A contribution to understand STEM students' difficulties with mathematics

Andrà, C.; Brunetto, D.; Pini, A.

Collezione dei *Quaderni di Dipartimento,* numero **QDD 231**
Inserito negli *Archivi Digitali di Dipartimento* in data 13-11-2018

A contribution to understand STEM students' difficulties with mathematics

Chiara Andrà[1], Domenico Brunetto[1], Alessia Pini[2,3]
[1] Department of Mathematics, Polytechnic of Milan, Italy,
[2] Department of Statistics, Umeå University, Sweden,
[3] Department of Statistical Sciences,
Catholic University of Sacred Heart, Italy.

November 12, 2018

Abstract

Drop out during the first year at university STEM courses is a plague spreading all around the world: it has been estimated that, on average, 40% of freshmen abandon their studies before the end of the first academic year. Research in Mathematics Education has revealed that mathematics is one of the main causes for drop out: not only the students' mathematical knowledge, but also affective issues such as attitudes towards learning mathematics, views about mathematics itself, as well as emotions determine the students' success or failure in university career. On the one's hand, thus, it is important to develop suitable and reliable means for investigating both cognitive and affective dimensions, and on the other hand it becomes necessary to reflect on the kind of information the researcher can get from these means of investigation. One of the central issues is the private versus public dimension of learning mathematics. This is connected to the public and private nature of telling about one's emotions and views. We understand "public" versus "private" as identifiable versus anonymous questionnaires and tests, respectively. In this paper, we discuss gains and drawbacks of either approach. In doing so, we also investigate the intertwining of cognitive and affective dimensions in freshmen Engineering students attending a bridge course in mathematics at the beginning of the first semester at the Politecnico di Milano.

Keywords: Anonimous tests; Community analysis; Massive Open Online Courses (MOOC); Regression trees; Students' attitudes; Students' performance

1 Introduction

Every year, at the beginning of the first semester, in all the universities in the world, thousands and thousands of freshmen enrolled at STEM university courses come to attend their first lessons. We know that around 40% of them would not sit in the same classroom few months later, because of dropout. What kind of information can we get from these first days at university, which can help universities to reduce dropout? How do students arrive at university? What is their first impact with the courses, and with mathematics in particular?

In this paper, we aim at contributing to these big, overarching questions by: firstly, recalling the main findings concerning mathematics difficulties at first year STEM university courses; secondly, focusing on the factors that have revealed to be central to understand the issue, and thus relevant for policy-makers to intervene on dropout; and thirdly, narrowing our perspective on one particular issue, namely the distinction between private and public nature of mathematical learning. We, thus, aim at identifying sub-groups of students (we will call them "profiles") who need special and differentiated interventions. Literature review in the following paragraphs discusses briefly the students' difficulties with mathematics at first year university STEM courses and the role that gender gap, differences in the kind of school (read in terms of different mathematical backgrounds), and attitudes towards mathematics and its teaching and learning play.

Literature review, then, focuses of the private and public nature of mathematical thinking, and learning. Two datasets are considered: one coming from an anonymous investigation and the other coming from a non-anonymous one. In order to understand the role played by affective factors in mathematical understanding, we resort to a tool provided by a data mining approach: i.e., classification trees. Once relevant factors are pinpointed, we resort to network analysis in order to identify different profiles of students – who may need specific, differentiated intervention. Data analysis allows us to discuss, first of all, drawbacks and advantages of two different kinds of investigation. Secondly, we discuss the relationship between cognitive and affective factors in shaping the students' difficulties with mathematics at first year university STEM courses.

1.1 Mathematics and STEM university courses

University mathematics causes difficulties to students with STEM majors in general and to Engineering students in particular (Gómez-Chacón et al., 2015). These difficulties can be traced back to several aspects that generally concern differences between secondary school and university (Gueudet, 2008): for example, the different thinking modes that are required at university, as evidenced by all the studies on Advanced Mathematical Thinking (Tall, 1991); the different organization of knowledge and the intrinsic complexity of the new contents to be learnt (Robert, 1998); the different processes and activities that are at issue, proof for one (Moore, 1994); the different didactical contract (Bosch et al., 2004), or, more generally, university courses organization (Hoyles et al., 2001).

In their fundamental study, Clark and Lovric (2008) contend that at the basis of the leap between secondary and tertiary studies there is a shock: *from procedural mathematics to conceptual understanding that university mathematics entails*. According to Hibert and Lefevre (1986), conceptual knowledge describes knowledge of the principles and relations between pieces of information in a certain domain, while procedural is the knowledge of the ways in which to solve problems quickly and efficiently. Pettersson and Scheja (2008) discovered that students develop their knowledge in an algorithmic way, not because of misconceptions, but because it is more suitable for them and enables them to deal functionally and successfully with the presented tasks. The aforementioned findings reveal that not only it is the transition from procedural to conceptual to be sustained, but also that - for the students' sake - it is neither possible, nor advisable, to give away the former and focus only on the latter.

Clark and Lovric (2008) further suggest that transition should be smooth, and communication between the two institutions (school and university) should be improved. According to this view, preparatory courses are available from universities almost all around the world. The goal for such courses is to bridge the gaps between school and university, supporting freshmen students to recapitulate certain mathematical topics. In the sequel, we name this kind of courses "Bridge Courses".

The focus of this paper is on sub-groups of students who may find the transition more difficult, compared to their mates. The differences in mathematical attainments between groups of students, and across schools is a topic of crucial interest for both educators and policy-makers (see e.g. Masci et al., 2016). In the sequel, we briefly recall the main findings to this regard.

1.2 Gender issues

There is an increasing number of studies focusing on the crucial role of social and affective factors—besides the cognitive ones, in undergraduate mathematics learning. Masci et al. (2018), for example, underlie that the students' features –such as gender and attitude towards study– influence students' attainment. In particular, it is well acknowledged that women are under-represented in STEM disciplines (Fox, 2002; Fox and Weisberg, 2011; Pinheiro and Bates, 2000), and we refer to Deaux and Major (1987) model to capture stability and flexibility of gender differences in social behaviour. This model allows us to conceptualize gender as a component of ongoing interactions in which individuals emit expectancies, selves negotiate their own identities, and the context in which interaction occurs shapes the resultant behaviour. This model is distinctly social and psychological in its roots and takes into account both the social influences on boys and girls enrolled in STEM courses, and inner reflections and disposition.

1.3 Mathematical backgrounds

Students' views of mathematics take also a key role. Roesken et al. (2011)'s study has for us many sources of interest. First of all, it discusses from a theoretical point of view the concept of "view of mathematics" and the related concept of "beliefs about mathematics". The authors state that "students' beliefs, wants and feelings are part of their view of mathematics". Secondly, the authors argue about the key role of different school backgrounds, different math curricula and different views and expectations in freshmen students attending a Bridge Course. Within this perspective, we also consider Daskalogianni and Simpson (2001)'s study, which discusses the concept of "beliefs overhang": some beliefs, developed during school days, are carried forward in university, and this may cause difficulties. The study points out the crucial role of beliefs (about mathematics) in determining university success or failure. This is also confirmed by Andrà et al. (2013). Specific to the Italian context, Lombardo (2015) has proved that the kind of high school influences both cognitive and affective factors in the transition. Also Masci et al. (2018) proved that school-level characteristics influence the students' mathematical performances: however, they focused on single schools features such as their size, their Dean's views and management practices, while in our study we focus on the kind of mathematics the students experience at school. In fact, in the Italian context, students who enroll in STEM courses mostly come from three kinds of secondary school: scientific (LS), humanistic (HU), and technical (TE). LS is a type of secondary school with a strong curriculum in math and sciences, while HU is stronger in history, phylosophy, languages and arts. And, while LS and HU curricula are specifically designed to prepare students to go to university, TE one is mostly related to workplace, but it is not rare that students from this type of school enroll at university. A focus of this sort (Andrà et al., 2013; Lombardo, 2015) allows us to understand the role that both mathematical prerequisites (at cognitive level), and views about the importance of mathematics in real life (mirrored by the importance assigned to mathematics in each school type time schedule) may play in the transition to university STEM courses.

1.4 The digital turn

There is a further factor that is gaining more and more attention in the last years: this factor is related to distance learning and e-learning in general. In particular, the students' disposition toward on line teaching material plays a key role in our study, given the organisation of the Bridge Course under investigation. More generally, this factor is related to differences between conceptual and procedural aspects in mathematics, as we argue in what follows.

Some researchers (e.g., Gamer and Gamer, 2001) found that teacher-centred (or teacher-oriented) methods (TO) favour the development of procedural knowledge while student-centered (or student-oriented) methods (SO) favour the development of conceptual knowledge. A TO lesson provides the students with a linear and organised exposition of knowledge, while a SO one engages students in groupwork activities, classroom discussion and in the production of meanings that are inevitably other than final or "authorized"—they are personal and provisional, not universal and absolute. A Massive Open On line Course (MOOC) has a SO pedagogical format, in that the students are required to: (a) watch videos and get sense of their content (without any guidance from the teacher); (b) in case parts of the videos are not clear for the student, search for other sources in order to make sense of the content; (c) make interactive exercises and in some cases engage in forum discussions. All this entails a production of meanings that is personal and that emerges from the mathematical activity in which the student is engaged. MOOCs are becoming a teaching format common to many universities around the world. Also the Bridge Course under study takes on a blended learning format, as we will describe more into details in the section dedicated to the context of the research.

2 Anonymous or identifiable?

In the previous section, we highlighted the factors that shape and influence the students' transition from secondary school to university, and hence their success or failure in the latter. We label them as: gender, beliefs, secondary school type, and MOOC attendance.

In order to investigate their role at the beginning of STEM studies, both questionnaires and tests are necessary. The former allow the researcher to get an insight on affective and social aspects, and the latter give information about mathematical knowledge. If the students attend a Bridge Course that lasts some days, it can be interesting to administer a questionnaire in the first day and another questionnaire in the last day, plus some tests to measure also the students' understanding of the mathematical concepts delivered during the course. At this point, it comes the research question that frames this paper: should the questionnaires and tests be anonymous or identifiable?

The question has methodological implications, as we will discuss extensively in the next sections, but it has also foundational premises, as we briefly illustrate now. The categories of public and private seem to correspond to various important aspects and activities of mathematical learning. Among the private aspects of mathematical learning, for example, may be counted reflection, internalization, visualization, and the creation of mathematical meanings, which have been given considerable weight in mathematics education and mathematics education reform (e.g., Clarke, 2001; Skemp, 1987; Fried and Amit, 2003; Schoenfeld, 1992; National Council of Teachers of Mathematics, 2000). When learning is private, the students feel free to make mistakes, to express their doubts and to be creative (Fried and Amit, 2003).

In our use of the words 'public' and 'private', we have primarily in mind the way they are employed in political theory, where these terms refer to 'access' and, especially, 'accountability' (Benn and Gaus, 1983). By public activities, then, we mean those activities with regards to which one is accountable to teachers, peers, or co-workers; these are activities, therefore, in which one is bound by common practices and by the necessity of formal communication. Questionnaires and tests, where to show one's identity is requested, belong to this sphere.

By private activities, on the other hand, we mean those with regards to which one is not accountable to teachers, peers, or co-workers; these are activities in which one is free from the expectations and constraints of common practice, activities that, as one writer puts it, take place in 'a zone of immunity' (Duby, 1985, p. 10); here, one is free to explore, backtrack, and reflect. Anonymous tests and questionnaires belong to this sphere.

The distinction between identifiable and anonymous questionnaires and tests is tied to the more general consideration that different pedagogical practices or assessment regimes may cause a given mathematical activity to be termed private or public. Students' homework, for example, might in one pedagogical setting be discussed, collected, and marked, and, therefore, take on a public character; in a different pedagogical setting, homework might be given only to reinforce classroom material and never be seen by anyone except the students themselves. Some aspects of mathematical activity, on the other hand, seem to resist relocation from one sphere to another; for example, mathematical papers or projects as public and individual preliminary reflections as private would be hard to reclassify as private and public (Fried and Amit, 2003). Tests, quizzes and questionnaires in a Bridge Course can take a private or a public form, and the next section illustrates the methodological details.

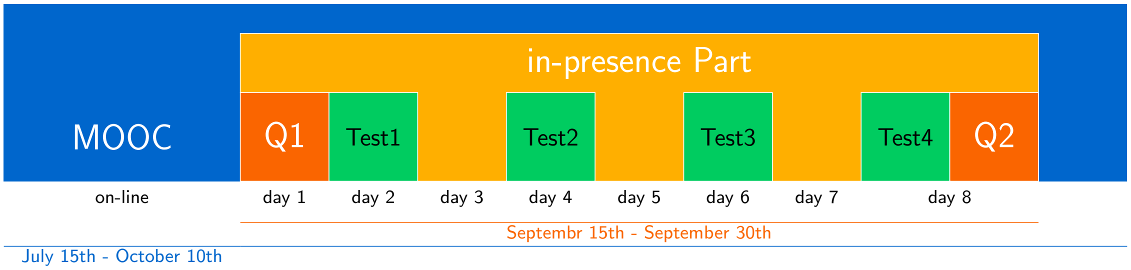


Figure 1: The timeline of the Bridge Course

3 The context of the research

The Bridge Course, delivered every year at the Polytechnic of Milan, is a preparatory course before the beginning of the first semester. The funding source of the course is the Polytechnic of Milan, which had no involvement in any phase of this research. Data analysis had been partly funded by the Italian National Project "Lauree Scientifiche" (STEM degrees), which aims at enhancing interest towards scientific career in young people. The Bridge Course recapitulates the basic math knowledge learned at school and is made of an e-Learning part (i.e., the Pre-Calculus MOOC on POK platform) and an attendance part. Students who enroll at university are invited to attend the MOOC course before the attendance one.

- In the e-learning part, the students are asked to recap essential mathematics on Pre-Calculus MOOC (`www.pok.polimi.it`), where they can watch videos on theory and exercises, and assess their basic knowledge in mathematics through quizzes. In addition, they can interact in a forum. The MOOC course is structured in 6 learning weeks, one for each of the following topics: arithmetics, algebra, geometry, logics, functions, probability.
- The in-presence part features the students in SO activities, such as group work activity and discussions built upon the syllabus of the Pre-Calculus MOOC. The attendance part consists in 32 hours of lessons, spread in the first 2 weeks of September.

With Niegemann et al. (2008), we maintain that the Bridge Course combines self-directed (i.e., MOOC) and externally-regulated (i.e., attendance) learning types of instructional formats. There's a need for the latter, since learners are new at the university, they have to acclimatise with the new learning environment and attendance helps them to familiarize with the new didactical contract and the new organisation of courses. There's a need for the former, since learners at university have to be more self-directed and e-Learning helps to adapt their learning behaviour (Mandl and Kopp, 2006). The first author is a researcher in Mathematics Education and designed the attendance part (both learning materials and schedule of the activities). Her role in the research has been also to identify relevant issues in Mathematics Education concerning difficulties for STEM students in the transition from school to university. The second author is a PhD student in Mathematical Modelling in Engineering and, with a team of researchers, realised the MOOC. His role in the research has been to perform network analysis. The third author is a researcher in Statistics and her role has been to conduct statistical analyses of the data. All the authors had been also tutors teaching the in-presence part.

The data for this study come from two questionnaires, which investigate affective factors, and four tests, which assess the students' knowledge on algebra, geometry and logics, calculus, and probability and statistics. The first questionnaire has been given to the students at the very beginning of the attendance part, while the second one was administered at the end of it, see Figure 1. The two questionnaires (referred to as Q1 and Q2) are composed by two main sections: 1) the personal data (Q0 in the sequel), and 2) the affective section (QA1 and QA2, respectively).

Q0 asked about: gender, school type and MOOC attendance.

QA1 is made of five questions. Question QA1.1 allows us to know how students see the math they have experienced at school, QA1.2 and QA1.3 open a window on the students' expectations about math at university, QA1.4 and QA1.5 investigate whether the students have been exposed to SO learning formats.

QA2 is made of 6 questions. Question QA2.1 asked whether they faced new math topics in the Bridge Course, while Question QA2.2 asked whether they saw exercises formulated in a different way. QA2.3 and QA2.4 were the same of QA1.3 and QA1.4 in the initial affective questionnaire QA1. QA2.5 and QA2.6 were dedicated to MOOC/course appreciation.

The students were also engaged in four math tests, made up of 10 multiple-choice math questions each. The tests provide information about the students' mathematical knowledge and skills and have been administered on the second day of the attendance course (algebra), on the fourth one (geometry and logics), on the sixth one (calculus), and on the eight and last one (probability and statistics). We refer the reader to Figure 1 for an overview of the entire timeline of the Bridge Course at the Polytechnic of Milan.

4 Background

From Section 3 we can see that the type of data that we would like to analyse are heterogeneous. Indeed, we have both quantitative variables measuring students' performances, and qualitative ones related to personal-level features and affective aspects. Moreover, there is a plea in Mathematics Education research for studies that do not assume linear correlations between variables that are complex in nature. This is especially true when affective aspects are under scrutiny.

For these reasons, we resort to methods that do not rely on strong modelistic assumptions on the structure of data, nor on linearity of connections. We employ classification trees to investigate the influence of personal-level features (i.e., gender, school type, and MOOC attendance) on mathematical test performances. We recall that previous studies in Mathematics Education have resorted to classification trees to investigate the interplay of cognitive and affective factors in determining students' performances. For example, Andrà et al. (2013) have analysed how taking the degree in Mathematics can be influenced by the same personal-level features considered here (gender, school type, and students' views of mathematics), plus other information coming from students' university career (such as, CFU and grades at exams earned at each session during the academic year).

Parallel to classification tree to understand the interplay of personal-level features and test performance, in the present study, in order to identify how students clusterize when they expose their views of mathematics, we resort to network analysis and specifically community detection. To our knowledge, despite it represents a powerful tool in many contexts such as in criminal networks (Calderoni et al., 2017), this method has never been employed in an educational context. One can wonder whether a more classical unsupervised clustering method was not employed for this purpose. We argue that in qualitative questionnaires the strong limitation of the latter approach is the need of defining a suitable metric to measure differences between students' answers, which can be avoided using a network analysis approach.

In the present section we describe the details of these two methods.

4.1 Classification and regression trees

The classification and regression trees (CART) are methods that aim at predicting the value of a target variable on the basis of several input variables, and selecting the input variables that explain the most the target variable. If the target variable is dichotomic (e.g., gender) or cathegorical (e.g., school), classification trees are used; if the target variable is numerical (e.g., a score ranging from 0 to 10), regression trees are used. For the analysis presented in this paper, we will use the tree for predicting the score of a test, that is a numerical variable. So, we will employ regression trees.

Specifically, a tree T is a set of successive splits that group the initial set into C groups, corresponding to the *leaves* of T. A tree is constructed by computing, for each factor to be considered, the information gain (with respect to the target variable) given by splitting the initial population into two groups, using some threshold value of the input variables. In the case of regression trees, the gain is computed as the amount of variance reduction of a split. Every possible split in terms of the input variables lead to a division of the sample units into two separate groups (i.e., their intersection is empty). For growing the tree, an iterative algorithm is used. The algorithm starts with a tree with a single node and successively splits it exploring all possible splits and performing the one that most reduces the variance (see Friedman et al., 2001).

4.2 Network analysis

While classification trees allow us to examine the relationship between the students' performance on tests (i.e., a measure of cognitive aspects) and personal characteristics of the students like gender, school type and MOOC attendance, we need a different mathematical tool to identify clusters within the set of students who answered the questionnaires, which represent a very big and complex set of data. Since the data are qualitative and not quantitative we decided to use network analysis and one of its most challenging area of investigation (Boccaletti et al., 2006; Barrat et al., 2008; Newman, 2010), that is the *community analysis*. Community analysis reveals possible sub-networks (i.e., groups of nodes called communities, or clusters, or modules) characterised by comparatively large internal connectivity, namely the nodes that tend to connect much more with the other nodes of the group than with the rest of the network.

Hence, we use community analysis to recognise clusters of students and figure out students' profiles according to their attitudes. To that end, two students' networks are designed, one for the set of answers to Q1 and one for the set of answers to Q2. The method proposed by (Calderoni et al., 2017) is used to design the networks, where two nodes are linked if they co-participate at the same 'meeting'. In this work, the nodes of the network are the students while the 'meetings' are represented by the same answer to the questions of the affective section QA1 and QA2: the more answer have in common the stronger the link between two nodes, Figure 2 exemplifies such an idea. The personal data collected in Q1 and Q2 represent further attributes of the nodes.

As a consequence of this approach, the two networks are undirected and weighted, the first one (N1 in continuation of the section) is associated to the answers to QA1, while the other one (N2 in the continuation of the section) is associated to QA2.

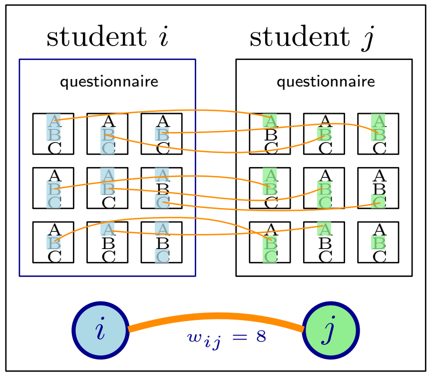


Figure 2: Schema for the design of students' network: students i and j gave 8 same responses to the questionnaire, hence there exists a link between nodes i and j, and its weight is 8.

Since we are interested in identifying sub-networks of students according to their attitudes, we seek for a specific partition of the set of nodes induced by a certain measurable quantity. To that end, we adopt the so-called "Louvain method" (Blondel et al., 2008) based on the optimization of the *modularity* Q. Roughly speaking, given a partition $\{C_1, C_2, \ldots, C_K\}$ of the network, modularity Q is the (normalized) difference between the total weight of links internal to the sub-graphs C_k, and the expected value of such a total weight in a randomized "null network model" suitably defined (Newman, 2006).

To evaluate the goodness or triviality of each community we adopt the *persistence probability* α_k, that measures the 'cohesiveness' of the sub-graph C_k. Radicchi et al. (Radicchi et al., 2004) argue that a sub-network which has $\alpha_k > 0.5$ is defined as *community*. Obviously, the larger α_k, the larger the internal cohesiveness of C_k. Notice that, since α_k tends to grow with the size N_k of C_k it is necessary to test the non triviality of the community (Calderoni et al., 2017; Fortunato and Hric, 2016) introducing the significance of α_k, identified by the standard z-score.

A.A.	Q1	T1	T2	T3	T4	Q2
September 2016	589	535	505	500	331	369
September 2017	231	193	163	181	136	38

Table 1: Number of questionnaires and tests answered by students in the two years under study.

A.A.	Total	Males	Females	LS	HU	TE	Other
September 2016	589	402	150	415	57	55	42
September 2017	231	159	72	128	42	39	22

Table 2: Number and characteristics of students in the two years.

5 Descriptive statistics

Every year, around 1200 freshmen at the Polytechnic of Milan attend the course. Table 1 shows the number of students who answered to 2 questionnaires and 4 tests in the two years under study. Data collected in 2016 were anonymous, while data collected in 2017 were not. This difference between 2016 and 2017 impacts the number of respondents, as Table 1 displays.

In 2016, when questionnaires and tests were anonymous, more students answered than in 2017, when the identity of the respondents was asked. Interestingly, in the non-anonymous setting, the dropout regarding Q2 is huge: only 38 students responded, which corresponds to 16% of the students who answered to Q1 in the same edition. In the anonymous setting (i.e., September 2016), there is a dropout as well, but the percentage of respondents is 63% of the ones who participated to Q1 in the same edition. A first, empirical observation is that, when the students are asked to provide a feedback about the course, being anonymous encourages them to respond. When they are asked to answer to math tests, instead, being anonymous or not seems not to affect their will to respond.

Table 2 shows the distribution of gender and school type of the students in September 2016 and 2017. We can see that the number of males is greater than the one of females, and that the students coming from LS high school represent the majority. This confirms a general trend in STEM studies.

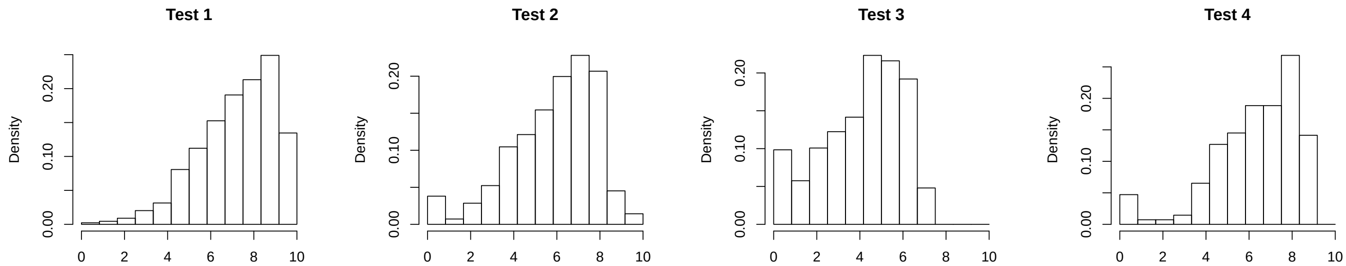


Figure 3: Histograms of the scores in the four tests

6 Results from anonymous investigation

Data were collected anonymously on September 2016. We firstly focus on the cognitive variables and see how the students performed in the tests, in the next subsection. At the same time, we also comment on how gender, school type and MOOC attendance impact test performance.

6.1 2016 test performances

The students' performances during the four mathematical tests is shown in Figure 3. Each test consisted of 10 multiple choice questions. Figure 3 shows the histograms of the number of correct answers out of 10 in each test. Even though it is clear from Figure 3 that some tests are harder for the students than others (e.g., the number of correct answers in test 3 seems to be generally lower than the number of correct answers in the other tests), the distributions in the four histograms is similar. In general we have an asymmetric distribution with a heavy left tail and centred in the medium-high range of the scores. Further, it is also clear that the distributions of the scores can not assumed to be Gaussian: first of all we have only a discrete set of 11 possible score values, and secondly the distributions are asymmetric. While it would be possible - applying some transformation - to symmetrise the data, it is definitely not possible to turn them from a discrete to a continuous distribution. Hence, we avoid to apply transformations to data and focus on nonparametric statistical tools, e.g., methods that do not assume the normality of data.

Since all tests were done in different days, and since data were anonymous, it is not possible to link the scores within students and the distributions of the scores gained in the four tests are considered separately. In order to understand how gender, school type and MOOC attendance influence the test scores, Figure 4 reports the boxplots indicating the relations between these variables and the test scores. The titles of the boxplots report the p-values of a Kruskal–Wallis rank test, that is a nonparametric test for comparing the medians of several groups. The score is significanlty related to the school in the first three tests, it is not related to the gender apart from test 2, and it is not related to the MOOC level.

Note that all statistical tests are performed separately on each factor. For instance, the test on the school does not take into consideration students' gender nor their MOOC level. To be able to consider all variables at the same time, we fit four regression trees to estimate the score of each test. We consider the test score as the target variable, which ranges between 0 and 10. We apply the classification tree method to single out which test score "characterize" different groups of students. We have at our disposal 3 input variables: gender, school type and MOOC attendance. The construction of the classification tree is controlled by the parameter γ, that is used to decide the minimum information gain to be considered for a split. Setting γ to zero would lead to consider all available factors to build the tree, which maximizes the predictive power on the available data, but is likely to be overfitting. Setting γ to values close to one would lead to a tree constituted by a very low number of factors and splits considered, with a very low predictive power. Usually, quite low values of γ are used, for instance performing a cut if it contributes to a decrease of $1\% - 5\%$ of the total variance. For exploratory reasons (for both 2016 and 2017 data) we set the threshold γ very low, and equal to 0.5%. Our choice has a practical reason, since if we set the threshold at 1% in all the threes we obtain only one split, which is not much informative. The choice of 0.5% means that the final leaves of the tree only contribute to a very small decrease of the variance. In all representations and for all nodes, we always include the average score and the percentage of students concerned.

The results of the four regression trees are similar, in terms of the order of the splits that are performed. Here we present and comment the tree obtained for the fourth test, that is the one characterized by a less significant relation between the covariates (school type, gender, and MOOC attendance) and the final score, at least when considering one covariate at a time. Our aim is to show that also in this situation, a regression tree is able to identify a relation between the covariates and the test score and to classify the students into groups with different characteristics. Furthermore, in the context of our research, the students who answered to test 4 were the ones who were present in the last day of the Bridge Course: in this way, we are somehow (and indirectly) able to exclude from the sample those who came to the

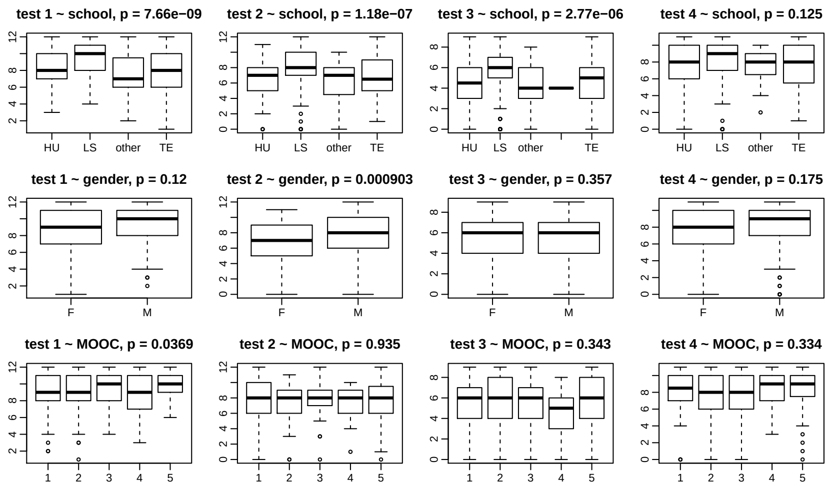


Figure 4: Boxplots indicating the relations between community, covariates (school type, gender and MOOC attendance), and test scores. The titles of each boxplot reports the p-values of a Kruskal–Wallis rank test for comparing the groups' medians.

Bridge Course only for one or two days at the beginning of it, and to consider those who actually attended a significant portion of the Bridge Course. The regression tree for test 4 is reported in Figure 5. The other three ones are provided as Supplementary Material.

From classification tree in Figure 5 we can see that the first split is determined by the school type: students from LS perform better than the ones coming from other school types. In the latter case, no further distinction is made and the average test score for these students is 7.6. Among the students who come from LS, a second split is given by MOOC attendance: those with high attendance perform better than those who attended the MOOC less. However, those who almost never attended the MOOC perform better than the students who partly attended it. Who are these students? Two groups of students are identified, at this stage: one group is made by those students who come from LS and dedicated time to watch the videos in the MOOC and to make exercises (these are the ones with the highest test average, namely 8.9); the other group is made of the students who come from LS, a school type where math curriculum is strong, and hence they do not feel the need to learn more math (on the MOOC). In fact, their performance is good (their average test scores is 8.3, which is higher than 7.6, namely the average of those who come from HU or TE school types). Among those students from LS who partly attended the MOOC, males perform much better than females.

From this analysis, we have seen that the school types is the most influencing variable in test scores, but within the same school type we can identify different sub-groups of students who have different attendance at the MOOC. These differences may be better understood by looking specifically at affective variables (beliefs, attitudes, ...). This is the aim of the next subsection.

6.2 Community analysis on Q1

Students' network N1

Let $N1$ be the undirected weighted network of the students' for Q1. The number of nodes of N1 is $n_1 = 589$ with $L_1 = 170533$ links, so the density of the network is $\rho_1 = 0.985$, which is very high for a social network. Despite this high density, community analysis allows us to identify a partition with 3 clusters (modularity $Q_1 = 0.0567$), whose details are reported in Table 3, that is the dimension of the community, the persistence probability α_k and the *z-score* z_k computed on the sub-graphs corresponding to the community. Table 3 shows that community C_3 is the biggest one, while communities C_1 and C_2 are of comparative sizes. All in all, the three communities are pretty big in size. Persistence probability α_k measures the probability that a random walker on the network would not get out from community C_k, once she has reached a node of it. Even though the low values of persistence probability ($\alpha_k < 0.5$) we can assume that the communities are not trivial due to the high values of the $z-$score. We recall that this score measures the significance of persistence probability, and if $z_k > 5$ we can conclude that the internal cohesiveness of community C_k is good. Therefore, three sub networks of the whole students' network are identified through the students' answers to QA1.

Figure 6 shows the communities' frequency of answers to Q0 in test Q1: we can notice that C_3 has more males in percentage, more LS students (and fewer students from other school type), and more students who did not attend the MOOC course. C_1 has relatively higher percentage of HU students, while community 2 has relatively higher percentage of TE ones. Performing χ^2-tests on the differences between the frequencies of answer to Q1 according to the identified

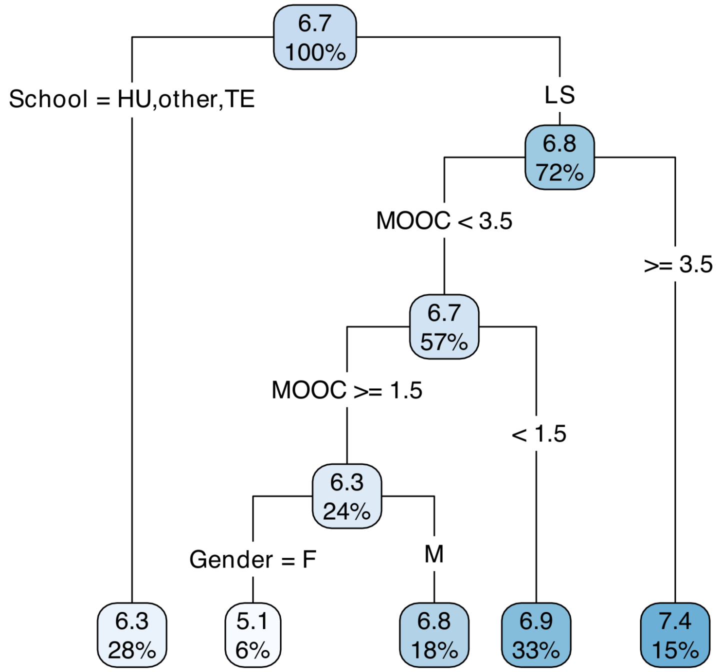


Figure 5: Regression tree for test 4

	N_k	N_k [%]	α_k	z_k
$C_{1,1}^{2016}$	170	28.86	0.321	8.732
$C_{1,2}^{2016}$	182	30.9	0.369	15.5
$C_{1,3}^{2016}$	237	40.24	0.477	18.04

Table 3: Results of max-modularity community analysis for students' network N1

communities gives the results reported in Table 4: it shows that high school is the only one, among the personal-level characteristics, that is statistically different in the three communities. At the same time, it shows that the students in the three communities have answered in significantly different ways at all the questions but QA1.3. The answers mostly given by the students belonging to the different communities to questionnaire QA1 allow us to characterise the three communities as follows.

For the students in C_1, math in high school (question QA1.1) had been mostly formulas to apply and reasoning, and their expectations about math at university (QA1.2) is that it will be used also in other subjects. The role of problems with respect to mathematical learning is, for the students in C_1, to get the students confused. From figure 6, we also see that the majority of these students comes from LS. Community C_1 is, thus, mostly made of students who have a strong mathematical curriculum but have been exposed to procedural mathematics: in fact, math is seen as "formulas", and problems that enhance conceptual understanding are not part of their everyday experience with math.

Community C_2 is made of students who experienced math in high school as problems to solve and reasoning, and their expectations towards math at university is to learn new topics. We can notice a more conceptual approach to mathematics from the students of this community. Even if there are students from LS, there's a significant percentage of students coming from HU and TE.

Finally, in community C_3 the relative percentage of males is higher with respect to the other two communities, and MOOC attendance is lower (see Figure 6). The most frequent answer to question QA1.1 (math in school) is reasoning, followed by problems to solve. Their expectations towards math at university is to deepen their knowledge —and this is not surprising given that the huge majority of them comes from LS, where math curriculum is strong. These students are aware that they have learned many mathematical topics, and their approach to math is rather conceptual.

6.3 Community analysis on Q2

The analysis of the students network N2 is the same of N1. The main properties of the network N2 are similar to the N1's ones, that is the density is $\rho_2 = 0.9822$ since the number of nodes of N2 is $n_2 = 369$ and the number of links is $L_2 = 66687$. The community analysis allows us to identify a partition with three cluster (modularity $Q_2 = 0.0650$),

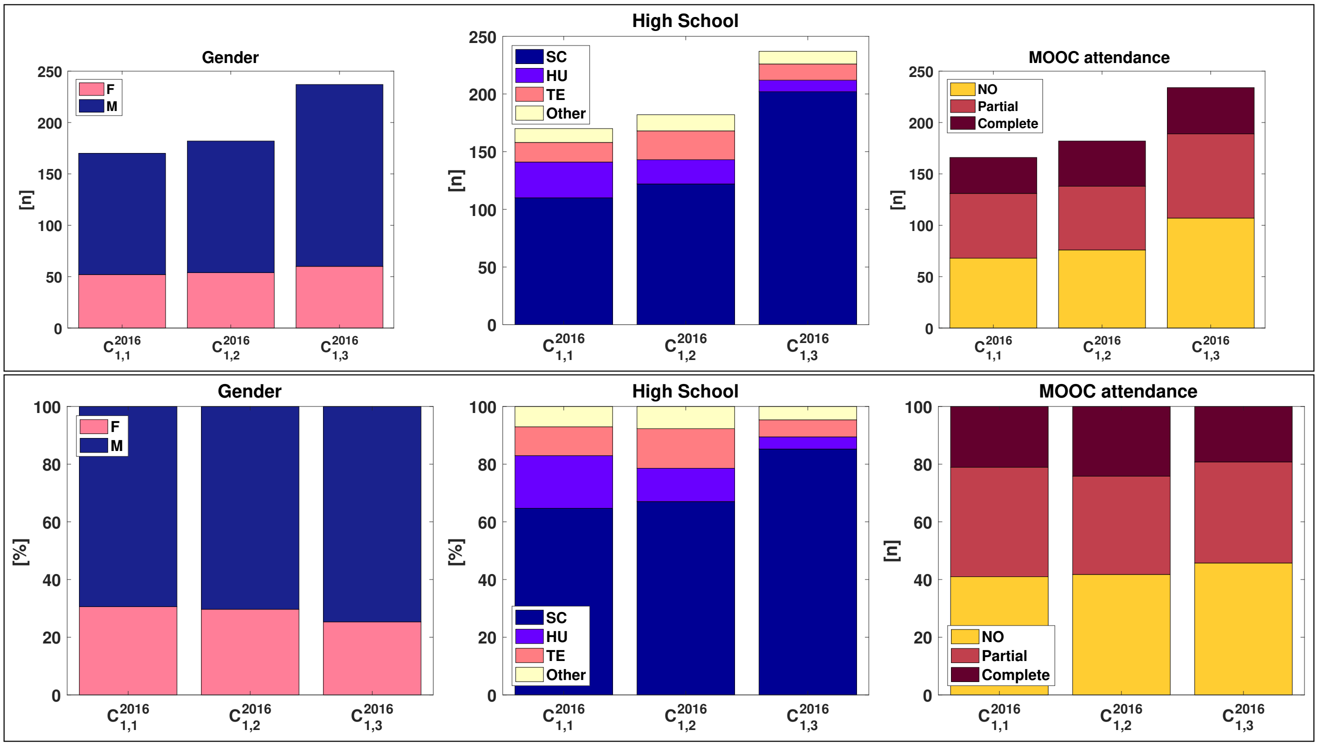


Figure 6: Students' network N1: frequencies (top panel) and percentage (bottom panel) of answers to personal data section of Q1 grouped by community.

variables	X-squared	df	p-value	relevant
Gender	1.647	2	0.4389	×
High School	34.425	6	$5.568 \cdot 10^{-6}$	✓
MOOC attendance	0.60899	2	0.7375	×
QA1.1	96.575	14	$2.143 \cdot 10^{-14}$	✓
QA1.2	584.62	8	$< 2.2 \cdot 10^{-16}$	✓
QA1.3	8.7484	4	0.06771	×
QA1.4	19.014	10	0.04008	✓
QA1.5	98.378	8	$< 2.2 \cdot 10^{-16}$	✓

Table 4: Chi-squared test of the frequencies of answers to Q1 for the three communities.

Community	N_k	N_k [%]	α_k	z_k
$C^2_{2,1}016$	79	21.409	0.250	8.647
$C^2_{2,2}016$	144	39.024	0.482	17.989
$C^2_{2,3}016$	146	39.566	0.455	12.413

Table 5: Results of max-modularity community analysis for students' network N2

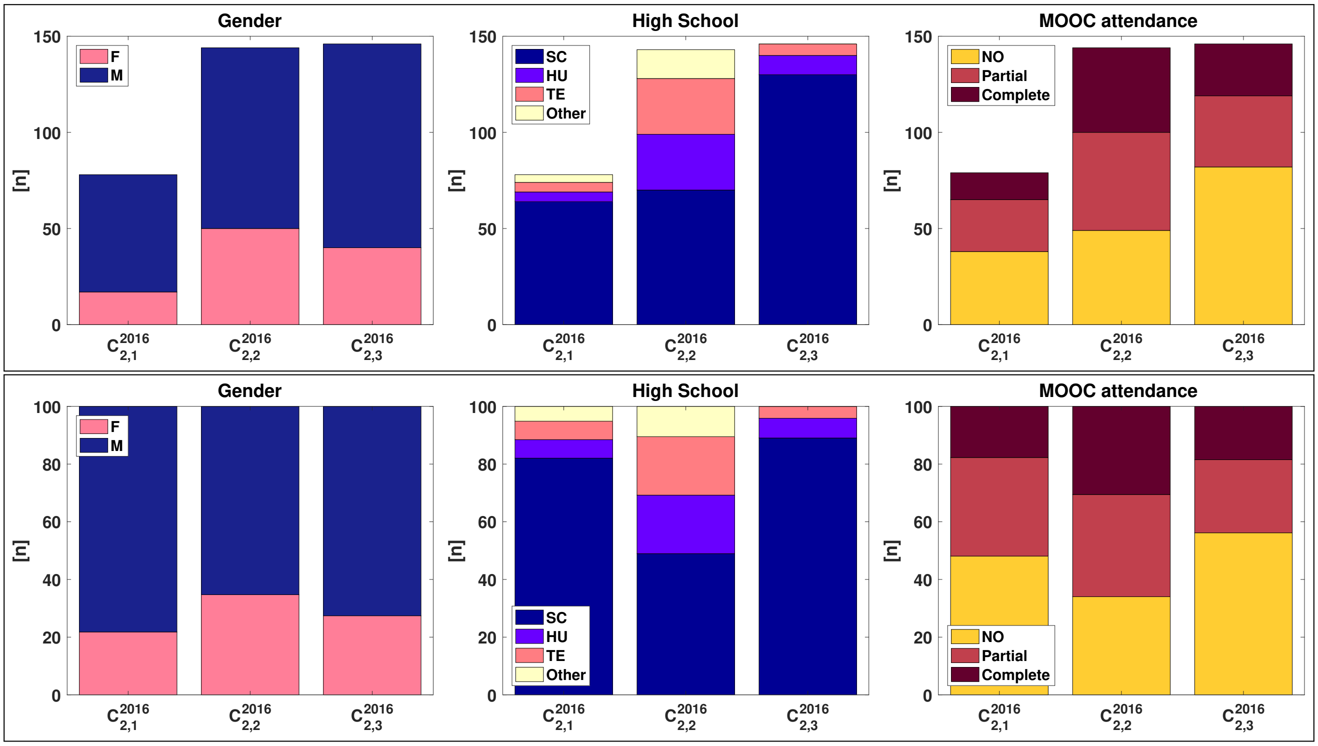


Figure 7: Students' network N2: frequencies (top panel) and percentage (bottom panel) of answers to personal data section of QA2 grouped by community.

whose details are reported in Table 5. The persistence probabilities are $\alpha_k < 0.5$ but the z-score for each one is very high ($z_k > 5$). Therefore three sub networks of the whole students' network are identified due to some some characteristics induced by the students' answers to the QA2.

Figure 7 shows the communities' frequency of answers to Q0 in test Q2: we can notice that community 3 has more males in percentage, more LS students (and fewer students from other school type), and more students who did not attend the MOOC course. Similarly, community 1 has relatively higher percentage of male and LS students, while community 2 has relatively lower percentage of LS ones and the MOOC attendance is almost uniformly distributed among the three level of attendance. Performing χ^2-tests on the differences between the frequencies of answer to Q2 according to the identified communities gives the results reported in Table 6: it shows that the three personal-level characteristics are statistically different in the three communities. At the same time, it shows that the students in the three communities have answered in significantly different ways at all the questions but QA2.3 and QA2.4. The answers mostly given by the students belonging to the different communities to questionnaire QA2 allow us to characterise the three communities as follows.

Profile P_1: The Profile is composed by students who claimed that they did not seen new topics during the course (QA2.1), however they declared that they saw some problems in some ways different from the ones they were used to (QA2.2). These students would appreciate *some extra tutoring as the same style of this course*. Furthermore, these students did not attend the Pre-Calculus MOOC at all and the majority of them came from LS. We can infer that the students of this profile belong to community C_1 that emerged from questionnaire Q1: coming from LS and, thus, having a strong math curriculum, they did not experience new math during the Bridge Course and, being used to procedural mathematics, they appreciated more the attendance part of the course, instead of the e-learning part, since the latter fosters conceptual understanding. The fact that they declare to have been exposed to problems different from the ones they were used to confirms their unfamiliarity with problem-based learning.

variables	X-squared value	degree freedom	p-value	relevant
Gender	21.214	2	$2.474 \cdot 10^{-5}$	✓
High School	50.309	4	$3.112 \cdot 10^{-10}$	✓
MOOC attendance	16.338	4	0.002598	✓
QA2.1	179.13	6	$< 2.2 \cdot 10^{-16}$	✓
QA2.2	217.76	6	$< 2.2 \cdot 10^{-16}$	✓
QA2.3	2.8962	2	0.235	×
QA2.4	11.251	14	0.6662	×
QA2.5	141.79	6	$< 2.2 \cdot 10^{-16}$	✓
QA2.6	38.878	6	$7.562 \cdot 10^{-7}$	✓

Table 6: Chi-square test of the frequencies of answers to QA2 for the three communities.

Profile P_2: The students who belong to this profile said that they have seen some new topics (QA2.1) and some problems were different from the ones they were used to (QA2.2), moreover they appreciated the bridging course (QA2.5) and would like a support on MOOC (QA2.6). The students' sample has a large part of females and the majority of them came from HU, TE and Other school, moreover two third of them attended the Pre-Calculus MOOC. This profile recalls community C_2 that emerged from questionnaire Q1.

Profile P_3: This profile is characterised by students who declared that they did not see any new topics (QA2.1) and problems posed in different ways (QA2.2). The sample is almost composed by LS students who have not attended the Pre-Calculus MOOC however half of them would like a future support to the math exam as the *some extra tutoring as the same style of this course* and even a *support on MOOC* (QA2.6). This profile somehow mirrors community C_3 in Q1.

These are the three profiles of students who attended the last day of attendance course of the bridging course. In the next subsection we come back to the classification tree and try to connect cognitive and affective variables.

6.4 Connections between affective questionnaires and cognitive tests in the anonymous setting

How do affective variables influence test performances of the students? If we go back to the classification tree shown in Figure 5, we can identify the three profiles that emerged from networks N1 and N2:

- after the first split, it emerges a group of students who come from HU, TE and Other schools and who have a lower test performance (average 7.6). These students can be identified by profile P_2: we know that they attended the MOOC more than the students belonging to the other two profiles (and they would like to find a support like a MOOC in the future), that math in high school for them has been mostly problems to solve, that they expect to learn new math topics (and they encountered new math even in the Bridge Course), and that a problem for them serves the purpose of practising and of giving a further example. Hence a procedural view of mathematics emerges from the students belonging to this profile.
- To the right of the split, LS students are identified and the ones with the highest test performance (average test score is 8.9) have also attended the MOOC almost entirely. This group of students, which corresponds to 15% of the sample, seem not to be identified within one of the profiles.
- Among the students who come from LS and attended the MOOC less, we see another split: the leaf of the tree with the students who attended almost no MOOC identifies profile P_1, whose average test score is pretty high (namely, 8.3). We know that these students had been used to a procedural approach to mathematics: the mixed nature of the Bridge Course, however, did not highlighted the weaknesses of their approach to mathematics and we claim that this is good, for two main reasons: the first one is that, among the purposes of the Bridge Course, there is the goal of not discouraging students; the second one, is that these students have a very high performance in mathematics, and hence they do not need to feel as weak learners. Our conjecture is that their rather strong mathematical knowledge, even if procedural in nature, will assist these students in the transition to more conceptual mathematics.

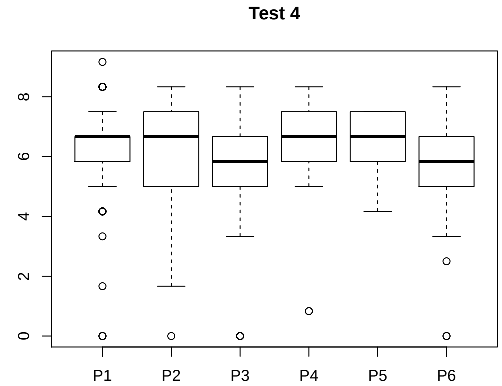


Figure 8: Students' network N1: Test performance by student profiles. P1:Male Students from LS with Partial MOOC attendance. P2: Students from HU, TE and Other with High MOOC attendance and Female students with Hihg MOOC attendance. P3: Students from LS with partial MOOC attendance. P4: Male students from LS with high MOOC attendance. P5: Female students from LS with low MOOC attendace. P6: Students from HU, TE and Other with low and partial MOOC attendance.

- The other leaf of the tree identifies the students coming from profile P_3: males are the majority and perform better that females in the test (average test score is 8.1 versus 6.1). We know that math for them had been mostly reasoning and that they expect to deepen their knowledge.

In an anonymous setting we cannot link students' answers to affective questionnaires with students' answers to mathematical tests. In our context, however, we were able to establish a connection between the two questionnaires and the tests by looking at the features of the students that *most characterize* the communities (i.e., gender, school type and MOOC attendance), and seeing if the same features influence the test scores. On one hand, this is a limitation since the three profiles emerging from the detected communities are not a partition of the students in terms of gender, school type, and MOOC attendance[1]. On the other hand, we claim that - even in a totally anonymous setting - it is possible to identify four overarching, general trends that at a gross grain give a representative picture of well-known phenomena related to dropout.

In order to address the aforementioned limitation, we now dwell on a non-anonymous setting and explore whether a different scenario emerges.

7 Results from non-anonymous investigation

Data collected in 2017 were not anonymous. We have already commented that this feature of the data collection had impacted the percentage of students who gave a feedback on the Bridge Course. We now see whether it is possible to identify profiles that are similar to the anonymous setting, or whether other characteristics emerge. Since, in this case, it is possible to link all the students' answers, we start with community analysis.

7.1 2017 community analysis

In 2017, there is only one network of students. We refer to this network as N2017. It comprises all the students who answered to at least one questionnaire or test during the Bridge Course. However, since only in Q1 we collected personal information, we can perform community analysis only on those nodes (i.e., the students) in the network that have replied to Q1.

The number of students involved is 395, but only 231 answered to Q1. Therefore network N2017 has 231 nodes, and the number of links is 26392 with a high density: $\rho = 0.9935$. Table 7 reports the data related to the community analysis, with 3 communities and modularity $Q = 0.0556$, $Q' = 0.667$, $Q/Q' = 0.083$

[1]Figure 8 shows test performances based on a true partition of the sample into six "profiles". We notice that, to the three main profiles P_1, P_2 and P_3 we can add other three minority profiles: P_4 corresponds to the split that appear in the classification tree and represent 15% of students with very high mathematical performance; P_5 and P_6 correspond to less than 5% of the sample.

	N_k	N_k [%]	α_k	z_k
$C^{2017}_{1,1}$	68	29.4	0.38	15.3
$C^{2017}_{1,2}$	78	33.8	0.37	6.1
$C^{2017}_{1,3}$	85	36.8	0.41	8.0

Table 7: Results of max-modularity community analysis for students' network N2017

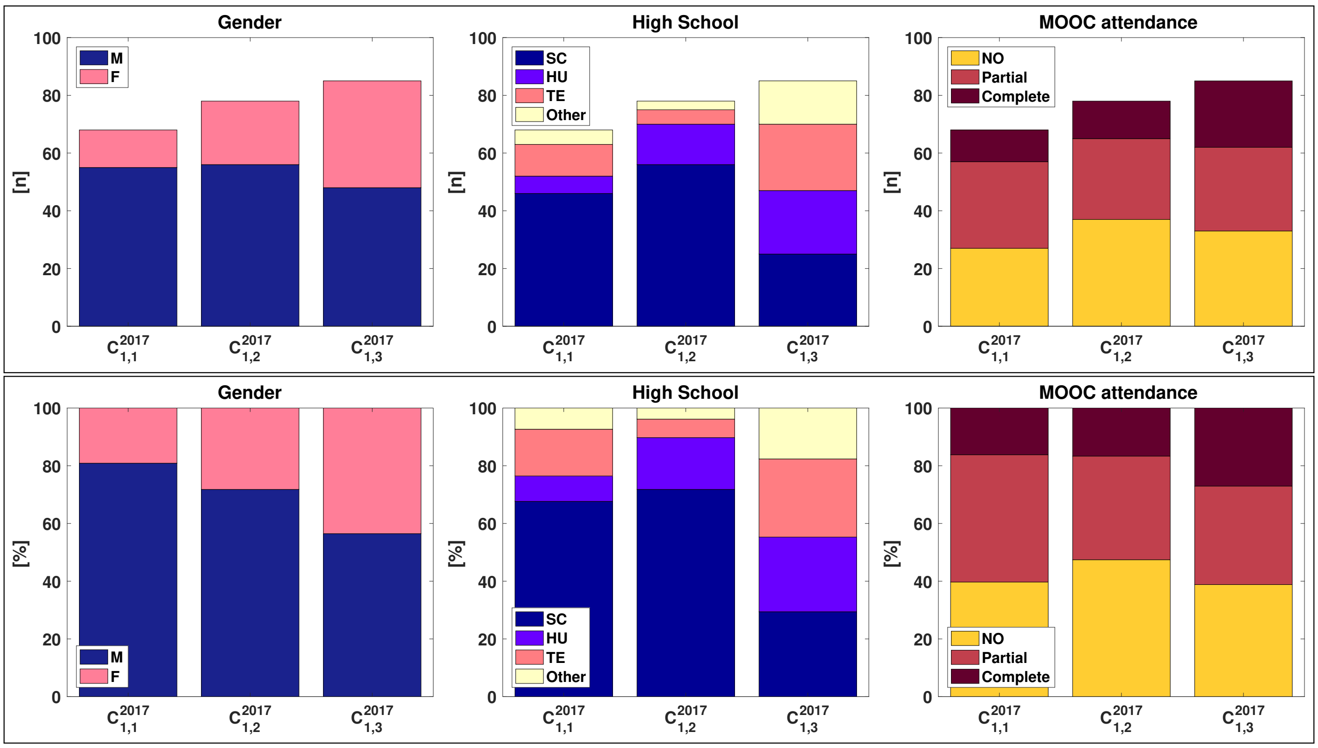


Figure 9: Students' network N2017: frequencies (top panel) and percentage (bottom panel) of answers to personal data section of Q1 grouped by community.

variables	**X-squared**	**df**	**p-value**	**relevant**
Gender	10.976	2	0.0041	✓
High School	40.89	6	$3.043{\cdot}10^{-7}$	✓
MOOC attendance	4.9214	4	0.2955	×
QA1.1	126	12	$< 2.2 \cdot 10^{-16}$	✓
QA1.2	136.76	12	$< 2.2 \cdot 10^{-16}$	✓
QA1.3	57.787	8	$1.264 \cdot 10^{-9}$	✓
QA1.4	258.54	12	$< 2.2 \cdot 10^{-16}$	✓
QA1.5	479.44	24	$< 2.2 \cdot 10^{-16}$	✓

Table 8: Network N2017: Chi-squared test of the frequencies of answers to QA1 for the three communities.

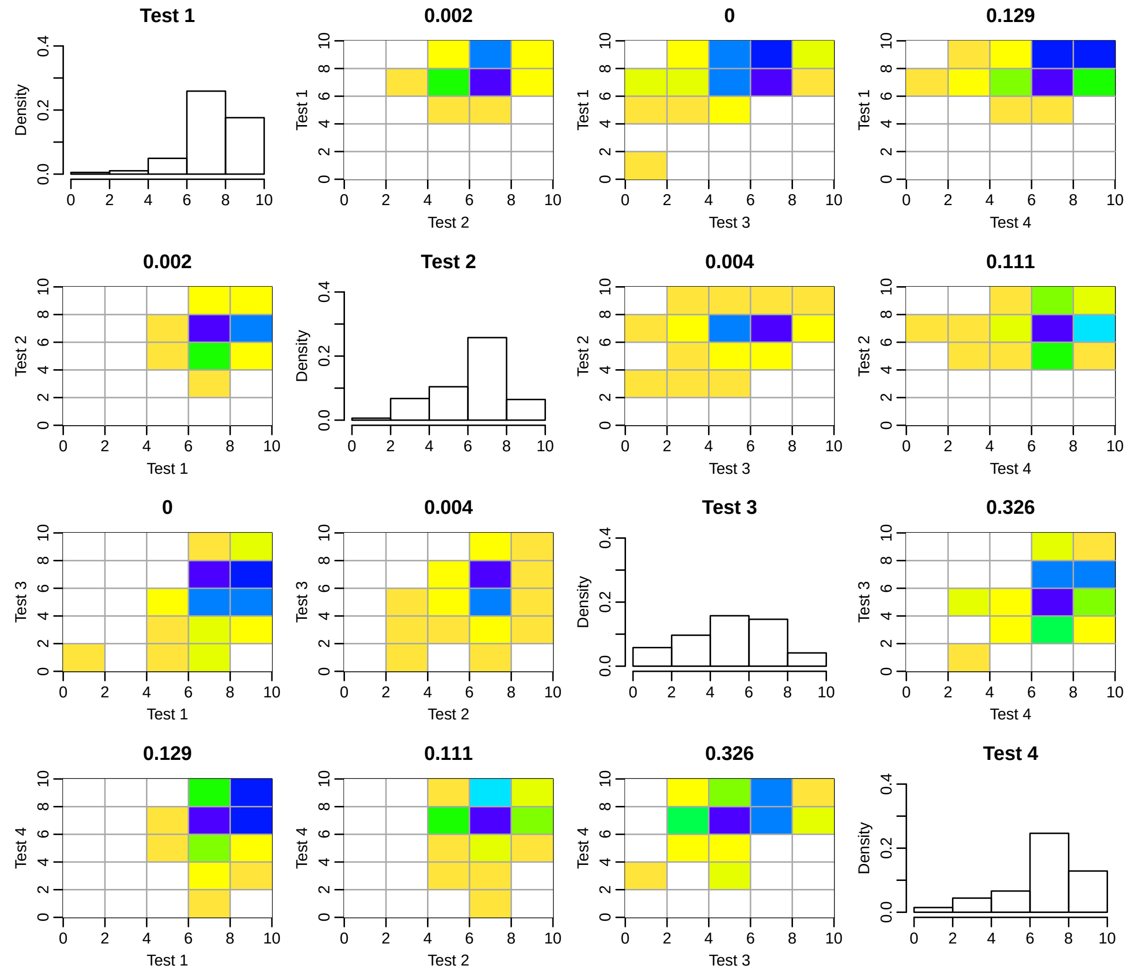


Figure 10: Diagonal: barplots of number of correct answers. Upper and lower diagonals: pairwise correlations among tests. The colour gradation of each cell is proportional to the number of students in the cell. The titles of each upper and lower diagonal plot reports the p-values of an exact-Fisher test of independence.

Student profiles

Profile 1: The students were exposed to a more **procedural** mathematics in high school, they used to resort to an easier exercise/problem/task when they deal with a task they are not able to solve. These students think that a similar exercise helps them to better understand the topic and improve their computational skills. They expect that the mathematics at university is more focused on reasoning and they will be exposed to many new topics. The group is composed by 80% of males, 70% of LS and about 20% of TE students.

Profile 2: The students were exposed to a more **conceptual** mathematics in high school, they used to resort to an easier exercise/problem/task when they deal with a task they are not able to solve. These students think that a similar exercise helps them to better understand the topic and improve their computational skills. They think that the mathematics at university is more focused on reasoning. The sample is composed by 70% of males, 70% of LS and about 20% of **HU** students.

Profile 3: The students were exposed to a more **procedural** mathematics in high school, they used to resort to an easier exercise/problem/task and to **web resources** when they deal with a task they are not able to solve. These students think that a similar exercise helps them to better understand the topic and improve their computational skills. They think that the mathematics at university is more conceptual. The sample is composed by 60% of males, the students come from all kind of high school (25% uniformly).

7.2 2017 test performances (and communities)

We now focus on students' performance on the four tests that were done during the course. The distributions and correlations of the scores of the four tests are displayed in Figure 10. The diagonal panels show the barplots of the

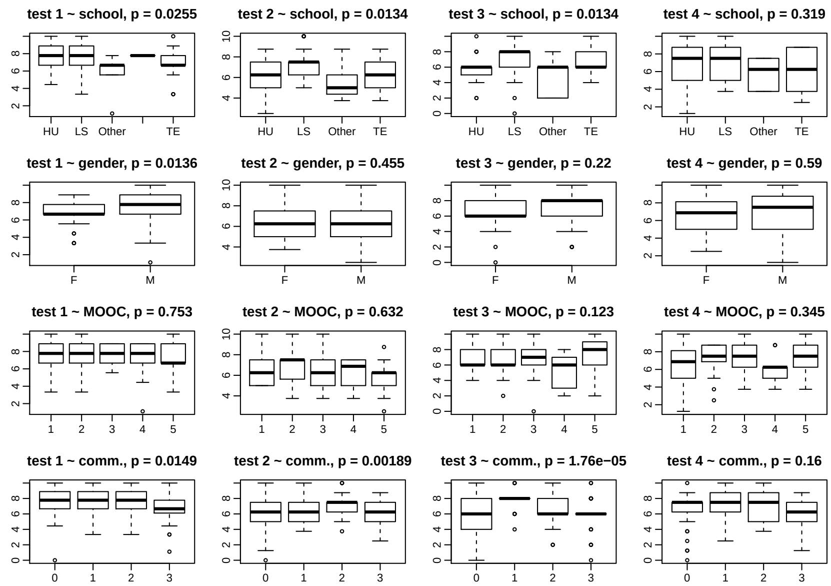


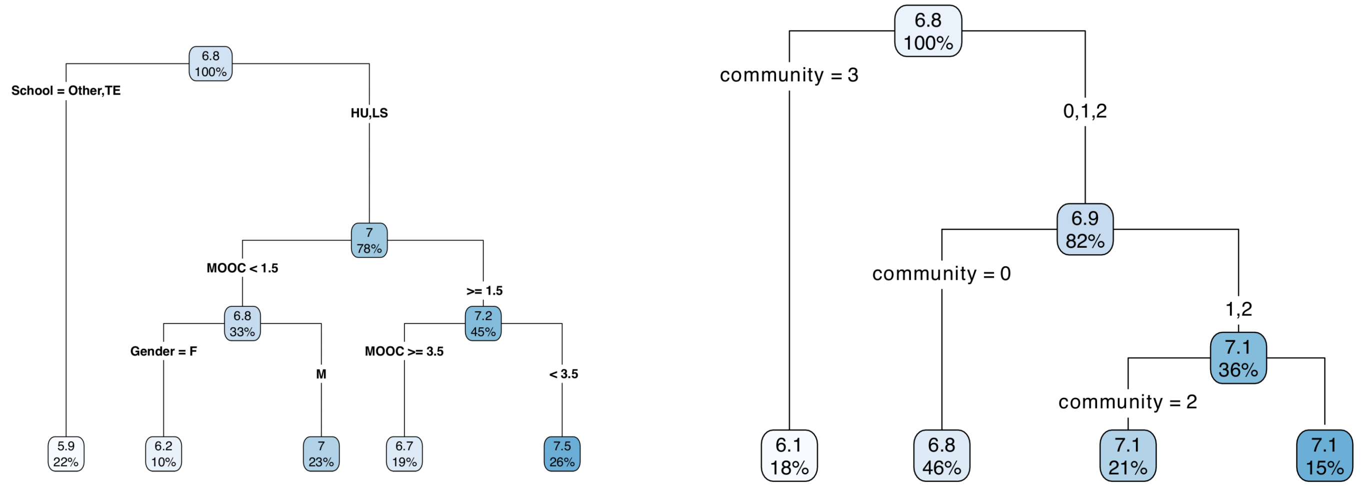
Figure 11: Boxplots indicating the relations between community, covariates (school type, gender and MOOC attendance), and test scores. The titles of each boxplot reports the p-values of a Kruskal–Wallis rank test for comparing the groups' medians.

histograms of the scores on each test. In addition, in this database we have information about which student did the tests. So, it is possible to relate the score of the four tests between them. In detail, the upper and lower diagonal panels show a representation of the correlation between the number of correct answers of students across the tests: the grey color gradation of each cell is proportional to the number of students in the cell. On the title of the plots we report the p-value of a Fisher-exact test of dependence between the scores of each pair of tests. Again, for computing the p-value we used a nonparametric test without assuming Gaussianity of data. Indeed, looking at the histograms on the diagonal, it is clear that Gaussianity cannot be assumed. Looking at the figure, we note that the distribution of the scores of different tests is similar. In addition, the scores of the first three tests are significantly dependent.

The relation between the scores and the other covariates are displayed in Figure 11. As for 2016 data analysis, we indicate in the title of each boxplot the p-value of a Kruskal-Wallis rank test. Furthermore, we include as covariate also the community, that was obtained from the previous network analysis on affect variables. Indeed, since here we can relate students to tests, it is also possible to consider the information about in which community they have been classified. Note that there is a non negligible number of students that performed at least one test but did not answered to the affect questionnaire. All those students are associated to the community "zero" (C_0). The results for school, gender, and MOOC attendance are very similar to the ones observed from 2016 data. In addition, we note that there is a significant relation between test score and community in the first three tests, while we have a high p-value for the fourth one. In general (with the possible exception of test 4), it can be seen that the test score is related to the community: students in communities 1 and 2 have generally a higher score than students in community 3. Students in community 0 have in many cases a higher variability of performances and in some cases a lower score. For the fourth test all p-values of tests on covariates are high, suggesting no significant relations between the score and the covariates, at least when considering one covariate at a time only.

For sake of homogeneity with the 2016 case, we focus on the regression tree for test 4, that is presented in Figure 12(a). The tree is similar to the one obtained from 2016 data. The first split regards the school, with the students from LS and HU performing better than the ones coming from other school types. Among the LS and HU students, a second split is given by MOOC attendance: those who attended the MOOC perform better than those who did not attended the MOOC. Among the students who did not attended the MOOC, males perform better than females, while among students who attended the MOOC, those who attended it only in part perform better than those who attended it completely.

Finally, since in this case students are identifiable, it is possible to see how the test score is affected by the community that was identified with the network analysis. From the boxplot and Kruskal-Wallis test we already saw that in case of test 4 there isn't a significant statistical relation between the test score and the community, but this was only the test 4 case (for all other test, the score is significantly dependent from the community). The lack of significance in case of test 4 might be due from the lower sample size and from the fact that the relation in this case is less strong. For exploratory purposes we present in Figure 12(b) the regression tree obtained for this test using the community as explanatory variable. For the sake of comparison, the trees for the other tests are reported as supplementary material. For test 4 we see that the first split of the tree separates community 3 from the others, with students in community 3 performing worse than students



(a) (b)

Figure 12: Regression tree for test 4 with: (a) students' features as explanatory variables, and (b) community as explanatory variable

in other communities. The second split of the tree separates community 0 (i.e., all students that did not answered to the affect test) from communities 1 and 2. Students in these latter two communities perform in average better than students in community zero. The final split is between communities 1 and 2, but the average score of these two communities is very similar.

8 Conclusions

In this paper, we aim at contributing to understand the phenomenon of drop out among first year STEM university students, phenomenon that is considered as a plague almost anywhere in the world. We, thus, recalled the main factors that can help decision-makers to activate resources in order to reduce drop out by identifying and then intervening on subgroups of students who need personalised intervention at the first year of STEM university studies.

Our findings reveal that three main communities can be identified. The first community is populated by students who had been exposed to a strong curriculum in high school, and who have a rather conceptual view of mathematics. They show good performance in mathematics and they declare that in the Bridge Course they encountered mathematical content that was familiar for them: in fact, their acquaintance with conceptual mathematics (e.g., reasoning) allows them to feel comfortable with the new context of university mathematics, and not to live it as a shock. Finally, they seem to be able to discern which online content is useful for them: indeed, they declared to have partly attended the MOOC and our interpretation is that, since these students are good in mathematics, they selected the contents they actually felt "useful" for them to recall—being able to discard the others and not loosing their time. As pertains this community of students, who represent the strongest group, our suggestion (following upon Clark & Lovric (Clark and Lovric, 2008)) for policy-makers at STEM university is to promote and reinforce their relationships with high schools, especially focusing on secondary school math teacher training, so that teachers will teach their students more conceptual math, in a student-oriented fashion, and their learners will enter the university "well equipped" to deal with the transition.

A second community is as well populated by students who had been exposed to a strong curriculum in mathematics, but with a procedural approach. These students perform less well than their mates in the first community, they did not attended the MOOC and they are able to appreciate only traditional ways of teaching. A majority of males is present in this group. We can further comment that their math performance in tests is good enough, and they declare not to be shocked by the Bridge Course, because their strong mathematical knowledge sustains them in the transition. However, these students seem not to be ready neither for a self-organised way of studying, nor for student-centred learning formats. We expect that these students will face difficulties in the first semester at university, as Andrà, Magnano & Morselli (Andrà et al., 2011) observed in a study conducted in a similar context. Andrà, Magnano & Morselli's (Andrà et al., 2011) findings reveal that these students have the highest probability of not taking the degree, with respect to the students with weakest mathematical curriculum in high school—namely, those belonging to the third community.

Students in the third community are aware that their mathematical knowledge is not enough to attend first year STEM university, and they start to work hard in order to bridge the gap: they attend the MOOC and they come to the Bridge Course. They appreciate the new format of learning, as well as they see novelty in the Bridge Course. According to (Andrà et al., 2011), these students have a probability of getting the degree on time that is comparable to the one of the students in the first community. This tells us that mathematical knowledge is important, but it is also important the student's awareness about her weaknesses. For this reason, we suggest policy-makers at university to make use of (or develop their own) questionnaires that help them detecting the students' attitudes towards mathematics, their beliefs about themselves as learners, and their resilience.

From our findings, it emerges a confirmation of well established findings in analogous contexts. However, there are two elements of novelty in this study: one is the taking into account the students' attitudes towards e-learning materials (MOOCs, in particular), the other one is the idea of clustering students with respect to both personal-level characteristics such as gender and school type, and their views of mathematics, as variables that can explain their mathematics performances. Such an approach allows the researcher to identify groups of students who need differentiated interventions, and hence to take on a decision-oriented approach.

Finally, our study aimed at comparing two ways of collecting data: anonymously, which leaves space for the student to think and to feel free to "be herself", but also encourages them to respond; or identifiable, which allows the researcher to establish stronger connections among data. This is another element of novelty for our research: questioning about better settings for respondents to provide data for the researcher seems to be an under-researched area—which may deserve some attention from a statisticians' perspective. In the identifiable setting the number of analyses that can be performed on data increases, and thus the amount of information that can be possibly gained from such an analysis also increases. Nonetheless, the number of students that are willing to answer to a non-anonymous questionnaire is lower, and results from this setting might be less strong due to the smaller sample size. In addition, nothing can be said about the selection bias in the identifiable setting: it is natural to assume that students willing to answer a non-anonymous questionnaire are not a random sample of all students, but they are likely to be the most interested to the course and the most comfortable about answering. In the case of the data analysed in this paper, taking on an overarching perspective of the phenomenon, we can comment that data collected in the two settings are consistent with each other. Hence we are prone to conclude that - especially in contexts when a finer grain analysis is not at hand - the anonymous setting is slightly better than the other one. In other words, to know the students' names did not add richness in data analysis. This is particularly important in the educational setting, since anonymous data on learners fit within ethical requirements about the accessibility to sensitive data concerning their knowledge, abilities and skills.

Acknoledgements: We wish to thank professor Giulio Magli for his strong support in all the phases of the design and realisation of the Bridge Course. We thank professor Anna Maria Paganoni for her time spent in giving us precious hints on the data analysis. We also thank all the tutors and the students who made this investigation possible by helping us collecting the data and by giving their answers. Above all, our special thanks goes to Giulia Cesari, who transcribed manually the students' anonymous answers to the six questionnaires/tests of 2016, once we discovered that optical reader was not working properly.

References

Andrà, C., Magnano, G., and Morselli, F. (2011). Dropout undergraduate students in mathematics: an exploratory study. In *Current state of research on mathematical beliefs XVII. Proceedings of the MAVI-17 conference*, pages 13–22.

Andrà, C., Magnano, G., and Morselli, F. (2013). Undergraduate mathematics students' career as a decision tree. In et al. (Eds.), M. H., editor, *Proceedings of the 18th Mathematical Views International Conference*, pages 135–146. University of Helsinki, Helsinki, FI.

Barrat, A., Barthelemy, M., and Vespignani, A. (2008). *Dynamical processes on complex networks*. Cambridge university press.

Benn, S. I. and Gaus, G. F. (1983). *Public and private in social life*. Taylor & Francis.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., and Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics reports*, 424(4):175–308.

Bosch, M., Fonseca, C., and Gascón, J. (2004). Incompletitud de las organizaciones matemáticas locales en las instituciones escolares. *Recherches en didactique des mathématiques*, 24(2.3):205–250.

Calderoni, F., Brunetto, D., and Piccardi, C. (2017). Communities in criminal networks: A case study. *Social Networks*, 48:116–125.

Clark, M. and Lovric, M. (2008). Suggestion for a theoretical model for secondary-tertiary transition in mathematics. *Mathematics Education Research Journal*, 20(2):25–37.

Clarke, D. (2001). Teaching/learning. In *Perspectives on practice and meaning in mathematics and science classrooms*, pages 291–320. Springer.

Daskalogianni, K. and Simpson, A. (2001). Beliefs overhang: the transition from school to university. *Proceedings of the British Society for Research into Learning Mathematics*, 21(2):97–108.

Deaux, K. and Major, B. (1987). Putting gender into context: An interactive model of gender-related behavior. *Psychological review*, 94(3):369.

Duby, G. (1985). *preface to Histoire de la vie privée: De l'Europe féodale à la Renaissance*, volume 1. Seuil.

Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1–44.

Fox, J. (2002). Linear mixed models. *Appendix to An R and S-PLUS Companion to Applied Regression.*

Fox, J. and Weisberg, S. (2011). *An R companion to applied regression.* Sage Publications.

Fried, M. N. and Amit, M. (2003). Some reflections on mathematics classroom notebooks and their relationship to the public and private nature of student practices. *Educational Studies in Mathematics*, 53(2):91–112.

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

Gamer, B. E. and Gamer, L. E. (2001). Retention of concepts and skills in traditional and reformed applied calculus. *Mathematics Education Research Journal*, 13(3):165–184.

Gómez-Chacón, I. M., Griese, B., Rösken-Winter, B., and Gónzalez-Guillén, C. (2015). Engineering students in spain and germany–varying and uniform learning strategies. In *CERME 9-Ninth Congress of the European Society for Research in Mathematics Education*, pages 2117–2123.

Gueudet, G. (2008). Investigating the secondary–tertiary transition. *Educational studies in mathematics*, 67(3):237–254.

Hibert, J. and Lefevre, P. (1986). Conceptual and procedural knowledge in mathematics: An introductory analyis. *Conceptual and procedural knowledge; The case of mathematics*, pages 1–23.

Hoyles, C., Newman, K., and Noss, R. (2001). Changing patterns of transition from school to university mathematics. *International Journal of Mathematical Education in Science and Technology*, 32(6):829–845.

Lombardo, V. (2015). Recuperare competenze matematiche all'ingresso di un percorso universitario, analisi di un'esperienza. *Quaderni di Ricerca in Didattica (Mathematics)*, 25.

Mandl, H. and Kopp, B. (2006). Blended learning: Forschungsfragen und perspektiven.

Masci, C., De Witte, K., and Agasisti, T. (2018). The influence of school size, principal characteristics and school management practices on educational performance: An efficiency analysis of italian students attending middle schools. *Socio-Economic Planning Sciences*, 61:52–69.

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

Moore, R. C. (1994). Making the transition to formal proof. *Educational Studies in mathematics*, 27(3):249–266.

National Council of Teachers of Mathematics, N. (2000). standards for school mathematics. reston, va: The national council of teachers of mathematics.

Newman, M. (2010). *Networks: an introduction.* Oxford university press.

Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582.

Niegemann, H. M., Domagk, S., Hessel, S., Hein, A., Hupfer, M., and Zobel, A. (2008). *Kompendium multimediales Lernen*. Springer-Verlag.

Pettersson, K. and Scheja, M. (2008). Algorithmic contexts and learning potentiality: A case study of students' understanding of calculus. *International Journal of Mathematical Education in Science and Technology*, 39(6):767–784.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer-Verlag New York.

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2658–2663.

Robert, A. (1998). Outils d'analyse des cors matématiques. *Recherches en didactique des mathématiques*, 8(2):139–190.

Roesken, B., Hannula, M. S., and Pehkonen, E. (2011). Dimensions of students' views of themselves as learners of mathematics. *ZDM*, 43(4):497–506.

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics. *Handbook for Research on Mathematics Teaching and Learning*, pages 334–370.

Skemp, R. R. (1987). *The psychology of learning mathematics*. Psychology Press.

Tall, D. (1991). *Advanced mathematical thinking*, volume 11. Springer Science & Business Media.

A Supplementary material

A.1 Regression trees for 2016 data

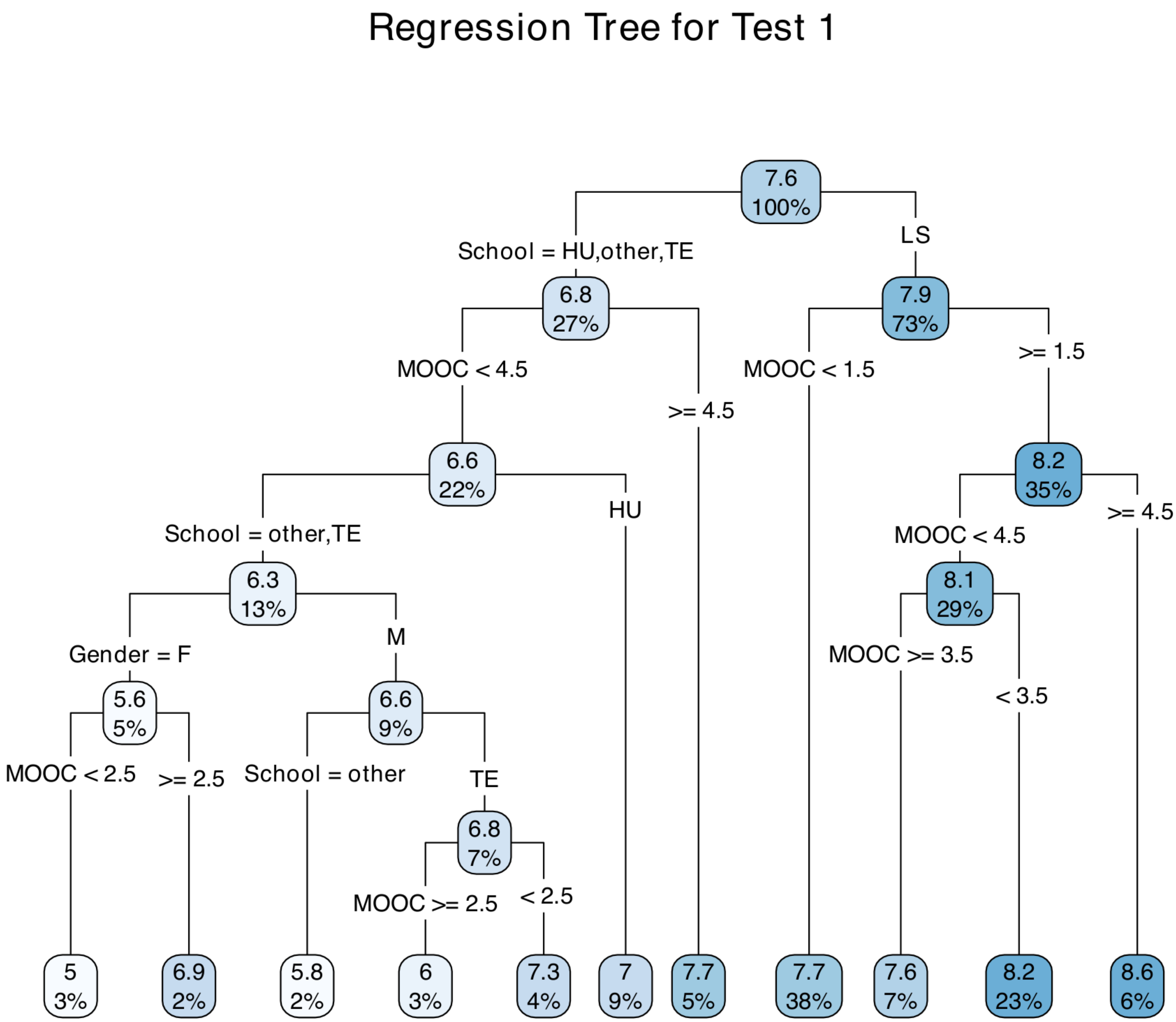


Figure 13: Regression tree for test 1 - 2016 data

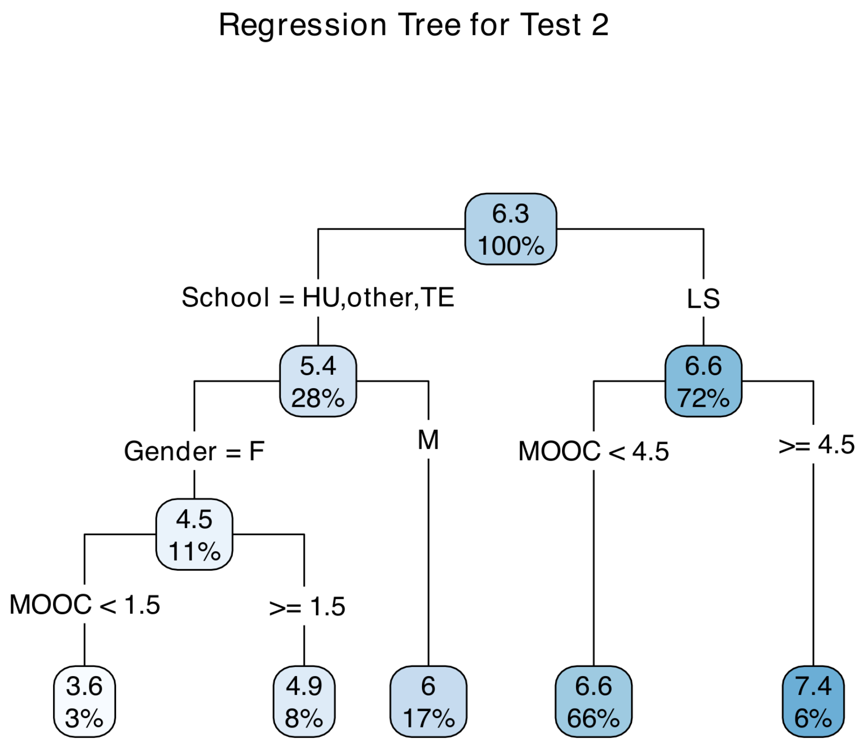


Figure 14: Regression tree for test 2 - 2016 data

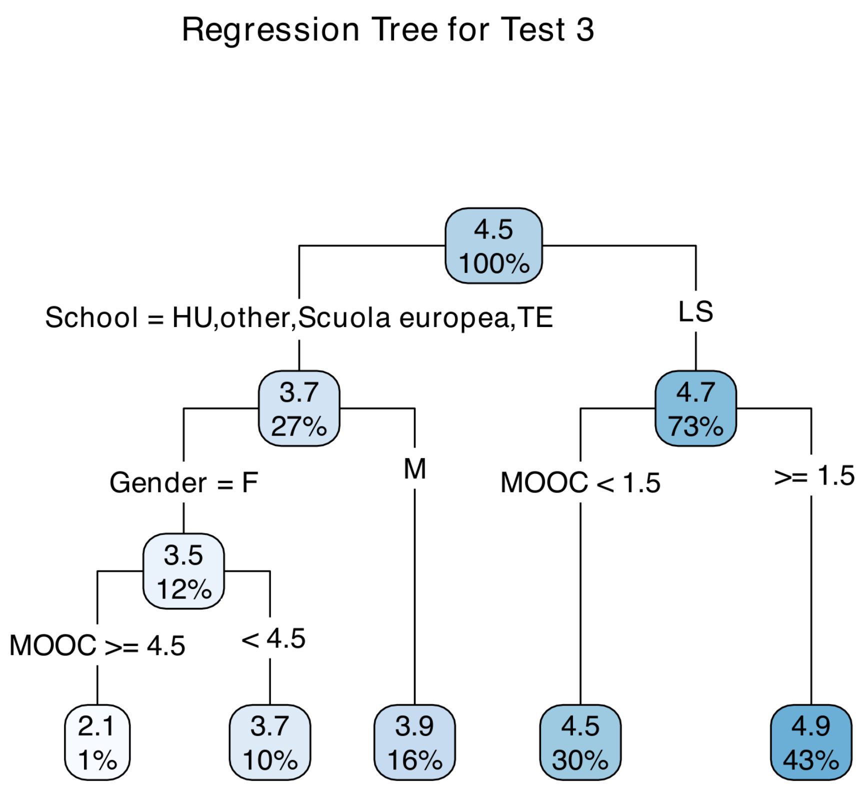


Figure 15: Regression tree for test 3 - 2016 data

A.2 Regression trees for 2017 data

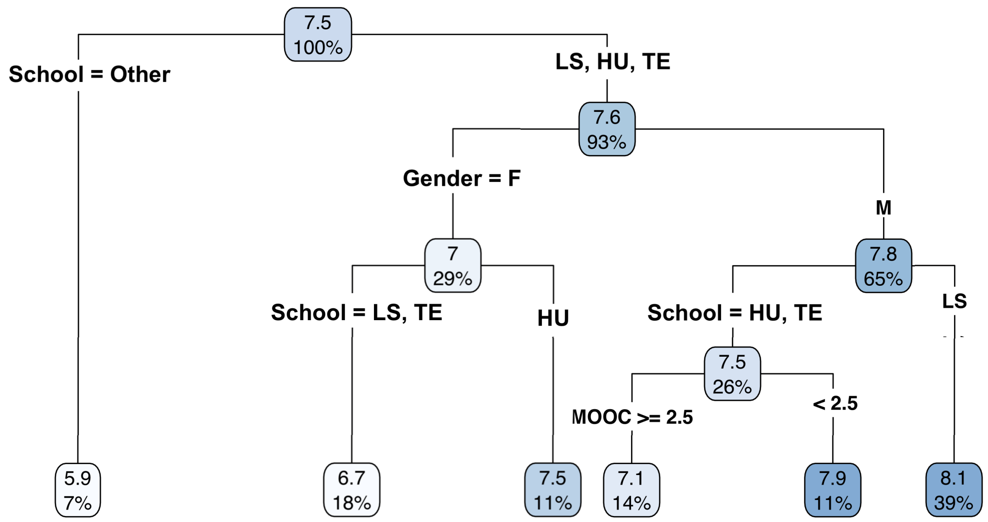


Figure 16: Regression tree for test 1 with students' features as explanatory variables - 2017 data

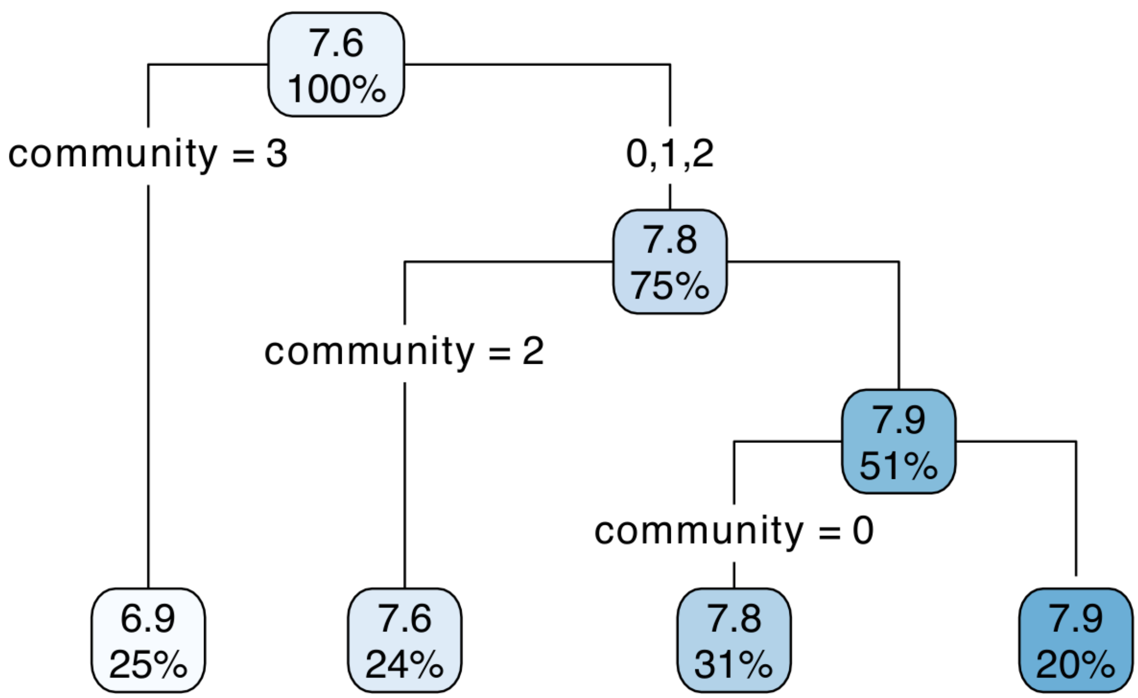


Figure 17: Regression tree for test 1 with community as explanatory variable - 2017 data

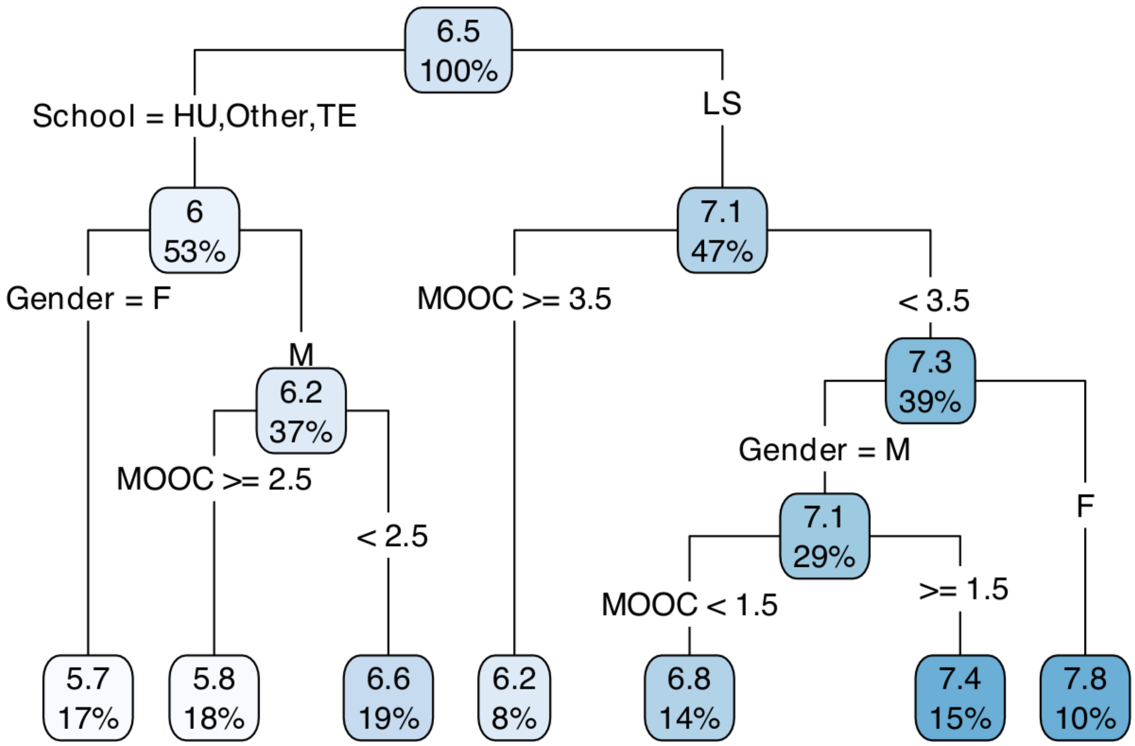


Figure 18: Regression tree for test 2 with students' features as explanatory variables - 2017 data

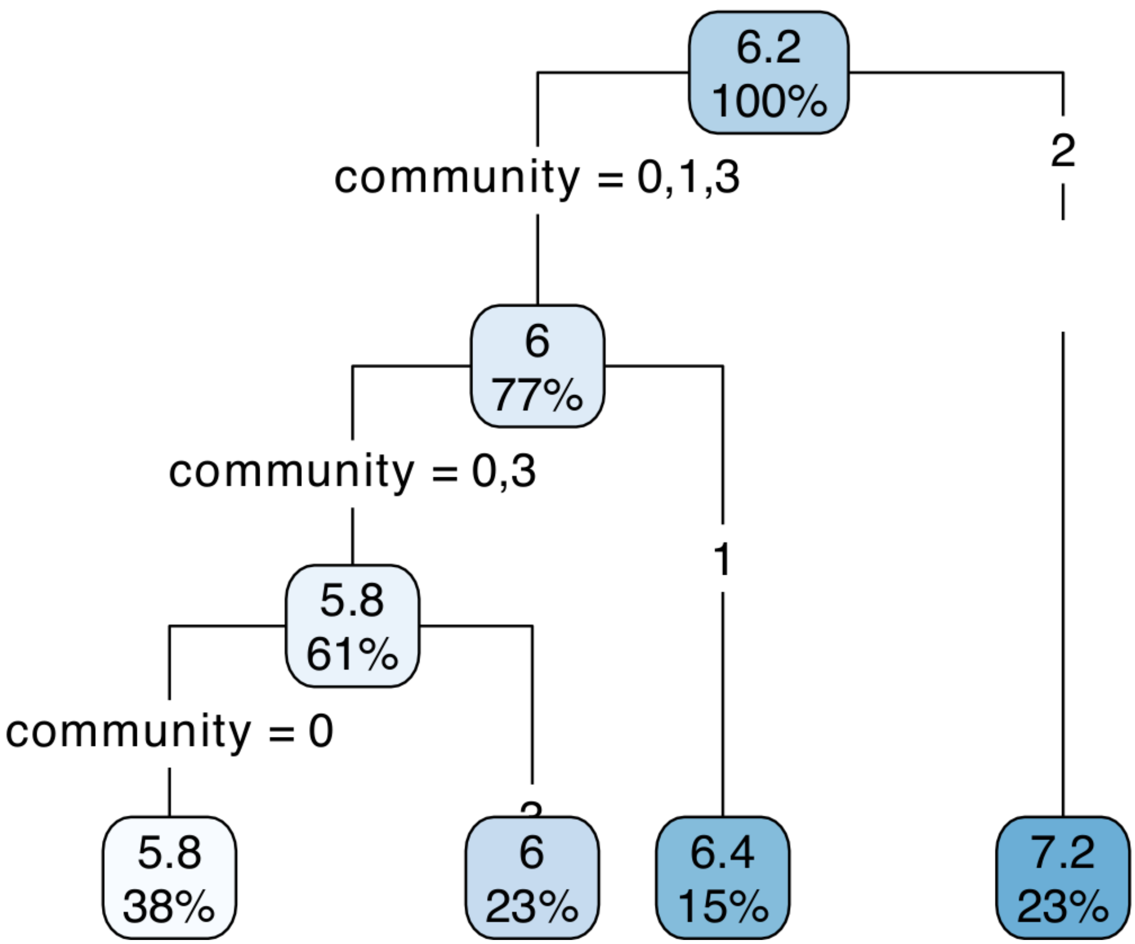


Figure 19: Regression tree for test 2 with community as explanatory variable - 2017 data

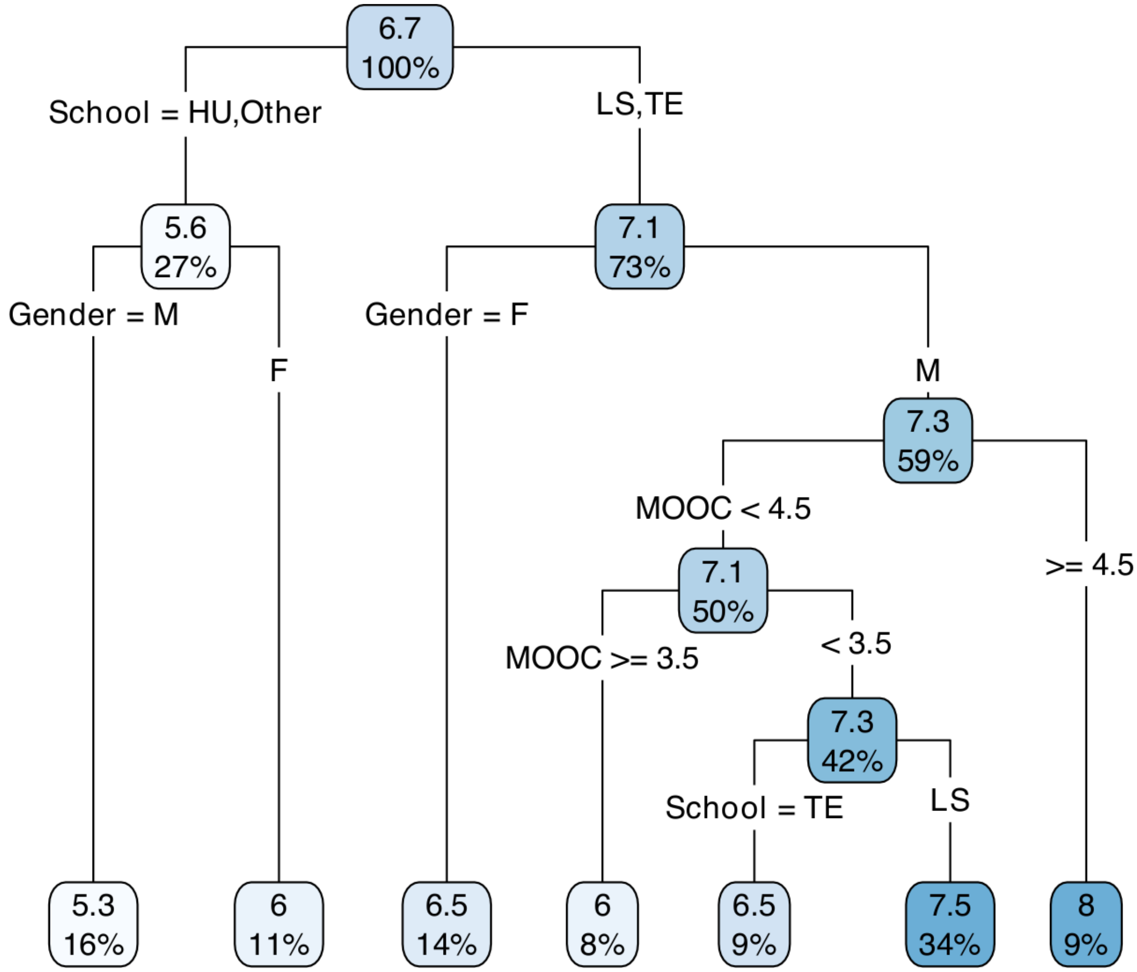


Figure 20: Regression tree for test 3 with students' features as explanatory variables - 2017 data

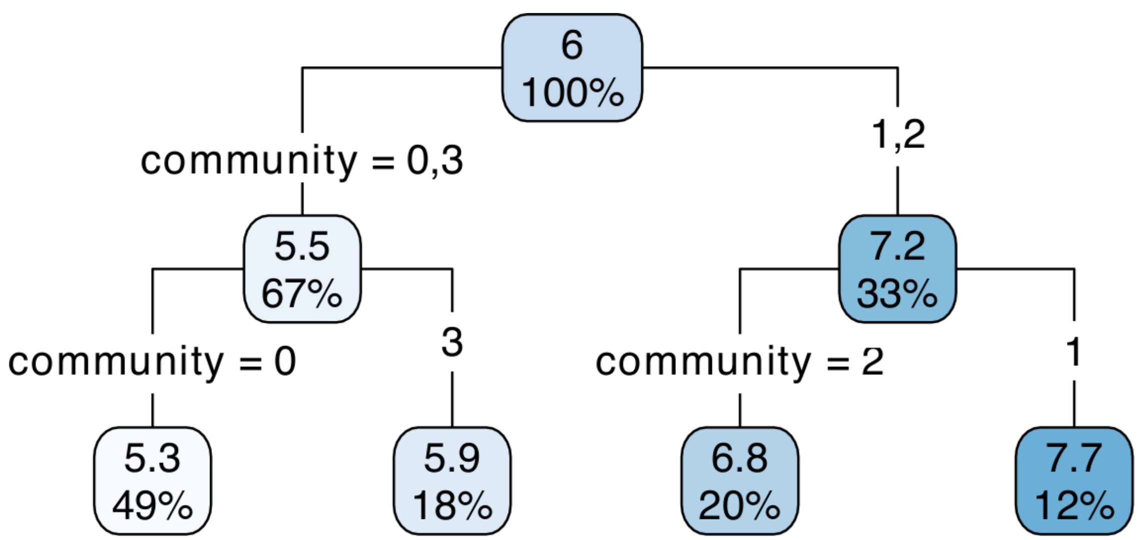


Figure 21: Regression tree for test 3 with community as explanatory variable - 2017 data