

Predictive data analysis with 👉 Je lève la main



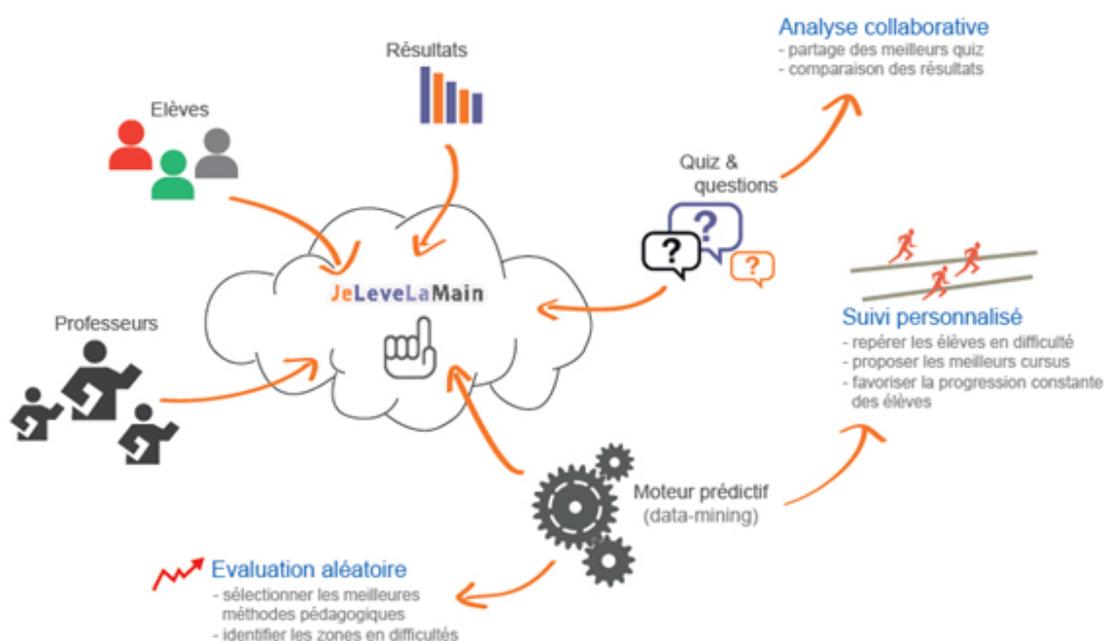
| | |
|-------------------------------------|-----------|
| Introduction | 2 |
| The case study | 4 |
| Creating the experiment | 6 |
| Qualitative analysis | 8 |
| Reduction of the number of criteria | 11 |
| Map of individuals | 15 |
| Clustering | 17 |
| First cluster analysis | 18 |
| Elimination of irrelevant criteria | 21 |
| Second cluster analysis | 23 |
| Quantitative analysis | 27 |
| Decision tree | 28 |
| Prediction | 30 |

Introduction

Predictive analysis has always been one of the principal motivations of our “Je Leve La Main” solution. In a blog post from 2013, entitled “On the road towards predictive analysis”, we were already presenting the principles:

“ Students and teachers connect to “Je Leve La Main” through a unique user name, throughout their schooling. The data is collected frequently and over time and the use that can be made of it goes well beyond traditional tick boxes.

This data can be used to better personalise the teaching of students, to better help them, to give them more suitable schemes of work, to guide them more reliably - even to predict - perhaps - which teacher, which teaching technique has the best chance of allowing this or that pupil to progress.



This makes it possible to develop new predictive evaluation systems (“learning analytics”).

The systematic use of this kind of technology, combined with data-mining techniques (statistical analysis of data), should make it possible to accurately assess students and to better help and guide them.

By applying these statistical techniques, teaching can be improved, by adapting subjects offered to the students in a more personalised way, the monitoring, the methods developed by teachers, etc.

It's about taking a step further making it possible to obtain, for example, the following results:

- *Intervene when students are in difficulty (compared to the average scores of other students) and be able to give them specific intervention*
- *Better personalise students' curricula, according to the results obtained*
- *Predict students' performance and guide them towards their "strong" subjects within the curriculum, while filling in their gaps in "weak" subjects.*

”

In summary, predictive analysis consists of analysing the impact of different criteria (sociological, technological, pedagogical, etc.) on learners' results, with a view to improving their learning process or better guiding them.

It draws on two types of data:

1. the value of the criteria on a population of learners
2. the results of this population in the same set of quizzes

Today, the "Je Leve La Main" application can collect all of this data:

1. the value of the criteria can be filled in either by the teachers (module "I manage my criteria") or directly by the learners (in their profile).
2. the results for the same set of quizzes can be obtained by having each teacher collect the quizzes on the portal and invite learners to answer them in a session (live or recorded).

The analysis of the data can then be carried out thanks to our statistical analysis tool, available at this address:



<https://statistics.jelevelamain.fr/>

A user manual already describes the basic operations of this tool (they will not be included in this document). It is available at this address:



<https://speechi-support.s3.amazonaws.com/JLLM-Virt/Outil-analyses-statistiques/Manuel-stats-JLLM-fr.pdf>

This document complements this manual by presenting advanced modules for predictive analysis: Principal Component Analysis (PCA), clustering, and the predictive module.

The orange framed parts are intended for non-specialist readers who want to deepen the statistical concepts, they can be ignored as part of a quick read.

With the exception of the graphics located in these framed parts, all the illustrations in the manual were automatically generated by the statistical tool "Je Leve La Main".

Tutorial

The green framed parts are intended for readers who wish to follow this manual while learning how to manipulate our statistical tool “Je Leve La Main”.

The “Méthode de lecture” study is available as a tutorial in the software and can be used to generate all the illustrations shown in this manual, based on real data which has been anonymized.

The case study

To illustrate our point, we will use a case study of Year 1 reading levels, in which the evaluation software “Je Leve La Main” allowed us to obtain the following information:

- the results of a panel of 500 Year 1 students in a reading comprehension quiz
- the values of 13 criteria for each student on the panel:
-

Criteria related to school (4 criteria):

| Name | Description | Field | Details |
|--------|---|-------|---|
| tclass | Class size | 0-5 | 0 : < 15 students 1 : 15 - 19 2 : 20 - 24 3 : 25 - 29 4 : 30 - 34 5 : more than 34 |
| pub | Public/ Private school | 0-1 | 0 : Private 1 : Public |
| xpprof | Number of years of Year 1 teaching experience | 0-25 | |
| mlect | Reading method used by the teacher | 0-3 | 0 : Global 1 : Global dominates 2 : Syllabic dominates 3 : Syllabic |

Criteria related to students (9 criteria):

| Name | Description | Field | Details |
|-------|---|-------|---------------------|
| nlivr | Number of books purchased or borrowed per month | 0-10 | |
| nhist | Number of stories read by parents per month | 0-30 | |
| sexe | Gender of student | 0-1 | 0 : Girl 1 : Boy |
| necr | Number of hours spent in front of a screen per day (tablet, television, computer ...) | 0-5 | |

| | | | |
|-------|--------------------------------------|-----|---|
| petu | Education level of the father | 0-5 | 0 : None/ GCSEs 1 : BTEC/ GNVQ 2 : 'A' levels 3 : Undergraduate degree 4 : Postgraduate degree 5 : PhD |
| metu | Education level of the mother | 0-5 | 0 : None/ GCSEs 1 : BTEC/ GNVQ 2 : 'A' levels 3 : Undergraduate degree 4 : Postgraduate degree 5 : PhD |
| mtrav | The student's mother works | 0-1 | 0 : No 1 : Yes |
| ptrav | The student's father works | 0-1 | 0 : No 1 : Yes |
| hdom | Time at which first parent gets home | 0-3 | 0 : <= 5pm 1 : 5:01pm - 6pm 2 : 6:01pm - 7pm 3 : > 7pm |

Creating the experiment

Before we begin, we need to define an experiment for all students who answered the reading comprehension quiz.

For this, we create a new "level assessment" experiment that we will call "reading method".

The screenshot shows the 'Je lève la main Statistics' website. At the top, there is a navigation bar with 'ACCUEIL', 'EXPÉRIENCES', and 'ACP'. Below this, the 'NOUVELLE EXPÉRIENCE' section is visible. It contains a form with a 'Nom' field containing 'méthode de lecture', a 'Type' dropdown menu set to 'Prise de niveau', and a green plus sign button. A text box below the form states: 'The first time you'll reach the "Expériences" section, the "Méthode de lecture" experiment will be automatically available. You can therefore ignore this step.'

We start this experiment and add the quiz “Reading comprehension” to the study’s quizzes.

← méthode de lecture (Prise de niveau) ✎ 🗑

The screenshot shows a 'Quiz' interface with two main sections: 'Quiz disponibles' and 'Quiz de l'étude'. In the 'Quiz disponibles' section, there is a search bar with the text 'Rechercher' and a magnifying glass icon. Below the search bar, a list of quizzes is displayed, each with a right-pointing arrow: 'Schématisation de circuits' (Par Emille Coquel - 0 questions), 'ENERGIE/PUISSANCE' (Par Antoine Usseault - 4 questions), 'WORD 2016 INITIATION' (Par Dominique SIKORA - 1 question), 'UDS TESTING' (Par Samir Lekbir - 8 questions), 'laon' (Par Mélanie Ducloy - 0 questions), and 'apple' (Par Gregory GUILLOU - 1 question). In the 'Quiz de l'étude' section, there is a left-pointing arrow and a selected quiz titled 'Compréhension lecture' (Par Critères JLLM - 20 questions).

To add the quiz “Compréhension lecture” which contains 20 questions, to the study quiz, do a name search in the search bar and click on the arrow on the right-hand side of the quiz.

This screenshot shows the 'Quiz' interface with the search bar in the 'Available quizzes' section containing the text 'compréhension lecture'. The search results list 'Compréhension lecture' (By Critères JLLM - 20 Questions) with a red circle around the right-pointing arrow. The 'Study quizzes' section is currently empty.

The group “Population interviewed”, which includes the persons who answered the study quizzes, is automatically updated. It has 500 people.

The group “Population interrogée” contains all the students that answered the selected quizzes.

The study can be completed by other professors evaluating new students, therefore the “population interrogée” group may be different when you read this manual. To be certain you are using the same data as the ones used while writing this manual, you may use the “étude méthode de lecture” group instead of the “population interrogée” one.

The screenshot shows the 'Groupes' interface with two main sections: 'Groupes disponibles' and 'Groupes de l'étude'. In the 'Groupes disponibles' section, there is a search bar with the text 'Rechercher' and a magnifying glass icon. Below the search bar, a list of groups is displayed, each with a right-pointing arrow: 'Population entière' (with a plus sign icon), 'Population interrogée' (with a person icon and '502 quizid'), and 'Étude méthode de lecture' (with a person icon and '500', and a red circle around the right-pointing arrow). In the 'Groupes de l'étude' section, there is a left-pointing arrow.

Qualitative analysis

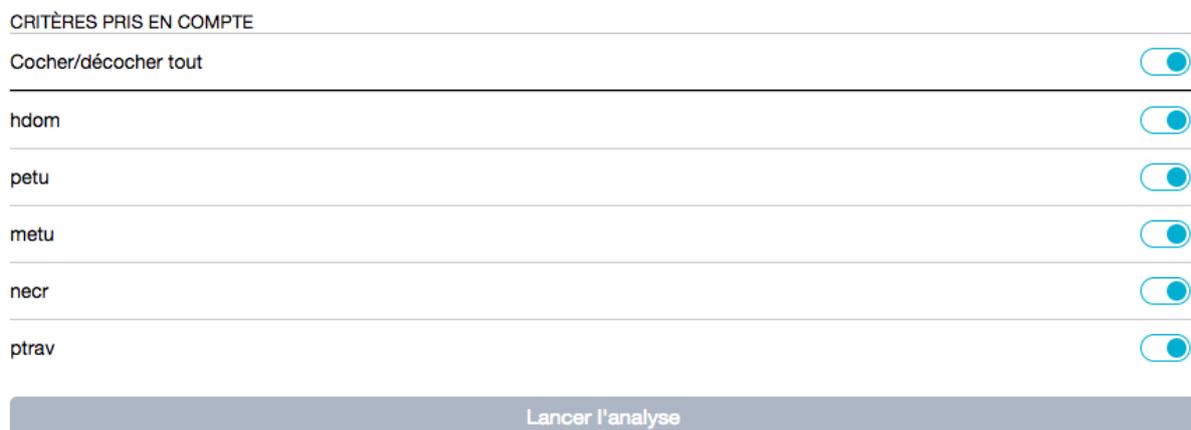
In the “PCA” tab, “Configuration of the analysis” section, we select the study “reading method” that we have just created. The group “Surveyed population” is selected by default, we then click “OK”.



Configuration de l'analyse ▾

Expérience Groupe OK

The list of criteria filled in for the surveyed population then appears. We find the 13 criteria of the study, which are ticked by default (we see here only the first criteria because it is possible to scroll through them vertically).



CRITÈRES PRIS EN COMPTE

| | |
|----------------------|-------------------------------------|
| Cocher/décocher tout | <input checked="" type="checkbox"/> |
| hdom | <input checked="" type="checkbox"/> |
| petu | <input checked="" type="checkbox"/> |
| metu | <input checked="" type="checkbox"/> |
| necr | <input checked="" type="checkbox"/> |
| ptrav | <input checked="" type="checkbox"/> |

Lancer l'analyse

When we click on “Start analysis”, a **Principal Component Analysis (PCA)** is performed.

The attribute of the PCA that we are going to use here is its ability to produce graphs providing a focused point of view of the data, such as the **circle of correlations** (which will allow us to better identify redundant sets of criteria) or the **map of individuals** (which will enable us to better separate individuals on the basis of criteria).

Principal Component Analysis (PCA)

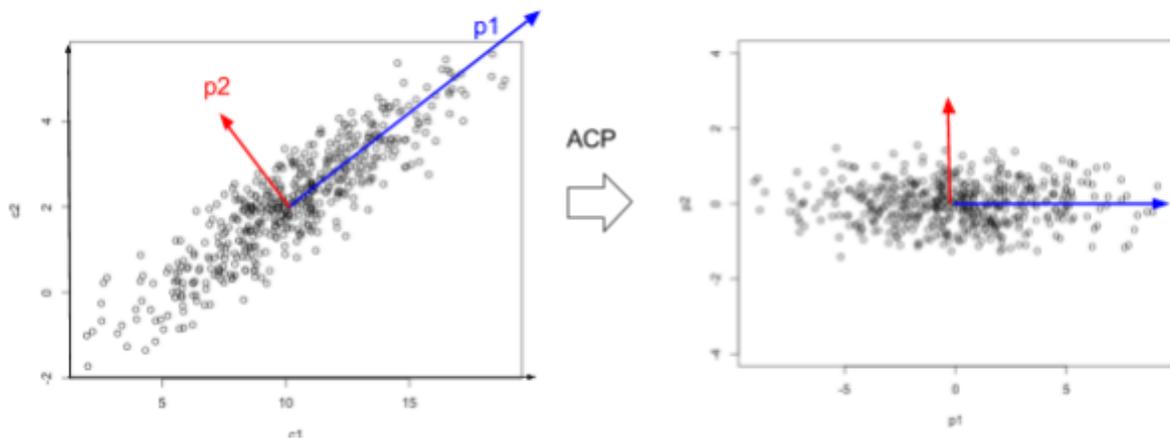
PCA is a technique used for problems involving many variables. It combines linked (or correlated) variables to create new uncorrelated variables called principal components.

The construction of the principal components can be explained geometrically.

Imagine that we know the criteria values c_1 and c_2 of a population. We can represent each individual on axis points c_1 and c_2 . Each individual is a point whose coordinates are the values for criteria c_1 and c_2 .

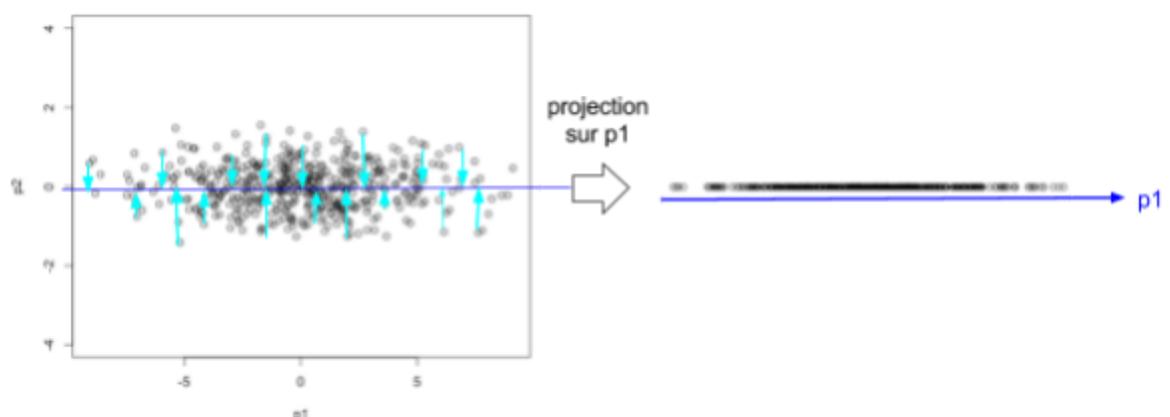
PCA makes a change of reference, in such a way that the cloud of points formed by the individuals spreads as much as possible according to a first axis p_1 (along the straight line described above), then according to a second p_2 , and so on. These new axes are the principal components.

The new point can be seen as a rotation of the initial point. On this new point, the data varies the most according to p_1 , then p_2 , and so on.



In our study, we will use the CPA's ability to provide us with a focused perspective of the data. As individuals vary the most on the first principal components, we can project the cloud of points on these axes by deforming it as little as possible (ie by preserving as far as possible the distances: two distant individuals will remain distant globally, and conversely).

If we take again the cloud of points in our example, we can note that it can be projected on the axis p_1 while keeping the majority of the information on the distances between individuals.



This projection process makes it possible to better highlight groups of individuals with similar values of criteria, called “**clusters**”. In the figure above we can for example identify a small group of individuals on the left, detached from the rest of the population.

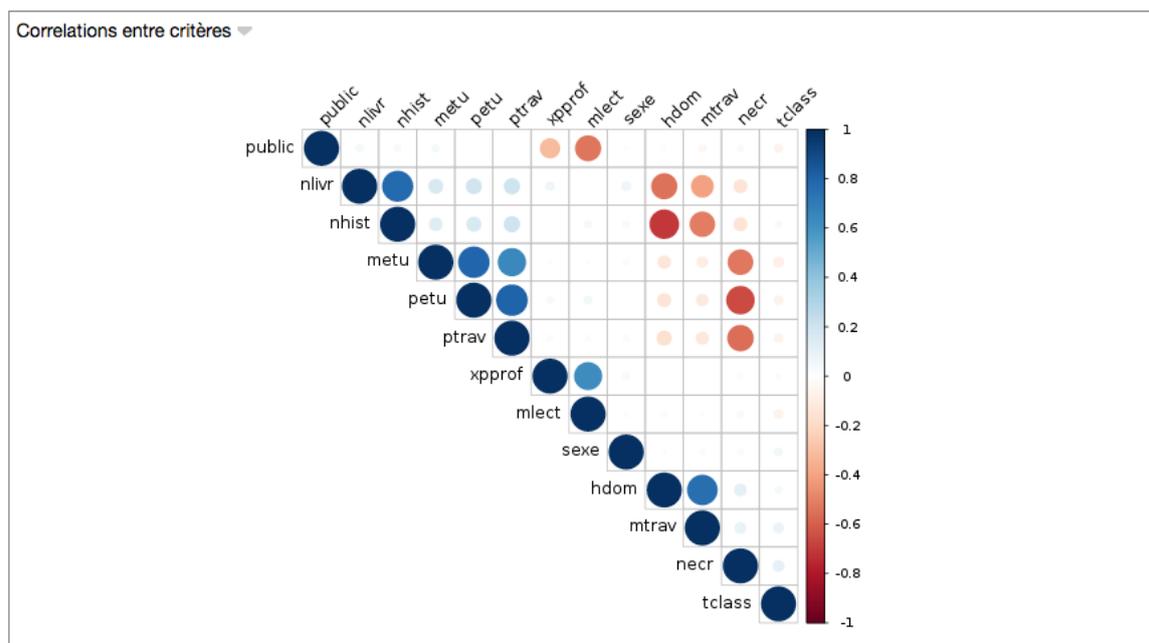
It also allows to observe the population in a convenient way when we have 3 criteria or more: as the visualization of the population in a 3 dimension space or more is not easy, it is useful to represent it in a 2 dimension space (a plan) being as faithful as possible to the initial distribution of the individuals.

This is done by projecting the data into a plan defined by the first two principal components. This projection is called “**map of individuals**” (see dedicated box for more information).

Reduction of the number of criteria

The “Correlations between criteria” section will help us eliminate any redundant criteria.

The first graph in this section is called “Correlation matrix”. For each pair of criteria, a colour point indicates whether these criteria are (linearly) linked or not. A blue dot indicates that both criteria are correlated (blue dot) or anti-correlated (red dot).



Two correlated criteria (correlation close to 1) move in the same direction. For example, the blue dot at the intersection of nivr and nhist means that if nivr increases, nhist increases. In other words, the more books at home, the more stories parents read. Some correlations are more subtle, the one between petu and metu, which indicates parents very often have the same level of education.

Two anti-correlated criteria (correlation close to -1) move in the opposite direction. For example, the red dot at the intersection of nhist and hdom means that if nhist increases, hdom decreases (or conversely: if hdom increases, nhist decreases). In other words, the more parents return home late, the less they read stories to their child.

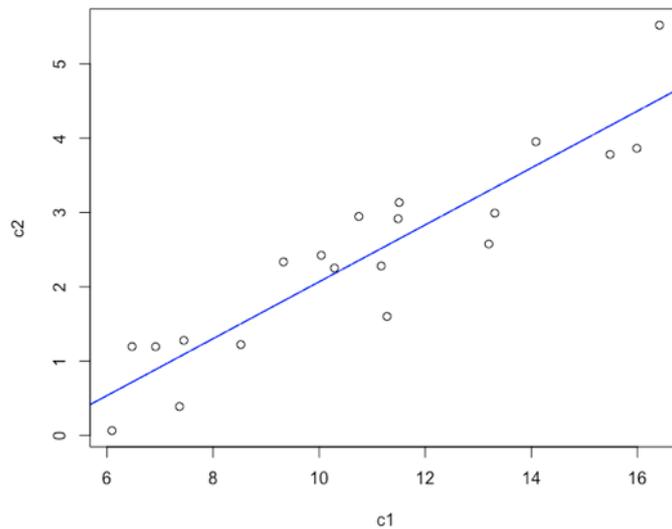
Linear correlation

The correlation coefficient we use in this study is the Pearson coefficient. It measures the strength of a linear link between two criteria.

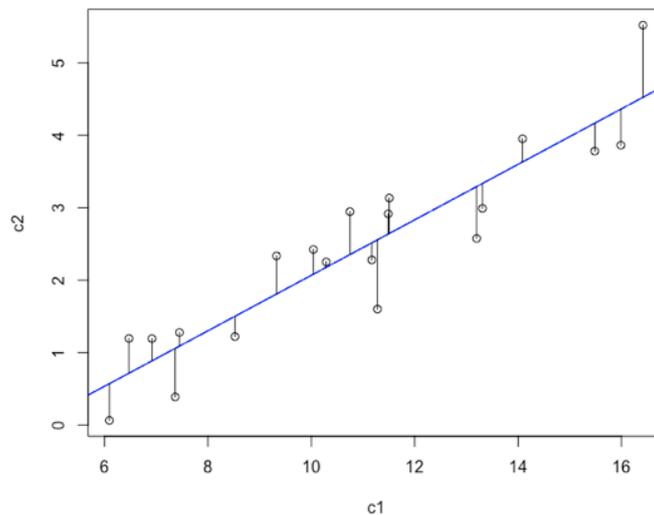
There is a linear link between two criteria c_1 and c_2 if the values of c_2 can be deduced from those of c_1 (and vice versa) by using a linear equation.

This can be seen qualitatively by representing the individuals of a population on a point on axis c_1 and c_2 (each individual is a point whose coordinates are the values for criteria c_1 and c_2).

If c_1 and c_2 are linearly correlated, then the point cloud formed by the population globally describes a straight line, illustrated in blue in the example below:



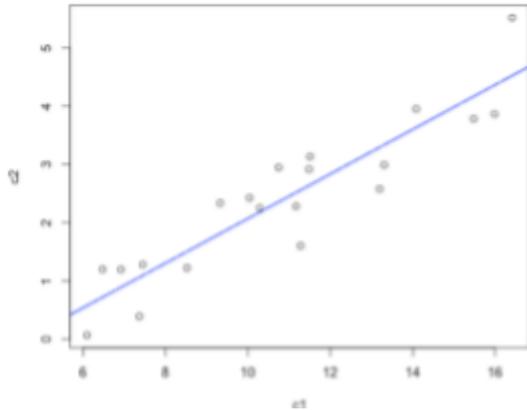
The correlation coefficient gives a quantitative measure of the linear relationship between the variables. It is calculated by measuring the distance between the real value of c_2 and the theoretical value given by the blue line (each distance is represented by a small black segment in the figure below).



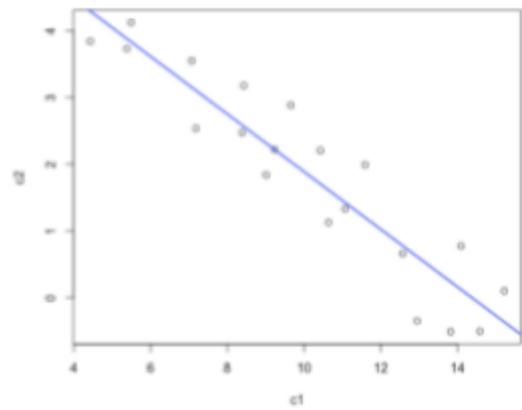
If this coefficient is close to 1 (in absolute value), then this link is strong. In the example above, criteria c_1 and c_2 have a correlation coefficient of approximately 0.91.

The sign of the coefficient indicates whether the criteria are moving in the same direction or not. If it is positive, c_1 and c_2 move in the same direction: if c_1 increases, c_2 increases (this is the case in our

example). If it is negative, they move in the opposite direction: if c_1 increases, c_2 decreases (and vice versa).



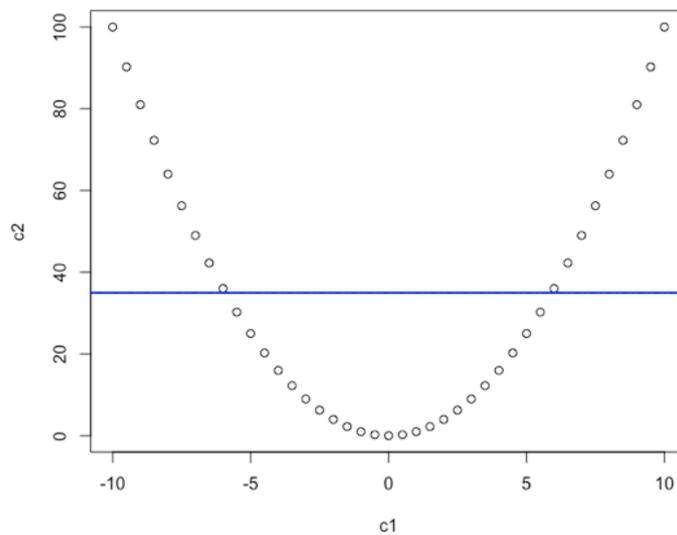
Coeff. correlation close to 1



Coeff. correlation close to -1

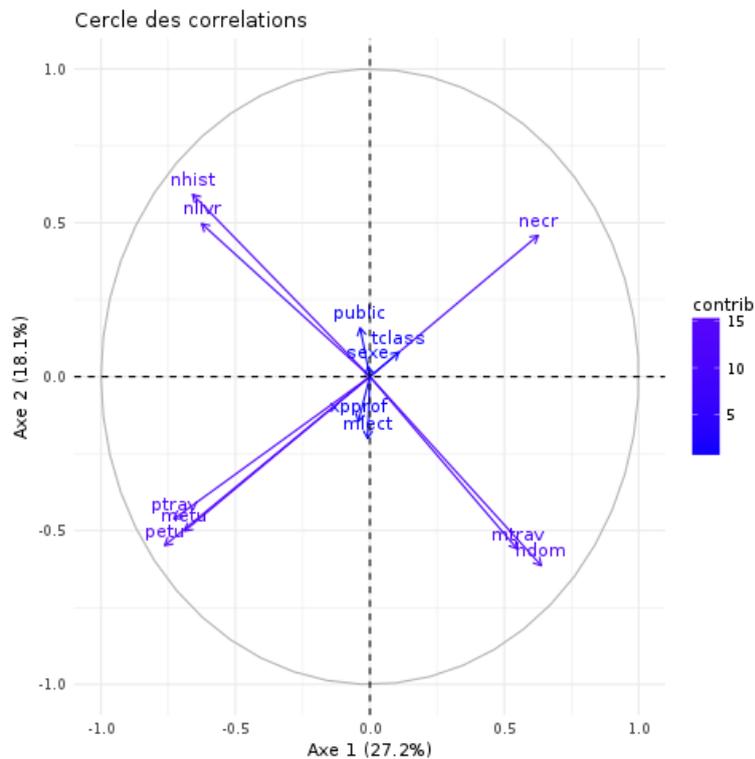
If the coefficient is close to 0, there is no linear relationship between the variables. It is important to note that this does not exclude that there are no other types of more complex links.

For example, in the figure below, there is a strong link between c_1 and c_2 , since we have exactly $c_2 = c_1 \times c_1$ (the points describe a parabola). But the mean (horizontal) line is a poor approximation, resulting in a correlation coefficient of 0.



It's noticeable that many criteria are linked in pairs, but it is not easy to identify sets of redundant criteria. To do this, a more appropriate view is the circle of correlations.

In this representation, the criteria that can be grouped are aligned (they have the same direction, but not necessarily the same meaning). Two criteria in the same direction are correlated (for example petu and ptrav) and two criteria in the opposite direction are anti-correlated (for example petu and necr). It can easily be verified in the correlation matrix.



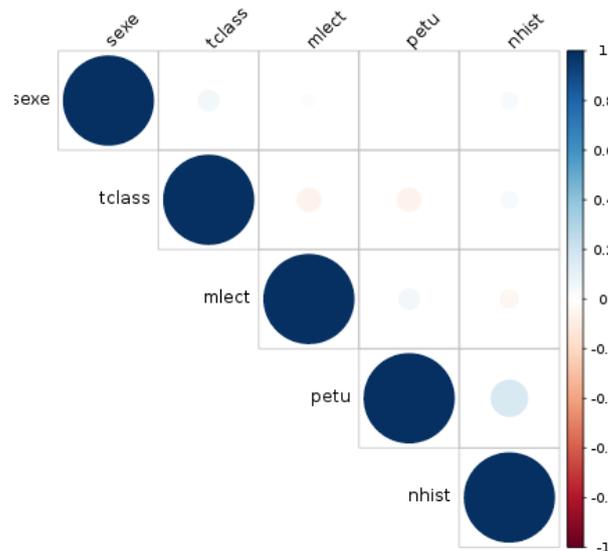
The circle of correlations clearly highlights two areas of study:

1. An axis "socio- professional status of parents" (spro), oriented at 45 °, in which we find petu , ptrav , metu , and necr;
2. an axis "accompanied reading at home" (lecta) oriented at -45°, in which one finds mtrav , hdom , nivr and nhist .

To simplify the study, we will keep only one variable per axis (the others can be deduced with very limited error given the strong correlations):

- the variable petu for the axis "spro"
- the nhist variable for the "lecta" axis

By repeating the operation several times (by unchecking the criteria already processed), we finish with the following set of criteria: petu, nhist, sexe, tclass and mlect. We can verify that the retained criteria is only weakly correlated:



Map of individuals

As described in the “Principal Components Analysis” box, each individual can be represented by a set point using the value of their criteria. A difficulty arises when there are more than 2 criteria: the points are then in a 3 dimension space or more, it is then necessary to project them on a plan to be able to observe them.

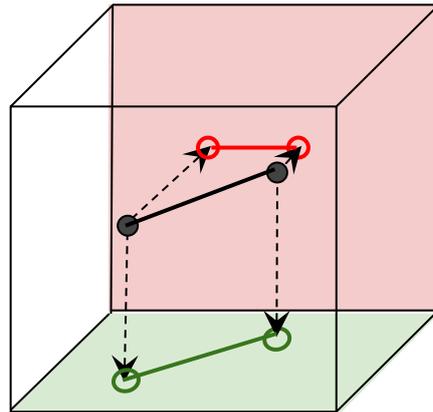
In the map of individuals, individuals are projected on a plan that is defined by the first two principal components (see the “PCA” box for a definition). This allows to spread the most the cloud of points formed by individuals, and thus facilitate the visualization of the clusters presented in the next part.

Map of individuals

Let’s consider that each individual in a population is represented by a point whose coordinates have the value of its criteria. The representation of this population is a cloud of points in a space with n dimensions, where n is the number of criteria.

When $n > 2$, the point cloud must be projected into a 2 dimension space (a plan) in order to be able to visualise it. This projection involves an alteration of the distances that initially separated the points. To illustrate this, let’s imagine two points A and B in dimension 3. The projection of these points on the red plan alters the distances much more than the projection on the green plan.

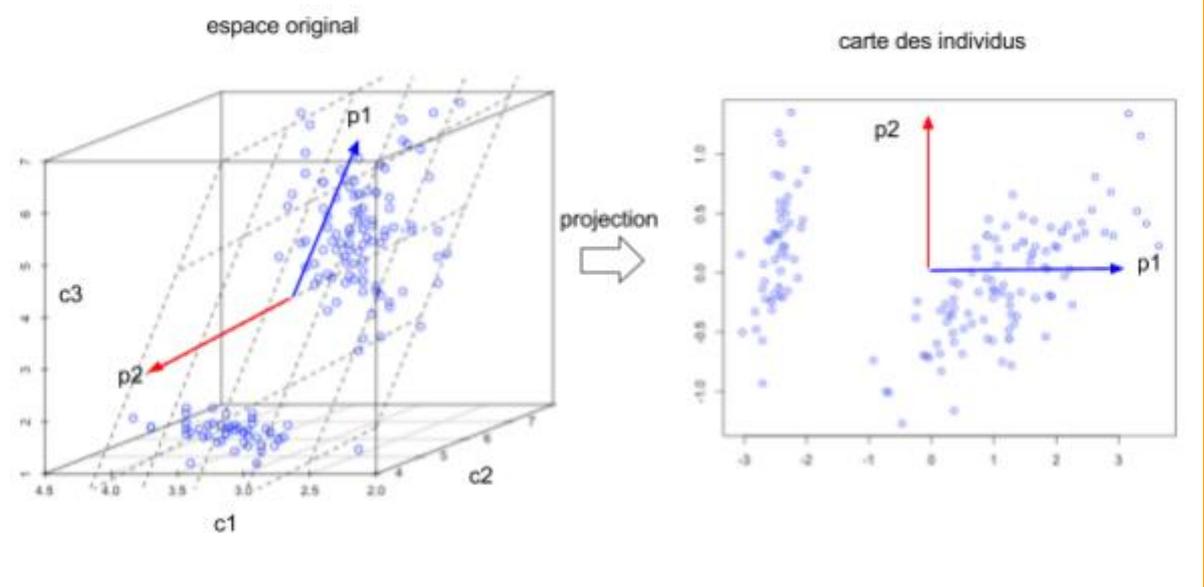
The green plan thus seems a better choice than the red to optimally differentiate the points.



As we saw in the PCA box, the principal components are calculated in such a way that the point cloud spreads the most according to the first principal component p_1 , then the second p_2 , and so on until p_n . The plan defined by the first two principal components p_1 and p_2 is the one that will result in a minimal deformation of the points cloud.

The projection of the population in this particular plan is called “map of individuals”. It makes it possible to better distinguish the possible groupings of points (called clusters).

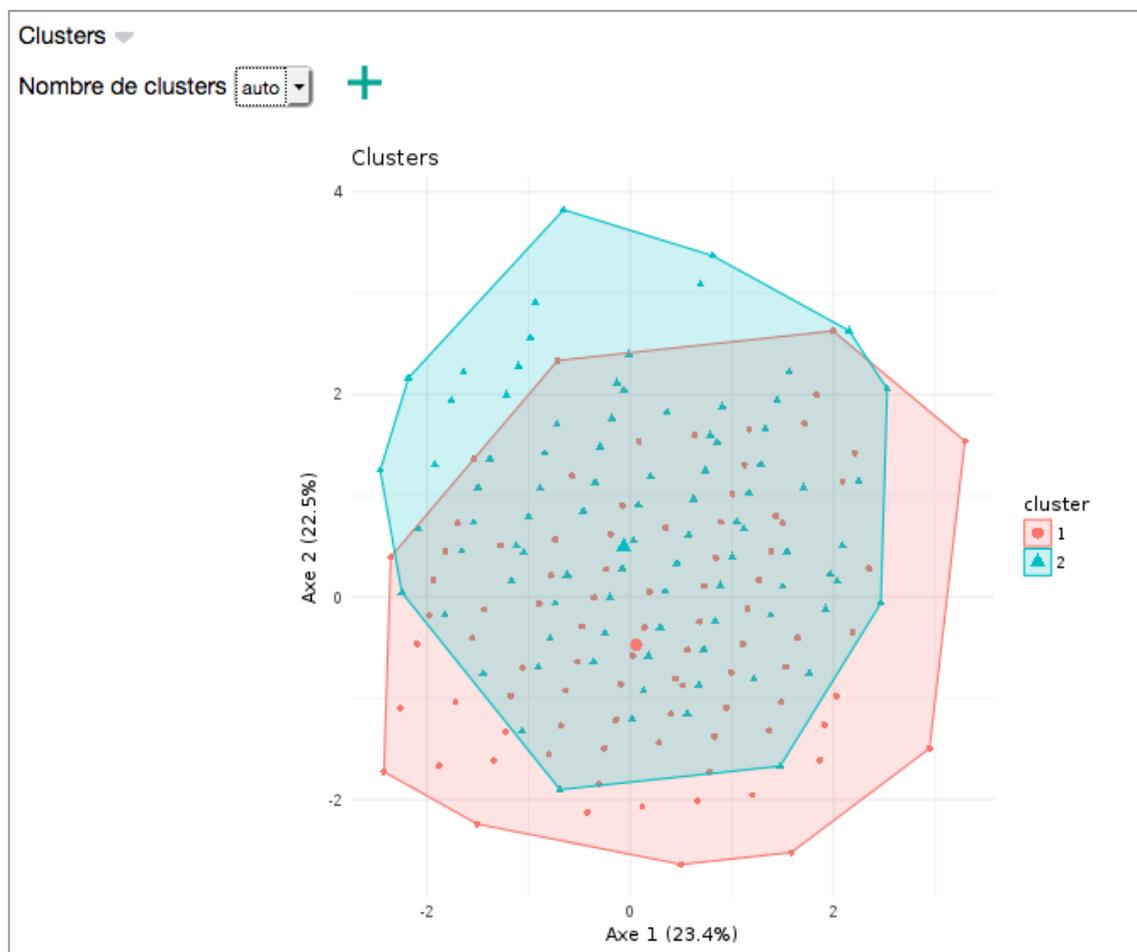
In the example below, we consider a population evaluated on 3 criteria c_1 , c_2 , c_3 . The principal components p_1 and p_2 are illustrated by coloured arrows, and the corresponding dotted plan. After projection on this plan, the two groups of points shown in 3D remain clearly distinguishable.



Clustering

Clustering allows to form groups as homogeneous as possible in terms of criteria values. These groups are called clusters. Individuals belonging to the same cluster will have similar criteria values.

The figure below shows the clustering of the population performed by the tool based on the values of the 5 selected criteria. By default, the tool automatically calculates the number of clusters (here 2).



The clusters are totally distinct (their intersection is empty). But in 2 dimensions, clusters seem to overlap because they are projected in a 2 dimension space from a larger space (5 dimensions here). Representing the clusters in the map of individuals limits these overlays but does not make it possible to eliminate them.

We will add this clustering as a group in our “reading method” experiment. This will allow us to operate on this group in the “Experiments” tab.

Group concept

A group is a set of “Je Leve La Main” users satisfying a set of constraints within certain criteria. In the “Experiments” tab, for example, we can create groups by geographical constraint on different scales: all the students of a school, a city, a county, etc. Criteria (personal or shared by other users) can also be used to create groups, such as grouping students by gender, socio-professional category of parents, and so on. See the manual for more details.

The “clustering” part of the PCA module enables groups to be automatically created by grouping individuals by similarity within the criteria. A group is created by cluster.

To do this, we click on , we keep the default name “clustering1” and click OK.

Nombre de clusters  Ajouter à l'expérience sous le nom:

First cluster analysis

The clustering that we have just generated will allow us to attempt to establish a link between the value of the criteria that characterises these clusters and the level of reading of the individuals belonging to these clusters.

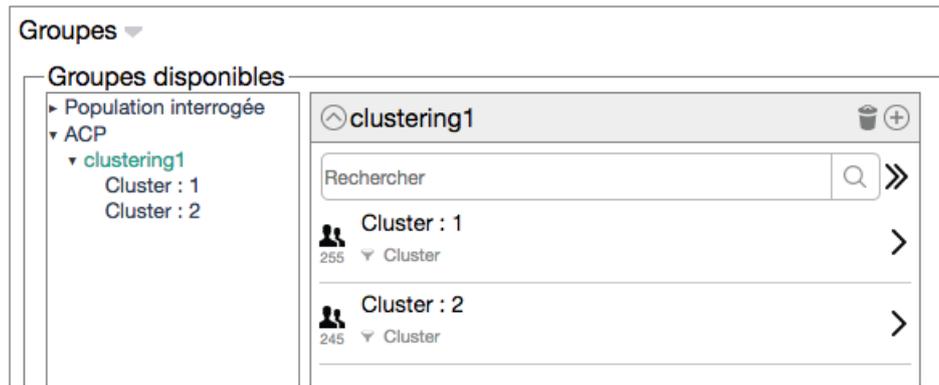
Experimental analysis method

The objective of an PCA analysis with “Je Leve La Main” is indeed to obtain clusters to highlight differences in levels.

In this case, this means that the criteria selected to obtain these clusters is relevant and is related to the level attained by the students. In the opposite case, the PCA does not show any link between criteria and level reached - which means that the experiment is negative.

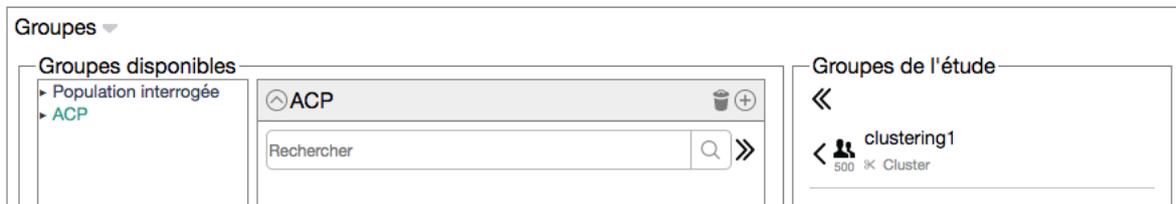
For this, we go to the tab “Experiments” and select the study “reading method”.

In the “Groups” section, we see that a new section has appeared: “PCA”. This part contains all the clusterings calculated in the PCA section - for the moment only the clustering named “clustering1”.

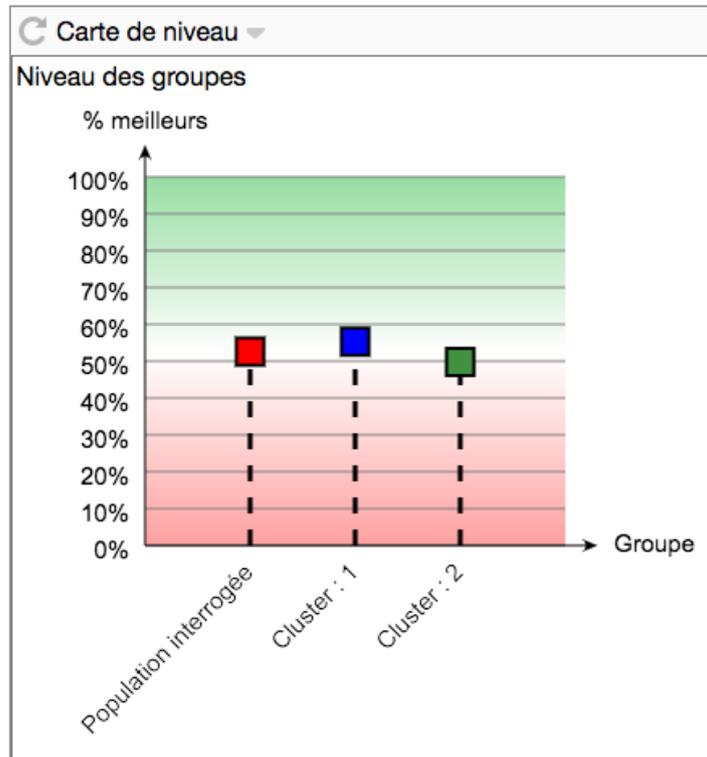


We note that the group “clustering1” is a division of the population surveyed. It contains all the individuals of the surveyed population (500 people), and is divided into 2 subgroups “Cluster: 1” (255 people) and “Cluster 2” (245 people). We have $255 + 245 = 500$.

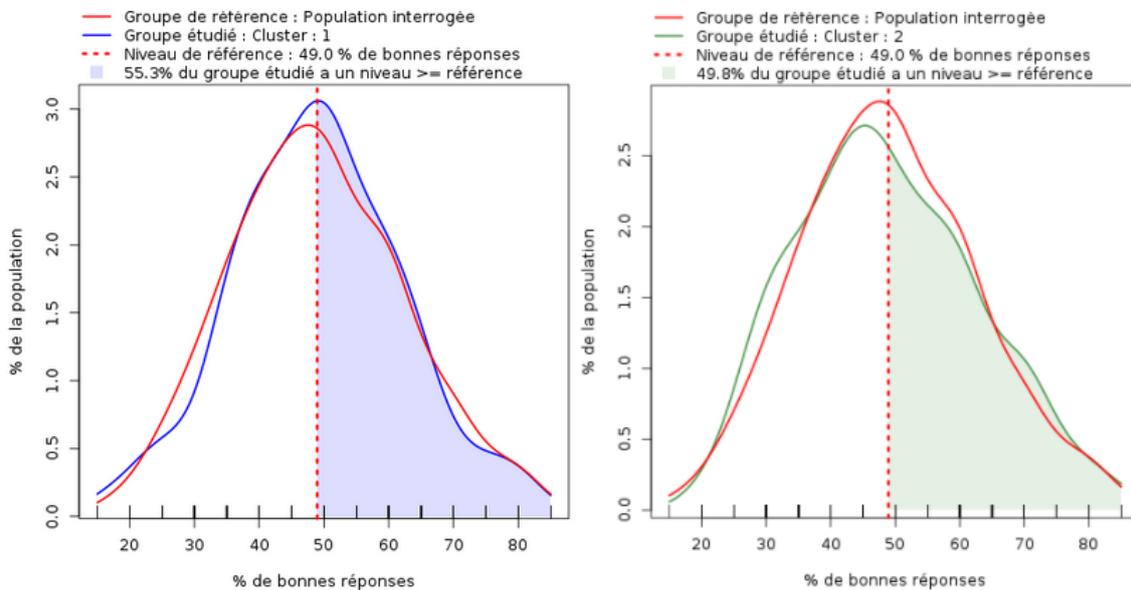
We can evaluate the results of “clustering1” in the reading comprehension quiz by adding this group to the study groups:



Then we click on the icon  from the “Level Map” section and we get the following result:



We find that there is no significant difference in level between the 2 clusters. If we look at things in a little more detail (by clicking on the blue and then green square), we can see that the distribution of levels in these clusters is very similar to that of the population surveyed:



Elimination of irrelevant criteria

Standardised value

The criteria can have orders of magnitude that are very heterogeneous according to their nature. For example, we can see that the “tclass” criterion varies between 0 and 5, while the “gender” criterion varies between 0 and 1. So we have a factor of 5 between these criteria.

Under these conditions, a difference in value of one criterion does not at all have the same importance according to the field in which this criterion varies. For example, a difference of 1 in “tclass” is much less important than a difference of 1 in “gender”. Indeed, when we speak about “tclass”, “a difference of 1” implies a “difference of 1 out of 5 possible levels”.

It is therefore more relevant to speak of a *relative* difference of 1/5 or 20% than an *absolute* difference of 1. By reasoning this way, two individuals having a difference of 1 in the “gender” criterion are 100% different.

Absolute values turn to relative values through a normalisation procedure which consists of the following calculation:

standardised value = (initial value - minimum value) / (maximum value - minimum value)

This calculation results in a criterion varying over the range [0,1] (0 = 0%, 1 = 100% of the initial value of the criterion) irrespective of the initial field of variation of this criterion.

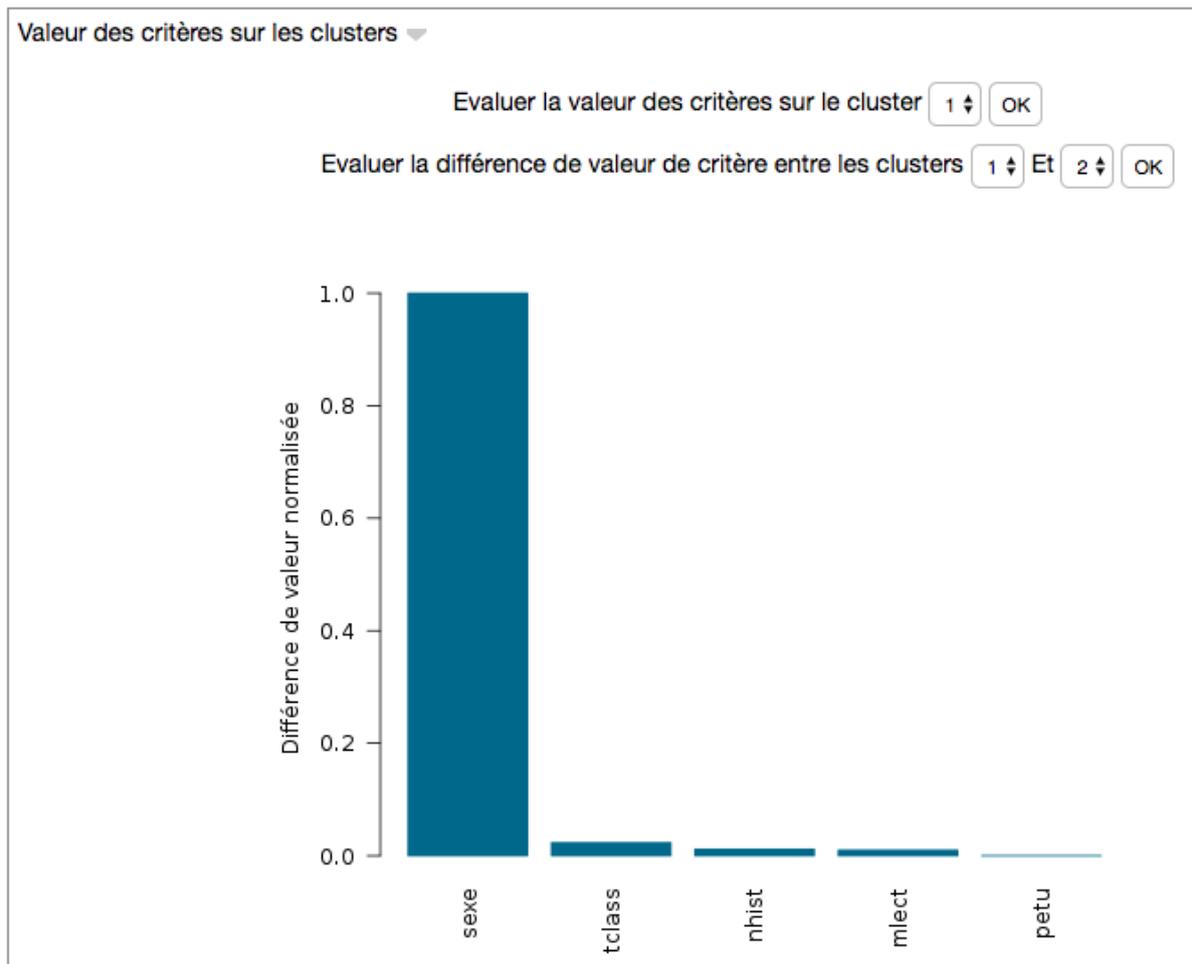
For example, if the criterion “year of birth” is considered, varying on the range [1950, 2000], the following standardised values are obtained on some examples of initial values:

| Initial value | Standardised value |
|---------------|--------------------|
| 1950 | 0 |
| 1952 | 0,04 |
| 1965 | 0,3 |
| 1974 | 0,48 |
| 1988 | 0,76 |
| 2000 | 1 |

The values obtained are called “standardised values”.

At this stage, our experiment has failed as the clusters obtained tell us nothing about the level of reading of the students.

Let's analyse what differentiates clusters 1 and 2 the most. For this, we go back to the "PCA" tab, "Value of criteria in clusters" section, and we ask to evaluate the difference in the value of criteria between clusters 1 and 2:



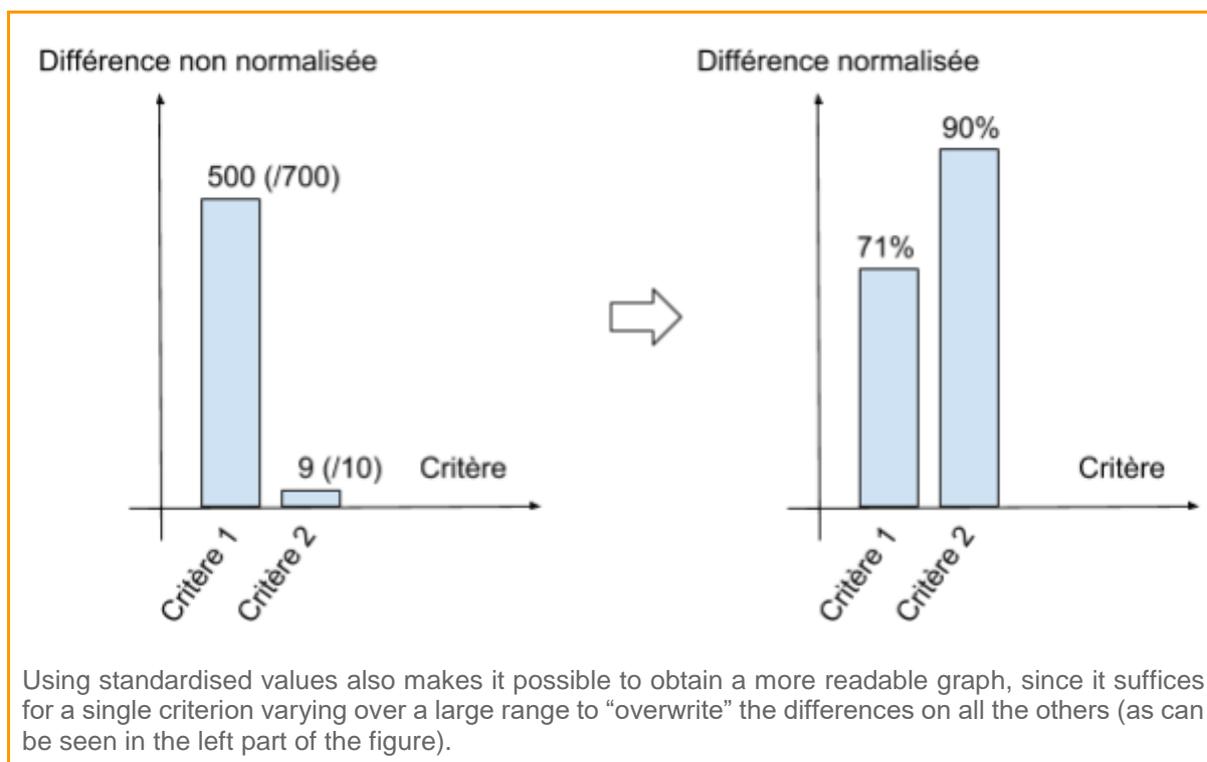
Comparing clusters procedure/ Difference from standardised value

To compare the value of clusters within criteria, the tool proceeds as follows:

1. It calculates the average value of each cluster for each criterion;
2. it standardises each value to bring it back to within the range [0,1] (see box "standardised value" for more details);
3. it calculates the difference (in absolute value) of these values.

By proceeding in this manner, the difference in value between two clusters for a criterion will necessarily be between 0 (difference of 0%) and 1 (difference of 100%), for whatever the nature of the criterion.

This makes it possible to give the same importance to each criterion, which makes it possible to better identify the criteria that vary the most (relative to their order of magnitude). For example, in the following figure, the use of standardised values makes it possible to see that criterion 2 (and not criterion 1) varies the most within clusters.



We find that the clusters are totally different on the “gender” criterion (the value 1 corresponds to a difference of “100%”): one cluster contains only girls, and the other only boys.

We also found that both clusters had equivalent reading levels, so we can deduce that the “gender” criterion has no impact on the reading level.

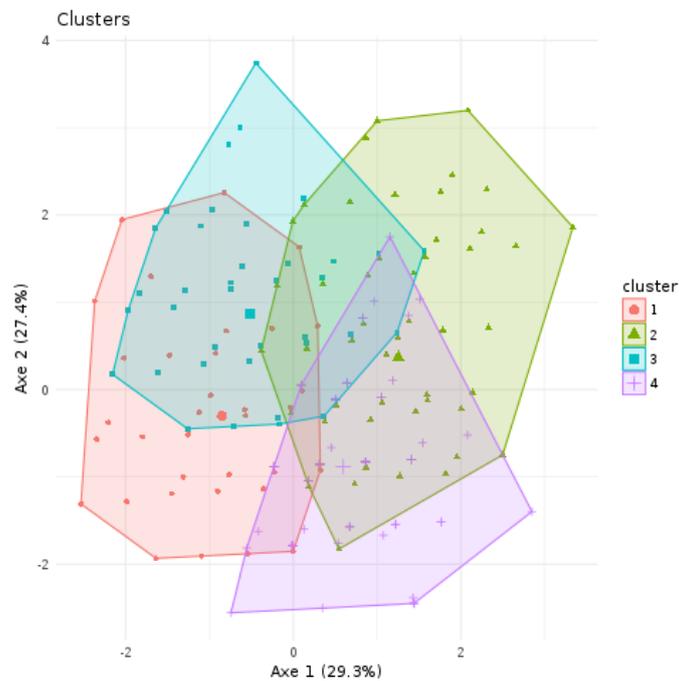
Second cluster analysis

As the “gender” criterion has no impact on the level of reading, we can exclude it from our study to try to identify the criteria that have an impact on the level among the remaining ones.

This exclusion is done by deselecting the “gender” criterion from the list of criteria in the configuration of the analysis:

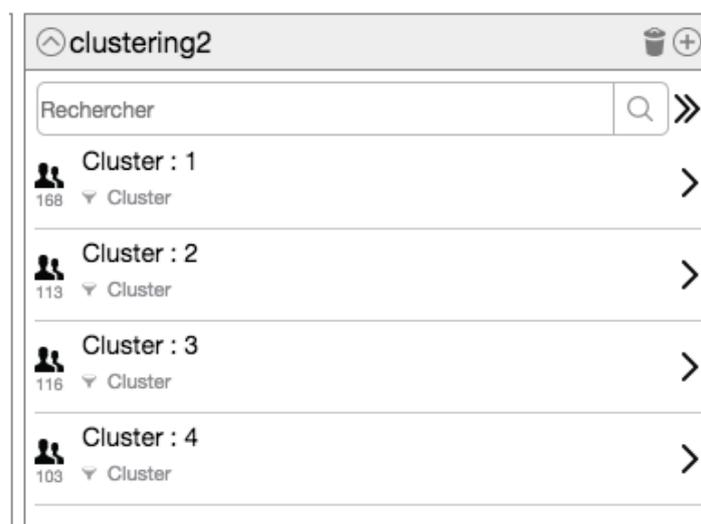
| | |
|--------|-------------------------------------|
| nhist | <input checked="" type="checkbox"/> |
| sexe | <input type="checkbox"/> |
| tclass | <input checked="" type="checkbox"/> |

We recalculate the clusters by fixing the number of clusters to 4 to separate the subgroups while keeping subgroups of reasonable size (a high number of clusters can lead to subgroups containing very few individuals).

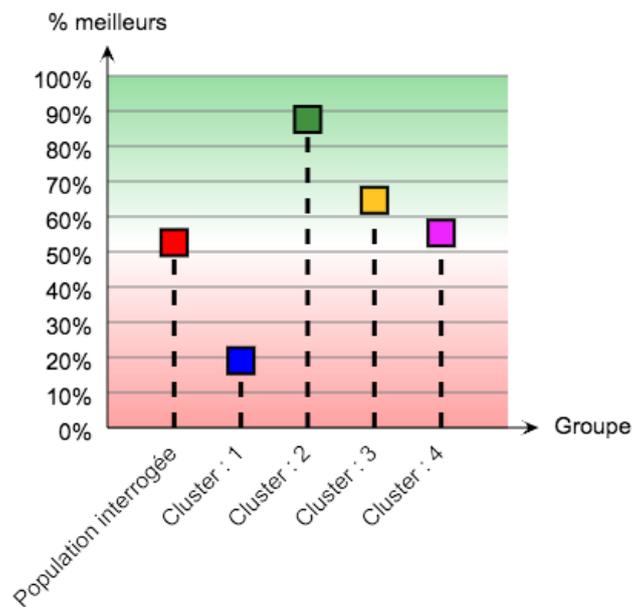


We add this new clustering to the study under the name “clustering2”.

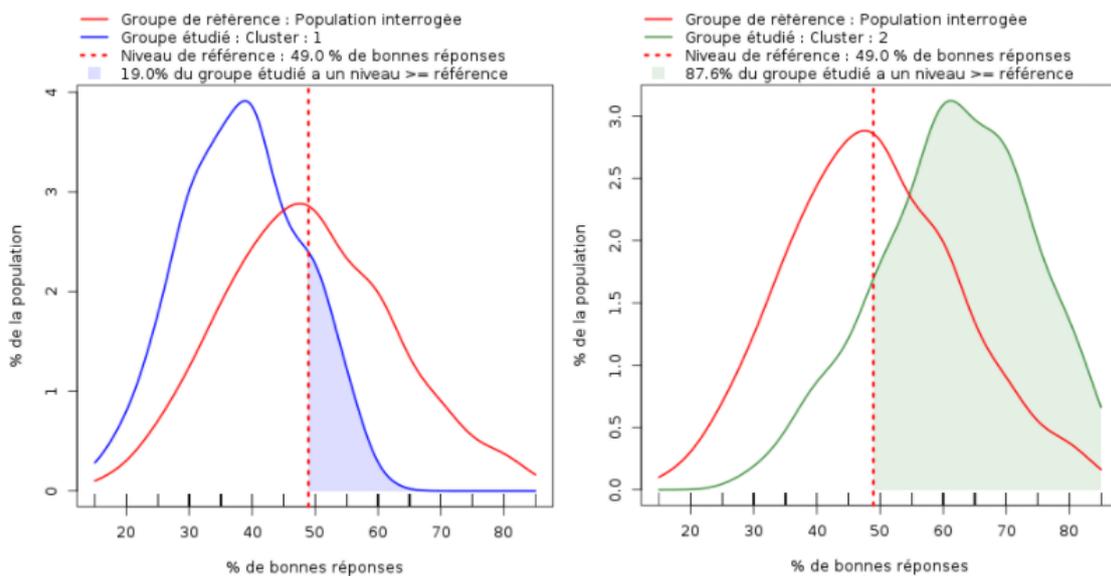
If we look at the composition of this clustering, we find that the subgroups are of homogeneous size and that they contain at least 100 people.



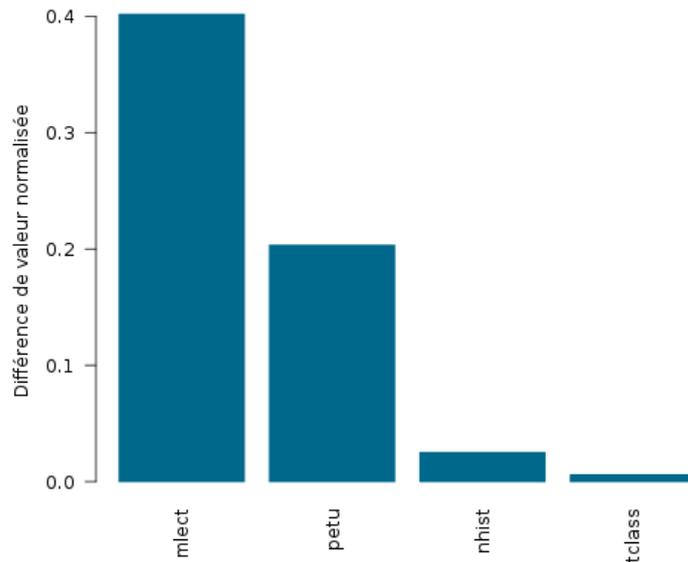
If we ask for the results of this clustering on the comprehension quiz, we find a significant difference in level between clusters 1 and 2.



This time, the distribution of levels between the two clusters is clearly different:



We also observe that what most differentiates these two clusters is the criterion “petu”, then “mlect”. The other criteria are almost identical in these clusters, their difference in value being less than 5%.

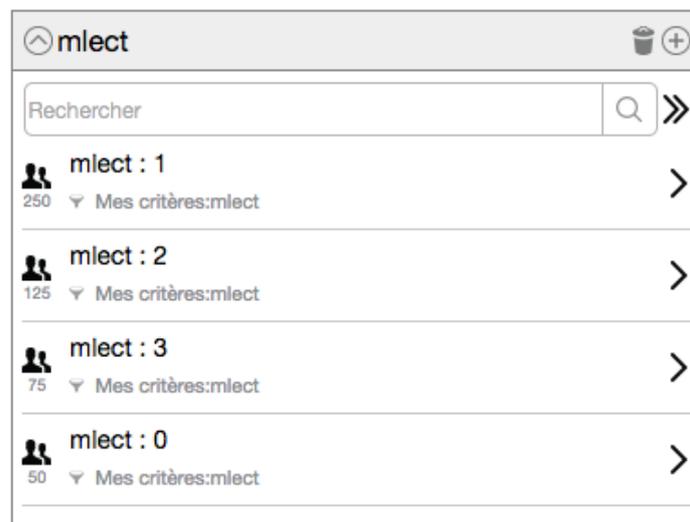


We can deduce that the criteria “mlect” and “petu” have an impact on the level, but we can not say more. In particular, it is not because the clusters differ most on the “mlect” criterion that this criterion has the greatest impact on the level.

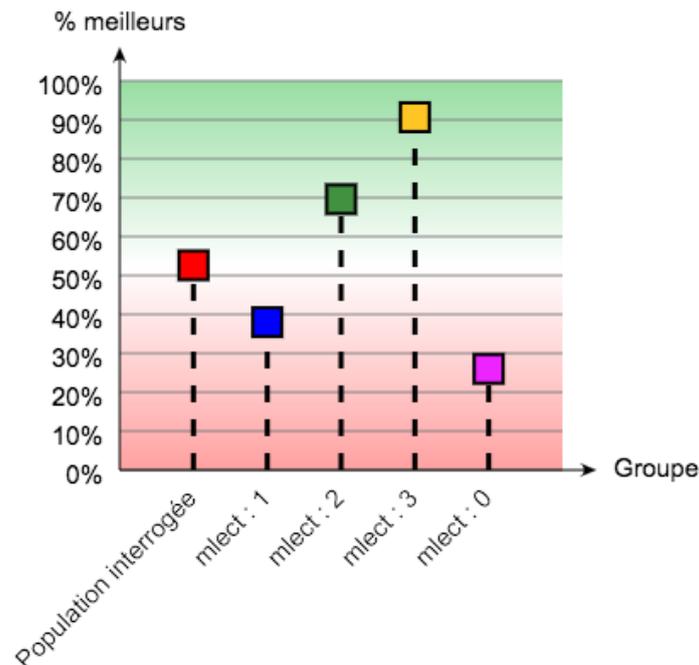
We can, however, get an idea of the progress of the level according to the criteria by using the division function of the “Experiments” section.

For example, to observe the impact of “mlect” on the reading level, we can create a “mlect” division based on the “mlect” criterion (see the manual of the statistical analysis tool for the details of operations).

The “mlect” division contains as many subgroups as reading methods. And as before, all the members of these subgroups constitute the surveyed population.



If we evaluate the results of the “mlect” division in the reference quiz, we obtain the following levels:



We clearly see that the level moves in the same direction as mlect, which makes it possible to identify a trend but not to quantitatively explain (let alone predict) results. This is the purpose of the predictive module that we will use in the next section.

Quantitative analysis

The PCA section makes it possible to study a population from the point of view of the criteria independently of its results in quizzes.

As we have seen above, it is possible to attempt to establish links between criteria and results using clusters. However, this procedure requires much back and forth between the “Experiment” and “PCA” tabs, and only allows for qualitative relationships to be established.

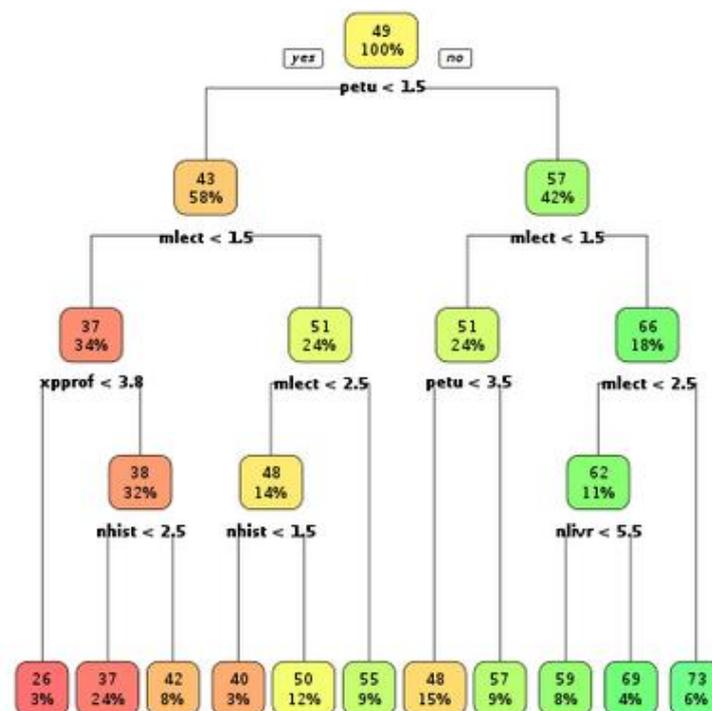
The “Experiment” tab contains a module whose purpose is to establish quantitative relationships between criteria and level. This is the “Predictive Module” section. This module is qualified as predictive because it allows not only to explain the level according to the criteria, but also to predict the level of a student from the known values on the various criteria of the study.

Decision tree

When we press , the module begins by separating the surveyed population into two subpopulations: a learning population (80% of the population surveyed) and a test population (20%).

The learning population is used by the module to build a decision tree. This tree summarizes the observed distributions of levels for the learning population according to thresholds in the criteria. The figure below shows the tree obtained from the data of our study:

Arbre de décision



On this tree, the nodes contain in the upper part an average level (% of correct answers) and the percentage of the population having this level. The nodes are coloured using a colour gradient from green (level: 100% correct answers) to red (0%).

The root of the tree corresponds to the average level of the learning population: 49% correct answers.

We have a first choice according to the value of the “petu” criterion. If its value is less than 1.5 (corresponding to the “No degree/GCSE and “BTEC/ GNVQ” levels of study), the observed average level is about 43% correct answers (58% of the learning population).

If the value is greater than 1.5 (which corresponds to a level of education “ ‘A’ levels or higher”), the average level observed is about 57% correct answers (42% of the learning population). Then, other choices allow to refine the value of the level. It should be noted that the same criterion can be the subject of several successive choices.

For example, if we go to the right to choice “mlect <1.5”, this means that “mlect <1.5” is false, ie mlect >= 1.5. This corresponds to the values 2 (syllabic dominates) and 3 (syllabic). Then there is another choice, “mlect <2.5” to distinguish these two cases. If we go to the left, it means that “mlect <2.5” is true, so necessarily mlect = 2 = “syllabic dominates”. Otherwise, mlect = 3 = “syllabic”.

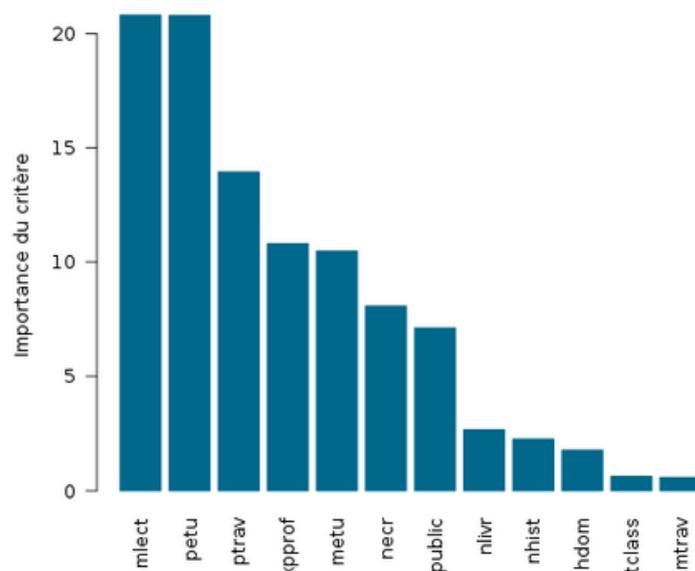
We notice that the first two selection criteria are “petu” and “ mlect”, which is consistent with our previous conclusions . However, it is not because “petu” is located at the root of the tree that it is the criterion having the most impact on the level.

Indeed, the importance of a criterion is determined by its contribution to the construction of each leaf of the tree (nodes at the bottom). For example, the criterion “mlect” occurs twice in the construction of the lower right leaf (73/6%) and “petu” just once.

Assessing the importance of the criteria that can be complex and often counterintuitive, the module has a “relative importance of criteria” section. There is talk of relative importance because these results are associated with *this* decision tree.

The software provides a graph showing the relative importance of the criteria. We are able to notice that the relative importance of the criteria “petu” and “MECT” is the same, which was difficult to estimate by observing the tree.

Importance relative des critères



We can therefore deduce that the reading level of students is mainly impacted, and in equal proportions by the level of parental education and reading method.

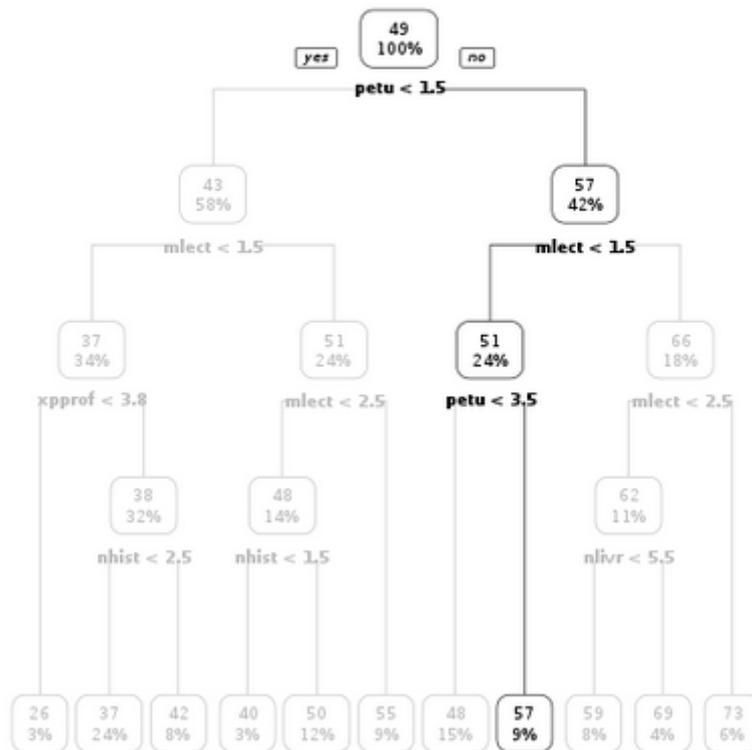
Prediction

Finally, we can predict the level of a student from the knowledge we have on the student's criteria values.

For this, we enter the name of a student in the "prediction of a person's level" section and click on "Predict".

Prédiction du niveau d'une personne

Niveau prédit : 56.6216
 Niveau réel : 45
 Erreur de prédiction : 11.6216



The module then navigates through the decision tree based on the values of criteria known on the student selected, until a leaf is attained (the missing values are completed using known values on the rest of the population). The level of the leaf corresponds to the prediction of the level for this student.

The module predicts a level of 56.6% correct answers whereas the actual level is 45%, an absolute error of about 11.6%. It's remarkably accurate when you consider the limited information used to make this prediction: only the values of the criteria "petu" (used twice) and "mlect".

However, we must remain cautious because on one hand, the results can vary greatly from one individual to another, and secondly, take care not to use in individuals belonging to the learning group to assess the quality of predictions.

Indeed, the decision tree was constructed to give the best possible results on the learning group. Under certain circumstances (too small learning group or not representative of the total population, for example), the tree can "stick" too much to the learning data. Under these conditions, we will achieve great results on the learning population but poor results on completely new individuals.

This is why it is usual to test the quality of predictions on individuals that the module has never seen during the learning phase. This is the objective of the test group.

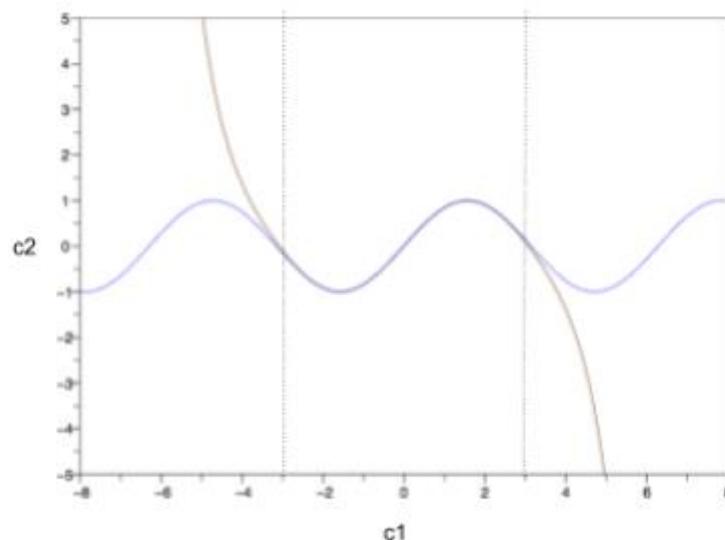
For the "Méthode de lecture" experiment, you can predict the level of a student by doing a name search (type "Test" in the search bar and choose a student).

Prédiction du niveau d'une personne

| | |
|------------------|---------|
| Test | Predict |
| Test Student1015 | |
| Test Student1016 | |
| Test Student1017 | |

Test group

If we consider a c2 criterion changing periodically according to a c1 criterion within a population (as shown by the blue curve in the figure below) and that the learning group contains only c1 examples belonging to the range [-3,3].



At the end of the learning process, the system could indeed lead to a model illustrated by the black curve. For c_1 belonging to $[-3,3]$, the curve picked up by the system and the actual curve are almost confused. Thus, for the learning population, the prediction error will be close to 0.

But it is clear that for individuals with a c_1 value outside $[-3,3]$, the prediction will become worse as c_1 increases (in absolute value).

This can be detected using individuals not belonging to the learning group.

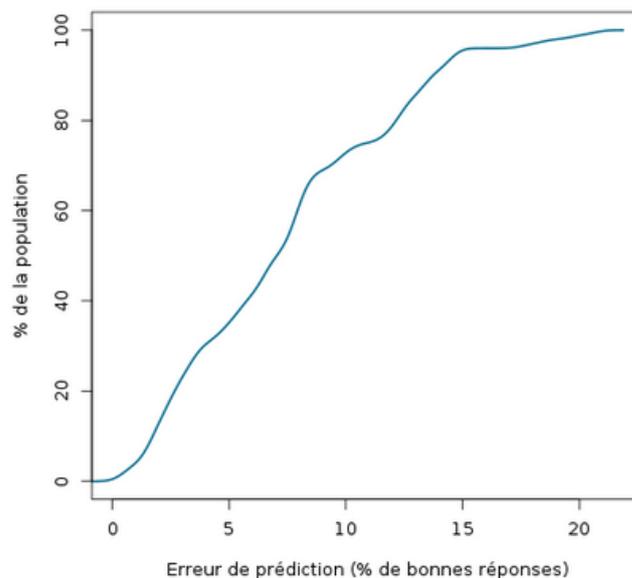
This group is called “**test group**”. It is only in this test group that prediction error will be relevant because the quality of the predictions is evaluated on individuals that the system has never seen during its learning phase. So, we minimize the risk that the system has learned “by heart” answers for these individuals. It checks that the system has properly generalised the concept linking the criteria (a correct generalisation here would be a sinus for example).

In practice, we use 80% of the initial population to constitute the learning group and 20% for the test group.

In our example, we constituted the learning group in the range $[-3,3]$ to highlight the problem of over-learning. But in practice, it is necessary that the learning group and the test group are both representative samples of the initial population. It should therefore pick individuals in the range capturing the whole c_1 variability (the range $[-8, 8]$ here). The important thing is that people in these two groups are different.

The “prediction error” shows the percentage of the test group that is most certain to have a prediction error.

Erreur de prédiction



Évalué sur le groupe de test (100 individus)

We can for example observe that the maximum prediction error is about 22% in the test group, or indeed an error over 12% was obtained for 80% of the test group.