Projet Cogmont

Azat Aleksanyan, Lucas Chabeau, Matthieu François, Etienne Hamard

March 13, 2019

Contents

Introduction

Le programme de notre 1ère année de Master prévoit un projet tutoré faisant appel à nos compétences acquises en statistique et en sciences des données. C'est dans ce contexte que nous avons travaillé en collaboration avec Marie-Caroline Crozet et Adeline Leclercq-Samson sur un sujet mêlant sciences cognitives, statistique et fouille de données.

Depuis 4 ans, une equipe de chercheurs de l'institut des sciences cognitives (ISC) de Lyon ont testé les capacités en mathématiques des élèves d'une école maternelle située dans une zone REP+ (Réseau d'éducation prioritaire) de l'agglomération Lyonnaise. Une partie de ces élèves ont suivi une méthode d'éducation conventionnelle et les autres ont suivi des enseignements suivant la méthode Montessori. Notre objectif est de

voir s'il y a ou non une différence entre le niveau en mathématiques des élèves ayant suivi les enseignements conventionnels et ceux ayant suivi les enseignements Montessori.

Nous avons reçu les résultats des tests sous forme de tableur que nous avons triés pour ne garder que les résultats des élèves de moyenne section. Nous avons ensuite néttoyé nos données et commencé l'étude sur cette population.

1. Contexte

Ce projet nous a été proposé par l'Institut des sciences cognitives - Marc Jeannerod spécialisé en neurosciences. L'UMR 5304 créée en 2007 est l'un des deux laboratoires de l'Institut des Sciences Cognitives – Marc Jeannerod. L'UMR 5304 est un laboratoire interdisciplinaire qui intègre l'expertise de chercheur des Sciences de la Vie (psychologie cognitive, neurosciences) et de médecine (pédo-psychiatrie, neuro-pédiatrie) avec celle de chercheur des Sciences Humaines et Sociales (linguistique computationelle et théorique et philosophie) pour étudier la nature et la spécificité de l'esprit humain.

1.1 Les besoins d'un changement éducatif

Le député et mathématicien Cédric Villani a publié un rapport pour renforcer l'apprentissage des mathématiques à l'école. Les élèves français ont aujourd'hui un niveau insuiétant dans cette discipline. Pourtant, jusqu'en 1985, l'enseignement des mathématiques en France était reconnu comme l'un des meilleurs. L'étude internationale "Trends in International Mathematics and Science Study" (TIMSS) 2015 qui mesure les performances en mathématiques et en sciences des élèves en fin de CM1 classe la France dernière des pays de l'Union européenne. Elle obtient même un score en dessous de la moyenne internationnale. Pour mettre un terme à cette tendance inquiétante de la dégradation du niveau des élèves français en mathématiques, le gouvernement est maintenant ouvert à de nouvelles pédagogie d'enseignement des mathématiques.

1.2 Les pedagogies étudiées

Pédagogie Montessori : La pédagogie Montessori est une méthode d'éducation créée en 1907 par Maria Montessori.
La pédagogie se base sur quelques principes :
-**La liberté** : les enfants sont libres de choisir l'activité qu'ils souhaitent faire parmi celles qui leur sont proposées.
-**L'autodiscipline** : les enfants sont invités à repérer eux-même leurs erreurs.
-**L'action en périphérique** : les professeurs vont préférer agir sur l'environnement de l'enfant plutôt que directement sur lui. Ils vont chercher à inciter l'enfant à faire une activité plutôt que de directement lui demander de faire cette activité.
-**Le respect du rythme de chacun** : Le rythme de l'enfant est respecté tant qu'il est concentré.
-**L'apprentissage par l'expérience** : Favoriser la pratique pour s'approprier un concept.
-**L'activité individuelle** : La plupart des activités se font en individuel.
-**L'éducation, une aide à la vie** : L'éducation est faite pour préparer l'enfant à une vie dans une société harmonieuse basée sur le respect de l'autre.

Pédagogie traditionnelle : La pédagogie traditionnelle (celle que nous connaissons aujord'hui dans nos écoles publiques) est celle du modèle transmissif. Selon le triangle pédagogique de Jean Houssaye, cette pédagogie privilégie la relation entre l'enseignant et le savoir. Autrement dit, l'enseignant expose un savoir sous forme de cours magistral, généralement suivi d'exercices ou/et de leçons à apprendre. L'élève doit intégrer et appliquer le savoir exposé par l'enseignant.

1.3 Présentation des données de départ

Notre jeu de données est composé de trois fichiers Excel (.xlsx), avec les résultats de chaque promotion au test cognitif mis en place par l'équipe de recherche l'institut des sciences cognitives. Un quatrième fichier du même acabit pour l'année 2018/2019 nous est parvenu au milieu de l'étude.

MathsJetons_2015-2016.xlsx : Pour l'année 2015/2016.

MathsJetons_2015-2016.xlsx : Pour l'année 2016/2017.

MathsJetons_2016-2017.xlsx : Pour l'année 2017/2018.

Jetons2019.xlsx : Pour l'année 2018/2019.

Chaque jeu de données représente les résultats questions par questions (en comptant les sous-questions) des élèves ainsi que leurs catégories pédagogiques et des informations telles que l'encadrant, le niveau scolaire, la langue natale, l'âge, le type de classe (mélangé entre plusieurs section ou pas), l'année de passage du test et leur école. Il y à 10 questions divisées en sous questions, ce qui fait un total de 34 réponses. Chaque question est indépendante et pour répondre à la sous question suivante il faut une bonne réponse à la sous-question précédente, sauf pour la question 4, toutes ses sous-questions sont indépendantes. Une bonne réponse correspond à un 1 et une mauvaise réponse à un 0, sauf pour la réponse à la question 1 qui est la valeur de comptage maximale de l'enfant. Ici les élèves viennent tous d'une ecole située en REP+.

1.4 Nettoyage des données

Les données ayant déjà été travaillées lors d'un précédent stage, le travail de nettoyage nécessaire n'a pas été excessif. Il nous a fallu tout de même renommer certaines variables pour les rendre plus lisibles et cohérentes entre elles, supprimer certaines questions car elles n'avaient été posées qu'à certaines classes... Les questions étant posées de manière à ce qu'au sein d'une même tache, il faille réussir les questions dans l'ordre pour passer à celles plus dures nous avions beaucoup de valeurs manquantes dès qu'une question était un peu difficile. Nous avons donc décidé de changer ces valaurs manquantes et les considérer comme une question que l'élève n'aurait pas réussie (donc remplacer NA par 0). Et pour ne pas passer à coté de l'information : "n'a pu aller plus loin", nous avons créé deux jeux de données : un vectoriel (chque question devient le regroupement des résultats au sous questions. ex : Q2.1 = 1; Q2.2 = 0 devient Q2 = 10), un composé de scores correspondant à la somme des questions au sein d'une même tâche (pour le même exemple que le jeu de données "vectoriel" nous aurions Q2 = 1+0 = 1).

1.5 Recodage des variables

Afin de ne pas influencer notre jugement sur nos résultats, nous avons dans un premier temps décidé de rendre anonyme le pédagogie enseignée pour chaque classe. Chaque pédagogie fut donc renommée en "P1" et "P2". De ce fait nous n'avons pas pu privilégier une pédagogie plus qu'une autre subjectivement parlant. Aux 3 quarts du projet environ, nous avons reçu les données de nouveaux individus, une cinquantaine, et avons par la même occasion décidé d'enlever cet anonymat. Notre travail étant déjà réalisé, seul l'interprétation sur le jeu de données comportant les nouveaux individus en plus importe. Comme dit précédemment les questions sont divisées en sous questions, ces questions sont regroupables en groupe : un groupe que nous appellerons "variable au-dela", un groupe "variable outil" puis un groupe "variable objet". Chacun de ces groupes fait appelle à une tache pédagogique en particulier.

variable Au-delà : est calculée en faisant la somme du nombre de bonnes réponses aux questions 2.3, 3.2, 4.2a, 4.2b, 4.2c, 4.2d, 5.2, 6.2, 8.6, 8.7, 8.8 et 8.9. La première question concernant la capacité à compter loin est prise en compte, les individus sachant compter au-delà de 7 ont un point en plus.

variable Outil : est calculée en faisant la somme du nombre de bonnes réponses aux questions des taches 4, 5 et 6 soit les questions des taches sur le surcomptage, la création d'une collection et la comparaison de deux collections.

variable Objet : est calculée en faisant la somme du nombre de bonnes réponses aux questions des taches 1, 2, 3 et 8 soit les questions des taches sur la capacité à savoir compter, à dénombrer une collection et à constituer une collection d'objets. La question 1 est gérée par paliés. Nous avons séparé les individus en 5 groupes : ceux qui savent compter jusqu'à 3, puis de 4 à 7, de 8 à 10, de 11 à 16, enfin ceux qui savent compter au delà de 16. Respéctivement pour chaque catégorie nous leur avons donnée un score de 0, 0.3, 0.6, 0.9 et 1.2.

Jetons2019.xlsx : Pour l'année 2018/2019.

2. Méthodologie

Afin de répondre au mieux à notre problématique nous avons fait le choix d'utiliser plusieurs méthodes statistiques différentes pour analyser nos données. Pour cela nous avons dans un premier temps utilisé une méthode qui permet de résumer l'information globale du jeu de données : l'analyse factorielle, et la classification ascendante hiérarchique (pour faire des regroupement de variables). Puis dans un but prédictif nous avons utilisé la régression logistique et les arbres de régressions. Plusieurs tests ont été fait en parallèle, comme celui du chi2, de student..

2.1 Analyse factorielle

Les méthodes d'analyse factorielle que nous avons utilisé ici sont l'analyse des correspondances multiples (ACM) et l'analyse en composantes principales (ACP), qui sont des méthodes de synthétisation du nombre de dimensions pour les données qualitatives et quantitatives. Cela nous permet d'appréhender plus rapidement le jeu de données, et avoir une première idée de ce qui diffère les individus entre eux (ou ce qui les rapproche). L'ACM permet dans un nuage à N dimensions, en cherchant les plans orthogonaux qui maximisent la variance entre les individus, à résumer celles ci en 4 voire 5 dimensions. l'ACM a été réalisée sur les données vectorisées, celles ci ont été prises comme variables actives (celles qui définissent le placement des individus sur le graphe) et les variables portant sur la pédagogie, la question 1, et l'âge en illustratives (ajoutée après le placement des individus sur le graphe). Le principe était le même pour l'ACP qui a été faite ensuite.

2.2 Classification Ascendante Hiérachique

N'ayant aucune information au préalable sur le thème des questions, leur regroupement... etc Mais sachant que certaines questions faisaient appelle aux mêmes compétences. Nous avons utilisé une variante de la classification ascendante hiérarchique (CAH) afin de partitionner nos variables. Nous avons aussi utilisé la CAH classique qui consiste à regrouper les individus selon leur points communs cela en partant d'une inertie interclasse maximale, pour arriver à une inertie interclasse de 0. Nous avons utilisé la CAH classique afin de partitionner nos individus dans un but descriptif.

2.3 Tests statistiques

Un test statistique est une procédure de décision entre deux hypothèses. Il s'agit d'une démarche consistant à rejeter ou à ne pas rejeter une hypothèse statistique, appelée hypothèse nulle, en fonction d'un jeu de données. Pour le projet nous avons utilisé le test paramétrique. Un test paramétrique est un test pour lequel on fait une hypothèse paramétrique sur la distribution des données sous H0. Les hypothèses du test

concernent alors les paramètres de cette distribution. En fonction du type de données, nous avons utilisé le t-test de Student et le test de proportion de succès.Trois différents types de test statistique ont été effectué pour cette analyse. Premièrement un test du d'indépendance Chi2 qui permet de déterminer l'existence d'une relation de dépendance entre deux variables au sein d'un effectif. Si il a dépendance il ne peut en aucun cas indiquer le sens de cette relation. Nous l'avons utiliser pour déterminer si il y avait une relation entre chaque question. Deuxièmement un test de proportionnalité, qui permet de tester une différence de proportion entre deux effectifs. Nous avons effectuer ce test pour vérifier la proportion de bonne réponse chez les élèves de pédagogie 1 et pédagogie 2. Et finalement le test de student qui permet de vérifier si la moyenne de deux échantillons est significativement différente. Nous avons utilisé ce test pour vérifier la moyenne entre les deux pédagogie pour certains regroupements de notes.

2.4 Regression Logistique

Notre problématique étant de voir s'il existe un lien entre la façon d'enseigner et les réponses au test, nous avons voulu essayer de prédire la méthode d'enseignement à l'aide des réponses des élèves avec la régression logistique. Cette méthode permet de modéliser une classification, à l'aide notamment de l'odds ratio. Cela revient à calculer la probabilité : $P(1|X) = \frac{e^{b_0+b_1x_1+\ldots+b_jx_j}}{1+e^{b_0+b_1x_1+\ldots+b_jx_j}}$. avec $b_0 = ln\frac{p(1)}{p(0)} + a_0$ et $b_j = a_j$.

2.5 Arbre de regression

L'arbre de régression est une technique d'apprentissage supervisé, qui permet en analysant un grand nombre de données, de prédire une variable à expliquer. Ils sont beaucoup utilisés dans le domaine du marketing, et plus récemment dans le domaine du machine learning (apprentissage automatique). Dans un premier temps il s'agit d'exprimer la variable à expliquer en fonction d'un maximum de variables explicatives, puis d'élaguer l'arbre afin de minimiser l'erreur, soit l'écart entre la valeur prédite et la valeur réelle. Cela revient donc à faire une régression logistique sur les données, puis d'appliquer l'algorithme de construction d'arbre à partir des résultats.

2.6 LME / LMM (to be checked)

La procédure des modèles mixtes linéaires développe le modèle linéaire général pour permettre aux données de présenter des variabilités en corrélation et des variabilités non constantes. Le modèle linéaire mixte offre donc une flexibilité pour modéliser non seulement les moyennes des données, mais également leurs variances et covariances.

Les modèles mixtes linéaires sont une extension des modèles linéaires simples permettant des effets fixes et aléatoires. Ils sont particulièrement utilisés lorsqu'il n'y a pas d'indépendance dans les données, telle qu'elle résulte d'une structure hiérarchique.

2.7 GLMM (to be checked)

Les GLMM (pour Generalized Linear Mixed Models) sont des modèles linéaires généralisés à effets mixtes. Ils sont employés pour analyser des données de comptages, des réponses binaires (*notre cas*) et lorsque les données ne sont pas indépendantes (ça c'est pour la partie mixte)! En général, un GLMM (Generalized Linear Mixed Model ou modèle linéaire généralisé mixtes) est un GLM avec une fonctionnalité supplémentaire qui lui permet de prendre en considération la non indépendance des données.

Un GLMM est dit "mixte", car il comporte au moins un effet dit "fixe" (la variable dont on souhaite évaluer l'effet, ici les *Pédagogie*, *Age* et *Année Scolaire*) et au moins un effet dit "aléatoire" (la variable de regroupement, ici *newClasse* ou *Group*). Les effets aléatoires ne sont pas évalués, ils servent seulement à indiquer au modèle que les données ne sont pas indépendantes pour une boite donnée. C'est ce qui permet à

la déviance résiduelle d'être bien estimée, et ainsi à l'erreur standard des paramètres de ne pas être biaisée, et aux final d'obtenir des résultats fiables.

3. Analyse exploratoire

3.1 Univariée | Bivariée

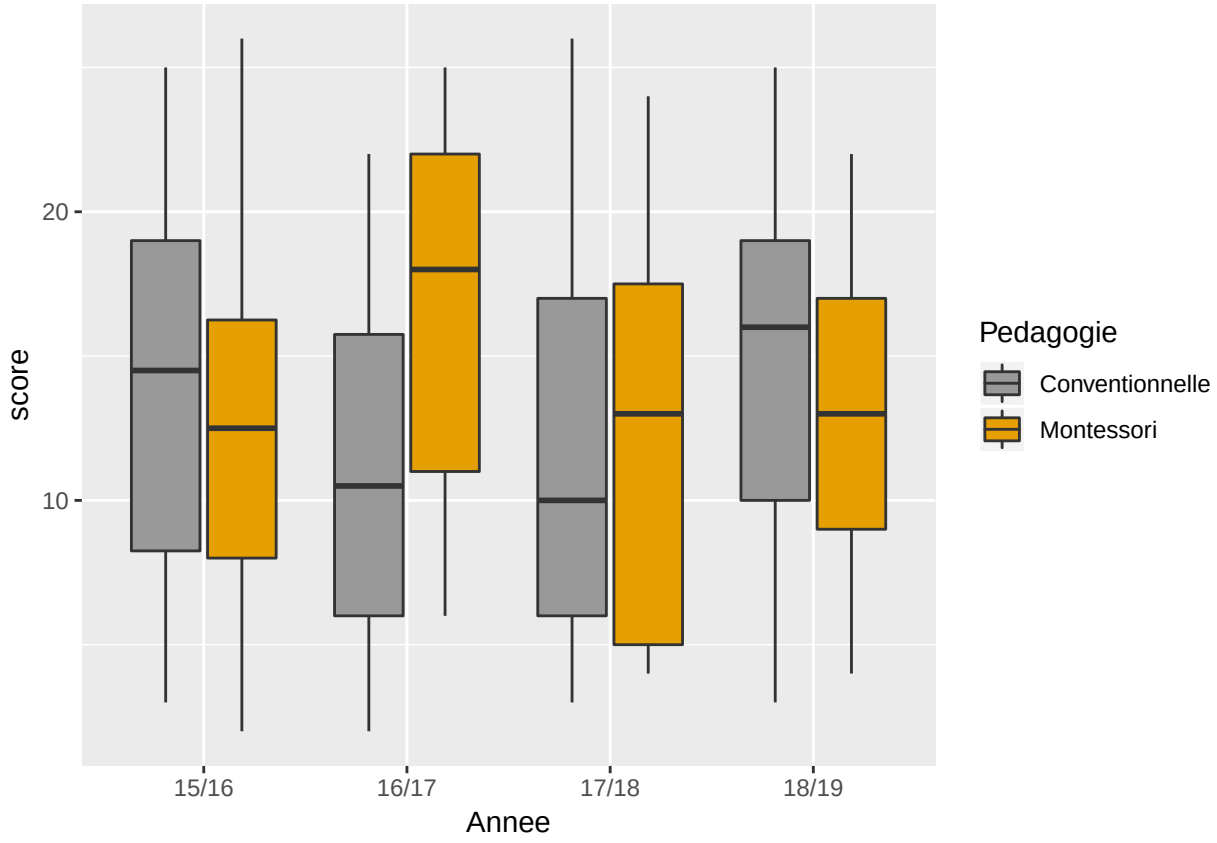


Figure 1: Score total par année par pédagogie

En ne regardant que certaines années on peut voir clairement une différence entre la pédagogie montessorienne et la pédagogie conventionnelle vis à vis du nombre de bonnes réponses. Toutefois lorsqu'on regarde la vue d'ensemble, on peut voir que celon les années, une fois la pédagogie montessorienne est supérieur à la conventionnelle, parfois c'est l'inverse. On peut donc s'attendre à ce qu'on ne puisse pas prédire quelle pédagogie permet d'obtenir un meilleur score global.

3.2 Multivariée

Afin de traiter l'information présente dans le jeu de données de la meilleure façon, nous avons procédé à 2 analyses multivariées : 1 sur le jeu de données qualitatif sous forme de vecteurs, et 1 sur le jeu de données quantitatif sous forme de scores.

3.2.1 Analyse des Correspondances Multiples

La réalisation d'une ACM comme première approche sur le jeu de données a permis de mieux comprendre ce qui différencie les individus dans notre jeu de données et à la fois d'avoir un premier résultat sur la différence entre les deux pédagogies selon cette méthode.

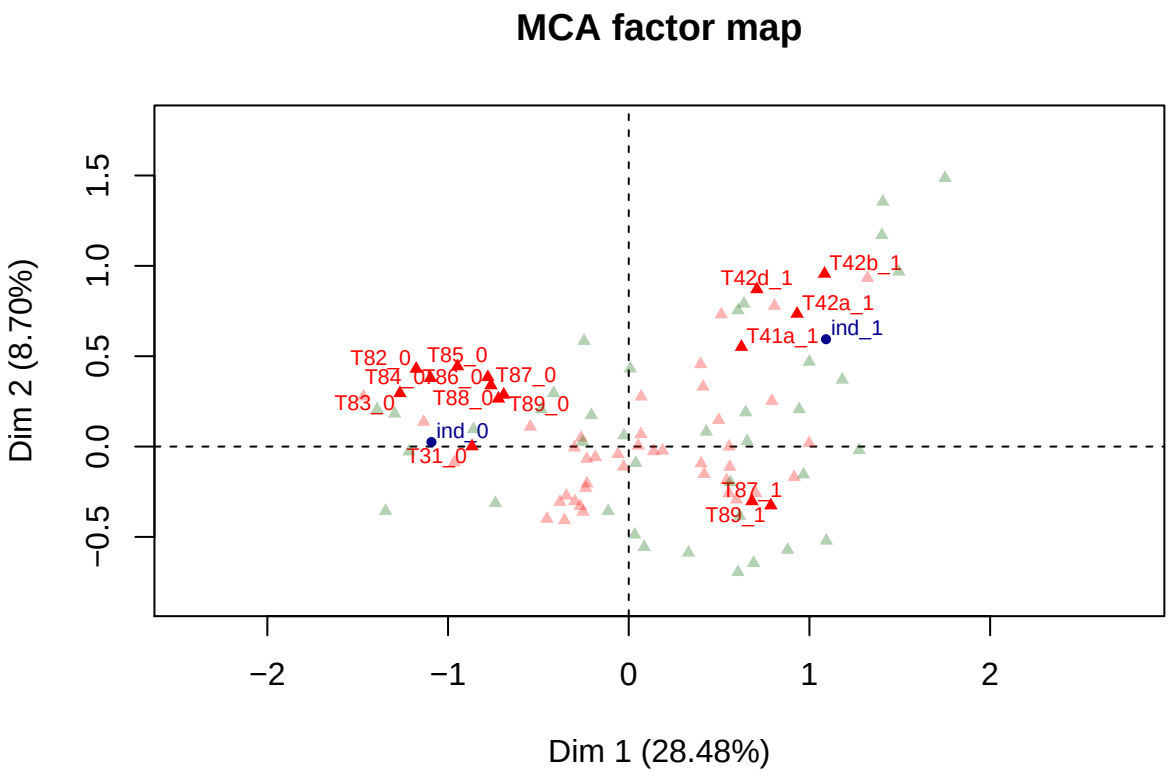


Figure 2: Graphe des modalités sur le plan principal

Le graphe précédent représente les 15 modalités qui contribuent le plus au placement des individus sur le plan principal. On peut donc voir que la dimension 1 oppose des modalités qui concernent la réussite à une question (avec un "_1" à la fin), à droite, à des modalités qui concernent l'échec d'une question (avec un "_0"), à gauche. Plus un élève réussira le questionnaire, plus il se trouvera à droite sur le graphe des individus. Cette interprétation est confirmée par l'ajout de deux individus fictifs : ind_1 et ind_0, qui comporte respectivement des succès à toutes les questions et des échecs à toutes les questions. L'individu ayant réussi en totalité le questionnaire se trouve à droite alors que l'individu ayant raté en totalité le questionnaire se trouve à gauche. De plus nous pouvons voir que les questions qui discriminent le plus la réussite ou non de l'examen sont les questions 4 et 8 (de part leur forte contribution). Toutefois cette première analyse n'aura pas permis de différencier les deux pédagogies, la variable projetée en supplémentaire sur le plan principal n'est pas significativement liée à celui ci.

Dans un second temps nous avons refait une ACM mais cette fois ci sur les données vectorielles. Cela afin de prendre en compte la succession de certaines questions qui se regroupent en "compétences".

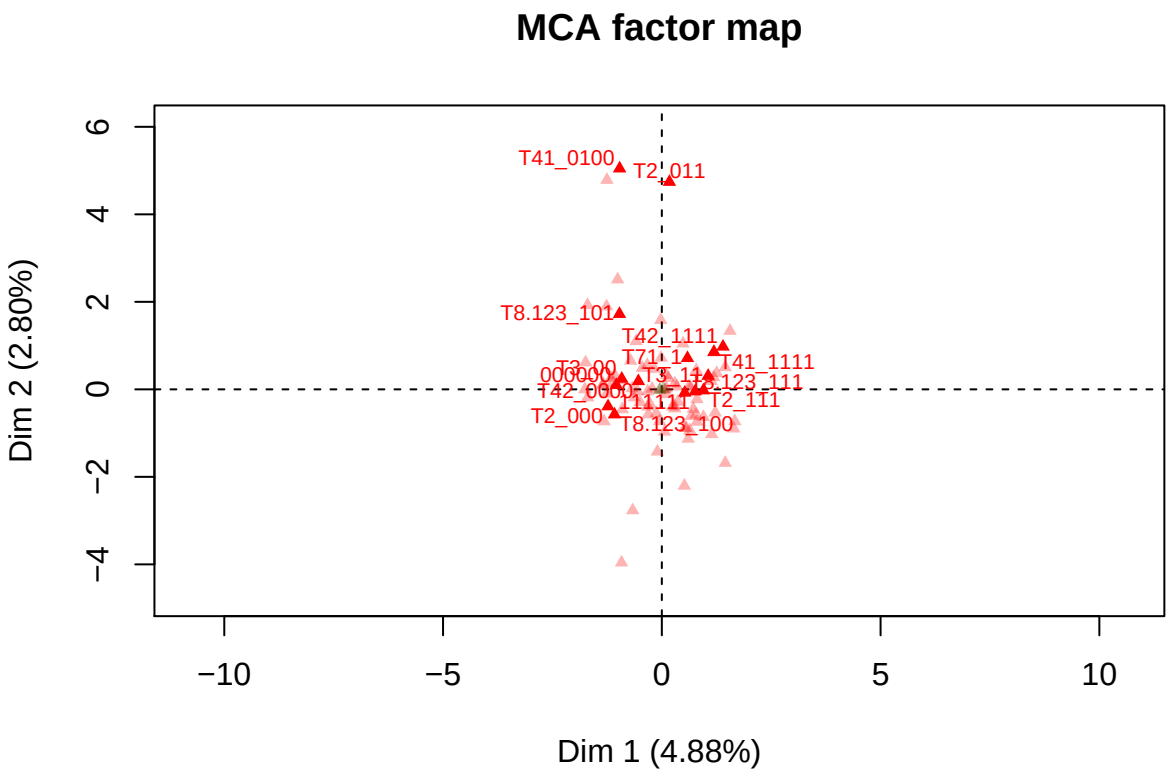


Figure 3: Graphe des modalités sur le plan principal

A nouveau la première dimension oppose les individus ayant réussit les totalités (ou la majorité pour certaines) des question à droite à ceux qui n'en ont réussi aucune à gauche. Nous ne pouvons pas non plus observer de différence significative entre les 2 pédagogies. On peut voir cette fois ci avec plus de précision les questions qui discriminent la réussite au questionnaire. Ce sont Les question 4, 3, 8 et 5.

Dans ces deux cas nous avons pu aussi observer un lien significatif entre les variables qualitatives, portant sur les réponses à la question 1 et l'age de l'élève, et le placement des individus sur la première dimension. En conséquent on peut dire qu'il existe un lien entre la réussite à l'examen et le fait qu'un enfant sache compter "loin" et dans une moindre mesure, qu'il soit âgé.

Enfin nous avons voulu voir si en classifiant les individus suite à l'ACM nous obtenions des groupes d'individus propre à une pédagogie ou non. Pour cela nous avons utilisé la classification ascendante hiérarchique (CAH).

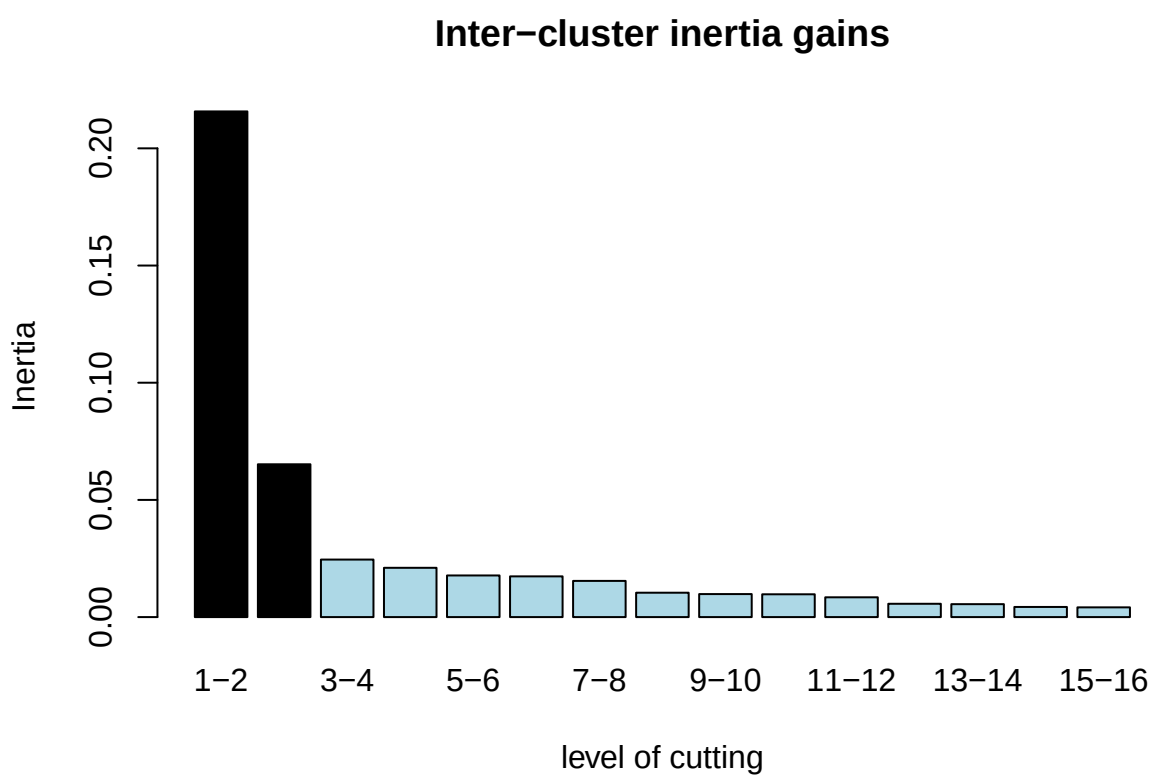


Figure 4: Diagramme des gains d'inertie

Nous pouvons observer un "saut" à la troisième classe, donc nous faisons le choix de retenir trois classes pour la CAH. Mais leur composition ne montre aucune sureprésentation d'une pédagogie plus que l'autre. Le test de chi2 entre la variable concernant la pédagogie et celle concernant la classe n'est pas significatif. Une fois de plus cela ne permet donc pas de montrer une liaison entre la pédagogie et ce qui discrimine nos classe. Au final nous obtenions une classe d'individus qui a une majorité d'échecs, une d'individus qui échouent sur les questions 4.2, et une qui d'individus qui réussissent globalement.

3.2.2 Analyse en Composantes Principales

La réalisation d'une ACP faisant suite à l'ACM a pour but d'étudier le jeu de données différemment. En effet nous avons étudié cette fois ci le jeu de données concernant les scores. Soit un jeu de données quantitatif. Afin de ne pas perdre d'informations nous avons dans un premier temps observé la matrice des corrélations entre les variables de notre jeu de données. Car plus les variables seront corrélées entre elles, plus l'ACP ne montrera que celles ci.

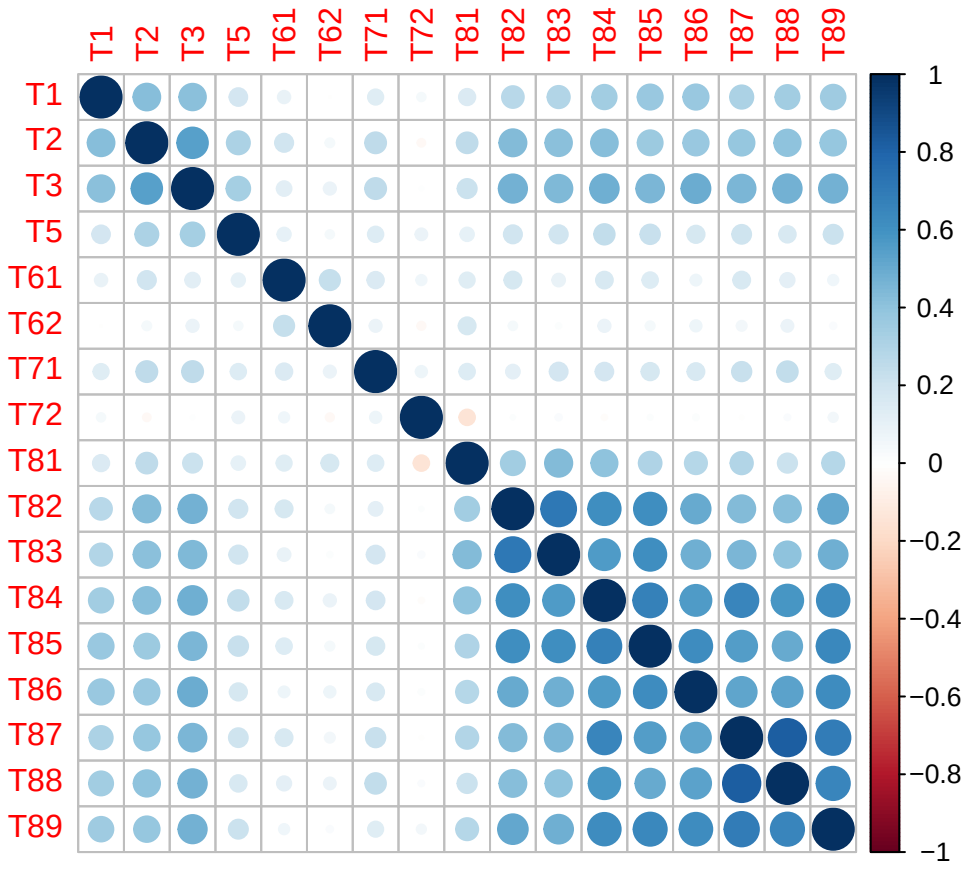


Figure 5: Matrice des corrélations

Nous avons donc fait le choix de regrouper les questions 8 en 2 groupes, aux vues des résultats. Un groupe comprenant les questions 8.1, 8.2 et 8.3, et un comportant les autres questions 8.

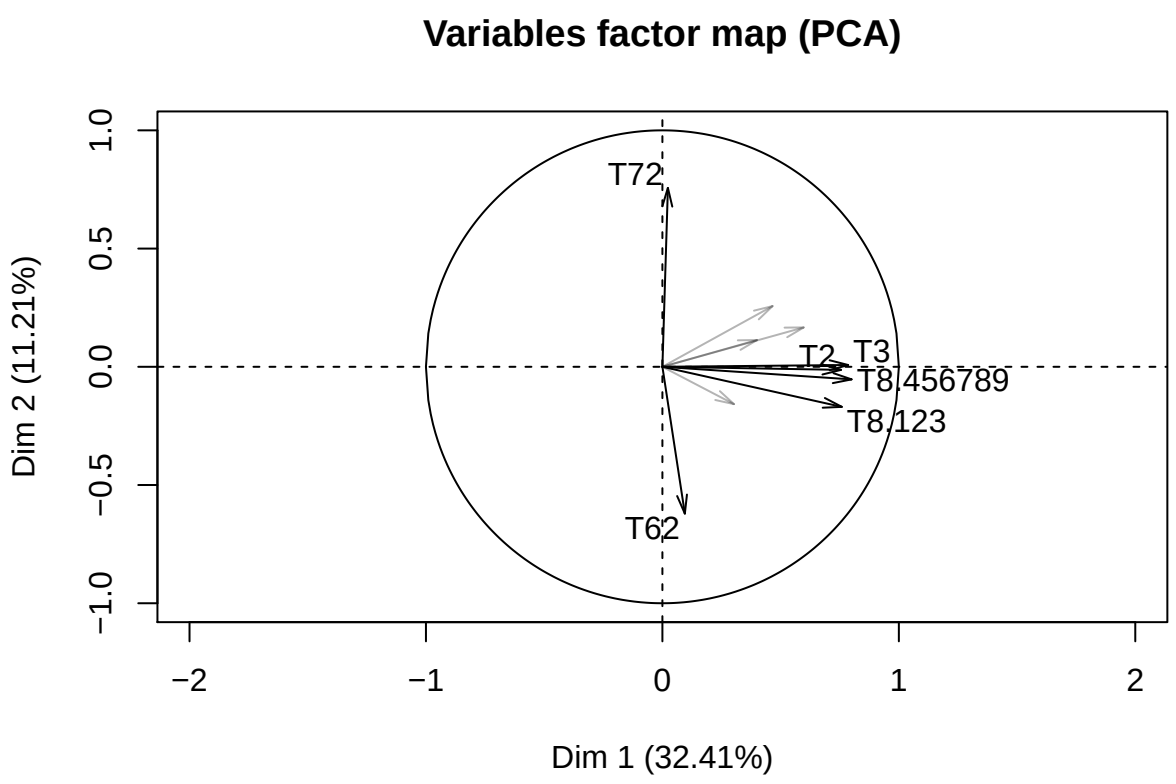


Figure 6: Cercle des corrélations du plan principal

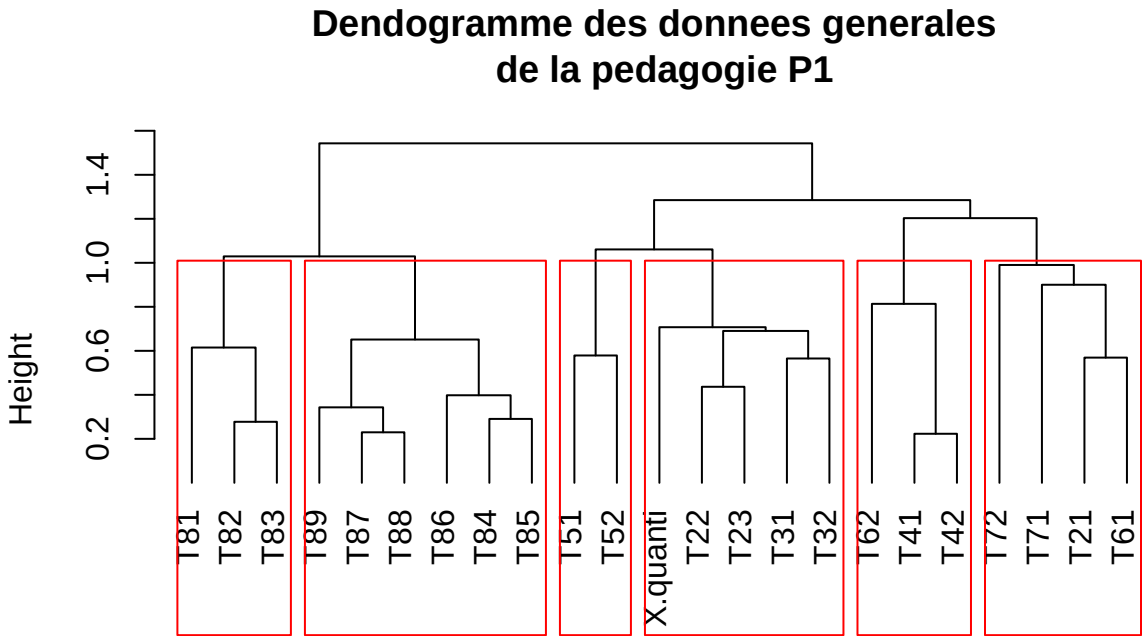
La réalisation de l'ACP nous permet donc de voir que les individus se diffèrent sur la première dimension selon les question 2,3, et 8. Alors que la dimension 2 les différents selon les question 7.2 et 6.2. A nouveau nous n'observons pas de lien significatif entre la pédagogie enseignée et le placement des individus sur les axes factoriels.

3.3 Classification Ascendante Hiérarchique des variables

N'ayant au début de notre analyse, aucune information sur le thème des questions, leur regroupement... etc Mais sachant que certaines questions faisaient appellent aux mêmes compétences. Nous avons utilisé une variante de la classification ascendante hiérarchique (CAH) afin de partitionner nos variables. La CAH est une méthode de classification qui permet de regrouper des individus sein d'une même classe et qu'ils soient le plus semblables possibles tandis que les classes soient elles ,le plus dissemblables possibles.

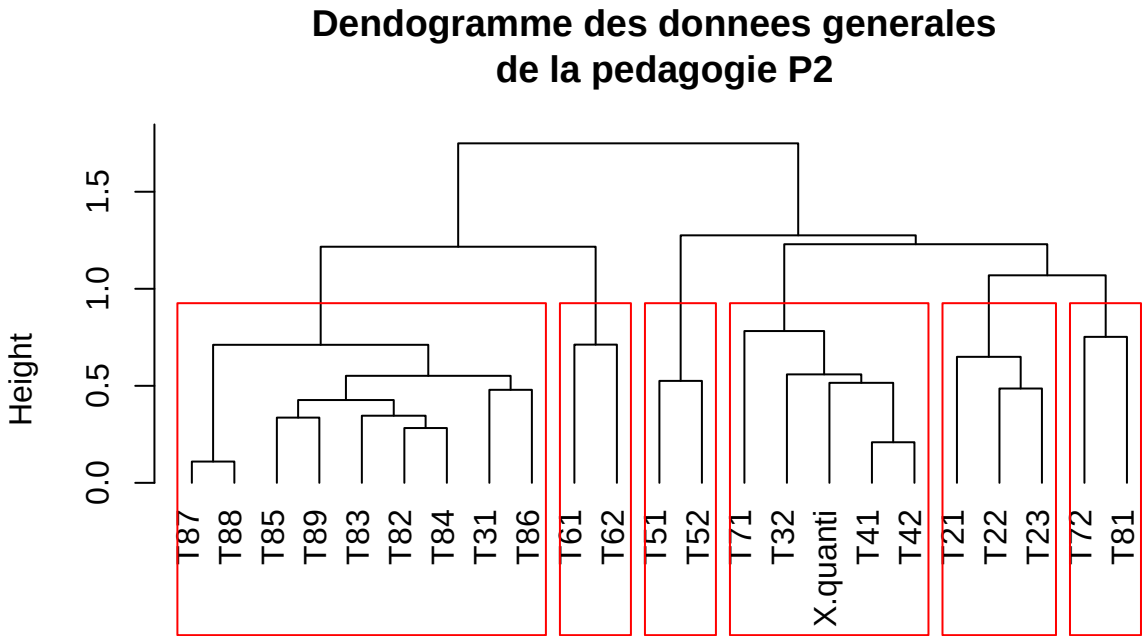
Nous allons appliquer cette méthode sur les jeux de données en isolant les pédagogies pour comparer les regroupements.

Procédons d'abord à l'isolation de la Pédagogie 1.



Ici X.quanti correspond a T1. Nous pouvons observer ici plusieurs regroupement redondant. Le regroupement des questions T81, T82 et T83 et celui des questions T84, T85, T86, T87, T88, T89. Les question T1, T2 et T3 sont aussi fortement attirées, on retrouve en parti la variable “Objet”.

Comparons maintenant avec les regroupements de la Pédagogie 2.



Nous observons en regroupement des questions T82, T83, T84, T85, T86, T87, T88, T89. La T81 étant séparée du reste. On voit aussi apparaitre deux couples de questions, T61 et T62 ainsi que T51 et T52. Ici nous ne voyons aucunes variable prédéfinie ressortir véritablement.

On retrouve plus de similarités entre les classification de la pédagogie 2 qu'entre celles de la Pédagogie 1. Et on remarque que les regroupement qui sont stables entre les changements de jeu de données sont principalement ceux liés aux sous questions de la T8. Pour résumer, les groupes qui ressortent sont pour la Pédagogie 1: (T81, T82 , T83), (T84, T85, T86, T87, T88, T89) et (T1,T2, T3) Et pour la Pédagogie 2: (T82, T83, T84, T85, T86, T87, T88, T89), (T61, T62) et (T51, T52).

Seul la variable 'Objet' est en parti retrouvée et seulement dans le cas de la Pédagogie 1.