribution given by $(c_{1,1}, \dots, c_{1,M})$ and $(w_{1,1}, \dots, w_{1,M})$ represents the first response-specific latent structure among groups, while the marginal distribution given by $(c_{2,1}, \dots, c_{2,K})$ and $(w_{2,1}, \dots, w_{2,K})$ represents the second response-specific one. The estimation of the parameters $\boldsymbol{\beta}$, $(\mathbf{c}_{11}, \dots, \mathbf{c}_{MK})$, $(w_{11}, \dots, w_{MK})$ and $\boldsymbol{\Sigma}$ is performed through the maximization of the likelihood function , mixture by the discrete distribution of random effects,

$$\begin{aligned} L(\boldsymbol{\beta}, \mathbf{c}_{mk}, \boldsymbol{\Sigma} | \mathbf{y}) = & \sum_{m=1}^{M} \sum_{k=1}^{K} \frac{w_{mk}}{\sqrt{|det(2\pi\boldsymbol{\Sigma})|^J}} \times \\ & \times \exp \left\{ \sum_{i=1}^{N} \sum_{j=1}^{n_i} -\frac{1}{2} \begin{pmatrix} y_{1,ij} - c_{1,1m} - \beta x_{1,ij} - c_{1,2m} z_{1,ij} \\ y_{2,ij} - c_{2,1k} - \beta x_{2,ij} - c_{2,2k} z_{2,ij} \end{pmatrix}^T \right. \\ & \left. \boldsymbol{\Sigma}^{-1} \begin{pmatrix} y_{1,ij} - c_{1,1m} - \beta x_{1,ij} - c_{1,2m} z_{1,ij} \\ y_{2,ij} - c_{2,1k} - \beta x_{2,ij} - c_{2,2k} z_{2,ij} \end{pmatrix} \right\} \end{aligned} \tag{4.4}$$

with respect to $\boldsymbol{\beta}$, the distribution of the coefficients of random effects $(\mathbf{c}_{mk}, w_{mk})$, for $m = 1, \dots, M$ and $k = 1, \dots, K$, and $\boldsymbol{\Sigma}$, respectively.

4.1.2 The MSPEM algorithm

The EM algorithm that we propose to estimate the parameters of the model in (4.3) is the generalization for the bivariate case of the SPEM algorithm, presented in the previous chapter. It alternates two steps: the expectation step (E step) in which we compute the conditional expectation of the likelihood function with respect to the random effects, given the observations and the parameters computed in the previous iteration; and the maximization step (M step) in which we maximize the conditional expectation of the likelihood function. At each iteration, the EM algorithm updates the parameters in order to increase the likelihood in Eq. (4.4) and it continues until the convergence. The update of the parameters is the following:

$$w_{mk}^{(up)} = \frac{1}{N}\sum_{i=1}^{N} W_{imk} \qquad \text{for } m = 1,\dots,M, \quad k = 1,\dots,K \tag{4.5}$$

and

$$\begin{aligned}(\boldsymbol{\beta}^{(up)}, \mathbf{c}_{mk}^{(up)}, \boldsymbol{\Sigma}^{(up)}) = \\ \underset{\beta,\mathbf{c}_{mk},\boldsymbol{\Sigma}}{\arg\max} \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{i=1}^{N} W_{imk}\, \ln p(\mathbf{y}_i|\boldsymbol{\beta},\boldsymbol{\Sigma},\mathbf{c}_{mk})\end{aligned} \tag{4.6}$$

where

$$W_{imk} = \frac{w_{mk}\, p(\mathbf{y}_i|\boldsymbol{\beta},\boldsymbol{\Sigma},\mathbf{c}_{mk})}{\sum_{m=1}^{M}\sum_{k=1}^{K} w_{mk}\, p(\mathbf{y}_i|\boldsymbol{\beta},\boldsymbol{\Sigma},\mathbf{c}_{mk})} \tag{4.7}$$

and

$$\begin{aligned}p(\mathbf{y}_i|\beta,\boldsymbol{\Sigma},\mathbf{c}_{mk}) =& \frac{1}{\sqrt{|det(2\pi\boldsymbol{\Sigma})|^{n_i}}}\times \\ &\times \exp\left\{\sum_{j=1}^{n_i} -\frac{1}{2}\begin{pmatrix} y_{1,ij} - c_{1,1m} - \beta x_{1,ij} - c_{1,2m} z_{1,ij} \\ y_{2,ij} - c_{2,1k} - \beta x_{2,ij} - c_{2,2k} z_{2,ij}\end{pmatrix}^T \right. \\ &\left. \boldsymbol{\Sigma}^{-1}\begin{pmatrix} y_{1,ij} - c_{1,1m} - \beta x_{1,ij} - c_{1,2m} z_{1,ij} \\ y_{2,ij} - c_{2,1k} - \beta x_{2,ij} - c_{2,2k} z_{2,ij}\end{pmatrix}\right\}.\end{aligned} \tag{4.8}$$

The coefficient W_{imk} represents the probability of $\mathbf{b}_i$ being equal to $\mathbf{c}_{mk}$ conditionally to observations $\mathbf{y}_i$ and given the fixed coefficient $\boldsymbol{\beta}$ and the variance/covariance matrix $\boldsymbol{\Sigma}$. Indeed, since $w_{mk} = p(\mathbf{b}_i = \mathbf{c}_{mk})$, then

$$W_{imk} = \frac{w_{mk}\, p(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{c}_{mk})}{\sum_{m=1}^{M}\sum_{k=1}^{K} w_{mk}\, p(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{c}_{mk})} = \frac{p(\mathbf{b}_i = \mathbf{c}_{mk})\, p(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{c}_{mk})}{p(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma})} =$$
$$= \frac{p(\mathbf{y}_i, \mathbf{b}_i = \mathbf{c}_{mk}|\boldsymbol{\beta}, \boldsymbol{\Sigma})}{p(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma})} = p(\mathbf{b}_i = \mathbf{c}_{mk}|\mathbf{y}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}). \tag{4.9}$$

Therefore, in order to compute the point $\mathbf{c}_{mk}$ for each group i, for $i = 1, \ldots, N$, we maximize the conditional probability of $\mathbf{b}_i$ given the observations $\mathbf{y}_i$, the coefficient $\boldsymbol{\beta}$ and the error variance/covariance matrix $\boldsymbol{\Sigma}$. So that, the estimation of the coefficients $\mathbf{b}_i$ of the random effects for each group i is obtained maximizing W_{imk} over m and k, that is

$$\hat{\mathbf{b}}_i = \mathbf{c}_{\tilde{mk}} \quad \text{where} \quad \tilde{mk} = \arg\max_{m,k} W_{imk} \quad i = 1, \ldots, N. \tag{4.10}$$

The maximization in Eq. (4.6) involves two steps and it is done iteratively. In the first step, we compute the *arg-max* with respect to the support points $\mathbf{c}_{mk}$, keeping $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ fixed to the last computed values. In this way, we can maximize the expected log-likelihood with respect to all support points $\mathbf{c}_{mk}$ separately, that means

$$\mathbf{c}_{mk}^{(up)} = \arg\max_{\mathbf{c}} \sum_{i=1}^{N} W_{imk}\, \ln p(\mathbf{y}_i|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \mathbf{c}_{mk}) \qquad m = 1, \ldots, M$$
$$k = 1, \ldots, K. \tag{4.11}$$

Since we are considering the linear case, the maximization step is done in closed-form[5]. In the second step, we fix the support points of the random effects distribution computed in the previous step and we compute the *arg-max* in Eq. (4.6) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. Again, this step is done in closed-form.

The initialization of the support points of the discrete distribution P^* and the criteria for the convergence of the EM algorithm are the direct extension of the ones chosen in the SPEM algorithm for the bivariate case. In particular, the algorithm starts considering N support points for the coefficients of random effects and a starting estimate for the coefficient of the fixed effects, for both the response variables. These parameters are chosen in the following way:

- random effects: for each response variable, the starting N support points are obtained fitting a simple linear regression within each group and estimating the couple of parameters (both the intercept and the slope) for each one of the N groups. The weights are uniformly distributed on these $N \times N$ support points;

[5]Closed-form calculations of model parameters can be found in the previous chapter.

- fixed effects: the starting values of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are estimated by fitting a unique bivariate linear regression on the entire population (i.e. without considering the nesting of the observations within groups).

Nonetheless, if the number of starting support points N is extremely large, the algorithm is relatively slow and using N starting support points becomes not strictly necessary. In this case, the initialization of the support points of the random effects distribution is done in the following way:

- we choose a number N* < N of support points, that is the same for both the two response variables;
- for each response variable, we extract N* points from a uniform distribution with support on the entire range of possible values for each parameter, that is estimated by fitting N distinct linear regressions for each one of the N groups, as before, and identifying the minimum and the maximum values;
- we uniformly distribute the weights on these $N^* \times N^*$ support points.

The $M \times K$ matrix of weights, that is composed by the elements w_{mk} previously described, represents the joint distribution of groups across the subpopulations and, by summing over rows and columns respectively, it represents the marginal distribution of the groups across the subpopulations, for each single response variable.
During the iterations, the EM algorithm performs the support reduction of the discrete distribution of random effects, in order to identify $M \times K$ mass points (starting from $N \times N$ mass points), where both M and K are smaller than N. The support reduction is made standing on two criteria. The former is that we fix a threshold value D and if two mass points are closer, in terms of euclidean distance, than D, they collapse to a unique point. This procedure is separately applied to the subpopulations related to the first and second response variable respectively. In particular, considering, for example, the case of the first response variable, if two mass points $\mathbf{c}_{1,h}$ and $\mathbf{c}_{1,g}$, for $h, g = 1, \ldots, M$, are closer than D, they collapse to a unique point $\mathbf{c}_{1,(hg)}$, where $\mathbf{c}_{1,(hg)} = \frac{\mathbf{c}_{1,h}+\mathbf{c}_{1,g}}{2}$. Consequently, $M^{new} = M^{old} - 1$, the new marginal weight is obtained as $w_{1,(hg)} = w_{1,h} + w_{1,g}$ and the joint weights $w_{(hg)k} = w_{hk} + w_{gk}$, for $k = 1, \ldots, K$. The same criterion applies to the subpopulations related to the second response variable. The first two masses collapsing to a unique point are the two masses with the minimum euclidean distance, among the couples of masses with euclidean distance less than D, and so on so forth. Note that the threshold parameter D can be settled equal to different values D_1 and D_2 when considering the first and the second response variable respectively. Anyway, even if D_1 is equal to D_2, the procedure might lead to different number of mass points M and K. The latter is that, starting from a given iteration up to the end, we fix a threshold value $\tilde{w}$ and

we remove mass points with marginal weights $w_{1,m} \leq \tilde{w}$, for $m = 1, \ldots, M$ and $w_{2,k} \leq \tilde{w}$, for $k = 1, \ldots, K$ or that are not associated to any subpopulation. Again, $\tilde{w}$ can be settled equal to different values $\tilde{w}_1$ and $\tilde{w}_2$when considering the first and the second response variable respectively. D and $\tilde{w}$ are two tuning parameters that tune the estimates of the subpopulations. Further insights on the choice of these parameters are discussed in the previous chapter.

The sketch of the MSPEM algorithm is shown in Algorithm 2 in Appendix. At each iteration a, the algorithm, given the estimated number of mass points, estimates all the parameters in Eq. (4.3) in an iterative way, updating the coefficients related to both fixed and random effects, until convergence or until it reaches the maximum number of sub-iterations fixed a priori for this stage (`itmax`). At the beginning of the iterative process, the algorithm performs the dimensional reduction of the mass points standing only on the distance D between the mass points. When the estimates are stable, meaning that all the differences between the estimates of the parameters at two consecutive iterations are smaller than fixed tolerance values, or after a given number of iterations `it1`, the algorithm continues performing the dimensional reduction of the support points standing also on the criterion of the minimum weight $\tilde{w}$. The final convergence is reached when all the differences between the estimates of the parameters at two consecutive iterations are smaller than fixed tolerance values, or after a given number of iterations `it`. In particular, we fix the tolerance values for the estimates of both the parameters of fixed and random effects to `tollF` and `tollR` respectively, which depend on the scale of the parameters. The usage of the maximum number of iterations `it`, `it1` and `itmax` is merely to avoid an infinite loop and their values depend on the complexity of the data and on the consequent convergence rate.

4.1.3 Simulation study

In this section, we test the performance of the MSPEM algorithm simulating four situations in which the two response variables are related to each other in four different ways, facing both structural correlation/uncorrelation between the subpopulations distribution and correlation/uncorrelation between the errors of the linear model.

We generate $1,000$ bivariate observations that are nested within 100 groups in the following way:

$$\begin{pmatrix} \mathbf{y}_{1,i} \\ \mathbf{y}_{2,i} \end{pmatrix} = \begin{pmatrix} c_{1,1m} \\ c_{2,1k} \end{pmatrix} + \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \mathbf{x}_i + \begin{pmatrix} c_{1,2m} \\ c_{2,2k} \end{pmatrix} \mathbf{z}_i + \boldsymbol{\epsilon}_i \qquad \begin{matrix} i = 1, \dots, 100 \\ m = 1, \dots, M \\ k = 1, \dots, K \end{matrix}$$

$$\boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i} \end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind. \tag{4.12}$$

in which we set $M = 3$ and $K = 2$. Without loss of generality, we set $n_i = 100$, for $i = 1, \dots, 100$, and we make the following choice of parameters[6] $\mathbf{c}_{mk}$, for $m = 1, \dots, 3$ and $k = \{1, 2\}$:

	First response parameters	Second response parameters
$i = 1, \dots, 33$	$c_{1,11} = 5$ $c_{1,21} = 10$ $\beta_1 = 3$	$c_{2,11} = 3$ $c_{2,21} = 1$ $\beta_2 = 2$
$i = 34, \dots, 66$	$c_{1,12} = 2$ $c_{1,22} = 5$ $\beta_1 = 3$	$c_{2,11} = 3$ $c_{2,21} = 1$ $\beta_2 = 2$
$i = 67, \dots, 100$	$c_{1,13} = 0$ $c_{1,23} = -2$ $\beta_1 = 3$	$c_{2,12} = 0$ $c_{2,22} = -3$ $\beta_2 = 2$

Table 4.1: Coefficients used to simulate data in Eq. (4.12). The intercepts and the coefficients of $\mathbf{z}$ differ across subpopulations, while the coefficients of x (β) are fixed. Colors highlight the different subpopulations related to each response variable. We impose a structure with three subpopulations in the first response (M=3) and two subpopulations in the second one (K=2).

Besides the coefficients, we sample the observations of the variables $\mathbf{x}$, $\mathbf{z}$ and $\boldsymbol{\epsilon}$ in the following way[7]:

$$\begin{aligned} \mathbf{z}_i &\sim \mathcal{N}(0.10, 0.4^2) \quad i = 1, \dots, 33 \\ \mathbf{z}_i &\sim \mathcal{N}(0.12, 0.4^2) \quad i = 34, \dots, 66 \\ \mathbf{z}_i &\sim \mathcal{N}(0.08, 0.4^2) \quad i = 67, \dots, 100 \end{aligned} \tag{4.13}$$

[6] Note that this choice of parameters is finalized to the simulation study and it is driven only from the aim of a simple and clear visualization of the results. Any other choice of parameters is possible.

[7] Again, different choices of values for variables $\mathbf{x}$ and $\mathbf{z}$ are possible and they are also allowed to be different between first and second response variables (i.e. $\mathbf{x}_{1,i} \neq \mathbf{x}_{2,i}$).

$$\begin{aligned} \mathbf{x}_i &\sim \mathcal{N}(0.30, 0.4^2) \quad i = 1, \ldots, 33 \\ \mathbf{x}_i &\sim \mathcal{N}(0.28, 0.4^2) \quad i = 34, \ldots, 66 \\ \mathbf{x}_i &\sim \mathcal{N}(0.27, 0.4^2) \quad i = 67, \ldots, 100 \end{aligned} \tag{4.14}$$

and

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}_2\Big(\mathbf{0}, \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\Big) \quad i = 1, \ldots, 100. \tag{4.15}$$

Since we choose three different sets of parameters $(\mathbf{c}, \boldsymbol{\beta})$ to generate the data of the first response and two different sets to generate the ones of the second response, the data related to the first response are clustered within three subpopulations (M=3), while the ones related to the second one are clustered within two subpopulations (K=2). Figure 4.1 shows the data simulated with the set of parameters in Table 4.1.

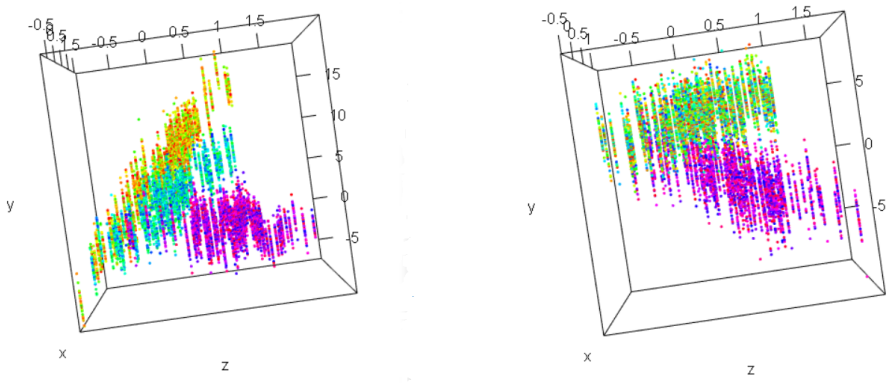


Figure 4.1: Data simulated with the set of parameters shown in Table 4.1 and values of $\mathbf{x}$, $\mathbf{z}$ and $\boldsymbol{\epsilon}$ shown in Eq. (4.13), (4.14) and (4.15) respectively. Figure on the left panel represents the first response and figure on the right panel represents the second one. It is possible to identify the presence of three and two subpopulations in the first and in the second response respectively. Colors are automatically assigned by the software R.

The eventual correlation among the two response variables depends on the subpopulations distribution that we use to generate them (i.e. on the choice of $\mathbf{c}_{mk}$) and on the correlation/uncorrelation between the errors. In this perspective, the parameters distribution shown in Table 4.1 imposes a structural correlation among the subpopulations of the two response variables, since the bivariate distribution of $\mathbf{c}_{mk}$ follows a precise structure among the groups. Regarding the distribution of the errors, the covariance of the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ in Eq. (4.15) is set to zero, implying the absence of any further correlation among the two responses.

We apply the MSPEM algorithm 100 times to the simulated dataset, choosing $D_1 = D_2 = 1$, $\tilde{w}_1 = \tilde{w}_2 = 0.01$ and $\texttt{tollR} = \texttt{tollF} = 10^{-2}$. On average, the algorithm converges in 6 iterations and it always identifies the correct number of subpopulations for both the two response variables, whose estimated parameters, averaged on the 100 runs of the simulation, are shown in Table 4.2.

First response parameters				
	$\hat{c}_{1,1m}$	$\hat{c}_{1,2m}$	$\hat{\beta}_1$	$\hat{w}_1$
m=1	4.997	10.014		0.33
m=2	2.011	4.922	2.999	0.33
m=3	0.010	−2.023		0.34

Second response parameters				
	$\hat{c}_{2,1k}$	$\hat{c}_{2,2k}$	$\hat{\beta}_2$	$\hat{w}_2$
k=1	3.010	1.013	1.994	0.66
k=2	−0.007	−2.983		0.34

Table 4.2: Coefficients of Eq. (4.12) estimated by the MSPEM algorithm, averaged on the 100 runs of the simulation. Colors represent the different subpopulations identified by the algorithm. The algorithm identifies three subpopulations (M=3) for the first response and two subpopulations for the second one (K=2).

Figure 4.2 shows the data with the regression planes identified by the algorithm, for both the two response variables.

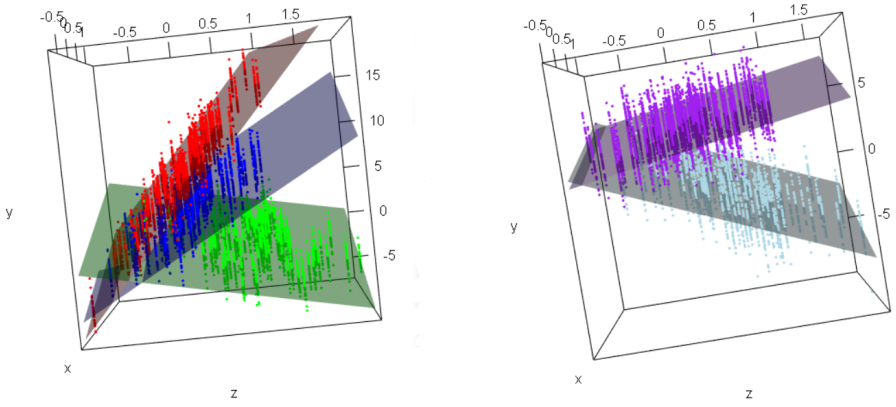


Figure 4.2: Simulated data with the regression planes identified by the MSPEM algorithm in one of the 100 runs. Colors represent the different subpopulations: three for the first response (figure on the left panel) and two for the second response (figure on the right panel). Coefficients of the regression planes are shown in Table 4.2.

The algorithm assigns the correct subpopulation to both the response variables of each group i, for $i = 1, \ldots, 100$, that means that assigns to each group i, for $i = 1, \ldots, 100$, the correct subpopulation (m, k). The estimates of the $(M \times K)$-dimensional matrix of weights W and of $\mathbf{\Sigma}$, averaged on the 100 runs, are the following:

$$\hat{W} = \begin{pmatrix} 0.33 & 0.00 \\ 0.33 & 0.00 \\ 0.00 & 0.34 \end{pmatrix} \quad \hat{\Sigma} = \begin{pmatrix} 1.016 & 0.001 \\ 0.001 & 0.969 \end{pmatrix}. \tag{4.16}$$

By looking at the matrix $\hat{W}$, we can identify the distribution of the groups on the support, composed by the 6 mass points (of which 3 have zero weight). Since we impose a structural correlation between the subpopulations distribution of the two response variables (see the coefficients in Table 4.1), the estimated distribution of the weights w_{mk} is not uniform on the $M \times K$ masses, but it is possible to recognize the pattern that we used to generate the data. Regarding the variance/covariance matrix $\hat{\mathbf{\Sigma}}$, the covariance is correctly estimated as null and the two estimated variances are also close to 1.

The case just shown represents only the particular situation in which the subpopulations distribution is not uniform on the mass points and the errors are not correlated, but it can also be the case that the two response variables do not present correlated subpopulations or even present correlated errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$. In order to test the performance of the MSPEM algorithm in these other cases, we modify the values of $\mathbf{c}_{mk}$ and $\boldsymbol{\epsilon}$ in order to simulate four different situations:

- Case 1: structural correlation among subpopulations of the two response variables and independence between the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ (case seen above);
- Case 2: structural correlation among subpopulations of the two response variables and dependence between the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$;
- Case 3 : not structural correlation among subpopulations of the two response variables and independence between the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$;
- Case 4: not structural correlation among subpopulations of the two response variables and dependence between the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$.

In order to not impose a structural correlation among the subpopulations of the two response variables (Case 3 and 4), i.e. in order to have a subpopulations distribution uniform on the mass points, we randomly shuffle the order of the parameters shown in Table 4.1 across the 100 groups, so that there are no definite patterns on the parameters c_{mk} between the two responses. In order to impose the dependence among the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ (Case 2 and 4), we set the covariance

of the variance/covariance matrix $\mathbf{\Sigma}$ equal to 0.5. In particular, we set $\Sigma = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}$.

We apply the MSPEM algorithm 100 times to these four different types of simulated dataset, with the same choice of parameters $D_1 = D_2 = 1$, `tollR` $=$ `tollF` $= 10^{-2}$. The algorithm is able to identify the correct subpopulations distribution in all the four situations. The visualizations of the results in all the four cases are similar to the one shown in Figure 4.2 and the estimates of the parameters $\mathbf{c}_{m,k}$, for $m = 1, \ldots, 3$ and $k = 1, 2$ and $\boldsymbol{\beta}$ in the four cases are in line with the ones shown in Table 4.2. What changes across the four cases are the estimates of the weights matrices W and of $\mathbf{\Sigma}$, that are shown in Table 4.3. The weights matrix W that we use to generate the data in the case of the not structural correlation among subpopulations are $W = \begin{pmatrix} 0.25 & 0.08 \\ 0.21 & 0.12 \\ 0.20 & 0.14 \end{pmatrix}$ and $W = \begin{pmatrix} 0.23 & 0.10 \\ 0.21 & 0.12 \\ 0.22 & 0.12 \end{pmatrix}$ for $\boldsymbol{\epsilon}_1 \not\perp\!\!\!\perp \boldsymbol{\epsilon}_2$ and $\boldsymbol{\epsilon}_1 \perp\!\!\!\perp \boldsymbol{\epsilon}_2$ respectively.

	Structural correlation among subpopulations	Not structural correlation among subpopulations
$\boldsymbol{\epsilon}_1 \not\perp\!\!\!\perp \boldsymbol{\epsilon}_2$	$\hat{W} = \begin{pmatrix} 0.33 & 0.00 \\ 0.33 & 0.00 \\ 0.00 & 0.34 \end{pmatrix}$	$\hat{W} = \begin{pmatrix} 0.25 & 0.08 \\ 0.21 & 0.12 \\ 0.20 & 0.14 \end{pmatrix}$
	$\hat{\Sigma} = \begin{pmatrix} 0.506 & 0.506 \\ 0.506 & 0.507 \end{pmatrix}$	$\hat{\Sigma} = \begin{pmatrix} 0.499 & 0.499 \\ 0.499 & 0.499 \end{pmatrix}$
$\boldsymbol{\epsilon}_1 \perp\!\!\!\perp \boldsymbol{\epsilon}_2$	$\hat{W} = \begin{pmatrix} 0.33 & 0.00 \\ 0.33 & 0.00 \\ 0.00 & 0.34 \end{pmatrix}$	$\hat{W} = \begin{pmatrix} 0.23 & 0.10 \\ 0.22 & 0.12 \\ 0.21 & 0.12 \end{pmatrix}$
	$\hat{\Sigma} = \begin{pmatrix} 1.016 & 0.001 \\ 0.001 & 0.969 \end{pmatrix}$	$\hat{\Sigma} = \begin{pmatrix} 0.993 & 0.009 \\ 0.009 & 1.023 \end{pmatrix}$

Table 4.3: Estimates of the weights matrix W and of the variance/covariance matrix $\mathbf{\Sigma}$ of model in Eq. (4.12) for the four different cases of values of $\mathbf{c}_{mk}$ and $\boldsymbol{\epsilon}$.

From Table 4.2, we see that the model is identifiable, it is able to distinguish the correlation among the two response variables that is given by a structural correlation among subpopulations distribution (showed in W) from the correlation imposed by dependent errors (showed in $\mathbf{\Sigma}$). In cases 3 and 4 (last column in Table 4.2), where we do not impose a structural correlation among subpop-

ulations, the distribution of the weights , less than small variations, uniformly distributed on the mass points.is

The only parameter that significantly influences the results of the MSPEM algorithm is the threshold distance D. In order to give an idea of the sensitivity of the algorithm to the values of D, in the cases seen above, the algorithm gives the same result for each value of D_1 and D_2 between 0.5 and 2. For values of D_1 and D_2 less than 0.5, the MSPEM algorithm is too sensitive to the variability among the data and identifies more that 6 mass points, while for values of D_1 and D_2 bigger than 2, the algorithm does not entirely catch the variability among the data and it identifies less than 6 mass points.

4.2 Case study: application of the MSPEM algorithm to INVALSI data

In this section, we present the INVALSI 2016/2017 dataset and we apply the MSPEM algorithm to it in order to identify subpopulations of classes, standing on their different effects on mathematics and reading student achievements.

4.2.1 The INVALSI 2016/2017 dataset

During the academic year 2016/2017, INVALSI tested Italian students at grades II and V of primary school (grade 2 and 5 respectively), at grade III of junior secondary school (grade 8) and at grade II of upper secondary school (grade 10), both in their skills in reading and mathematics. In this case study, we consider students attending grade III of junior secondary school in year 2016/2017. About these students, besides their results of the INVALSI tests in reading and mathematics at grade 8 (`read8` and `math8` respectively), we consider other three variables: the socio-economic index (`ESCS`) that is an index built by INVALSI by considering parents' occupation and educational titles and the possession of certain goods at home (for instance, computer or the number of books); the INVALSI test scores in reading and mathematics of these students three years before, i.e. at the last year of primary school (`read5` and `math5` respectively). The INVALSI test score is a continuous variable that takes values between 0 and 100, while the ESCS is built as a continuous variable with mean equal to 0 and variance equal to 1. Table 4.4 reports the five student level variables at student level used in the analysis with their descriptive statistics.
In addition to the information at student level, INVALSI collects information about the class and the teachers by means of a questionnaire. This questionnaire collects information about the class body composition and the approach of the teachers to INVALSI tests, personal information of teachers (age, education, gender), teaching practices and available materials in the class. Among the

Variable	Mean	sd	Median	IQR
math8	53.2001	20.036	52.489	29.322
read8	64.491	17.278	66.392	23.001
math5	68.475	16.641	70.000	26.001
read5	66.608	16.736	68.965	24.138
ESCS	0.1473	0.991	0.069	1.323

Table 4.4: Student level variables of the INVALSI database used in the analysis with their explanation.

entire set of variables, we select for the analysis the ones that are reported in Tables 4.5 and 4.6.

Variable	Type	Explanation
Teachers general questions (for both maths and reading teachers)		
`updated techniques`	y/n	the teacher applies new techniques learned at refreshment courses
`team work or research`	y/n	the teacher organizes team work or research in groups for students
`extra activities`	y/n	the teacher organizes extra scholastic activities for student reinforcement
`computer/internet`	y/n	the teacher uses media support in class
`refresher courses`	num	number of refreshment courses the teacher had in the last two years
`contacts among teachers`	y/n	teacher exchanges views with other teachers
Teacher's personal information (for both maths and reading teachers)		
`num years of teaching here`	$1:4$	since how many years the teacher teaches in the actual school. 1: one year or less; 2: 2-3 years; 3: 4-5 years; 4: $>$ than 5 years.
`permanent job`	y/n	the teacher has a permanent contract
`gender`	y/n	y= male; n = female.
`age`	num	age of the teacher
`education`	$1:3$	higher level of education of the teacher 1: less than degree; 2: degree; 3: phd/master
Questions about school principals (for both maths and reading teachers)		
`princ refreshment courses`	y/n	the school principal encourages teachers to follow refreshment courses
`princ lineup teach`	y/n	the school principal organizes lineup meetings for teachers
`princ evaluate`	y/n	the school principal evaluates the teachers in their job
Only for mathematics teachers		
`num mathematics hours`	num	number of hours of maths lesson per week
`main teaching method`	cat	'a': teach definitions and theorems so that students can apply to new problems 'b': favor the maths language and the capacity of using formulas written in symbols 'c': favor meanings of maths symbols 'd': favor the capacity of build concepts, models and theories 'e': follow only the book
`oral individ exam`	y/n	the teacher tests students by means of oral individual exams
`oral group exam`	y/n	the teacher tests students by means of oral exams for groups of students

Table 4.5: Teacher and class level variables of the INVALSI database used in the analysis with their explanation.

Variable	Type	Explanation
`teacher written exam`	y/n	the teacher tests students by means of written exam made by him/herself
`book written exam`	y/n	the teacher tests students by means of written exam taken by the book
calculations alone	y/n	the teacher teaches students to make calculations without the support of the calculator
`table diagram graph`	y/n	the teacher teaches students to interpret tables, diagrams and graphs
`maths memory`	y/n	the teacher asks students to memorize maths rules and theorems
`graphs for problems`	y/n	the teacher teaches students to analyze graphs to solve maths problems
Only for reading teachers		
`num reading hours`	num	number of hours of reading lesson per week
`programmed oral exam`	y/n	the teacher tests students by means of programmed oral exam
`not programmed oral exam`	y/n	the teacher tests students by means of not programmed oral exam
`grouped oral exam`	y/n	the teacher tests students by means of oral exam for groups of students
`teacher close test`	y/n	the teacher tests students by means of written close questions tests made by him/herself
`teacher open test`	y/n	the teacher tests students by means of written open questions tests made by him/herself
`teacher book test`	y/n	the teacher tests students by means of written tests taken by the book
`summarize text`	y/n	the teacher trains students to summarize texts
`write reflections`	y/n	the teacher trains students to write texts about their reflections and thinking
`read newspaper`	y/n	the teacher trains students to read newspapers and journals
Class information and body composition		
`area geo`	cat	Northern/Central/Southern Italy
`Nstud`	num	number of students
`% stud antic`	num	percentage of early-enrolled students
`% stud postic`	num	percentage of late-enrolled students
`% stud S1`	num	percentage of first generation immigrants
`% stud S2`	num	percentage of second generation immigrants

Table 4.6: Teacher and class level variables of the INVALSI database used in the analysis with their descriptive statistics.

We include in the analysis only classes that have at least ten students. The dataset of interest regards $18,242$ students nested within $1,082$ classes.

4.2.2 MSPEM algorithm applied to INVALSI data

The semi-parametric two-level linear model applied to INVALSI data, considering students (level 1) nested within classes (level 2), takes the following form:

$$\mathbf{Y}_i = \begin{pmatrix} c_{1,1m} \\ c_{2,1k} \end{pmatrix} \mathbf{1} + \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \mathbf{x}_i + \begin{pmatrix} c_{1,2m} \\ c_{2,2k} \end{pmatrix} \mathbf{z}_i + \boldsymbol{\epsilon}_i \qquad \begin{matrix} i = 1, \dots, N \\ m = 1, \dots, M \\ k = 1, \dots, K \end{matrix}$$

$$\boldsymbol{\epsilon}_i = \begin{pmatrix} \boldsymbol{\epsilon}_{1,i} \\ \boldsymbol{\epsilon}_{2,i} \end{pmatrix} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}) \qquad ind. \tag{4.17}$$

where i is the class index and N is the total number of classes. $\mathbf{Y}_i = \begin{pmatrix} \texttt{math8}_i \\ \texttt{read8}_i \end{pmatrix}$ is the bivariate vector of the INVALSI test scores of students attending grade 8, in mathematics and reading, $\mathbf{x}$ is the `ESCS` and $\mathbf{z}$ is the INVALSI test score of the same students three years before (at grade 5), that differs across the two response variables, being `math5` for the first response variable (`math8`) and `read5` for the second one (`read8`). In particular, we standardize the variables `math8`, `read8`, `math5`, `read5` and `ESCS`, so that they all have mean equal to 0 and variance equal to 1. Our interest is to see how the association between the INVALSI test score at the end of the primary school/beginning of the junior secondary school and the INVALSI test score at the end of the junior secondary school does change across students attending different classes, adjusting for the socio-economic index, both in reading and mathematics. The period between grade 5 and grade 8 is the entire period of the junior secondary school and this association represents a kind of class effect, seen as the impact that the class has on the evolution of its student achievements. With this modeling, we identify subpopulations of classes within which class impacts are similar and across which they are different. The bivariate nature of the modeling allows to do that both for reading and mathematics achievements, considering also the joint effect of the class on the two school subjects. We apply the MSPEM algorithm with the following choice of parameters: $D_1 = D_2 = 0.5$, $\tilde{w}_1 = \tilde{w}_2 = 0.01$, `tollR` = `tollF` = 10^{-2}, `it`=40, `itmax`=20, `it1`=20. The algorithm converges in 20 iterations and identifies $M = 3$ mass points for the random effects distribution related to the first response (mathematics) and $K = 4$ mass points for the one related to the second response (reading). Estimates of the identified parameters are shown in Table 4.7.

In Table 4.7, $\hat{w}_1$ and $\hat{w}_2$ are the estimated weights related to the marginal distributions of the two random effects; $\hat{\beta}_1$ and $\hat{\beta}_2$ are the coefficients of fixed effects (variable `ESCS`) and therefore their estimates are stable across the subpopulations; $\hat{\mathbf{c}}_{1,m}$, for $m = 1, \dots, 3$ and $\hat{\mathbf{c}}_{2,k}$, for $k = 1, \dots, 4$ are the estimates of

First response variable				
	$\hat{c}_{1,1}$	$\hat{c}_{1,2}$	$\hat{w}_1$	$\hat{\beta}_1$
m=1	−1.652	0.113	0.037	
m=2	0.251	0.628	0.675	0.106
m=3	−0.354	0.387	0.288	
Second response variable				
	$\hat{c}_{2,1}$	$\hat{c}_{2,2}$	$\hat{w}_2$	$\hat{\beta}_2$
k=1	0.233	0.572	0.712	
k=2	−1.199	0.213	0.029	0.095
k=3	−0.270	0.405	0.239	
k=4	−2.696	−0.099	0.020	

Table 4.7: Estimates of the coefficients of Eq. (4.17) obtained by the MSPEM algorithm.

the coefficients of random effects. In order to visualize these results, Figure 4.3 reports the regression planes identified for both the two response variables, projected on the 2-dimensional plane identified by the answer variable and the random covariate.

By looking at the estimated parameters in Table 4.7 and the regression lines in Figure 4.3, it is possible to make considerations about the identified subpopulations of classes. Let start considering the results of mathematics (left panel of Figure 4.3). The three subpopulations are well identified and they almost do not cross, except for a small overlap between two subpopulations in correspondence of small values of z. Subpopulation 1 ($m = 1$ in Table 4.7), that contains 3.7% of the classes, can be interpreted as the subpopulation of the worst set of classes. Indeed, it is represented by a low intercept and a very low slope with respect to the others. This means that students in these classes have on average very low predicted score at grade 8, even if they had higher results at grade 5. On the other side, subpopulation $m = 2$, that contains 67.5% of the classes, represents the subpopulation of the best classes, since, for almost each value of previous score z, students within this subpopulation have the highest predicted value of y. Regarding the results of reading, the four identified subpopulations are also very well distinct. The subpopulation of the worst classes corresponds to subpopulation $k = 4$ (containing 2% of the classes), that is characterized by a very low intercept and a slightly negative slope: students attending classes that belong to this subpopulation have a low predicted value of INVALSI score, regardless of the fact that they had high or low scores at grade 5. On the opposite, subpopulation $k = 1$ (containing 71.2% of the classes) contains the set of the best classes since for all values of previous score z between -3 and 2, i.e. for

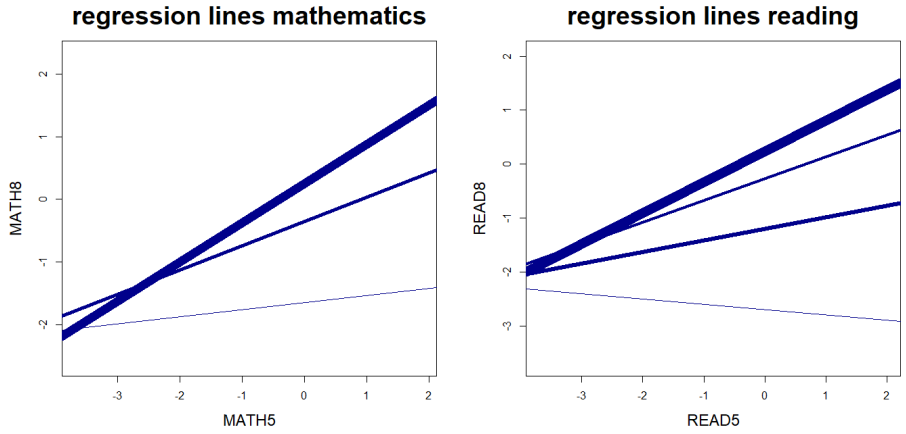


Figure 4.3: Regression planes projected on the 2-dimensional plane identified by the answer variable and the random covariate, identified by the parameters of Eq. (4.17) estimated by the MSPEM algorithm and whose parameters are shown in Table 4.7. Panel on the left reports the results for the first response, while panel on the right reports the results for the second one. The algorithm identifies $M = 3$ mass points for the first response and $K = 4$ mass points for the second one. For a better visualization, we do not represent all the observations but only the identified regression planes. Line widths are proportional to the marginal weights w_1 and w_2.

almost the entire range of values of the random covariate, the predicted value of y is higher that the ones of the other subpopulations of classes. Subpopulation $k = 3$ (containing 23.9% of the classes) is the second one in terms of high values of predicted score y, while subpopulation $k = 2$ (containing 2.9% of the classes) have predicted values of y lower than the ones of subpopulations $k = 1$ and $k = 3$ but higher than the ones of subpopulation $k = 4$.

The interpretations of these subpopulations are also supported by the average values of the standardizes variables across them, reported in Table 4.8. Regarding mathematics, subpopulation 1 contains classes where the average score of `math5` is the highest ($\overline{\texttt{math5}}_1 = 0.344$), but where the average score of `math8` is the lowest ($\overline{\texttt{math8}}_1 = -1.683$), confirming the negative effects of the classes that belong to this subpopulation. Subpopulation 2, interpreted as the subpopulation containing classes with the best positive effect, is characterized by the lowest average score of `math5` ($\overline{\texttt{math5}}_2 = -0.052$), but with the highest average score of `math8` ($\overline{\texttt{math8}}_2 = 0.223$). This subpopulation is the one with the highest average `ESCS` of students. Speaking about reading, subpopulation 1, interpreted as the one containing the best classes, is indeed characterized by the lowest average value of `read5` ($\overline{\texttt{read5}}_1 = -0.065$) and the highest average score of `read8` ($\overline{\texttt{read8}}_1 = 0.207$). Also here, this subpopulation is characterized by the highest average value of `ESCS`. On the other side, subpopulation 4, associated to a negative class effect, has the highest average value of `read5`

($\overline{\texttt{read5}}_4 = 0.391$) and the lowest average score of `read8` ($\overline{\texttt{read8}}_4 = -2.78$).

First response variable			
	$\overline{\texttt{math8}}$	$\overline{\texttt{math5}}$	$\overline{\texttt{ESCS}}$
m=1	−1.683	0.344	−0.360
m=2	0.223	−0.052	0.045
m=3	−0.374	0.026	−0.144
Second response variable			
	$\overline{\texttt{read8}}$	$\overline{\texttt{read5}}$	$\overline{\texttt{ESCS}}$
k=1	0.207	−0.065	0.026
k=2	−1.166	0.271	−0.213
k=3	−0.289	0.056	−0.153
k=4	−2.78	0.391	−0.085

Table 4.8: Average values of the standardized variables used in the model within the identified subpopulations, for mathematics and reading.

The $M \times K$ matrix of the joint weights W and the variance/covariance matrix $\boldsymbol{\Sigma}$ are estimated as follows:

$$\hat{W} = \begin{pmatrix} 0.009 & 0.011 & 0.010 & 0.007 \\ 0.606 & 0.010 & 0.053 & 0.006 \\ 0.097 & 0.008 & 0.176 & 0.007 \end{pmatrix} \quad \hat{\Sigma} = \begin{pmatrix} 0.479 & 0.175 \\ 0.175 & 0.448 \end{pmatrix}. \tag{4.18}$$

The covariance and the correlation among the errors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are 0.175 and 0.377, respectively. Considering the two marginal distributions of the class effects, we observe from Table 4.7 that, in the case of mathematics (first response variable), classes are divided into three subpopulations, one containing more than half of the total number of classes (67.5%), one a bit smaller containing the 28.8% of the classes and a very small one containing the 3.7% of the classes. The distribution of the class effects in reading on the four subpopulations also sees a very numerous subpopulation containing the 71.2% of the classes, followed by a subpopulation containing about the 23.9% of the classes and by two very small subpopulations containing the remaining 4.9% of the classes. By looking at the matrix W of the joint weights, we see that the joint distribution of the class effects on reading and mathematics is not uniform on the 12 mass points, even if we adjust for the two marginal distributions, but it is mainly concentrated on certain mass points. This result further highlights the utility and the advantage of the bivariate modeling. The subpopulation $(2, 1)$ contains the 60.6% of the classes. This subpopulation is characterized by a coherent effect of the class in

mathematics and reading since both subpopulation 2 for mathematics and subpopulation 1 for reading are interpreted as the best subpopulations, that are the ones in which the predicted values of student achievements are very high. Being this subpopulation the most numerous one and the one containing classes with high positive effects, it can actually be interpreted as the reference subpopulation. The algorithm identifies the deviations from this reference subpopulation of groups of classes that perform poorly, considering both their effects in mathematics and reading. In particular, the reference subpopulation is followed, in terms of frequency, by the subpopulation $(3, 3)$, with the 17.6% of classes, where the trend of student achievements between mathematics and reading is still similar and the predicted scores at grade 8 are still high, but lower than the ones of the reference subpopulation. Therefore, subpopulations $(2, 1)$ and $(3, 3)$ do not present substantial differences, but, both of them are characterized by positive effects of the classes in mathematics and reading. What deserves attention are the subpopulations of classes that significantly differ from the reference subpopulation in their effect, because they are characterized by negative class effects on both maths and reading or by opposite effects on the two school subjects. In this perspective, subpopulations $(1, 2)$ and $(1, 4)$, containing 1.3% of the classes, are associated to classes that have both negative effects on mathematics and reading. Regarding the subpopulations with incoherent effects, subpopulation $(1, 1)$ (with about 1% of the classes) contains classes with a very positive effects in reading, but a very negative one in mathematics, while subpopulations $(2, 2)$ and $(2, 4)$ (with 1.6% of the classes) contain those classes that have a very positive effects in mathematics, but a negative effect in reading. In particular, since we are interested in identifying the behaviors that significantly differ in their effects on student achievements from the reference subpopulation, we select 4 subpopulations that deserve attention:

- S_{ref} = subpopulation (2,1) - the reference subpopulation. It contains classes with positive impacts both in mathematics and reading.

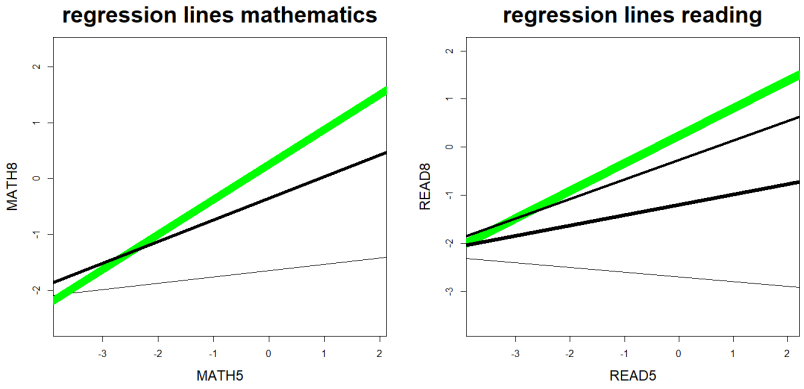


Figure 4.4: Regression planes projected on the 2-dimensional plane identified by the answer variable and the random covariate, identified by the parameters of Eq. (4.17) estimated by the MSPEM algorithm and whose parameters are shown in Table 4.7. Colored lines identify the subpopulation ($m = 2, k = 1$).

- S_2 = union of subpopulations (1,2) and (1,4). It contains classes with negative impacts both in mathematics and reading.

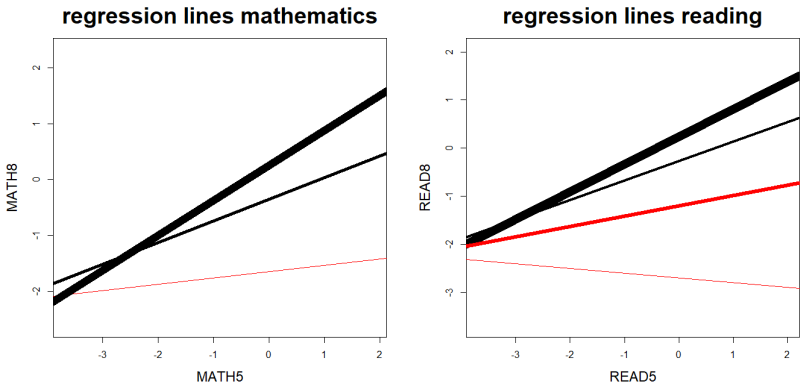


Figure 4.5: Regression planes projected on the 2-dimensional plane identified by the answer variable and the random covariate, identified by the parameters of Eq. (4.17) estimated by the MSPEM algorithm and whose parameters are shown in Table 4.7. Colored lines identify the subpopulations ($m = 1, k = \{2, 4\}$).

- S_3 = union of subpopulations (2,2) and (2,4). It contains classes with a positive impact in mathematics and a negative one in reading.

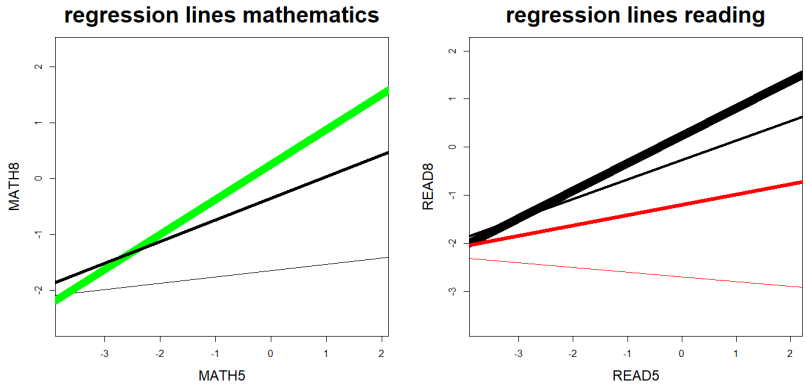


Figure 4.6: Regression planes projected on the 2-dimensional plane identified by the answer variable and the random covariate, identified by the parameters of Eq. (4.17) estimated by the MSPEM algorithm and whose parameters are shown in Table 4.7. Colored lines identify the subpopulations ($m = 2, k = \{2, 4\}$).

- S_4 = subpopulation (1,1). It contains classes with a negative impact in mathematics and a positive one in reading.

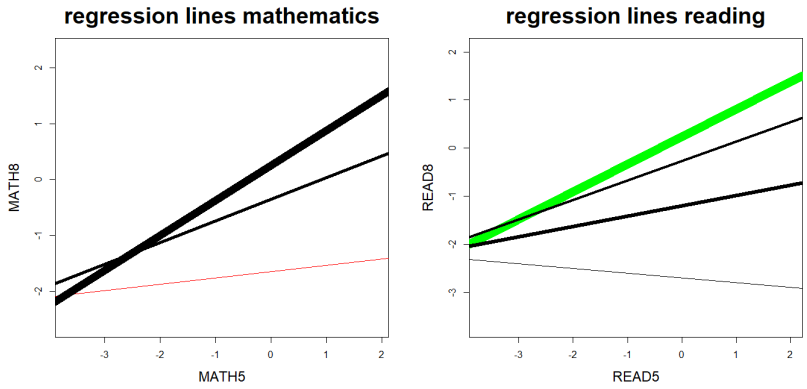


Figure 4.7: Regression planes projected on the 2-dimensional plane identified by the answer variable and the random covariate, identified by the parameters of Eq. (4.17) estimated by the MSPEM algorithm and whose parameters are shown in Table 4.7. Colored lines identify the subpopulation ($m = 1, k = 1$).

The correlation among the class effects in mathematics and reading that emerges in this analysis is, in some way, in contrast with the results obtained in Masci et al. (2017a), where the class effects in the two school subjects result to be not correlated. Nonetheless, while in this analysis we consider only one level of grouping, i.e. students nested within classes, in Masci et al. (2017a), we consider two levels of grouping, i.e. students nested within classes, in turn nested within schools. Therefore, the school effects in the two schools subjects

that in Masci et al. (2017a) are disentangled by the class ones and that result to be correlated, in this case are not disentangled by the class ones, meaning that part of the correlation that we identify here among the class effects migth be due to the schools in which classes are nested.

4.2.3 Characterization of the subpopulations of classes

The presence of subpopulations of classes that differ in their effect on mathematics and reading student achievements might be the consequence of different class body-compositions, peers, teachers or teaching practices. These aspects may influence the class effect in reading, mathematics or both of them. Moreover, having a disadvantaged situation in one school subject learning may favor student learning in the other school subject and viceversa. Therefore, we are interested in investigating whether there are some class and teacher level variables associated to the 4 heterogeneous identified subpopulations. To this end, we apply a multinomial logit model by treating the class and teacher levels characteristics as covariates and the belonging of classes to the 4 subpopulations (S_{ref}, S_2, S_3, S_4) as outcome variable.

Considering S_{ref} as the reference subpopulation, for each class $i = 1, \ldots, N$ and each subpopulation $l = \{S_2, S_3, S_4\}$, the model takes the following form:

$$\ln\Big(\frac{P(Y_i = l)}{P(Y_i = S_{ref})}\Big) = \beta_{0l} + \sum_{q=1}^{Q} \beta_{ql} X_{iq}. \tag{4.19}$$

Y_i represents the cluster of belonging of class i, for $i = 1, \ldots, N$, X is the $N \times Q$ matrix of class and teacher levels covariates shown in Tables 4.5 and 4.6, where Q is the total number of covariates. Since the number of class and teacher levels covariates is very high and we do not expect all of them to be significant, we perform a lasso multinomial logit regression (Tibshirani, 1996; Lokhorst, 1999) in order to select the significant covariates, addressing multicollinearity issues, and to estimate their association with the response variable. By using cross-validation, we select the penalization term λ of the lasso regression in order to minimize the mean-squared error.

The results of the lasso multinomial logit model, with the best selected choice of λ, are obtained by using the R package `glmnet` (Friedman et al., 2010) and are shown in Table 4.9[8].

[8]We do not report in the Table all the covariates shown in Tables 4.5 and 4.6, but only the ones whose coefficient is not shrinked to zero in the lasso regression.

Variable name	S_2	S_3	S_4
Reading teacher general questions			
`extra activities`	0.075		
`refresher courses`			0.042
Maths teacher's personal information			
`age`			0.043
Reading teacher's personal information			
`permanent job`		−0.150	
`gender`			0.560
Questions about school principals (reading teachers)			
`princ evaluate`			−0.087
Only for mathematics teachers			
`main teaching method 'e'`	2.392		
`main teaching method 'd'`			−0.098
`teacher written exam`	−1.275		
Only for reading teachers			
`num reading hours`			0.064
`teacher open test`		−0.053	
`read newspaper`	0.308		
Class information and body composition			
`area geo South`	0.453		

Table 4.9: Results of the lasso multinomial logit regression in Eq. (4.19). We report in the table only the variables at class and teacher levels that result to be significant in the model.

According to the results of the multinomial logit model shown in Table 4.9, classes in which the reading teacher organizes extra scholastic activities for student reinforcement (variable `extra activities`) and asks students to read newspapers or journals in class (variable `read newspaper`), the mathematics teacher follows the `teaching method 'e'` (i.e. follows only the book) and does not choose by him/herself the questions of the written exams for students (variable `teacher written exam`) and that are in Southern Italy (variable `area geo South`) are more likely to belong to subpopulation S_2 (negative class effect both in reading and mathematics). Classes in which the reading teacher

does not have a permanent job in the school in which he/she teaches (variable `permanent job`) and does not test students by means of open questions tests are more likely to belong to subpopulation S_3 (classes with a positive effect on mathematics and a negative one in reading). Lastly, classes in which the reading teacher is a male (variable `gender`) and follows refresher courses for improving his/her teaching skills (variable `refresher courses`), the school principal does not evaluate the work of reading teachers, the mathematics teacher does not follow the `teaching method 'd'` (i.e. does not favor the capacity of build concepts, models and theories) and is elder (variable `age`) and the number of hours of reading lessons per week is high are more likely to belong to subpopulation S_4 (classes with a positive effect in reading and a negative one in mathematics). Besides the geographical area or the number of hours of lesson per week, these results reflect the fact that personal and working characteristics of teachers are in some way associated to student learning. For instance, being a "not proactive" teacher, who simply follows the book and who does not make personalized tests in mathematics, or who does not ask students to articulate their answers by means of open questions tests in reading, has a negative effect in both the school subjects. Also not having a permanent job, with the likely consequence of not teaching to the same students for consequent years, has a negative impact on student performances. On the contrary, attending refresher courses for teachers improves the positive effect of a teacher on student skills.

4.3 Conclusions

In this chapter, we develop a multivariate, i.e. that allows a multivariate response variable, semi-parametric mixed-effects model, together with an EM algorithm for estimating its parameters (MSPEM algorithm), for hierarchical data and we apply it to INVALSI data 2016/2017 for performing a classification of Italian classes. The MSPEM algorithm is the extension to the multivariate case of the SPEM algorithm presented in the previous chapter. We assume the random coefficients of the mixed-effects model to follow a discrete distribution, where the numbers of support points of the coefficients distribution related to the multiple responses are unknown and are allowed to be different. In doing so, the algorithm identifies a latent structure among the higher level of hierarchy. Each group, i.e. observation at the higher level of hierarchy, is assigned to one of the subpopulations identified, that characterizes the effect of the group related to the multiple response variables. Considering the case of a bivariate response, the novelty and the advantage of this modeling is twofold. First, the MSPEM algorithm identifies two latent structures among the higher level of hierarchy, one related to the first response and one related to the second one. As stated in the previous chapter, identifying patterns within complex data where we do not

have any prior about the number of existing latent subpopulations is already a big advantage. Second, the joint modeling reveals two natures of the correlation between the two response variables: one is the correlation among the distribution of the subpopulations, that can be seen in the matrix of weights W, that tells us how groups are distributed on the $M \times K$ mass points; the second correlation is among the unexplained variance of the two response variables, i.e. Σ_{12}, that tells us whether in the variance of the two response variables that we are unable to explain with the model there is still correlation or not. In this perspective, the algorithm, as shown in the simulation study, is completely identifiable, in the sense that it is able to disentangle the correlation related to these two sources.

The MSPEM algorithm is unique in the literature and can be applied in many classification problems, with the aim of individuating latent patterns within data or also for confirming the presence of a theoretically known number of subpopulations.

Applying the MSPEM algorithm to INVALSI data, considering students as level 1 and classes as level 2, we jointly model the impact of the class on both mathematics and reading student achievements. In this case, we interpret the impact of a class as the linear relation between previous (grade 5) and current (grade 8) INVALSI test scores of students within a class, adjusting for the students socio-economic index. The algorithm reveals the presence of 4 different trends in mathematics and 3 different ones in reading. The distribution of classes on these 4×3 mass points is not uniform but it is possible to identify some more common behaviors. In particular, we distinguish classes that have a positive impacts on student achievements in both maths and reading, from the ones that have a negative one, from the ones that have opposite impacts on the two school subjects. Interested in characterizing the identified subpopulations of classes, we apply, in a second step, a lasso multinomial logit model to explain the belonging of classes to the subpopulations by means of teacher and class levels variables. It emerges that, more than the classical information about class body composition or peers, there are certain teacher practices or characteristics that are associated to different class impacts. In particular, the attitude, the pro-activeness and the preparation of teachers result to be effective on student learning.

Appendix

Algorithm 2: EM algorithm for bivariate semi-parametric mixed-effects models

input : Initial estimates for $(\mathbf{c}_{11}^{(0)}, \ldots, \mathbf{c}_{MK}^{(0)})$ and $(w_{11}^{(0)}, \ldots, w_{MK}^{(0)})$, with $M = N$ and $K = N$;
Initial estimates for $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\Sigma}^{(0)}$;
Tolerance parameters D_1, D_2, $\tilde{w}_1$, $\tilde{w}_2$, `tollR`, `tollF`, `it`, `it1`, `itmax`.
output: Final estimates of $\mathbf{c}_{mk}^{(a)}$, $w_{mk}^{(a)}$, for $m = 1, \ldots, M$, $k = 1, \ldots, K$, $\boldsymbol{\beta}^{(a)}$ and $\Sigma^{(a)}$.
a=1; conv1=0; conv2=0;
while *(conv*1 == 0 *or conv*2 == 0 & $a <$ `it`*)* **do**
  compute the distance matrices DIST1 and DIST2 for both the subpopulations distribution (where, e.g.,for the first response variable, $DIST1_{st} = \sqrt{(c_{1,1s} - c_{1,1t})^2 + (c_{1,2s} - c_{1,2t})^2}$ is the euclidean distance between each couple of mass points $s, t \; \forall s, t = 1, \ldots, M, s \neq t$);
  if *(*$DIST1_{st} < D_1$ & $DIST1_{st} = min(DIST1) \quad (\forall s, t = 1, \ldots, M, s \neq t)$*)* **then**
    collapse marginal masses s and t to a unique mass point;
  if *(*$DIST2_{st} < D_2$ & $DIST2_{st} = min(DIST2) \quad (\forall s, t = 1, \ldots, K, s \neq t)$*)* **then**
    collapse marginal masses s and t to a unique mass point;
  compute the new distance matrices DIST1 and DIST2;
  if *conv*1 == 1 *or* $a \geq$ `it1` **then**
    if $w_{1,m}^{(a)} \leq \tilde{w}_1 \quad (\forall m = 1, \ldots, M)$ **then**
      delete marginal mass point m;
      reparameterize the weights;
    if $w_{2,k}^{(a)} \leq \tilde{w}_2 \quad (\forall k = 1, \ldots, K)$ **then**
      delete marginal mass point k;
      reparameterize the weights;
    if *no changes are done* **then**
      *conv*2 = 1;
  given $\mathbf{c}_{mk}^{(a-1)}, w_{mk}^{(a-1)}$ for $m = 1, \ldots, M$ and $k = 1, \ldots, K$, $\boldsymbol{\beta}^{(a-1)}$ and $\Sigma^{(a-1)}$, compute the matrix W according to Eq. (4.9);
  update the weights $w_{11}^{(a)}, \ldots, w_{MK}^{(a)}$ according to Eq. (4.5);
  $\boldsymbol{\beta}^{(a,0)} = \boldsymbol{\beta}^{(a-1)}$;
  $\Sigma^{(a,0)} = \Sigma^{(a-1)}$;
  $\mathbf{c}_{mk}^{(a,0)} = \mathbf{c}_{mk}^{(a-1)}$;
  $w_{mk}^{(a,0)} = w_{mk}^{(a-1)}$;
  keeping $\boldsymbol{\beta}^{(a,0)}$ and $\Sigma^{(k,0)}$ fixed, update the $M \times K$ support points $\mathbf{c}_{11}^{(a,1)}, \ldots, \mathbf{c}_{MK}^{(a,1)}$ according to Eq. (4.6);
  keeping $\mathbf{c}_{mk}^{(a,1)}, w_{mk}^{(a,0)}$ for $m = 1, \ldots, M$ and $k = 1, \ldots, K$ fixed, update $\boldsymbol{\beta}^{(a,1)}$ and $\Sigma^{(a,1)}$ according to Eq. (4.6);
  j=1;
  while $(|\boldsymbol{\beta}^{(a,j-1)} - \boldsymbol{\beta}^{(a,j)}| \geq tollF \quad or \quad |\Sigma^{(a,j-1)} - \Sigma^{(a,j)}| \geq tollF \quad or \quad |\mathbf{c}_{mk}^{(a,j-1)} - \mathbf{c}_{mk}^{(a,j)}| \geq tollR) \quad \& \quad j \leq$ `itmax` **do**
    j=j+1;
    keeping $\boldsymbol{\beta}^{(a,j-1)}$ and $\Sigma^{(a,j-1)}$ fixed, update the $M \times K$ support points $\mathbf{c}_{11}^{(a,j)}, \ldots, \mathbf{c}_{MK}^{(a,j)}$ according to Eq. (4.6);
    keeping $\mathbf{c}_{mk}^{(a,j)}, w_{mk}^{(a,j-1)}$ for $m = 1, \ldots, M$ and $k = 1, \ldots, K$ fixed, update $\boldsymbol{\beta}^{(a,j)}$ and $\Sigma^{(a,j)}$ according to Eq. (4.6);
  set $\mathbf{c}_{mk}^{(a)} = \mathbf{c}_{mk}^{(a,j)}$ for $m = 1, \ldots, M$ and $k = 1, \ldots, K$, $\boldsymbol{\beta}^{(a)} = \boldsymbol{\beta}^{(a,j)}$, $\Sigma^{(a)} = \Sigma^{(a,j)}$;
  estimate subpopulation mk for each group i according to Eq. (4.10);
  if $(\boldsymbol{\beta}^{(a)} - \boldsymbol{\beta}^{(a-1)} < tollF)$ & $(\Sigma^{(k)} - \Sigma^{(k-1)} < tollF)$ & $(\mathbf{c}_{mk}^{(a)} - \mathbf{c}_{mk}^{(a-1)} < tollR)$ **then**
    *conv*1 = 1;
  a= a+1;

Conclusions

The content of this thesis contributes both to the statistical literature about mixed-effects models and to the research about student learning and educational providers effectiveness. We proposed novel statistical methods with the aim of applying them to educational administrative databases in order to address new and interesting research questions. Relaxing the parametric assumptions on both fixed and random effects of mixed-effects models, we delineated new models able to identify unexplored and interesting patterns within the complex educational data. These models represent an innovative contribution in learning analytics. The common strength of the semi-parametric mixed-effects models that we proposed in the four chapters is their flexibility. They are suited to be applied to several problems in which data have such a complex structure that parametric assumptions would be inadequate or restrictive and where the a priori knowledge about the patterns within data is very limited. When applied to educational data, that collect a huge amount of features about a huge number of individuals and that are characterized by a hierarchical structure where numerous variables interact among each others, these models resulted to be highly informative, to well adjust the data and to extract worthwhile information. In particular, the results that emerged in the four chapters are of extreme interest in a policy perspective. Stressing the importance of disentangling the levels of the data hierarchical structure and of the interactions among the variables, the models outcomes give a new gateway to determine students, classes, schools or universities performances, that can be easily interpreted and drive effective policy implications.

The first topic we covered in this thesis (presented in Chapter 1 and Chapter 2) regards the development of mixed-effects regression and classification trees and their application to educational data. Inspired by the flexibility and the capability in modeling the interactions among the variables of tree-based methods, we started applying the existing mixed-effects regression trees (see Sela and Simonoff (2012)) to OECD-PISA data and, in view of its potential, we continued in the framework of tree-based methods for clustered data developing a method that extends classification trees to handle clustered data, the GMET algorithm, and applying it to the ERASMUS$^+$ SPEET data. Our proposed

GMET algorithm is new in the literature, it can handle any answer variable in the exponential family and it can be applied in many classification problems when data are nested within groups. Its predictive accuracy resulted to be comparable to similar well established statistical methods, but having the advantage of providing easily interpretable results, outlining complex interactions among the input variables.

We showed that regression and classification mixed-effects models where the classical linear function of the fixed-effects is substituted by a tree structure, when applied to educational data, give new insights in the framework of student learning. The main advantage consists in their flexible structure of fixed effects, that allows them to reveal complex patterns within data, pointing out non-linear relationships and individuating the range of values of the inputs associated to different outputs, and, mostly, to model interactions among the input variables. These qualities lead to the identification of both student and school level variables interactions that affect the output, finding out patterns otherwise not identifiable. Moreover, the fact that mixed-effects regression and classification trees are graphically displayable and easy to be interpreted is a worthwhile advantage in a policy perspective, when results need to be communicated to stakeholders with the aim of improving educational systems.

The second topic we covered in this thesis (presented in Chapter 3 and Chapter 4) regards the development of univariate and multivariate linear mixed-effects models where the random effects coefficients follow a discrete distribution with an a priori unknown number of support points, together with EM algorithms to estimate their parameters (SPEM and MSPEM algorithms, respectively). These methods represent a novelty in the context of mixed-effects models and their ability in identifying a latent structure, without knowing a priori the number of latent subpopulations, among the higher level of grouping of a two-levels structure data is completely new in the literature. Moreover, when considering a multivariate response variable, the MSPEM algorithm models the correlation among the latent subpopulations related to the multiple responses, extracting further new information from the data. When applied to INVALSI data, the SPEM algorithm identifies latent subpopulations of schools (when considering students nested within schools) that differ in their effect on student achievements in one school subject, i.e. mathematics. The MSPEM algorithm, additionally, models the effect of classes (when considering students nested within classes) on their student achievements in two school subjects, i.e. reading and mathematics, investigating also the correlation among the subpopulations related to the two effects. In this perspective, a model that identifies clusters of schools or classes standing on their heterogeneous effect on student achievements is new to the literature and enriches the research about school and classes effectiveness. The identified subpopulations are useful in a policy perspective when their characterization is able to suggest the variables on which it is possible to act to improve

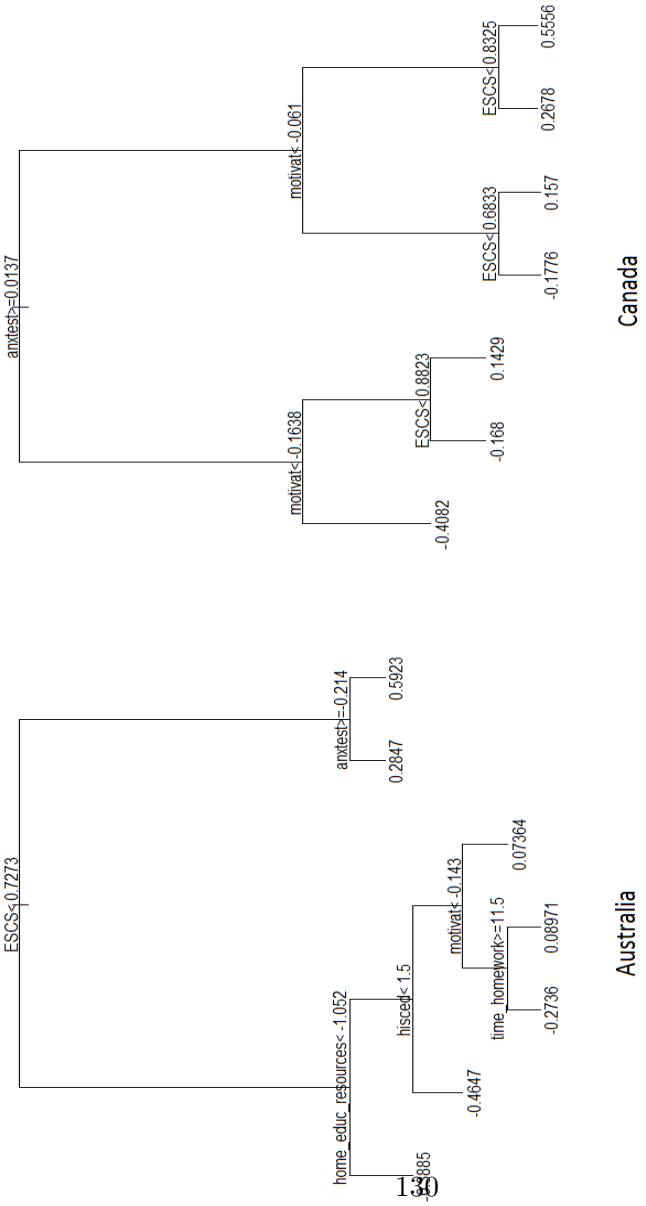
the school or class effectiveness.

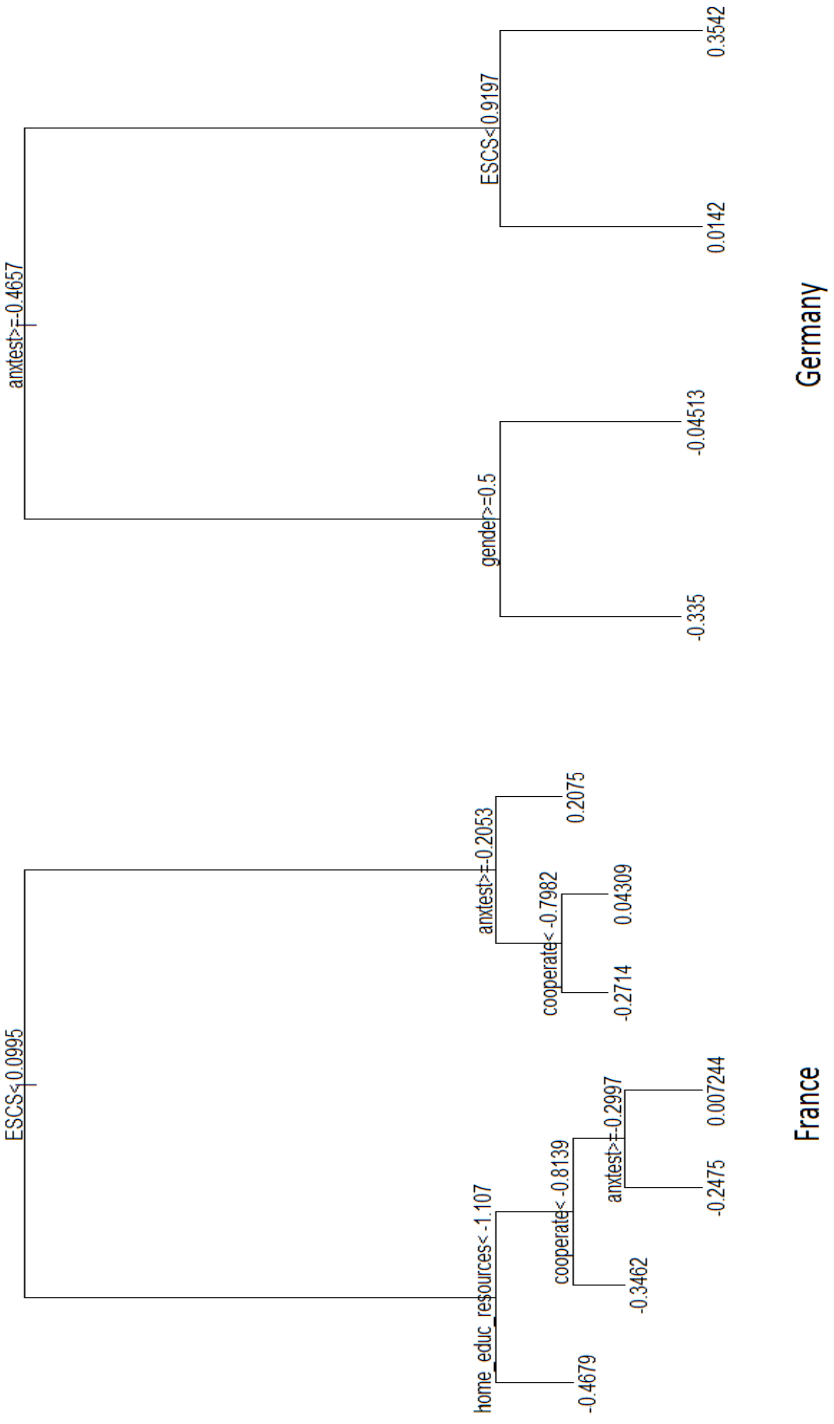
These algorithms are of great utility when applied to any type of data having a hierarchical structure, especially when they are complex and the a priori knowledge about them is very limited, with the aim of classifying the groups at the higher level of hierarchy. The identification of subpopulations within which data have different behaviors, in the era of the performance assessment, can be of interest for measuring the effectiveness or the performance of firms, schools, universities, hospitals and so on so forth, being of impact in real life.

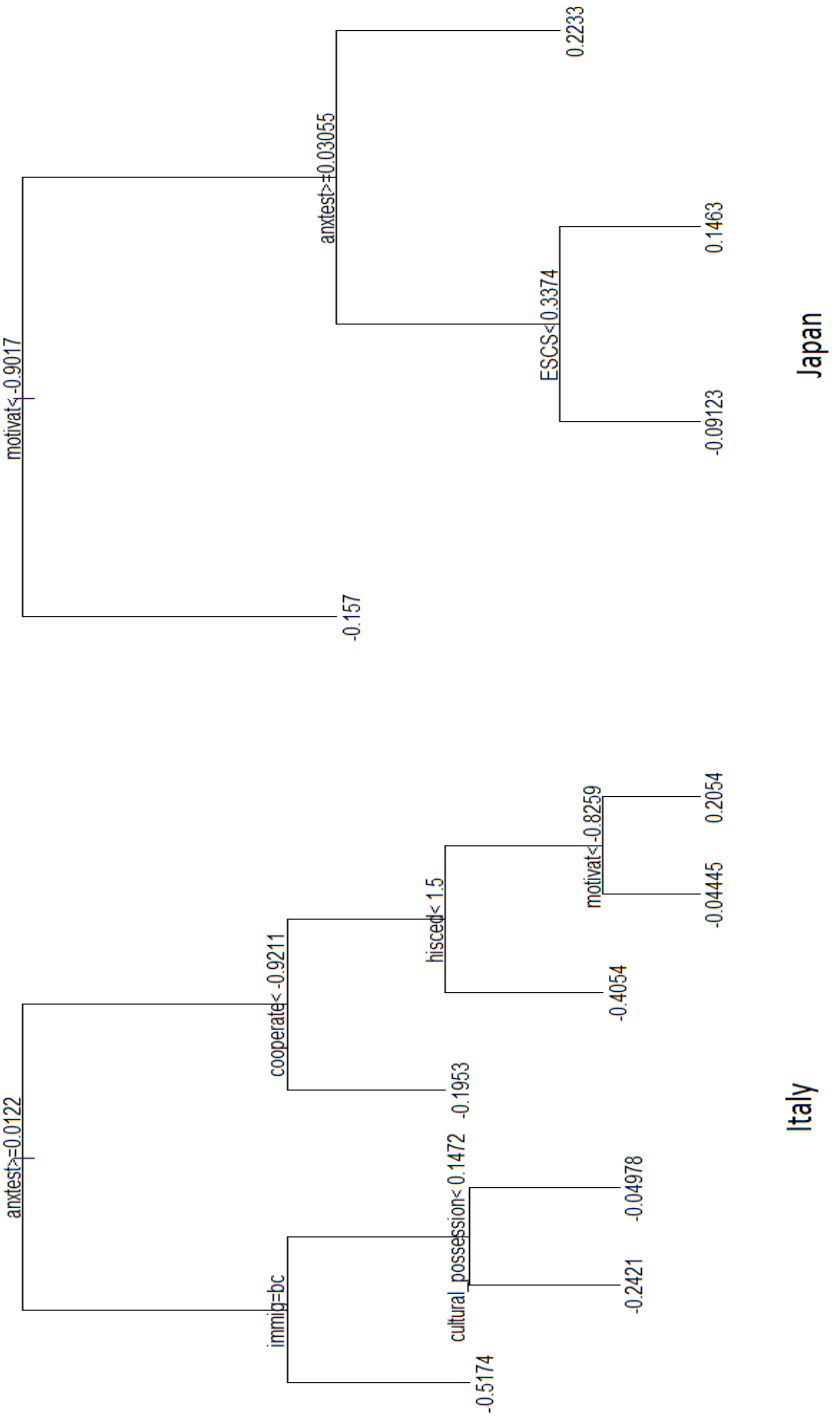
Therefore, the thesis gives evidences that the presented semi-parametric mixed-effects models, besides being innovative in the statistical scenario, enriching the literature about mixed-effects models, when applied to real world educational data, give informative insights that describe educational processes and that, especially, can be used to make policy implications to improve the effectiveness of educational providers.

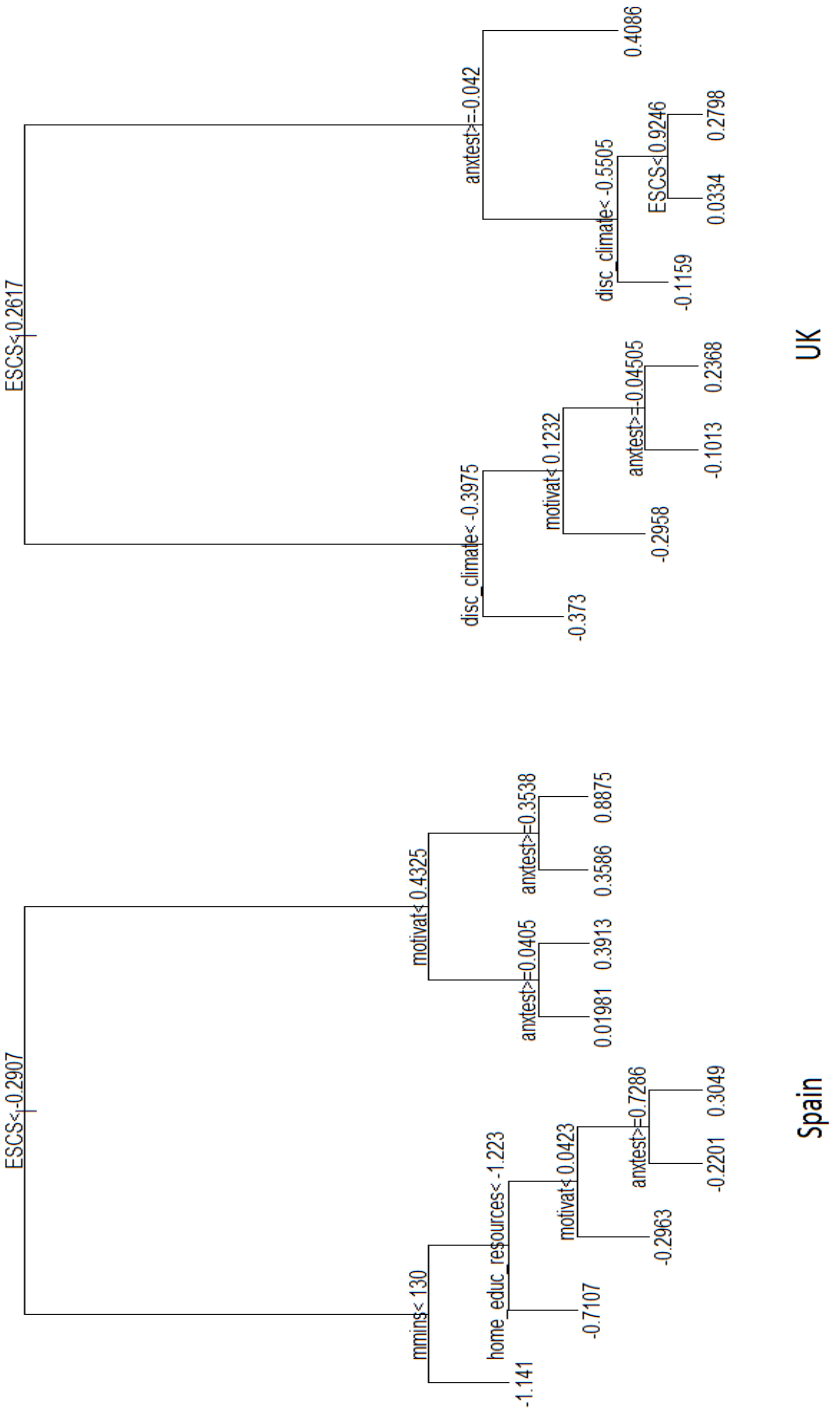
Appendix A

OECD-PISA results




anxtest>=-0.4657
ESCS<0.9197
0.3542
0.0142
gender>=0.5
-0.04513
-0.335
Germany
ESCS<0.0995
anxtest>=-0.2053
0.2075
cooperate<-0.7982
0.04309
-0.2714
home_educ_resources<-1.107
cooperate<-0.8139
anxtest>=-0.2997
0.007244
-0.2475
-0.3462
-0.4679
France


Japan
motivat<-0.9017
anxtest>=0.03055
0.2233
ESCS<0.3374
0.1463
-0.09123
-0.157
Italy
amxtest>=0.0122
cooperate<-0.9211
hisced<1.5
motivat<-0.8259
0.2054
-0.04445
-0.4054
-0.1953
immig=bc
cultural_possession<0.1472
-0.04978
-0.2421
-0.5174


ESCS< 0.2617
anxtest>=-0.042
0.4086
disc_climate< -0.5505
ESCS<0.9246
0.2798
0.0334
-0.1159
disc_climate< -0.3975
motivat< 0.1232
anxtest>=-0.04505
0.2368
-0.1013
-0.2958
-0.373
UK
ESCS<-0.2907
motivat< 0.4325
anxtest>=0.3538
0.8875
0.3586
anxtest>=0.0405
0.3913
0.01981
mmins< 130
home_educ_resources< -1.223
motivat< 0.0423
anxtest>=0.7286
0.3049
-0.2201
-0.2963
-0.7107
-1.141
Spain

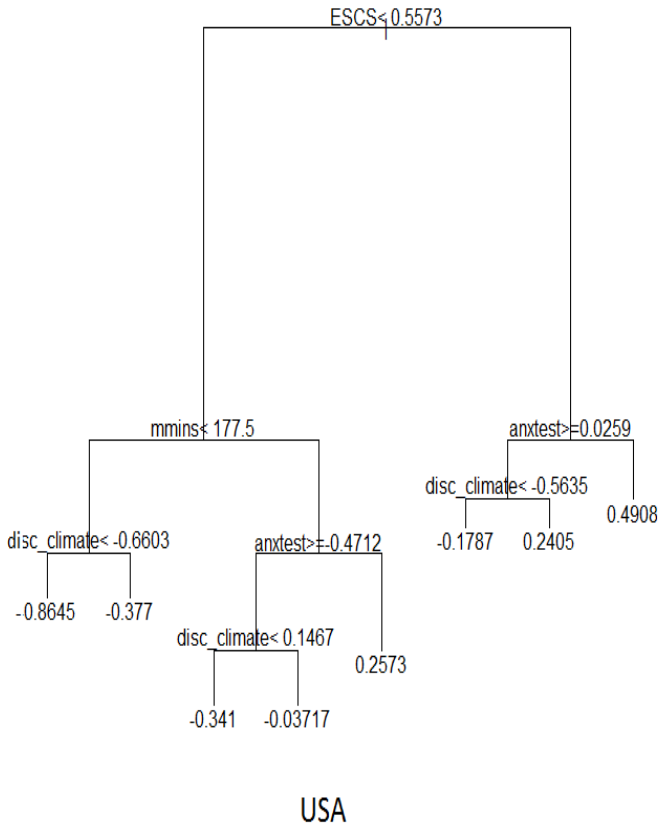
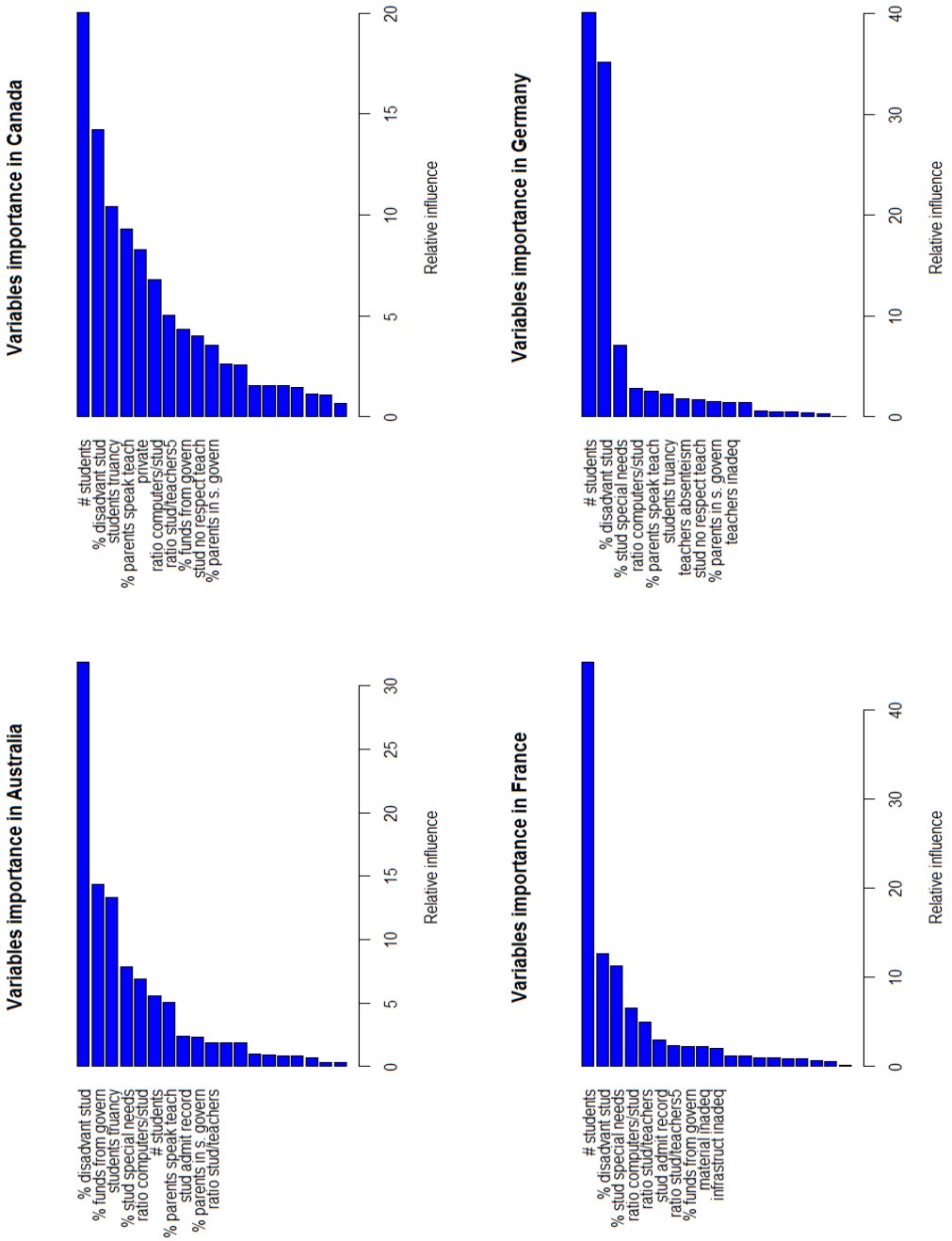
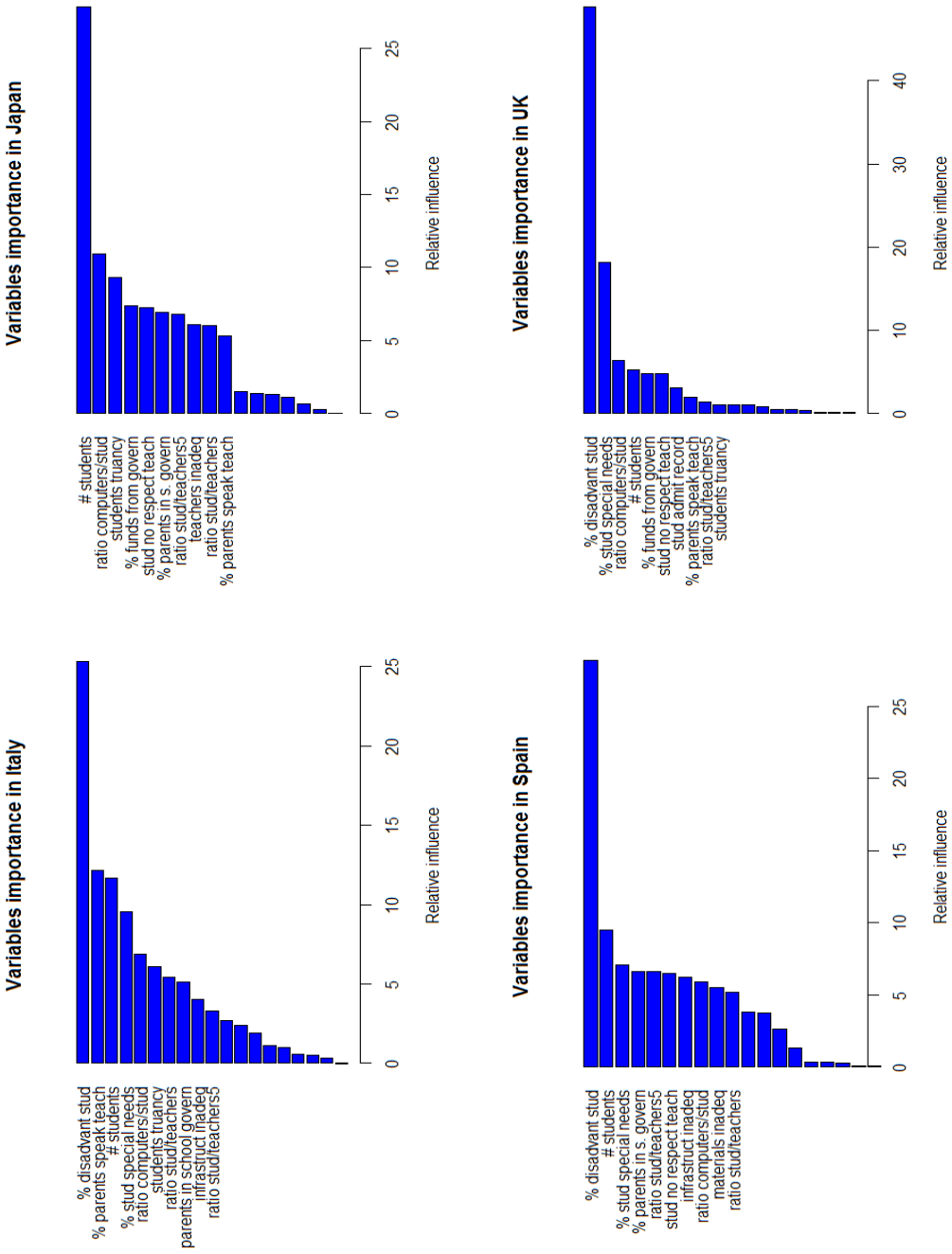


Figure A.1: Fixed effect trees of first stage analysis (RE-EM tree in model 1.5) in the 9 countries.


Variables importance in Canada
students
% disadvant stud
students truancy
% parents speak teach
private
ratio computers/stud
ratio stud/teachers5
% funds from govern
stud no respect teach
% parents in s. govern
0
5
10
15
20
Relative influence
Variables importance in Germany
students
% disadvant stud
% stud special needs
ratio computers/stud
% parents speak teach
students truancy
teachers absenteism
stud no respect teach
% parents in s. govern
teachers inadeq
0
10
20
30
40
Relative influence
Variables importance in Australia
% disadvant stud
% funds from govern
students truancy
% stud special needs
ratio computers/stud
students
% parents speak teach
stud admit record
% parents in s. govern
ratio stud/teachers
0
5
10
15
20
25
30
Relative influence
Variables importance in France
students
% disadvant stud
% stud special needs
ratio computers/stud
ratio stud/teachers
stud admit record
ratio stud/teachers5
% funds from govern
material inadeq
infrastruct inadeq
0
10
20
30
40
Relative influence


Variables importance in Japan
students
ratio computers/stud
students truancy
% funds from govern
stud no respect teach
% parents in s. govern
ratio stud/teachers5
teachers inadeq
ratio stud/teachers
% parents speak teach
0
5
10
15
20
25
Relative influence
Variables importance in UK
% disadvant stud
% stud special needs
ratio computers/stud
students
% funds from govern
stud no respect teach
stud admit record
% parents speak teach
ratio stud/teachers5
students truancy
0
10
20
30
40
Relative influence
Variables importance in Italy
% disadvant stud
% parents speak teach
# students
% stud special needs
ratio computers/stud
students truancy
ratio stud/teachers
parents in school govern
infrastruct inadeq
ratio stud/teachers5
0
5
10
15
20
25
Relative influence
Variables importance in Spain
% disadvant stud
# students
% stud special needs
% parents in s. govern
ratio stud/teachers5
stud no respect teach
infrastruct inadeq
ratio computers/stud
materials inadeq
ratio stud/teachers
0
5
10
15
20
25
Relative influence

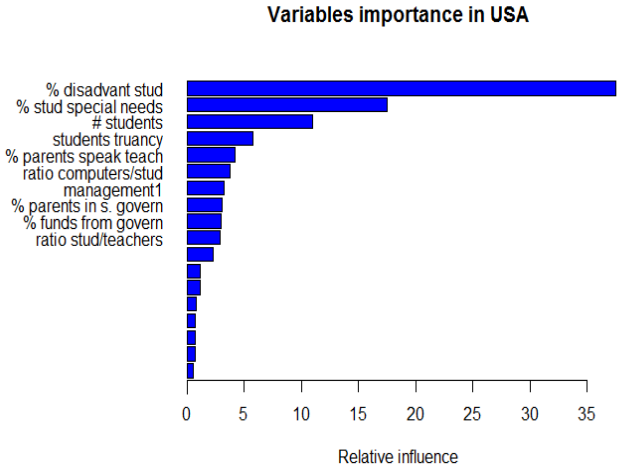
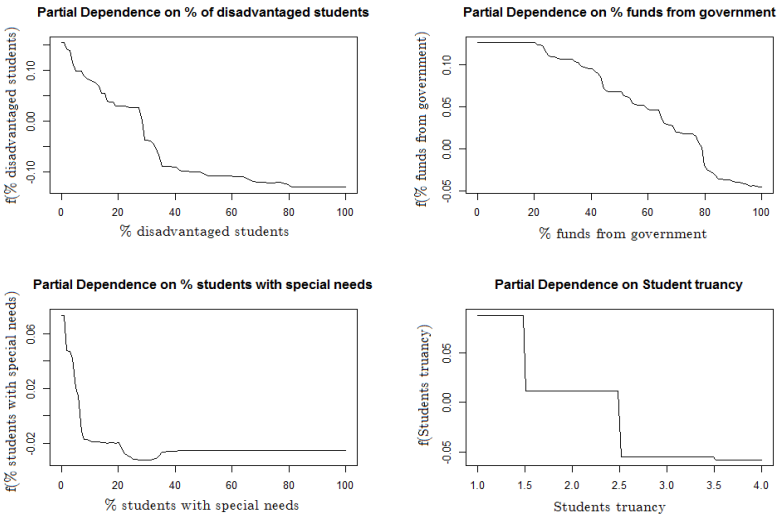
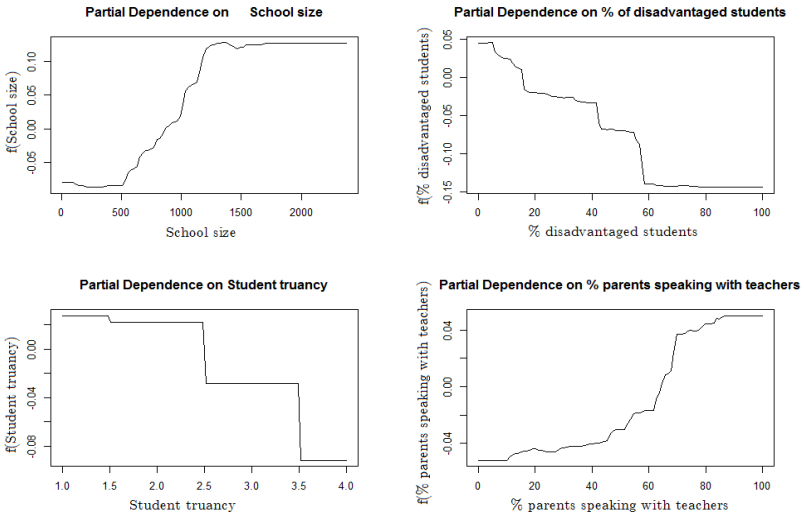


Figure A.2: School level variables importance ranking in the second stage of the analysis. For each country, Boosting creates a ranking of the relative influences of the covariates on the outcome variable (school value-added). To lighten the reading, we report here only the first ten most important variables within each country (where the most important variable is the one able to catch the bigger part of variability in the outcome).

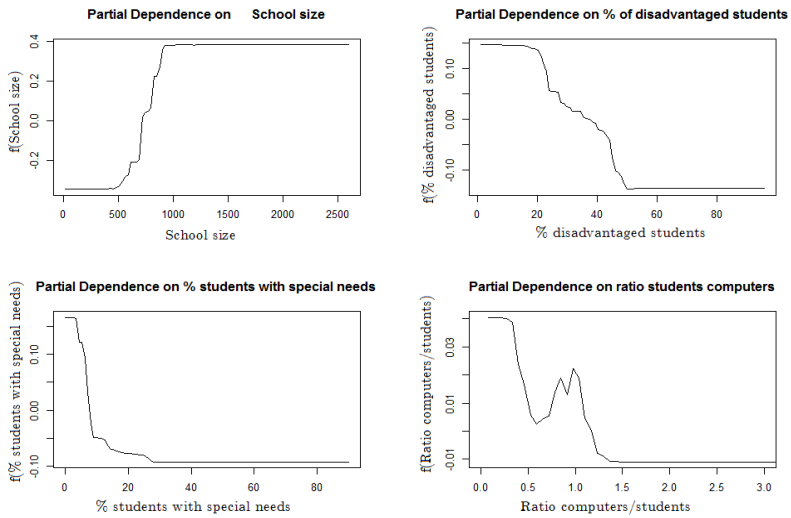
Australia



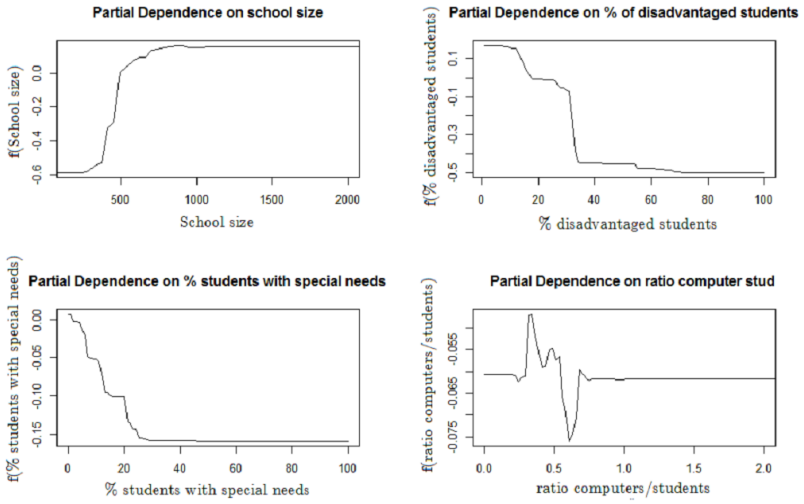
Canada



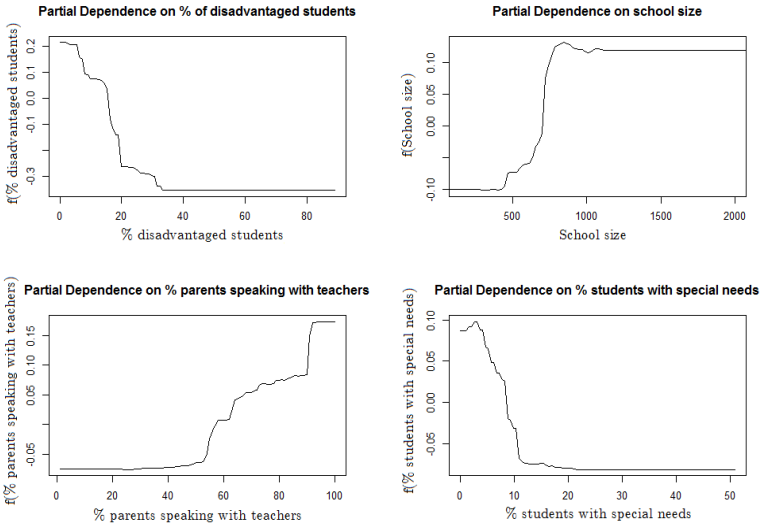
France



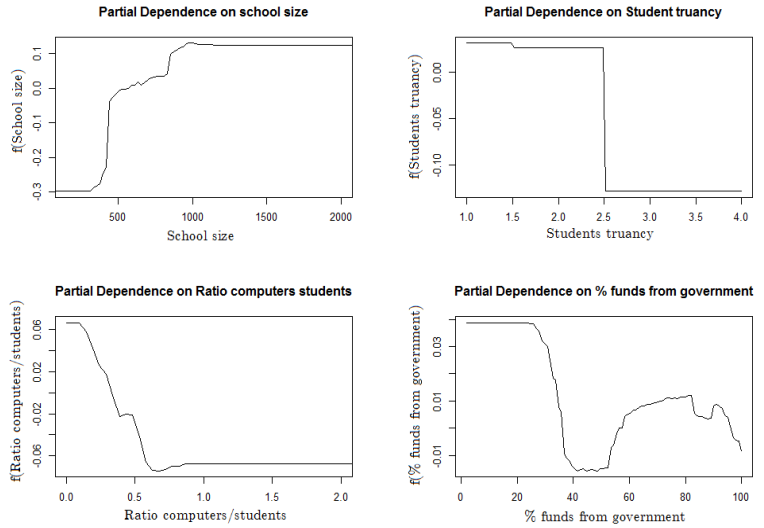
Germany



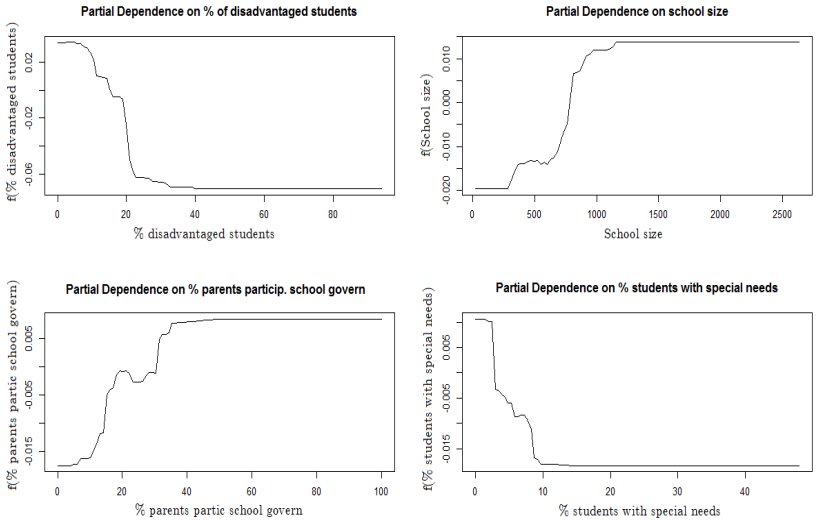
Italy


Partial Dependence on % of disadvantaged students
f(% disadvantaged students)
% disadvantaged students
Partial Dependence on school size
f(School size)
School size
Partial Dependence on % parents speaking with teachers
f(% parents speaking with teachers)
% parents speaking with teachers
Partial Dependence on % students with special needs
f(% students with special needs)
% students with special needs

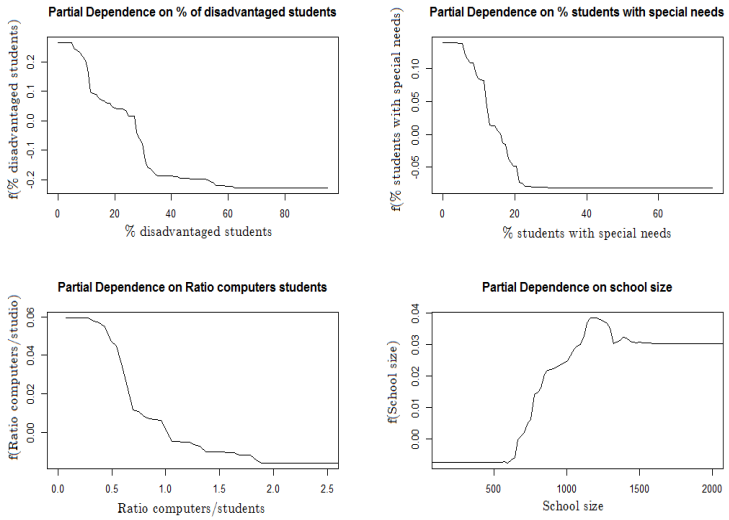
Japan


Partial Dependence on school size
f(School size)
School size
Partial Dependence on Student truancy
f(Students truancy)
Students truancy
Partial Dependence on Ratio computers students
f(Ratio computers/students)
Ratio computers/students
Partial Dependence on % funds from government
f(% funds from government)
% funds from government

Spain


Partial Dependence on % of disadvantaged students
f(% disadvantaged students)
% disadvantaged students
Partial Dependence on school size
f(School size)
School size
Partial Dependence on % parents particip. school govern
f(% parents partic school govern)
% parents partic school govern
Partial Dependence on % students with special needs
f(% students with special needs)
% students with special needs

UK


Partial Dependence on % of disadvantaged students
f(% disadvantaged students)
% disadvantaged students
Partial Dependence on % students with special needs
f(% students with special needs)
% students with special needs
Partial Dependence on Ratio computers students
f(Ratio computers/studio)
Ratio computers/students
Partial Dependence on school size
f(School size)
School size

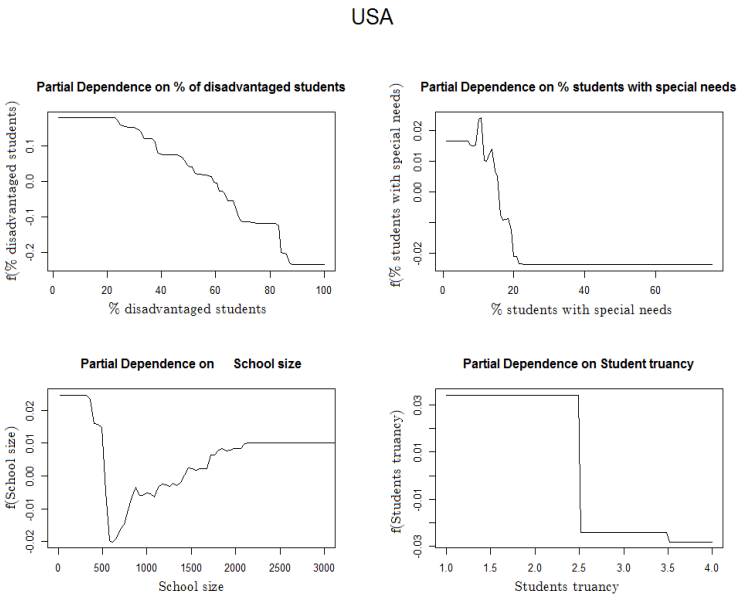
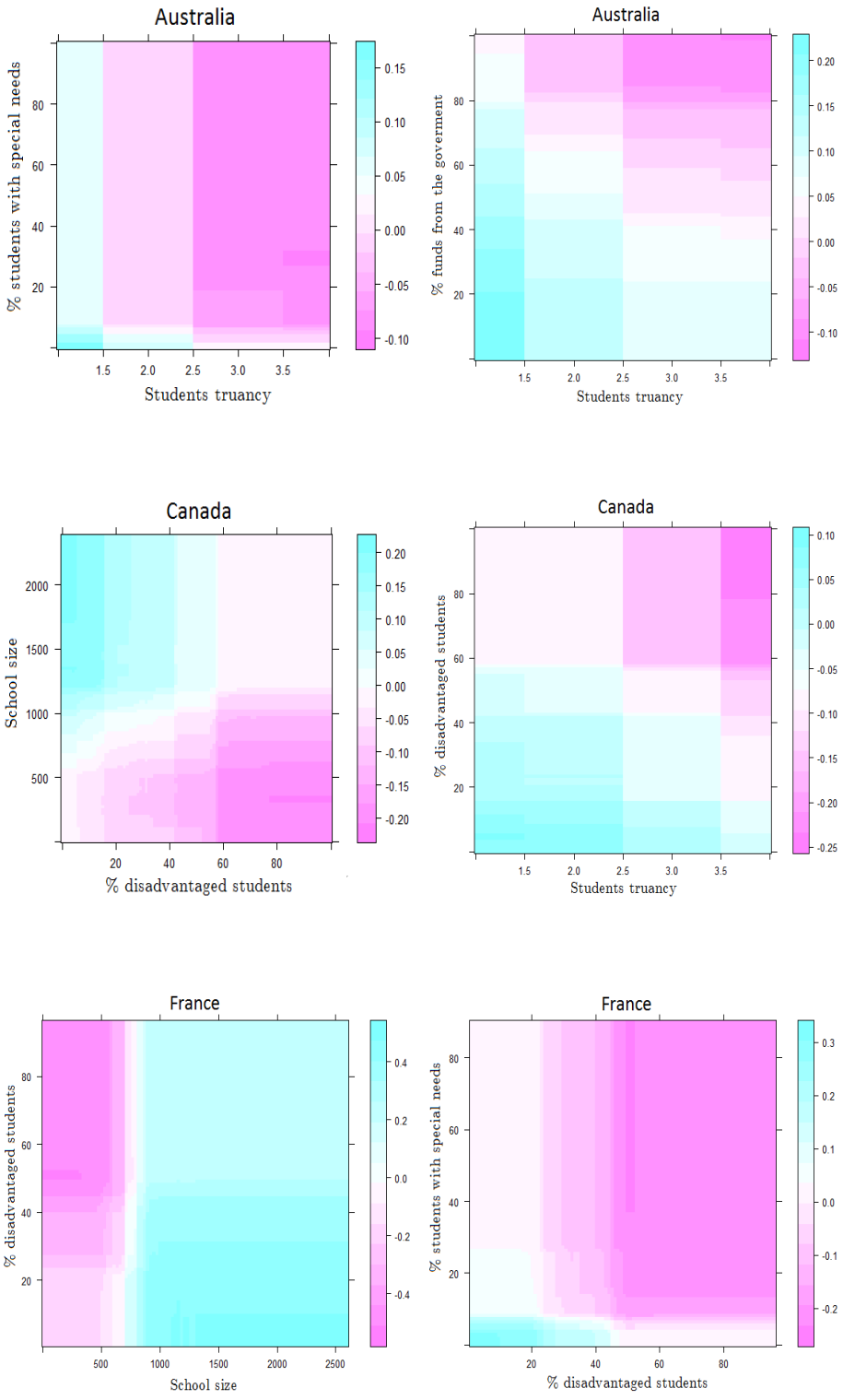
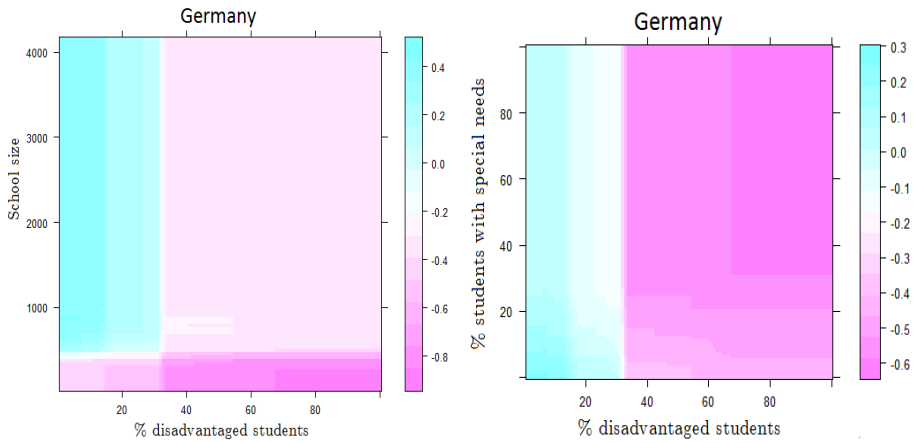
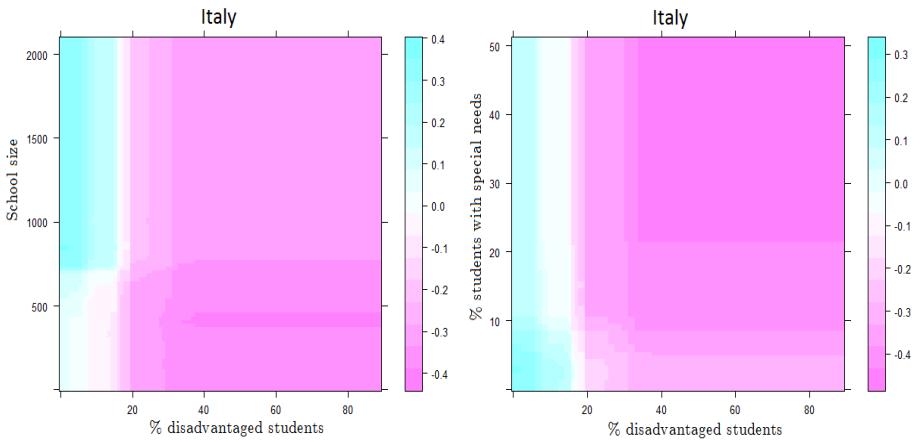
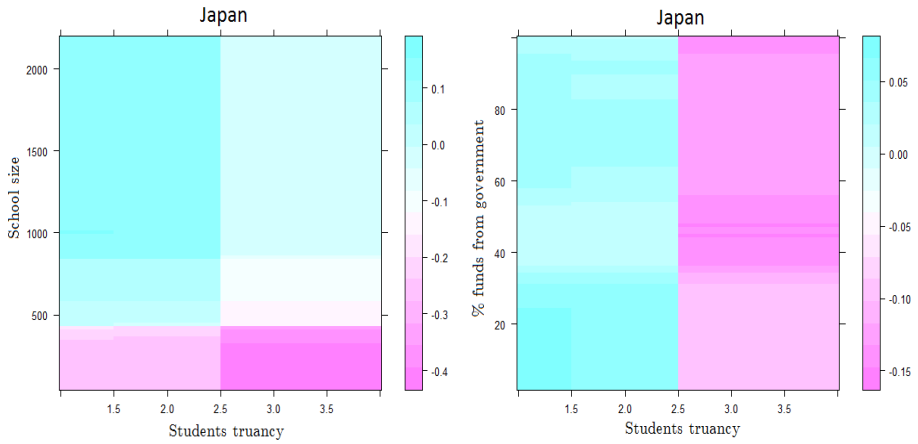


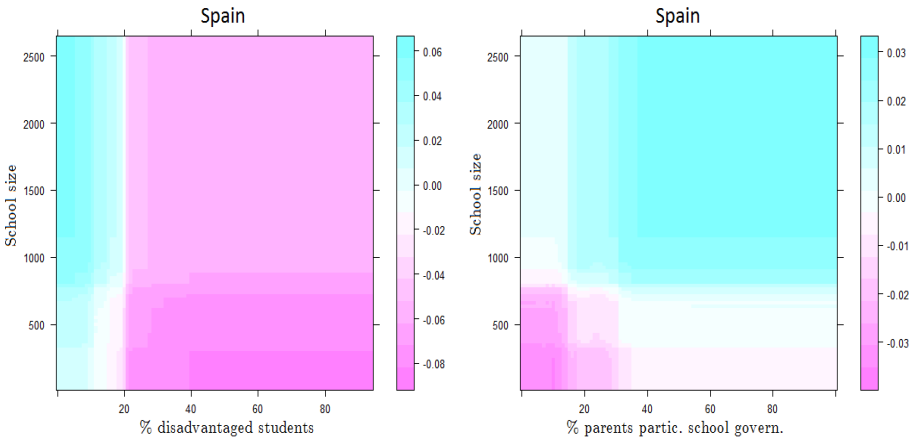
Figure A.3: Partial plot of the four most important school level variables in the association with school value-added, in each country. Note: the selection of the four most significant variables within each country is taken from Figure A.2 and the explanation of each school level covariate is given in Table 1.2.

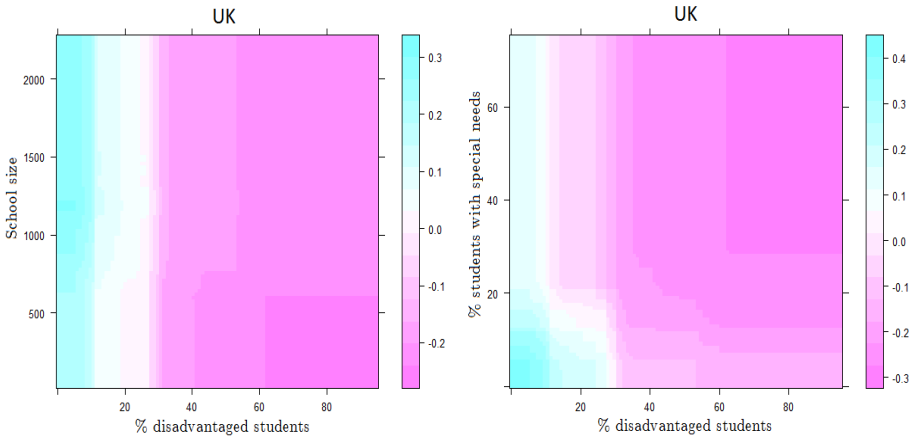

Australia
% students with special needs
Students truancy
Australia
% funds from the goverment
Students truancy
Canada
School size
% disadvantaged students
Canada
% disadvantaged students
Students truancy
France
% disadvantaged students
School size
France
% students with special needs
% disadvantaged students


Germany
School size
% disadvantaged students
Germany
% students with special needs
% disadvantaged students


Italy
School size
% disadvantaged students
Italy
% students with special needs
% disadvantaged students


Japan
School size
Students truancy
Japan
% funds from government
Students truancy


Spain
School size
2500
2000
1500
1000
500
20
40
60
80
% disadvantaged students
0.06
0.04
0.02
0.00
-0.02
-0.04
-0.06
-0.08
Spain
School size
% parents partic. school govern.
0.03
0.02
0.01
0.00
-0.01
-0.02
-0.03


UK
School size
2000
1500
1000
500
20
40
60
80
% disadvantaged students
0.3
0.2
0.1
0.0
-0.1
-0.2
UK
% students with special needs
60
40
20
% disadvantaged students
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3

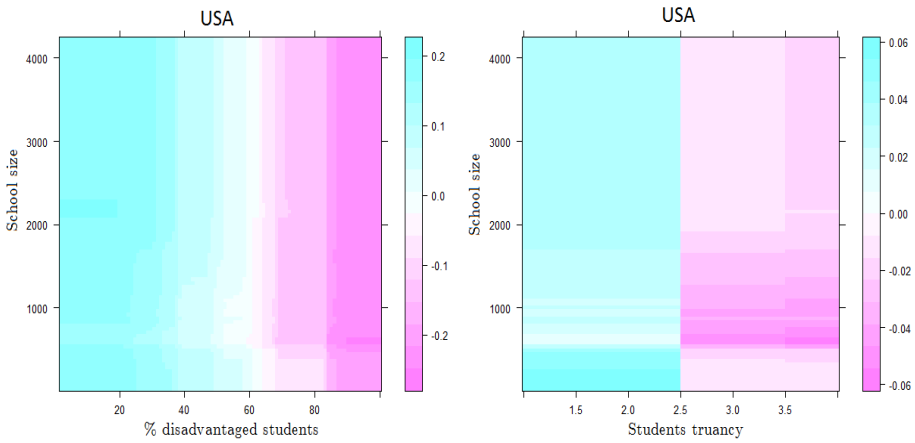


Figure A.4: Joint partial plot of the most important school level variables in association with school value-added, in each country. Notes: 1. Colors represent the scale of the values of the response (school value-added). 2. The selection of variables is based on the group of the variables that turn out to be significant in previous steps.

Bibliography

Agasisti, T., Ieva, F., Masci, C., Paganoni, A. M., and Soncin, M. (2017a). Using statistical analytics to study school performance through administrative datasets. *Data Analytics Applications in Education*, page 181.

Agasisti, T., Ieva, F., and Paganoni, A. M. (2017b). Heterogeneity, school-effects and the north/south achievement gap in italian secondary education: evidence from a three-level mixed model. *Statistical Methods & Applications*, 26(1):157–180.

Agasisti, T. and Vittadini, G. (2012). Regional economic disparities as determinants of student's achievement in italy. *Research in Applied Economics*, 4(2):33.

Agresti, A. (2003). *Categorical data analysis*, volume 482. John Wiley & Sons.

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and computing*, 6(3):251–262.

Anderson, D. (1993). Public schools in decline: Implications of the privatization of schools in australia. *Restructuring schools*, pages 184–199.

Angrist, J. D. and Lavy, V. (1999). Using maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, 114(2):533–575.

Asparouhov, T. and Muthen, B. (2008). Multilevel mixture models. *Advances in latent variable mixture models*, pages 27–51.

Azzimonti, L., Ieva, F., and Paganoni, A. M. (2013). Nonlinear nonparametric mixed-effects models for unsupervised classification. *Computational Statistics*, 28(4):1549–1570.

Barbu, M., Vilanova, R., Lopez Vicario, J., Pereira, M. J., Alves, P., Podpdora, M., Ángel Prada, M., Morán, A., Torreburno, A., Marin, S., et al. (2017). Data mining tool for academic data exploitation: literature review and first architecture proposal. *Projecto SPEET-Student Profile for Enhancing Engineering Tutoring*.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bock, R. D. (2014). *Multilevel analysis of educational data*. Elsevier.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Browne, W. J., Subramanian, S. V., Jones, K., and Goldstein, H. (2005). Variance partitioning in multilevel logistic models that exhibit overdispersion. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(3):599–613.

Bryk, A. S. and Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education*, 97(1):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*.

Cortez, P. and Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. *EUROSIS*.

Elith, J., Leathwick, J. R., and Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4):802–813.

Fitzpatrick, T. and Mues, C. (2016). An empirical comparison of classification algorithms for mortgage default prediction: evidence from a distressed mortgage market. *European Journal of Operational Research*, 249(2):427–439.

Fontana, L., Masci, C., Ieva, F., and Paganoni, A. M. (2018). Performing learning analytics via generalized mixed-effects trees. *MOX-Report 43*.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA:.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Gabriel, F., Signolet, J., and Westwell, M. (2017). A machine learning approach to investigating the effects of mathematics dispositions on mathematical literacy. *International Journal of Research & Method in Education*, pages 1–22.

Goldstein, H. (1987). *Multilevel models in education and social research.* Oxford University Press.

Goldstein, H. (2011). *Multilevel statistical models*, volume 922. John Wiley & Sons.

Goldstein, H., Browne, W., and Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding statistics: statistical issues in psychology, education, and the social sciences*, 1(4):223–231.

Goldstein, H. and Spiegelhalter, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pages 385–443.

Grayson, J. P. (1997). Academic achievement of first-generation students in a canadian university. *Research in higher Education*, 38(6):659–676.

Grilli, L. and Rampichini, C. (2007). A multilevel multinomial logit model for the analysis of graduates' skills. *Statistical Methods and Applications*, 16(3):381–393.

Hajjem, A., Bellavance, F., and Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & probability letters*, 81(4):451–459.

Hajjem, A., Larocque, D., and Bellavance, F. (2017). Generalized mixed effects regression trees. *Statistics & Probability Letters*, 126:114–118.

Hanushek, E. A. (2008). Education production functions. *The New Palgrave Dictionary of Economics. Basingstoke: Palgrave Macmillan.*

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.

Hanushek, E. A. and Woessmann, L. (2010). *The High Cost of Low Educational Performance: The Long-Run Economic Impact of Improving PISA Outcomes.* ERIC.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences.* Sage Publications, Inc.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 6. Springer.

Lin, L. I. et al. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine*, 19(2):255–270.

Lindsay, B. G. et al. (1983a). The geometry of mixture likelihoods: a general theory. *The annals of statistics*, 11(1):86–94.

Lindsay, B. G. et al. (1983b). The geometry of mixture likelihoods, part ii: the exponential family. *The Annals of Statistics*, 11(3):783–792.

Loh, W.-Y., Eltinge, J., Cho, M., and Li, Y. (2016). Classification and regression tree methods for incomplete data from sample surveys. *arXiv preprint arXiv:1603.01631*.

Lokhorst, J. (1999). The lasso and generalised linear models. *Honors Project, The University of Adelaide, Australia*.

Ma, X. (2005). Growth in mathematics achievement: Analysis with classification and regression trees. *The Journal of Educational Research*, 99(2):78–86.

Marginson, S. (1993). *Education and public policy in Australia*. Cambridge University Press.

Martínez, J. F. (2012). Consequences of omitting the classroom in multilevel models of schooling: an illustration using opportunity to learn and reading achievement. *School Effectiveness and School Improvement*, 23(3):305–326.

Masci, C., De Witte, K., and Agasisti, T. (2016a). The influence of school size, principal characteristics and school management practices on educational performance: An efficiency analysis of italian students attending middle schools. *Socio-Economic Planning Sciences*.

Masci, C., Ieva, F., Agasisti, T., and Paganoni, A. M. (2016b). Does class matter more than school? evidence from a multilevel statistical analysis on italian junior secondary school students. *Socio-Economic Planning Sciences*, 54:47–57.

Masci, C., Ieva, F., Agasisti, T., and Paganoni, A. M. (2017a). Bivariate multilevel models for the analysis of mathematics and reading pupils' achievements. *Journal of Applied Statistics*, 44(7):1296–1317.

Masci, C., Ieva, F., Agasisti, T., and Paganoni, A. M. (2018a). Bivariate semiparametric mixed-effects models for the classification of italian classes. *MOX-Report, work in progress*.

Masci, C., Ieva, F., and Paganoni, A. M. (2017b). Semi-parametric mixed-effects models for unsupervised classification of italian schools. *MOX-Report 63*.

Masci, C., Johnes, G., and Agasisti, T. (2018b). Student and school performance across countries: A machine learning approach. *European Journal of Operational Research*, 269(3):1072–1085.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

McCulloch, C., Lin, H., Slate, E., and Turnbull, B. (2002). Discovering subpopulation structure with latent class mixed models. *Statistics in medicine*, 21(3):417–429.

McCulloch, C. E. and Searle, S. R. (2004). *Generalized, linear, and mixed models*. John Wiley & Sons.

Mullainathan, S., Spiess, J., et al. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Muthén, B. (2004). Latent variable analysis. *The Sage handbook of quantitative methodology for the social sciences*, 345:368.

Muthén, B. and Asparouhov, T. (2015). Growth mixture modeling with non-normal distributions. *Statistics in Medicine*, 34(6):1041–1058.

Muthén, B. and Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*, 55(2):463–469.

Nagin, D. S. (1999). Analyzing developmental trajectories: a semiparametric, group-based approach. *Psychological methods*, 4(2):139.

OECD (2016). Pisa 2015 results (volume ı).

Orfield, G., Kucsera, J., and Siegel-Hawley, G. (2012). E pluribus... separation: Deepening double segregation for more students. *UCLA: The Civil Rights Project/Proyecto Derechos Civiles. Retrieved from: http://escholarship.org/uc/item/8g58m2v9*.

Owens, T. L. (2013). Thinking beyond league tables: A review of key pisa research questions. *PISA, power, and policy: The emergence of global educational governance*, pages 27–49.

Peña-López, I. et al. (2016). Pisa 2015 results (volume i). excellence and equity in education.

Pinheiro, J. C. and Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, pages 3–56.

Plewis, I. (2011). Contextual variations in ethnic group differences in educational attainments. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 174(2):419–437.

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. *Statistics in Medicine*, 26(10):2229–2245.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raudenbush, S. and Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of education*, pages 1–17.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, 13(2):85–116.

Raudenbush, S. W. and Willms, J. (1995). The estimation of school effects. *Journal of educational and behavioral statistics*, 20(4):307–335.

Rivkin, S. G., Hanushek, E. A., and Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458.

Romero, C. and Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.

Rumberger, R. W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American educational Research journal*, 32(3):583–625.

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. *Journal of applied quantitative methods*, 6(4):88–99.

Sarrico, C. S., Rosa, M. J., and Manatos, M. J. (2012). School performance management practices and school achievement. *International Journal of Productivity and Performance Management*, 61(3):272–289.

Savona, R. (2014). Hedge fund systemic risk signals. *European Journal of Operational Research*, 236(1):282–291.

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*, pages 149–171. Springer.

Schiltz, F., Masci, C., Agasisti, T., and Horn, D. (2018). Using regression tree ensembles to model interaction effects: a graphical approach. *Applied Economics*, 50(58):6341–6354.

Sela, R. J. and Simonoff, J. S. (2012). Re-em trees: a data mining approach for longitudinal and clustered data. *Machine learning*, 86(2):169–207.

Shen, J., Leslie, J. M., Spybrook, J. K., and Ma, X. (2012). Are principal background and school processes related to teacher job satisfaction? a multilevel study using schools and staffing survey 2003-04. *American Educational Research Journal*, 49(2):200–230.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of educational research*, 75(3):417–453.

Snijders, T. A. (2011). *Multilevel analysis*. Springer.

Speiser, J. L., Wolf, B. J., Chung, D., Karvellas, C. J., Koch, D. G., and Durkalski, V. L. (2018). Bimm tree: a decision tree method for modeling clustered and longitudinal binary outcomes. *Communications in Statistics-Simulation and Computation*, pages 1–20.

Stacey, K. (2015). The international assessment of mathematical literacy: Pisa 2012 framework and items. In *Selected regular lectures from the 12th International Congress on Mathematical Education*, pages 771–790. Springer.

Sun, L., Bradley, K. D., and Akers, K. (2012). A multilevel modelling approach to investigating factors impacting science achievement for secondary school students: Pisa hong kong sample. *International Journal of Science Education*, 34(14):2107–2125.

Therneau, T., Atkinson, B., and Ripley, B. (2017). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-11.

Thomas, E. H. and Galambos, N. (2004). What satisfies students? mining student-opinion data with regression and decision tree analysis. *Research in Higher Education*, 45(3):251–269.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, 28(2):3–27.

Vermunt, J. K. (2011). Mixture models for multilevel data sets. In *Handbook of advanced multilevel analysis*, pages 67–90. Routledge.

Vermunt, J. K. and Magidson, J. (2002). Latent class cluster analysis. *Applied latent class analysis*, 11:89–106.

Willms, J. D. and Raudenbush, S. W. (1989). A longitudinal hierarchical linear model for estimating school effects and their stability. *Journal of educational measurement*, 26(3):209–232.

Wiseman, A. W., Meyer, H.-D., and Benavot, A. (2013). Policy responses to pisa in comparative perspective. *PISA, power, and policy: The emergence of global educational governance*, pages 303–322.

Word, E. et al. (1990). Student/teacher achievement ratio (star) tennessee's k-3 class size study. final summary report 1985-1990.