
Evaluation d'un outil de simulation d'examen en ligne

Auteur : Nadin, Boris

Promoteur(s) : Haesbroeck, Gentiane

Faculté : Faculté des Sciences

Diplôme : Master en sciences mathématiques, à finalité didactique

Année académique : 2020-2021

URI/URL : <http://hdl.handle.net/2268.2/12045>

Avertissement à l'attention des usagers :

Tous les documents placés en accès ouvert sur le site le site MatheO sont protégés par le droit d'auteur. Conformément aux principes énoncés par la "Budapest Open Access Initiative"(BOAI, 2002), l'utilisateur du site peut lire, télécharger, copier, transmettre, imprimer, chercher ou faire un lien vers le texte intégral de ces documents, les disséquer pour les indexer, s'en servir de données pour un logiciel, ou s'en servir à toute autre fin légale (ou prévue par la réglementation relative au droit d'auteur). Toute utilisation du document à des fins commerciales est strictement interdite.

Par ailleurs, l'utilisateur s'engage à respecter les droits moraux de l'auteur, principalement le droit à l'intégrité de l'oeuvre et le droit de paternité et ce dans toute utilisation que l'utilisateur entreprend. Ainsi, à titre d'exemple, lorsqu'il reproduira un document par extrait ou dans son intégralité, l'utilisateur citera de manière complète les sources telles que mentionnées ci-dessus. Toute utilisation non explicitement autorisée ci-avant (telle que par exemple, la modification du document ou son résumé) nécessite l'autorisation préalable et expresse des auteurs ou de leurs ayants droit.



Université de Liège
Faculté des sciences
Département de Mathématique

Évaluation d'un outil de simulation d'examen en ligne

Boris Nadin
(Promotrice : Gentiane Haesbroeck)

Mémoire en vue de l'obtention du grade de Master en
sciences mathématiques, à finalité didactique

Année académique 2020–2021

Chapitre 1

Introduction

Dans le cadre d'un cours de physique générale donné à des étudiants de première année de bachelier en médecine, un outil en ligne, nommé « simulateur d'examen » a été proposé aux étudiants à partir de l'année académique 2015–2016, qui est l'année qui sera étudiée ici ¹.

Cet outil leur a permis, pour chacun des cinq chapitres étudiés dans le cadre du cours de physique, de se voir proposer une série de questions leur permettant de vérifier leurs connaissances sur la matière. Ces questions sont bien entendu classées par thème, mais aussi par difficulté : des membres de l'équipe pédagogique du cours de physique ont évalué la difficulté de chaque question sur une échelle de 1 à 4.

Pour chacun des cinq chapitres, trois niveaux de difficultés étaient proposés dans l'outil, dénommés *bronze*, *argent* et *or* par ordre croissant de difficulté. L'étudiant qui souhaitait utiliser l'outil choisissait un thème et un niveau de difficulté, et se voyait alors proposer une série de questions générées automatiquement hors d'un pool de questions sur le thème choisi, principalement issues d'examens des années précédentes, et dont la difficulté variait selon le niveau de difficulté choisi. Lorsqu'un étudiant réussissait au moins la moitié des questions d'un test ainsi généré, il obtenait la « médaille » correspondante (un badge virtuel attestant de la réussite d'un test pour le thème et le niveau de difficulté correspondants).

Le dispositif proposait par ailleurs un incitant : les étudiants qui obtenaient les quatre premières médailles d'or (autrement dit, ceux qui avaient réussi un test du plus haut niveau de difficulté pour chacun des quatre premiers thèmes vus dans l'année au moins une fois) avant une date limite fixée au début de l'année avaient le droit de présenter un examen blanc en présen-

1. La présentation qui suit du dispositif est également résumée par les auteurs du dispositif dans un acte de conférence (Marique et al., 2016).

tiel, dans les conditions de l'examen².

Pour l'année académique 2015–2016, qui est la seule année qui sera exploitée dans cette étude, deux ensembles de données ont été collectés. Les premières données, que l'on appellera *données d'entrée*, sont des informations sur le profil socioéconomique et le parcours scolaire des étudiants. Ces données ont été collectées par une enquête qui était proposée aux étudiants lors de leur première utilisation de l'outil en ligne. Les autres données, que l'on appellera *données de travail*, sont des informations sur l'utilisation de l'outil qui a été faite par chaque étudiant. On y trouve des informations sur les médailles obtenues ainsi que le nombre de tentatives effectuées par chaque étudiant, pour chaque thème et niveau de difficulté. Ces deux ensembles de données seront décrits en détail à l'aide de méthodes exploratoires dans les deux chapitres suivants. En plus de ces deux bases de données, la note de chaque étudiant à l'examen de janvier, qui sanctionne la réussite de ce cours, est fournie.

Comme mentionné dans le paragraphe précédent, il a été choisi de ne s'intéresser qu'à la seule année académique 2015–2016, alors que des données ont également été collectées pour les années suivantes. En effet, il n'est pas aisé de transposer l'étude faite pour chacune des années car les données d'entrée ont une forme différente d'année en année, due à l'ajout de questions dans le questionnaire d'entrée. De plus, à partir de l'année académique 2017–2018, le public est modifié, en raison de l'apparition de l'examen d'entrée en médecine. Il s'ensuit que la méthodologie utilisée dans cette étude devrait être adaptée afin de l'appliquer pour les autres années académiques. Notons également que les données de travail pour l'année 2015–2016 avaient l'avantage d'être immédiatement disponibles lors du début de cette étude, ce qui n'était pas le cas des autres années qui devaient encore être traitée par l'équipe en charge du simulateur.

1.1 Problématique

De nombreuses problématiques peuvent être posées pour évaluer l'outil proposé par l'équipe de P.-X. Marique et al. La problématique qui sera étudiée ici³ est celle qui est centrale pour déterminer l'utilité de l'outil : le fait d'avoir utilisé l'outil en ligne améliore-t-il significativement la note obtenue à l'examen de janvier ?

Cette question, qui peut paraître simple à première vue, nécessite en

2. Un autre test formatif en présentiel était proposé au milieu du quadrimestre à tous les étudiants.

3. D'autres idées de problématiques sont formulées dans les conclusions.

réalité de surmonter quelques difficultés liées à l'aspect rétrospectif de l'étude. En effet, et notamment pour des raisons éthiques, il n'est pas concevable de mettre en place un groupe test (n'ayant pas accès à l'outil) pour le comparer avec le groupe des étudiants y ayant eu accès. Les deux groupes sont donc constitués selon le souhait de l'étudiant de participer ou non, ce qui implique qu'ils ne sont a priori pas constitués aléatoirement. De plus, il n'est pas non plus possible de comparer les résultats des étudiants avec ceux des années précédentes, car malgré toute la bonne volonté donnée par l'enseignant, la méthode d'évaluation n'est pas strictement identique d'année en année.

Il est donc nécessaire de mettre en place des techniques permettant de quantifier le biais induit par le choix de l'étudiant d'utiliser ou non l'outil afin de déterminer s'il est légitime de comparer des étudiants ayant utilisé l'outil avec ceux qui ne l'ont pas utilisé.

1.2 Revue de la littérature

Ce sujet de recherche se place dans le contexte des *learning analytics*, qui consistent à collecter et traiter des données de manière à améliorer la qualité et la performance de pratiques d'enseignement. Des revues systématiques consacrées à ce sujet, telles que celles de Papamitsiou & Economides (2014), ou de Wong (2017) ont été consultées. Il en ressort que très peu d'études ont été effectuées sur ce sujet précis.

La plupart des études effectuées utilisent comme méthodes statistiques une classification (telle que déterminer tôt dans l'année si un étudiant est à risque ou non d'échouer), du clustering (déterminer des profils d'étudiants sur base d'une diversité de données liées à leur apprentissage), ou des régressions (afin de déterminer par exemple l'impact de certains facteurs sur la note).

Beaucoup d'études s'intéressent aux MOOC (cours ouverts en ligne), aux forums en ligne, aux méthodes employées par l'étudiant lors de son étude ou à la détermination de la réussite d'un étudiant en fonction de certains critères étudiés, mais un tel dispositif de questionnaire en ligne, à mettre en perspective avec des données d'entrée, ne semble pas revenir dans les revues systématiques précitées.

La principale source qui est exploitée ici, particulièrement du point de vue méthodologique, est l'article assez récent de Dehon et al. (2019) où les auteurs s'intéressent à une situation assez similaire où des étudiants, également en Belgique, se sont vus proposer des tests en ligne. Notons que cet article n'a lui-même pas beaucoup de références. Les auteurs ont collecté les données des tests et en ont tiré des nouvelles variables qualifiant l'utilisation de l'outil par chaque étudiant. Ils utilisent alors une régression ordinale pour faire le lien

entre ces variables proposées et des données socioéconomiques et liées au parcours scolaire collectées. À l'aide du modèle d'Heckman, ils déterminent que les données socioéconomiques et liées au parcours scolaire ne biaisent pas la note lors de la comparaison des étudiants n'ayant pas utilisé l'aide en ligne avec ceux qui l'ont utilisé. Ils utilisent alors des régressions linéaires pour montrer que l'aide en ligne a bien influencé positivement la note des étudiants.

C'est cette approche méthodologique qui sera employée dans cette étude : après avoir décrit les données d'entrée dans le chapitre 2, les données de travail sont exploitées dans le chapitre 3 afin de créer des groupements d'étudiants ayant plus ou moins utilisé l'outil en ligne. Le chapitre 4 fait le lien entre l'utilisation de l'outil et les données d'entrée à l'aide d'une régression ordinale. Le chapitre 5 présente les résultats, en détectant l'absence de biais à l'aide d'un modèle d'effet de traitement (qui est une extension du modèle d'Heckman utilisé par Dehon et al.) afin de comparer les étudiants ayant utilisé l'outil avec ceux qui ne l'ont pas utilisé, et conclut à l'aide d'une régression linéaire.

Chapitre 2

Données d'entrée

Les données d'entrée contiennent des informations sur la situation socioéconomique des étudiants ainsi que sur leur parcours scolaire antérieur. Ces données ont été obtenues par une enquête¹, à laquelle il était demandé aux étudiants de répondre lors de leur premier accès au simulateur d'examen.

Ce chapitre décrit les variables disponibles dans cette base de données ainsi que la manière dont les données ont été préparées. De plus, des groupes d'étudiants y sont constitués sur base de ces différentes variables, de telle manière à ce que les étudiants d'un même groupe soient les plus semblables possibles, et les plus différents possibles des étudiants des autres groupes. Cette classification permettra de se donner une idée de l'information que recèle cette base de données.

Initialement, cette base de données comporte 555 observations (étudiants) et 14 variables d'intérêt.

2.1 Variables des données d'entrée

Dans cette base de données, les variables sont toutes qualitatives ou quantitatives discrètes, si on omet la variable d'identification des étudiants (qui permet de faire la jointure avec la base de données de travail, mais ne recèle aucune information en tant que telle²).

1. Le détail des questions posées lors de cette enquête est fourni en document annexe, fichier `questionnaire_entree_details.docx`. Il s'agit d'un document de travail obtenu au début de l'étude, fourni par les concepteurs de l'outil de simulation d'examen (et donc de l'enquête). Notons que ce document est plus complet que nécessaire, et contient des informations en lien avec les enquêtes posées les années suivantes, ce qui est ici sans intérêt.

2. Les données ayant été anonymisées afin de respecter le règlement général de protection des données, cette variable d'identification est un numéro aléatoire unique pour chaque étudiant.

Les variables qualitatives ayant bien souvent trop de modalités peu utilisées, un regroupement de certaines modalités a été opéré. De même, les variables quantitatives discrètes ont été divisées en classes d'effectifs suffisants afin de les traiter de manière qualitative. La manière dont les variables ont été réécrites est détaillée dans la présentation exhaustive des variables disponibles qui suit.

2.1.1 Données socioéconomiques

Les données socioéconomiques de cette base de données sont fournies par les variables suivantes :

- Le genre de l'étudiant : féminin (282 étudiants)³ ou masculin (134).
- L'âge de l'étudiant : cette variable est en réalité quantitative discrète, mais sera exploitée de manière qualitative en regroupant des valeurs en classes : moins de 17 ans (62), 18 ans (202), 19 ans (94) ou plus de 20 ans (58).
- La nationalité de l'étudiant : belge (342) ou autre (74).
- La langue maternelle de l'étudiant : le français (364) ou autre (52).
- Le domicile de l'étudiant pendant l'année académique : avec sa famille (296) ou autre (120). Cette deuxième modalité regroupe les étudiants qui logent en kot (89), en colocation (19) ou autre (12).
- Le plus haut diplôme obtenu par les parents : selon que les deux parents sont diplômés du supérieur (167), un seul du supérieur (140) ou autre (109). À nouveau, il existait davantage de modalités moins représentées qui ont été regroupées dans la modalité *autre* : les deux parents diplômés du secondaire (46) ou un seul (36), rien de plus que du primaire (10) et autre (17).

2.1.2 Données liées au parcours des étudiants

Les données liées au parcours scolaire des étudiants de cette base de données sont fournies par les variables suivantes :

3. Dans cette liste, le nombre entre parenthèses représente à chaque fois l'effectif de la modalité décrite. Cet effectif est comptabilisé après le traitement des données manquantes, détaillé à la section suivante.

- Ce que l'étudiant faisait comme études l'année précédente : dans l'enseignement secondaire (288) ou supérieur (128). Alors que la variable originelle comprenait davantage de modalités, à savoir deuxième terminale à l'étranger (12), haute-école (12) et classes préparatoires (8), en plus de secondaire (276) et université (108), les très faibles effectifs des trois premières modalités sont mieux étudiés sous l'aspect binaire secondaire ou supérieur⁴.
- La situation de l'année actuelle dans le parcours en enseignement supérieur de l'étudiant : s'il est primant (326) ou non (90). La variable initiale précisait si les étudiants étaient répétants (90) ou réorientés (26). Il a été choisi de regrouper les étudiants réorientés *avec les étudiants primants*, ce qui revient à définir un étudiant primant comme un étudiant inscrit pour la première fois dans cette section précise. En effet, ce choix permet de donner un intérêt à cette variable comme indicatrice de si les étudiants ont déjà suivi le cours ou non, plutôt que d'être redondante avec la précédente⁵.
- Si l'étudiant a redoublé en secondaire : non (358) ou oui (58). La variable initiale précisait également que 5 étudiants avaient redoublé plus d'une fois ; ceux-ci font donc partie des 58 étudiants ayant redoublé en secondaire.
- Le type d'enseignement suivi en secondaire : en Fédération Wallonie-Bruxelles (358) ou non (58). La variable initiale précisait, parmi les étudiants ayant suivi un enseignement secondaire en Fédération Wallonie-Bruxelles, ceux qui étaient inscrits en technique de transition (16) ou autre que le général ou le technique de transition (6) mais les effectifs de ces deux dernières modalités sont trop faibles pour pouvoir en tirer la moindre conclusion.
- Le nombre d'heures de physique par semaine suivies en secondaire est une variable quantitative discrète regroupée en classes comme suit :

4. Il est à noter que le questionnaire était tel qu'il n'était pas possible pour les étudiants d'indiquer qu'ils ne faisaient pas d'études l'année précédente. Il est possible que les quelques observations manquantes dans cette variable soient dues à cela, mais il est aussi possible que certaines observations aient été mal encodées de par ce fait. Il est cependant peu probable que cela soit impactant car il n'est pas attendu qu'une telle situation soit fréquente.

5. Ce choix permet aussi de parer à une éventuelle mauvaise compréhension du sens du terme *primant* par les étudiants, qui peut être compris aussi bien au sens d'un étudiant inscrit pour la première fois à l'université, ou bien pour la première fois dans le supérieur, ou bien pour la première fois dans cette section précise selon les interprétations.

une heure (52), deux heures (160), trois heures ou plus (204).

- Le nombre d'heures de mathématiques par semaine suivies en secondaire est également une variable quantitative discrète regroupée en classes comme suit : moins de six heures (166) ou six heures ou plus (250), les effectifs étant bien trop concentrés dans les modalités quatre et six pour pouvoir faire un découpage plus fin.
- Le nombre d'heures de sciences est une variable du même type, qui a été découpée en : cinq heures ou moins (59), six heures (106), sept heures (125) et huit heures ou plus (126).
- Enfin, une dernière variable indique si l'étudiant a suivi des activités préparatoires avant de s'inscrire dans le supérieur (151) ou non (265). La variable initiale précisait également que 9 étudiants avaient suivi plus d'une activité préparatoire et quelles activités avaient été suivies par chaque étudiant qui en avait suivi au moins une, mais ces informations étaient peu utilisables dans cette étude.

À ces données d'entrée s'ajoute la note que les étudiants ont obtenu à l'examen de janvier. Il s'agit d'une variable quantitative continue entre 0 et 20, qui peut aussi être considérée comme qualitative si on considère les modalités réussite ou échec de l'examen selon que cette note vaut plus ou moins que 10/20 respectivement.

2.2 Traitement des données manquantes

Comme précisé précédemment, les effectifs des différentes variables qui viennent d'être présentées à la section précédentes sont ceux après traitement des données manquantes. Cette section précise la manière dont elles ont été traitées.

La plupart des données manquantes sont concentrées dans des étudiants n'ayant pas du tout pris part à l'enquête. D'autres proviennent du fait que certains étudiants n'ont pas répondu à l'une ou l'autre question.

2.2.1 Étudiants n'ayant pas participé à l'enquête

Les données manquantes proviennent pour la plus grande part du fait qu'un nombre assez important d'étudiants (111 étudiants, à savoir très exactement 20% de la population étudiée) n'ont pas participé à l'enquête. La

plupart d'entre eux (100 étudiants) n'ont jamais accédé à l'outil de simulation d'examen⁶, et ne se sont donc pas vu proposer l'enquête. Les 11 autres étudiants n'ont sans doute pas rempli l'enquête malgré le fait qu'elle leur ait été proposée.

La seule analyse que l'on puisse faire pour ces étudiants est d'étudier l'écart de leur note par rapport à ceux qui ont rempli le formulaire.

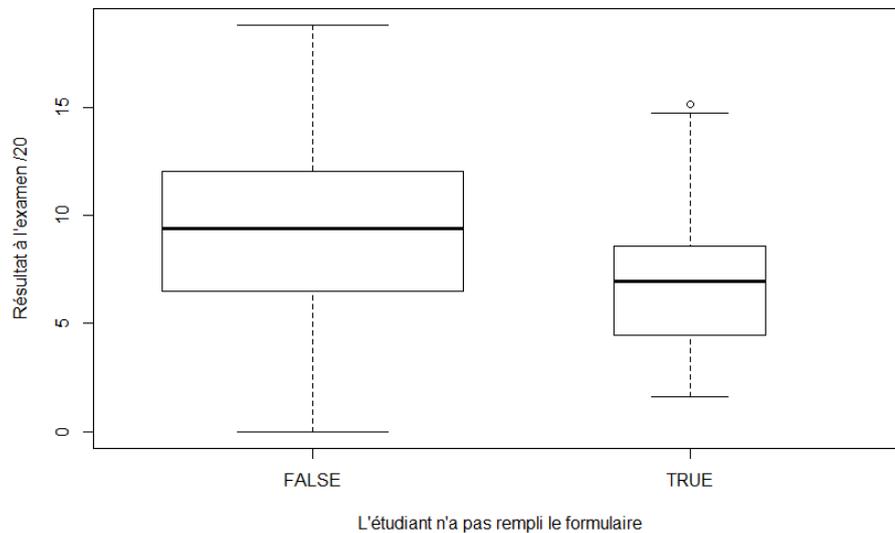


FIGURE 2.1 – Boîte à moustaches de la note des étudiants ayant rempli le formulaire (à gauche) comparée à celle de la note des étudiants n'ayant pas rempli le formulaire (à droite). L'épaisseur des boîtes (largeur le long de l'axe horizontal) est proportionnelle à la racine carrée du nombre d'observations dans le groupe en question.

Les boîtes à moustaches présentées en figure 2.1 montrent que 75% des étudiants n'ayant pas rempli le formulaire ont obtenu une note inférieure à 10/20 ; cette note est plus faible que la moitié des étudiants l'ayant rempli. Les étudiants n'ayant pas rempli le formulaire ont donc globalement moins bien réussi que ceux qui l'ont rempli. Ceci pourrait par exemple s'expliquer par le fait que les élèves n'ayant pas rempli le formulaire, et n'ayant donc pas utilisé l'outil en ligne, étaient moins impliqués dans leurs études, ont

6. Ce nombre de 100 étudiants est le nombre d'étudiants n'ayant eu aucune activité sur l'outil de simulation d'examen, il n'est cependant pas à exclure que certains aient accédé à l'outil sans rien y faire pour autant.

abandonné, ou correspondent à un milieu socioéconomique moins propice à la réussite scolaire.

La plupart des informations utiles pour ces étudiants n'étant pas disponibles, ceux-ci ne seront dès lors pas considérés par la suite. Ceci réduit notre base de données à 444 observations.

2.2.2 Autres données manquantes

Sans considérer ces étudiants n'ayant pas rempli du tout le formulaire, il reste tout de même quelques données manquantes dispersées dans la base de données, dont on peut voir une représentation en figure 2.2. Dans cette figure, chaque colonne représente une variable et chaque ligne, une observation de la base de données. Un trait noir indique une donnée manquante, pour la variable et l'observation concernée.

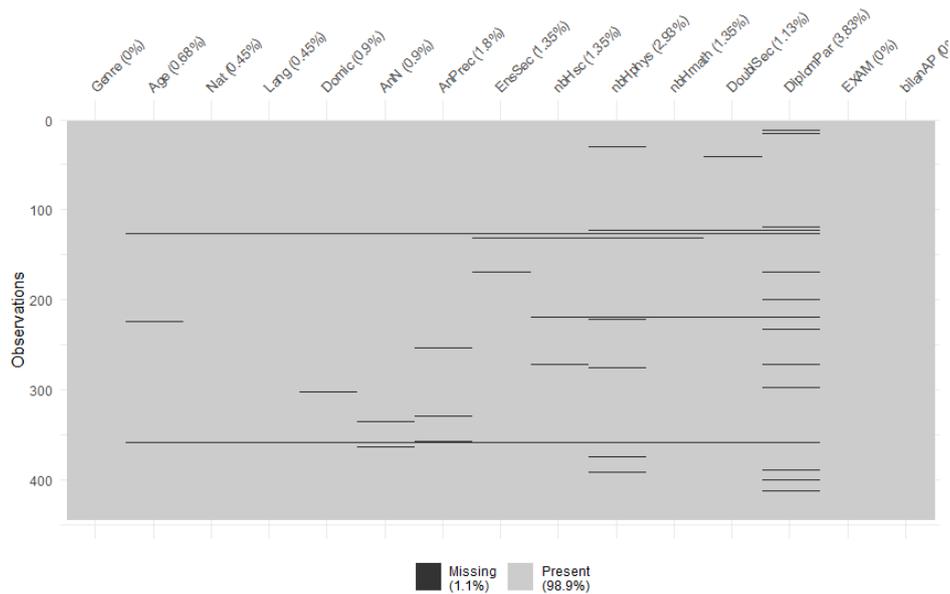


FIGURE 2.2 – Visualisation des données manquantes dans la base de données, où les étudiants pour lesquels la plupart des données étaient manquantes sont exclus. Les rangées concernent les observations (étudiants) et les colonnes les variables. Un trait noir signale une donnée manquante pour une variable, pour une observation donnée.

On constate sur la figure 2.2 que deux étudiants n'ont pas donné plus d'information que leur genre. De plus, deux autres étudiants semblent avoir omis la plupart des informations en lien avec les cours suivis en secondaire.

Du point de vue des variables, la variable détaillant le plus haut diplôme des parents est la variable qui détient le plus de valeurs manquantes (13 étudiants, en ne comptant pas les deux étudiants mentionnés précédemment). Le nombre d'heures de physique en secondaire suit, avec 9 étudiants. Il y a également quelques valeurs manquantes dans les variables sur le parcours scolaire (année en cours et année précédente) ainsi que pour le nombre d'heures de sciences suivies en secondaire.

Certaines méthodes statistiques utilisées par la suite, telles que l'analyse en composantes principales ou la régression ordinale, posent des difficultés lorsqu'il y a des données manquantes dans la base de données⁷. Il est donc nécessaire de traiter ces données manquantes.

Par simplicité, il a été choisi de travailler en cas complet, c'est-à-dire d'éliminer de l'étude les observations qui possèdent des données manquantes. En effet, les données manquantes sont en faible nombre et dispersées entre les différentes variables. De plus l'imputation simple qui consisterait, pour ces variables qualitatives, à remplacer les données manquantes par la modalité la plus fréquente, est problématique car elle modifie la distribution des modalités de la variable de manière peu justifiée. Remarquons que d'autres approches plus complexes, telles que l'imputation multiple, auraient pu être employées.

Avant de travailler en cas complet, il est toutefois nécessaire de s'assurer que le retrait des étudiants pour lesquels une donnée manquait n'introduit pas un biais dans notre base de données. Afin de quantifier ce biais, il est utile de s'intéresser à la note de l'examen, étant donné que l'utilité de cette base de données d'entrée dans cette étude est, au final, de déterminer si ces données d'entrée biaisent la note lors de l'étude de l'utilité de l'outil sur l'amélioration de la note.

En particulier, il est utile de s'intéresser aux données manquantes de la variable en lien avec le plus haut diplôme des parents. En effet, cette variable contient un certain nombre de données manquantes, et une hypothèse qui pourrait être formulée est que les étudiants qui n'ont pas rempli cette case du formulaire aient des parents peu diplômés, ou soient en tout cas dans une situation socioéconomique moins favorable à la réussite scolaire.

Comme on peut le voir en figure 2.3, le fait que la donnée manque pour cette variable ne semble pas avoir d'influence majeure sur la note. En effet, bien que la médiane de la note des étudiants n'ayant pas renseigné le diplôme

7. Il est en fait possible d'adapter l'ACP afin d'utiliser une matrice de variance-covariance dont chaque élément de la matrice est calculé sur base des données non manquantes pour les variables concernées. Il en résulte néanmoins une matrice de variance-covariance qui parfois n'est même pas semi-définie positive, ce qui est un écueil qu'il est préférable d'éviter.

des parents soit légèrement plus basse que pour les autres étudiants, cela est contrebalancé par la grande variabilité dans ces données. Les moyennes des notes ne sont pas bien différentes non plus : 8,73/20 pour les étudiants n'ayant pas renseigné cette variable à comparer à 9,31/20 pour les autres⁸. Il devrait donc être possible d'éliminer ces étudiants dont la donnée est manquante pour cette variable sans trop biaiser la base de données, en tout cas en ce qui concerne la note à l'examen.

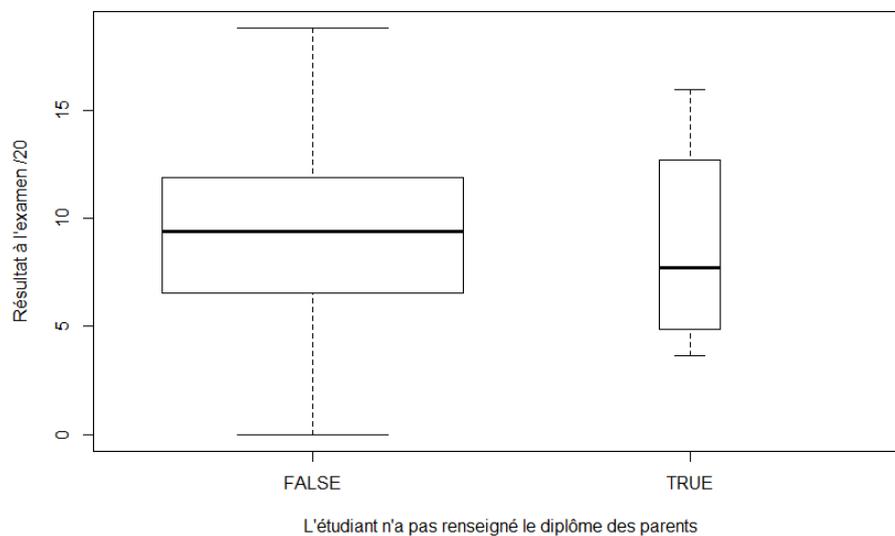


FIGURE 2.3 – Boîte à moustaches de la note des étudiants ayant renseigné le plus haut diplôme des parents (à gauche) comparée à celle de la note des étudiants ne l'ayant pas fait (à droite). L'épaisseur des boîtes est proportionnelle à la racine carrée du nombre d'observations dans le groupe en question.

Pour les autres variables, le nombre de données manquantes est assez faible, et il est donc peu probable que l'utilisation de cas complets induise un biais majeur. On se permettra d'éliminer les quelques étudiants concernés par une de ces données manquantes pour la suite.

La base de données en cas complet, constituée uniquement des étudiants

8. Le test de Wilcoxon ne rejette pas l'égalité des médianes, avec une p-valeur de 0,45. Le test t de Welch (adaptation du test t de Student au cas hétéroscédastique) ne rejette pas plus l'égalité des moyennes, avec une p-valeur de 0,52. L'emploi de ce dernier test est justifié car le test de Shapiro-Wilk ne rejette pas la normalité dans les deux échantillons comparés (étudiants n'ayant pas renseigné le diplôme des parents et étudiants l'ayant renseigné) avec des p-valeurs de 0,12 et 0,068 respectivement.

pour lesquels toutes les données sont présentes, est ainsi constituée de 405 observations.

2.3 Liens entre différentes variables d'entrée

Cette section étudie les liens entre les différentes variables qui composent cette base de données. Pour ce faire, un clustering hiérarchique⁹ des variables avec une mesure de dissimilarité basée sur la corrélation va être utilisé. Cette méthode permettra d'obtenir des groupes de modalités corrélées entre elles, et non corrélées (ou peu corrélées) avec les modalités des autres groupes, ce qui permettra d'explorer les liens linéaires entre les variables.

En détail, les différentes variables décrites précédemment ont été remplacées par des variables binaires. Pour ce faire, on considère chaque modalité de chaque variable comme une variable binaire qui vaut 1 lorsque la variable initiale prend cette modalité, et 0 sinon. Une variable à m modalités fournit donc m variables binaires. Il est à noter que la m^{e} variable binaire est redondante, car celle-ci vaut 1 si et seulement si les $m - 1$ autres valent 0. Il a toutefois été décidé de garder à chaque fois cette dernière modalité afin de permettre à toutes les modalités d'apparaître dans le résultat final de la classification et de permettre une meilleure interprétation des groupes qui suivent. Cette approche implique l'obtention d'une matrice de variance-covariance singulière, ce qui est parfois problématique mais ne pose pas de problème pour effectuer un clustering hiérarchique. On obtient alors, hors des 15 variables étudiées, 34 variables binaires dont la liste est proposée en annexe 1.

Le clustering hiérarchique est en général employé pour constituer des groupes d'observations. Il est cependant possible d'utiliser cette même méthode pour classifier des variables en appliquant la méthode à la base de données transposée, où les observations deviennent les variables et vice-versa.

Cette méthode se base sur une matrice de dissimilarité. Afin de répondre à l'objectif de ce chapitre qui est de déterminer les éventuels liens entre les variables, la mesure de dissimilarité basée sur les corrélations semble appropriée¹⁰, et comme la base de données est transposée, cette matrice de dissimilarité est une matrice 34×34 dont les cellules sont les mesures de la dissimilarité entre chaque couple de variables binaires.

9. Härdle & Simar (2015), pp. 393–396.

10. La mesure de dissimilarité basée sur les corrélations entre x et y est $1 - \text{cor}(x, y)$, où $\text{cor}(x, y)$ est la corrélation entre les variables x et y . Cette mesure est donc proche de 0 si les deux variables sont corrélées, proche de 1 si elles ne le sont pas et proche de 2 si elles sont anti-corrélées.

Le clustering est effectué à l'aide de la fonction `hclust` de R. Celle-ci prend en paramètres la matrice de dissimilarité et l'algorithme utilisé. L'algorithme utilisé ici est la méthode des liens complets (paramètre `complete` de la fonction de R). Cette méthode, qui est par ailleurs celle par défaut de R, est une méthode très commune de clustering, où la dissimilarité entre deux groupes est déterminée par la plus grande mesure de dissimilarité entre des éléments de ces groupes.

Cette fonction permet de produire l'arbre de la figure 2.4.

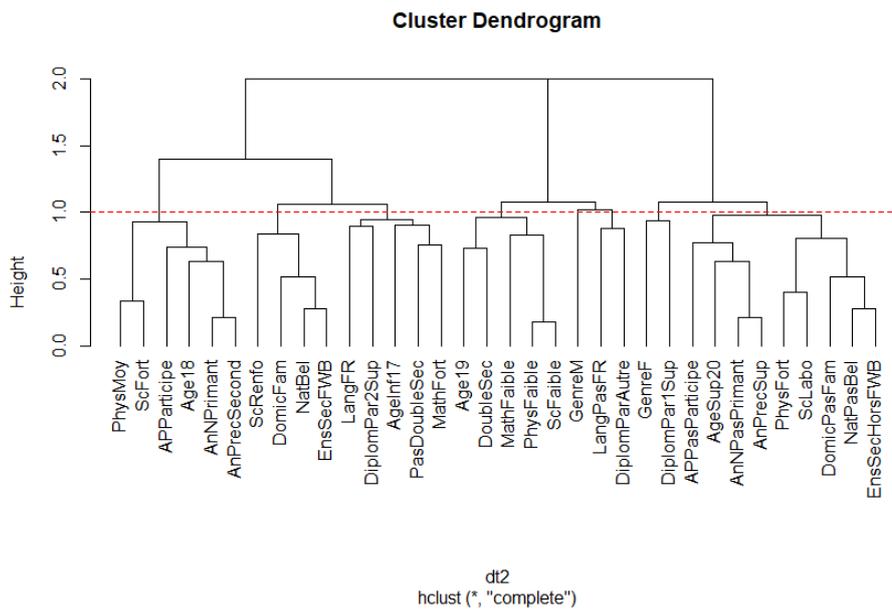


FIGURE 2.4 – Représentation graphique (sous forme d'arbre) du regroupe-ment hiérarchique des modalités des variables étudiées.

Sur cet arbre, les modalités qui sont regroupées à des hauteurs faibles sont les plus corrélées. Cette représentation permet donc de faciliter la lecture des corrélations entre les différentes variables ¹¹.

En coupant à la hauteur médiane (qui vaut 1, c'est-à-dire là où les groupes constitués ne sont plus considérés comme corrélés ¹²), comme indiqué sur la

11. La matrice des corrélations entre les variables binaires est par ailleurs présentée en annexe 2.

12. Vu que la méthode des liens complets choisit la plus grande valeur pour cette mesure de dissimilarité, alors si deux groupes sont reliés à une hauteur supérieure à 1, cela signifie que chacun de ces deux groupes contient une variable qui est telle que la corrélation entre ces deux variables est inférieure ou égale à 0.

figure, on obtient une idée des modalités qui sont souvent liées. Voici une description des groupes constitués, de gauche à droite :

- Les étudiants qui sortent de secondaire sont souvent primants, et âgés de 18 ans. Ils ont de plus plus souvent participé aux activités préparatoires. Le fait d'avoir suivi 2 heures de physique par semaine est lié au fait d'avoir suivi 6 heures de sciences par semaine, et ces deux variables sont plus faiblement liées aux précédentes.
- Les étudiants belges ont souvent fait leurs secondaires en Fédération Wallonie-Bruxelles. Il y a également un certain lien entre ces modalités et le fait d'être domicilié avec sa famille (sans doute par comparaison avec les étudiants étranger) et un lien plus faible avec le fait d'avoir suivi 7 heures de sciences par semaine en secondaire.
- Dans le troisième groupe en partant de la gauche, on voit qu'il y a un certain lien entre le fait de parler français, d'avoir deux parents diplômés du supérieur, d'avoir 17 ans ou moins, de ne pas avoir doublé en secondaire, et le fait d'avoir suivi plus de 6 heures de mathématiques par semaine en secondaire, mais toutes ces corrélations sont relativement faibles.
- Le fait d'avoir 19 ans est corrélé avec le fait d'avoir doublé en secondaire. De plus, le fait d'avoir suivi peu de sciences, peu de mathématiques et peu de physique en secondaire est lié. Toutes ces modalités sont également faiblement corrélées.
- Le fait d'être un homme n'est lié à aucune autre modalité. Ne pas parler français est faiblement corrélé avec des parents non diplômés du supérieur.
- Le fait d'être une femme est faiblement corrélé avec le fait d'avoir un seul parent diplômé du supérieur.
- Les modalités de ne pas avoir participé aux activités préparatoires, d'être âgé de 20 ans ou plus, d'être répétant et d'avoir déjà fait une année dans le supérieur sont toutes corrélées. Il s'agit en quelque sorte du symétrique des liens observés dans le premier groupe.
- Enfin, les étudiants ayant suivi beaucoup de physique en secondaire ont aussi suivi beaucoup de sciences. Les étudiants étrangers ont tendance à ne pas être domicilié avec leur famille et à avoir fait leurs secondaires hors de la Fédération Wallonie-Bruxelles, ce qui est à nouveau le relatif symétrique du deuxième groupe de modalités.

Cette analyse exploratoire permet d’avoir un premier aperçu des liens linéaires entre les variables et sera notamment utile lors de la sélection des variables pertinentes dans les méthodes linéaires qui seront employées par la suite.

2.4 Classification initiale des étudiants

Afin d’établir des profils d’étudiants en fonction des données d’entrée, il est à nouveau possible d’effectuer un clustering hiérarchique, cette fois sur les observations (c’est-à-dire avec la base de données non transposée).

On souhaite, à l’aide de cette méthode, classer les 405 étudiants de la base de données d’entrée dans des groupes, selon les 15 variables détaillées ci-dessus. À nouveau, les variables sont utilisées sous leur forme binaire. Par cohérence avec ce qui a été fait précédemment, et puisque cela ne pose pas de problème particulier, les modalités redondantes sont de nouveau conservées. Les variables se décomposent donc toujours en 34 variables binaires. La mesure de dissimilarité et l’algorithme employés diffèrent cependant de ceux qui ont été employés dans la section précédente.

2.4.1 Constitution de groupes d’étudiants

Comme précédemment, il est tout d’abord nécessaire de choisir une mesure de dissimilarité appropriée, ainsi qu’un algorithme d’agglomération. La mesure de dissimilarité choisie est celle de Jaccard, qui est davantage appropriée pour les variables binaires. L’algorithme d’agglomération utilisé est la méthode de Ward, dans une version généralisée aux mesures de dissimilarité autres que la distance euclidienne.

Mesure de dissimilarité de Jaccard. La mesure de dissimilarité de Jaccard entre deux observations (dont les attributs sont binaires) est définie comme le nombre d’attributs qui diffèrent d’une observation à l’autre divisé par le nombre d’attributs où la valeur 1 apparaît pour l’une ou l’autre des observations, autrement dit,

$$\frac{M_{01} + M_{10}}{M_{01} + M_{10} + M_{11}}$$

où $M_{v_1 v_2}$ (avec $v_1, v_2 \in \{0, 1\}$) est le nombre d’attributs qui valent v_1 pour la première observation et v_2 pour la seconde¹³. Cette expression n’a pas de

13. Härdle & Simar (2015), pp. 388–389, présentent la *mesure de similarité de Jaccard*. On obtient simplement la *mesure de dissimilarité de Jaccard*, présentée ici, et directement

sens si les deux observations considérées ont 0 pour chaque attribut (car le dénominateur est alors nul). Il est cependant possible de poser la dissimilarité valant 0 dans ces cas-là, ce qui a du sens étant donné que les deux individus ont alors obtenu exactement les mêmes valeurs pour chaque variable. C'est d'ailleurs la convention qui est suivie par le logiciel R.

Cette mesure de dissimilarité entre des variables binaires est asymétrique : en effet, elle ne donne pas le même poids aux valeurs 0 qu'aux valeurs 1. Comme M_{00} n'apparaît pas au dénominateur, cela implique que si deux observations ont pour le même attribut la valeur 0, cela ne réduit pas la dissimilarité entre les deux observations (sans pour autant l'augmenter), ce qui aurait été le cas si ces deux observations avaient pour ce même attribut la valeur 1. Cette propriété est utile dans ce cas concret, étant donné qu'il est préférable de regrouper des étudiants davantage sur base de leurs points communs que sur base de leurs différences.

Algorithme de Ward. Contrairement à la section précédente, où la méthode des liens complets était utilisée, l'algorithme choisi ici pour ses résultats plus convaincants est l'algorithme de Ward¹⁴. L'idée de cet algorithme est de constituer des groupes les plus homogènes possibles, en regroupant des éléments de manière à ne pas trop augmenter une certaine mesure d'hétérogénéité I_R , définie pour chaque groupe R par

$$I_R = \frac{1}{n_R} \sum_{i=1}^{n_R} d^2(x_i, \bar{x}_R)$$

où d est une mesure de dissimilarité, n_R est le nombre d'observations dans le groupe R , x_1, \dots, x_{n_R} les n_R observations du groupe R et \bar{x}_R la moyenne des observations du groupe R .

Cette méthode est également très utilisée, mais est souvent présentée comme ne fonctionnant qu'avec pour mesure de dissimilarité d la distance euclidienne¹⁵. Cependant, Batagelj (1988) montre que l'algorithme de Ward peut être généralisé à d'autres mesures de dissimilarité. En particulier, il montre que la formule de Lance-William-Jambu reste valide pour toute mesure de dissymétrie d . Or, comme l'indiquent Murtagh & Legendre (2014), il s'agit de la formule implémentée dans la méthode nommée `ward.D` pour

implémentée dans R, en prenant son complémentaire : si J est la mesure de similarité de Jaccard calculée entre deux observations, alors la mesure de dissimilarité de Jaccard d_J entre ces deux mêmes observations est obtenue par $d_J = 1 - J$.

14. Härdle & Simar (2015), pp. 396–399

15. Sans interdire explicitement l'utilisation d'une autre mesure de dissimilarité, Härdle & Simar (2015) n'en utilisent pas d'autre.

la fonction `hclust` dans R. On en conclut qu'il est possible d'utiliser l'algorithme de Ward pour notre mesure de dissimilarité, à condition d'utiliser la méthode `ward.D` de `hclust` dans R.

Regroupement hiérarchique. La fonction `hclust` de R, avec pour matrice de dissimilarité la matrice des mesures de dissimilarité de Jaccard entre chaque couple d'étudiants, et comme algorithme d'agglomération `ward.D` permet d'obtenir l'arbre de la figure 2.5. Contrairement à l'arbre précédent, les feuilles de l'arbre qui sont les observations ne sont pas affichées par question de lisibilité. En principe, au bas de chaque branche devrait se trouver le numéro de l'observation qui lui est associé, mais vu le grand nombre d'observations (405 étudiants), les afficher provoque des chevauchements qui rendent le tout illisible.

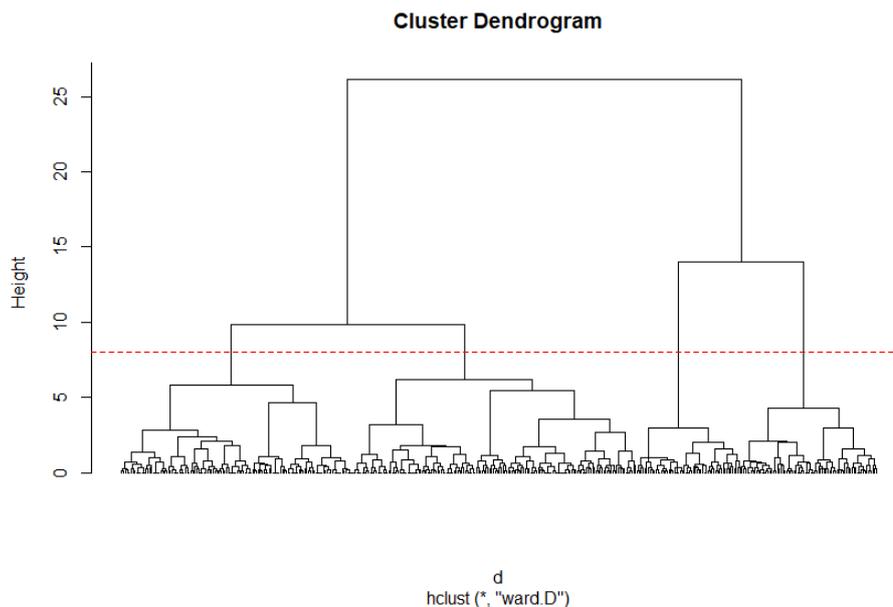


FIGURE 2.5 – Représentation graphique (sous forme d'arbre) du regroupement hiérarchique des étudiants, sur base de leurs données d'entrée.

Afin d'obtenir un regroupement, il est nécessaire de choisir la hauteur à laquelle couper l'arbre. Empiriquement, il est souvent proposé de couper à la hauteur médiane (dans ce cas, à peu près 13). Cependant, couper à cette hauteur provoque la création d'un énorme groupe et de deux groupes beaucoup plus petits. C'est pourquoi il a été choisi de couper à une hauteur plus faible (8), ce qui permet de diviser le plus gros groupe en deux plus petits groupes et d'obtenir des effectifs plus comparables. Précisément, les effectifs

obtenus par ce découpage sont présentés dans la table 2.1 dans l'ordre de gauche à droite de la figure 2.5.

Groupe	3	2	1	4
Effectif	155	117	62	82

TABLE 2.1 – Table des effectifs des groupes d'entrée constitués par clustering hiérarchique.

2.4.2 Description des groupes

Afin d'interpréter les groupes qui ont été formés par ce clustering hiérarchique, il est possible de commencer par regarder comment les groupes se répartissent dans le plan principal. Cette analyse par composantes principales guidera vers la constitution des profils-types de chaque groupe, qui seront obtenus formellement en étudiant les effectifs de chaque groupe.

Analyse en composantes principales. L'analyse en composantes principales consiste à réexprimer les 34 variables binaires¹⁶ à l'aide de 34 nouvelles composantes dites principales, toutes non corrélées et construites de façon à ce que chacune des composantes exprime la plus grande part de variance possible de la base de données qui n'ait pas déjà été exprimée par une autre composante principale.

L'objectif est de pouvoir résumer la base de données en le moins de composantes possibles. Ainsi, plus la part de variance est grande pour les premières composantes principales, et moins de composantes sont nécessaires pour avoir une bonne représentation de la variabilité dans la base de données.

Dans ce cas, plutôt que d'utiliser la matrice de variance-covariance de notre base de données pour effectuer cette analyse en composantes principales, la matrice de corrélation sera utilisée, ce qui revient à normaliser les variables et permet de donner plus d'importance aux liens entre les variables plutôt qu'à la variabilité de chacune d'entre elles.

Commençons alors par procéder à une analyse par composantes principales sur base des 405 observations complètes, et pour les 15 variables binarisées (en conservant les redondances, qui à nouveau, ne posent pas de problème avec cette méthode). La figure 2.6 indique la quantité de variance qu'il reste à exprimer dans la base de données selon le nombre de composantes principales que l'on a choisi. On y observe que deux composantes donnent déjà de très

16. À nouveau, on se permet de conserver les variables binaires redondantes, étant donné que la singularité de la matrice de variance-covariance n'est pas un problème dans ce cas.

bons résultats, mais il est avantageux d'en ajouter une troisième. Un plateau apparaît alors, et il semble peu utile d'ajouter davantage de composantes principales pour obtenir de l'information.

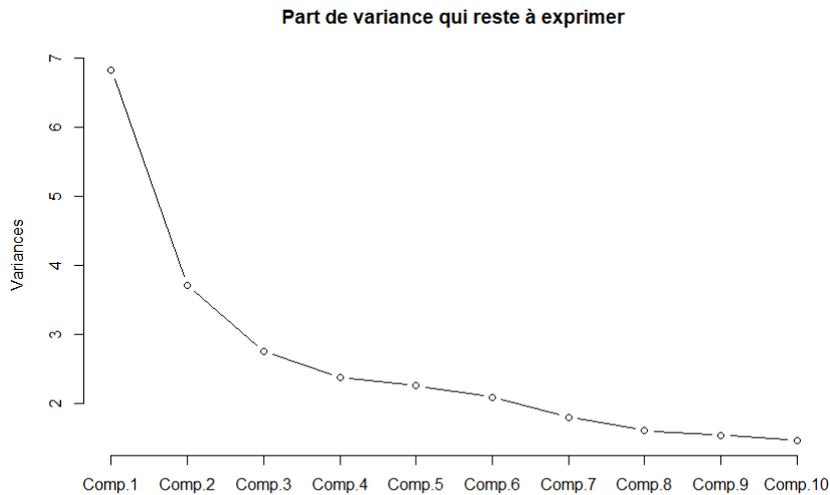


FIGURE 2.6 – Quantité de variance qu'il reste à exprimer dans la base de données selon le nombre de composantes principales que l'on choisit.

Par facilité, commençons par regarder le plan principal, en deux dimensions. Celui-ci est présenté en figure 2.7. Les observations y sont représentées selon leurs deux premières composantes principales (la première sur l'axe des x et la seconde sur l'axe des y). Les différents points sont des observations qui ont été colorées selon leur groupe d'appartenance.

On observe que tous les groupes, à l'exception des groupes 2 et 3, semblent bien séparés sur le plan principal, ce qui semble un bon signe quant à leur interprétabilité.

Pour ce qui en est des groupes 2 et 3, on peut davantage les distinguer lorsque l'on regarde le plan constitué de la première composante principale en abscisse et de la troisième composante principale en ordonnée, comme on le voit en figure 2.8. Il n'est pas étonnant qu'il soit plus difficile de distinguer ces deux groupes, étant donné que ceux-ci auraient été un seul et même groupe si l'on avait coupé l'arbre hiérarchique à la hauteur médiane.

Loadings des composantes principales. Il est tout d'abord possible de consulter les *loadings* des trois composantes principales utilisées. Il s'agit

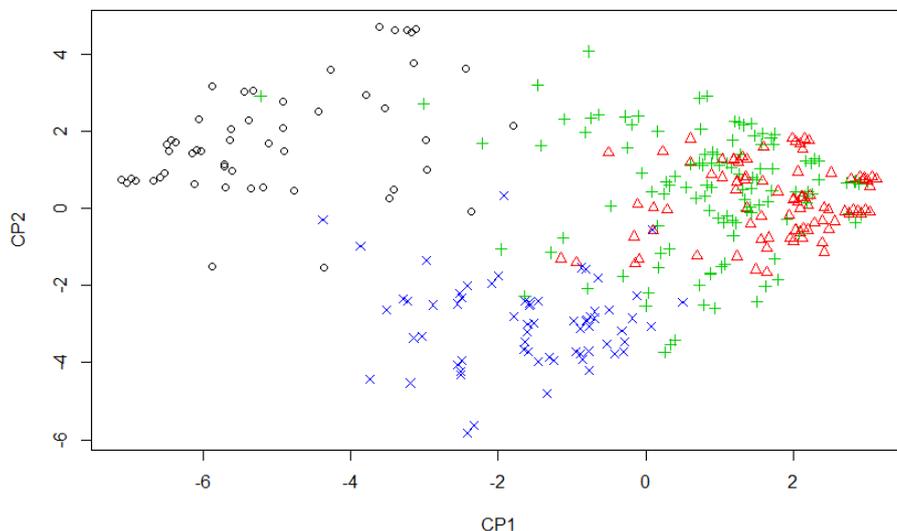


FIGURE 2.7 – Représentation des observations dans le plan principal. Les cercles noirs sont des étudiants du groupe 1, les triangles rouges du groupe 2, les + verts du groupe 3 et les X bleus du groupe 4.

des coefficients donnés à chaque variable afin de constituer les composantes principales comme combinaisons linéaires des variables initiales.

Ces *loadings* pour les trois composantes principales considérées peuvent être visualisés dans la figure 2.9, où les points bleus indiquent que lorsque la variable binaire vaut 1, la composante principale considérée augmente, tandis que les points rouges indiquent que lorsque la variable binaire vaut 1, la composante principale considérée diminue.

L'analyse de ces *loadings*, ainsi que les données de la table 2.2 qui présente la répartition des modalités de chacune des 15 variables d'entrées en fonction de ses modalités au sein de chaque groupe permettent alors de se donner une idée de l'individu-type de chacun de ces groupes. Notons que cette description est empreinte d'une part de subjectivité et qu'elle n'est le fruit que de méthodes exploratoires des données.

Groupe 1. Les individus de ce groupe sont caractérisés par des valeurs négatives de la première composante principale, et des valeurs positives de la deuxième.

En consultant les *loadings*, on s'aperçoit que cela signifie que ces individus sont typiquement de nationalité étrangère, ne se domicilient pas avec leur

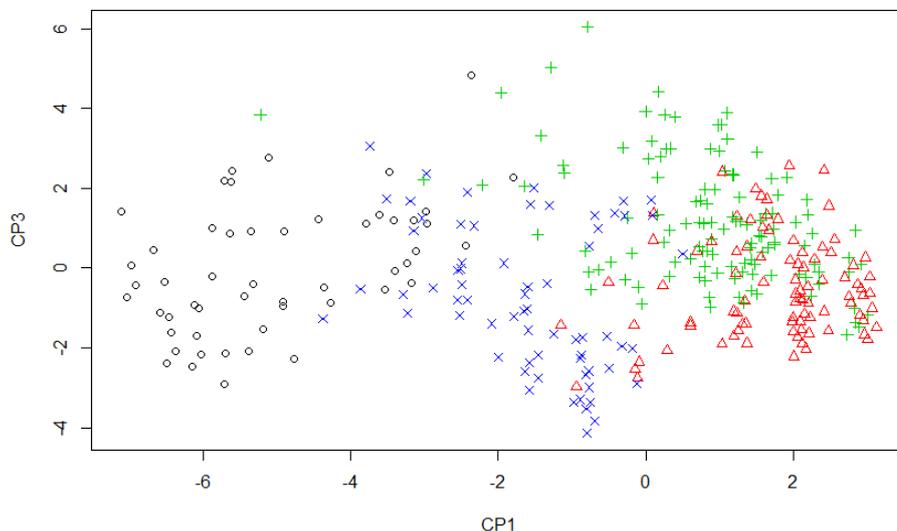


FIGURE 2.8 – Représentation des observations dans le plan constitué de la première et de la troisième composantes principales. Les cercles noirs sont des étudiants du groupe 1, les triangles rouges du groupe 2, les + verts du groupe 3 et les X bleus du groupe 4.

famille, ont effectué leurs études secondaires hors de la Fédération Wallonie-Bruxelles et ont suivi plus d’heures de sciences et de physique en secondaire.

Le tableau confirme ces informations, et permet également de voir que ces individus sont typiquement plus âgés (au moins 19 ans), sont plus souvent répétants que les groupes 2 et 3 et n’ont très majoritairement pas suivi d’activités préparatoires.

En conclusion, ce groupe recense majoritairement des étudiants étrangers, ayant eu une formation scientifique en secondaire.

Groupe 2. Les étudiants de ce groupe sont caractérisés par des valeurs positives de la première composante principale et des valeurs négatives de la troisième composante principale.

En consultant les *loadings*, on s’aperçoit que cela signifie que ces individus sont typiquement jeunes (18 ans ou moins), Belges francophones, domiciliés en famille, primants, sortent de secondaire, en Fédération Wallonie-Bruxelles, ont suivi 2 heures de physique, 6 heures de sciences et plus de 6 heures de mathématiques par semaine en secondaire. Leurs parents sont davantage diplômés du supérieur.

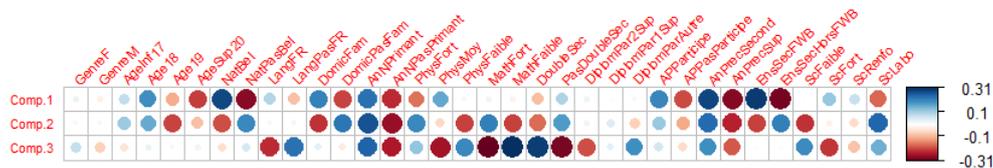


FIGURE 2.9 – Représentation des *loadings* des trois premières composantes principales de la base de données contenant les variables binaires (avec redondance). Un point bleu indique que la variable contribue positivement à la composante alors qu’un point rouge indique que la variable y contribue négativement.

Le tableau confirme ces informations. On constate cependant que ce qui est important pour le plus haut diplôme des parents est qu’au moins un des deux soit diplômé du supérieur. On constate également que comparative- ment aux autres groupes, très peu d’étudiants de ce groupe ont doublé en secondaire.

En conclusion, ce groupe recense majoritairement des étudiants belges sortant du secondaire, qui n’ont jamais doublé dans le secondaire, et dont les parents sont diplômés du supérieur. En secondaire, ils ont typiquement choisi plus d’heures de mathématiques au détriment d’heures de sciences.

Groupe 3. Le groupe 3 est le groupe au plus grand effectif et est assez similaire au précédent (qui avait le deuxième plus grand effectif). La distinction de ce groupe du précédent provient du choix de couper à une hauteur plus faible que la hauteur médiane dans l’arbre hiérarchique et il avait fallu utiliser la troisième composante principale pour voir une différence entre ces deux groupes.

En comparaison avec le groupe précédent, on y retrouverait donc, si on lit les *loadings* de la troisième composante principale, plus d’étudiants dont la langue maternelle n’est pas le français et d’étudiants ayant doublé en secondaire. Ils ont privilégié, en secondaire, des heures de science plutôt que de mathématiques, et leurs parents ne sont typiquement pas diplômés du supérieur.

Le tableau permet de remarquer que, si ce groupe contient davantage d’étudiants dont les parents ne sont pas diplômés du supérieur, il en reste quand même une grande proportion qui le sont.

En conclusion, ce groupe recense majoritairement des étudiants belges sortant du secondaire et dont les parents sont moins diplômés que ceux du groupe 2. En secondaire, ils ont typiquement choisi plus d’heure de sciences

au détriment des heures de mathématiques. La proportion de membres de ce groupe ayant doublé dans le secondaire est plus grande que dans le groupe 2, et comparable aux autres groupes.

Groupe 4. Les membres du quatrième groupe sont caractérisés par une deuxième composante principale négative.

En regardant les *loadings*, on peut en conclure que leurs membres sont plus âgés, Belges, domiciliés en famille, sont plutôt répétants et ont suivi peu de cours scientifiques ou de mathématiques en secondaire.

Le tableau nous permet de vérifier que ce groupe a le plus grand nombre de répétants en son sein. De plus, aucun étudiant de ce groupe vient du secondaire et tous ont fait une année de supérieur auparavant. Toutes les modalités en lien avec les choix de cours en secondaire sont équilibrées, ce qui implique que les modalités en lien avec une formation moins scientifique sont surreprésentées dans ce groupe par rapport aux autres groupes.

En conclusion, ce groupe recense les étudiants répétants ou réorientés, qui ont moins souvent que les autres suivi une formation scientifique en secondaire.

2.5 Liens entre les données d'entrées et la note obtenue à l'examen

Il peut être intéressant de voir comment la note obtenue à l'examen se comporte en fonction des variables d'entrée. Il s'agit ici de premiers résultats proposés dans un but descriptif; les liens entre la note et les autres données disponibles (autant ces données d'entrée que les données de médailles présentées au chapitre 3) seront davantage étudiés dans le chapitre 5.

Note des étudiants en fonction du groupe. Comme des groupes ont été formés à l'aide d'un clustering hiérarchique, il est naturel de vouloir comparer les notes dans les différents groupes.

La figure 2.10 permet de comparer les notes des différents groupes à l'aide des boîtes à moustaches.

On constate que les notes obtenues ne semblent pas très différentes d'un groupe à l'autre. Tout au plus peut-on constater que le groupe 4 (le groupe des répétants) est le seul pour lequel plus de la moitié des étudiants ont réussi l'épreuve.

Effectif	Gr1 56	Gr2 125	Gr3 151	Gr4 73	Pop 405
Femme	60.71	71.20	68.21	67.12	67.90
Homme	39.29	28.80	31.79	32.88	32.10
17 ans ou moins	5.36	24.80	16.56	0.00	14.57
18 ans	17.86	66.40	61.59	15.07	48.64
19 ans	37.50	4.80	17.88	50.68	22.47
20 ans ou plus	39.29	4.00	3.97	34.25	14.32
Belge	7.14	96.00	91.39	94.52	81.73
Autre nationalité	92.86	4.00	8.61	5.48	18.27
Parle français à la maison	80.36	96.80	83.44	84.93	87.41
Parle une autre langue	19.64	3.20	16.56	15.07	12.59
Domicilié avec sa famille	5.36	74.40	84.77	84.93	70.62
Autre domicile pendant l'année	94.64	25.60	15.23	15.07	29.38
2 parents diplômés du sup.	39.29	42.40	41.72	39.73	41.23
1 seul parent diplômé du sup.	44.64	43.20	25.17	31.51	34.57
Parents non diplômés du sup.	16.07	14.40	33.11	28.77	24.20
Primant (ou réorienté)	55.36	99.20	100.00	16.44	78.52
Répétant	44.64	0.80	0.00	83.56	21.48
Secondaire l'année précédente	25.00	94.40	98.01	0.00	69.14
Supérieur l'année précédente	75.00	5.60	1.99	100.00	30.86
A participé aux AP	3.57	49.60	49.01	8.22	35.56
Pas d'activité préparatoire	96.43	50.40	50.99	91.78	64.44
Secondaire en FWB	5.36	99.20	98.01	98.63	85.68
Secondaire hors FWB	94.64	0.80	1.99	1.37	14.32
N'a pas doublé en secondaire	78.57	96.80	82.12	79.45	85.68
A doublé en secondaire	21.43	3.20	17.88	20.55	14.32
Options suivies en secondaire					
6h ou plus de math	64.29	70.40	56.95	45.21	60.00
Moins de 6h de math	35.71	29.60	43.05	54.79	40.00
3h ou plus de physique	92.86	4.00	73.51	39.73	48.64
2h de physique	7.14	80.00	20.53	31.51	39.01
1h de physique	0.00	16.00	5.96	28.77	12.35
8h ou plus de sciences	83.93	2.40	43.05	12.33	30.62
7h de sciences	3.57	18.40	46.36	34.25	29.63
6h de sciences	5.36	63.20	3.97	21.92	25.68
Moins de 6h de sciences	7.14	16.00	6.62	31.51	14.07

TABLE 2.2 – Pourcentage d'étudiants vérifiant les différentes modalités des variables d'entrée au sein de chaque groupe d'entrée constitué par clustering hiérarchique, ainsi que dans la population étudiée entière. Les effectifs considérés comme importants dans un groupe par rapport aux autres (de manière assez subjective) ont été mis en évidence.

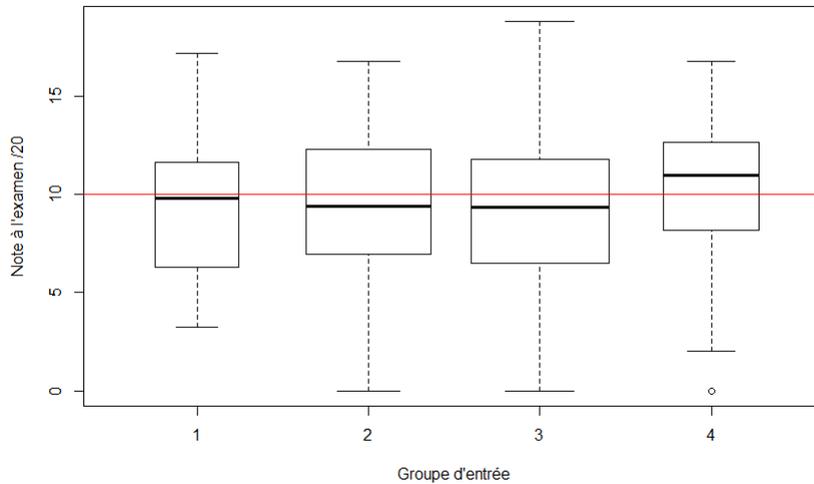


FIGURE 2.10 – Boîtes à moustaches de la note obtenue à l’examen en fonction du groupe d’appartenance. L’épaisseur des boîtes est proportionnelle au nombre d’observations dans le groupe considéré. La ligne horizontale rouge indique le seuil de réussite.

Réussite des étudiants en fonction des variables d’entrée. Il est également possible, pour chacune des 34 variables binaires constituées précédemment, de déterminer la proportion d’étudiant vérifiant cette modalité qui ont réussi l’examen. Cette information est présentée en table 2.3.

Lors de la constitution de cette table, une régression logistique de la réussite à l’examen en fonction de chaque variable d’entrée séparément a été effectuée. Les variables qui y étaient significatives sont suivies d’un astérisque. Il est à noter que la multiplication des tests augmente le risque qu’une variable soit donnée significative alors qu’elle ne le serait pas. Il est donc nécessaire d’appliquer une correction au seuil de significativité afin de prendre compte du risque réel. Cette correction n’a cependant pas été appliquée ici, étant donné qu’il s’agit simplement, de manière exploratoire, de mettre en évidence quelques variables qui pourraient avoir un rôle plus important dans la réussite de l’étudiant. Une étude plus poussée de la réussite de l’étudiant en fonction des données d’entrée (et de travail) à l’aide de modèles linéaires sera effectuée dans le chapitre 5.

En particulier, on constate que les facteurs propices à la réussite sont le fait de parler le français à la maison, d’avoir deux parents diplômés du supérieur, d’être répétant, d’avoir déjà fait une année dans le supérieur, de

n'avoir jamais doublé en secondaire et d'avoir suivi six heures ou plus de mathématiques par semaine.

On constate également qu'avoir suivi peu de sciences et de physique sont des facteurs qui influencent négativement la réussite ; mais les autres modalités de ces variables ne sont pas significatives. Ceci est dû à la multiplication des modalités dans la variable initiale. Un phénomène similaire apparaît avec la variable du diplôme des parents, où avoir un seul parent diplômé du supérieur n'est pas significatif alors que les deux autres modalités le sont, ce qui est a priori absurde.

	Note moy.	% réuss.	
Femme	9.18	44.73	
Homme	9.79	46.15	
17 ans ou moins	10.12	50.85	
18 ans	9.29	44.67	
19 ans	9.19	40.66	
20 ans ou plus	9.19	48.28	
Belge	9.47	45.62	
Autre nationalité	8.95	43.24	
Parle français à la maison	9.52	47.46	*
Parle une autre langue	8.38	29.41	*
Domicilié avec sa famille	9.39	45.80	
Autre domicile pendant l'année	9.33	43.70	
2 parents diplômés du sup.	10.61	58.68	*
1 seul parent diplômé du sup.	8.98	38.57	
Parents non diplômés du sup.	7.83	31.63	*
Primant (ou réorienté)	9.03	39.62	*
Répétant	10.63	65.52	*
Secondaire l'année précédente	9.12	40.36	*
Supérieur l'année précédente	9.96	56.00	*
Participation aux AP	9.70	47.22	
Pas d'activité préparatoire	9.20	44.06	
Secondaire en FWB	9.37	44.67	
Secondaire hors FWB	9.43	48.28	
N'a pas doublé en secondaire	9.77	50.43	*
A doublé en secondaire	7.03	13.79	*
6h ou plus de math	10.32	58.02	*
Moins de 6h de math	7.96	25.93	*
3h ou plus de physique	9.33	45.18	
2h de physique	9.76	49.37	
1h de physique	8.34	32.00	*
8h ou plus de sciences	9.16	44.35	
7h de sciences	9.91	50.83	
6h de sciences	9.53	47.12	
Moins de 6h de sciences	8.43	31.58	*

TABLE 2.3 – Note moyenne sur 20 et pourcentage de réussite pour les étudiants vérifiant chacune des modalités de chaque variable d'entrée. Les taux de réussite mis en évidence avec une étoile signifient que lors d'une régression logistique de la réussite en fonction de cette seule variable, la variable y est significative.

Chapitre 3

Données de travail

Dans la base de données de travail, pour chacune des cinq matières étudiées dans le cours de physique (optique, mécanique, électricité, fluides et imagerie médicale), et pour chacun des niveaux de difficulté (définis comme étant de *bronze*, d'*argent* et d'*or*, par ordre croissant de difficulté), des informations sont données sur l'utilisation qui a été faite du simulateur d'examen par chacun des 555 étudiants¹. Ces données peuvent être décomposées en deux grandes catégories.

D'une part, des données sur l'obtention des *médailles* attribuées aux étudiants pour avoir obtenu une note d'au moins 10/20 à la simulation d'examen. Ces médailles se décomposent par matière et par niveau de difficulté. Pour chacune des cinq matières, il est donc possible d'obtenir trois médailles (de bronze, d'argent et d'or, par ordre croissant de difficulté).

D'autre part, des données détaillant les tentatives effectuées par les étudiants, pour chaque matière et chaque niveau de difficulté, sur le simulateur d'examen. Ces tentatives sont décomposées selon le taux d'omission de l'étudiant lors de la tentative. Si celui-ci est faible (fixé à moins de 20%), on considère la tentative comme *valide* : l'étudiant a tenté de répondre à la plupart des questions. S'il est de 100% (l'étudiant n'a répondu à aucune question), on considère la tentative comme *de visionnage* (car l'étudiant a probablement juste voulu regarder les questions, sans tenter d'y répondre). Entre les deux, on considère la tentative comme *intermédiaire*. De plus, une moyenne du résultat de chaque étudiant *pour les tentatives valides* qu'il a effectuées est donnée. Cette variable n'est donc disponible que si l'étudiant

1. Contrairement au chapitre précédent, on travaille ici avec la base de données complète, qui contient également des lignes correspondant aux étudiants éliminés de la base de données précédente pour cause de données manquantes. En effet, il n'y a aucune donnée manquante dans cette base de données. Ces étudiants seront à nouveau éliminés dans les chapitres suivants, lorsqu'il sera nécessaire de fusionner les deux bases de données.

considéré a effectué au moins une tentative valide pour la matière et le niveau de difficulté correspondants.

3.1 Données d'obtention des médailles

Le graphique présenté en figure 3.1 donne, pour chaque niveau de difficulté et chaque matière, le pourcentage d'étudiants ayant reçu la médaille correspondante (autrement dit, ayant réussi la simulation d'examen au moins une fois). Les matières sont par ailleurs disposées par ordre chronologique de leur développement au cours.

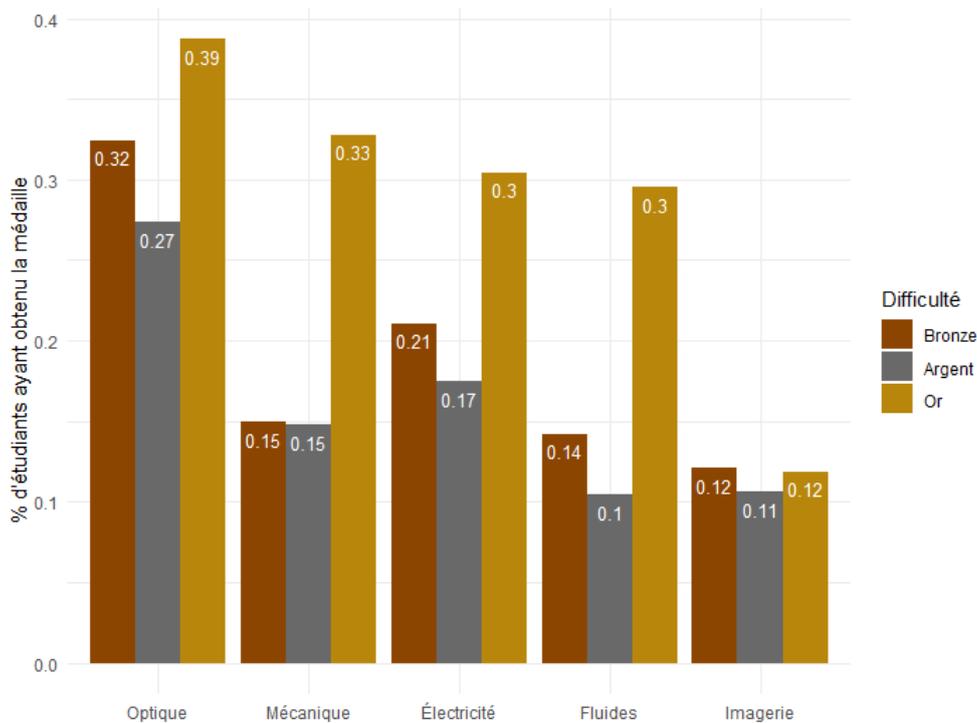


FIGURE 3.1 – Pourcentage d'étudiants ayant obtenu chacune des médailles pour chaque matière proposée dans le simulateur.

Il peut y être observé qu'à part pour le dernier sujet (l'imagerie médicale), le pourcentage d'étudiants ayant obtenu la médaille d'or est chaque fois plus élevée que pour les deux autres médailles. Ceci est très certainement dû au fait que les étudiants obtenaient le droit de présenter un examen blanc (en présentiel) s'ils avaient obtenu ces quatre premières médailles d'or, ce qui a pu les motiver à rechercher celles-là en particulier. Pour le cinquième thème,

qui n'était pas lié à cet avantage, le pourcentage d'obtention des médailles d'or est plus en ligne avec celui des deux autres médailles.

En détail, un peu plus d'un quart des étudiants ont obtenu les médailles d'or pour les quatre premières matières (27%) et on donc pu passer l'épreuve formative en présentiel. Ce pourcentage retombe à 10% si l'on ajoute la cinquième médaille pour l'imagerie médicale. Pour ce qui en est des médailles de bronze et d'argent, après un certain engouement au début de l'année pour la première matière (qui affecte également les médailles d'or), et un léger regain d'intérêt lors du chapitre sur l'électricité, la proportion d'étudiants les ayant réussies reste comprise entre 10% et 15%. Il n'y a que 2,7% des étudiants qui ont récolté les 15 médailles. De plus, près de la moitié (48%) des étudiants n'ont obtenu aucune médaille.

3.1.1 Constitution de groupes d'étudiants selon les médailles obtenues

Il est intéressant d'établir des profils d'étudiants en fonction des médailles qu'ils ont obtenues. Pour ce faire, il est possible d'utiliser une méthode de décomposition en clusters hiérarchique², sur base des données d'obtention ou non de chacune des 15 médailles pour chaque étudiant.

Avant de procéder au regroupement, remarquons, comme cela a déjà été précisé, que près de la moitié des étudiants (268 étudiants, soit 48%) n'ont obtenu aucune médaille. Lors de la composition des profils d'étudiants, il paraît donc opportun de commencer par mettre dans un même groupe, que nous désignerons comme le groupe 0, ces étudiants qui n'ont obtenu aucune médaille. Il peut d'ailleurs être vérifié que procéder comme suit avec les 555 étudiants mène bien à l'obtention de ce groupe, sans que la composition des autres groupes ne soit même modifiée : cette hypothèse préalable ne modifie donc en rien les résultats.

Les données des médailles consistent donc en 15 variables binaires indiquant, pour chaque médaille (des trois difficultés, pour les cinq matières), si l'étudiant a obtenu ou non cette médaille, et ce pour chacun des 287 étudiants qui n'ont pas déjà été regroupés dans le groupe 0.

À nouveau, la mesure de dissimilarité de Jaccard convient à notre base de données. Son côté asymétrique est particulièrement important, en mettant de l'importance sur les médailles obtenues plutôt que celles qui ne l'ont pas été.

2. Härdle & Simar (2015), pp. 393–396

Regroupement hiérarchique. Le regroupement hiérarchique peut à nouveau s'effectuer à l'aide de la fonction `hclust`. À nouveau, on effectue le regroupement sur base de la matrice des mesures de dissimilarité de Jaccard (cette fois, des médailles obtenues ou non), et on utilise à nouveau l'algorithme de Ward, à l'aide de la méthode `ward.D` compatible avec notre mesure de dissimilarité. On obtient alors l'arbre représenté en figure 3.2.

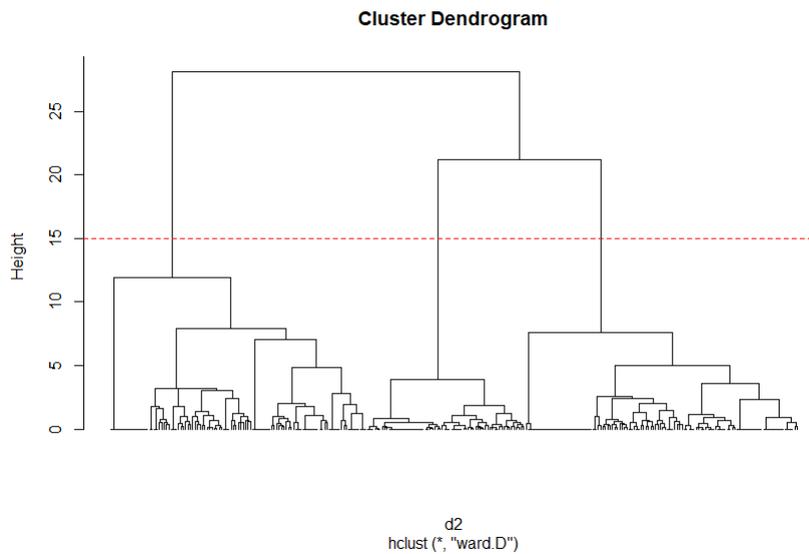


FIGURE 3.2 – Représentation graphique (sous forme d'arbre) du regroupement hiérarchique des étudiants, sur base des médailles obtenues.

Il est à nouveau nécessaire de choisir une hauteur à laquelle couper l'arbre. On choisit, comme cela est souvent recommandé, de couper à la hauteur médiane, c'est-à-dire à la hauteur 15, comme indiqué par le trait rouge. On obtient alors trois groupes (à savoir, quatre groupes au total en comptant le groupe 0 des étudiants n'ayant obtenu aucune médaille) dont les effectifs sont les suivants :

Groupe de médailles	0	1	2	3
Effectif	268	114	66	107

3.1.2 Interprétation des groupes obtenus

Il est possible de donner du sens au regroupement qui a été effectué, en regardant les médailles obtenues dans chaque groupe. Tout d'abord, rappelons que le groupe 0 (qui contient 48% des étudiants) est constitué des étudiants n'ayant obtenu aucune médaille.

Premier groupe. Pour ce qui est du premier groupe (qui contient 20,5% des étudiants), la figure 3.3 présente le pourcentage d'obtention de chaque médaille dans le groupe (à gauche) ainsi que le nombre de médailles obtenues par les membres du groupe selon le nombre de médailles considérées (à droite). Dans ce second graphique, chaque ligne brisée représente un étudiant. Pour chaque médaille considérée une par une, la ligne de l'étudiant évolue en croissant si l'étudiant a obtenu cette médaille, ou de manière constante s'il ne l'a pas obtenue.

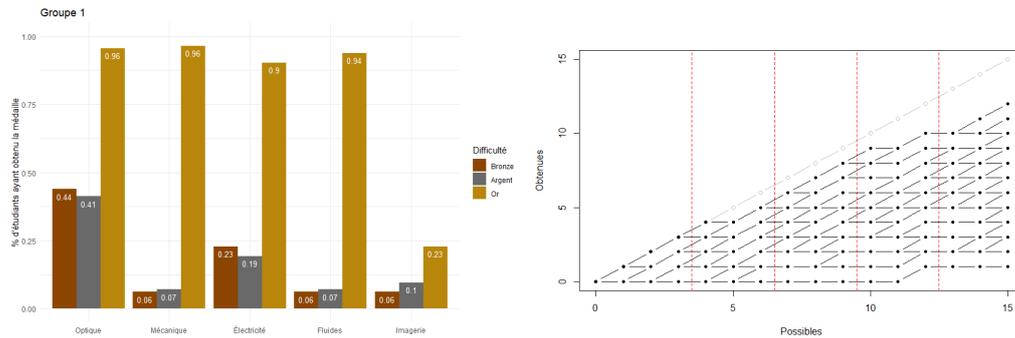


FIGURE 3.3 – Pourcentage d'étudiants du groupe 1 ayant obtenu chaque médaille (à gauche) et évolution de l'obtention des médailles pour chaque étudiant en fonction du nombre de médailles considérées, celles-ci étant considérées dans l'ordre chronologique de l'année (à droite).

On constate que ce groupe est constitué d'étudiants ayant majoritairement obtenu les quatre premières médailles d'or. Les autres médailles, ainsi que celles liées au dernier thème, semblent délaissées par ce groupe en comparaison. De plus, aucun étudiant de ce groupe n'a obtenu toutes les médailles (comme il peut être vu sur la figure 3.3, où aucune ligne d'étudiant ne suit la droite représentant la fonction identité), et ce même avant d'ajouter les trois médailles du dernier thème.

Il pourrait être supposé que les étudiants de ce groupes sont ceux qui ont souhaité avant tout obtenir les médailles d'or, afin de passer la simulation d'examen en présentiel. Notons tout de même que certains d'entre eux n'ont même pas obtenu quatre médailles lorsque l'on considère les médailles en lien avec les quatre premiers thèmes. Ce groupe contient donc quelques étudiants qui ont obtenu peu de médailles au total.

Deuxième groupe. Pour le deuxième groupe (11,9% des étudiants), des représentations similaires sont présentées en figure 3.4.

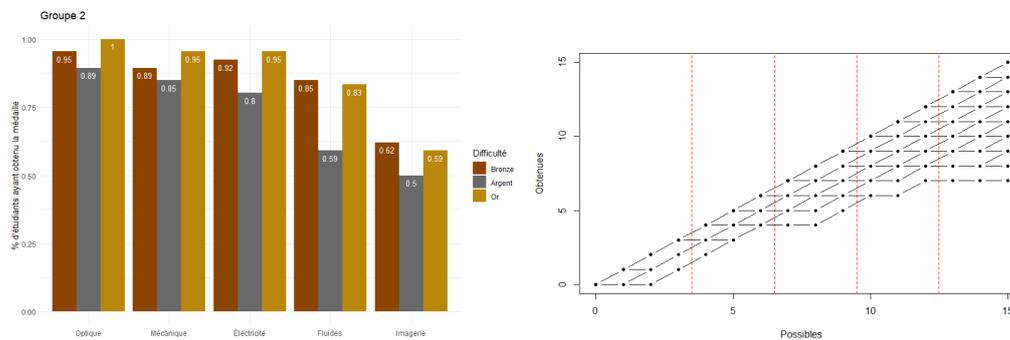


FIGURE 3.4 – Pourcentage d'étudiants du groupe 2 ayant obtenu chaque médaille (à gauche) et évolution de l'obtention des médailles pour chaque étudiant en fonction du nombre de médailles considérées, celles-ci étant considérées dans l'ordre chronologique de l'année (à droite).

On constate, dans ce groupe, des pourcentages d'obtention des trois types de médailles bien plus homogènes, avec des taux d'obtention élevés pour tous les types de médailles. On retrouve dans ce groupe tous les étudiants ayant obtenu toutes les médailles (ceux qui suivent la droite identité dans le graphique de gauche de la figure 3.4) et tous les étudiants ont obtenu un nombre de médailles élevés (le plus faible nombre de médailles obtenues par un élève de ce groupe est 7).

Il s'agit sans doute des étudiants ayant souhaité tirer le plus profit de l'outil, en cherchant à obtenir la plupart des médailles, ou en tout cas en tentant de les obtenir dans l'ordre, ce qui expliquerait l'obtention plus grande des médailles plus faciles que l'on observe pour les trois derniers thèmes.

Troisième groupe. Enfin, les graphiques en lien avec le troisième groupe (19,3% des étudiants) sont présentés en figure 3.5.

Ce groupe contient des étudiants ayant globalement obtenu peu de médailles (même si les profils semblent très divers, et que certains semblent tout de même en avoir obtenu un certain nombre), celles-ci étant assez dispersées, sans stratégie apparente.

On observe que le taux d'obtention des médailles est comparativement plus élevé pour le premier chapitre, ce qui pourrait montrer un intérêt initial pour l'outil, en tout cas pour une partie des étudiants de ce groupe, qui ne s'est pas prolongé.

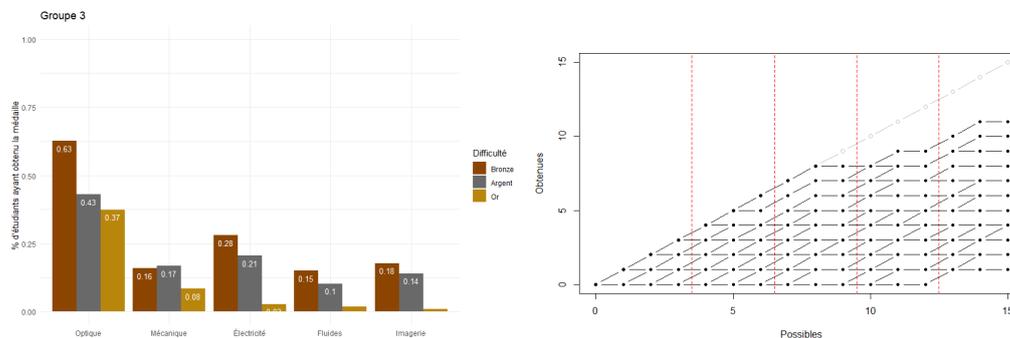


FIGURE 3.5 – Pourcentage d’étudiants du groupe 3 ayant obtenu chaque médaille (à gauche) et évolution de l’obtention des médailles pour chaque étudiant en fonction du nombre de médailles considérées, celles-ci étant considérées dans l’ordre chronologique de l’année (à droite).

3.1.3 Notes obtenues dans chaque groupe

La figure 3.6 permet de se donner une idée des notes obtenues à l’examen par les membres de chacun des groupes.

Ainsi, les membres des groupes 1 et 2 ont des notes assez similaires et tendent à réussir, tandis que le groupe 3 a des résultats autour de la moyenne (la plupart des étudiants de ce groupe ont tout de même échoué). Le groupe 0 contient majoritairement des étudiants qui ont raté l’examen, même si quelques étudiants isolés le réussissent.

Remarquons déjà que tant les membres du groupe 1 que ceux du groupe 2, ont majoritairement pu réaliser l’examen blanc en présentiel, ce qui est un biais qu’il ne sera pas possible d’éliminer. De plus, il est toujours impossible à ce stade de déterminer si l’utilisation qui a été faite de l’outil par les étudiants de ces deux groupes (qui ont mieux réussi) est responsable de leur réussite, ou si l’outil tend simplement à attirer davantage les étudiants qui auraient de toute manière réussi.

3.1.4 Stabilité vis-à-vis du dernier chapitre

Comme on a pu l’observer dans les comparaisons des différents groupes, les médailles obtenues pour le dernier chapitre vu au cours, l’imagerie médicale, semblent se comporter très différemment que pour les autres thèmes. Cela peut être dû tant au fait que le chapitre a été vu à la toute fin de l’année qu’au fait que sa médaille d’or n’était pas requise pour passer l’examen blanc en présentiel.

Il est dès lors intéressant de déterminer l’impact qu’ont eues les médailles

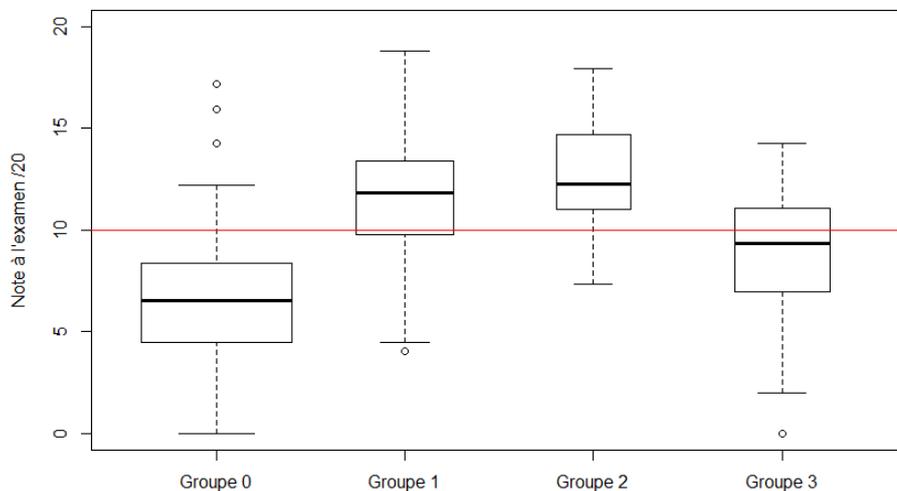


FIGURE 3.6 – Boîte à moustache de la note obtenue à l’examen par les étudiants de chaque groupe constitué. L’épaisseur des boîtes est proportionnelle à la racine carrée du nombre d’observations dans le groupe considéré.

obtenues pour ce chapitre sur la composition des différents groupes.

En appliquant la même méthode que précédemment afin de constituer les groupes (à nouveau, on place préalablement dans un groupe 0 les étudiants n’ayant obtenu aucune médaille ; il y en a 271 à présent), on obtient l’arbre de regroupement hiérarchique présenté en figure 3.7.

En coupant à nouveau l’arbre à la hauteur 15, on obtient à nouveau 3 groupes (auxquels il faut ajouter notre groupe 0), dont les effectifs sont les suivants :

Groupe de médailles	0	1	2	3
Effectif	271	112	68	104

On constate immédiatement que les effectifs des quatre groupes sont très similaires à ceux constitués avec le cinquième chapitre compris. Il reste donc à savoir si les membres des différents groupes n’ont pas été mélangés. Pour vérifier ceci, étudions la table de contingence des étudiants en fonction des deux groupements, avec ou sans le cinquième chapitre utilisé pour construire les groupes. Cette table de contingence est présentée en table 3.1.

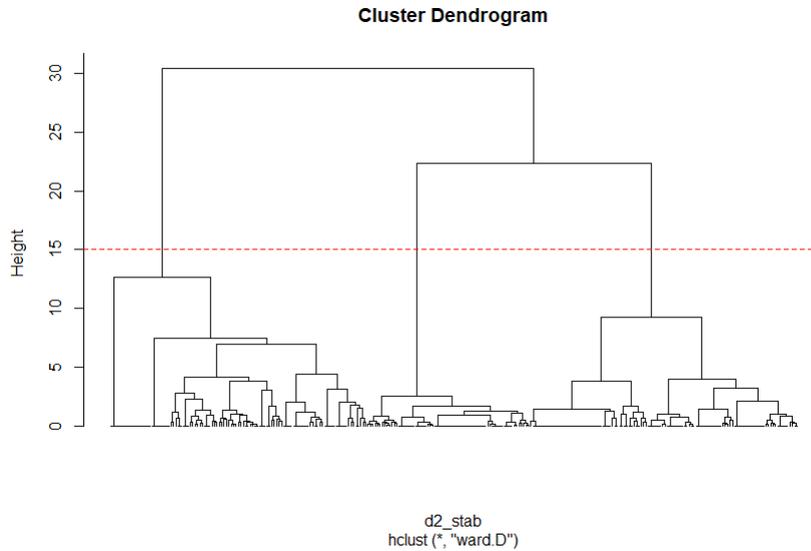


FIGURE 3.7 – Représentation graphique (sous forme d’arbre) du regroupement hiérarchique des étudiants, sur base des médailles obtenues lors des quatre premiers chapitres.

Groupe modifié	Groupe initial			
	1	2	3	4
0	0	268	3	0
1	108	0	3	1
2	5	0	0	63
3	1	0	101	2

TABLE 3.1 – Table de contingence des groupements d’étudiants selon leurs médailles obtenues, comparés aux groupes « modifiés » construits sans les données en lien avec le chapitre 5.

On constate, au vu de la table 3.1, que les différences selon que l’on a construit les groupes avec ou sans les données en lien avec le dernier chapitre, sont minimales. On en conclut que l’impact du cinquième chapitre sur les groupes constitués sur base des médailles est minime, et que ce regroupement d’étudiants est donc parfaitement stable vis-à-vis de la présence du cinquième chapitre.

3.2 Données concernant les tentatives

Les données sur la fréquence d'utilisation de l'outil consistent en le nombre de tentatives de visionnage, intermédiaires et valides de chaque étudiant, pour chaque matière et chaque niveau de difficulté.

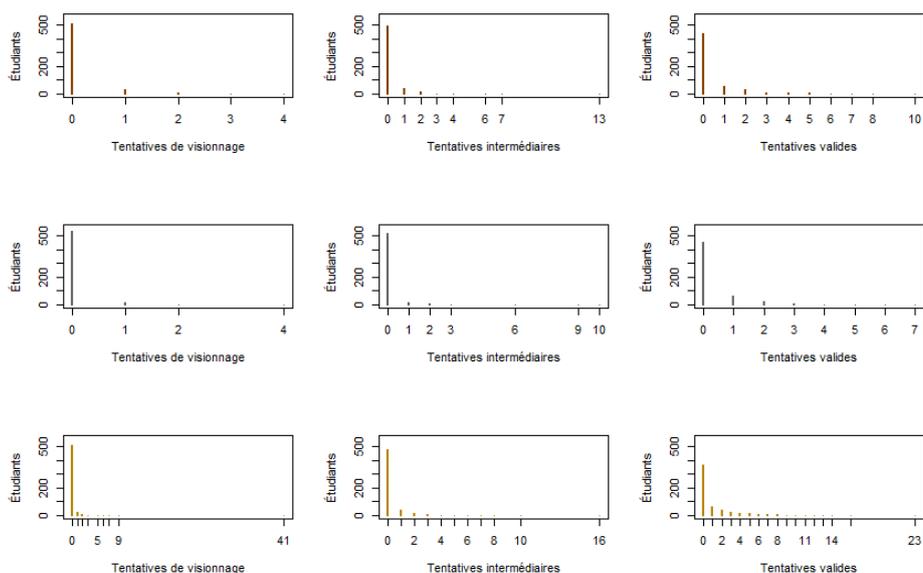


FIGURE 3.8 – Nombre d'étudiants ayant effectué chaque nombre de tentatives pour le chapitre de mécanique. Les graphiques de gauche concernent les tentatives de visionnage, ceux du milieu celles intermédiaires et ceux de droite les tentatives valides. De plus, la première ligne concerne la médaille de bronze, celle du milieu la médaille d'argent, et celle du bas, la médaille d'or.

La figure 3.8 permet de se représenter la distribution du nombre de tentatives de chaque type, en s'intéressant au cas particulier du chapitre sur la mécanique.

On y observe que les tentatives de visionnage sont peu fréquentes : peu d'étudiants sont concernés et lorsqu'ils le sont, ils ont effectué une ou deux tentatives. La situation est semblable pour les tentatives intermédiaires. Dans le cas des tentatives valides, on obtient une distribution un peu plus variée, en particulier pour la médaille d'or, ce qui pourrait être le signe de multiples essais effectués afin de l'obtenir, ou d'un choix de l'étudiant pour s'entraîner étant donné qu'il s'agit du niveau de difficulté le plus proche de celui de l'examen.

Il semblerait donc que seules les tentatives valides, et principalement pour la médaille d'or semblent avoir une diversité de valeurs suffisante et un nombre d'étudiants concernés suffisants pour être une donnée intéressante à analyser. Celles-ci sont très fortement concentrées sur les petites valeurs (en particulier 0 et 1) et la moyenne sur les tentatives valides, disponible lorsqu'au moins une d'entre elle a été effectuée, pourrait apporter un éclairage sur la raison pour laquelle l'étudiant a effectué un certain nombre de tentatives.

Considérant la difficulté de synthétiser ces données, il n'a donc été choisi de les exploiter, en se contentant des données de médailles qui témoignent suffisamment bien de l'investissement de l'étudiant dans l'outil pour pouvoir poursuivre l'étude.

3.3 Moyenne sur les tentatives valides

En plus du nombre de tentatives de chaque type, la moyenne des résultats sur les tentatives valides a été récoltée pour chaque matière et pour chaque niveau de difficulté, pour tous les étudiants ayant présenté au moins une tentative valide.

Cette information est assez peu exploitable en réalité, car elle manque à chaque fois qu'un étudiant n'a pas présenté de tentative valide pour une matière et un niveau donnés. Comme moins de 3% des étudiants ont obtenu toutes les médailles pour toutes les matières, il n'est donc pas étonnant que presque tous les étudiants ont cette valeur manquante pour au moins une matière et un niveau.

L'intérêt principal de ces variables pourrait être de discriminer, parmi les étudiants ayant obtenu une médaille précise, ceux qui l'ont obtenue facilement, de ceux qui auraient eu besoin de nombreuses tentatives infructueuses avant de l'obtenir (ces derniers auraient alors très certainement une moyenne inférieure à 10). Ainsi, ces variables pourraient apporter un certain éclairage lors de l'interprétation du nombre de tentatives valides effectuées, entre une forme d'acharnement afin de décrocher la médaille et une forme de dévouement à vouloir tirer le maximum de l'outil.

La figure 3.9 permet de se donner une idée de la dispersion des moyennes pour les niveaux or. On observe que leur médiane est toujours supérieure à la moitié, mais que la dispersion est très grande, ce qui témoigne de résultats très différents d'un étudiant à l'autre.

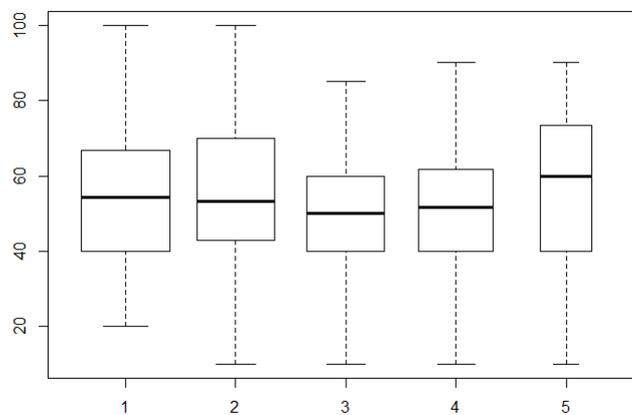


FIGURE 3.9 – Dispersion des moyennes sur les tentatives valides pour le niveau or de chacune des matières dans l'ordre chronologique de l'année (de 1 : optique à 5 : imagerie médicale). Seuls les étudiants ayant effectué une tentative valide pour la matière concernée apparaissent, ce qui justifie que le nombre d'étudiants repris est différent pour chaque matière. L'épaisseur de chaque boîte est proportionnelle à la racine carrée du nombre d'observations sur lesquelles elle est basée.

Chapitre 4

Liens entre les données d'entrée et de travail

À présent que les deux bases de données ont été présentées, il semble naturel de vouloir déterminer si les données de ces deux bases de données (en particulier, les données d'entrée et de médailles) sont liées d'une certaine manière ou non.

Cette question est d'autant plus centrale, qu'il est nécessaire afin de répondre à la principale question de recherche de déterminer si le lien éventuel entre les données d'obtention des médailles et la note à l'examen qui sera étudiée dans la section suivante est biaisé par les données d'entrées.

Vu que des observations (étudiants) ont été retirées de la base de données d'entrée en raison de valeurs manquantes pour certaines variables, ce qui n'est le cas d'aucune observation dans la base de données de travail, cette seconde base de données contient plus d'observations et il sera nécessaire lors de la jointure d'ignorer ces observations pour lesquelles les données d'entrée n'ont pas été utilisées.

4.1 Comparaison des groupes d'entrée et des groupes de médailles

Étant donné que quatre groupes ont été constitués pour chacune des bases de données, commençons par établir s'il existe un lien entre l'appartenance à un groupe d'entrée et celle à un groupe de médailles.

La table de contingence comparant les deux types de groupements est présentée en table 4.1.

Le test du chi-carré de Pearson rejette l'indépendance des groupes d'entrée par rapport aux groupes de médailles avec une p-valeur de 0,0064. Il y

Gr. méd.	Groupe d'entrée			
	1	2	3	4
0	19	49	64	18
1	11	38	35	14
2	10	10	17	20
3	16	28	35	21

TABLE 4.1 – Table de contingence comparant les groupes d'entrée et les groupes de médailles.

aurait donc bien un lien entre le groupe d'entrée d'un étudiant et son groupe de médaille.

Afin de déterminer quelles cellules ont déterminé le rejet de l'indépendance dans le test du chi-carré de Pearson, il est possible de regarder les résidus (élevés au carré) du test. Ces résidus sont définis, pour chaque cellule de la table 4.1, par la différence entre la valeur de la cellule et la valeur attendue pour cette cellule (afin qu'on ait l'indépendance), standardisée en divisant par la racine carré de la valeur attendue. La somme des résidus de chaque colonne au carré donne la statistique chi-carré de Pearson, qui est comparée au quantile 0,95 de la distribution χ^2 (à 9 degrés de liberté dans notre cas), dont la valeur approximative est 16,92. Si la somme des résidus au carré est plus grande que cette valeur, le test conclut au rejet de l'indépendance des groupes.

Gr. méd.	Groupe d'entrée			
	1	2	3	4
0	0.15	0.16	1.17	3.02
1	0.48	1.99	0.06	0.76
2	0.57	3.28	0.85	9.21
3	0.34	0.27	0.14	0.49

TABLE 4.2 – Résidus du chi-carré de Pearson élevés au carré, pour chaque cellule de la table de contingence présentée en table 4.1.

La table 4.2 fournit les résidus pour chaque cellule de la table de contingence. La somme des résidus au carré (la statistique de test) est de 22,93 (à comparer au quantile 0,95 de la χ^2 , 16,92).

On constate que la surabondance¹ d'étudiants du groupe 2 de médailles (ayant recherché presque toutes les médailles) dans le groupe 4 d'entrée

1. Bien que cette distinction soit perdue dans cette table où les résidus sont affichés au carré, la valeur du résidu était en effet positive, ce qui indique bien un surplus d'étudiant dans cette cellule. Au contraire, le résidu de la cellule en lien avec les étudiants dans le

(constitué majoritairement de répétants) suffit à elle seule à déterminer le résultat du test.

Ce surplus d'étudiants du groupe de médailles 2 dans le groupe d'entrée 4 se compense par un manque d'étudiants du groupe de médailles 2 dans le groupe d'entrée 2 et par un manque d'étudiants du groupe d'entrée 4 dans le groupe de médailles 0 par rapport à ce qui serait attendu dans les cas où les groupes étaient tous homogènes.

4.2 Impact des données d'entrée sur le groupe de médailles

La section précédente permet déjà de conclure à un certain lien entre les deux bases de données, à savoir que le groupe d'entrée 4 (caractérisé par la présence d'étudiants ayant déjà fait une année dans le supérieur) tend à se retrouver dans le groupe de médailles 2 (caractérisé par un intérêt pour la récolte du plus grand nombre de médailles). Autrement dit, les étudiants ayant déjà fait une année dans le supérieur sont plus nombreux à rechercher toutes les médailles par rapport aux autres étudiants, sortant du secondaire.

L'objectif de cette section est d'affiner cette recherche des liens entre les bases de données en déterminant, parmi les variables d'entrée, quels facteurs sont en lien avec la manière qu'ont eu les étudiants d'utiliser l'outil. Pour ce faire, de manière similaire à Dehon et al. (2019), une régression ordinale va être utilisée.

Avant de pouvoir appliquer le modèle, il est cependant recommandé d'opérer une sélection des variables explicatives qui paraissent les plus utiles dans le modèle.

4.2.1 Sélection des variables explicatives

Bien qu'il soit techniquement possible d'introduire autant de variables que souhaité dans le modèle, chaque variable ajoutée augmente la complexité du modèle en augmentant le nombre de degrés de liberté et rend plus difficile l'interprétation du modèle. Il est donc en pratique suggéré d'investiguer les variables potentiellement utiles afin d'éliminer toute variable qui ne paraisse pas pertinente dans le modèle².

groupe de médailles 2 et le groupe d'entrée 2 est négative et il y a donc moins d'étudiants dans cette cellule qu'attendu; il en va de même pour la cellule des étudiants du groupe d'entrée 4 et du groupe de médailles 0.

2. Comme suggéré par Neter et al. (1989), p. 418.

De plus, il faut veiller à éviter la multicolinéarité dans le modèle. En effet, si des variables explicatives sont corrélées ou associées, ou si une variable explicative est combinaison linéaire d'autres variables explicatives, le coefficient de régression des variables explicatives varie quand d'autres variables explicatives sont introduites dans le modèle ; ces coefficients de régression ne reflètent alors qu'une partie de l'effet que la variable explicative a sur la variable dépendante³. Remarquons que dans le cas de variables qualitatives, ce qui est le cas ici, à chaque variable qualitative à m modalités correspond $m - 1$ variables binaires⁴, et donc $m - 1$ coefficients à estimer.

Détail des variables sélectionnées

Il a tout d'abord été choisi de limiter le nombre de modalités des variables donnant le nombre d'heures de sciences et de physique suivies par les étudiants en secondaire. En effet, comme vu dans la table 2.3 qui comparait la note des étudiants selon les variable d'entrée, où seule la variable indiquant si les étudiants étaient en science faible (ou physique faible) semblait importante par rapport aux autres, la multiplication des modalités indiquant un nombre plus important d'heures semble créer des effets de compensation tant et si bien qu'aucune de ces modalités ne se démarque, sans que la pertinence de ce découpage ne soit claire⁵, en particulier quant à mettre ces variables en relation avec l'investissement de l'étudiant. Il s'ensuit que malgré des effectifs plus faibles dans les modalités science faible et physique faible (57 et 50 étudiants respectivement), les découpages sciences faibles versus sciences fortes et physique faible versus physique forte semblent receler la majorité de l'information présente dans la variable.

Pour des raisons similaires, la variable du plus haut diplôme obtenu par les parents a été réduite à deux modalités, selon qu'au moins un parent est diplômé du supérieur ou non. En effet, il ne semblait pas y avoir de différence particulière entre les étudiants ayant un ou deux parents diplômés du supérieur, et il s'agit à nouveau de modalités qui, si elles sont présentées comme exclusives, risquent de masquer certains effets de cette variable sur la variable dépendante.

Afin d'éviter la corrélation, quelques variables ont été éliminées à cause de leur corrélation évidente (les corrélations calculées peuvent être consultées

3. Neter et al. (1989), p. 277

4. On ne conserve donc pas ici la dernière variable binaire redondante, qui causerait bien évidemment de la multicolinéarité.

5. Il avait d'ailleurs été effectué de manière tout à fait subjective lors du premier traitement des données. Or, les observations faites sur la base de données, comme expliqué ici, semblent remettre en cause la pertinence de ce découpage.

en Annexe 2) :

- Étant donné la manière dont les choix d'options sont présentés en secondaire, où le cours de physique est une partie de l'option sciences, les variables donnant le nombre d'heures de physique et de sciences suivies en secondaire sont évidemment corrélées. Cet effet est empiré par la simplification qui a été effectuée plus haut, où l'on perd l'éventuelle distinction du choix d'avoir pris des laboratoires (nombre d'heures de sciences élevé). La variable du nombre d'heures de physique a donc été retirée.
- De plus, il semble évident que la variable indiquant si l'étudiant sort de secondaire ou a déjà fait une année dans le supérieur est corrélée avec le fait d'être primant ou répétant. En fait, seul le fait d'avoir regroupé primant et réorienté distinguait ces deux variables. La variable indiquant si l'étudiant est primant ou répétant a été retirée.
- Ensuite, vu la manière dont elle avait été réécrite dès le début, la variable indiquant si l'étudiant a fait ses secondaires en Fédération Wallonie-Bruxelles ou non est corrélée avec la nationalité de l'étudiant. Cette variable a donc été retirée en faveur de la nationalité.
- Enfin, il a été décidé de retirer également la variable de l'âge de l'étudiant car la plupart de l'information de cette variable peut être retrouvée dans d'autres variables, en particulier le fait d'avoir doublé ou non et le fait de sortir du secondaire ou d'être déjà dans le supérieur.

Grâce à cette première sélection de variables, une étude des VIF⁶ des variables ne permet pas de déterminer d'effet de multicolinéarité trop important dans les variables restantes.

Après cette sélection, il reste 10 variables qui possèdent toutes deux modalités. Dans la liste qui suit, la modalité de référence, autrement dit celle à laquelle l'autre sera comparée, est en gras :

- Genre : **masculin** ou féminin ;
- Nationalité : **étranger** ou belge ;
- Langue parlée à la maison : **pas le français** ou le français ;
- Domicile pendant l'année académique : **sans famille** ou avec la famille ;

6. *Variance Inflation Factor*, méthode de détection de la multicolinéarité présentée par exemple dans Neter et al. (1989), p. 391–393.

- Études suivies l'année précédente : **secondaire** ou supérieur ;
- Mathématiques suivies en secondaire : **faibles** ou fortes ;
- Sciences suivies en secondaire : **faibles** ou fortes ;
- Fait d'avoir doublé en secondaire : **a doublé** ou n'a pas doublé ;
- Plus haut diplôme des parents : **pas du supérieur** ou du supérieur ;
- Activités préparatoires : **n'a pas participé** ou a participé.

Ces dix variables binaires seront les variables explicatives du modèle de régression ordinale qui suit.

4.2.2 Régression ordinale

Le modèle de régression ordinale est un modèle linéaire généralisé, pour lequel la variable dépendante est ordinale. En particulier, ce modèle va ici permettre de déterminer les effets linéaires des variables d'entrée (qui seront les variables explicatives du modèle) sur le groupe de médailles, considéré comme une variable ordinale.

En effet, il a été choisi d'ordonner les quatre groupes de médailles du groupe des étudiants les moins investis dans l'outil à celui des étudiants qui s'y étaient le plus investi de manière à obtenir un modèle plus informatif.

L'ordre qui a été choisi est évidemment un peu subjectif mais semble justifié. Dans l'ordre croissant, le groupe le moins investi est le groupe 0 des étudiants qui n'ont pas utilisé l'outil. Suit le groupe 3, des étudiants qui ont utilisé l'outil pour obtenir une ou l'autre médaille sans aucune régularité. Ensuite, vient le groupe 1 des étudiants ayant recherché la plupart des médailles d'or, mais pas vraiment les autres types de médailles. Enfin, le groupe des étudiants s'étant le plus investi dans l'outil est le groupe 2 des étudiants ayant recherché le plus de médailles.

Description du modèle

Par définition, le modèle de régression ordinale modélise via une fonction de lien g la probabilité qu'une variable ordinale Y à K niveaux soit plus petite ou égale à un niveau k compris entre 1 et $K - 1$ en fonction des p variables explicatives x_1, \dots, x_p (regroupées dans le vecteur x) afin que

$$g(\mathbb{P}[Y \leq k | x]) = \theta_k + \beta_1 x_1 + \dots + \beta_p x_p$$

où les ordonnées à l'origine θ_k pour chaque niveau $k \in \{1, \dots, K-1\}$ et les coefficients β_1, \dots, β_p sont les paramètres du modèle à estimer.

La fonction de lien qui va être utilisée dans le modèle de régression ordinaire est la fonction logistique, définie par

$$\text{logit}(x) = \ln\left(\frac{x}{1-x}\right),$$

qui est une fonction de lien très souvent choisie pour ses bonnes propriétés et pour la qualité de son interprétabilité.

On suppose alors, comme Donneau (2013), que la variable ordinaire Y sous-tend une variable latente continue Y^* , qui serait telle que Y soit un découpage en K classes de Y^* , et dont les bornes sont $\theta_1 < \dots < \theta_{K-1}$. On note également $\theta_0 = -\infty$ et $\theta_K = +\infty$.

Comme Y^* est une variable continue, il est possible de décrire le modèle linéaire classique avec les mêmes variables explicatives : $Y^* = x^T\beta + \epsilon$, où x^T est le vecteur des valeurs des variables explicatives transposée et ϵ un terme d'erreur dont on note la fonction de répartition F .

On a alors, comme présenté par Donneau (2013), pour tout niveau $k \in \{1, \dots, K-1\}$ de la variable ordinaire Y ,

$$\begin{aligned} \mathbb{P}[Y \leq k] &= \mathbb{P}[Y^* \leq \theta_k | x] \\ &= \mathbb{P}[x^T\beta + \epsilon \leq \theta_k | x] \\ &= \mathbb{P}[\epsilon \leq \theta_k - x^T\beta] \\ &= F(\theta_k - x^T\beta). \end{aligned}$$

Notons qu'il est également possible de paramétriser le modèle avec un signe $+$ avant le terme en $x^T\beta$, à condition de changer les signes de β . Le modèle est alors identique, mais l'interprétation serait différente. Toutefois, il a été choisi de conserver ici le signe $-$.

Si la réciproque de F , F^{-1} , est la fonction de lien, on obtient

$$F^{-1}(\mathbb{P}[Y \leq k | x]) = \theta_k - x^T\beta. \quad (4.1)$$

Dans le cas logistique qui sera à présent le seul considéré, on choisit comme F^{-1} la fonction logit dont la réciproque est la fonction F définie par

$$F(\epsilon) = \frac{\epsilon}{1 + \exp(\epsilon)}.$$

En particulier, la probabilité que la variable ordinaire Y soit de niveau inférieur ou égal à k (avec $k \in \{1, \dots, K-1\}$) si les variables explicatives ont

pour vecteur de modalités a est donnée dans le cas logistique par

$$\begin{aligned}\mathbb{P}[Y \leq k | x = a] &= F(\theta_k - a^T \beta) \\ &= \frac{\theta_k - a^T \beta}{1 + \exp(\theta_k - a^T \beta)}.\end{aligned}$$

Le modèle ainsi défini est appelé modèle des cotes proportionnelle, et n'est justifié que si l'hypothèse dite des cotes proportionnelles peut être raisonnablement admise.

Cotes et hypothèse des cotes proportionnelles

La cote (*odds* en anglais) d'un événement étant définie comme le quotient de la probabilité que l'événement survienne par la probabilité qu'il ne survienne pas, la cote du fait qu'une variable ordinale soit de niveau inférieur ou égal à k pour $k \in \{1, \dots, K - 1\}$ se détermine par la formule

$$\begin{aligned}\text{odds}(Y \leq k | a) &= \frac{\mathbb{P}(Y \leq k | a)}{1 - \mathbb{P}(Y \leq k | a)} \\ &= \exp(\theta_k - a^T \beta)\end{aligned}$$

où a est un vecteur de valeurs possibles pour le vecteur des variables explicatives x .

Le rapport de cotes cumulé compare les cotes d'un événement pour plusieurs valeurs que peuvent prendre les variables explicatives. Autrement dit, on a en toute généralité

$$\begin{aligned}\text{OR}_k(a, b) &= \frac{\text{odds}(Y \leq k | a)}{\text{odds}(Y \leq k | b)} \\ &= \exp[\theta_k - a^T \beta - (\theta_k - b^T \beta)] \\ &= \exp[(b - a)^T \beta],\end{aligned}$$

où a et b sont des vecteurs de valeurs possibles pour le vecteur de variables explicatives x , pour chacun des $K - 1$ niveaux si la variable ordinale Y est à K niveaux.

On constate donc que pour ce modèle des cotes proportionnelles, le rapport de cotes cumulé est indépendant de k . Un seul rapport de cotes cumulé suffit donc par variable, ce qui simplifie l'interprétation du modèle. Il est toutefois nécessaire, comme précisé précédemment, que l'hypothèse des cotes proportionnelles soit raisonnable pour les données à modéliser pour pouvoir utiliser ce modèle.

L'interprétation habituelle d'un rapport de cotes cumulé de niveau k entre a et b est que la probabilité de passer du niveau k au niveau $k + 1$ de la variable ordinale est multipliée par ce rapport de cote lorsque les variables explicatives valent b plutôt que a . L'hypothèse des cotes proportionnelles est donc vérifiée s'il est raisonnable de considérer que cette probabilité est identique peu importe le niveau $k \in \{1, \dots, K - 1\}$ de la variable ordinale considérée.

Dans le cas où l'hypothèse n'est pas vérifiée, il est alors nécessaire de déterminer tous les rapports de cotes cumulés pour chaque niveau (sauf le dernier), ce qui complique l'interprétation. Cependant, il est aussi possible, à condition de déterminer les variables à cause desquelles l'hypothèse n'est pas vérifiée, de ne considérer tous les niveaux que pour ces variables, et de considérer l'hypothèse des cotes proportionnelles pour les autres.

Cette variante du modèle est parfois appelé le modèle de régression ordinal avec cotes partiellement proportionnelles et ajoute dans l'équation (4.1) un terme dont les coefficients dépendent du niveau k pour chaque variable pour laquelle l'hypothèse n'est pas vérifiée. Autrement dit, si une seule variable \tilde{x} est concernée, le modèle des cotes partiellement proportionnelles est défini par ⁷

$$\text{logit}(\mathbb{P}[Y \leq k | x]) = \theta_k - x^T \beta + \tilde{x} \tilde{\beta}_k, \quad (4.2)$$

où x est le ici vecteurs des variables explicatives différentes de \tilde{x} .

Application du modèle et résultats

Lorsque l'on applique ce modèle à nos données, la variable ordinale est la version ordonnée du groupe de médailles (qui est à quatre niveaux) et les variables explicatives sont celles qui ont été sélectionnées parmi les données d'entrée.

Commençons par vérifier que l'hypothèse des cotes proportionnelles est vérifiée pour ces données. Pour ce faire, il est possible d'effectuer un test du rapport de vraisemblance qui pour chaque variable, compare le modèle des cotes proportionnelles avec le modèle des cotes partiellement proportionnelles où la variable sélectionnée est considérée sans l'hypothèse. Si le test retourne une différence significative entre les deux modèles, alors il y a des raisons d'envisager un modèle des cotes partiellement proportionnelles, où cette variables se verrait attribuer un rapport de cotes cumulé pour chaque niveau.

7. La paramétrisation du modèle proposée ici est celle implémentée dans R pour la fonction `clm` du package `ordinal`, ainsi que cela est décrit dans la documentation de R (Christensen, 2018).

En l'occurrence, les différents rapports de vraisemblance ont été effectués à l'aide de la fonction `nominal_test` de la bibliothèque `ordinal` de R, ce qui donne le résultat présenté en table 4.3.

	Df	LRT	Pr(>Chi)	
Genre_	2	0.72	0.6979	
Nat_	2	6.24	0.0442	*
Lang_	2	0.97	0.6168	
Domic_	2	0.02	0.9890	
AnPrec_	2	11.33	0.0035	**
Math_	2	3.51	0.1728	
Sc_	2	3.06	0.2169	
DoubleSec_	2	0.36	0.8355	
DiplomPar_	2	3.60	0.1652	
AP_	2	2.80	0.2470	

TABLE 4.3 – Résultat du test de rapports de vraisemblance proposé par la fonction `nominal_test` du package `ordinal` de R.

On constate donc que deux variables sont susceptibles de ne pas vérifier l'hypothèse des cotes proportionnelles, ce pourquoi il a été choisi de considérer le modèle des cotes partiellement proportionnelles, avec `AnPrec_`, la variable qui détermine si l'étudiant était en secondaire ou dans le supérieur l'année précédente, considérée comme ayant une cote différente pour chaque niveau de groupes de médailles.

Modifier ainsi le modèle ne rend plus la variable `Nat_` significative lors du test de vraisemblance. Il n'est donc pas nécessaire de considérer également les cotes de cette variable sur tous les niveaux de groupes de médailles.

Résultats

Le modèle qui a été utilisé est donc le modèle des cotes partiellement proportionnelles avec toutes les variables considérées comme ayant des cotes proportionnelles sauf `AnPrec_`.

La fonction `clm` de la bibliothèque `ordinal` de R permet d'utiliser ce modèle, en lui indiquant en formule la variable ordinaire dépendante (les groupes de médailles ordonnés comme indiqué précédemment) ainsi que les variables explicatives pour lesquelles l'hypothèse des cotes proportionnelles est valable ; de plus, le paramètre `nominal` de la fonction permet d'ajouter des variables qui doivent être considérées pour tous les niveaux de la variable dépendante. On y met donc `AnPrec_`.

Le résultat obtenu est le suivant :

	Estimate	Std. Error	z value	Pr(> z)
Genre_F	0.04684	0.20096	0.233	0.81571
Nat_Bel	0.13292	0.29760	0.447	0.65514
Lang_FR	0.81928	0.30469	2.689	0.00717 **
Domic_Famille	0.45578	0.23788	1.916	0.05536 .
Math_Fort	0.44924	0.19764	2.273	0.02303 *
Sc_Fort	0.65170	0.27723	2.351	0.01874 *
DoubleSec_PasDouble	0.81641	0.29597	2.758	0.00581 **
DiplomPar_Sup	0.63769	0.22466	2.838	0.00453 **
AP_Participe	-0.01708	0.21335	-0.080	0.93618

Threshold coefficients:

	Estimate	Std. Error	z value
0 3.(Intercept)	2.9140	0.5578	5.224
3 1.(Intercept)	3.9165	0.5707	6.862
1 2.(Intercept)	5.7128	0.6084	9.389
0 3.AnPrec_Sup	-1.1086	0.2730	-4.061
3 1.AnPrec_Sup	-0.6893	0.2540	-2.713
1 2.AnPrec_Sup	-1.5158	0.3189	-4.753

On constate que la langue parlée à la maison, les choix d'option en secondaire (math et sciences), le fait d'avoir doublé en secondaire et le diplôme des parents ont une influence sur l'investissement dans l'outil de simulation d'examen.

Les rapports des cotes cumulés peuvent également être utilisés pour interpréter les résultats. En effet, le rapport de cotes cumulé lorsqu'une variable binaire x passe de 0 (modalité de base) à 1 (modalité alternative) est égal à l'exponentielle du coefficient lié à cette variable :

$$\begin{aligned}
 \text{OR}_k(0, 1) &= \frac{\text{odds}(Y \leq k | x = 0)}{\text{odds}(Y \leq k | x = 1)} \\
 &= \frac{\exp(\theta_k - \beta \times 0)}{\exp(\theta_k - \beta \times 1)} \\
 &= \exp(\theta_k + \beta - \theta_k) = \exp(\beta)
 \end{aligned}$$

pour tout niveau k possible de la variable ordinaire, lorsque l'on considère l'hypothèse des cotes proportionnelles.

Le rapport de cotes cumulé peut donc être obtenu simplement en exponentiant les coefficients des différentes variables explicatives pour lesquelles l'hypothèse des risques proportionnelles est vraie. Il peut alors être interprété comme étant le facteur multipliant la chance de passer d'une catégorie

de médailles à la suivante si l'on changeait de modalité pour cette variable explicative, toutes autres variables explicatives restant égales.

	Gr0	Gr3	Gr1	Gr2	OR
Homme	35.38	26.15	26.15	12.31	-
Femme	37.82	24.00	23.27	14.91	1.05
Pas Belge	37.84	32.43	14.86	14.86	-
Belge	36.86	22.96	26.28	13.90	1.14
Parle une autre langue	54.90	25.49	11.76	7.84	-
Parle français à la maison	34.46	24.58	25.99	14.97	2.27 *
Pas domicilié avec sa famille	38.66	25.21	23.53	12.61	-
Domicilié avec sa famille	36.36	24.48	24.48	14.69	1.58
Secondaire l'année précédente	41.79	22.50	26.43	9.29	-
Supérieur l'année précédente	26.40	29.60	19.20	24.80	-
Moins de 6h de math	42.59	29.63	19.14	8.64	-
6h ou plus de math	33.33	21.40	27.57	17.70	1.57 *
Moins de 6h de sciences	45.61	31.58	15.79	7.02	-
Plus de 6h de sciences	35.63	23.56	25.57	15.23	1.92 *
A doublé en secondaire	55.17	25.86	13.79	5.17	-
N'a pas doublé en secondaire	34.01	24.50	25.94	15.56	2.26 *
Pas de parent diplômé du sup.	43.70	25.21	20.59	10.50	-
Parent diplômé du sup.	27.54	23.95	29.34	19.16	1.89 *
Pas d'activité préparatoire	36.78	25.29	21.84	16.09	-
Participation aux AP	37.50	23.61	28.47	10.42	0.98

TABLE 4.4 – Répartition dans les groupes de médailles (en pourcentages) des étudiants vérifiant chaque variable binaire d'entrée (parmi celles sélectionnées lors de la régression ordinale). La dernière colonne présente les odds ratio obtenus lors de la régression ordinale. Ceux-ci sont suivis d'un astérisque si la variable y était significative.

Les rapports de cotes cumulés pour toutes les variables sauf `An_Prec` sont présentées dans la colonne OR de la table 4.4 (les autres valeurs du tableau sont disponibles pour référence, mais sans lien avec le modèle). On peut ainsi y lire que parler français à la maison, avoir eu plus de 6h de sciences par semaine en secondaire, ne pas avoir doublé en secondaire et avoir des parents diplômés du supérieur sont toutes des variables qui doublent à peu près les chances de passer d'une catégorie de médailles à la suivante par rapport à l'autre modalité respective, si toutes les autres variables explicatives restent égales. Le fait d'avoir eu plus de 6h de mathématiques par semaine en secondaire améliore plus modérément les chances de passer à un autre groupe de médailles, d'un facteur de 1,57.

Pour ce qui en est de la variable détaillant ce que l'étudiant faisait comme études l'année précédente, il est également possible d'interpréter son rapport de cotes cumulé en exponentiant ses coefficients. Simplement, au lieu d'un simple coefficient β , cette variable possède trois coefficients pour chacune des valeurs de $k = 1, 2, 3$. De plus, étant donné que le modèle utilisé par R utilise un signe + devant ces coefficients, il est nécessaire d'exponentier l'opposé des β_k plutôt que les β_k eux-mêmes si l'on souhaite obtenir l'augmentation de probabilité de passer dans la catégorie suivante de médailles lorsque l'on a déjà fait une année dans le supérieur (modalité alternative), par cohérence avec les autres variables.

Les valeurs obtenues sont les suivantes : en ayant déjà fait une année dans le supérieur, on a 3,03 fois plus de chances d'être dans le groupe 3 plutôt que dans le groupe 0 ; 2 fois plus de chances d'être dans le groupe 1 plutôt que dans le groupe 3 ; et 4,55 fois plus de chances d'être dans le groupe 2 plutôt que dans le groupe 1. Ces résultats confirment la surreprésentation de étudiants ayant déjà fait une année dans le supérieur dans le groupe des étudiants ayant utilisé le plus l'outil, qui avait déjà été observée lors de l'analyse des résidus du test du chi carré d'indépendance.

Chapitre 5

Impact de l'utilisation de l'outil sur la note

Les chapitres précédents ont permis d'explorer les deux bases de données de manière exploratoire et d'en tirer les liens de manière plus formelle. Ces premières observations étant faites, il est maintenant possible de tenter de répondre à la problématique avec les données disponibles, qui était de savoir si utiliser l'outil avait un impact significatif sur la note de l'étudiant lors de l'examen de janvier.

Il est toutefois possible que l'utilisation de l'outil soit corrélée avec la note uniquement parce que les étudiants qui utilisent l'outil sont ceux qui tendent à réussir pour d'autres raisons. Autrement dit, il existe peut-être un biais de sélection, qui peut être expliqué par les données d'entrées. Un modèle d'effet de traitement peut être utilisé pour déterminer l'existence d'un tel biais de sélection¹, et il sera utilisé par la suite pour déterminer si les données d'entrée sont susceptible de biaiser l'étude de la note en fonction de l'utilisation de l'outil.

Il est à noter qu'un biais de sélection non expliqué par les données d'entrée pourrait toujours être présent. Cependant, le choix des variables lors de la constitution de la base de données d'entrée a été faite dans le but de recenser les variables les plus susceptible d'influencer la note. Cette étude pourrait cependant toujours être améliorée en ajoutant d'autres questions dans l'enquête qui a mené à la constitution de cette base de données.

Comme il sera vu que la note n'est pas biaisée par les données d'entrée lorsque l'on compare les étudiants n'ayant pas utilisé du tout l'outil à ceux qui l'ont utilisé un tant soit peu, une régression linéaire multiple sera effectuée

1. Cette approche est inspirée par l'article de C. Dehon et al. (2019), qui utilisent le modèle d'Heckman, dont dérive le modèle d'effet de traitement qui sera utilisé ici.

afin d'évaluer l'impact de cette variable déterminant si l'outil a été utilisé ou non sur la note.

5.1 Biais induit par les données d'entrée

Commençons donc par déterminer si les données d'entrée biaisent (significativement) la note, lorsque l'on souhaite étudier l'influence de l'utilisation de l'outil. Pour ce faire, il est possible d'utiliser un modèle d'effet de traitement, qui est une extension du modèle d'Heckman². Cette section présente rapidement le modèle d'Heckman avant de détailler le modèle d'effet de traitement, qui sera employé afin de répondre à cette question.

5.1.1 Modèle d'Heckman

Le modèle d'Heckman est un modèle en deux temps qui vise à déterminer si, lorsque l'on a récolté des données dont la variable cible ne peut être observée pour certains individus, le fait de ne travailler qu'avec les individus dont la variable est disponible induit un biais significatif.

Pour ce faire, le modèle passe par deux équations : d'une part, une équation de régression linéaire classique, et d'autre part, une équation de sélection visant à modéliser la disponibilité de la variable dépendante.

La première équation est une régression linéaire, modélisant pour chaque observation i la variable dépendante cible y_i en fonction des variables explicatives dont les valeurs sont contenues dans le vecteur ligne x_i , *chaque fois que la y_i est disponible* :

$$y_i = x_i\beta + \epsilon_i$$

où β est un vecteur de coefficients de vecteurs à estimer, et ϵ_i est un terme d'erreur.

L'équation de sélection est un modèle probit qui modélise, pour toute observation i , une variable endogène w_i^* en fonction des variables explicatives de la disponibilité de y_i contenues dans le vecteur ligne z_i :

$$w_i^* = z_i\gamma + u_i,$$

où u_i est un terme d'erreur, et γ est un vecteur de coefficients à estimer, tel que y_i soit disponible si $w_i^* > 0$, et indisponible dans le cas contraire avec pour probabilités

$$\mathbb{P}(w_i = 1 | z_i) = \Phi(z_i\gamma) \quad \text{et} \quad \mathbb{P}(w_i = 0 | z_i) = 1 - \Phi(z_i\gamma)$$

2. Voir le chapitre 4 de (Guo & Fraser, 2014) pour une introduction du modèle d'Heckman. Le modèle d'effet de traitement y est détaillé en pages 96–98.

où Φ est la fonction de répartition de la normale centrée réduite, et où $w_i = 1$ si y_i est disponible, et $w_i = 0$ sinon. Autrement dit, on a $w_i = 1$ si $w_i^* > 0$, et $w_i = 0$ dans le cas contraire.

On suppose alors que les termes d'erreur ϵ_i et u_i sont normaux bivariés et de matrice de variance-covariance

$$\begin{pmatrix} \sigma_\epsilon & \rho \\ \rho & 1 \end{pmatrix}$$

où σ_ϵ et ρ sont à estimer.

Guo & Fraser (2014) proposent deux méthodes pour estimer les coefficients des deux équations. La première est la méthode d'estimation en deux étapes d'Heckman, qui exploite la méthode des moindres carrés et est détaillée dans leur article. La seconde est la technique maximum de vraisemblance, qui, d'après eux, demande davantage de puissance de calcul (même si ce n'est aujourd'hui plus un problème) et amène des résultats similaires. Ils renvoient à un article de Greene (1995) pour les détails sur cette méthode.

Pour ce qui en est de la variance des ϵ_i , σ_ϵ et de la corrélation entre les ϵ_i et u_i , Guo & Fraser (2014) citent des articles de Greene (1981, 2003) où de tels estimateurs sont construits.

Une fois que ces paramètres sont estimés, on vérifie si ρ est significativement différent de 0. Si tel est le cas, on peut conclure que la troncature (le fait que certaines valeurs manquent pour la variable cible) biaise significativement la régression linéaire modélisant la variable cible en fonction de ses variables explicatives.

5.1.2 Modèle d'effet de traitement

Le modèle qui sera utilisé ici est le modèle d'effet de traitement, qui est inspiré par le modèle d'Heckman et est à nouveau présenté par Guo & Fraser (2014). Dans ce modèle, on pose une variable de traitement w_i qui vaut 1 si l'observation est dans le groupe *traité*, et 0 sinon. L'objectif de ce modèle est de déterminer, lorsque la répartition des individus dans les deux groupes (traités et non traités) n'a pas été réalisée aléatoirement, si les deux groupes d'individus peuvent tout de même être considérés comme répartis aléatoirement dans le but de comparer une variable cible, ou si au contraire, les deux groupes sont biaisés de par leur constitution.

En particulier, et contrairement au modèle d'Heckman, le modèle d'effet de traitement contient la variable indicatrice de la présence du traitement w_i qui peut valoir 0 ou 1 dans l'équation de régression :

$$y_i = x_i\beta + w_i\delta + \epsilon_i$$

où en plus des notations précédentes, δ est un paramètre supplémentaire à estimer : cette équation est donc identique au modèle d'Heckman dans le cas non traité, mais possède un terme supplémentaire dans le cas traité. Pour ce qui en est de l'équation de sélection, elle est identique. Avec les mêmes notations que précédemment, la seconde équation s'écrit toujours

$$w_i^* = z_i\gamma + u_i.$$

avec pour but que l'individu i soit traité (et donc qu'on ait $w_i = 1$) si $w_i^* > 0$ et non traité ($w_i = 0$) sinon, avec pour probabilités

$$\mathbb{P}(w_i = 1 | z_i) = \Phi(z_i\gamma) \quad \text{et} \quad \mathbb{P}(w_i = 0 | z_i) = 1 - \Phi(z_i\gamma).$$

On suppose à nouveau que les termes d'erreur ϵ_i et u_i sont normaux bivariés et de matrice de variance-covariance

$$\begin{pmatrix} \sigma_\epsilon & \rho \\ \rho & 1 \end{pmatrix}$$

où σ_ϵ et ρ sont à estimer.

À nouveau, Guo & Fraser (2014) indiquent que deux méthodes sont possibles afin d'estimer les coefficients, et cette fois, ils détaillent la méthode du maximum de vraisemblance dans leur article. Pour l'estimation en deux étapes via les moindres carrés, ils renvoient à l'article de Maddala (1983).

Application du modèle d'effet de traitement

Afin de détecter la présence d'un biais significatif induit par les données d'entrée lors de la détermination de l'impact de l'utilisation de l'outil sur la note, il est possible d'utiliser le modèle d'effet de traitement.

Le modèle d'effet de traitement nécessite de définir le caractère « traité » d'un étudiant comme une variable binaire. Il n'est donc pas possible de considérer la variable d'appartenance à un groupe de médailles qui possède quatre modalités. Comme Dehon et al. (2019), on choisira de définir le traitement comme le fait d'avoir utilisé un tant soit peu l'outil (ce qui correspond à tous les groupes de médailles sauf le 0) et le non traitement comme le fait de ne pas avoir utilisé du tout l'outil (groupe de médailles 0).

Dans le modèle d'effet de traitement, on considère donc

- dans l'équation de régression, la note en fonction du traitement et des données d'entrée³ : il s'agit d'un modèle de régression linéaire qui per-

3. Comme on est ici à nouveau en présence d'un modèle linéaire, il est à nouveau nécessaire d'effectuer une sélection de variables pour les raisons exposées au chapitre 4. La même sélection qu'effectuée en section 4.2.1 a été reprise ici.

met de répondre à la principale problématique de cette étude, en regardant si l'utilisation de l'outil est significative dans la modélisation linéaire de la note ;

- et dans l'équation de sélection, le traitement en fonction des données d'entrée, car l'intuition est que ces données d'entrées peuvent influencer le traitement (à savoir, l'utilisation de l'outil).

La bibliothèque `sampleSelection` de R permet notamment de modéliser l'effet de traitement à l'aide de la fonction `treatReg`. Cependant, comme dans le détail du modèle présenté ci-dessus, le modèle ne prend en charge que deux modalités possibles pour le traitement : traité ou non traité. Le choix qui a été fait est de considérer l'étudiant comme traité s'il a fait une quelconque utilisation de l'outil, autrement dit, s'il appartient à un autre groupe de médailles que le groupe 0.

La fonction `treatReg` prend en arguments l'équation de sélection (la variable dépendante est la variable de traitement et les variables explicatives, celles qui avaient été sélectionnées pour la régression ordinaire du chapitre précédent), puis l'équation de régression (la variable dépendante est la note, et les variables explicatives sont les mêmes auxquelles on ajoute la variable de traitement), et enfin la méthode d'estimation (celle choisie est celle par défaut : maximum de vraisemblance). On obtient le résultat suivant :

Probit selection equation:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.35504	0.35910	-3.773	0.000187	***
Genre_F	-0.05217	0.14322	-0.364	0.715849	
Nat_Bel	-0.07731	0.21095	-0.366	0.714210	
Lang_FR	0.51482	0.20553	2.505	0.012668	*
Domic_Famille	0.27268	0.16737	1.629	0.104104	
AnPrec_Sup	0.61591	0.16576	3.716	0.000233	***
Math_Fort	0.16125	0.14005	1.151	0.250315	
Sc_Fort	0.24234	0.19403	1.249	0.212440	
DoubleSec_PasDouble	0.51700	0.19196	2.693	0.007388	**
DiplomPar_Sup	0.27577	0.15605	1.767	0.077999	.
AP_Participe	0.05210	0.14939	0.349	0.727490	

Outcome equation:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.3560	0.7128	4.708	3.5e-06	***
treatment	3.2420	1.9680	1.647	0.100312	
Genre_F	-0.5996	0.2887	-2.077	0.038445	*
Nat_Bel	0.5404	0.4263	1.268	0.205638	

Lang_FR	-0.1191	0.5568	-0.214	0.830710
Domic_Famille	0.2005	0.3850	0.521	0.602925
AnPrec_Sup	0.9498	0.5267	1.803	0.072118 .
Math_Fort	1.7429	0.3018	5.775	1.6e-08 ***
Sc_Fort	0.3516	0.4309	0.816	0.415022
DoubleSec_PasDouble	1.3376	0.5418	2.469	0.013994 *
DiplomPar_Sup	1.2733	0.3752	3.394	0.000762 ***
AP_Participe	0.4348	0.3076	1.414	0.158290

Error terms:

	Estimate	Std. Error	t value	Pr(> t)
sigma	2.64735	0.10388	25.484	<2e-16 ***
rho	0.06738	0.44561	0.151	0.88

On y constate, en lisant les résultats de l'équation de sélection probit, que parler le français à la maison, avoir déjà fait une année en supérieur, et ne pas avoir doublé en secondaire est lié avec le fait d'avoir utilisé l'outil; cependant, ce lien n'est pas suffisant pour rejeter l'égalité à 0 de la corrélation ρ entre les deux termes d'erreur (p-valeur de 0,88). Le modèle ne détecte donc pas de biais induit par les données d'entrée dans l'impact sur la note du fait d'utiliser ou non l'outil de simulation.

En revanche, si on avait défini le groupe de traitement comme étant les étudiants des groupes de médailles 1 et 2 seulement, en considérant que les étudiants du groupe 3 n'ont pas assez utilisé l'outil que pour être considérés comme traités, alors l'égalité à 0 de la corrélation entre les termes d'erreur ρ est rejetée (avec une p-valeur de 0,0057) :

Probit selection equation:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.765485	0.424438	-6.516	2.3e-10 ***
Genre_F	-0.009194	0.142541	-0.065	0.948605
Nat_Bel	0.315501	0.221360	1.425	0.154895
Lang_FR	0.494386	0.235456	2.100	0.036414 *
Domic_Famille	0.141207	0.165339	0.854	0.393616
AnPrec_Sup	0.487551	0.166347	2.931	0.003584 **
Math_Fort	0.318116	0.144770	2.197	0.028595 *
Sc_Fort	0.551606	0.210021	2.626	0.008978 **
DoubleSec_PasDouble	0.424546	0.216487	1.961	0.050601 .
DiplomPar_Sup	0.574742	0.172352	3.335	0.000938 ***
AP_Participe	0.007248	0.155820	0.047	0.962922

Outcome equation:

	Estimate	Std. Error	t value	Pr(> t)
--	----------	------------	---------	----------

(Intercept)	3.5883	0.9841	3.646	0.000303	***
treatm	0.5079	1.4731	0.345	0.730461	
Genre_F	-0.6652	0.3228	-2.061	0.040006	*
Nat_Bel	0.4083	0.4977	0.820	0.412522	
Lang_FR	0.4164	0.5177	0.804	0.421793	
Domic_Famille	0.4721	0.3923	1.203	0.229596	
AnPrec_Sup	1.5556	0.4330	3.592	0.000371	***
Math_Fort	1.8521	0.3713	4.988	9.29e-07	***
Sc_Fort	0.5650	0.5025	1.124	0.261604	
DoubleSec_PasDouble	1.8761	0.4935	3.802	0.000167	***
DiplomPar_Sup	1.5153	0.4394	3.448	0.000627	***
AP_Participe	0.5095	0.3443	1.480	0.139750	
Error terms:					
	Estimate	Std. Error	t value	Pr(> t)	
sigma	2.9887	0.3337	8.955	< 2e-16	***
rho	0.6274	0.2255	2.782	0.00567	**

En consultant les résultats de l'équation probit de ce second modèle, on constate qu'en plus des variables d'entrée signalées précédemment, le fait d'avoir suivi des options de mathématiques fortes et de sciences fortes en secondaire, ainsi que d'avoir des parents diplômés du supérieur, amènent plus souvent à être dans le groupe traité dans cette seconde configuration.

Lorsque l'on ignore les étudiants n'ayant pas utilisé l'outil (groupe 0) et que l'on compare les étudiants du groupe 3 aux étudiants des deux autres groupes, on obtient un résultat similaire. On peut en conclure que l'impact sur la note du type d'utilisation de l'outil est significativement biaisé par les données d'entrée.

En conclusion, les prédispositions socioéconomiques et scolaires des étudiants n'influencent pas significativement le fait que ceux-ci utilisent ou non l'outil⁴; mais lorsque les étudiants ont utilisé l'outil, ces prédispositions influencent l'intensité de l'utilisation de l'outil par l'étudiant.

Il est donc possible de comparer la note du groupe des étudiants n'ayant pas utilisé l'outil avec celle des autres étudiants (ayant donc utilisé l'outil), sans trop s'attendre à ce que ces résultats soient dûs aux prédispositions socioéconomiques et scolaires des étudiants. Par contre, lors de la comparaison de la note d'étudiants de différents groupes de médailles autres que 0, il faut s'attendre à ce qu'au moins une partie de l'explication soit due aux données d'entrée, et non à la seule utilisation de l'outil.

4. En tout cas, le modèle d'effet de traitement sur nos données ne permet pas d'affirmer le contraire.

5.2 Impact de l’outil sur la note

Il est à présent possible de répondre à la problématique principale qui est de déterminer l’impact de l’utilisation de l’outil sur la note. Il est possible de répondre à cette question à l’aide d’un modèle linéaire. Ensuite, quelques résultats supplémentaires seront obtenus à l’aide d’un modèle linéaire avec interactions et l’utilisation des groupes d’entrée constitués dans le chapitre 2.

5.2.1 Modélisation linéaire de la note

Avant de s’intéresser aux données de médailles, commençons par effectuer une régression linéaire multiple visant à déterminer la note en fonction des données d’entrée. Pour ce faire, on utilise la fonction de base de R `lm` qui effectue un modèle linéaire par la méthode des moindres carrés ordinaire, avec pour variable dépendante la note et comme variables explicatives celles qui ont été sélectionnées pour la régression ordinaire du chapitre 4. On obtient alors le tableau 5.1.

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.3927	0.8471	4.01	0.0001	*
Genre_F	-0.6650	0.3399	-1.96	0.0511	
Nat_Bel	0.4564	0.5032	0.91	0.3650	
Lang_FR	0.4923	0.4935	1.00	0.3191	
Domic_Famille	0.5026	0.4025	1.25	0.2125	
AnPrec_Sup	1.6370	0.3822	4.28	0.0000	*
Math_Fort	1.9177	0.3359	5.71	0.0000	*
Sc_Fort	0.6472	0.4659	1.39	0.1655	
DoubleSec_PasDouble	1.9495	0.4690	4.16	0.0000	*
DiplomPar_Sup	1.6030	0.3774	4.25	0.0000	*
AP_Participe	0.5043	0.3623	1.39	0.1647	

TABLE 5.1 – Résultats de la régression linéaire de la note en fonction des données d’entrée.

Le modèle linéaire de la note par rapport aux données d’entrées a un R^2 ajusté de 0,22. Il explique donc 22% de la note et davantage de variables explicatives sont nécessaires. Dans ce modèle, tous les étudiants partent avec 3,39/20. Le fait d’avoir fait une année dans le supérieur, d’avoir des parents diplômés du supérieur, d’avoir suivi l’option mathématiques fortes en secondaire, de même que le fait de ne pas avoir doublé en secondaire sont tous des facteurs qui augmentent la note d’un peu moins de 2 points, et les seuls qui soient significatifs dans ce modèle.

Ainsi, d'après ce modèle, un étudiant qui a déjà fait une année dans le supérieur, a suivi beaucoup d'heures de mathématiques et n'a jamais doublé en secondaire, et a des parents diplômés du supérieur obtiendrait une note de 10,5/20, là où un étudiant ayant les caractéristiques opposées n'aurait que 3,4/20.

Comme il a été montré à la section précédente, la note des étudiants n'ayant pas utilisé l'outil peut être comparée à celle de ceux qui ont utilisé l'outil sans avoir peur du fait que l'utilisation de l'outil soit influencée par d'autres variables d'entrée. Ajoutons donc une variable `Sim_Participe` qui vaut 1 si l'étudiant a participé à l'outil de simulation de l'examen, et 0 sinon dans le modèle linéaire. On obtient les résultats du tableau 5.2.

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.3526	0.7223	4.64	0.0000	*
Genre_F	-0.5937	0.2899	-2.05	0.0412	*
Nat_Bel	0.5480	0.4291	1.28	0.2023	
Lang_FR	-0.1745	0.4243	-0.41	0.6811	
Domic_Famille	0.1731	0.3443	0.50	0.6154	
AnPrec_Sup	0.8875	0.3316	2.68	0.0078	*
Math_Fort	1.7270	0.2869	6.02	0.0000	*
Sc_Fort	0.3249	0.3981	0.82	0.4150	
DoubleSec_PasDouble	1.2822	0.4036	3.18	0.0016	*
DiplomPar_Sup	1.2434	0.3231	3.85	0.0001	*
AP_Participe	0.4285	0.3089	1.39	0.1662	
Sim_Participe	3.5356	0.2897	12.20	0.0000	*

TABLE 5.2 – Résultats de la régression linéaire de la note en fonction des données d'entrée et de `Sim_Participe`.

Ce second modèle voit son R^2 ajusté augmenté à 0,43. Ce modèle explique donc 43% de la note, ce qui est mieux mais toujours perfectible. On constate que notre nouvelle variable qui détermine si l'étudiant a employé l'outil est significative et permet d'obtenir en moyenne 3,54 points supplémentaire par rapport à un étudiant pour lequel toutes les variables sont identiques sauf celle-ci.

En compensation, les points que font gagner les autres variables significatives est un peu diminué; le plus marquant étant le fait d'avoir déjà fait une année en secondaire qui ne fait plus que gagner 0,89 points (contre 1,63) pour le précédent modèle. Une partie de la note qui était expliquée par le fait d'avoir fait une année dans le supérieur est donc maintenant expliquée par le fait d'avoir participé à l'outil de simulation d'examen. Ceci confirme

ce qui avait été vu auparavant concernant cette variable. On avait vu que les étudiants venant du supérieur tendaient à utiliser l'outil en ligne (ils étaient même surreprésentés dans le groupe 2 des étudiants ayant récupéré le plus de médailles). Par ailleurs, cette variable était également significative dans le modèle probit du traitement (identique à `Sim_Participe`) en fonction des données d'entrées, ce qui semble aller dans le sens qu'il y a un certain lien entre le fait d'avoir fait une année dans le supérieur et de participer au simulateur d'examen, même si ce lien n'était pas suffisant pour que le modèle d'effet de traitement montre un biais significatif.

Dans ce second modèle, l'étudiant qui a déjà fait une année dans le supérieur, a suivi beaucoup d'heures de mathématiques et n'a jamais doublé en secondaire, et a des parents diplômés du supérieur obtiendrait une note de 12,03/20 s'il a utilisé l'outil, mais seulement 8,49/20 s'il n'a pas utilisé l'outil.

Terminons cette section en regardant tout de même le modèle linéaire qui modélise la note en fonction des données d'entrée ainsi que du groupe de médailles, en ayant bien conscience qu'il y a un certain biais des données d'entrée qui interfère dans le nombre de point réellement attribué au fait d'être dans un groupe de médailles. Celui-ci est présenté en table 5.3.

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.3562	0.6982	6.24	0.0000	*
Genre F	-0.6451	0.2744	-2.35	0.0192	*
Nat_Bel	0.3180	0.4073	0.78	0.4355	
Lang_FR	-0.2843	0.4012	-0.71	0.4790	
Domic_Famille	0.1151	0.3265	0.35	0.7246	
AnPrec_Sup	0.7057	0.3193	2.21	0.0277	*
Math_Fort	1.4761	0.2735	5.40	0.0000	*
Sc_Fort	0.0506	0.3782	0.13	0.8936	
DoubleSec_PasDouble	1.1768	0.3817	3.08	0.0022	*
DiplomPar_Sup	0.9704	0.3077	3.15	0.0017	*
AP_Participe	0.4929	0.2921	1.69	0.0923	
Groupe 1	4.3234	0.3421	12.64	0.0000	*
Groupe 2	4.9655	0.4247	11.69	0.0000	*
Groupe 3	2.2445	0.3335	6.73	0.0000	*

TABLE 5.3 – Résultats de la régression linéaire de la note en fonction des données d'entrée et des groupes de médailles.

Les résultats sont en fait assez similaires. Le R^2 ajusté est un peu meilleur, à 0,49. À peu près la moitié de la note est donc expliquée par ces variables.

Le fait de passer du groupe 0 à n'importe lequel des groupes garantit de gagner 2,24 au minimum (dans le cas du groupe 3), ce qui est très équivalent au nombre trouvé précédemment étant donné que l'ordonnée à l'origine vaut à peu près un point de plus que dans le modèle précédent.

On constate également que être dans le groupe 1 ou 2 augmente encore beaucoup les points, permettant d'obtenir 4 à 5 points de plus par rapport à un étudiant n'ayant pas utilisé l'outil. Cependant, ce dernier résultat est à mettre en perspective avec le fait que ces groupes sont biaisés, et qu'il y a probablement d'autres facteurs que le groupe d'appartenance qui fasse que les étudiants des groupes 1 et 2 gagnent tant de points par rapport aux autres.

5.2.2 Prise en compte des interactions

En extension du modèle linéaire classique, il est possible d'explorer les modèles linéaires avec interactions, en introduisant dans le modèle de nouvelles variables explicatives qui consistent en la multiplication de deux variables explicatives initiales.

Une recherche exhaustive consistant à ajouter des termes constitués de produits de variables d'entrées ou de la variable `Sim_Participe`, indicatrice de l'utilisation du simulateur, ne semble pas justifier l'utilité d'ajouter d'interaction, car aucune des nouvelles variables explicatives constituées n'est significative.

Cependant, lorsque la même opération est faite, non pas avec l'indicatrice de la participation au simulateur, mais avec les indicatrices d'appartenance à chacun des groupes de médailles, certaines interactions sont significatives et le modèle final voit un R^2 ajusté bien meilleur, à 0,57. Malheureusement, il n'est pas rigoureux d'interpréter ce nouveau modèle, à cause du biais connu des données d'entrée dans l'appartenance au groupe lors de la modélisation de la note.

Afin de se donner une idée de l'information qui pourrait être contenue dans ce modèle, une idée pourrait être d'utiliser les groupes d'entrée qui ont été constitués dans un but exploratoire dans le chapitre 2. En effet, l'idée de ces groupes est qu'ils sont, d'une certaine manière, homogènes, car ils ont été constitués afin de maximiser la similarité entre les individus d'un même groupe et la dissimilarité entre des individus de deux groupes différents.

On constitue alors, toujours à l'aide de la fonction `ml` de R, le modèle linéaire avec interactions de la note en fonction du groupe d'entrée, du groupe de médailles, et des interactions entre les groupes d'entrées et de médailles. On obtient le tableau 5.4.

Ce nouveau modèle a un R^2 ajusté de 0,43, ce qui est plus faible que

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	6.7218	0.6135	10.96	0.0000	*
grE2	-0.2074	0.7227	-0.29	0.7743	
grE3	0.3633	0.6987	0.52	0.6034	
grE4	-0.1902	0.8796	-0.22	0.8289	
grM1	4.6127	1.0132	4.55	0.0000	*
grM2	4.3917	1.0448	4.20	0.0000	*
grM3	3.6360	0.9074	4.01	0.0001	*
grE2 :grM1	0.3503	1.1665	0.30	0.7641	
grE3 :grM1	0.0656	1.1587	0.06	0.9549	
grE4 :grM1	1.3060	1.3909	0.94	0.3483	
grE2 :grM2	2.1089	1.3974	1.51	0.1321	
grE3 :grM2	2.0464	1.2743	1.61	0.1091	
grE4 :grM2	1.3077	1.3588	0.96	0.3365	
grE2 :grM3	-0.0067	1.1067	-0.01	0.9952	
grE3 :grM3	-3.1573	1.0674	-2.96	0.0033	*
grE4 :grM3	-0.5379	1.2495	-0.43	0.6671	

TABLE 5.4 – Résultats de la régression linéaire avec interactions de la note en fonction du groupe d'entrée, du groupe de médailles, et des interactions entre les groupes d'entrées et de médailles. L'abréviation grE indique un groupe d'entrée, et grM un groupe de médailles. Les groupe de médailles 0 et groupe d'entrée 1 sont les groupes de référence.

le modèle précédent où les groupes de médailles étaient également pris en compte. Il ne considère aucun groupe d'entrée séparément comme significatif, ce qui est différent des modèles précédents où certaines données d'entrées l'étaient.

Le fait que les groupes de médailles soient significatifs est à prendre en compte avec la même prudence que précédemment. En effet, là où il est acceptable, vu ce qui a été vu précédemment, de comparer le fait de n'avoir pas utilisé l'outil avec le fait d'être dans n'importe lequel des trois autres groupes, les valeurs différentes d'un groupe de participants à l'autre ne peuvent pas réellement être interprétées.

Cependant, dans ce modèle, on constate que le nombre de points apportés par les trois groupes de médailles n'est pas aussi différent. Avec au départ un score de 6,72/20 dans ce modèle, on ajoute au moins 3,63 points dans le pire des cas (groupe 3) aux étudiants ayant utilisé l'outil : ce qui signifierait la plupart du temps une réussite. Cependant, ce modèle nous apporte également l'information que les étudiants du groupe d'entrée 3 qui sont dans le groupe de médailles 3 perdent en moyenne 3,16 points.

Autrement dit, un étudiant gagne au moins 3,63 points pour avoir utilisé l'outil, sauf s'il est membre du groupe d'entrée 3⁵ et du groupe de médailles 3, auquel cas il ne gagne que 0,48 points, ce qui équivaut d'après le modèle à un échec. Cet impact sur la note très différencié chez ce profil d'étudiant entre une utilisation faible (groupe 3 de médailles) et une utilisation plus intense (groupes 1 et 2) de l'outil pourrait s'expliquer par l'impact de l'outil chez ce type d'étudiant, mais pourrait aussi s'expliquer par le fait qu'aucun étudiant du groupe 3 n'a eu accès à l'examen blanc en présentiel (dont la condition d'accès était d'avoir quatre médailles d'or), au contraire des membres des groupes 1 et 2 qui y ont presque tous eu accès.

Notons enfin que ce dernier résultat est hautement tributaire de la composition des groupes d'entrées par clustering hiérarchique, qui contient une part d'arbitraire. Pour cette raison, ainsi que la précédente, il pourrait être utile de mener une autre étude plus axée sur ce profil d'étudiant (qui est le plus courant dans la population étudiée) afin de déterminer cet effet particulier de l'utilisation de l'outil, c'est-à-dire de la nécessité de l'utiliser de manière suffisamment intense, chez ce type d'étudiant.

5. Pour rappel, ce groupe est principalement constitué d'étudiants sortant du secondaire, et ayant suivi une option scientifique, mais moins de mathématiques. Leurs parents sont également moins souvent diplômés du supérieur.

Chapitre 6

Conclusion

Ce dernier chapitre rappelle les principaux résultats obtenus dans ce mémoire et propose quelques pistes de recherches pour approfondir le sujet.

6.1 Principaux résultats

Ce mémoire visait avant tout à répondre à une problématique, qui était de déterminer si l'utilisation de l'outil améliorait la note des étudiants, en tenant compte d'éventuels biais pouvant influencer la note sans que cela ne soit en lien avec l'utilisation du simulateur.

Il a été déterminé, dans le chapitre 5, que l'utilisation du simulateur apporte bien un avantage, de l'ordre de 3,5 points sur 20, aux étudiants lorsqu'on compare les étudiants ayant utilisé un tant soit peu le simulateur à ceux qui ne l'ont pas utilisé du tout, et ce résultat n'est pas biaisé par les données d'entrée.

Quand on compare les étudiants ayant peu utilisé l'outil à ceux qui l'ont plus utilisé, on voit également une amélioration de la note chez ceux qui ont le plus utilisé l'outil ; cependant, ce résultat est manifestement biaisé par les données d'entrée. Toutefois, un certain profil d'étudiants, caractérisés par une formation plus scientifique au détriment de mathématiques en secondaire, sortant du secondaire, et ayant des parents moins diplômés semblent ne réellement bénéficier de l'outil que s'ils l'utilisent suffisamment, au moins pour chercher les principales médailles d'or.

D'autres résultats moins importants obtenus tout au long de ce mémoire retiennent tout de même l'attention. Tout d'abord, les chapitres 2 et 3, biens que purement exploratoires, permettent de se donner une idée des profils d'étudiants inscrits en médecine l'année étudiée et de ce qui les différencie, pour le chapitre 2 ; quant au chapitre 3, il permet de comprendre les différents

comportements qu’ont pu prendre les étudiants face à l’outil de simulation d’examen.

Par ailleurs, dans le chapitre 4, on a pu se rendre compte de l’utilisation particulière de l’outil qui est faite par les étudiants ayant déjà fait une année dans le supérieur, qui ont tendance à utiliser beaucoup plus l’outil que les étudiants qui sortent du secondaire.

6.2 Pistes de recherche

De nombreuses questions restent à explorer dans ce domaine, et même sur ce sujet en particulier.

Tout d’abord, l’étude d’éventuels biais dans la note obtenue induite par d’autres éléments que l’utilisation de l’outil est restée cantonnée aux données à notre disposition dans la base de données d’entrée. Il serait cependant possible que la note soit biaisée par d’autres données qui n’étaient pas disponibles pour cette étude. Ajouter d’autres variables dans la base de données d’entrée est donc une première piste d’amélioration du modèle final.

En particulier, il n’y a pas dans cette base de données d’entrée de variable indiquant si l’étudiant a suivi du latin en secondaire, qui est une variable souvent liée dans les études avec la réussite dans le supérieur. Cette dernière variable a d’ores et déjà été ajoutée pour les années suivantes (non étudiées ici), mais là encore, les modalités proposées dans le questionnaire ne sont pas adéquates¹. Par ailleurs, comme remarqué dans cette étude, les variables quantifiant le nombre d’heures d’une certaine matière (mathématiques, physique, sciences) en secondaire sont assez problématiques à utiliser telles quelles et il a été nécessaire de les regrouper. Il serait peut-être profitable, dans une future enquête, de proposer directement des modalités plus qualitatives (option « forte » versus option « moins forte », avec une explication de ce qu’on entend par là) afin d’éviter d’avoir de nombreuses modalités dont la distinction n’est pas réellement porteuse de sens.

Ensuite, cette étude ne s’est basée que sur la seule année 2015–2016, mais il pourrait être intéressant de la répéter pour d’autres années afin de voir une éventuelle évolution, mais aussi de déterminer si les différentes variables qui ont été ajoutées dans les données d’entrée les années suivantes améliorent le modèle. Par ailleurs, il serait intéressant de voir l’évolution du comportement des répétants, qui se retrouveraient donc dans deux bases de données d’années distinctes, afin de confirmer ou d’infirmes des hypothèses sur la raison pour

1. La variable ajoutée en 2018 propose à l’étudiant d’indiquer s’il a suivi 2 années ou 6 années de latin, ce qui l’empêche de répondre par exemple qu’il a suivi 4 années de latin. Il s’ensuit que les réponses risquent d’être aléatoires pour les étudiants concernés.

laquelle les étudiants qui ont déjà fait une année dans le supérieur optent pour un comportement particulier vis-à-vis de l'outil.

Ce dernier point, mais aussi bien d'autres, pourraient d'ailleurs être abordés lors d'entretiens réalisés auprès d'étudiants afin de comprendre leur intérêt pour l'outil, la manière qu'ils ont eu de l'utiliser, et l'utilité qu'ils pensent en avoir tiré.

Enfin, des questions plus précises sur la manière de tirer le meilleur profit de l'outil mériteraient d'être éclairées. Ainsi, il pourrait être intéressant de déterminer si utiliser l'outil sans chercher à bien répondre aux questions (mais de juste les consulter) apporte quelque chose. Les données pour y répondre étaient disponibles, mais difficiles à exploiter. Il pourrait également être intéressant de déterminer si le moment d'utilisation de l'outil a une importance : les étudiants ayant utilisé l'outil dès que la matière avait été vue ont-ils un avantage comparativement à ceux qui l'ont utilisé au dernier moment ? Les données pour répondre à cette question sont disponibles, mais deviennent très complexe par l'ajout de profils temporels pour chaque étudiant et chaque test. Dernièrement, il serait tout de même intéressant de déterminer si l'intensité d'utilisation de l'outil a réellement un impact sur les résultats ; un autre protocole devrait cependant sans doute être mis en place afin de répondre à cette question afin d'éliminer le biais de sélection qui empêche d'y apporter une réponse définitive.

Chapitre 7

Annexes

7.1 Annexe 1 : Noms des variables binaires

La table 7.1 en page suivante présente le nom et la description des variables binaires telles qu'elles ont été encodées dans le logiciel, qui peuvent apparaître à divers endroits dans ce mémoire. Dans le corps de texte, la description aura cependant été utilisée autant que possible.

Variable	Description
GenreF	Femme
GenreM	Homme
AgeInf17	17 ans ou moins
Age18	18 ans
Age19	19 ans
AgeSup20	20 ans ou plus
NatBel	Belge
NatPasBel	Autre nationalité
LangFR	Parle français à la maison
LangPasFR	Parle une autre langue
DomicFam	Domicilié avec sa famille
DomicPasFam	Autre domicile pendant l'année
DiplomPar2Sup	2 parents diplômés du sup.
DiplomPar1Sup	1 seul parent diplômé du sup.
DiplomParAutre	Parents non diplômés du sup.
AnNPrimant	Primant (ou réorienté)
AnNPasPrimant	Répétant
AnPrecSecond	Secondaire l'année précédente
AnPrecSup	Supérieur l'année précédente
APParticipe	Participation aux AP
APPasParticipe	Pas d'activité préparatoire
EnsSecFWB	Secondaire en FWB
EnsSecHorsFWB	Secondaire hors FWB
PasDoubleSec	N'a pas doublé en secondaire
DoubleSec	A doublé en secondaire
MathFort	6h ou plus de math
MathFaible	Moins de 6h de math
PhysFort	3h ou plus de physique
PhysMoy	2h de physique
PhysFaible	1h de physique
ScLabo	8h ou plus de sciences
ScRenfo	7h de sciences
ScFort	6h de sciences
ScFaible	Moins de 6h de sciences

TABLE 7.1 – Nom et description des variables binaires telles qu'elles ont été encodées dans le logiciel. Celles-ci peuvent apparaître à divers endroits dans ce mémoire.

7.2 Annexe 2 : Matrice de corrélation des variables binaires

La matrice de corrélation des variables binaires de la base de données d'entrée est proposée en figure 7.1 via une représentation colorée. Afin d'en améliorer la lisibilité, les variables ont été ajoutées de manière non redondante : ainsi, la dernière modalité de chaque variable initiale n'est pas représentée.

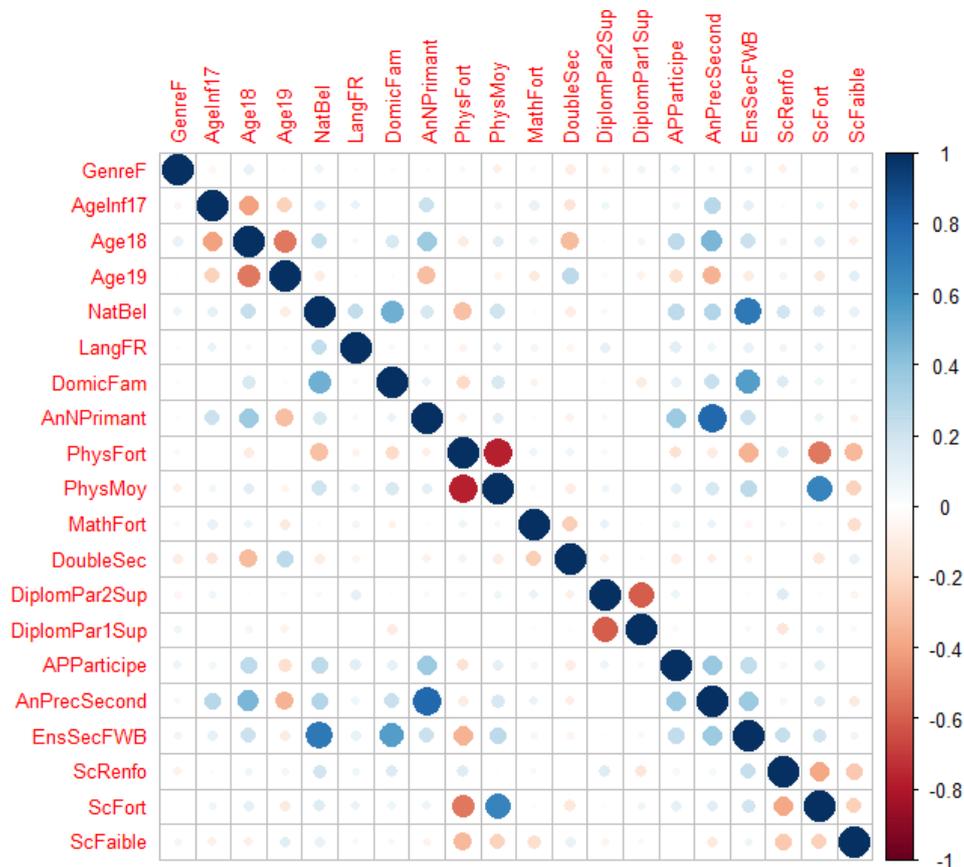


FIGURE 7.1 – Représentation de la matrice de corrélation de la base de données d'entrée binarisée, sans les variables redondantes. Un grand cercle bleu indique une corrélation positive, un grand cercle rouge, une corrélation négative.

Bibliographie

- [1] Vladimir Batagelj, *Generalized Ward and related clustering problems*, Classification and Related Methods of Data Analysis (1988), 67–74.
- [2] R. H. B. Christensen, *Cumulative link models for ordinal regression with the R package ordinal*, disponible via l'URL <https://cran.r-project.org/web/packages/ordinal/vignettes/clm_article.pdf>, documentation de R, 2018.
- [3] C. Dehon, P. Emplit, et E. Van Lierde, *A case study of learning analytics within a statistics course for undergraduate students in economics*, Decision making based on data (Kuala Lumpur, Malaysia) (S. Budgett, éd.), Proceedings of the Satellite conference of the International Association for Statistical Education, août 2019.
- [4] Anne-Françoise Donneau, *Contribution to the statistical analysis of incomplete longitudinal ordinal data*, Thèse de doctorat, Université de Liège, avril 2013.
- [5] Shenyang Guo et Mark W. Fraser, *Propensity score analysis*, Sage Publications, septembre 2014.
- [6] Wolfgang Karl Härdle et Léopold Simar, *Applied multivariate statistical analysis*, 4^e éd., Springer-Verlag, Berlin, 2015.
- [7] Pierre-Xavier Marique, Jean-François Van de Poël, et Maryse Hoebeke, *Quel outil d'entraînement pour des étudiants en médecine évalués par QCM en physique ?*, Actes du 28^e colloque de l'ADMEE europe (Lisbonne, Portugal), janvier 2016.
- [8] Fionn Murtagh et Pierre Legendre, *Ward's hierarchical agglomerative clustering method : Which algorithms implement Ward's criterion ?*, Journal of Classification **31** (2014), n^o 3, 274–295.
- [9] John Neter, William Wasserman, et Michael H. Kutner, *Applied linear regression models*, Irwin, Homewood, Illinois, 1989.

- [10] Zacharoula Papamitsiou et Anastasios Economides, *Learning analytics and educational data mining in practice : A systematic literature review of empirical evidence*, Educational Technology & Society **17** (2014), 49–64.
- [11] Billy Tak Ming Wong, *Learning analytics in higher education : an analysis of case studies*, Asian Association of Open Universities Journal **12** (2017), n° 1, 21–40.

Table des matières

1	Introduction	1
1.1	Problématique	2
1.2	Revue de la littérature	3
2	Données d'entrée	5
2.1	Variables des données d'entrée	5
2.1.1	Données socioéconomiques	6
2.1.2	Données liées au parcours des étudiants	6
2.2	Traitement des données manquantes	8
2.2.1	Étudiants n'ayant pas participé à l'enquête	8
2.2.2	Autres données manquantes	10
2.3	Liens entre différentes variables d'entrée	13
2.4	Classification initiale des étudiants	16
2.4.1	Constitution de groupes d'étudiants	16
2.4.2	Description des groupes	19
2.5	Liens entre les données d'entrées et la note obtenue à l'examen	24
3	Données de travail	29
3.1	Données d'obtention des médailles	30
3.1.1	Constitution de groupes d'étudiants selon les médailles obtenues	31
3.1.2	Interprétation des groupes obtenus	32
3.1.3	Notes obtenues dans chaque groupe	35
3.1.4	Stabilité vis-à-vis du dernier chapitre	35
3.2	Données concernant les tentatives	38
3.3	Moyenne sur les tentatives valides	39
4	Liens entre les données d'entrée et de travail	41
4.1	Comparaison des groupes d'entrée et des groupes de médailles	41
4.2	Impact des données d'entrée sur le groupe de médailles	43
4.2.1	Sélection des variables explicatives	43

4.2.2	Régression ordinale	46
5	Impact de l'utilisation de l'outil sur la note	54
5.1	Biais induit par les données d'entrée	55
5.1.1	Modèle d'Heckman	55
5.1.2	Modèle d'effet de traitement	56
5.2	Impact de l'outil sur la note	61
5.2.1	Modélisation linéaire de la note	61
5.2.2	Prise en compte des interactions	64
6	Conclusion	67
6.1	Principaux résultats	67
6.2	Pistes de recherche	68
7	Annexes	70
7.1	Annexe 1 : Noms des variables binaires	70
7.2	Annexe 2 : Matrice de corrélation des variables binaires	72