



AIX MARSEILLE UNIVERSITE

ECOLE DOCTORALE EN MATHEMATIQUES ET
INFORMATIQUE DE MARSEILLE (ED184)

UFR SCIENCES

INSTITUT DE MATHEMATIQUES DE MARSEILLE (I2M) UMR 7373

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline : Mathématiques et Informatique

Spécialité : Statistiques

Pierre MICHEL

Sélection d'items en classification non-supervisée et questionnaires
informatisés adaptatifs : applications à des données de qualité de vie
liée à la santé

Soutenue le 13/12/2016 devant le jury :

Christophe BIERNACKI	Université Lille 1	Rapporteur
Iven VAN MECHELEN	Université de Louvain	Rapporteur
Liliane BEL	Université Paris-Sud	Examinatrice
Pascal AUQUIER	Aix Marseille Université	Examineur
Laurent BOYER	Aix Marseille Université	Directeur de thèse
Badih GHATTAS	Aix Marseille Université	Directeur de thèse



Cette oeuvre est mise à disposition selon les termes de la Licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 3.0 France.

Résumé

Cette thèse s'inscrit dans le cadre de la classification non supervisée et s'intéresse plus particulièrement au problème de la hiérarchisation de variables. Elle s'articule autour de nombreuses applications dans le domaine de la santé publique, notamment l'analyse des données issues des questionnaires de qualité de vie liée à la santé pour les patients atteints de maladies chroniques. Elle comporte des aspects techniques, méthodologiques et pratiques, et est structurée en trois parties : développement de questionnaires adaptatifs, classification non supervisée de données qualitatives et sélection de variables en apprentissage non supervisé.

Contexte et motivations

Cette thèse s'intéresse aux propriétés statistiques et psychométriques des mesures subjectives dites "modernes" de la qualité de vie. Dans la pratique clinique, les mesures de la qualité de vie sont considérées comme d'importants indicateurs pour de nombreuses prises de décisions (évaluation des interventions de santé, contrôle des symptômes liés à une pathologie ou un traitement). Nous nous intéressons ici à la théorie de réponse à l'item, qui utilise des modèles paramétriques pour établir une relation entre le trait latent d'un individu (son niveau de qualité de vie par exemple) et un item donné. Nous utilisons ces modèles pour sélectionner des sous-ensembles d'items présentant des propriétés psychométriques satisfaisantes, au regard de la qualité d'ajustement du modèle. Cette démarche, appelée *item banking* ou calibration, permet de construire une banque d'items, c'est à dire un ensemble de questions à plusieurs modalités de réponses ayant pour but de mesurer un même trait latent. Ces items, une fois collectés, peuvent être administrés à un individu de manière séquentielle. Cependant, il est aussi possible de considérer un questionnaire informatisé adaptatif, basé sur la banque d'items calibrée et sur un algorithme de sélection d'items. Dans ce type de questionnaire, le score reflétant le niveau de trait latent de l'individu est estimé après chaque réponse à un item. L'item suivant est sélectionné par l'algorithme en fonction du score actuel de l'individu, comme étant l'item le plus informatif pour cet individu (l'information d'un item est une fonction du trait latent). Il est possible de n'administrer qu'un sous-ensemble des items de la banque avec cette procédure, en définissant un critère d'arrêt basé sur un nombre d'items fixe, ou sur un indice de précision du score obtenu.

Dans la pratique clinique, ce type de questionnaires est utilisé pour fournir une mesure valide de la qualité de vie des patients, tout en réduisant le fardeau ressenti par le patient lorsque le nombre d'items à compléter est trop élevé, ou que certaines questions dans la banque ne sont pas adaptées à son niveau de trait latent. Cependant, cette approche est très dépendante de la théorie de réponse

à l'item, et il est possible qu'en pratique un ensemble d'items donné ne soit pas conforme aux standards d'ajustement de ces modèles paramétriques. En effet, la théorie de réponse à l'item repose sur des hypothèses fondamentales (unidimensionalité, indépendance locale) qui peuvent être invérifiables sur certains jeux de données réelles. De plus, le développement et la mise à jour de banques d'items sont des processus coûteux et difficiles à mettre en place dans la pratique clinique.

Développement de questionnaires adaptatifs

Nous proposons donc une alternative non-paramétrique pour le développement de questionnaires adaptatifs, basée sur la méthode CART, les arbres de décisions de Breiman. Elle consiste à construire un arbre de régression afin de prédire la valeur d'une variable aléatoire Y représentant un score (de qualité de vie par exemple) en fonction de p variables explicatives $X_{.j}$, $j \in \{1, \dots, p\}$, qui sont des variables ordinales représentant les items de la banque. Cette approche présente certains avantages : elle ne repose sur aucune hypothèse, elle peut s'adapter aux jeux de données contenant des données manquantes, elle requiert moins de temps de calcul lors de l'administration des items et elle permet aussi d'évaluer un potentiel comportement différentiel des items, en ajoutant des variables démographiques dans le modèle.

En terme de résultats, nous présentons la calibration d'une banque d'items mesurant la qualité de vie liée à la santé mentale pour des patients atteints de sclérose en plaques. Cette banque est constituée d'un regroupement d'items issus de deux questionnaires. Nous présentons des simulations pour le développement des versions adaptatives du MusiQoL et du SQoL, qui sont deux questionnaires multidimensionnels de qualité de vie spécifiques à la sclérose en plaques et à la schizophrénie respectivement. Enfin, nous présentons différentes simulations démontrant la pertinence de l'approche basée sur les arbres de décision. L'approche classique est comparée à cette méthode, cette dernière fournit des estimations de scores de qualité de vie plus satisfaisantes au regard d'un score de référence.

Classification non supervisée de données qualitatives

La deuxième partie de cette thèse présente une méthode de classification hiérarchique non supervisée, basée sur des arbres de décisions binaires, appelée *clustering using unsupervised binary trees* (CUBT). CUBT est une méthode descendante qui comprend trois étapes pour obtenir une partition optimale d'un jeu de données. La première étape construit un "arbre maximal" en divisant récursivement le jeu de données en deux sous-ensembles d'observations, les subdivisions faites visant à minimiser un critère d'hétérogénéité, appelé *déviance*, calculé sur les sous-ensembles obtenus. La *déviance* est basée sur la trace de la matrice variance-covariance de chaque sous-ensemble d'observations. La deuxième étape consiste à élaguer l'arbre, elle permet de regrouper les paires de noeuds terminaux de l'arbre. Pour chaque paire de noeuds terminaux issue d'une même division bi-

naire, une mesure de similarité inter-noeuds est calculée. Les deux noeuds sont agrégés au sein d'un même noeud (i.e le noeud parent) si leur dissimilarité est inférieure à un certain seuil noté *mindist*. La mesure de dissimilarité utilisée dans CUBT est basée sur la distance euclidienne. Enfin, la troisième étape permet aussi d'élaguer l'arbre, en considérant également l'aggrégation des noeuds terminaux qui ne sont pas issus d'une même division. Pour cette étape, on peut choisir d'utiliser soit le critère d'hétérogénéité, soit la mesure de dissimilarité. La méthode CUBT a les mêmes avantages que CART, elle peut produire des partitions optimales pour une grande variété de données et possède de bonnes propriétés de convergence. Elle est aussi interprétable grâce à la lecture des divisions binaires obtenues dans l'arbre.

Cette dernière propriété est mise en évidence dans quelques applications en santé publique, en utilisant CUBT avec des scores issus de questionnaires de qualité de vie. Cette approche nous permet d'apporter une aide à l'interprétabilité et à la prise de décision vis à vis de ces scores, la structure d'arbre de décision permettant aisément d'obtenir des scores-seuils pour définir l'appartenance des patients à différentes sous-groupes.

Une des limitations techniques de la version initiale de CUBT est le fait que le critère d'hétérogénéité et la mesure de dissimilarité utilisés sont spécifiques aux données quantitatives continues. Nous proposons donc des extensions de CUBT pour l'adapter au cas de données ordinales (de type item) et nominales. Nous suggérons de nouveaux critères, basés sur l'information mutuelle et l'entropie de Shannon. Différents modèles de simulation de données sont présentés pour expérimenter cette nouvelle version de CUBT et la comparer à d'autres approches non supervisées. Nous définissons aussi quelques heuristiques concernant le choix des paramètres de CUBT.

Sélection de variables en apprentissage non supervisé

Nous nous intéressons ensuite au problème de sélection de variables en classification non supervisée. Un arbre de classification construit avec CUBT permet d'identifier les variables qui prennent part activement à la construction de l'arbre. Cependant, bien que certaines variables peuvent être non pertinentes pour la construction de l'arbre, elles peuvent être compétitives pour les variables actives dans les différentes divisions binaire de l'arbre. Dans de nombreuses applications d'analyse de données ou de modélisation statistique, il est essentiel de classer les variables selon un score d'importance afin de déterminer leur pertinence dans un modèle donné. La sélection de variables permet ainsi de réduire la complexité des modèles que l'on utilise, ou de réduire le bruit présent dans les données, afin d'obtenir un gain de précision et d'interprétabilité du modèle. Nous présentons donc une méthode pour mesurer l'importance des variables dans le cadre de la classification non-supervisée. Cette méthode, inspirée de CART, utilise CUBT et la notion de divisions binaires compétitives pour définir un score d'importance des variables. Nous analysons l'efficacité et la stabilité de ce nouvel indice, en le

comparant à d'autres méthodes classiques de scores d'importance de variables. Nous considérons des modèles de simulation de données pour comparer notre approche, en ajoutant des variables non pertinentes dans les jeux de données obtenus. Cette méthode montre des résultats satisfaisants en termes d'efficacité et de stabilité. Ce nouveau critère peut être utilisé pour obtenir une hiérarchie des variables d'un jeu de données, et développer un algorithme performant de sélection de variables.

Cette thèse a donné lieu à plusieurs publications dans des revues et dans des conférences internationales. En voici la liste :

Publications issues de la thèse

- Pierre Michel, Badih Ghattas et Laurent Boyer (2016) *Assessing variable importance in clustering. A new method based on unsupervised binary decision trees*. (soumis)
- Pierre Michel, Karine Baumstarck, Christophe Lançon, Badih Ghattas, Anderson Loundou, Pascal Auquier et Laurent Boyer (2016) *Modernizing quality of life assessment : development of a multidimensional computerized adaptive questionnaire for patients with schizophrenia*. (en révision)
- Badih Ghattas, Pierre Michel et Laurent Boyer (2016) *Clustering nominal data using Unsupervised Binary decision Trees : Comparisons with the state of the art methods*. (en révision)
- Pierre Michel, Karine Baumstarck, Badih Ghattas, Jean Pelletier, Anderson Loundou, Mohamed Boucekine, Pascal Auquier et Laurent Boyer (2016) *A Multidimensional Computerized Adaptive Short-Form Quality of Life Questionnaire Developed and Validated for Multiple Sclerosis : The MusiQoL-MCAT*. *Medicine*, 95(14) :e3068.
- Pierre Michel, Pascal Auquier, Karine Baumstarck, Anderson Loundou, Badih Ghattas, Christophe Lançon et Laurent Boyer (2015) *How to interpret multidimensional quality of life questionnaires for patients with schizophrenia ?* *Quality of Life Research*, 24(10).
- Pierre Michel, Pascal Auquier, Karine Baumstarck, Jean Pelletier, Anderson Loundou, Badih Ghattas et Laurent Boyer (2015) *Development of a cross-cultural item bank for measuring quality of life related to mental health in multiple sclerosis patients*. *Quality of Life Research*, 24(9).
- Pierre Michel, Karine Baumstarck, Laurent Boyer, Oscar Fernandez, Peter Flachenecker, Jean Pelletier, Anderson Loundou, Badih Ghattas et Pascal

Auquier (2014) *Defining Quality of Life Levels to Enhance Clinical Interpretation in Multiple Sclerosis*. Medical care.

- Pierre Michel, Karine Baumstarck, Pascal Auquier, Xavier Amador, Rémy Dumas, Jessica Fernandez, Christophe Lancon et Laurent Boyer (2013) *Psychometric properties of the abbreviated version of the Scale to Assess Unawareness in Mental Disorder in schizophrenia*. BMC Psychiatry, 13(1) :229.

Participations à des conférences internationales

- Pierre Michel, Karine Baumstarck, Laurent Boyer, Anderson Loundou, Badih Ghattas et Pascal Auquier (2016) *Development of a new quality of life computerized adaptive testing algorithm based on regression trees*. 23rd Annual Conference of the International Society for Quality of Life Research, Copenhagen, Denmark.
- Pierre Michel et Badih Ghattas (2016) *Variable Importance in Clustering Using Binary Decision Trees*. COMPSTAT 2016, 22nd International Conference on Computational Statistics, Oviedo, Spain.
- Pierre Michel, Zeinab Hamidou, Laurent Boyer, Badih Ghattas et Pascal Auquier (2015) *Clustering based on unsupervised binary trees to define levels of symptom severity in cancer*. 22nd Annual Conference of the International Society for Quality of Life Research, Vancouver, Canada.
- Pierre Michel et Badih Ghattas (2014) *Clustering ordinal data using binary decision trees*. COMPSTAT 2014, 21st International Conference on Computational Statistics, Geneva, Switzerland.

Mots clés : sélection de variables, banques d'items, questionnaires adaptatifs, classification non supervisé, arbres de décision

Abstract

This thesis deals with unsupervised classification (commonly called clustering) and its main focus is about the problem of variables ranking. The document presents several applications in the field of public health, particularly the statistical analysis of health-related quality of life data. These data are from quality of life questionnaires for patients with chronic diseases. The thesis consists in technical, methodological and practical aspects, and is structured in three parts : development of computerized adaptive tests, clustering of qualitative data and variable selection in unsupervised learning.

Context and motivations

This thesis focus on the statistical and psychometric properties of so-called "modern" subjective quality of life measures. In clinical practice, the quality of life measures are considered as important indicators for many clinical decisions (assessment of care provided, control of symptoms due to some pathology or treatment). We herein focus on item response theory, that uses parametric models that consider a link between an individual's latent trait (for example his level of quality of life) and a given item. We use these models to select subsets of items that have satisfactory psychometric properties, regarding the model fit. This approach, named item banking, or item calibration aims to construct an item bank, a set of items having several response choices that measure a same latent trait. Once collected, these items can be administered sequentially to an individual. However, a computerized adaptive test, based on the calibrated item bank and an algorithm of adaptive administration, can be considered. In this type of tests, the score reflecting the individual's latent trait level is estimated after each response to an item. The next item is selected by the algorithm regarding the current individual's score, it the most informative for this individual (information is a function of the latent trait). A computerized adaptive test administers only a subset of items from the item bank, defining a stopping criterion based on a fixed number of items, or based on an index of score precision.

In clinical practice, this type of questionnaires is used to provide a valid measure of patients' quality of life, reducing the burden felt by the patient when the number of items to complete is too high, or when some items in the item bank are not adapted to his latent trait level. This approach is very dependent of item response theory, but in practice, a set of items could not be in conformity with the fit of these parametric models. Indeed, item response theory have three fundamental assumptions (uni dimensionality, local independence and monotonicity) that can be not valid on several real datasets. Moreover, development and updating of large item banks are costly processes that are difficult to undertake in some clinical settings.

Development of computerized adaptive tests

We propose a non-parametric alternative for the development of adaptive tests, which is based on the CART method (Breiman's decision trees). This method consists in constructing a regression tree in order to predict the value of random variable Y representing a (quality of life) score in function of p input variables $X_{.j}, j \in \{1, \dots, p\}$, which are ordinal variables representing the item from an item bank. This approach shows some advantages : it does not rely on any assumption, it can be adapted to datasets with missing data, it needs less computation time during the administration of the items and it allows us to assess the potential differential item functioning, by adding some confounding factors to the model.

In terms of results, we present the calibration of an item bank measuring mental health-related quality of life for patients with multiple sclerosis. This item bank contains items issued from two questionnaires. We present some simulations for the development of the adaptive version of the questionnaires MusiQoL (respectively SQoL), which is a multidimensional quality of life questionnaires specific to multiple sclerosis (respectively schizophrenia). Finally, we present some simulations to show the interest to use the approach based on decision trees. The "classical" approach, based on item response theory, is compared to our new method, which shows more satisfactory quality of life scores estimates, with regard to a reference score.

Clustering of qualitative data

The second part of this thesis presents a hierarchical clustering method, based on binary decision trees, named *clustering using unsupervised binary trees* (CUBT). CUBT is a top-down method that consists in three stages to get an optimal partition of the data. The first stage aims to construct the "maximal tree" by recursively splitting the dataset in two subsets of observations, the binary splits aim to minimize a heterogeneity criterion, called the *deviance*, computed on the subsets obtained. The *deviance* is based on the covariance matrix of each subset of observations. The second stage consists in pruning the tree, it allows us to aggregate the pairs of terminal nodes of the tree. For each pair of terminal nodes issued from a same binary split, a dissimilarity measure between-nodes is computed. The two nodes are aggregated within a same node (i.e the parent node) if their dissimilarity is below some threshold denoted *mindist*. The dissimilarity measure used in CUBT is based on the euclidean distance. Finally, the third stage allows us to prune the tree, considering the aggregation of terminal nodes that are not issued from a same split. In this stage, we can choose to use either the heterogeneity criterion, or the dissimilarity measure. CUBT has the same advantages as CART, it can produce optimal partitions for a large variety of data structures et shows good convergence properties. It can be interpretable by following the succession of binary splits obtained in the tree.

This property has been highlighted in some application in public health, using CUBT on scores from quality of life questionnaires. This approach allows us to help the clinicians to interpret the data and improve clinical decision-making regarding these scores, the tree structure can be used to easily get threshold-scores to define a categorization of the patients in different subgroups.

One of the technical limitations of the initial version of CUBT is that the heterogeneity criterion and the dissimilarity measure are specific to continuous data. We propose some extensions of this method to adapt it to the case of ordinal and nominal data. We suggest new criteria, based on the mutual information and Shannon's entropy. Different data simulation models are presented to experiment this new version of CUBT and compare it to other unsupervised approaches. We also define some heuristics concerning the choice of the CUBT parameters.

Variable selection in unsupervised learning

We focus on the problem of variable selection in clustering. A classification tree obtained with CUBT can identify the variables that take part directly in the construction of the tree. However, although some variables may be irrelevant for the construction of the tree, they can be competitive of the active ones in different binary splits of the tree. In many applications of data analysis and statistical modeling, it is essential to get a hierarchy of the variables, defined on an importance score, to determine their relevance in a given model. Variable selection allows us to reduce the complexity of the model, to reduce the noise in the data, and hence to gain in model precision and interpretability. We present a method to assess variable importance in clustering. This method, inspired from CART, use CUBT and the notion of competitive binary splits to define a score of variable importance. We analyze the efficiency and the stability of this new index, comparing it with other classical methods of variable importance scoring. We consider data simulation models to compare our approach, adding irrelevant variables in the simulated datasets. This method shows satisfactory results in terms of efficiency and stability. This new criterion may be used to get a ranking of the variables in a dataset, and develop a new algorithm of variable selection.

This thesis resulted in several publications in international revues and conferences. Here is the list :

Publications resulting from the thesis

- Pierre Michel, Badih Ghattas et Laurent Boyer (2016) *Assessing variable importance in clustering. A new method based on unsupervised binary decision trees.* (soumis)
- Pierre Michel, Karine Baumstarck, Christophe Lançon, Badih Ghattas, Anderson Loundou, Pascal Auquier et Laurent Boyer (2016) *Modernizing quality of life assessment : development of a multidimensional computerized adap-*

tive questionnaire for patients with schizophrenia. (en révision)

- Badih Ghattas, Pierre Michel et Laurent Boyer (2016) *Clustering nominal data using Unsupervised Binary decision Trees : Comparisons with the state of the art methods. (en révision)*
- Pierre Michel, Karine Baumstarck, Badih Ghattas, Jean Pelletier, Anderson Loundou, Mohamed Boucekine, Pascal Auquier et Laurent Boyer (2016) *A Multidimensional Computerized Adaptive Short-Form Quality of Life Questionnaire Developed and Validated for Multiple Sclerosis : The MusiQoL-MCAT. Medicine, 95(14) :e3068.*
- Pierre Michel, Pascal Auquier, Karine Baumstarck, Anderson Loundou, Badih Ghattas, Christophe Lançon et Laurent Boyer (2015) *How to interpret multidimensional quality of life questionnaires for patients with schizophrenia ? Quality of Life Research, 24(10).*
- Pierre Michel, Pascal Auquier, Karine Baumstarck, Jean Pelletier, Anderson Loundou, Badih Ghattas et Laurent Boyer (2015) *Development of a cross-cultural item bank for measuring quality of life related to mental health in multiple sclerosis patients. Quality of Life Research, 24(9).*
- Pierre Michel, Karine Baumstarck, Laurent Boyer, Oscar Fernandez, Peter Flachenecker, Jean Pelletier, Anderson Loundou, Badih Ghattas et Pascal Auquier (2014) *Defining Quality of Life Levels to Enhance Clinical Interpretation in Multiple Sclerosis. Medical care.*
- Pierre Michel, Karine Baumstarck, Pascal Auquier, Xavier Amador, Rémy Dumas, Jessica Fernandez, Christophe Lancon et Laurent Boyer (2013) *Psychometric properties of the abbreviated version of the Scale to Assess Unawareness in Mental Disorder in schizophrenia. BMC Psychiatry, 13(1) :229.*

Contributions in international conferences

- Pierre Michel, Karine Baumstarck, Laurent Boyer, Anderson Loundou, Badih Ghattas et Pascal Auquier (2016) *Development of a new quality of life computerized adaptive testing algorithm based on regression trees. 23rd Annual Conference of the International Society for Quality of Life Research, Copenhagen, Denmark.*
- Pierre Michel et Badih Ghattas (2016) *Variable Importance in Clustering Using Binary Decision Trees. COMPSTAT 2016, 22nd International Conference on Computational Statistics, Oviedo, Spain.*

- Pierre Michel, Zeinab Hamidou, Laurent Boyer, Badih Ghattas et Pascal Auquier (2015) *Clustering based on unsupervised binary trees to define levels of symptom severity in cancer*. 22nd Annual Conference of the International Society for Quality of Life Research, Vancouver, Canada.
- Pierre Michel et Badih Ghattas (2014) *Clustering ordinal data using binary decision trees*. COMPSTAT 2014, 21st International Conference on Computational Statistics, Geneva, Switzerland.

Keywords : variable selection, item banking, computerized adaptive testing, clustering, binary decision trees.

Remerciements

Je tiens tout d'abord à remercier les deux rapporteurs de ma thèse, Christophe Biernacki et Iven Van Mechelen, qui ont accepté de relire et juger mon travail.

Je remercie également Liliane Bel, qui a accepté de présider le jury de cette thèse.

Je remercie le professeur Pascal Auquier, qui fait également partie de ce jury, pour m'avoir accueilli dans son laboratoire et pour avoir supervisé ma thèse tout au long de ces trois années, mais également pour m'avoir fait confiance en tant qu'ingénieur d'études, ou encore en tant qu'enseignant. J'ai appris beaucoup de ces quatre années passées à la Timone sous sa direction.

Je remercie Badih Ghattas, mon directeur de thèse, qui est à l'initiative de ce projet. Badih, je vous remercie de m'avoir fait confiance depuis le début, que ce soit pour la thèse, pour me permettre d'obtenir les meilleures conditions pour la mener à bout, mais également pour les années précédentes à l'université où vous avez été un enseignant de référence dans mon parcours.

Je remercie finalement Laurent Boyer, mon co-directeur de thèse, de s'être intéressé à mes travaux, pour m'avoir permis de participer à de nombreuses études, depuis que je le connais. Laurent, merci pour toutes ces opportunités, j'espère que l'on continuera à travailler ensemble sur de nouvelles études à l'avenir. Merci de m'avoir encadré durant cette thèse, et de m'avoir transmis tes compétences de méthodologiste.

Table des matières

Résumé	7
Abstract	12
Remerciements	13
Liste des figures	16
Liste des tableaux	17
1 - Introduction générale	18
1.1 Les mesures subjectives de la qualité de vie	18
1.1.1 La qualité de vie liée à la santé : une variable latente	18
1.1.2 Les questionnaires de qualité de vie	19
1.1.3 Notions de psychométrie	20
1.1.4 Validité d'un questionnaire	21
1.2 La modernisation des mesures	22
1.2.1 La théorie de réponse à l'item	22
1.2.2 Banques d'items et questionnaires adaptatifs	25
1.2.3 Apprentissage non-supervisé sur les données de qualité de vie	27
1.3 Objectifs et structure de la thèse	28
2 - Classification non supervisée de données qualitatives	29
2.1 Introduction	29
2.2 Classification non-supervisée basée sur les arbres binaires	32
2.2.1 Construction de l'arbre maximal	32
2.2.2 Elagage de l'arbre maximal	35
2.2.3 Regroupement des feuilles	37
2.3 Applications aux données de qualité de vie liée à la santé	39
2.3.1 Etude sur des patients atteints de sclérose en plaques	41
2.3.2 Etude sur des patients atteints de schizophrénie	47
2.4 Extensions de la méthode pour des données qualitatives	48
2.4.1 Extension aux données ordinales	48
2.4.2 Extension aux données nominales	48
2.4.3 Expérimentations	50
2.4.4 Application à des données réelles : les maladies du soja	67
2.4.5 Conclusion	68
3 - Sélection de variables en apprentissage non supervisé	71

3.1	Introduction	72
3.2	Importance des variables en apprentissage non supervisé	74
3.2.1	Les forêts aléatoires non-supervisées	74
3.2.2	L'algorithme TW- k -means	74
3.2.3	Le score Laplacien	75
3.2.4	Le score "leave-one-variable-out"	75
3.3	Importance des variables avec la méthode CUBT	76
3.3.1	Rappels sur la méthode CUBT	76
3.3.2	Mesure de l'importance des variables dans CUBT	78
3.4	Expérimentations	79
3.4.1	Stabilité de l'importance des variables	79
3.4.2	Efficacité de l'importance des variables	84
3.4.3	Conclusion	90
4	- Développement de questionnaires adaptatifs	92
4.1	Développement d'une banque d'items mesurant la QV liée à la santé mentale pour des patients atteints de sclérose en plaques	93
4.1.1	Introduction	93
4.1.2	Méthodes	94
4.1.3	Résultats	99
4.1.4	Conclusion	106
4.2	Questionnaires adaptatifs multidimensionnels basés sur la théorie de réponse à l'item	108
4.2.1	Introduction	108
4.2.2	Le MusiQoL-CAT : Développement d'un questionnaire adaptatif multidimensionnel mesurant la qualité de vie chez les patients atteints de sclérose en plaques	108
4.2.3	Le SQoL-CAT : Développement d'un questionnaire adaptatif multidimensionnel mesurant la qualité de vie chez les patients atteints de schizophrénie	109
4.3	Questionnaires adaptatifs basés sur les arbres de décision binaires	127
4.3.1	Introduction	127
4.3.2	Méthodes	129
4.3.3	Résultats	132
4.3.4	Conclusion	135
	Conclusion et perspectives	138

Liste des figures

1.1	Exemple de courbes caractéristiques d'items suivant le modèle Rasch.	24
1.2	Schéma d'un questionnaire adaptatif.	27
2.1	Exemple de division binaire d'un noeud t en deux noeuds enfants t_g et t_d , obtenue avec CUBT.	33
2.2	Exemple d'arbre maximal obtenu avec la première étape de CUBT (<i>growing</i>).	35
2.3	Utilité de la troisième étape de CUBT (<i>joining</i>).	37
2.4	Structure de l'arbre correspondant à la partition donnée pour illustrer le regroupement des feuilles.	38
2.5	Structure à trois classes.	45
2.6	Scores des dimensions du MusiQoL, pour chaque classe de niveau de qualité de vie.	45
2.7	Représentation graphique de la structure à trois classes.	46
2.8	Structure de l'arbre utilisée pour le modèle de simulation M2.	54
2.9	Relation entre la déviance et l'utilité des classes (en utilisant l'information mutuelle comme mesure de dissimilarité).	59
2.10	Relation entre la déviance et l'utilité des classes (en utilisant la distance de Hamming comme mesure de dissimilarité).	60
2.11	Relation entre les taux d'erreur obtenus avec la distance de Hamming et l'information mutuelle.	62
2.12	Taux d'erreur de mauvaise classification en fonction de la valeur du paramètre <i>minsize</i> .	64
2.13	Structures des arbres obtenus avec CUBT (gauche) et DIVCLUS-T (droite) pour le jeu de données Soybean.	68
3.1	Structure de l'arbre obtenu par CUBT avec le jeu de données Iris.	73
3.2	Résultats de stabilité pour le jeu de données Toys.	81
3.3	Résultats de stabilité pour le jeu de données Toys.	82
3.4	Importance de chacune des 4 variables du jeu de données Iris en fonction des valeurs du <i>minsize</i> .	83
3.5	Importance des variables du jeu de données Iris pour une valeur optimale <i>minsize</i> = 16.	84
3.6	Modèle de simulation M1.	85
3.7	Modèle de simulation M2.	86
3.8	Modèle de simulation M2.	87
4.1	Courbe de l'alpha de Cronbach.	101
4.2	Courbes caractéristiques des 22 items de la banque.	102

4.3	Distribution du score issu de la banque d'items et courbe d'information de la banque.	104
4.4	Distribution des scores issus des 41 items du SQoL.	118
4.5	Taux d'exposition des items lors des simulations.	122
4.6	Structure de l'arbre obtenu avec la méthode CART.	130
4.7	Questionnaire adaptatif basé sur la théorie de réponse à l'item.	133

Liste des tableaux

2.1	Caractéristiques de la population.	44
2.2	Caractéristiques cliniques et sociodémographiques des trois classes.	47
2.3	Résultats de prédiction	66
2.4	Résultats d'apprentissage	69
3.1	Les quatre variables du jeu de données Iris et leurs scores d'importance respectifs.	73
3.2	Taux de vrais positifs (TVP) et plus hauts rangs (PHR) pour chaque modèle de simulation de données et chaque méthode	89
3.3	Taux de vrais positifs (TVP) et plus hauts rangs (PHR) pour le jeu de données Toys.	90
4.1	Analyse descriptive des items de la banque.	100
4.2	Etape de recodage et qualité d'ajustement du modèle.	102
4.3	Seuils de difficulté des items et ajustement des items après recodage.	103
4.4	Résultats relatifs au comportement différentiel des items.	105
4.5	Comparaison des scores de qualité de vie liée à la santé mentale en fonction de données sociodémographiques, clinique, et de qualité de vie.	106
4.6	Contenu des 41 items du SQoL en français et en anglais.	111
4.7	Caractéristiques descriptives de l'échantillon d'étude.	116
4.8	Caractéristiques des items.	117
4.9	Matrice de corrélations des huit dimensions du SQoL.	118
4.10	Comportement différentiel des items en fonction du genre, de la présence de symptômes paranoïdes et de l'insight.	120
4.11	Score moyen, indicateurs de précision et nombre d'items moyen administrés pour chaque niveau de précision testé.	121
4.12	Validité externe du SQoL-MCAT.	123
4.13	Comparaisons des deux approches.	134

1- Introduction générale

Dans ce chapitre introductif, nous présentons certains concepts essentiels à la bonne compréhension du manuscrit, notamment les concepts liés à la qualité de vie, qui est le domaine d'application de cette thèse. Dans un premier temps, nous définissons le concept de qualité de vie, et ses mesures dites subjectives. Nous nous intéressons aux questionnaires de qualité de vie, leurs propriétés psychométriques et les indicateurs permettant d'évaluer leur validité clinique. Dans un second temps, nous nous penchons sur les méthodes modernes permettant d'améliorer la qualité de ces mesures, via l'utilisation des modèles de réponse à l'item, permettant de définir des ensembles homogènes d'items mesurant des concepts identiques (le développement de banques d'items), ainsi que des algorithmes d'administrations adaptatifs. Nous considérons également l'utilisation de l'apprentissage non-supervisé comme outil d'aide à la décision clinique, ou comme moyen d'améliorer l'interprétabilité des données de qualité de vie. Enfin, nous présentons les différents objectifs de cette thèse.

1.1 Les mesures subjectives de la qualité de vie

La qualité de vie est un concept "flou", non-observable objectivement (contrairement aux mesures physiques habituelles). Pour déterminer les attributs de la qualité de vie, on a cependant besoin d'éléments observables, subjectifs ou objectifs. C'est le cas d'autres mesures subjectives, telles que la douleur, la satisfaction ou encore le bien-être.

1.1.1 La qualité de vie liée à la santé : une variable latente

La qualité de vie liée à la santé d'un patient est une *variable latente*, elle n'est pas observable et ne peut donc pas être mesurée directement. Ce type de variable est communément appelé *trait latent*. Les mesures de la qualité de vie, ou plus généralement les résultats auto-rapportés par les patients (en anglais *patient-reported outcomes*), sont des outils de mesure qui proviennent directement de l'évaluation de la santé des patients ou de l'efficacité et de la tolérance des médicaments par le patient lui-même [3]. Ces dernières années, des progrès significatifs ont été observés dans l'utilisation de ces mesures dans les essais cliniques. L'agence américaine des produits alimentaires et médicamenteux (*Food and Drug Administration*, FDA), a officiellement reconnu la qualité de vie comme un critère de jugement essentiel. L'appréciation de la qualité de vie peut reposer sur des approches qualitatives, souvent peu opérationnelles, ou bien quantitatives, par le biais de questionnaires standardisés [17].

1.1.2 Les questionnaires de qualité de vie

La plupart des instruments de mesures de la qualité de vie couramment utilisés sont des questionnaires "papier-crayon". Ces questionnaires contiennent un nombre fixe de questions, appelées *items*. Ces questionnaires peuvent aborder différents concepts sous-jacents de la qualité de vie, c'est-à-dire des *dimensions* de la qualité de vie. Ils ont pour but de fournir aux cliniciens un score reflétant le niveau de qualité de vie des patients, ou un ensemble de sous-scores reflétant les différents niveaux de qualité de vie relatifs aux dimensions considérées. Dans le but d'assurer une comparabilité suffisante de ces scores entre les patients, il est demandé aux patients de répondre à tous les items du questionnaire. Cependant, dans la pratique clinique, l'administration de certains items inadaptés (par leur formulation, leur difficulté) peut provoquer un sentiment de fardeau chez les patients et les réponses à certains items peuvent être biaisées (réponses extrêmes, non-informatives) ou bien manquantes. La principale approche permettant de répondre à ce problème est le développement de formes courtes des questionnaires existants. A partir de réponses observées contenus dans un ou plusieurs échantillons de patients, il est possible de réduire la taille d'un questionnaire en sélectionnant uniquement les items les plus informatifs au regard de certaines propriétés psychométriques.

Nous fournissons ici deux exemples de questionnaires mesurant la qualité de vie liée à la santé, qui sont spécifiques à deux pathologies : la sclérose en plaques et la schizophrénie. Nous décrivons également un des questionnaires génériques les plus largement utilisés dans les études de la qualité de vie, le questionnaire SF-36. Ces trois questionnaires sont disponibles en langue française.

1.1.2.1 Le questionnaire MusiQoL

Le MusiQoL [155], développé au laboratoire de santé publique de Marseille, est un questionnaire validé, spécifique à la sclérose en plaques, qui comprend 31 items, et qui fournit un score global de qualité de vie et neuf sous-scores représentant les dimensions suivantes : activités de la vie quotidienne (8 items), bien-être psychologique (4 items), symptômes (3 items), relations avec les amis (4 items), relations avec la famille (3 items), relation avec le système de soins (3 items), vie sentimentale et sexuelle (2 items), coping (2 items) et rejet (2 items). Le score de chaque dimension est obtenu en calculant la moyenne des scores d'items, les réponses manquantes étant remplacées par la moyenne des réponses non-manquantes. Dans ce questionnaire les scores varient de 0 à 100, 100 représentant le meilleur niveau de qualité de vie.

1.1.2.2 Le questionnaire SQoL

Le SQoL [12] est un questionnaire validé, spécifique à la schizophrénie, qui comprend 41 items, fournissant un score global de qualité de vie, ainsi que huit

sous-scores représentant les différentes dimensions suivantes : bien-être psychologique (10 items), estime de soi (6 items), relations avec la famille (5 items), relations avec les amis (5 items), résilience (5 items), bien-être physique (4 items), autonomie (4 items) et la vie sentimentale (2 items). Les scores varient de 0 à 100, 100 étant le score maximum, représentant le meilleur niveau de qualité de vie.

La version courte de ce questionnaire, le SQoL-18 [28], comprend 18 items, qui décrivent les huit mêmes dimensions : bien-être psychologique (3 items), estime de soi (2 items), relations avec la famille (2 items), relations avec les amis (2 items), résilience (3 items), bien-être physique (2 items), autonomie (2 items) et la vie sentimentale (2 items).

1.1.2.3 Le questionnaire SF-36

Le SF-36 [167, 168, 103] est un questionnaire générique qui décrit huit dimensions : fonctionnement physique (10 items), fonctionnement social (2 items), limitations dues à l'état physique (4 items), limitations dues à l'état émotionnel (3 items), santé mentale (5 items), vitalité (4 items), souffrance physique (2 items) et santé globale (5 items). Deux scores composites (physique et mental) sont calculés en fonction des différents scores de dimensions. Le questionnaire SF-36 fournit des scores sur une échelle de 0 à 100 (0 étant le score représentant le plus faible niveau de qualité de vie, 100 le plus élevé).

1.1.3 Notions de psychométrie

On considère ici que l'on dispose d'un questionnaire comprenant p items, que l'on a administré à un ensemble de n individus. On note X_i le vecteur contenant les p réponses de l'individu i , $i \in \{1, \dots, n\}$, et $X_{\cdot j}$ le vecteur contenant les n réponses à l'item j , $j \in \{1, \dots, p\}$. En psychométrie classique [111], le score de l'individu i , noté X_{ij} , est la réponse fournie par ce dernier à l'item j d'un questionnaire. Ce score observé est considéré comme une somme de deux termes, un vrai score T_{ij} , et une erreur ϵ_{ij} . On considère donc un modèle linéaire :

$$X_{ij} = T_{ij} + \epsilon_{ij}, \quad (1.1)$$

avec $E(\epsilon_{ij}) = 0$ et $cov(X_{ij}, T_{ij}) = 0$.

Une mesure essentielle étudiée en psychométrie est la *fiabilité* du score, définie comme le rapport entre la variance du vrai score $\sigma_{T_{\cdot j}}^2$ et la variance du score observé $\sigma_{X_{\cdot j}}^2$, pour $X_{\cdot j}$, elle est notée :

$$\rho_j^2 = \frac{\sigma_{T_{\cdot j}}^2}{\sigma_{X_{\cdot j}}^2} \quad (1.2)$$

L'indice de fiabilité le plus communément utilisé dans l'analyse des données

de qualité de vie est le coefficient α de Cronbach [49]. On définit le score total comme un vecteur $T = \sum_{j=1}^p X_{.j}$ (qui contient la somme des scores de chaque individu) et le coefficient α de Cronbach est défini de la manière suivante :

$$\alpha = \frac{p}{p-1} \left(1 - \frac{\sum_{j=1}^p \sigma_{X_{.j}}^2}{\sigma_T^2} \right), \quad (1.3)$$

où σ_T^2 est la variance des scores totaux observés.

Cet indice est utilisé lors de la validation psychométrique d'un instrument de mesure. Dans la pratique clinique, un coefficient supérieur à 70% est attendu pour établir la bonne fiabilité d'un questionnaire [35]. Une extension a été développée [94], elle consiste à calculer le coefficient en enlevant de manière séquentielle l'item dont la suppression résulte en une augmentation maximale du coefficient. Cette approche a été proposée pour valider l'hypothèse d'unidimensionnalité qui est vérifiée lors de la validation d'un questionnaire. L'unidimensionnalité est une hypothèse importante qui consiste à supposer que le trait latent mesuré par les items est un scalaire (et non un vecteur, ce qui reflèterait un trait latent multidimensionnel).

1.1.4 Validité d'un questionnaire

D'autres indicateurs existent pour valider un questionnaire, la plupart sont basés sur une analyse descriptive des items (taux de données manquantes faibles, effets plancher et plafond, asymétrie de la distribution des modalités). La validation des différentes dimensions (si le questionnaire est multidimensionnel) est basée sur la contribution des items vis-à-vis de leur dimension respective. Pour cela on peut vérifier que le score de chaque item est plus corrélé avec le score de sa dimension respective (en omettant cet item du calcul) qu'avec les scores des autres dimensions. On parle alors de *cohérence interne* des items lorsque la corrélation du score de l'item avec celui de sa dimension est supérieur à 0.4 [35]. On peut également vérifier que le score d'un item est plus corrélé avec le score d'un item de sa dimension qu'avec les scores des items des autres dimensions. On parle alors de *validité discriminante* des items [34]. La dimensionnalité d'un questionnaire peut également être étudiée en utilisant des méthodes d'analyse factorielle exploratoire [58] (par exemple l'analyse en composantes principales) ou confirmatoire (par exemple les modèles d'équations structurelles [96]) afin de représenter les différents facteurs, et obtenir une classification des différents items en dimensions. Les méthodes précédemment citées sont utilisées pour vérifier la validité interne d'un questionnaire.

Il existe un autre aspect de la validité d'un questionnaire, qui permet d'évaluer le comportement du score en fonction de variables externes. On parle alors de *validité externe*. Dans la pratique clinique, ces variables sont de type socio-démographique, clinique, ou bien elles peuvent correspondre à des scores issus

d'autres questionnaires. En corrélant les scores du questionnaire à valider avec des scores issus d'autres questionnaires, mesurant des dimensions similaires, on peut vérifier que le questionnaire mesure bien les dimensions qu'il est sensé mesurer. On parle alors de *validité convergente*. De la même manière, on peut également vérifier que les scores des questionnaires ne mesurent pas ce qu'ils ne sont pas sensés mesurer. On parle alors de *validité divergente*. Ces deux aspects sont également basés principalement sur l'analyse de corrélations [142].

Un exemple de validation psychométrique d'un instrument de mesure est proposé [119], il décrit les propriétés psychométriques d'un instrument de mesure de l'*insight* en schizophrénie, c'est-à-dire le degré de conscience de la maladie du patient. Cette étude a permis de valider une forme courte du test SUMD (*scale to assess unawareness in mental diseases*), et a fait l'objet d'une publication [119]. Bien que ce questionnaire ne soit pas auto-rapporté (en effet c'est le clinicien qui évalue la conscience de la maladie du patient en répondant lui-même aux items), les méthodes de validité restent les mêmes que celles présentées ci-dessus.

1.2 La modernisation des mesures

Les notions abordées dans la section 1.1.3 sont issues de la théorie dite classique, en opposition à la théorie moderne. Les premiers questionnaires standardisés ont été validés métriquement sur la base de stratégies méthodologiques discutables. Les questionnaires les plus récents ont recouru à des techniques plus adaptées basées sur la théorie classique de la mesure [48, 131] ainsi que sur une approche plus moderne, la *théorie de réponse à l'item* [82, 106, 26, 56, 144]. La théorie de réponse à l'item offre plus d'information concernant les items et plus de flexibilité sur la manière d'obtenir un score à partir d'un ensemble d'items.

1.2.1 La théorie de réponse à l'item

La théorie de réponse à l'item [14, 106] désigne un ensemble de modèles paramétriques ou non-paramétriques, très utilisés en sciences éducationnelles et en psychologie, pour évaluer la façon dont les individus répondent aux items d'un questionnaire [161]. Dans la suite du document, nous considérerons uniquement les modèles de réponse à l'item paramétriques. Ces modèles permettent de modéliser la probabilité de réponse à un item en fonction du trait latent d'un individu (noté θ) à l'aide d'un lien logistique. On distingue généralement quatre paramètres qui définissent les items dans ces modèles : la difficulté de l'item, le pouvoir discriminant de l'item, et deux paramètres de pseudo-hasard (qui ne sont pas pris en compte dans les applications en qualité de vie). Les modèles de réponse à l'item sont des extensions du modèle de Rasch [141], qui est un modèle à un paramètre adapté aux items dichotomiques (c'est-à-dire des items

à deux modalités de réponses), défini de la façon suivante :

Soit $X_{ij} \in E = \{0, 1\}$ une variable aléatoire dichotomique, on définit la probabilité qu'un individu i réponde correctement à l'item j (c'est-à-dire $X_{ij} = 1$) comme suit :

$$P(X_{ij} = 1) = \frac{e^{\theta_i - \beta_j}}{1 + e^{\theta_i - \beta_j}}, \quad (1.4)$$

où θ_i est le niveau de trait latent de l'individu i et β_j est la difficulté de l'item j . Cette probabilité (qui est une fonction de θ_i) peut être représentée graphiquement pour chaque item en utilisant la courbe caractéristique de l'item. Un exemple de courbes caractéristiques d'items est fourni à la figure 1.1. Pour obtenir ces courbes, nous avons généré 10000 réponses dichotomiques pour cinq items avec $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (-2, -1, 0, 1, 2)$.

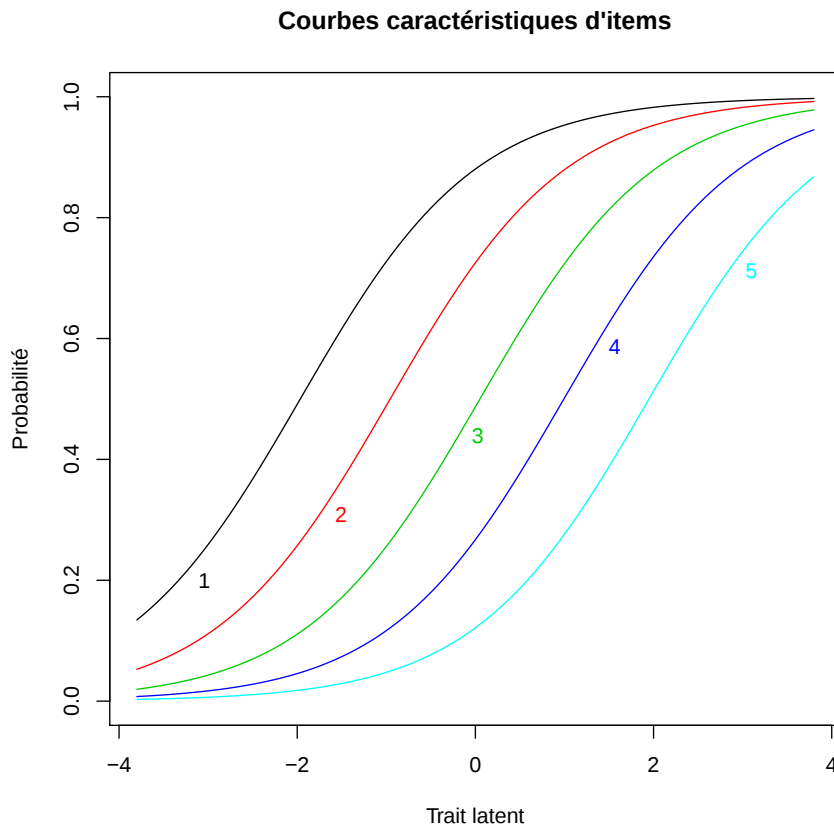


Figure 1.1 – Exemple de courbes caractéristiques d'items suivant le modèle Rasch. Les cinq items ont des paramètres de difficulté différents, avec $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) = (-2, -1, 0, 1, 2)$.

La courbe caractéristique d'un item dichotomique représente la probabilité de répondre positivement à un item en fonction du trait latent. Ainsi, on peut illustrer la hiérarchie d'un groupe d'items grâce à leurs paramètres de difficulté. La valeur du paramètre de difficulté est égale à la valeur de trait latent pour lequel un individu a 50% de chance de répondre correctement à l'item, autrement dit, c'est l'abscisse correspondant au point d'inflexion du sigmoïde. Cette notion de difficulté s'illustre intuitivement quand il s'agit d'étudier des items mesurant un domaine physique unidimensionnel de la QV (telles que la fatigue ou le fonctionnement physique d'un individu) comme on peut s'en apercevoir dans de nombreuses études [85]. Les paramètres de discrimination (notés α_j) sont représentés par la pente de la courbe au niveau du point d'inflexion, c'est à dire que plus la pente est élevée plus l'item est discriminant. Dans l'exemple de la figure, les paramètres de discrimination des items sont fixés à 1.

Dans le cas des items polytomiques (c'est-à-dire des items à plus de deux modalités de réponses), on ne parle plus de paramètres de difficulté mais de seuils de difficulté. Ces seuils correspondent à des niveaux de traits latents où la probabilité de répondre une certaine modalité est égale à la probabilité de répondre à la modalité suivante. En d'autres termes, un item ayant m modalités de réponse a $m - 1$ seuils. Des modèles de réponse pour items polytomiques tels que le modèle de crédit partiel [116] ou le modèle d'échelle de notation [10] permettent de prendre en compte ces seuils de difficulté.

Le fait de ne mesurer qu'un trait latent unidimensionnel limite l'interprétation des scores obtenus et ne permet pas de prendre en compte les interactions qu'il peut y avoir entre les différentes dimensions qui découlent de la qualité de vie. Dans un cadre plus technique, les méthodes utilisées pour mesurer un concept unidimensionnel sont déjà bien connues dans le domaine de la psychométrie et reposent sur des hypothèses fondamentales (unidimensionnalité, monotonie, indépendance locale). En revanche les modèles IRT prenant en compte le caractère polytomique, ordinal et multidimensionnel des items d'un questionnaire, représentent encore aujourd'hui un large champ de recherche et les méthodes statistiques d'estimation utilisées dans ce cadre sont encore à développer et à améliorer.

La théorie de réponse à l'item multidimensionnelle, qui s'affranchit des trois hypothèses fondamentales (unidimensionnalité, indépendance locale, monotonie), est une extension récente de la théorie de réponse à l'item, où l'on suppose que les réponses aux items dépendent de plus d'un trait latent [106]. On y fait l'hypothèse que le trait latent mesuré est un vecteur de dimension supérieure à 1, chaque composante du vecteur représentant un score d'une dimension. La plupart des modèles cités précédemment peuvent être adaptés au cas multidimensionnel; on dispose ainsi des mêmes paramètres d'items, spécifiques à chaque dimension. Cette approche est très utile si l'on veut estimer des traits latents multiples, correspondant à des scores de dimensions expliquant un même trait latent multidimensionnel (la qualité de vie par exemple).

1.2.2 Banques d'items et questionnaires adaptatifs

Dans la littérature, de nombreuses démarches ont été entreprises ces dernières années pour mettre en application la théorie de réponse à l'item pour développer des banques d'items. Les banques d'items sont des collections d'items qui sont calibrées, c'est-à-dire qui sont sous-jacentes à un modèle paramétrique, et qui mesurent tous un même trait latent. Par définition, une banque d'items est unidimensionnelle. Cependant, en fonction de la multidimensionnalité d'un trait latent (typiquement la qualité de vie), on peut développer une banque d'items pour chaque dimension considérée. Par exemple, un projet américain (PROMIS, [38]) oeuvre à développer depuis plusieurs années des banques d'items mesurant différents domaines de la qualité de vie, générique ou spécifique à différentes pathologies ou différentes populations. Les items servant à alimenter une banque d'items peuvent être créés par un comité d'experts du domaine. Ils peuvent aussi être issus de questionnaires existants, mesurant le même concept que la banque d'items. Une fois qu'un ensemble initial d'items est récolté, la calibration consiste à sélectionner un ensemble d'items candidats présentant des propriétés psychométriques satisfaisantes, ou supprimer un certain nombre d'items non pertinents. Cette démarche est très proche et repose sur les mêmes fondements que la validation d'un questionnaire. Cependant, dans une démarche de développement de banques d'items, la théorie de réponse à l'item a l'avantage de fournir aux cliniciens une mesure précise de la difficulté des items, leur pouvoir discriminant et leur proportion d'information totale. L'estimation des paramètres de ces modèles est basée sur des méthodes de maximisation de la vraisemblance [95] ou sur des approches bayésiennes [110]. La principale application que l'on peut faire des banques d'items calibrées consiste à développer des questionnaires adaptatifs. Dans ce cadre, on cherche à évaluer la structure factorielle de la banque d'items ainsi que la qualité d'ajustement d'un modèle de réponse à l'item [177].

L'administration adaptative d'un sous-ensemble d'items contenus dans une banque s'appuie tout d'abord sur une population de référence. Cette population de référence correspond aux individus qui ont permis de développer la banque d'items. La mise en place de la procédure adaptative repose sur les choix méthodologiques faits lors de l'étape de calibration de la banque. La sélection de l'item à administrer lors de la procédure adaptative est souvent basée sur la théorie de réponse à l'item. L'algorithme d'administration adaptative vise à maximiser l'information [170], et est également possible par le biais des réseaux bayésiens [51].

Trois critères majeurs sont à prendre en compte lors de la procédure adaptative. Ces trois critères sont : un critère initial ("quel item doit-être administré en premier?"), un critère de sélection ("en tenant compte des réponses précédentes, quel item doit être administré?") et enfin un critère d'arrêt ("quand, et selon quel critère la procédure adaptative doit-elle s'arrêter?"). Ces différents critères peuvent être fixés et comparés afin d'évaluer les avantages et faiblesses des

différents modes d'administration [136, 45]. On peut schématiser en détails la procédure d'administration d'un questionnaire auto-adaptatif en six étapes [22], comme le montre la figure 1.2.

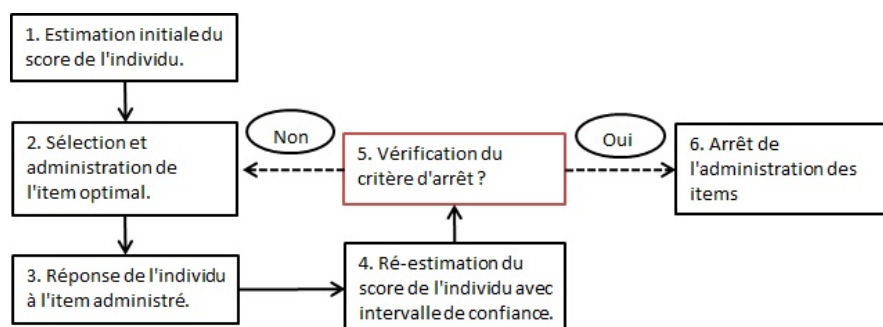


Figure 1.2 – Schéma d'un questionnaire adaptatif.

1.2.3 Apprentissage non-supervisé sur les données de qualité de vie

L'apprentissage non-supervisé désigne un ensemble de méthodes statistiques où la variable Y à expliquer est absente, ces méthodes pouvant être utilisés pour estimer la densité d'une distribution empirique ou encore classifier un ensemble d'observations. La sélection de variables non supervisée, dans le cadre des données de qualité de vie constitue un élément majeur dans l'analyse de ces données. Les variables explicatives sont associées aux items des questionnaires. La détection des items les plus pertinents pour évaluer la qualité de vie peut améliorer le choix des items à administrer à un patient donné. De plus, les questionnaires utilisés sont souvent constitués de sous-groupes de questions décrivant des dimensions différentes (mais souvent complémentaires) de la qualité de vie. L'interprétation de l'aspect multidimensionnel de la qualité de vie est un défi pour la recherche en santé publique puisque les méthodes statistiques et psychométriques existantes (originellement utilisées dans d'autres domaines tels que les sciences éducationnelles) ont connu de grandes avancées durant ces dernières décennies, à l'image des modèles de la théorie de réponse à l'item qui fait partie des méthodes psychométriques modernes les plus utilisées dans ce champ.

Un score utilisé pour évaluer la qualité de vie est modélisé en fonction des items sur lesquels son calcul est basé. Les items sont les variables explicatives, et il peut être intéressant d'évaluer l'apport de chaque item à un modèle donné, ou encore d'évaluer une hiérarchie sur l'ensemble des items. L'interprétabilité des scores de qualité de vie est souvent fastidieuse dans la pratique clinique, et les études utilisent souvent des scores-seuil pour classifier les patients en différents niveaux de qualité de vie, ce qui améliore l'interprétabilité des scores. Cependant, ces scores-seuils sont souvent définis sur des populations de référence

utilisées pour valider les questionnaires, dont la représentativité vis-à-vis d'un échantillon d'étude peut être discutable. Ils sont donc très souvent inadaptés en pratique et il est nécessaire de proposer des alternatives sur la production de tels scores-seuils. Il est envisageable de définir des scores-seuils à partir d'une population de référence en utilisant des méthodes de classification non-supervisée. De plus les méthodes de classification non supervisées hiérarchiques peuvent permettre d'obtenir une partition des individus plus interprétable.

1.3 Objectifs et structure de la thèse

Cette thèse a trois objectifs principaux, chaque chapitre vise à proposer une ou plusieurs contributions permettant de répondre à chacun d'eux.

Dans un premier temps, le chapitre 2 présente une méthode récente de classification non supervisée basée sur les arbres de décision binaires (appelée CUBT, [64]). Nous verrons par le biais d'applications sur des données de qualité de vie que cette méthode d'apprentissage non supervisée peut améliorer l'interprétabilité des scores de qualité de vie. Nous proposerons également une extension de cette méthode pour le cas des données qualitatives.

Ensuite, le chapitre 3 vise à proposer une méthode permettant de sélectionner des variables dans un contexte non supervisé, par le biais d'un score d'importance des variables. Ce score sera basé sur la méthode CUBT. Nous analyserons sa stabilité ainsi que son efficacité au travers de diverses simulations. Ses performances seront également comparées avec celles de méthodes plus classiques de hiérarchisation de variables en classification automatique.

Enfin, nous cherchons à proposer une alternative à la théorie de réponse à l'item pour le développement de questionnaires adaptatifs. Le chapitre 4 porte sur plusieurs applications de méthodes utilisées dans le champ de la qualité de vie pour développer des banques d'items et des questionnaires adaptatifs. Il porte également sur la présentation d'une nouvelle méthode non-paramétrique pour développer un algorithme d'administration adaptative d'items, basée sur les arbres de décision binaires. Chaque section de ce chapitre est autonome et restitue certains résultats issus d'articles rédigés pendant la thèse.

2- Classification non supervisée de données qualitatives

Dans ce chapitre, nous nous intéressons aux méthodes de classification non supervisées applicables aux données qualitatives. En apprentissage statistique, la classification non supervisée (appelée communément *clustering*) consiste à créer une partition des données, c'est-à-dire à classer un ensemble d'individus en différents sous-groupes, de manière à ce que chaque sous-groupe obtenu (appelé *cluster*) soit aussi homogène que possible, au regard d'une certaine mesure de dissimilarité. La majorité des méthodes de clustering consistent à construire une partition d'un ensemble de n observations en k clusters, k pouvant être spécifié a priori ou automatiquement déterminé par la méthode.

Nous proposons une extension d'une méthode récente, *clustering using unsupervised binary trees* (CUBT, [64]), pour les données qualitatives. CUBT est une méthode de clustering hiérarchique adaptée aux données continues, inspirée par la méthode CART, qui utilise trois étapes pour estimer une partition optimale des données. La division des noeuds est basée sur un critère de type covariance. L'étape d'élagage utilise une mesure de dissimilarité robuste basée sur la distance euclidienne.

Dans un premier temps, nous décrivons la méthode CUBT, initialement prévue pour des données continues. Cette méthode a été appliquée dans différentes études en santé publique. Nous présentons les principaux résultats de ces travaux. Ensuite, nous proposons une extension au cas des données ordinales et nominales. Nous utilisons ici des critères d'hétérogénéité et des mesures de dissimilarités basés sur l'information mutuelle, l'entropie de Shannon, et la distance de Hamming. Nous proposons et justifions certains choix de paramètres pour la méthode CUBT basés sur des heuristiques.

2.1 Introduction

Les méthodes de partitionnement de données ont pour objectif de constituer des groupes d'observations les plus homogènes possibles, c'est-à-dire regrouper des individus dans des groupes de façon à ce que les individus d'un même groupe soient plus similaires entre eux que par rapport aux individus d'un autre groupe. Ces méthodes visent donc à maximiser l'homogénéité des individus au sein d'un groupe tout en maximisant l'hétérogénéité entre les groupes. On distingue les méthodes de classification dites supervisées de celles dites non supervisées [84]. Dans le premier cas, les labels de classe (c'est-à-dire le groupe d'appartenance des individus) sont connus. Cela permet notamment d'évaluer le taux de mau-

vais classement. Un algorithme hiérarchique de classification supervisée bien connu est CART [29]. En revanche, les labels ne sont pas connus en classification non supervisée. Il est donc plus délicat d'évaluer la qualité d'une méthode étant donné qu'il n'existe pas de critère universel, seulement des heuristiques.

Il y a deux principaux types de méthodes de clustering : les méthodes hiérarchiques et les méthodes non hiérarchiques. Les algorithmes de clustering hiérarchiques [128] permettent d'obtenir un ensemble de clusters qui peut être représenté sous la forme d'un arbre, appelé dendrogramme. Bien souvent, cet arbre est binaire, et chaque noeud de l'arbre (à l'exception de la racine de l'arbre, qui contient l'échantillon d'apprentissage) représente l'union de deux classes. Un exemple récent d'algorithme de clustering hiérarchique, appelé DIVCLUS-T, est proposé par [41]. Cet algorithme repose sur la minimisation d'un critère d'inertie, comme l'algorithme classique de classification hiérarchique avec le critère de Ward. La principale différence avec ce dernier est qu'il permet de produire une interprétation naturelle et intuitive des clusters obtenus, en utilisant des valeurs-seuils sur les variables qui expliquent les clusters.

Les méthodes non hiérarchiques (ou partitionnelles) ont pour but de définir les k classes simultanément. Il y a deux types de méthodes non hiérarchiques : les approches de partitionnement, parmi lesquelles la plus populaire n'est autre que k -means [113], et les approches basées sur la densité. La méthode k -means produit une partition des données en k clusters, qui minimise la variance intra-classe. Les méthodes basées sur la densité visent à construire des régions d'observations de haute densité, séparées par des régions de faible densité. Ces méthodes sont particulièrement utiles pour des données contenant du bruit ou des points aberrants. Un algorithme de ce type bien connu est DBSCAN [57]. D'autres méthodes de clustering existent (voir par exemple [147]), basées sur différents principes, telles que les approches probabilistes, dans lesquelles les clusters sont modélisés par un mélange de distributions paramétriques, permettant d'obtenir des probabilités d'appartenance aux clusters (voir par exemple [50]), comme les méthodes de clustering basées sur l'analyse des classes latentes (latent class analysis, LCA [163]). De la même façon, il existe des méthodes de clustering dites "floues", dans lesquelles chaque observation d'un échantillon d'apprentissage peut appartenir à plus d'un cluster à la fois (voir par exemple l'algorithme c -means flou [54]). Récemment, différentes méthodes de clustering ont été proposées par de nombreux auteurs [135, 134, 65, 165, 20], voir notamment [66] qui propose une revue de la littérature consacrée aux méthodes de clustering robustes.

Dans ce chapitre, nous nous focalisons sur les méthodes de clustering hiérarchiques basées sur les arbres de décision. A notre connaissance, les premiers travaux concernant l'utilisation des arbres de décision pour le clustering est l'algorithme C0.5 [138], basé sur l'induction descendante d'arbres de décision logiques [23]. Ce type d'arbres est communément appelé "arbres de clustering prédictifs". Ils sont basés sur l'algorithme de classification C4.5 de Quinlan [137]

et ont été largement utilisés dans la littérature. Un autre algorithme de clustering basé sur les arbres de décision, CLTree [108], est parfaitement capable d'ignorer les points aberrants. Pour cela, l'algorithme introduit des points "non-existants" dans l'ensemble des données, mais il n'est applicable qu'à des jeux de données continues. Un autre algorithme, plus récent, exclusivement applicable aux jeux de données nominales est nommé CG-CLUS [181]. Cet algorithme possède des propriétés satisfaisantes en termes d'efficacité et de performance, comparé à l'algorithme bien connu COBWEB [60]. Toutes les méthodes basées sur les arbres de décision que nous venons de citer se réfèrent à une catégorie de méthodes appelée classification non supervisée conceptuelle (terme très employé en apprentissage automatique). Le clustering conceptuel vise à former des groupes d'objets (ou individus) qui sont représentés par des paires de valeurs-attributs (c'est-à-dire des variables descriptives nominales), et ne prend en compte aucune notion d'ordinalité des modalités des variables. L'avantage principal des ces approches réside dans l'interprétabilité directe des arbres résultants. Dans ce cadre, une mesure courante, l'utilité des classes (*category utility*) [47] souvent utilisée en théorie de l'information, permet de mesurer la qualité d'une partition. Cette mesure est formellement équivalente à l'information mutuelle [73].

CUBT est une méthode de clustering hiérarchique descendante inspirée par la méthode CART [29] qui s'articule en trois étapes. La première étape (*growing*) permet de construire un "arbre maximal", en divisant de manière récursive l'échantillon d'apprentissage en plusieurs sous-ensembles d'observations, et en minimisant un critère d'hétérogénéité. Ce critère d'hétérogénéité, appelé la *déviance*, est basé sur la trace de la matrice variance-covariance de chaque sous-ensemble d'observations. Cette étape permet de minimiser la déviance au sein des clusters finaux. La deuxième étape permet d'élaguer l'arbre, c'est à dire réduire le nombre de feuilles qu'il contient ; les feuilles sont les noeuds terminaux qui se situent à la fin de chaque branche de l'arbre de décision. Pour chaque paire de feuilles issues d'un même noeud parent, une mesure de dissimilarité entre les deux feuilles est calculée. Si cette mesure est inférieure à un certain seuil noté *mindist*, alors les deux feuilles sont agrégées au sein d'une nouvelle même feuille (i.e le noeud parent). La mesure de dissimilarité utilisée dans CUBT est basée sur la distance euclidienne. La dernière étape (*joining*) est également une étape d'élagage, dans laquelle la contrainte de parenté directe des feuilles n'est plus requise. Dans cette étape, l'utilisateur peut utiliser soit le critère d'hétérogénéité utilisé à la première étape, soit la mesure de dissimilarité utilisée à la deuxième étape.

CUBT est une méthode de clustering très simple qui possède des propriétés similaires à celles de CART. En effet, CUBT est flexible et efficace, c'est-à-dire que cette méthode produit un bon partitionnement pour une grande famille de structures de données, elle est interprétable (du fait des divisions binaires obtenues) et elle possède de bonnes propriétés de convergence, c'est-à-dire la partition effectuée sur l'échantillon converge vers la partition de la population dont est issu

l'échantillon.

Contrairement à CART, où le critère de division prend en compte les labels des observations, dans CUBT la recherche de la meilleure division d'un noeud de l'arbre n'utilise que l'information contenue dans les données observées. De plus, la phase d'élagage de l'arbre dans CUBT se fait en deux étapes (*pruning* et *joining*). La section 2.2 donne une description détaillée de la méthode CUBT. La section 2.3 présente des applications de CUBT sur des données de qualité de vie liée à la santé. Enfin, la section 2.4 propose une extension de la méthode CUBT pour les données qualitatives.

2.2 Classification non-supervisée basée sur les arbres binaires

La méthode CUBT se compose en trois étapes qui se succèdent : dans un premier temps, l'arbre maximal est construit, il est ensuite élagué, c'est-à-dire réduit, de manière à fusionner les noeuds terminaux qui se ressemblent. Nous commençons par définir certaines notations utiles pour la suite de cette section. Soit $\mathbf{X} \in \mathbb{R}^p$ un vecteur aléatoire réel p -dimensionnel, dont les coordonnées sont notées $X(j)$, avec $j \in \{1, \dots, p\}$, défini sur un espace de probabilité $(\Omega, \mathcal{A}, P)$ tel que $\mathbb{E}(\|\mathbf{X}\|^2) < \infty$. Les données représentent un ensemble $\mathbf{S} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$, c'est-à-dire n réalisations aléatoires indépendantes et identiquement distribuées de $\mathbf{X}$. Nous distinguons la version populationnelle de l'algorithme, basée sur le vecteur aléatoire $\mathbf{X}$, de la version empirique, basée sur l'échantillon $\mathbf{S}$. Le noeud d'un arbre est désigné par t . Chaque noeud t de l'arbre peut être regardé comme un sous-ensemble de $\mathbb{R}^p$, c'est-à-dire $t \subset \mathbb{R}^p$. On note $\hat{t}$ l'intersection entre $\mathbf{S}$ et t , c'est-à-dire l'ensemble d'observations obtenues à partir de l'échantillon $\mathbf{S}$. Par la suite, nous utiliserons toujours t au lieu de $\hat{t}$ pour simplifier les notations de la version empirique de l'algorithme.

2.2.1 Construction de l'arbre maximal

Etant donné que la méthode CUBT est descendante, l'échantillon $\mathbf{S}$ est assigné au premier noeud de l'arbre (la racine). A chaque étape, toutes les divisions binaires possibles d'un noeud terminal t sont considérées. Un exemple de division binaire est illustré dans la figure 2.1. Le noeud t est divisé en deux noeuds enfants, le noeud gauche t_g et le noeud droit t_d , si une certaine condition est vérifiée. Une règle de division binaire est de la forme $x(j) \leq a$, où $x(j)$ représente la variable j , et a est une valeur-seuil. Ainsi on peut définir les noeuds enfants t_g et t_l ,

$$\begin{aligned} t_g &= \{x \in \mathbf{S} : x(j) \leq a\}, \\ t_d &= \{x \in \mathbf{S} : x(j) > a\}. \end{aligned} \tag{2.1}$$

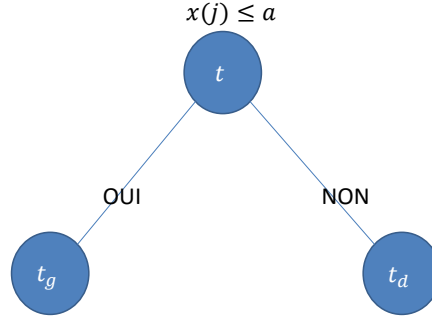


Figure 2.1 – Exemple de division binaire obtenue avec CUBT.

Soit $\mathbf{X}_t$ la restriction de $\mathbf{X}$ au noeud t , c'est-à-dire $\mathbf{X}_t = \mathbf{X}|_{\{\mathbf{X} \in t\}}$, et $\alpha_t = P(\mathbf{X} \in t)$. On définit alors une mesure d'hétérogénéité pour le noeud t , notée $R(t)$, de la manière suivante,

$$R(t) = \alpha_t \text{trace}(\text{Cov}(\mathbf{X}_t)), \tag{2.2}$$

où $\text{Cov}(\mathbf{X}_t)$ est la matrice variance-covariance de $\mathbf{X}_t$. Cette mesure, appelée la *déviance*, peut être vue comme une mesure de concentration du vecteur aléatoire $\mathbf{X}$ au noeud t , pondérée par la masse du noeud t . Dans la version empirique de l'algorithme, α_t et $\text{Cov}(\mathbf{X}_t)$ sont remplacés par leurs estimateurs. Ainsi, en notant n_t le cardinal de l'ensemble t , $n_t = \sum_{i=1}^n \mathbf{1}_{\{\mathbf{X}_i \in t\}}$ (où $\mathbf{1}_A$ est la fonction indicatrice de l'ensemble A), ainsi que la probabilité estimée $\hat{\alpha}_t = \frac{n_t}{n}$, on peut définir l'estimateur de $E(\|\mathbf{X}_t - \mu_t\|^2)$ par :

$$\frac{\sum_{\{\mathbf{X}_i \in t\}} \|\mathbf{X}_i - \bar{\mathbf{X}}_t\|^2}{n_t},$$

où $\bar{\mathbf{X}}_t$ est la moyenne empirique des observations contenues dans le noeud t et l'estimateur de la déviance est,

$$\hat{R}(t) = \frac{\sum_{\{\mathbf{X}_i \in t\}} \|\mathbf{X}_i - \bar{\mathbf{X}}_t\|^2}{n} \tag{2.3}$$

A chaque étape de la construction de l'arbre, chaque noeud terminal t est testé pour être divisé en deux sous-noeuds t_g et t_d . La meilleure division pour un noeud

t est définie par un couple $(j, a) \in \{1, \dots, p\} \times \mathbb{R}$, où j représente la variable à partir de laquelle la partition est définie et a est une valeur-seuil. Cette division doit maximiser la perte en déviance, c'est-à-dire la quantité suivante,

$$\Delta(t, j, a) = R(t) - R(t_g) - R(t_d). \quad (2.4)$$

Il est facile de vérifier que $\Delta(t, j, a) \geq 0$ pour tout triplet (t, j, a) , qui est aussi vérifiable par tous les critères de division utilisés dans la méthode CART.

Remarque : Unicité De la même façon que dans la méthode CART, le maximum du critère défini dans 2.4 n'est pas forcément unique. S'il ne l'est pas, nous utilisons des règles arbitraires suivantes. Si le maximum est atteint pour des divisions définies par des variables différentes, nous choisissons la division définie sur la variable j ayant le plus petit index. Si le maximum est atteint pour des divisions définies par une même variable j mais différentes valeurs pour le seuil a , nous choisissons la division avec la plus petite valeur de a .

Deux paramètres sont fixés pour définir l'arrêt des divisions, τ et $mindev \in (0, 1)$. A partir du noeud racine de l'arbre, chaque noeud terminal t est divisé de manière récursive jusqu'à ce qu'une des deux règles suivantes soit vérifiée :

1. $\alpha_t < \tau$,
2. $\Delta(t, j, a) < mindev \times \Delta(\mathbf{S}, j_0, a_0)$,

où $\mathbf{S}$ est l'échantillon d'apprentissage (contenant les n observations), et (j_0, a_0) le couple définissant la meilleure division du noeud racine. Dans la version empirique, on remplace α_t par $\hat{\alpha}_t$ et $\Delta(t, j, a)$ par $\hat{\Delta}(t, j, a) = \hat{R}(t) - \hat{R}(t_g) - \hat{R}(t_d)$, et on note $minsize = \lceil \tau n \rceil$ la taille minimale souhaitée d'un noeud terminal. Les deux paramètres $minsize$ et $mindev$ sont des paramètres de réglage fixés par l'utilisateur.

Une fois que la division des noeuds est stoppée, un label de classe est assigné à chaque feuille (ou noeud terminal), et l'arbre ainsi obtenu est l'arbre maximal. Une partition optimale de l'échantillon d'apprentissage est obtenue, dans laquelle chaque feuille de l'arbre est associée à un cluster. Dans le meilleur des cas, l'arbre contient au moins le même nombre de clusters que la population, bien qu'en pratique il puisse avoir trop de clusters, et une étape supplémentaire d'agglomération des clusters peut être appliquée, comme dans la méthode CART. Il est important de remarquer que si le nombre de clusters k est connu, le nombre de feuilles doit être supérieur ou égal à k . Des faibles valeurs de $mindev$ donneront lieu à de plus grands arbres qui contiennent plus de feuilles. De plus, si le nombre de feuilles dans l'arbre est égal au nombre de clusters, il n'est pas nécessaire d'appliquer les étapes suivantes de la méthode CUBT. En guise d'exemple, la figure 2.2 donne la structure d'un arbre obtenu avec CUBT, à partir d'un échantillon de données générées grâce à un modèle de simulation, contenant 400 observations décrites par 2 variables. Dans les sous-sections 2.2.2

et 2.2.3, nous décrivons les deux algorithmes permettant d'obtenir un regroupement des feuilles de l'arbre maximal. Le premier (*pruning*) permet d'élaguer l'arbre et le deuxième (*joining*) permet d'agréger les feuilles non-adjacentes.

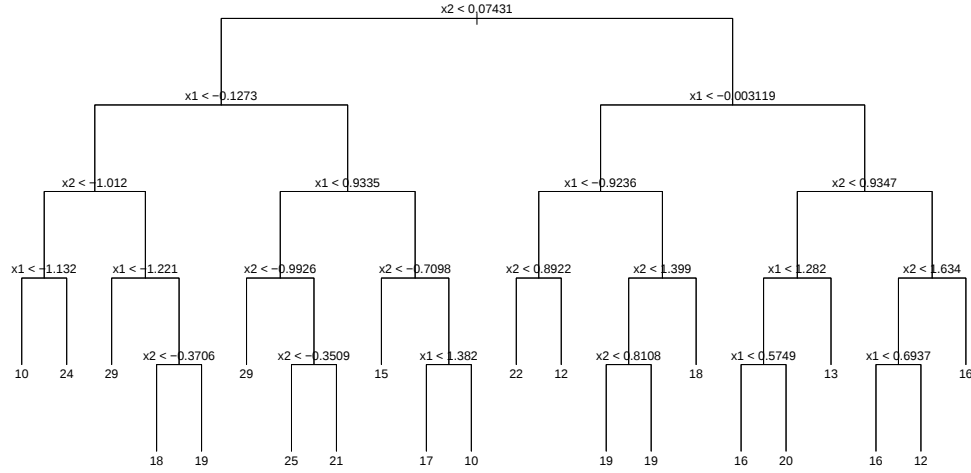


Figure 2.2 – Exemple d'arbre maximal obtenu avec la première étape de CUBT (*growing*).

2.2.2 Elagage de l'arbre maximal

Soient t_g et t_d une paire de noeuds terminaux issus d'un même noeud ascendant t . Nous définissons (en termes populationnels) les variables aléatoires $W_{g(d)} = D(\mathbf{X}_{t_g}, \text{sop}(\mathbf{X}_{t_d}))$ ($\text{sop}(Z)$ désigne le support de la variable aléatoire Z), représentant la distance euclidienne entre les vecteurs aléatoires de $\mathbf{X}_{t_g}$ et le support de $\mathbf{X}_{t_d}$, et respectivement $W_{r(l)} = D(\mathbf{X}_{t_d}, \text{sop}(\mathbf{X}_{t_g}))$.

Par définition, on a $\text{sop}(\mathbf{X}_{t_g}) \subset t_g$ et $\text{sop}(\mathbf{X}_{t_d}) \subset t_d$. Ces deux supports sont disjoints (avec une probabilité égale à 1), c'est-à-dire $P(\text{sop}(\mathbf{X}_{t_g}) \cap \text{sop}(\mathbf{X}_{t_d})) = 0$. S'il y a plus d'un cluster dans le sous-ensemble t , on peut s'attendre à ce que de petits quantiles des variables aléatoires $W_{g(d)}$ et $W_{d(g)}$ soient relativement larges, sinon ils seraient très proches de 0. Pour chacune de ces variables aléatoires on définit,

$$\begin{aligned} \Delta_{g(d)} &= \int_0^\delta q_\nu(W_{g(d)}) d\nu, \\ \Delta_{d(g)} &= \int_0^\delta q_\nu(W_{d(g)}) d\nu, \end{aligned} \tag{2.5}$$

où $q_\nu(W_{g(d)})$ représente la fonction quantile, $(\mathbb{P}(W_{g(d)} \leq q_\nu) = \nu)$ et δ est une proportion, $\delta \in (0, 1)$.

Finalement, nous proposons comme mesure de dissimilarité entre les deux ensembles t_g et t_d :

$$\Delta_{gd} = \max\{\Delta_{g(d)}, \Delta_{d(g)}\}.$$

Pour une valeur fixée $\epsilon > 0$, un noeud t est élagué si $\Delta_{gd} < \epsilon$, c'est-à-dire les deux noeuds t_g et t_d sont remplacés par leur union $t_g \cup t_d$ dans la partition.

Etant donné que $\Delta_{g(d)}$ et $\Delta_{d(g)}$ sont les moyennes des quantiles de $D(\mathbf{X}_{t_g}, \text{sop}(\mathbf{X}_{t_d}))$ et $D(\mathbf{X}_{t_d}, \text{sop}(\mathbf{X}_{t_g}))$ respectivement, Δ_{gd} peut être vu comme une version plus "résistante" de la distance entre les supports des vecteurs aléatoires $\mathbf{X}_{t_g}$ et $\mathbf{X}_{t_d}$.

Dans la version empirique de l'algorithme, nous utilisons de façon naturelle un estimateur de la mesure de dissimilarité Δ . Soit n_g (respectivement n_d) la taille du noeud $\hat{t}_g$ (respectivement $\hat{t}_d$). Pour tout $\mathbf{X}_i \in \hat{t}_g$ et $\mathbf{X}_j \in \hat{t}_d$, on considère les vecteurs $\tilde{d}_i = \min_{\mathbf{x} \in \hat{t}_g} d(\mathbf{X}_i, \mathbf{x})$, $\tilde{d}_j = \min_{\mathbf{x} \in \hat{t}_d} d(\mathbf{X}_j, \mathbf{x})$ et leurs versions ordonnées, notées d_i et d_j . Pour $\delta \in [0, 1]$, on pose,

$$\bar{d}_g^\delta = \frac{1}{\delta n_g} \sum_{i=1}^{\delta n_g} d_i,$$

$$\bar{d}_d^\delta = \frac{1}{\delta n_d} \sum_{j=1}^{\delta n_d} d_j.$$

On peut alors calculer la mesure de dissimilarité entre t_g et t_d de la façon suivante :

$$d^\delta(g, d) = d^\delta(t_g, t_d) = \max(\bar{d}_g^\delta, \bar{d}_d^\delta),$$

et à chaque étape de l'algorithme, les feuilles t_g et t_d sont agrégées au sein de leur noeud ascendant t si $d^\delta(g, d) \leq \epsilon$ où $\epsilon > 0$.

L'élagage basé sur une dissimilarité minimale prend en compte deux paramètres, δ et ϵ , ce dernier pouvant également être noté *mindist*. En termes populationnels, il suffit que ϵ soit inférieur à la distance entre les supports des deux clusters disjoints, mais cette information est inconnue et cette étape peut être évitée. Ce n'est plus un problème puisque si deux noeuds partageant le même ascendant sont proches, ils pourront toujours être agrégés dans la troisième étape de CUBT (*joining*), et la partition finale ne sera pas modifiée. Le paramètre δ est un bon moyen de gérer la possible présence de points aberrants. Quand δ tend vers 0 la mesure de dissimilarité devient la distance classique entre les deux supports $\text{sop}(\mathbf{X}_{t_g})$ et $\text{sop}(\mathbf{X}_{t_d})$.

2.2.3 Regroupement des feuilles

Le but de l'étape de regroupement des feuilles (*joining*) est d'agréger des noeuds terminaux qui ne partagent pas forcément le même noeud ascendant, c'est-à-dire des feuilles non-adjacentes. Le critère utilisé dans cette étape est le même qu'à l'étape précédente, il peut également être basé sur la perte en déviance, définie à l'équation 2.4. La figure 2.3 illustre l'utilité de cette étape. En effet, dans cet exemple, on dispose d'un nuage de points (décrits par deux variables $X(1)$ et $X(2)$) dans lequel on peut distinguer quatre groupes distincts. Un premier groupe se situe en haut à gauche du graphique (C3), un deuxième groupe en bas à droite (C2). Enfin, deux groupes de points, linéairement distribués, se situent respectivement en bas à gauche (C1) et en haut à droite (C4) du graphique. Les divisions binaires sont illustrées par des lignes séparant les différents sous-groupes.

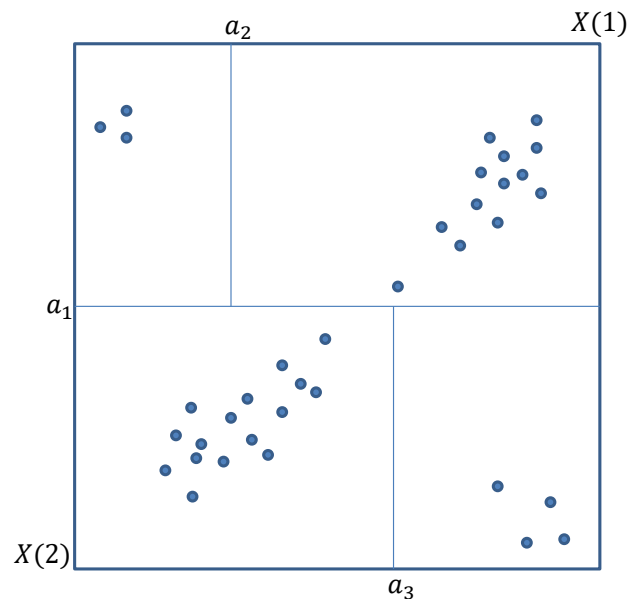


Figure 2.3 – Utilité de la troisième étape de CUBT (*joining*).

La structure de l'arbre correspondant à cette partition est donnée dans la figure 2.4, pour plus de compréhension. Dans cet exemple simple, on peut intuitivement considérer les deux sous-groupes de points C2 et C3 comme des

sous-groupes de points aberrants, à l'inverse des autres points (appartenant aux groupes C1 et C4), ceux-ci ne sont pas linéairement distribués. Il serait alors naturel d'agréger les deux sous-groupes C1 et C4. Cependant comme le montre la succession des divisions binaires effectuées, les noeuds terminaux auxquels sont assignés les sous-groupes C1 et C4 ne sont pas issus d'une même division binaire (ils ne partagent pas le même noeud ascendant). L'étape de regroupement des feuilles permettrait donc ici d'obtenir une partition en trois clusters, comprenant les deux sous-groupes C2 et C3, ainsi que l'agrégation des sous-groupes C1 et C4.

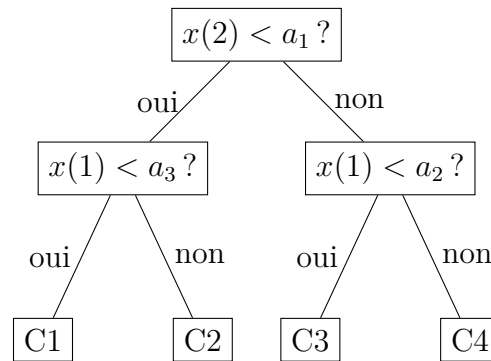


Figure 2.4 – Structure de l'arbre correspondant à la partition donnée pour illustrer le regroupement des feuilles.

Dans cette étape, toutes les paires de noeuds terminaux (adjacents ou non) t_i et t_j sont comparées en calculant la distance $d^\delta(i, j)$. De la même manière qu'en classification hiérarchique non supervisée classique, les paires de noeuds terminaux sont agrégées de façon successive, en commençant par la paire i, j ayant la valeur minimale de $d^\delta(i, j)$. Ainsi, à chaque agrégation, on obtient un cluster en moins. On considère ici deux critères d'arrêt, un pour le cas où le nombre de clusters k est connu, un autre pour le cas où k ne l'est pas. Notons m le nombre de feuilles contenues dans l'arbre après élagage. Si k est connu, on répète l'instruction suivante jusqu'à ce que $m \leq k$:

- Pour chaque paire (i, j) , $1 \leq i < j \leq m$, soit $(\tilde{i}, \tilde{j}) = \arg \min_{i,j} \{d^\delta(i, j)\}$. On remplace alors $t_{\tilde{i}}$ et $t_{\tilde{j}}$ par leur union $t_{\tilde{i}} \cup t_{\tilde{j}}$, on pose $m = m - 1$, et ainsi de suite.

Si k n'est pas connu :

- Si $d_{\tilde{i}\tilde{j}}^\delta < \eta$, on remplace $t_{\tilde{i}}$ et $t_{\tilde{j}}$ par leur union $t_{\tilde{i}} \cup t_{\tilde{j}}$, où $\eta > 0$ est une constante donnée, et on continue jusqu'à ce que cette condition ne soit plus vérifiée.

Dans le premier cas (k connu), le critère d'arrêt est simplement le nombre de clusters k , tandis que dans le deuxième cas (k non connu) une valeur seuil de η pour la distance $d^\delta(i, j)$ doit être spécifiée.

Nous avons présenté une nouvelle méthode de clustering basée sur les arbres de décision, appelée CUBT, qui est semblable aux arbres de classification et de régression, par plusieurs aspects. Cette méthode définit les clusters en termes de divisions binaires sur les variables descriptives des données. De la même manière que la méthode CART, cette méthode peut être très utile en pratique dans de nombreuses applications. Comme la structure de l'arbre est définie à partir des variables initiales, elle peut être utilisée pour déterminer quelles variables sont importantes pour la construction des clusters. De plus, l'arbre permet la classification de nouvelles observations. Cet algorithme en trois étapes est simple et ne nécessite pas un temps de calcul trop important. Aucune restriction n'est faite concernant la dimensionnalité des données, c'est-à-dire le nombre de variables initiales. Les travaux initiaux sur CUBT ont démontré que cette méthode est consistante et produit de bons résultats (au regard de simulations et d'applications sur des données réelles), en comparaison à d'autres méthodes de clustering classiques.

2.3 Applications aux données de qualité de vie liée à la santé

Les agences de réglementation telles que la Food and Drug Administration ou l'Agence Européenne de Médecine recommandent largement l'utilisation des mesures de la qualité de vie chez les patients atteints de maladies chroniques [3, 4]. Malgré le besoin grandissant exprimé de considérer les indicateurs de la qualité de vie dans la pratique clinique, les techniques d'interprétation et d'aide à la décision n'ont pas encore été parfaitement implémentées dans la pratique courante [16, 75, 27]. Certaines limites pratiques ont été décrites [76, 74, 53], en effet des études démontrent que l'interprétation des données de qualité de vie n'est pas forcément intuitive pour les cliniciens, ne leur permettant pas de choisir des interventions cliniques appropriées [81, 80]. L'implémentation de ces outils dans la pratique clinique impliquerait que des directives concernant l'interprétation des données soient disponibles pour les cliniciens [16]. Cependant, une des difficultés majeures rencontrées par les cliniciens vient du manque de telles directives lors de l'interprétation des scores de qualité de vie [16, 88]. En général, seuls les scores de qualité de vie issus des populations de référence, décrits dans les publications de validation des questionnaires, sont disponibles et peuvent être considérés comme des normes par les cliniciens. Cette approche simpliste ne permet pas de catégoriser les individus en termes de niveaux de qualité de vie, étant donné qu'aucune indication n'est disponible pour considérer simultanément les scores de plusieurs dimensions, alors que la multidimensionnalité des questionnaires est présentée comme un atout. La classification non supervisée est une technique courante pour l'analyse statistique des données, qui permet des applications dans de nombreux domaines, notamment en santé publique et

en épidémiologie. La classification non-supervisée de patients en différents sous-groupes de niveaux de qualité de vie est une approche qui peut être utilisée par les cliniciens pour interpréter les scores de qualité de vie issus de questionnaires multidimensionnels. Les méthodes de classification non supervisées (à l'inverse des méthodes supervisées, qui nécessitent de connaître l'appartenance des individus à des classes) ne requièrent aucune information a priori concernant les individus, seules les observations des variables sont utiles pour obtenir une partition. A notre connaissance, aucune approche non supervisée n'a été entreprise dans le domaine de la qualité de vie.

Plusieurs applications de la méthode CUBT [64] ont été entreprises en collaboration avec les chercheurs du laboratoire de santé publique de Marseille (site Timone), sur des données de qualité de vie, ces travaux ont donné lieu à trois publications dans des revues scientifiques, voir [121, 120]. Plus particulièrement, nous nous sommes intéressés aux scores des sous-dimensions de la qualité de vie sur différents questionnaires spécifiques à différentes pathologies : la sclérose en plaques et la schizophrénie. En appliquant la méthode CUBT aux scores des dimensions issus des questionnaires MusiQoL et SQoL, nous proposons un outil de décision binaire non-supervisé, permettant de découvrir des sous-groupes de patients, présentant des niveaux de qualité de vie similaires. Cette approche a permis aux cliniciens d'améliorer l'interprétabilité ainsi que l'aide à la décision au regard de scores-seuils, qui permettent de définir l'appartenance des individus aux différents clusters. Les sections suivantes présentent ces différentes études et leurs résultats, elles ont toutes pour but de définir des clusters de niveaux de qualité de vie à partir d'un questionnaire spécifique multidimensionnel, en utilisant la méthode CUBT. La validité clinique des partitions ainsi obtenues est testée en prenant en compte d'autres variables sociodémographiques, cliniques et de qualité de vie.

Une attention particulière est également portée sur la validité statistique de la classification non-supervisée, notamment la reproductibilité (ou la stabilité) et la discrimination (la séparabilité des clusters) de la partition. La validité d'une méthode de clustering [86, 87], ou d'une partition, est essentielle puisque la partition obtenue n'a pas forcément une structure interprétable. Différentes approches ont été proposées dans la littérature, et sont basées principalement sur l'utilisation d'information externe (c'est-à-dire de l'information non-utilisée pour créer la partition, mais qui peut apporter plus de connaissance sur la partition), l'utilisation de tests de significativité de la structure (par exemple des tests d'homogénéité), la comparaison de partitions obtenues avec différentes méthodes de classification non-supervisée, l'utilisation d'indices de validation, la mesure de la stabilité (en utilisant des techniques de ré-échantillonnage ou de validation croisée) et l'exploration visuelle (permettant d'illustrer la séparabilité des clusters). Dans nos travaux nous utilisons une technique de ré-échantillonnage de type bootstrap, et une analyse discriminante linéaire pour projeter les individus classifiés sur les deux premiers axes.

2.3.1 Etude sur des patients atteints de sclérose en plaques

Les mesures de la qualité de vie sont considérées de plus en plus comme des moyens de prédire l'invalidité des patients à moyens et longs termes [15], d'évaluer les traitements et les soins fournis aux patients atteints de sclérose en plaques [124], qui est la plus commune des maladies neurodégénératives et démyélinisantes. En plus de l'invalidité physique, une multitude d'autres symptômes ont été décrits pour la sclérose en plaques, notamment la fatigue, l'anxiété, la dépression and les troubles cognitifs, affectant de manière sévère la qualité de vie des patients [123, 150, 130]. De nombreux cliniciens se retrouvent confrontés à des difficultés d'interprétation des données de qualité de vie issus des patients atteints de sclérose en plaques et au manque de directives méthodologiques. En effet, dans le champ de la sclérose en plaques, les seules normes souvent disponibles pour les cliniciens sont les scores issus des populations de référence. La classification non supervisée pourrait donc être un bon moyen de définir de telles normes, permettant une interprétation plus aisée des données, et la distinction de différents niveaux de qualité de vie. Le but de cette étude est de définir des catégories de patients, définissant des niveaux de qualité de vie, à partir d'un questionnaire spécifique à la sclérose en plaques, le MusiQoL, pour des patients atteints de sclérose en plaques, en utilisant la méthode CUBT, et de tester la validité de la classification obtenue, au regard de différents critères cliniques et fonctionnels.

2.3.1.1 Méthodes

Conception de l'étude Cette application porte sur le même échantillon de patients décrit dans le développement de la banque d'items mesurant la qualité de vie liée à la santé mentale spécifique à la sclérose en plaque, dans la section 4.1.2.1.

Population Les critères d'inclusion sont identiques à ceux présentés dans la section 4.1.2.2.

Récolte des données Les données cliniques ont été récoltées par les cliniciens en utilisant un formulaire. Les données sociodémographiques, ainsi que les données de qualité de vie ont été récoltées en utilisant des questionnaires "papier-crayon", remplis par les patients dans les salles d'attente des différents centres hospitaliers. Les données suivantes ont été récoltées :

- Données sociodémographiques : sexe, âge, niveau scolaire, état civil et activité professionnelle.
- Données cliniques : sous-type de sclérose [112], durée de la maladie, invalidité due à la sclérose mesurée avec l'échelle d'invalidité étendue EDSS [99] et troubles cognitifs mesurés avec l'échelle MMSE [63].

- La qualité de vie est mesurée avec les questionnaires MusiQoL [155] et SF-36 [103]. Ces deux questionnaires sont décrits dans la section 1.1.2.

Analyse statistique

1. Classification non-supervisée de patients en différents groupes de niveaux de qualité de vie

Les patients sont classifiés à l'aide de la méthode CUBT [64]. L'algorithme est lancé en utilisant la première version du package R *cubt*. L'arbre de clustering est construit à partir des observations des 9 scores des dimensions du questionnaire MusiQoL. La méthode est entreprise en n'utilisant que les observations sans données manquantes. Les trois étapes de CUBT sont appliquées de manière à obtenir un arbre optimal. Nous choisissons une structure d'arbre à trois classes de niveaux de qualité de vie, de manière à simplifier l'interprétation de la partition obtenue. Les paires de noeuds terminaux (dont la taille minimale est $minsize = 50$) sont agrégés successivement jusqu'à vérification du critère d'arrêt. D'autres scénarios sont testés en faisant varier le nombre de clusters finaux (4, 5 et 6) et le paramètre *minsize* (25, 100 et 200).

2. Validité statistique de la partition

Nous utilisons une analyse discriminante canonique en utilisant les neuf variables représentant les scores des dimensions du MusiQoL, pour évaluer graphiquement la séparabilité (ou la discrimination) des classes obtenues. La procédure CANDISC, disponible avec le logiciel SAS, est utilisée pour représenter les individus de chaque cluster dans un graphique en deux dimensions, dont les axes sont les variables canoniques.

La stabilité de la partition est mesurée en utilisant une technique de bootstrap. Pour 1000 échantillons bootstrap issus de l'échantillon d'étude, nous construisons 1000 arbres en utilisant la même méthode décrite précédemment. Nous calculons les proportions d'échantillons bootstrap ayant la même variable dans le même premier noeud, et les mêmes variables dans le deuxième et troisième noeuds conditionnellement au premier noeud. Nous calculons également la moyenne du seuil de chaque division binaire de l'arbre. Nous utilisons l'indice de Rand, l'indice de Rand ajusté [92] et l'erreur de mauvaise classification [64] (voir section 2.4.3.5) afin de comparer les différentes partitions obtenues avec la partition initiale (c'est-à-dire celle obtenue sur l'échantillon d'étude).

3. Validité clinique de la partition

Afin de mesurer la validité clinique de la partition obtenue, nous comparons entre les trois classes d'individus, les mesures suivantes : les scores du questionnaire SF-36, l'âge, le score EDSS, la durée de la maladie en utilisant une analyse de la variance, le genre, le niveau scolaire, le statut

		N=1361
Age, years (n=1327) M $\pm$ SD		42.2 $\pm$ 11.9
Sex (n=1345) N (%)	Female	925 (68.8)
	Male	420 (31.2)
Country (n=1361) N(%)	Argentina	17 (1.2)
	Canada	47 (3.5)
	France	106 (7.8)
	Germany	130 (9.6)
	Greece	55 (4.0)
	Israel	32 (2.4)
	Italy	331 (24.3)
	Lebanon	12 (0.9)
	Norway	65 (4.8)
	Russia	92 (6.8)
	South Africa	38 (2.8)
	Spain	162 (11.9)
	Turkey	171 (12.6)
	UK	26 (1.9)
	USA	77 (5.7)
Educational level (n=1087) N (%)	<12 years	697 (64.1)
	$\geq$ 12 years	390 (35.9)
Marital status (n=1103) N (%)	Single	180 (16.3)
	Not single	923 (83.7)
Occupational status (n=1048) N (%)	Worker or student	668 (63.7)
	Not working	380 (36.3)
MS subtype (n=1329) N (%)	Relapsing Remitting	947 (71.3)
	Primary Progressive	94 (7.1)
	Secondary Progressive	266 (20.0)
	Clinically Isolated Syndrome	22 (1.6)
EDSS score (n=1335) m [IQR]		3.0 [1.5 – 5.0]
Cognitive issue (n=1091) N (%)	Present	122 (11.2)
	Absent	969 (88.8)
Disease duration, years (n=1309) M $\pm$ SD		7.5 $\pm$ 6.5
MusiQoL* M $\pm$ SD	ADL	54.2 $\pm$ 26.8
	PWB	55.8 $\pm$ 23.8
	RFr	63.0 $\pm$ 25.6
	SPT	66.6 $\pm$ 23.4
	RFa	75.4 $\pm$ 22.9
	RHCS	78.2 $\pm$ 19.8
	SSL	60.8 $\pm$ 31.5
	COP	62.7 $\pm$ 30.4
	REJ	75.7 $\pm$ 25.4
	Global index	65.8 $\pm$ 14.7
SF-36* M $\pm$ SD	PF (n=1346)	55.3 $\pm$ 31.2
	SF (n=1352)	67.9 $\pm$ 25.7
	RP (n=1329)	42.2 $\pm$ 41.6
	RE (n=1313)	55.9 $\pm$ 42.4
	MH (n=1350)	62.5 $\pm$ 20.6
	Vi (n=1350)	48.3 $\pm$ 22.3
	BP (n=1346)	63.1 $\pm$ 24.6
	GH (n=1331)	49.5 $\pm$ 22.3
	PCS (n=1278)	39.6 $\pm$ 10.4
	MCS (n=1278)	45.9 $\pm$ 11.2

Table 2.1 – Caractéristiques de la population.

marital, l'activité professionnelle, la sous-type de sclérose et les fonctions cognitives en utilisant un test du Khi-2.

2.3.1.2 Résultats

Echantillon d'étude et caractéristiques de la population 1992 patients ont participé à cette étude. 1361 patients ont rempli complètement les items du questionnaire, de manière à avoir neuf scores. Le nombre de patients dans chaque pays varie de 12 (0.9%, Liban) à 331 (24.3%, Italie). Plus de deux tiers des patients sont des femmes, l'âge moyen est de 42.2 ans. Près de 70% des patients ont une sclérose en plaques de type cyclique. Les plus faibles scores observés sont sur la dimension activité de la vie quotidienne et les plus hauts sont observés pour la dimension relations avec le système de soins. La table 2.1 présente les caractéristiques de la population.

Classification non-supervisée de patients en différents groupes de niveaux de qualité de vie La structure d'arbre à trois classes a classé 87 patients dans le groupe "faible niveau de qualité de vie", 1173 dans le groupe "niveau modéré de qualité de vie" et 101 dans le groupe "haut niveau de qualité de vie" (voir Figure 2.5). La première variable qui discrimine le plus les patients est la dimension vie sentimentale et sexuelle. Les feuilles de l'arbre maximal sont constituées par les trois aspects de la qualité de vie du MusiQoL, de la manière suivante : aspect social (vie sentimentale et sexuelle et les relations avec les amis), aspect psychologique (bien-être psychologique, coping, rejet), et aspect physique (activités de la vie quotidienne). Les scores de dimensions du MusiQoL pour les trois classes sont fournis dans la Figure 2.6.

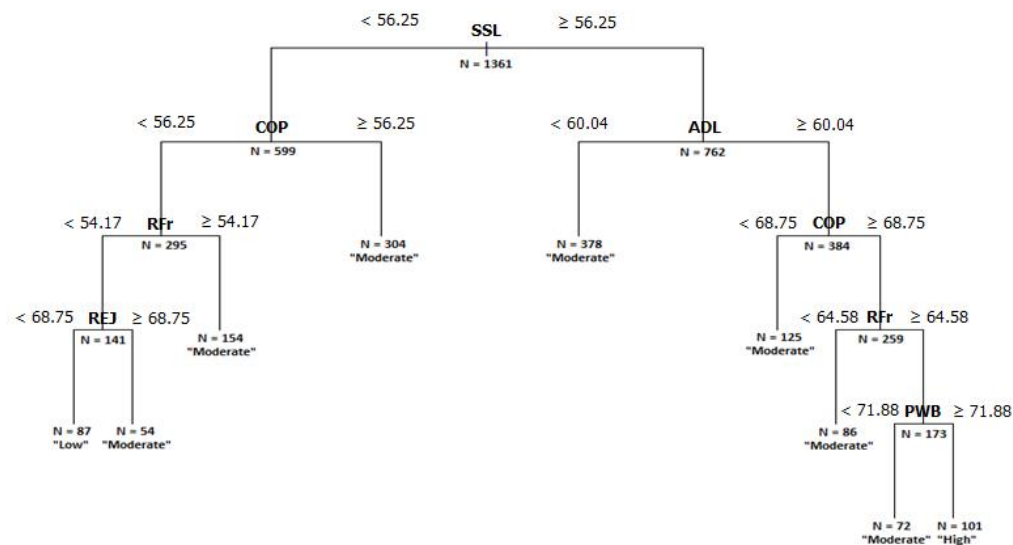


Figure 2.5 – Structure à trois classes obtenue avec les scores des dimensions du questionnaire MusiQoL.

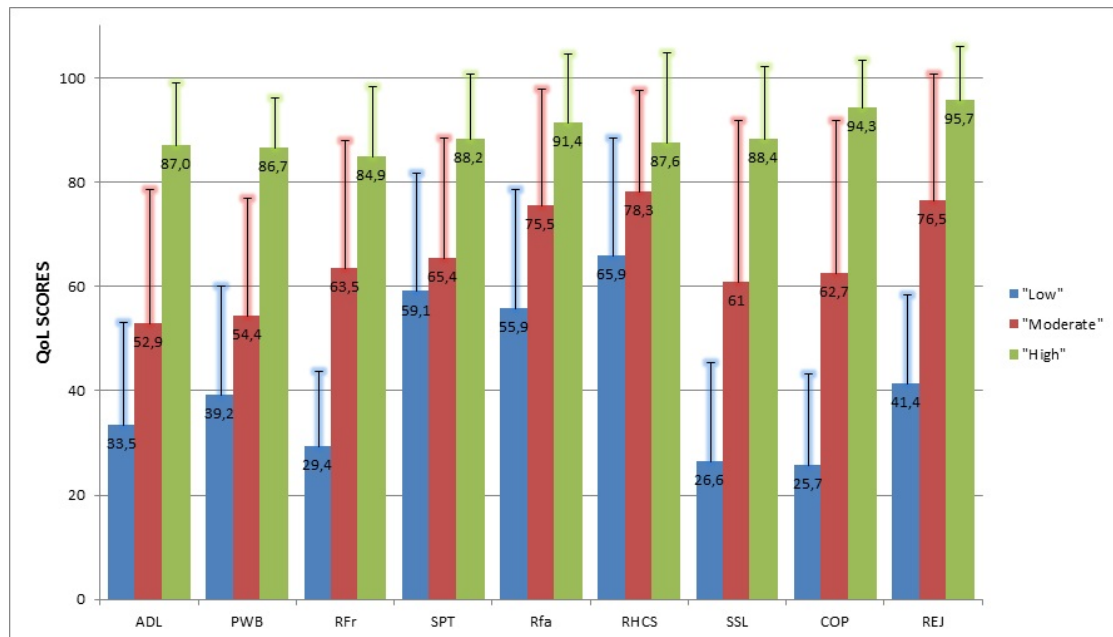


Figure 2.6 – Scores des dimensions du MusiQoL, pour chaque classe de niveau de qualité de vie.

Validité statistique de la partition Les résultats de l'analyse discriminante canonique sont fournis à la Figure 2.7, qui montre une représentation graphique de la structure à trois classes. La première variable canonique est représentée principalement par les trois dimensions suivantes (au regard des valeurs absolues des coefficients canoniques) : relations avec les amis (0.57), vie sentimentale et sexuelle (0.37) et coping (0.39). la deuxième variable canonique est représentée par les dimensions suivantes : activités de la vie quotidienne (-0.55), bien-être psychologique (-0.69) et rejet (0.77). Le groupe de points à droite du graphique représente le groupe à "haut niveau de qualité de vie" et le groupe à gauche représente le groupe "faible niveau de qualité de vie". Graphiquement, la séparabilité des classes semble satisfaisante.

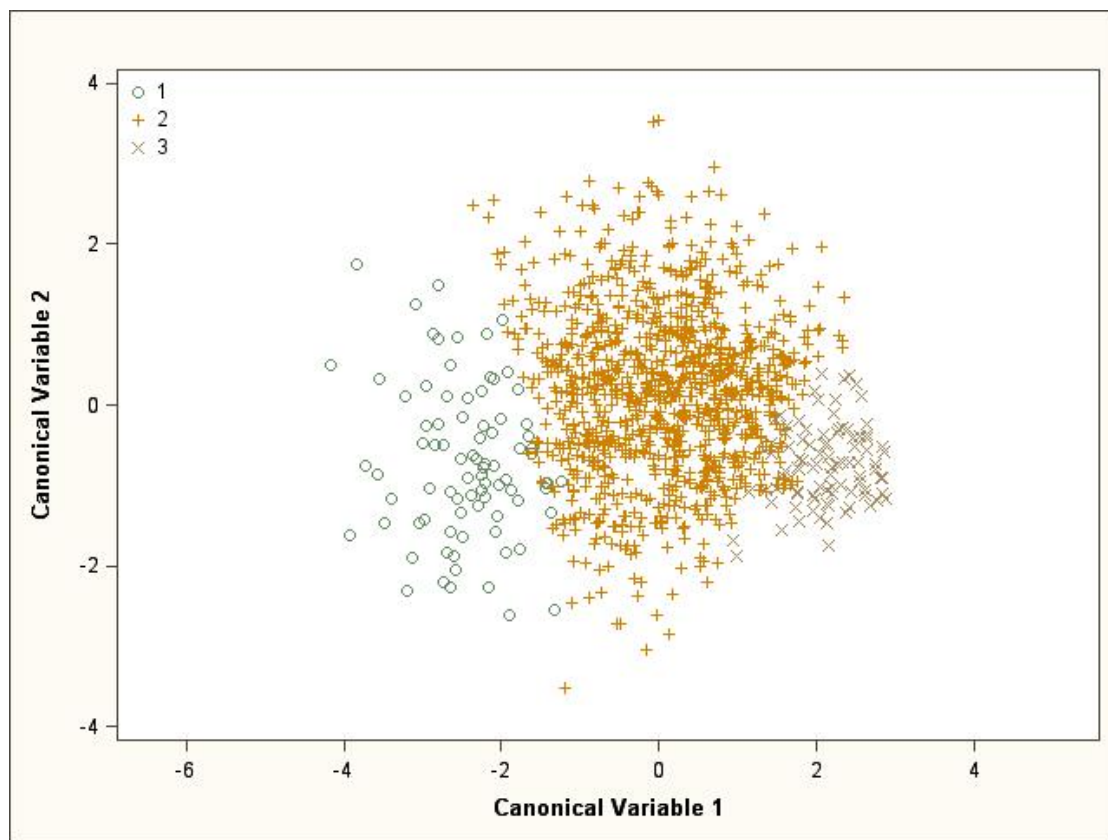


Figure 2.7 – Représentation graphique de la structure à trois classes.

Les proportions d'échantillons bootstrap qui ont utilisé la vie sentimentale et sexuelle pour déterminer la première division binaire de l'arbre, le coping pour la seconde (conditionnellement à la première) et les activités de la vie quotidienne pour la troisième (conditionnellement à la première) sont 71.6%, 99.3% et 74.5% respectivement. Pour ces trois premières divisions, les moyennes des valeurs-seuils utilisées pour les définir sont 56.82, 54.54 et 61.07 respectivement. La valeur moyenne de l'indice de Rand est 0.95, celle de l'indice de Rand ajusté est 0.58 et celle de l'erreur de mauvaise classification est 0.35, indiquant une stabilité satisfaisante de la partition.

Validité clinique de la partition En accord avec les résultats de l'étude de la validation du questionnaire MusiQoL, les résultats attendus ont été confirmés (voir Table 2.2). Les patients du groupe "haut niveau" sont ceux qui ont les plus hauts scores pour les dimensions du SF-36 (et ses scores composites). Les moyennes des scores du MusiQoL diffèrent significativement entre les trois groupes. Les trois classes sont également significativement différentes au regard de critères sociodémographiques et cliniques. L'âge moyen des patients du

		Cluster 1 Low QoL N=87	Cluster 2 Moderate QoL N=1173	Cluster 3 High QoL N=101	p*	p 1 vs 2	p 1 vs 3	p 2 vs 3
Age, years M±SD		40.3± 9.2	42.6± 12.0	38.8± 12.0	0.003	0.249	1	6E-3
Sex N(%)	Male	27(32.1)	365(31.5)	28(27.7)	0.726			
	Female	57(67.9)	795(68.5)	73(72.3)				
Educational level N(%)	<12 years	61(83.6)	604(64.3)	32(42.7)	<10-3	1E-3	<10-3	<10-3
	≥ 12 years	12(16.4)	335(35.7)	43(57.3)				
Marital status N(%)	Not single	61(81.3)	798(83.8)	64(84.2)	0.847			
	Single	14(18.7)	154(16.2)	12(15.8)				
Occupational status N(%)	Worker	36(53.7)	567(62.6)	65(86.6)	<10-3	0.15	<10-3	<10-3
	Not working	31(46.3)	339(37.4)	10(13.4)				
MS subtype N(%)	RR	55(65.5)	807(70.2)	85(88.5)	<10-3	0.30	<10-3	<10-3
	PP	4(4.7)	87(7.6)	3(3.1)				
	SP	25(29.8)	237(20.6)	4(4.2)				
	CIS	0(0.0)	18(1.6)	4(4.2)				
EDSS score m [IQR]		3.5 [2.5-6.0]	3.0 [2.0-5.0]	1.5 [1.0-2.5]	<10-3	<10-3	<10-3	<10-3
Cognitive issue N(%)	Present	8(10.8)	112(11.9)	2(2.7)	0.052			
	Absent	66(89.2)	832(88.1)	71(97.3)				
Disease duration, years M±SD		8.4±6.7	7.5±6.4	7.3±7.0	0.413			
MusiQoL M±SD		Index	41.9±8.6	65.6±12.4	89.4±19.7	<10-3	<10-3	<10-3
SF-36 M±SD	PF	34.2±26.6	54.2±42.4	86.4±19.7	<10-3	<10-3	<10-3	<10-3
	SF	43.4±21.4	67.5±24.9	92.6±13.7	<10-3	<10-3	<10-3	<10-3
	RP	15.8±27.9	40.5±40.9	82.8±31.3	<10-3	<10-3	<10-3	<10-3
	RE	24.1±33.5	55.4±42.4	88.6±22.4	<10-3	<10-3	<10-3	<10-3
	MH	41.1±17.0	62.0±19.5	86.3±10.4	<10-3	<10-3	<10-3	<10-3
	Vi	31.3±14.0	47.0±21.2	76.9±15.5	<10-3	<10-3	<10-3	<10-3
	BP	53.0±23.6	62.2±24.5	82.1±15.8	<10-3	2E-3	<10-3	<10-3
	GH	34.0±8.5	48.6±21.4	73.4±18.7	<10-3	<10-3	<10-3	<10-3
	PCS	34.8±8.5	39.1±10.2	49.7±7.1	<10-3	<10-3	<10-3	<10-3
	MCS	35.3±9.4	45.7±10.9	56.3±5.7	<10-3	<10-3	<10-3	<10-3

Table 2.2 – Caractéristiques cliniques et sociodémographiques des trois classes.

groupe "haut niveau" est plus bas que dans les deux autres groupes. Le niveau scolaire est la proportion d'actifs sont plus élevés dans le groupe "haut niveau". La proportion de scléroses progressives secondaires est plus grande dans le groupe "bas niveau". Le genre, le statut marital, la durée de la maladie et la présence de trouble cognitif ne diffèrent pas entre les trois groupes.

2.3.2 Etude sur des patients atteints de schizophrénie

Une étude similaire à celle présentée dans la section 2.3.1 a été entreprise sur des données de qualité de vie issues d'un échantillon de patients atteints de schizophrénie. Dans cette étude nous appliquons la même méthode pour définir des groupes de niveaux de qualité de vie, à partir des scores des dimensions du questionnaire SQoL. Ces travaux ont donné lieu à une publication [120].

2.4 Extensions de la méthode pour des données qualitatives

Une des limitations de CUBT réside dans le fait que cette méthode utilise des critères pour construire et élaguer l'arbre (la déviance et la mesure de dissimilarité) qui sont spécifiques aux données continues. Nous proposons donc dans cette section des extensions de CUBT pour les données ordinales et nominales. Pour cela, pour chaque étape de CUBT (*growing*, *pruning*, *joining*), nous définissons de nouveaux critères basés sur l'information mutuelle ou sur l'entropie

de Shannon. Nous fournissons quelques heuristiques pour le choix initial des paramètres utilisés dans chacune des étapes.

2.4.1 Extension aux données ordinales

La version ordinale de l'extension de la méthode CUBT a donné lieu à la participation à une conférence internationale (COMPSTAT 2014, Genève, Suisse). Les détails de cette extension et des simulations effectuées sont décrites dans un article court [70].

2.4.2 Extension aux données nominales

Nous décrivons ici les nouveaux critères utilisés dans les trois étapes de la méthode CUBT, pour pouvoir appliquer cette dernière aux données nominales.

2.4.2.1 Notations

Soit $X \in E = \prod_{j=1}^p \mathcal{X}_j$, un vecteur nominal p -dimensionnel dont les coordonnées sont notées $X_{.j}$, $j \in \{1, \dots, p\}$, et soit $m_j \in \mathbb{N}$ le nombre de modalités de la variable j , et $\mathcal{X}_j$ est l'ensemble des modalités de la variable j . Nous considérons un ensemble S de n vecteurs aléatoires, notés X_i , avec $i \in \{1, \dots, n\}$. Finalement, on note X_{ij} l'observation i de la composante j de X . Des notations similaires sont utilisées avec des minuscules pour noter les réalisations de ces variables : x , x_i, x_j et x_{ij} .

2.4.2.2 Critère d'hétérogénéité

Pour tout noeud t (un sous-ensemble de S), soit $X^{(t)}$ la restriction de X au noeud t , c'est-à-dire $X^{(t)} = X|_{\{X \in t\}}$, et on définit $R(t)$, la mesure d'hétérogénéité du noeud t définie de la manière suivante :

$$R(t) = \sum_{j=1}^p H(X_{.j}^{(t)}) = - \sum_{j=1}^p \sum_{k=1}^{m_j} p_{kj}^{(t)} \log_2 p_{kj}^{(t)}, \quad (2.6)$$

où $H(X_{.j}^{(t)})$ est l'entropie de Shannon de $X_{.j}^{(t)}$ au sein du noeud t et $p_{kj}^{(t)}$ est la probabilité pour la composante j de X de prendre la valeur k au sein du noeud t . Cette mesure est la somme des entropies de chaque variable. Elle peut aussi être écrite de la manière suivante :

$$R(t) = \text{trace}(\mathbf{MI}(X^{(t)})), \quad (2.7)$$

où $\mathbf{MI}$ est la matrice d'information mutuelle de $X^{(t)}$, et l'information mutuelle entre deux variables Y et Z est définie par :

$$MI(Y, Z) = \sum_{y,z} \log_2 \frac{p(y, z)}{p(y)p(z)}, \quad (2.8)$$

où y et z sont les modalités de Y et Z , respectivement.

Le choix de ce critère d'hétérogénéité peut être justifié en quelques mots. L'entropie est un critère naturel pour mesurer l'hétérogénéité pour des mesures qualitatives. Elle est souvent utilisée dans les algorithmes de classification supervisée comme CART. L'information mutuelle de deux variables mesure la dépendance entre les variables, elle est comparable à la covariance dans le cas continu, les deux statistiques étant liées lorsque les variables sont gaussiennes.

2.4.2.3 Construction de l'arbre maximal

Initialement, le noeud primaire (la racine) de l'arbre contient toutes les observations de l'ensemble S . L'échantillon est ensuite divisé récursivement en deux sous-échantillons disjoints, en utilisant des divisions binaires de la forme $x_{.j} \in \mathcal{A}_j$, où $j \in \{1, \dots, p\}$ et $\mathcal{A}_j$ est un sous-ensemble des modalités de $x_{.j}$. Ainsi, la division binaire d'un noeud t en deux noeuds enfants t_g et t_d est définie par une paire $(j, \mathcal{A}_j)$. Les noeuds t_g et t_d sont définis de la manière suivante :

$$\begin{aligned} t_g &= \{x \in E : x_{.j} \in \mathcal{A}_j\}, \\ t_d &= \{x \in E : x_{.j} \notin \mathcal{A}_j\} \end{aligned} \quad (2.9)$$

Le split optimal de t en deux noeuds enfants t_g et t_d est défini par :

$$\operatorname{argmax}_{(j, \mathcal{A}_j) \in \{1, \dots, p\} \times \{1, \dots, m_j\}} \{\Delta(t, j, \mathcal{A}_j)\}, \quad (2.10)$$

avec $\Delta(t, j, \mathcal{A}_j) = R(t) - R(t_g) - R(t_d)$.

Les nouveaux noeuds ainsi créés sont alors divisés à leur tour, récursivement. La construction de l'arbre maximal s'arrête lorsque l'un des critères d'arrêt est vérifié. Les critères d'arrêt utilisés sont les mêmes que dans la version continue de l'algorithme (voir section 2.2.1).

2.4.2.4 Elagage de l'arbre et regroupement des feuilles

Le principe des deux étapes de réduction de l'arbre maximal (*pruning* et *joining*) est très similaire à celui des versions continues. La seule différence ici est la mesure de dissimilarité qui doit être basée sur une distance autre que la distance euclidienne, étant donné que les variables sont nominales.

Pour l'élagage, on note t_g et t_d deux noeuds adjacents issus de la division binaire d'un noeud t . Soit n_g (respectivement n_d) la taille (c'est-à-dire le nombre d'observations) du noeud t_g (respectivement t_d) et $\delta \in [0, 1]$. Pour tout $x_i \in t_d$ et $x_j \in t_g$, avec $i, j \in \{1, \dots, n\}$, on pose :

$$\begin{aligned}\tilde{d}_i &= \min_{x \in t_g} d(x, x_i), \\ \tilde{d}_j &= \min_{x \in t_d} d(x, x_j),\end{aligned}\tag{2.11}$$

et leurs versions ordonnées notées d_i et d_j . Ici, $d(y, z)$ peut être soit la distance de Hamming (notée d_{Ham}), soit l'information mutuelle entre deux observations. La distance de Hamming est définie de la manière suivante :

$$d_{Ham}(y, z) = \sum_{i=1}^n \mathbf{1}_{\{y_i \neq z_i\}}\tag{2.12}$$

La suite du calcul de la mesure empirique de dissimilarité ainsi que le critère d'arrêt de l'étape d'élagage sont similaires à ceux de la version continue (voir section 2.2.2).

Le regroupement des feuilles dans le cas des données nominales ne présente aucune différence avec la version continue, exceptée la prise en compte des nouveaux critères définis ci-dessus (critère d'hétérogénéité et mesure de dissimilarité). Les critères d'arrêt sont les mêmes, que le nombre de groupes soit connu ou inconnu (voir section 2.2.3).

2.4.3 Expérimentations

Dans cette section, nous présentons quelques simulations qui utilisent différents modèles. Nous comparons les résultats de la méthode CUBT avec ceux obtenus par six méthodes adaptées aux données nominales, notamment la classification hiérarchique non-supervisée [128], ainsi que les algorithmes k -modes [91], DBSCAN [57], COBWEB [60], DIVCLUS-T [41] et LCA [163]. L'erreur de mauvaise classification ainsi que l'indice de Rand ajusté sont utilisés pour ces comparaisons. Une mise à jour du package R CUBT a été effectuée pour prendre en compte les nouveaux critères décrits dans la section précédente.

2.4.3.1 Méthodes de classification non-supervisée

Nous fournissons ici une description de plusieurs méthodes classiques de classification non supervisée, adaptées aux données qualitatives, que nous utilisons dans les simulations, afin de les comparer à la méthode CUBT.

Classification hiérarchique La classification hiérarchique a pour but de construire un dendrogramme, étant donnée une distance d , et peut être ascendante ou descendante. En classification ascendante hiérarchique, la hiérarchie se construit par agrégation itérative de paires de clusters proches. La distance entre deux clusters peut être définie de différentes manières, les plus populaires étant le

saut minimum (minimum des distances inter-individus, [62]), le saut maximum (maximum des distances inter-individus), le lien moyen (moyenne des distances inter-individus) et la distance de Ward (qui maximise l'inertie interclasse). Dans ces expérimentations, nous utilisons le saut maximum pour agréger les clusters. Nous utilisons l'information mutuelle comme mesure de dissimilarité entre les observations. L'algorithme est lancé en utilisant la fonction *hclust* de R.

***k*-modes** L'algorithme *k*-modes est une extension du bien connu algorithme *k*-means [113]. Cet algorithme séquentiel s'initialise en choisissant *k* centres de clusters arbitrairement, il assigne chaque observation au centre le plus proche, et calcule le nouveau centre de chaque classe en utilisant les observations assignées à cette classe. Etant donné que la partition obtenue dépend fortement des centres initiaux, il est recommandé de lancer l'algorithme plusieurs fois, et de conserver la partition qui minimise la somme des carrés des clusters, donnée par $\sum_{i=1}^n \sum_{j=1}^k \|\mathbf{X}_i - c_j\|^2 \mathbf{1}_{\mathbf{x}_i \in G_j}$, où G_j est la *j*-ième classe et c_j est le centre correspondant.

L'algorithme *k*-modes [91] est une méthode de clustering pour les données catégorielles qui étend l'algorithme *k*-means. Il recherche une partition des observations en *k* groupes tels que la distance entre les observations et les modes de chaque groupe soit minimale. Nous utilisons la distance appelée *simple matching coefficient* (SMC) pour mesurer la dissimilarité entre les paires d'observations. Le package R *klaR* [169] est utilisé pour appliquer cette méthode.

Clustering basé sur la densité DBSCAN [57] est un algorithme non hiérarchique basé sur la densité. Il utilise deux paramètres : une distance de joignabilité (c'est-à-dire un rayon) ϵ et un nombre minimal de points, noté *MinPts*. Considérons un point fixe arbitraire (une observation) issu des données, le voisinage défini par ϵ est construit autour de ce point, il représente l'ensemble des points qui sont à une distance inférieure à ϵ de ce point. S'il y a au moins *MinPts* points dans ce ϵ -voisinage, ce point (qui est défini comme dense) ainsi que tout son voisinage constituent alors un cluster, sinon ce point est identifié comme du bruit. Tous les points denses qui sont trouvés dans le ϵ -voisinage sont également ajoutés au cluster ainsi que leur propres voisinages. Une fois que tous les points denses ont été détectés, un nouveau point "non-visité" par l'algorithme est alors recherché et le processus pour explorer de nouveaux clusters est répété. Pour cette méthode, nous utilisons l'information mutuelle comme mesure de dissimilarité. Les paramètres ϵ et *MinPts* sont fixés par l'utilisateur. La fonction *dbscan* du package R *fpc* est utilisée pour utiliser cette méthode.

COBWEB L'algorithme COBWEB [60] est une approche de clustering conceptuel basé sur une mesure appelée l'utilité des classes (category utility, [47]).

Cette procédure hiérarchique incrémentale est adaptée aux données qualitatives. Elle construit un arbre de manière dynamique, en insérant un individu à la fois dans la construction de l'arbre. A chaque insertion d'un nouvel individu, COBWEB a quatre options disponibles : insérer l'individu dans un cluster existant, créer un nouveau cluster (représenté par le nouvel individu), agréger deux noeuds ou diviser un noeud. Pour chacune de ces options, l'utilité des classes est calculée pour chaque partition correspondante. La partition produisant la valeur maximale de cette mesure est alors sélectionnée. Ensuite, l'individu suivant dans les données est inséré dans la construction de l'arbre. Le critère d'arrêt est une valeur minimale de l'utilité des classes, qui ne peut être dépassée. Nous utilisons le seuil par défaut dans les simulations (0.02). L'algorithme est lancé via la fonction *Cobweb* du package R *RWeka* [79].

DIVCLUS-T La méthode DIVCLUS-T [41] est une méthode de clustering hiérarchique descendante. Elle est monothétique, c'est-à-dire que les sous-ensembles d'observations dans le jeu de données, sont divisés en n'utilisant qu'une seule variable. Son but est d'optimiser le même critère qu'en classification hiérarchique classique basée sur la méthode de Ward. Cette méthode est adaptée aux données quantitatives, qualitatives, et aux données mixtes. Pour calculer le critère de division, c'est-à-dire l'inertie intra-classe, DIVCLUS-T utilise la distance euclidienne pour le cas quantitatif, et la distance du Khi-2 pour le cas qualitatif. Le principal avantage de cette méthode, en comparaison avec les autres méthodes hiérarchiques, est, qu'en plus de fournir un dendrogramme, chaque noeud de l'arbre est désigné par sa règle de division binaire. Ainsi le dendrogramme obtenu par DIVCLUS-T peut être lu comme un arbre de décision binaire. DIVCLUS-T est donc conceptuellement la méthode la plus proche de CUBT. L'algorithme est lancé en utilisant la fonction *divclust* du package R *divclust*.

LCA L'analyse de classes latentes (latent class analysis, LCA, [163]) est une méthode de clustering basée sur des modèles de mélanges. Dans cette approche, les observations appartenant à une même classe sont similaires au regard des variables observées et sont supposées être issues d'une même distribution de probabilité (dont les paramètres sont inconnus). Cette méthode est adaptée à l'analyse des données catégorielles multivariées. Un modèle LCA est un modèle de mélange fini [5] dans lequel les distributions sont des tables de classification croisée. Appliquée à des données catégorielles, cette méthode est très similaire à un mélange de modèles issus de la théorie de réponse à l'item. Ainsi, le modèle suppose que les variables sont conditionnellement indépendantes. Les paramètres du modèle sont estimés par maximisation de la vraisemblance en utilisant les algorithmes Espérance-Maximisation et de Newton-Raphson. L'algorithme est utilisé via la fonction *lca* [107] du package R *poLCA*.

2.4.3.2 Modèles de simulation de données qualitatives

Nous considérons cinq modèles de simulation de données : le modèle "combinaison linéaire" (M1), deux modèles basés sur des structures d'arbres (M2 et M3) et deux modèles basés sur la théorie de réponse à l'item. Pour chacun des modèles, nous testons différentes tailles d'échantillons $n \in \{100, 300, 500\}$, et les différents clusters sont de taille égale.

M1 : Modèle de combinaison linéaire Dans ce modèle, chaque variable $X_{.j}$, $j \in \{1, \dots, p = 9\}$ a $m = 5$ modalités. Nous définissons $k = 3$ clusters, chacun est caractérisé par une haute fréquence d'une des modalités. Pour les observations du cluster 1, on a $P(X_{.j} = 1) = q$, et une probabilité uniforme est utilisée pour les autres modalités, c'est-à-dire $P(X_{.j} = l) = \frac{1-q}{m-1}$ pour $l \neq 1$. Pour les clusters 2 et 3, les modalités fréquentes sont 3 et 5, respectivement, et on utilise les mêmes probabilités. Dans les simulations, on fixe la probabilité $q = 0.8$. Ce modèle de simulation est très difficile pour CUBT étant donné que $\sum_{j=1}^p X_{.j}$ est une variable parfaitement discriminante pour les clusters, particulièrement pour de grandes valeurs de q ; de ce fait $1 - q$ peut être vu comme un indice de chevauchement des clusters.

M2 : Modèle basé sur une structure d'arbre 1 Nous utilisons ici un modèle basé sur la structure d'un arbre. Nous fixons la dimensionnalité (c'est-à-dire le nombre de variables) $p = 3$ et le nombre de groupes $k = 4$. Chaque variable $X_{.j}, j \in \{1, \dots, p\}$, a $m = 4$ modalités. Chaque modalité est codée comme un entier, et nous distinguons les modalités paires et les modalités impaires. La partition utilisée pour la simulation est présentée à la figure 2.8. Les clusters sont définis de la manière suivante :

1. C1 : $x_{.1}$ et $x_{.2}$ ont des modalités impaires, et $x_{.3}$ est arbitraire
2. C2 : $x_{.1}$ a des modalités impaires, $x_{.2}$ a des modalités paires, et $x_{.3}$ est arbitraire
3. C3 : $x_{.1}$ a des modalités paires, $x_{.3}$ a des modalités impaires, et $x_{.2}$ est arbitraire
4. C4 : $x_{.1}$ et $x_{.3}$ ont des modalités paires, et $x_{.2}$ est arbitraire

Ce modèle produit des clusters qui devraient être découverts facilement par la méthode CUBT, étant donné que chaque cluster est caractérisé par quelques modalités de chaque variable. Cependant, comme les modalités sont distribuées uniformément, leur contribution à l'entropie est élevée, rendant les divisions binaires optimales difficiles à retrouver.

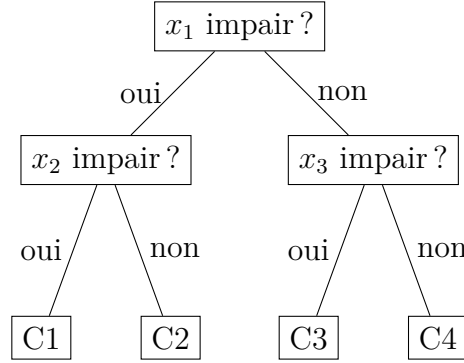


Figure 2.8 – Structure de l'arbre utilisée pour le modèle de simulation M2.

M3 : Modèle basé sur une structure d'arbre 2 Nous utilisons un modèle avec la même structure d'arbre que dans le modèle précédent (voir figure 2.8). Nous fixons la dimensionnalité $p = 3$ et le nombre de groupes $k = 4$. Ici, chaque variable X_j , $j \in \{1, \dots, p\}$ a $m = 4$ modalités. La seule différence avec le modèle précédent est que les modalités des variables ne sont pas distribuées uniformément dans chaque cluster. Nous considérons un paramètre p_0 qui permet de contrôler la non-uniformité de la distribution des modalités. Dans nos simulations, nous fixons $p_0 = 0.8$. Les clusters sont définis de la manière suivante :

1. C1 : x_1 et x_2 ont des modalités impaires avec $P(x_1 = 1) = P(x_2 = 1) = p_0$, et x_3 est arbitraire
2. C2 : x_1 a des modalités impaires, x_2 a des modalités paires avec $P(x_1 = 1) = P(x_2 = 2) = p_0$, et x_3 est arbitraire
3. C3 : x_1 a des modalités paires, x_3 a des modalités impaires avec $P(x_1 = 2) = P(x_3 = 1) = p_0$, et x_2 est arbitraire
4. C4 : x_1 et x_3 ont des modalités paires avec $P(x_1 = 2) = P(x_3 = 2) = p_0$, et x_2 est arbitraire

Ce modèle permet de générer des clusters qui sont plus faciles à découvrir que ceux du modèle précédent. En effet, les modalités au sein des clusters sont distribuées non-uniformément, pour variable intervenant dans les divisions de l'arbre. Cela implique une minimisation de la contribution des variables à l'entropie.

M4 : Modèle nominal de réponse à l'item Nous utilisons ici un modèle de réponse à l'item adapté aux données nominales. Nous fixons la dimensionnalité $p = 9$ et le nombre de groupes $k = 3$. Chaque variable a $m_j = 5$ modalités. Dans ce modèle, on suppose que les variables représentent des items à choix multiples. Le modèle de réponse nominal [24] est une spécialisation du modèle général pour réponses multinomiales et il est défini de la manière suivante :

Soit θ un niveau de trait mesuré par les réponses d'un ensemble d'items. La probabilité qu'un individu de niveau θ réponde la modalité k à l'item j est donnée par :

$$\Psi_{jk_j}(\theta) = \exp[z_{jk_j}(\theta)] / \sum_{h=1}^{m_j} \exp(z_{jh}(\theta)), \quad (2.13)$$

où $z_{jh}(\theta) = c_{jh} + a_{jh}\theta$ avec $h = 1, 2, \dots, k_j, \dots, m_j$, et c_{jh} et a_{jh} des paramètres d'items associés à la h -ème de l'item j .

Nous pouvons générer aléatoirement des jeux de données en utilisant ce modèle, en simulant des distributions de traits latents pour chaque groupe. Pour $c \in \{1, 2, 3\}$, nous simulons un vecteur contenant des valeurs de trait latent pour chaque groupe c , suivant la distribution $N(\mu_c, \sigma^2)$, avec $\mu = (-3, -1, 1, 3)$ et $\sigma^2 = 0.2$. Pour $j \in \{1, \dots, p\}$, les valeurs de c_{jh} varient uniformément entre -2 et 2 alors que les valeurs a_{jh} suivent la distribution $N(1, 0.1)$. Ces simulations sont effectuées avec la fonction *NRM.sim* du package R *mcIRT*.

M5 : Modèle ordinal de réponse à l'item Nous utilisons encore la théorie de réponse à l'item. Ces modèles permettent de mesurer la probabilité d'observer une modalité pour chaque item en fonction d'un niveau de trait latent. Le trait latent est une variable continue inobservable qui définit la capacité d'un individu, mesurée par des variables observable. Dans le cadre de la théorie de réponse à l'item, les variables, appelées items, sont ordinales. Les observations peuvent être binaire ou polytomiques. Ici, nous présentons un modèle réponse à l'item polytomique, afin de générer des données de manière probabiliste. Le modèle de crédit partiel généralisé [127] est un modèle réponse à l'item adapté aux données ordinales. C'est une extension du modèle logistique à deux paramètres pour données dichotomiques. Le modèle est défini de la manière suivante :

$$p_{jx}(\theta) = P(X_{ij} = x|\theta) = \frac{\exp \sum_{k=0}^x \alpha_j(\theta_i - \beta_{jk})}{\sum_{r=0}^{m_j} \exp \sum_{k=0}^r \alpha_j(\theta_i - \beta_{jk})}, \quad (2.14)$$

où θ_i représente le niveau de trait latent de l'individu i . β_{jk} est un paramètre de seuil de difficulté pour la modalité k de l'item j . Pour $j \in \{1, \dots, p\}$, β_j est un vecteur de dimension $m - 1$. α_j est un paramètre de discrimination représenté par un scalaire.

Nous pouvons générer aléatoirement des jeux de données grâce à ce modèle, en simulant des valeurs de trait latent pour les trois groupes d'individus. Pour $c \in \{1, 2, 3\}$, on simule un vecteur contenant des valeurs de trait latent pour chaque classe c en considérant la distribution $N(\mu_c, \sigma^2)$, avec $\mu = (-3, 0, 3)$ et $\sigma^2 = 0.2$. Pour $j \in \{1, \dots, p\}$, α_j est distribué selon $N(1, 0.1)$, et β_j est un vecteur de valeurs ordonnées qui sont distribuées uniformément entre -2 et 2. Les simulations sont effectuées avec la fonction *rmvordlogis* du package R *ltm* [146].

2.4.3.3 Contrôle de la séparabilité des clusters

Pour chaque modèle de simulation, nous proposons deux configurations de séparabilité des clusters : faible et élevée. Pour le modèle de combinaison linéaire (M1), la probabilité q représente la modalité la plus fréquente dans chaque cluster, c'est elle qui contrôle la séparabilité des clusters. Nous fixons $q = 0.8$ pour obtenir des clusters fortement séparés, et $q = 0.4$ pour des clusters moins séparés.

Pour les deux modèles basés sur une structure d'arbre (M2 et M3), la séparabilité peut être réduite en ajoutant des variables "bruit" dans le jeu de données. Nous utilisons la configuration initiale du modèle pour obtenir des clusters séparés, et nous ajoutons six variables distribuées uniformément sur l'ensemble $\{1, \dots, m\}$ au jeu de données, pour obtenir des clusters moins séparés.

Pour les deux modèles basés sur la théorie de réponse à l'item, la séparation des clusters peut être contrôlée en changeant la distribution du trait latent pour chaque groupe. Pour chaque classe $c \in \{1, 2, 3\}$, nous simulons un vecteur de valeurs de trait latent, avec la distribution $N(\mu_c, \sigma^2)$, avec $\mu = (-3, 0, 3)$ pour obtenir des clusters séparés et $\mu = (-1, 0, 1)$ pour des clusters moins séparés, avec $\sigma^2 = 0.2$ et des paramètres d'items identiques dans les deux cas.

2.4.3.4 Prédiction à l'aide des partitions

La méthode CUBT permet de prédire le cluster d'appartenance de toute nouvelle observation, en se basant sur la structure de l'arbre, suivant l'ensemble des divisions binaires à partir de la racine. Toute approche de clustering peut être utilisée pour faire des prédictions en assignant une nouvelle observation au proche cluster de la partition. Il est cependant nécessaire de définir une mesure de similarité à utiliser pour évaluer la distance entre une observation et un cluster. Nous utilisons une similarité de type lien moyen, c'est-à-dire la similarité moyenne entre la nouvelle observation et toutes les observations du cluster.

2.4.3.5 Critères de comparaison des méthodes

Etant donné que dans les simulations, les clusters sont connus a priori, nous pouvons mesurer la performance des différents algorithmes en utilisant l'indice de Rand ajusté [92] ainsi que l'erreur de mauvaise classification [64]. Soient $y_1, \dots, y_n$ les étiquettes de classe de chaque observation, et soient $\hat{y}_1, \dots, \hat{y}_n$ les étiquettes de classe assignées aux n observations par un algorithme de clustering. L'indice de Rand ajusté est un indice classique calculé à partir d'une table de contingence de y et $\hat{y}$. Il varie entre 0 et 1, les valeurs proches de 1 reflétant des partitions plus similaires. Notons Σ l'ensemble de toutes les permutations possibles de l'ensemble d'étiquettes, le taux d'erreur de mauvaise classification, appelé *matching error*, est défini de la manière suivante :

$$ME = \min_{\sigma \in \Sigma} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{y_i \neq \sigma(\hat{y}_i)\}} \quad (2.15)$$

Pour plus de sept modalités, cet indice peut être calculé de manière efficace en utilisant la méthode hongroise [132].

2.4.3.6 Paramétrage de la méthode

Dans cette section, nous discutons le choix de certains paramètres impliqués dans chaque étape de la méthode CUBT.

Des simulations supplémentaires sont entreprises pour analyser de façon expérimentale l'effet du paramètre *minsize* (voir section 2.2.1) sur la partition finale obtenue par la méthode CUBT. Pour chaque modèle, pour chaque taille d'échantillon et pour les deux configurations de séparabilité des clusters (élevée et faible), nous testons 15 valeurs de *minsize* sur une échelle de $\lfloor \log(n) \rfloor$ à $\lfloor n/4 \rfloor$. Pour chaque cas, 20 jeux de données sont générés pour former les échantillons d'apprentissage et les échantillons tests. Nous calculons la déviance, l'utilité des classes, l'erreur de mauvaise classification pour l'arbre maximal, pour l'arbre élagué et pour l'arbre obtenu après le regroupement des feuilles. Voici les principales conclusions de ces simulations :

- Pour tous les cas, la valeur optimale du *minsize* est la même en fonction de la déviance et l'utilité des classes. En effet, cela vient de la forte relation qu'il y a entre l'utilité des classes et la déviance [73]. Cette relation est illustrée dans les tables 2.9 et 2.10.

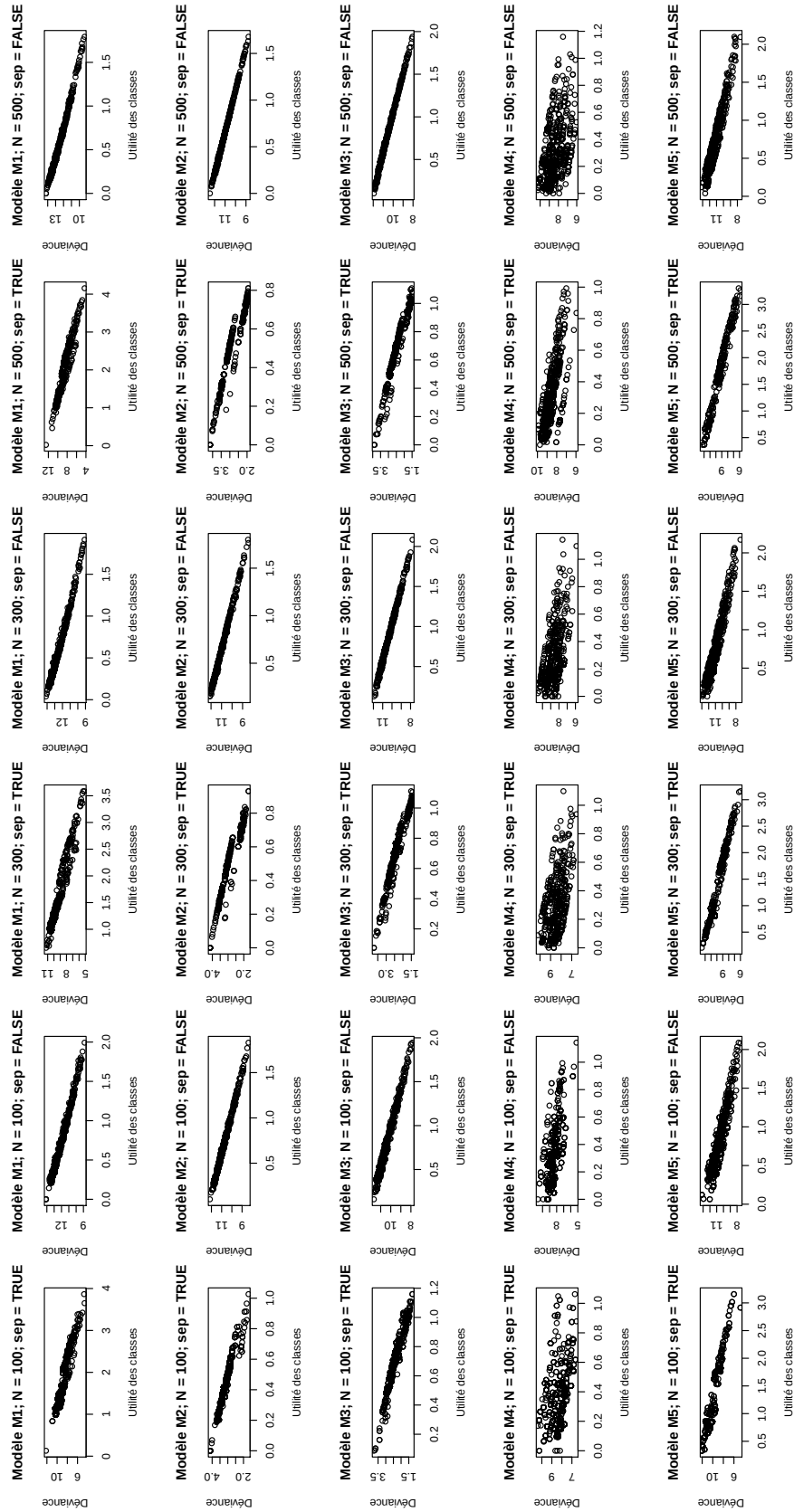


Figure 2.9 – Relation entre la déviance et l'utilité des classes (en utilisant l'information mutuelle comme mesure de dissimilarité). Chaque graphique correspond à l'analyse du lien entre les deux mesures, pour un modèle donné, une taille d'échantillon fixée, et une configuration de séparabilité des classes donnée.

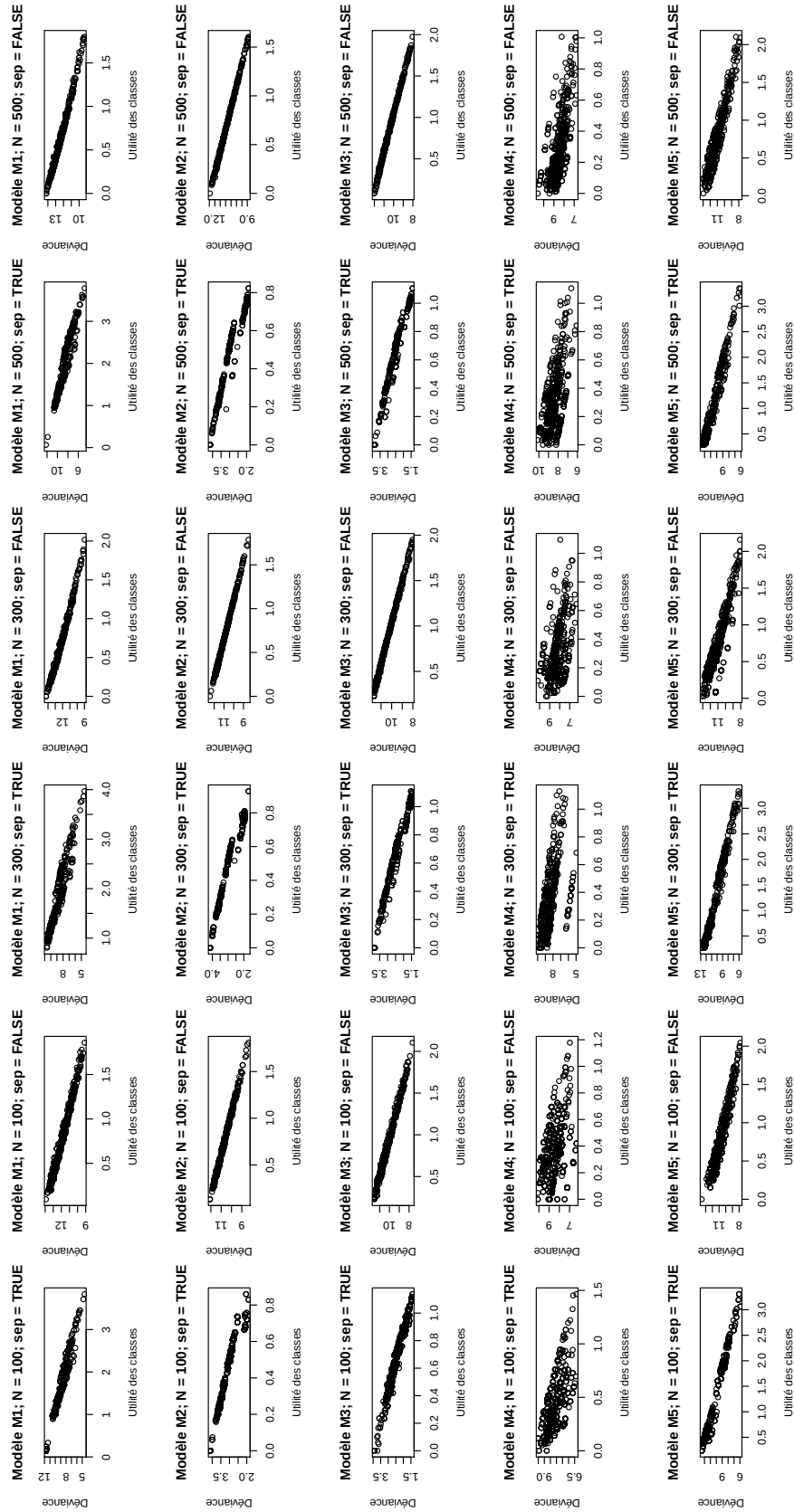


Figure 2.10 – Relation entre la déviance et l'utilité des classes (en utilisant la distance de Hamming comme mesure de dissimilarité). Chaque graphique correspond à l'analyse du lien entre les deux mesures, pour un modèle donné, une taille d'échantillon fixée, et une configuration de séparabilité des classes donnée.

- Une ANOVA a été effectuée pour évaluer la différence entre les taux d'erreur de mauvaise classification des arbres élagués en utilisant l'information mutuelle comme mesure de dissimilarité, et ceux des arbres élagués avec la distance de Hamming. Des différences significatives sont observées entre les modèles, mais il y a très peu de différence entre les deux configuration de séparabilité des clusters. Pour tous les cas ces différences ne dépassent jamais 0.05. Pour les modèles M1 et M5, l'élagage avec l'information mutuelle fournit de meilleurs résultats. Pour les autres modèles, l'élagage avec la distance de Hamming est meilleur, mais les différences avec l'information mutuelle sont très faibles (< 0.012). La figure 2.11 montre les liens entre le taux d'erreur obtenu avec la distance de Hamming et celui obtenu avec l'information mutuelle.

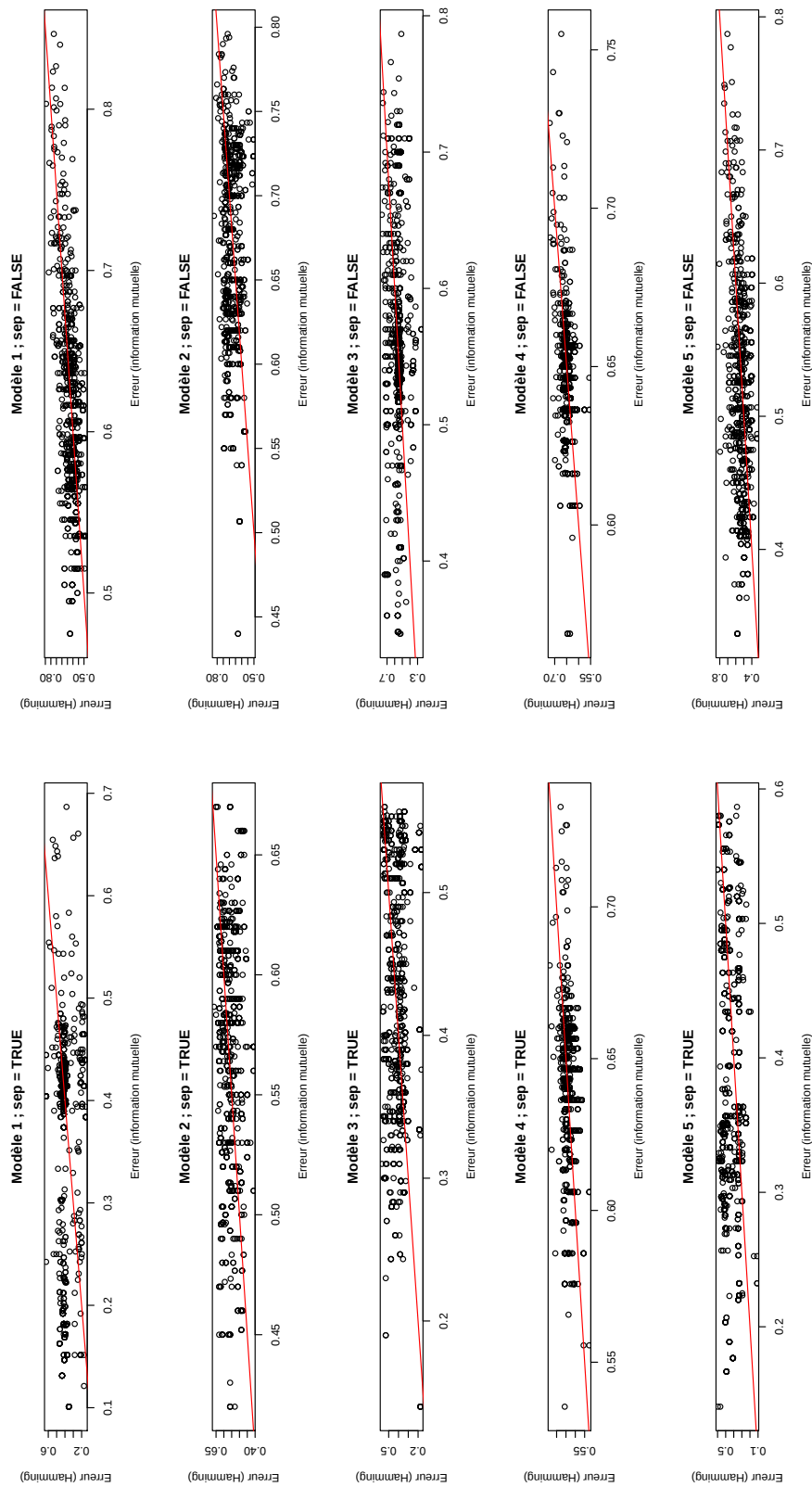


Figure 2.11 – Relation entre les taux d'erreur obtenus avec la distance de Hamming et l'information mutuelle. Chaque graphique correspond à l'erreur obtenue avec la distance de Hamming, en fonction de l'erreur obtenue avec l'information mutuelle, pour un modèle donné, une configuration de séparabilité des classes donnée, et pour toutes les tailles d'échantillon confondues.

- Les valeurs du paramètre *minsize* qui minimisent la déviance totale des arbres après regroupement des feuilles sur les échantillons test, sont très souvent les mêmes que celles minimisant l'erreur de mauvaise classification. Lorsqu'elles sont différentes, elles sont toujours très proches. Cela signifie que, dans nos simulations, optimiser la déviance totale sur les échantillons tests est équivalent à optimiser l'erreur de mauvaise classification. Ces résultats sont illustrés par la figure 2.12.

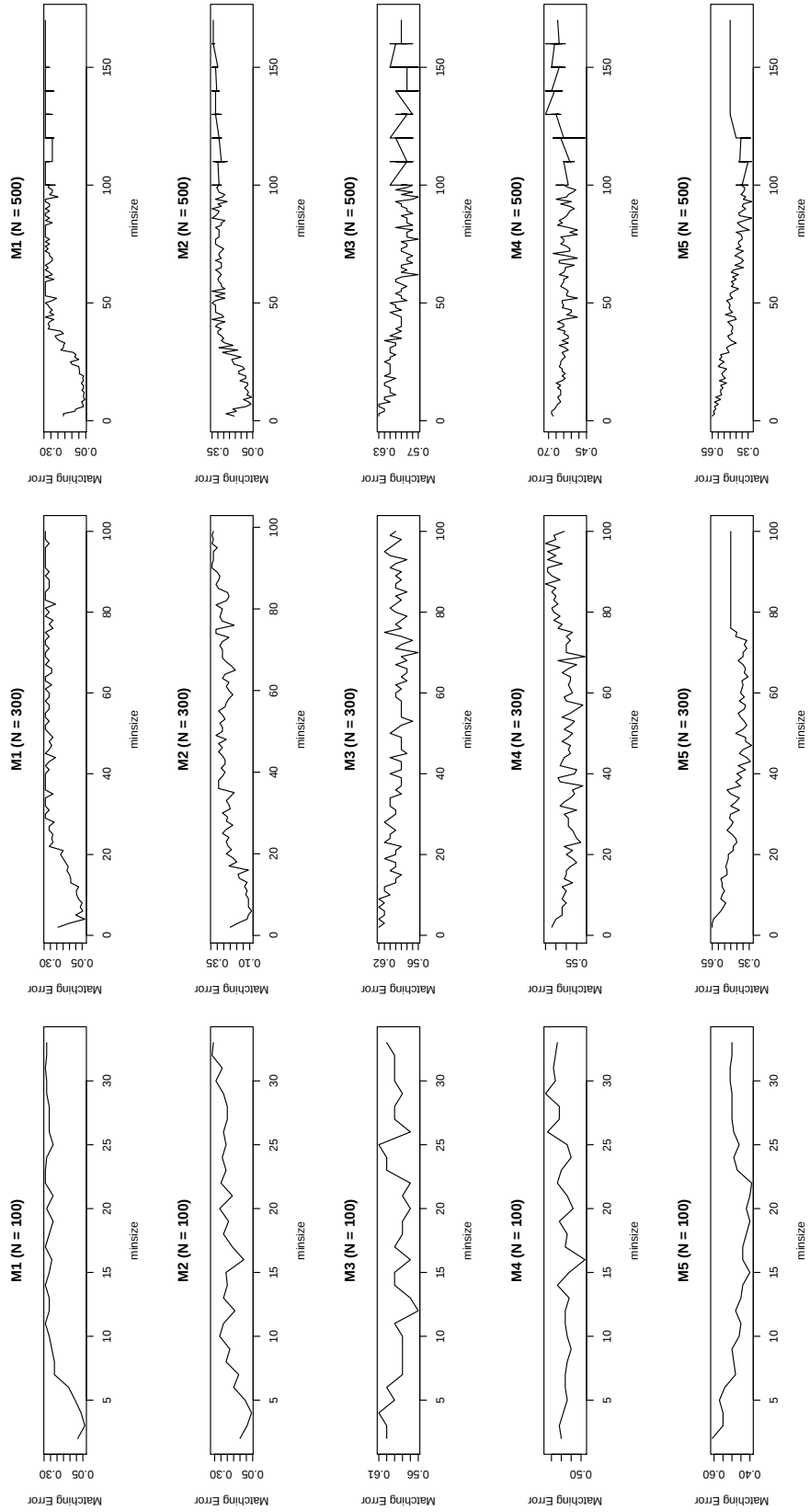


Figure 2.12 – Taux d'erreur de mauvaise classification en fonction de la valeur du paramètre *minsize*. Les résultats sont fournis pour chaque modèle et chaque taille d'échantillon

- Concernant la contribution de l'étape de regroupement des feuilles, les différences entre les taux d'erreur de mauvaise classification entre les arbres élagués et les arbres après regroupement sont souvent très significatives, quel que soit le modèle, la taille d'échantillon ou le niveau de séparabilité des clusters. Cela signifie que cette étape est efficace dans la plupart des cas.
- La valeur optimale du paramètre *minsize* augmente proportionnellement à la taille de l'échantillon dans tous les cas.

Finalement, nous choisissons les différents paramètres de CUBT de la façon suivante. Nous effectuons une validation croisée à 10 répétitions de la déviance totale de l'arbre obtenu après regroupement des feuilles, et nous choisissons la valeur de *minsize* qui minimise ce critère. Le paramètre *mindev* est fixé à 0.001. Pour l'étape d'élagage, nous fixons $\delta = 0.03$. Dans leur papier initial [64], les auteurs ont montré que CUBT est très sensible au paramètre *mindist* pour l'élagage. Ainsi, au lieu de calculer la mesure de dissimilarité pour toutes les paires de noeuds issus de la division d'un même noeud, au lieu de choisir une valeur de *mindist*, l'utilisateur peut spécifier une valeur pour le paramètre *qdist*, qui correspond au quantile de la distribution empirique des dissimilarités, sur l'arbre maximal. Le seul paramètre devant être fixé dans l'étape de regroupement des feuilles est le nombre de classes k , cela ne présente aucune différence avec la version précédente de la méthode CUBT (voir 2.2.3).

2.4.3.7 Résultats

La table 2.3 présente les résultats de prédiction (calculés sur les échantillons tests) pour tous les modèles de simulation de données, la table 2.4 présente les résultats d'apprentissage (calculés sur les échantillons d'apprentissage). Nous fournissons les taux d'erreur de mauvaise classification, sous la forme de pourcentage obtenu pour chaque méthode, moyennés sur les 100 répliques. Les résultats basés sur l'indice de Rand ajusté sont omis volontairement étant donné qu'ils apportent exactement les mêmes conclusions que le taux d'erreur.

En termes de prédiction, CUBT est souvent une des deux méthodes les plus performantes, pour les modèles M1, M3 et M5, quel que soit la taille d'échantillon et la configuration de la séparabilité des clusters. Pour le modèle M2, CUBT dépasse aussi les autres méthodes, sauf pour $N = 500$ et une séparabilité des clusters élevée, dans ce cas ce sont la méthode de classification hiérarchique ainsi que DIVCLUS-T qui sont les méthodes les plus performantes. Pour le modèle M4, la classification hiérarchique, k -modes et DIVCLUS-T sont parmi les deux méthodes les plus performantes, elles surpassent CUBT. Les deux méthodes pour élaguer (information mutuelle et distance de Hamming) donnent des résultats très similaires dans la majorité des cas.

Model	Sep	N	<i>k</i> -modes	HCA	DBSCAN	DIVCLUS-T	LCA	COBWEB	CUBT ^{MI}	CUBT ^{Ham}
M1	High	100	0.610	0.610	0.640	0.610	0.610	0.640	0.230	0.180
		300	0.630	0.640	0.660	0.630	0.630	0.660	0.180	0.160
		500	0.640	0.640	0.660	0.640	0.640	0.690	0.220	0.150
	Low	100	0.610	0.620	0.650	0.620	0.620	0.710	0.540	0.540
		300	0.640	0.640	0.680	0.640	0.640	0.740	0.570	0.550
		500	0.650	0.650	0.680	0.640	0.650	0.740	0.560	0.540
M2	High	100	0.650	0.650	0.750	0.650	0.660	0.670	0.610	0.580
		300	0.660	0.650	0.750	0.670	0.670	0.690	0.640	0.610
		500	0.670	0.650	0.750	0.660	0.670	0.680	0.670	0.630
	Low	100	0.670	0.670	0.730	0.670	0.670	0.740	0.660	0.670
		300	0.700	0.700	0.720	0.700	0.700	0.780	0.700	0.700
		500	0.710	0.710	0.720	0.710	0.710	0.780	0.700	0.710
M3	High	100	0.620	0.600	0.750	0.620	0.630	0.670	0.540	0.440
		300	0.630	0.600	0.750	0.610	0.630	0.690	0.560	0.500
		500	0.630	0.600	0.750	0.610	0.620	0.700	0.600	0.520
	Low	100	0.670	0.670	0.700	0.660	0.670	0.740	0.630	0.600
		300	0.690	0.690	0.710	0.660	0.690	0.770	0.600	0.590
		500	0.690	0.690	0.720	0.650	0.680	0.760	0.590	0.570
M4	High	100	0.610	0.610	0.640	0.620	0.610	0.650	0.590	0.590
		300	0.630	0.630	0.650	0.630	0.630	0.660	0.610	0.610
		500	0.640	0.640	0.660	0.640	0.640	0.660	0.610	0.610
	Low	100	0.610	0.610	0.640	0.610	0.610	0.650	0.610	0.610
		300	0.640	0.640	0.660	0.640	0.640	0.660	0.630	0.630
		500	0.640	0.640	0.660	0.640	0.640	0.660	0.640	0.640
M5	High	100	0.590	0.650	0.600	0.590	0.580	0.570	0.220	0.230
		300	0.590	0.650	0.620	0.600	0.600	0.590	0.220	0.200
		500	0.590	0.650	0.610	0.600	0.600	0.600	0.230	0.190
	Low	100	0.610	0.610	0.630	0.610	0.620	0.660	0.480	0.480
		300	0.630	0.630	0.650	0.630	0.640	0.670	0.460	0.470
		500	0.630	0.630	0.650	0.630	0.640	0.660	0.460	0.480

Table 2.3 – Résultats de prédiction : Moyenne des taux d'erreur sur 100 répliques, pour les cinq modèles de simulation, les différentes tailles d'échantillons (N), en fonction du niveau de séparabilité (sep) des clusters. Les valeurs en gras correspondent aux plus faibles taux d'erreur pour chaque cas. CUBT^{MI} et CUBT^{Ham} désignent CUBT en utilisant respectivement l'information mutuelle et la distance de Hamming pour l'élagage.

2.4.4 Application à des données réelles : les maladies du soja

Dans cette section, nous présentons une application de CUBT sur la version courte de la célèbre base de données *Soybean*, constituée par Michalski [118]. Ce jeu de données contient 47 observations de 35 variables nominales. Chaque ligne du jeu de données représente un échantillon de soja, qui est décrit par des attributs tels que la précipitation, la température ou encore l'état de la racine. Les observations sont réparties dans quatre groupes distincts, chaque groupe représentant une maladie du soja : Diaporthe (10 observations), Pourriture charbonneuse (10 observations), Rhizoctonie (10 observations), Phytophthora (17 observations).

Dans un premier temps, nous appliquons CUBT à l'échantillon total, en fixant la valeur du paramètre *minsize* = 10. Les quatre classes sont toutes découvertes par CUBT, et sont assignées à des noeuds terminaux de l'arbre maximal obtenu (voir partie gauche de la figure 2.13). La division binaire principale trouvée par CUBT permet de séparer les observations qui diffèrent au regard de la présence de chancres ($x_{21} \in \{0, 3\}$). Les deux autres divisions définies sur les deux sous-ensemble obtenus sont de la forme $x_3 \in \{0\}$ et $x_{22} \in \{1\}$. Ces divisions discriminent les observations en fonction du type de précipitation et en fonction de la couleur de la lésion du chancre respectivement. L'interprétation de cette partition est en accord avec les résultats de Fisher, basés sur l'algorithme COBWEB [60]. Par exemple, le cluster "Pourriture charbonneuse" est défini par Fisher par une absence de chancres, variable également prise en compte par CUBT dans la première division binaire. L'autre attribut utilisé par Fisher pour décrire ce cluster est une précipitation en-dessous de la normale, qui est également pris en compte par CUBT. La principale différence est que CUBT utilise seulement deux variables pour retrouver les clusters alors que huit variables sont nécessaires dans la description fournie par Fisher, basée sur l'algorithme COBWEB.

La partie droite de la figure 2.13 montre l'arbre de clustering obtenu avec DIVCLUS-T. Une partition parfaite est trouvée avec le même jeu de données, mais elle présente quelques différences au niveau de sa structure, avec la partition obtenue avec CUBT.

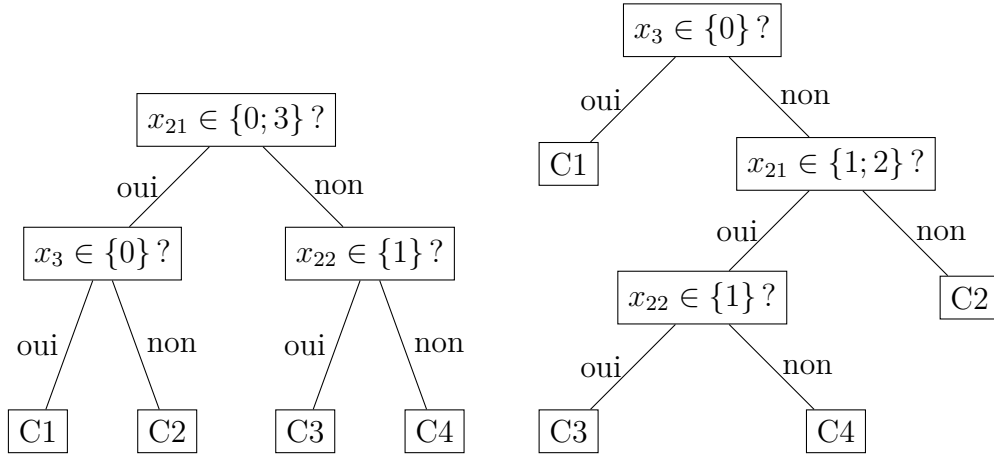


Figure 2.13 – Structures des arbres obtenus avec CUBT (gauche) et DIVCLUS-T (droite) pour le jeu de données Soybean.

On s'intéresse maintenant à la performance de CUBT, en utilisant l'information mutuelle et la distance de Hamming pour l'élagage, en termes d'apprentissage et de prédiction. On utilise encore le jeu de données Soybean, et on cherche à comparer les résultats obtenus par CUBT avec les méthodes de clustering utilisées dans la section 2.4.3. Nous générons 100 échantillons aléatoires stratifiés, en prenant à chaque itération deux tiers des observations pour former un échantillon d'apprentissage, les observations restantes sont utilisées pour former un échantillon test. Pour CUBT avec l'élagage basé sur l'information mutuelle (CUBT^{MI}) et la distance de Hamming (CUBT^{Ham}), nous sélectionnons la valeur de *minsize* qui minimise la déviance totale de l'arbre après regroupement des feuilles. Les valeurs obtenues sont respectivement 5 et 3. Les résultats d'apprentissage et de prédiction de ces expériences sont présentés à la table ???. L'élagage utilisant la distance de Hamming produit de meilleurs résultats que l'information mutuelle, pour l'apprentissage et la prédiction. CUBT^{Ham} et DIVCLUS-T sont les deux méthodes les plus performantes sur les échantillons d'apprentissage alors que CUBT^{Ham} et CUBT^{MI} sont les plus performantes sur les échantillons test (prédiction). Le taux d'erreur de mauvais classement vaut 0 pour CUBT^{Ham} en apprentissage, et 0.39 en prédiction, alors que les taux d'erreur pour les autres méthodes varient entre 0 et 0.87 pour l'apprentissage et de 0.39 à 0.56 pour la prédiction.

2.4.5 Conclusion

Nous avons présenté une extension de l'algorithme CUBT pour les données nominales, qui utilise des critères adaptés pour gérer ce type de données. Ces critères sont basés sur l'entropie de Shannon et l'information mutuelle. Nous

Model	Sep	N	k -modes	HCA	DBSCAN	DIVCLUS-T	LCA	COBWEB	CUBT ^{MI}	CUBT ^{Ham}
M1	High	100	0.019	0.051	0.460	0.084	0.059	0.490	0.160	0.097
		300	0.012	0.150	0.550	0.130	0.019	0.400	0.150	0.130
		500	0.011	0.270	0.520	0.130	0.019	0.340	0.190	0.130
	Low	100	0.440	0.570	0.640	0.490	0.470	0.920	0.510	0.510
		300	0.370	0.630	0.630	0.490	0.370	0.950	0.550	0.510
		500	0.330	0.640	0.640	0.480	0.290	0.960	0.540	0.520
M2	High	100	0.540	0.520	0.750	0.520	0.520	0.580	0.600	0.560
		300	0.580	0.510	0.750	0.550	0.550	0.610	0.630	0.610
		500	0.580	0.510	0.750	0.550	0.570	0.620	0.670	0.630
	Low	100	0.610	0.630	0.740	0.610	0.600	0.830	0.660	0.660
		300	0.650	0.690	0.710	0.660	0.640	0.850	0.690	0.690
		500	0.660	0.710	0.740	0.650	0.650	0.840	0.700	0.700
M3	High	100	0.410	0.460	0.750	0.460	0.450	0.620	0.530	0.430
		300	0.440	0.470	0.750	0.470	0.450	0.630	0.560	0.490
		500	0.430	0.490	0.750	0.480	0.450	0.630	0.600	0.520
	Low	100	0.550	0.600	0.690	0.500	0.560	0.810	0.610	0.580
		300	0.580	0.670	0.730	0.470	0.530	0.820	0.590	0.580
		500	0.580	0.690	0.740	0.480	0.500	0.820	0.590	0.570
M4	High	100	0.580	0.590	0.640	0.600	0.580	0.640	0.580	0.590
		300	0.600	0.630	0.650	0.600	0.590	0.650	0.590	0.600
		500	0.610	0.640	0.660	0.600	0.590	0.650	0.590	0.600
	Low	100	0.600	0.610	0.640	0.620	0.600	0.650	0.600	0.600
		300	0.630	0.640	0.660	0.630	0.630	0.650	0.630	0.630
		500	0.630	0.650	0.660	0.640	0.640	0.660	0.630	0.640
M5	High	100	0.063	0.270	0.140	0.061	0.005	0.210	0.077	0.089
		300	0.070	0.350	0.180	0.071	0.016	0.200	0.087	0.068
		500	0.064	0.370	0.150	0.074	0.002	0.190	0.079	0.110
	Low	100	0.430	0.540	0.560	0.400	0.290	0.550	0.370	0.380
		300	0.440	0.610	0.610	0.400	0.200	0.530	0.370	0.450
		500	0.440	0.620	0.640	0.400	0.170	0.500	0.390	0.460

Table 2.4 – Résultats d'apprentissage : Moyenne des taux d'erreur sur 100 répliques, pour les cinq modèles de simulation, les différentes tailles d'échantillons (N), en fonction du niveau de séparabilité (sep) des clusters. Les valeurs en gras correspondent aux plus faibles taux d'erreur pour chaque cas. CUBT^{MI} et CUBT^{Ham} désignent CUBT en utilisant respectivement l'information mutuelle et la distance de Hamming pour l'élagage.

avons comparé cette approche à d'autres méthodes classiques, en utilisant différents modèles de simulation de données. Parmi les méthodes comparées, CUBT montre une excellente performance et surpasse la majorité des méthodes auxquelles il est comparé, particulièrement sur les échantillons test. Dans tous les cas, CUBT présente toujours trois avantages pratiques sur les autres méthodes : la partition obtenue est interprétable grâce aux variables utilisées pour former les divisions binaires, l'arbre obtenu peut être utilisé pour assigner facilement des nouvelles observations aux clusters, et l'algorithme est parallélisable, donc utilisable sur de grands jeux de données.

Nous avons également fourni quelques heuristiques concernant le choix des paramètres de CUBT : la taille minimale des feuilles de l'arbre maximal et la distance minimale d'élagage. Nous avons justifié ces choix grâce à des simulations supplémentaires sur différents modèles. Ces méthodes de paramétrage peuvent également être utilisées pour des jeux de données continues. CUBT fournit également d'excellents résultats sur un jeu de données réelles, en comparaison à d'autres méthodes.

Les futurs développements de CUBT devraient se concentrer sur de nouveaux critères permettant de prendre en compte des données de type mixte, en utilisant des critères additifs adaptés aux données mixtes [93, 115]. Un autre champ de recherche est aussi de grande importance, c'est la mesure de l'importance des variables via CUBT. Le chapitre suivant est consacré à ce problème.

3- Sélection de variables en apprentissage non supervisé

Le problème de sélection de variables est un problème connu comme étant un problème d'optimisation combinatoire. Sa formulation générale qu'on retrouve également en dehors du cadre des statistiques peut être faite de la façon suivante :

On suppose que l'on dispose d'un ensemble de n observations indépendantes d'un vecteur aléatoire de dimension p , noté $Z = (Z_1, Z_2, \dots, Z_p)$, et d'un critère $J(Z)$, où J est une fonction à valeurs dans $\mathbb{R}$ et qui dépend des n observations de Z . On souhaite déterminer un sous-ensemble de taille $p' < p$ de composantes de Z , tel que la valeur du critère J calculé uniquement avec ces p' composantes soit proche de $J(Z)$. En statistiques, ce problème a été abordé essentiellement dans un cadre où Z est constitué d'un couple (X, Y) de variables aléatoires, où X est un vecteur de variables explicatives indépendantes dans $\mathbb{R}^p$ et Y une variable unidimensionnelle dépendant de X . C'est le contexte de l'apprentissage supervisé : le critère J est dans ce cas une fonction de perte associée à un estimateur de la fonction de régression f de Y sur X . Si Y est continue, J peut être l'erreur quadratique de la régression, et si Y est discret J peut être par exemple le taux d'erreur du prédicteur construit. Lorsque l'on ne dispose pas de variable dépendante Y , en apprentissage non supervisé, le critère J dépend de la technique utilisée. Le problème peut se poser essentiellement dans le cadre de la classification automatique.

Ce chapitre s'intéresse à la sélection de variables en apprentissage non supervisé, particulièrement aux différentes approches pour mesurer l'importance des variables en classification non supervisée. Nous proposons d'utiliser la méthode CUBT, une méthode de clustering hiérarchique adaptée aux données continues et nominales. Nous considérons une mesure de l'importance des variables pour cette méthode, similaire à celle utilisée dans les arbres de classification et de régression de Breiman. Le score d'importance des variables peut être utilisé pour obtenir un ordonnancement des variables d'un jeu de données, afin de déterminer quelles variables sont les plus importantes, ou à l'inverse pour détecter celles qui ne sont pas pertinentes. Nous étudions la stabilité et l'efficacité de ce score sur différents modèles de simulation de données, en présence de bruit, et nous le comparons à d'autres mesures classiques d'importance des variables. Les résultats de ces expériences montrent que notre approche est beaucoup plus efficace que d'autres, dans de nombreuses situations.

3.1 Introduction

Dans de nombreuses applications en modélisation statistique et en analyse de données, la mesure de l'importance des variables via l'attribution d'un score d'importance est essentielle. Celle-ci peut être utilisée pour diminuer la dimensionnalité des données ou pour la sélection de variables [109], afin de réduire la complexité des modèles, ou réduire le bruit présent dans les données, ou encore améliorer à la fois la précision et l'interprétabilité des modèles. Généralement, un score est calculé pour chaque variable, en fonction d'un modèle ou d'une tâche spécifique.

En apprentissage supervisé, l'importance d'une variable est souvent liée à sa corrélation (ou sa dépendance) avec la variable à expliquer Y . Elle peut être calculée pendant la phase d'apprentissage du modèle ou après que le modèle soit estimé. Les approches supervisées les plus connues proposant un tel indice sont les arbres de classification et de régression (CART, [29]), les forêts aléatoires (*Random Forests*, RF, [32]) et les machines à vecteurs supports (*Support Vector Machines*, SVM, [77, 140]), dans lesquelles l'importance des variables est fortement liée à la structure et à la précision du modèle.

L'apprentissage non supervisé concerne des applications telles que la classification non supervisée de données et l'estimation de la densité. Nous nous focalisons ici sur les méthodes de clustering visant à construire une partition d'un ensemble de n observations en k clusters, où k est spécifié a priori ou déterminé automatiquement par la méthode employée. En classification non supervisée, il y a un grand intérêt à déterminer quelles variables sont les plus importantes pour obtenir une partition d'un jeu de données.

Le but de ce chapitre est de définir l'importance des variables basée sur la méthode CUBT. Le rôle des variables importantes est analysé grâce à des modèles de simulation de données, et des comparaisons de CUBT avec d'autres méthodes classiques. Des résultats concernant la stabilité et l'efficacité du score d'importance des variables sont fournis. La section 3.2 présente quelques méthodes classiques pour mesurer l'importance des variables en classification non supervisée. La section 3.3 fournit une brève description de la méthode CUBT. La section 3.3 présente la nouvelle méthodologie pour mesurer l'importance des variables basée sur la méthode CUBT. Enfin, la section 3.4 présente des expérimentations, accompagnées de résultats, avec différents modèles de simulation de données et quelques méthodes de clustering.

La classification non supervisée basée sur les arbres de décision binaires (la méthode CUBT [ghattas_clustering_2016, 64]) est une méthode hiérarchique descendante non paramétrique, adaptée aux données continues et nominales. Un arbre de décision construit avec CUBT peut être utilisé pour identifier quelles variables du jeu de données sont actives et participent directement à la construction de l'arbre maximal. Cependant, bien que certaines variables du jeu de données puissent être inutiles pour la construction de l'arbre, elles peuvent être compé-

Method	<i>Sepal.Length</i>	<i>Sepal.Width</i>	<i>Petal.Length</i>	<i>Petal.Width</i>
CUBT	3.24	0.14	3.97	3.82
URF	17.33	15.46	17.38	12.57
LOVO-cubt	0.69	0.81	0.59	0.83
LOVO-km	0.55	0.53	0.57	0.53
TWKm	2.5×10^{-6}	1.3×10^{-5}	2.5×10^{-6}	0.99
LS	0.18	0.40	0.03	0.08

Table 3.1 – Les quatre variables du jeu de données Iris et leurs scores d'importance respectifs.

titives vis-à-vis des variables actives, pour différentes divisions binaires définies dans l'arbre. Pour de nombreuses applications, il peut être très utile d'identifier ces variables importantes. Un exemple d'arbre obtenu avec CUBT, sur le célèbre jeu de données Iris [fisher_the_1936] est présenté dans la figure 3.1. La variable classe représentant l'espèce d'Iris, à trois modalités (*setosa*, *virginica*, *versicolor*), n'est pas utilisée ici.

La table 3.1 présente le score d'importance de chaque variable, calculée avec la méthode CUBT, les forêts aléatoires non supervisées (URF), la méthode "leave-one-variable-out" basée sur CUBT et sur k -means (LOVO-CUBT et LOVO-KM), l'algorithme TW- k -means (TWKM) et le score laplacien. Ces méthodes sont décrites dans la suite du chapitre. En analysant les résultats obtenus avec CUBT, on voit que la seule variable active (*Petal.Length*) obtient le plus haut score. Ce résultat est également observé pour les méthodes URF et LOVO-KM. Les autres variables (*Sepal.Length*, *Sepal.Width* et *Petal.Width*) ne sont pas actives dans l'arbre de clustering, cependant deux d'entre elles ont des scores d'importance élevés.

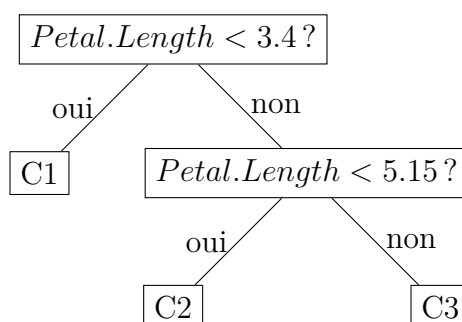


Figure 3.1 – Structure de l'arbre obtenu par CUBT avec le jeu de données Iris. La seule variable active est *Petal.Length*.

3.2 Importance des variables en apprentissage non supervisé

Dans cette section, nous présentons quelques approches permettant d'obtenir des scores d'importance pour les variables, dans le contexte de la classification non supervisée. Deux de ces approches sont très liées à la méthode de clustering sous-jacente, à savoir les forêts aléatoires non-supervisées et l'algorithme TW- k -means. Une troisième approche, le score laplacien, calcule un score d'importance des variables indépendamment de la méthode de clustering utilisée. Finalement, nous proposons une approche originale et intuitive, appelée "leave-one-variable-out" (LOVO), qui est adaptée à n'importe quelle méthode de clustering.

3.2.1 Les forêts aléatoires non-supervisées

Les forêts aléatoires non-supervisées [32] et leur mesure d'importance des variables sont une des méthodes supervisées les plus connues pour la régression et la classification. Breiman [31] a suggéré d'utiliser les forêts aléatoires dans le cadre de la classification non supervisée, en se rapportant à un problème de classification supervisée binaire : les observations disponibles sont assignées à un premier groupe. Le second groupe contient des observations qui sont générées par échantillonnage aléatoire indépendant à partir des distributions marginales de chaque variable du premier groupe. La classification non supervisée est alors transformée en un problème de classification supervisée binaire, et l'importance des variables est calculée aisément. Dans les forêts aléatoires, l'importance d'une variable correspond à la perte moyenne de l'indice de Gini à chaque division binaire de l'arbre, en utilisant cette variable dans chaque division de chaque arbre de la forêt. Les forêts aléatoires non supervisées sont adaptées aux données continues et nominales.

3.2.2 L'algorithme TW- k -means

L'algorithme pondéré TW- k -means (TWKM, [42]) est une méthode de classification non-supervisée qui peut assigner un poids à chaque variable du jeu de données. Cet algorithme est une extension de l'algorithme de clustering k -means pondéré basé sur l'entropie, qui lui-même est une extension de l'algorithme k -means [113]. Dans un premier temps, un ensemble de k centres est choisi aléatoirement. Les observations sont alors assignées à leur centre le plus proche, au regard d'une certaine mesure de dissimilarité (choisie par l'utilisateur en fonction des données traitées). Ensuite les nouveaux centres sont recalculés, en prenant en compte les observations ajoutées à chaque cluster. Des poids sont pris en compte dans le calcul de la mesure de dissimilarité choisie. Ce processus est répété jusqu'à convergence du critère d'hétérogénéité intra-classe.

3.2.3 Le score Laplacien

Le score Laplacien [19] est un indice populaire utilisé pour obtenir un classement des variables d'un jeu de données en apprentissage non-supervisé. On commence par construire une matrice de similarités inter-individus. On seuille ensuite cette matrice de manière à obtenir une matrice S telle que pour tous i, j , $s_{ij} = s(i, j)$ si $s(i, j) > s_0$ et $s_{ij} = 0$ sinon, où s est choisie selon la nature des données et s_0 est un seuil défini par l'utilisateur. La matrice S est donc une matrice d'adjacence. Soit $D = \text{diag}(S\mathbf{1})$ (où $\mathbf{1}$ est un vecteur identité) la matrice diagonale contenant pour chaque observation la somme des similarités avec les k plus proches voisins, c'est-à-dire $D = \sum_{i=1}^n E_i S \mathbf{1} e_i$, où E_i est une matrice $n \times n$ contenant la valeur 1 sur chaque case de sa diagonale, et 0 dans les autres, et e_i une ligne d'une matrice de dimension $1 \times n$ valant 1 pour la colonne i , et 0 sinon. La matrice Laplacienne est alors définie par $L = D - S$, et chaque variable $X_{.j}$, pour $j \in \{1, \dots, p\}$ est "centrée" de la manière suivante :

$$\tilde{X}_{.j} = X_{.j} - \frac{X'_{.j} D \mathbf{1}}{\mathbf{1}' D \mathbf{1}} \mathbf{1}. \quad (3.1)$$

Finalement, le score Laplacien (ou plus généralement l'importance) de la variable j est calculé de la façon suivante :

$$\text{Imp}(X_{.j}) = \frac{\tilde{X}'_{.j} L \tilde{X}_{.j}}{\tilde{X}'_{.j} D \tilde{X}_{.j}}. \quad (3.2)$$

Le score Laplacien assigne des scores élevés aux variables qui respectent le mieux la structure du graphe des plus proches voisins. Il est souvent utilisé dans les méthodes de filtrage de sélection de variables. Comparé à d'autres scores plus classiques, tels que le score de Fisher, le score Laplacien a montré de meilleures performances [180].

3.2.4 Le score "leave-one-variable-out"

Nous proposons une méthode intuitive d'évaluation de l'importance des variables, qui peut être utilisée avec n'importe quelle approche de classification non-supervisée. Admettons qu'une partition $\mathcal{P}$ soit obtenue avec une quelconque méthode, en utilisant toutes les variables contenues dans un jeu de données. On peut alors calculer un critère d'hétérogénéité intra-classes à partir de cette partition, noté $R(\mathcal{P})$. Notons ensuite $\mathcal{P}^{-j}$ la partition obtenue en utilisant la même méthode, mais après avoir supprimé la variable j du jeu de données, et $R(\mathcal{P}^{-j})$ la mesure d'hétérogénéité intra-classes correspondante. L'importance de la variable j peut être alors définie de la manière suivante :

$$\text{Imp}(X_{.j}) = R(\mathcal{P}^{-j}) \quad (3.3)$$

Nous appelons cette méthode LOVO pour "leave-one-variable-out", puisqu'elle consiste à recalculer le critère d'hétérogénéité pour chaque variable retirée une-à-une du jeu de données. Elle peut être utilisée aussi bien pour des données continues que qualitatives. Dans la suite de ce chapitre, cette méthode sera appliquée avec la méthode CUBT pour données continues et nominales, ainsi qu'avec les algorithmes k -means et k -modes.

3.3 Importance des variables avec la méthode CUBT

La méthode CUBT, qui est décrite de manière détaillée dans le chapitre précédent [64], est une méthode de classification non supervisée inspirée par la méthode CART [29]. Les partitions obtenues avec cette méthode peuvent être représentées par des arbres de décision binaires, interprétables grâce aux divisions sur les variables. CUBT peut être utilisée pour assigner de nouvelles observations aux différents clusters. De plus l'algorithme est parallélisable, ce qui peut être très avantageux lorsque l'on dispose de jeux de données volumineux.

Le score d'importance des variables avec la méthode CUBT est fortement lié à l'algorithme de construction de l'arbre maximal, et peut être calculé dans le cas continu et nominal. Nous commençons donc par en rappeler les éléments essentiels pour définir ce score.

3.3.1 Rappels sur la méthode CUBT

Soit $X \in E = \prod_{j=1}^p \text{sup}(j)$, un vecteur aléatoire p -dimensionnel dont les coordonnées sont $X_{.j}$, $j \in \{1, \dots, p\}$, et $\text{sup}(j)$ est le support de la variables j , c'est-à-dire l'ensemble des valeurs que la variable peut prendre. Nous considérons un ensemble S de n observations aléatoires de X , notées X_i , avec $i \in \{1, \dots, n\}$ et X_{ij} l'observation i de la composante j de X . Des notations similaires en lettres minuscules sont utilisées pour désigner les réalisations de ces variables : x , x_i , $x_{.j}$ and x_{ij} .

Pour tout noeud t (c'est-à-dire un sous-ensemble de S), soit n_t le nombre d'observations contenues dans t . Soit $X^{(t)}$ la restriction de X au noeud t , c'est-à-dire $X^{(t)} = \{X | X \in t\}$. On définit alors $R(t)$, la mesure d'hétérogénéité (appelée aussi *déviance*) de t pour les données continues comme suit :

$$R(t) = \frac{n_t}{n} \text{trace}(\text{cov}(X^{(t)})) = \frac{\sum_{X_i \in t} \|X_i - \bar{X}_t\|^2}{n}, \quad (3.4)$$

où $\text{cov}(X^{(t)})$ est la matrice variance-covariance de $X^{(t)}$ et

$$\bar{X}_t = \frac{\sum_{X_i \in t} X_i}{n_t}. \quad (3.5)$$

Pour les données nominales, la mesure d'hétérogénéité de t est définie de la manière suivante :

$$R(t) = \text{trace}(\mathbf{MI}(X^{(t)})) = - \sum_{j=1}^p \sum_{l \in \text{sup}(j)} p_{lj}^{(t)} \log_2 p_{lj}^{(t)}, \quad (3.6)$$

où $\mathbf{MI}(X^{(t)})$ est la matrice d'information mutuelle de $X^{(t)}$ et $p_{lj}^{(t)}$ est la probabilité que la coordonnée j de X prenne la valeur l au sein du noeud t .

La construction de l'arbre maximal commence par le partage du noeud racine de l'arbre contient toutes les observations de S . L'échantillon d'apprentissage est divisé de manière récursive en deux sous-échantillons disjoints, à l'aide de divisions binaires de la forme $x_{.j} \in \mathcal{A}_j$, où $j \in \{1, \dots, p\}$ et $\mathcal{A}_j$ est un sous-ensemble de $\text{sup}(j)$. Dans le cas de données continues, $\mathcal{A}_j$ peut être un intervalle réel de la forme $\mathcal{A}_j =]-\infty; a_j]$, avec $a_j \in \text{sup}(j)$. Dans le cas des données nominales, $\mathcal{A}_j$ peut être un ensemble dénombrable de m_j valeurs, où m_j est le nombre de modalités de la variable j .

On définit alors une division binaire, notée $s_j(t)$, d'un noeud t en deux sous-noeuds t_l et t_r , comme une paire $(j, \mathcal{A}_j)$ avec :

$$t_l = \{x \in E : x_{.j} \in \mathcal{A}_j\}, t_r = \{x \in E : x_{.j} \notin \mathcal{A}_j\}. \quad (3.7)$$

La division binaire optimale d'un noeud t en deux sous-noeuds t_l et t_r est définie par :

$$\text{argmax}_{(j, \mathcal{A}_j)} \{\Delta(t, j, \mathcal{A}_j)\}, \quad (3.8)$$

où

$$\Delta(t, j, \mathcal{A}_j) = R(t) - R(t_l) - R(t_r) \quad (3.9)$$

Les deux nouveaux sous-noeuds sont à leur tour divisés récursivement, et les critères d'arrêt de la construction de l'arbre maximal sont les mêmes que ceux utilisés dans la section 2.2. Une fois que l'algorithme est arrêté, une étiquette de classe est assignée à chaque feuille de l'arbre maximal. Une partition de l'échantillon d'apprentissage est donc obtenue, chaque feuille correspondant à une classe.

Les détails techniques concernant les étapes d'élagage de CUBT sont présentés dans les sections 2.2 et 2.4.2, étant donné que l'élagage de l'arbre maximal n'est pas une étape nécessaire au calcul des scores d'importance des variables.

3.3.2 Mesure de l'importance des variables dans CUBT

Afin de définir l'importance des variables dans CUBT, nous suivons quelques idées utilisées dans CART [29]. Pour cela nous définissons dans un premier temps les divisions de substitution pour chaque variable au sein de chaque noeud de l'arbre maximal.

3.3.2.1 Les divisions concurrentes

Soit s la division binaire optimale d'un noeud t d'un arbre T , définie sur la variable j_0 , qui divise t en deux sous-noeuds t_L et t_R . Soit $s_j = s_j(t)$ une division binaire de t définie sur n'importe quelle variable $j \neq j_0$, divisant t en deux sous-noeuds t'_L et t'_R .

La probabilité qu'une observation soit envoyée dans le sous-noeud de gauche pour les deux divisions binaires s et s_j est définie comme suit :

$$p(t_L \cap t'_L) = \frac{\text{Card}\{t_L \cap t'_L\}}{n_t} \quad (3.10)$$

Ensuite, la probabilité que les deux divisions binaires envoient une même observation dans le noeud de gauche est définie par :

$$p_{LL}(s, s_j) = \frac{p(t_L \cap t'_L)}{p(t)} \quad (3.11)$$

où $p(t)$ peut être estimée par $\frac{n_t}{n}$, et $p_{RR}(s, s_j)$ est définie de façon équivalente. La probabilité que s_j prédise bien s est définie par :

$$p(s, s_j) = p_{LL}(s, s_j) + p_{RR}(s, s_j) \quad (3.12)$$

La *division concurrente* pour s définie sur la variable j au sein du noeud t , notée $\tilde{s}_j$, est définie par :

$$p(s, \tilde{s}_j) = \max_{s_j \in S_j} p(s, s_j) \quad (3.13)$$

où S_j est l'ensemble de toutes les divisions binaires possibles définies sur la variable j .

Les divisions concurrentes sont utilisées pour mesurer l'importance des variables. Elles sont également utiles pour prédire la classe d'appartenance de nouvelles observations contenant des données manquantes.

3.3.2.2 L'importance des variables

Nous définissons l'importance de la variable j de la manière suivante :

$$\text{Imp}(X_{.j}) = \sum_{t \in T} \Delta(t, \tilde{s}_j(t)). \quad (3.14)$$

Le score d'importance correspond à la perte totale de *déviante* induite si chaque division binaire optimale dans l'arbre T est remplacée par la division concurrente définie sur $X_{.j}$.

L'importance des variables peut être mesurée en utilisant les données dont on dispose, cependant l'utilisation de technique ré-échantillonnage peut être utile pour obtenir des résultats plus stables au regard des scores obtenus [68, 69].

3.4 Expérimentations

Dans cette section, nous analysons la stabilité ainsi que l'efficacité des scores d'importance des variables basés sur CUBT. Tout d'abord, la stabilité du score obtenu est analysée en utilisant deux jeux de données où les classes d'appartenance des individus sont connues. Ces jeux de données sont Toys et Iris. Ensuite l'efficacité de la détection des variables importantes est analysée en utilisant plusieurs modèles de simulation de données, et CUBT est comparée à d'autres méthodes classiques de calcul de scores d'importance de variables.

3.4.1 Stabilité de l'importance des variables

Les arbres de décision sont réputés pour être instables lorsque les données sont sujettes à de légères perturbations [30]. En outre, CUBT, similairement à CART, est très sensible au choix du paramètre *minsize*, c'est-à-dire la taille minimale des feuilles de l'arbre. Nous nous intéressons ici à la stabilité de l'importance des variables, au regard de certaines modifications des données et du choix du paramètre *minsize*.

Dans cette analyse, nous utilisons deux jeux de données bien connus en apprentissage supervisé, le premier est le jeu de données Toys [171] et le deuxième est le jeu de données Iris [61].

Le jeu de données Toys contient deux groupes d'observations ($Y \in \{-1, 1\}$) et n observations peuvent être générées aléatoirement, pour un nombre de variables $p \geq 6$, de la manière suivante :

- Pour $j \in \{1, 2, 3\}$, $P(X_{.j} \sim N(y_j, 1)) = 0.7$ et $P(X_{.j} \sim N(0, 1)) = 0.3$,
- Pour $j \in \{4, 5, 6\}$, $P(X_{.j} \sim N(0, 1)) = 0.7$ et $P(X_{.j} \sim N(y(j - 3), 1)) = 0.3$,
- Si $p > 6$, $P(X_{.j} \sim N(0, 1)) = 1$.

Les six premières variables sont les vraies variables importantes, les autres ne sont pas pertinentes pour le modèle.

Le jeu de données Iris contient trois classes et 150 observations décrites par 4 variables continues.

Pour chaque jeu de données, nous effectuons 100 ré-échantillonnages en utilisant le modèle de simulation (décrit par Toys) ou du ré-échantillonnage de type

bootstrap (pour le jeu de données Iris). Pour chaque échantillon généré, un arbre est construit avec la CUBT, et l'importance de chaque variable est calculée. Ce calcul est répété pour différentes valeurs entières du paramètre *minsize*, variant de 2 à $\lfloor \frac{n}{4} \rfloor$, où $\lfloor \cdot \rfloor$ désigne la partie entière. Ainsi pour chaque valeur de *minsize* nous obtenons 100 estimateurs du score d'importance de chaque variable. Une valeur optimale du paramètre *minsize* peut être obtenue en utilisant l'heuristique proposée dans la section 2.4.3.

Pour le jeu de données Toys, nous fixons $n = 100$ et $p = 12$, c'est-à-dire qu'il y a autant de variables non pertinentes que de variables importantes.

La figure 3.2 (respectivement 3.4) montre des boîtes à moustache de l'importance des variables en fonction de la valeur du paramètre *minsize* pour le jeu de données Toys (respectivement Iris). La figure 3.3 (respectivement 3.5) fournit des boîtes à moustache représentant l'importance de chaque variable, pour une valeur optimale *minsize* = 8 (respectivement *minsize* = 16).

Le score d'importance des variables est très sensible à la valeur du paramètre *minsize*. Il décroît proportionnellement à la valeur de *minsize*. Comme le score d'importance est additif avec le nombre de noeuds de l'arbre, des valeurs plus élevées de *minsize* donneront des arbres contenant moins de noeuds, et donc des scores d'importance des variables moins élevés.

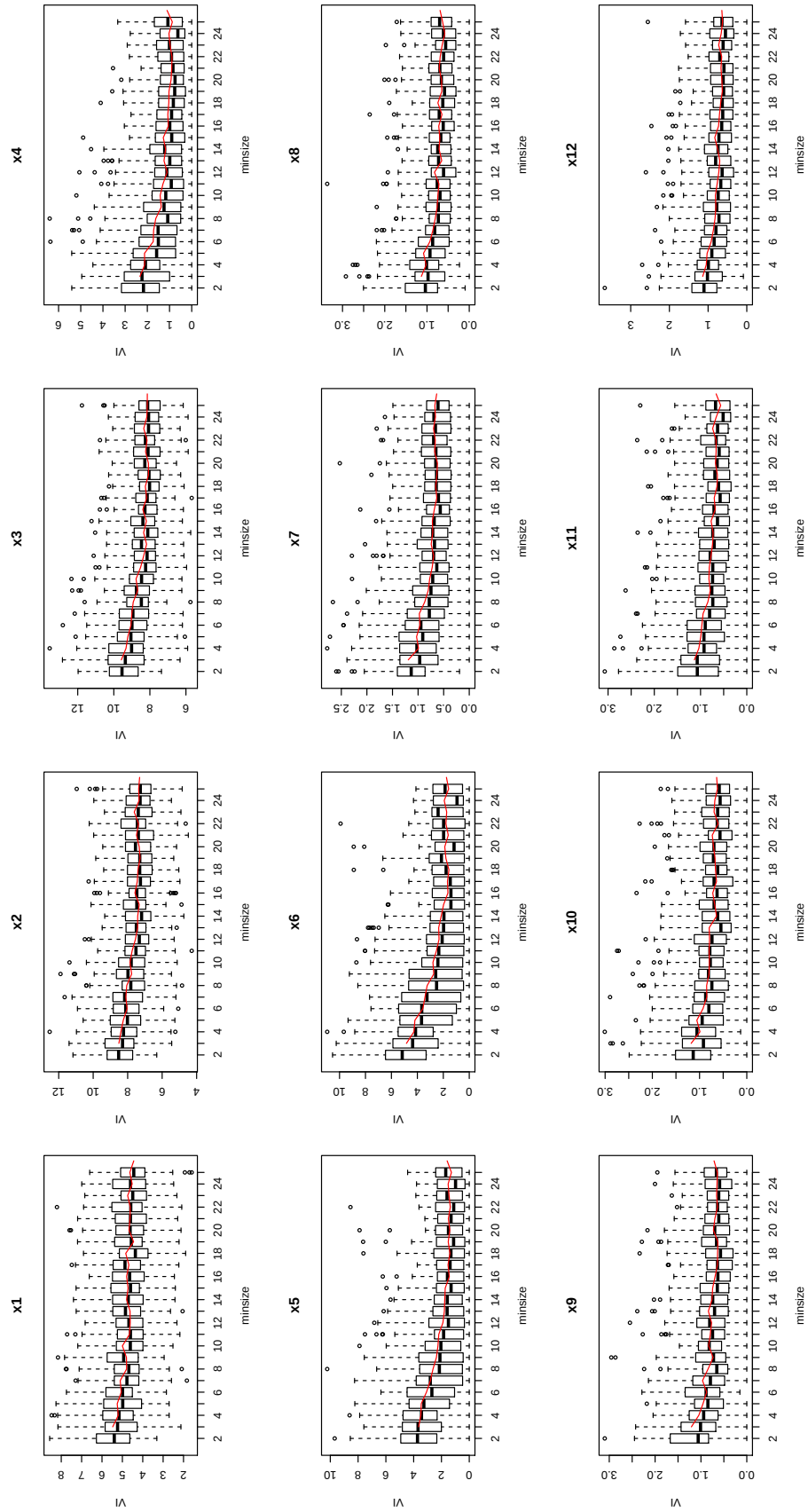


Figure 3.2 – Résultats de stabilité pour le jeu de données Toys. Chaque graphique montre des boîtes à moustache de l'importance d'une variable en fonction de la valeur du paramètre *minsize*.

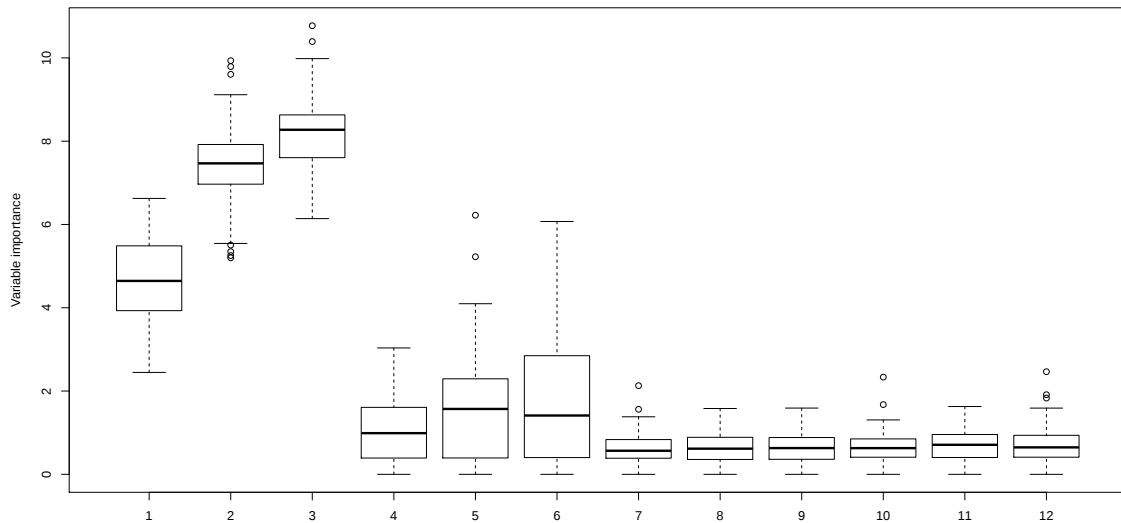


Figure 3.3 – Résultats de stabilité pour le jeu de données Toys. Le graphique montre des boîtes à moustache de l'importance de chaque variable pour une valeur optimale $minsize = 8$.

Pour le jeu de données Toys, les boîtes à moustache montrent que les trois premières variables sont toujours détectées comme les variables les plus importantes, les trois suivantes sont un peu moins importantes, ce qui est un résultat attendu étant donné le modèle de simulation de données sous-jacent. Les six variables restantes (les variables bruit) montrent des scores d'importance très stables, inférieurs aux scores des six premières variables. Les variables ayant les plus faibles scores d'importance ont aussi une plus faible dispersion.

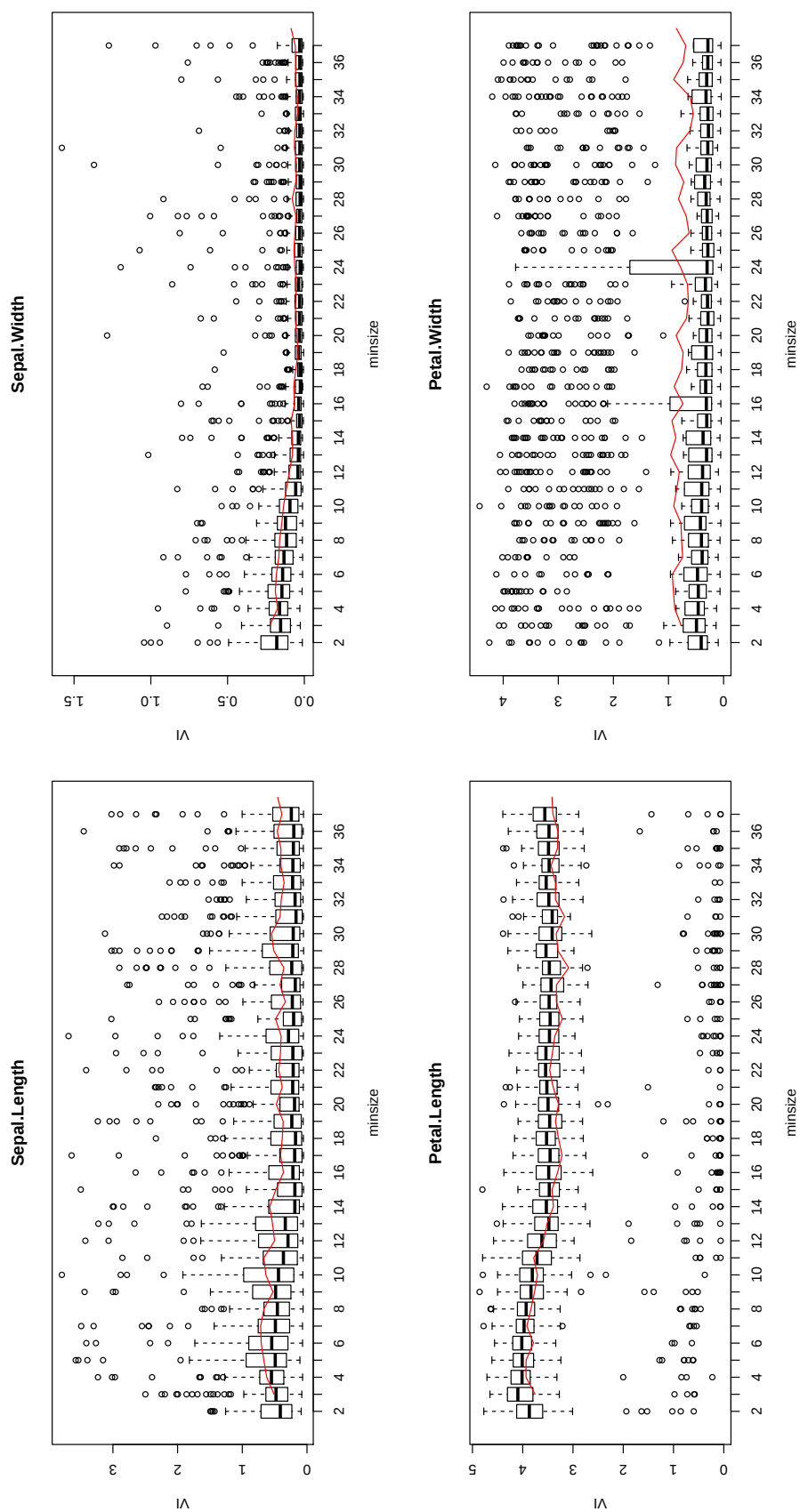


Figure 3.4 – Importance de chacune des 4 variables du jeu de données Iris en fonction des valeurs du *minsize*.

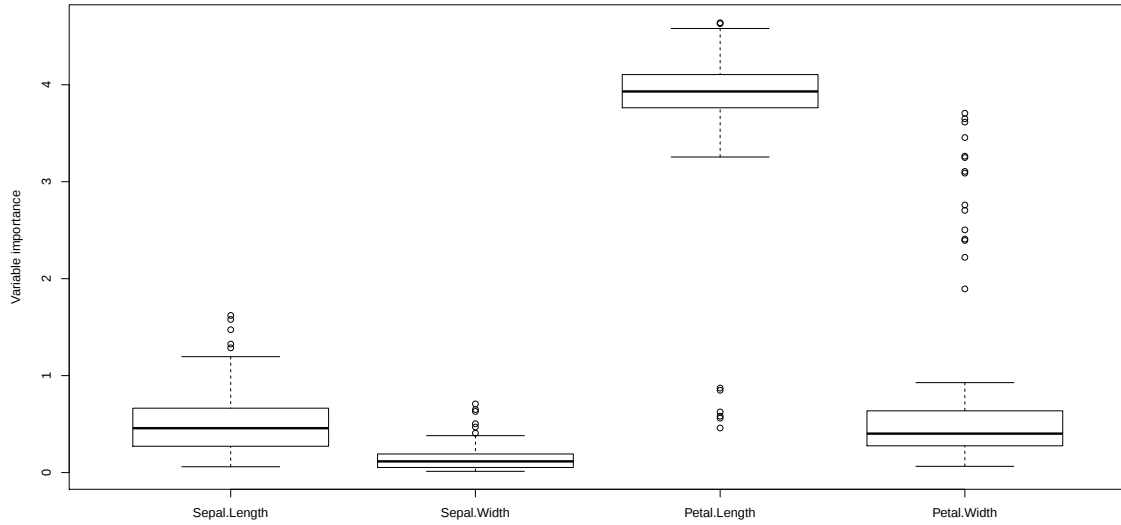


Figure 3.5 – Importance des variables du jeu de données Iris pour une valeur optimale $minsize = 16$.

Pour le jeu de données Iris, le score d'importance est plus stable pour la variable la moins importante (*Sepal.Width*), que les variables "intermédiaires" (*Sepal.Length*, *Petal.Width*) sont très dispersées. La variable *Petal.Length* est détectée comme la variable la plus importante dans ce jeu de données, ce qui confirme les résultats donnés dans la table 3.1.

3.4.2 Efficacité de l'importance des variables

Nous analysons maintenant l'efficacité du score d'importance des variables basé sur CUBT dans la détection de "vraies" variables importantes. Pour cela, nous utilisons quelques modèles de simulation de données dans lesquels les variables importantes qui définissent les classes sont connues. Nous utilisons les mêmes modèles de simulation qui ont été précédemment définis pour les données continues [64] ainsi que pour les données nominales (voir section 2.4.3.2). Nous analysons aussi l'efficacité de notre méthode en utilisant encore le jeu de données Toys.

3.4.2.1 Modèles de simulation de données

Nous considérons neuf modèles de simulation de données. Les quatre premiers modèles ont pour but de générer des données continues (modèles M1 à M4) et les cinq restants sont les mêmes que ceux présentés à la section 2.4.3.2 (modèles

M5 à M9). Nous présentons donc uniquement les modèles pour les données continues.

M1 : Modèle 2D Dans ce modèle, nous fixons le nombre de classes $k = 4$ et $X \in \mathbb{R}^2$ qui suit une distribution normale multivariée $N(\mu_l, \Sigma)$, $l \in \{1, \dots, k\}$, où $\mu_1 = (-1, 0)$, $\mu_2 = (1, 0)$, $\mu_3 = (0, -1)$ et $\mu_4 = (0, 1)$, et la matrice variance-covariance $\Sigma = \text{diag}(\sigma^2 \mathbf{1})$, avec $\sigma = 0.1$. Ce modèle de simulation de données est illustré à la figure 3.6.

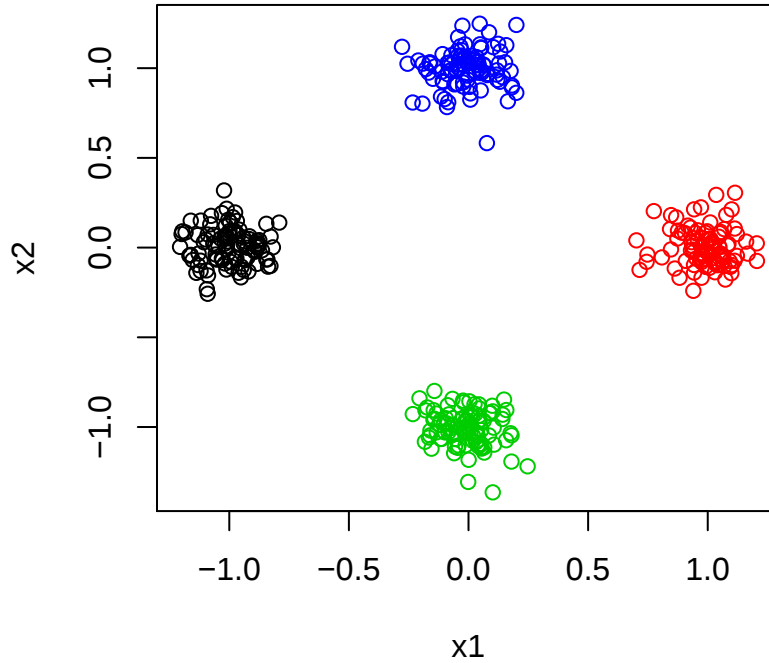


Figure 3.6 – Modèle de simulation de données M1, $n = 100$ et $\sigma = 0.1$.

M2 : Modèle 5D Ce modèle génère $k = 10$ groupes d'observations dans $\mathbb{R}^5$, suivant différentes distributions normales multivariées $N(\mu_l, \Sigma)$, $l \in [1, k]$, avec $\mu_1 = (1, 0, 0, 0, 0)$, $\mu_2 = (0, 1, 0, 0, 0)$, $\mu_3 = (0, 0, 1, 0, 0)$, $\mu_4 = (0, 0, 0, 1, 0)$, $\mu_5 = (0, 0, 0, 0, 1)$ et $\mu_i = -\mu_{i-5}$ pour $i \in \{6, \dots, 10\}$. La matrice variance-covariance est $\Sigma = \text{diag}(\sigma^2 \mathbf{1})$, avec $\sigma = 0.1$. Ce modèle de simulation de données est illustré à la figure 3.7.

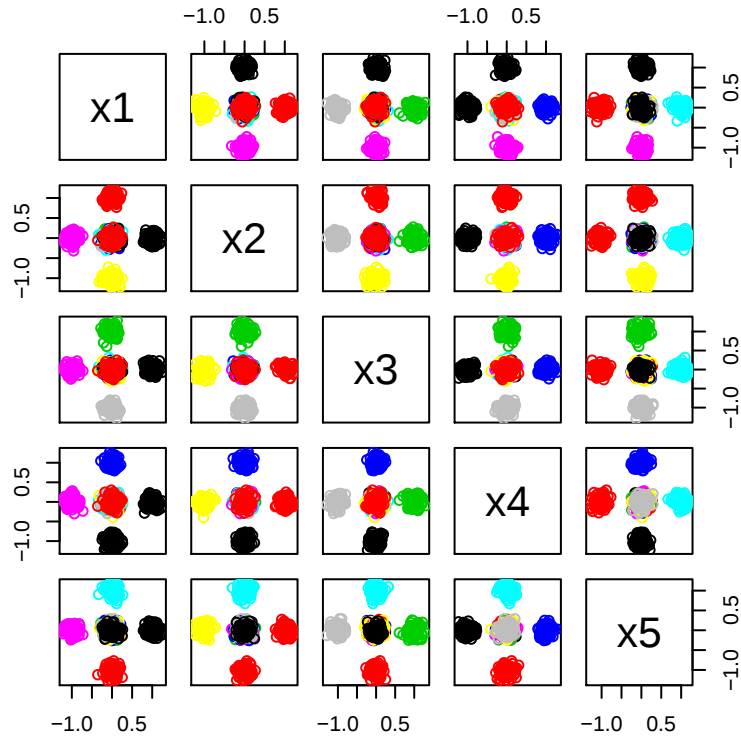


Figure 3.7 – Modèle de simulation de données M2, $n = 100$ et $\sigma = 0.1$.

M3 : Modèle à classes cocentriques Dans ce modèle, on considère deux classes cocentriques dans $\mathbb{R}^2$; chaque classe contient des observations distribuées uniformément entre deux cercles cocentriques centrés en $(0, 0)$. La première classe délimitée par des cercles ayant un rayon entre 50 et 80, la deuxième classe par des cercles de rayon entre 200 et 230. Ce modèle de simulation de données est illustré à la figure 3.8.

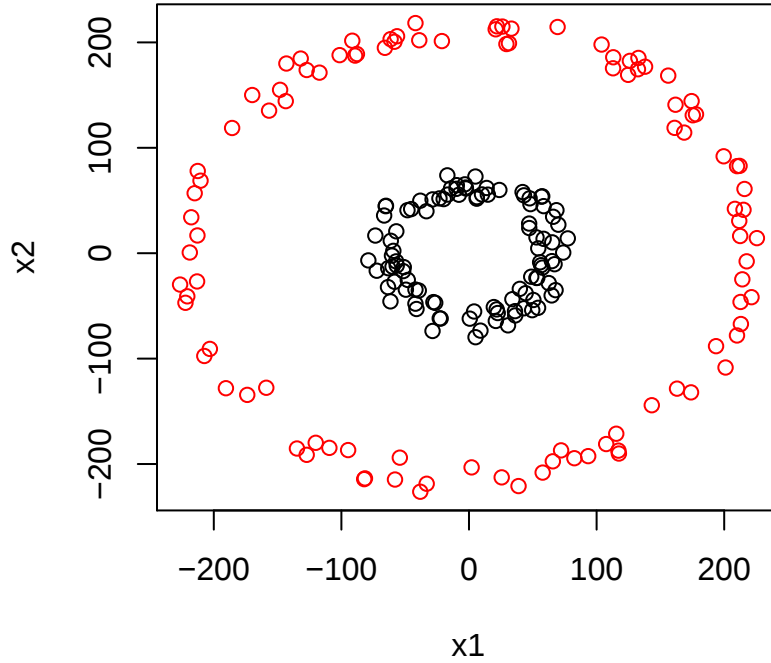


Figure 3.8 – Modèle de simulation de données M3, $n = 100$.

M4 : Modèle à grande dimension Dans ce modèle, on considère trois classes dans $\mathbb{R}^{50}$ normalement distribuées avec $\mu_1 = (-1, \dots, -1)$, $\mu_2 = (0, \dots, 0)$ et $\mu_3 = (1, \dots, 1)$, et la matrice variance-covariance est $\Sigma = \text{diag}(\sigma^2 \mathbf{1})$, avec $\sigma = 0.01$.

3.4.2.2 Ajout de bruit aux données simulées

Chacun des modèles de simulation de données proposés dans la section 3.4.2.1 est basé sur un nombre de variables p . Ces variables peuvent être considérées comme les vraies variables importantes au regard de l'identification des classes. Pour chaque cas, nous ajoutons aux données générées par le modèle un nombre p' de variables non pertinentes, ou variables bruit. Nous testons deux configurations : $p' = p$ et $p' = 2p$.

Pour les modèles de simulation de données continues M1 à M4, nous simulons $\lfloor \frac{p'}{2} \rfloor$ variables suivant une distribution normale $N(0, \frac{\sigma_0}{2})$, où $\sigma_0 = \min_{j \in \{1, \dots, p\}} \sigma_j$ est l'écart-type minimum observé sur l'ensemble initial des p variables. Les $\lfloor \frac{p'}{2} \rfloor$ variables restantes à ajouter ont une distribution uniforme $U(-\sigma_0 \sqrt{\frac{3}{4}}, \sigma_0 \sqrt{\frac{3}{4}})$.

Cette méthode nous permet de considérer deux types de distribution pour les variables bruit, qui ont la même variance.

Pour les modèles de simulation de données nominales M5 à M9, pour chacune des p vraies variables importantes, nous générons p' variables non pertinentes de la manière suivante : si la vraie variable importante a m modalités, sa variable bruit correspondante a la distribution $(p_1, \dots, p_m)$ sur son support. On pose $p_l = 0.8$ et $p_j = \frac{0.2}{m-1}$ pour $j \neq l$ et la modalité l est choisie arbitrairement. Cette approche nous permet de générer des variables bruit qui ont une faible entropie.

3.4.2.3 Résultats

Pour chaque modèle de simulation de données, nous faisons varier la taille de l'échantillon avec $n = 100, 300$ et 500 , et nous produisons 100 échantillons, en calculant le score d'importance des variables avec les méthodes suivantes : CUBT, URF, LOVO-cubt, LOVO-km, TWKM et LS. Pour les données continues et nominales, les classes contiennent le même nombre d'observations. Pour la méthode CUBT, le score d'importance des variables est calculé à partir de l'arbre optimal obtenu après élagage et regroupement des feuilles.

Pour mesurer l'efficacité de chaque méthode, nous calculons la proportion de vraies variables importantes apparaissant dans le top p de la hiérarchie de variables obtenues. Cet indice est similaire à un taux de vrais positifs (TVP). Des valeurs élevées de cet indice représentent une forte capacité de détection des variables importantes. Lorsqu'un taux de vrais positifs est inférieur à 100%, il peut être intéressant d'identifier la position des vraies variables importantes non détectées dans la hiérarchie et de voir quels scores les vraies variables non-détectées obtiennent. Nous considérons un autre indicateur, qui est le plus haut rang (PHR) observé parmi les p vraies variables importantes. Le premier indice (TVP) est reporté comme la moyenne des TVP obtenus sur les 100 échantillons, le deuxième (PHR) est reporté comme le rang calculé à partir de la moyenne du score d'importance de chaque variable sur les 100 réplicats.

La table 3.2 fournit les résultats pour les neuf modèles de simulation, pour deux niveaux de bruit dans les données. Elle présente les TVP ainsi que le PHR (valeurs entre parenthèses). Pour les modèles de simulation de données continues (M1 à M4), la méthode LOVO-km utilise la méthode k -means, et pour les autres modèles, elle utilise la méthode k -modes. La méthode TWKM étant adaptée uniquement aux données continues, on n'observe aucun résultat dans la table dans le cas des données nominales pour cette méthode. La table 3.3 présente les résultats pour le jeu de donnée Toys.

En termes de TVP, CUBT surpasse toutes les autres méthodes, quelle que soit la taille de l'échantillon (excepté pour le modèle M8), et obtient au moins 90% de vraies variables importantes correctement détectées. Pour les modèles de simulation de données continues (M1 à M4), les forêts aléatoires non-supervisées obtiennent des résultats peu satisfaisants, qui se dégradent en augmentant la

Modèle	n	CUBT	URF	LOVO-cubt	LOVO-km	LS	TWKM
M1 $p = 2$ $p' = 2$	100	100(2)	30.5(4)	100(2)	100(2)	0(4)	23(4)
	300	100(2)	18.5(4)	100(2)	99.5(2)	0(4)	56.5(4)
	500	100(2)	99(2)	100(2)	100(2)	0(4)	57(4)
M1 $p = 2$ $p' = 4$	100	100(2)	23(6)	100(2)	100(2)	0(4)	17.5(6)
	300	100(2)	16.5(6)	100(2)	100(2)	0(4)	54.5(6)
	500	99.5(2)	13.5(6)	95(2)	100(2)	0(4)	58.5(4)
M2 $p = 5$ $p' = 5$	100	100(5)	56.6(9)	93(5)	91.4(5)	20(9)	29.8(9)
	300	96(5)	58.4(8)	86.4(5)	96.2(5)	20(9)	75(9)
	500	100(5)	54(9)	85.2(5)	96.8(5)	20(9)	90.6(9)
M2 $p = 5$ $p' = 10$	100	100(5)	36.8(10)	86.4(5)	93.6(14)	0(14)	21.8(5)
	300	96.2(5)	39.2(6)	82.6(5)	94.2(5)	0(13)	24.2(6)
	500	100(5)	34.4(15)	79.4(5)	96.2(5)	0(10)	88.8(14)
M3 $p = 2$ $p' = 2$	100	97.5(2)	30(4)	42(4)	80.5(2)	50(4)	61(4)
	300	99.5(2)	8(4)	53(4)	78.5(2)	50(4)	55.5(4)
	500	100(2)	5.5(4)	81(2)	77.5(2)	50(4)	49(4)
M3 $p = 2$ $p' = 4$	100	97.5(2)	19(4)	40.5(3)	77(2)	50(6)	92(2)
	300	100(2)	18(5)	52.5(3)	71(2)	50(4)	59(4)
	500	100(2)	8(6)	89(2)	71.5(2)	50(4)	41(4)
M4 $p = 50$ $p' = 50$	100	100(50)	2.3(100)	99.9(97)	73.4(100)	0(100)	78.5(50)
	300	100(50)	0.3(100)	100(50)	72.4(100)	0(100)	79.3(50)
	500	100(50)	0(100)	100(50)	71.7(97)	0(100)	75.7(50)
M4 $p = 50$ $p' = 100$	100	100(50)	0(150)	99.8(136)	46(150)	0(147)	74.1(50)
	300	100(50)	0(146)	100(50)	43.8(149)	0(150)	76(50)
	500	100(50)	0(150)	100(50)	45.3(147)	0(150)	80(50)
M5 $p = 9$ $p' = 9$	100	100(9)	64.4(9)	57.8(14)	86.7(18)	37.8(18)	-
	300	100(9)	78.9(9)	55.6(16)	86.7(18)	28(18)	-
	500	100(9)	84.4(9)	57.8(18)	81.7(15)	18.1(18)	-
M5 $p = 9$ $p' = 18$	100	100(9)	58.9(9)	43.3(25)	51.1(23)	27.6(26)	-
	300	100(9)	71.7(9)	41.7(22)	63.9(25)	22.4(27)	-
	500	100(9)	80(9)	38.9(25)	65(27)	19.1(26)	-
M6 $p = 3$ $p' = 3$	100	96.7(3)	100(3)	56.7(6)	53.3(6)	78(13)	-
	300	100(3)	100(3)	28.3(6)	31.7(6)	12(6)	-
	500	100(3)	100(3)	21.7(6)	35(6)	0(6)	-
M6 $p = 3$ $p' = 6$	100	95(3)	100(3)	28.3(7)	38.3(7)	70.3(3)	-
	300	100(3)	100(3)	21.7(7)	26.7(9)	68(3)	-
	500	100(3)	100(3)	25(9)	20(8)	26(9)	-
M7 $p = 3$ $p' = 3$	100	96.7(3)	98.3(3)	43.3(6)	58.3(4)	45(6)	-
	300	100(3)	100(3)	55(6)	53.3(5)	1(6)	-
	500	100(3)	100(3)	41.7(6)	50(6)	0(6)	-
M7 $p = 3$ $p' = 6$	100	98.3(3)	100(3)	33.3(9)	45(3)	49(3)	-
	300	100(3)	100(3)	45(4)	46.7(9)	20(9)	-
	500	100(3)	100(3)	35(6)	45(5)	1(9)	-
M8 $p = 9$ $p' = 9$	100	73.9(9)	83.3(9)	51.7(18)	55(17)	62.7(15)	-
	300	76.1(9)	84.4(9)	53.3(16)	53.3(17)	60(14)	-
	500	76.1(9)	86.7(9)	50.6(15)	54.4(16)	66.9(9)	-
M8 $p = 9$ $p' = 18$	100	68.3(9)	77.2(9)	40(27)	36.7(27)	45.9(26)	-
	300	71.1(9)	85.6(9)	37.8(26)	31.7(23)	46.2(17)	-
	500	73.3(9)	85(9)	31.7(25)	34.4(27)	48.4(22)	-
M9 $p = 9$ $p' = 9$	100	100(9)	55.6(11)	66.7(14)	61.1(15)	21.6(18)	-
	300	100(9)	59.4(16)	52.2(18)	62.8(10)	3.8(18)	-
	500	100(9)	60(16)	55(16)	58.3(14)	0.3(18)	-
M9 $p = 9$ $p' = 18$	100	100(9)	43.9(13)	56.7(23)	40.6(23)	16(27)	-
	300	100(9)	47.8(26)	50.6(23)	47.2(24)	9.2(25)	-
	500	100(9)	56.1(22)	35(11)	53.3(27)	2.6(27)	-

Table 3.2 – Taux de vrais positifs (TVP) pour chaque modèle de simulation de données et chaque méthode. Les valeurs entre parenthèses représentent le plus haut rang (PHR) obtenu parmi les p vraies variables importantes. Les valeurs en gras correspondent aux méthodes les plus performantes. p est le nombre de vraies variables importantes et p' est le nombre de variables bruit ajoutées aux données.

	n	CUBT	URF	LOVO-cubt	LOVO-km	LS	TWKM
Toys	100	85.7(6)	34.3(12)	75.7(12)	73(9)	16.7(12)	83.7(12)
$p = 6$	300	95.5(6)	25.7(12)	65(11)	72.7(7)	6.7(12)	83.3(12)
$p' = 6$	500	95.7(6)	23.5(12)	58.7(9)	75.2(10)	2.5(12)	83.8(12)
Toys	100	80.5(6)	18.2(18)	68.8(17)	61.2(18)	11.2(18)	83.5(16)
$p = 6$	300	91.7(6)	11.2(18)	50.8(17)	62.3(11)	4.2(18)	83.5(17)
$p' = 12$	500	95.2(6)	8.2(17)	51.2(17)	62(17)	4.3(18)	83.3(18)

Table 3.3 – Taux de vrais positifs (TVP) pour les données Toys et chaque méthode. Les valeurs entre parenthèses représentent le plus haut rang (PHR) parmi les rangs des p vraies variables importantes. Les valeurs en gras correspondent aux méthodes les plus performantes. p est le nombre de vraies variables importantes et p' est le nombre de variables bruit ajoutées aux données.

taille de l'échantillon. Ce n'est pas le cas pour la méthode TWKM et le score Laplacien, dont les performances sont stables et s'améliorent lorsqu'on augmente la taille de l'échantillon. CUBT est souvent ex-aequo avec les deux approches LOVO pour le modèle M1 (ainsi que pour le modèle M4 pour LOVO-cubt).

Pour les modèles de simulation de données nominales (M5 à M9), les performances de CUBT sont très proches de celles des forêts aléatoires non-supervisées. Ces deux méthodes ont un TVP supérieur à 90% (excepté pour le modèle M8), et elles surpassent les approches LOVO. Les résultats relatifs au PHR sont complémentaires avec ceux déjà obtenus, une plus grande proportion de vraie variables importantes détectées implique des valeurs de PHR plus proches de p . Cependant, lorsque CUBT n'est pas la méthode la plus performante au regard du TVP (cela arrive dans 13 cas sur 54), elle reste aussi performante que les meilleures méthodes (LOVO ou forêts aléatoires non-supervisées) au regard du PHR. Bien que l'écart en termes de TVP des autres méthodes est souvent négligeable, le PHR appuie le fait que CUBT montre une bonne capacité de détection des variables importantes dans les jeux de données simulées.

Pour le jeu de données Toys (voir table 3.3), CUBT obtient les plus hautes valeurs de TVP (excepté dans le cas $p' = 2p$ et $n = 100$, où la méthode TWKM est plus performante au regard de cet indice) et les plus faibles valeurs de PHR. CUBT détecte parfaitement les p vraies variables importantes dans cet exemple.

3.4.3 Conclusion

Dans ce chapitre, nous avons présenté une nouvelle méthode pour mesurer l'importance des variables, basée sur la méthode CUBT. Nous avons comparé cette nouvelle approche à d'autres méthodes classiques de calcul de scores d'importance, en considérant plusieurs modèles de simulation de données, en pré-

sence de bruit. Le score d'importance des variables obtenu avec CUBT a montré d'excellentes performances comparé aux autres indices, dans la majorité des expérimentations. D'autres simulations sont à prévoir pour analyser l'efficacité des scores obtenus en fonction de la corrélation ou de la redondance entre les variables. Nous avons également analysé la stabilité de ce score en fonction de différentes données et des paramètres de réglage de CUBT (particulièrement le paramètre *minsize*), en utilisant des heuristiques définies dans les travaux précédents.

L'importance des variables en classification non supervisée basée sur les arbres de décision binaires peut être utilisée pour effectuer de la sélection de variables, de la même manière qu'en apprentissage supervisé. Les divisions concurrentes utilisées pour calculer le score d'importance peuvent également être utiles dans la méthode CUBT pour gérer les données manquantes potentielles, à la fois pour l'apprentissage mais également pour effectuer des prédictions. Ces idées sont actuellement prises en considération.

4- Développement de questionnaires adaptatifs

Dans le champ de la qualité de vie (QV), les méthodes basées sur la théorie de réponse à l'item sont couramment utilisées pour le développement de banques d'items unidimensionnelles et de questionnaires adaptatifs informatisés. Ces méthodes peuvent être adaptées pour surmonter les problèmes rencontrés lors du développement de formes courtes de questionnaires, où le nombre d'items est fixe. En effet, le développement de banques d'items permet d'obtenir une collection d'items qui mesurent un même trait latent. Combiné à une banque d'items, un questionnaire adaptatif peut donc permettre de n'administrer que les items qui offrent le plus de pertinence pour un individu donné, réduisant ainsi la longueur du questionnaire (c'est-à-dire le nombre d'items administrés) et temps de remplissage, tout en maintenant un niveau de précision du test satisfaisant. Une banque d'items contient une liste d'items, qui ont été préalablement *calibrés*. La calibration d'une banque d'items consiste à évaluer les propriétés psychométriques des items contenus dans la banque. Une fois calibrée, un algorithme d'administration adaptative des items de la banque peut être développé. Une autre démarche réside dans la prise en compte du caractère multidimensionnel du trait latent à mesurer (typiquement la qualité de vie). En utilisant des modèles multidimensionnels de la théorie de réponse à l'item, il est possible d'estimer plusieurs scores de dimensions en administrant qu'un sous-ensemble d'items. Ainsi, la réponse à un item permet l'évaluation de traits latents multiples, c'est la spécificité des questionnaires adaptatifs multidimensionnels. Enfin, une autre approche, plus récente, et non-paramétrique, consiste à considérer un arbre de décisions permettant de modéliser un questionnaire adaptatif. Cette méthode serait une avancée dans le développement des outils de mesure adaptatifs, étant donné que les arbres de décision sont plus simples à interpréter, et les algorithmes résultant plus aisés à implémenter.

Dans ce chapitre, nous présentons trois applications sur des données de qualité de vie liée à la santé. La section 4.1 présente le développement d'une banque d'items unidimensionnelle mesurant la qualité de vie liée à la santé mentale. La section 4.2 présente le développement de deux questionnaires adaptatifs multidimensionnels mesurant la qualité de vie, spécifiques à des patients atteints de sclérose en plaques, et de schizophrénie. Enfin, nous proposons dans la section 4.3 une alternative à la théorie de réponse à l'item à travers le développement d'un questionnaire adaptatif basé sur les arbres de décision binaires.

4.1 Développement d'une banque d'items mesurant la QV liée à la santé mentale pour des patients atteints de sclérose en plaques

Dans cette section, nous présentons le développement d'une banque d'items unidimensionnelle mesurant la santé mentale liée à la qualité de vie pour des patients atteints de sclérose en plaques. Cette application a donné lieu à une publication, et vise à regrouper des items issus de deux questionnaires différents couramment utilisés dans ce champs. La section 4.1.1 présente le contexte et les objectifs de cette étude. Les sections 4.1.2 et 4.1.3 détaillent la méthodologie et les principaux de cette étude.

4.1.1 Introduction

Chez les patients atteints de sclérose en plaques, l'incapacité physique due à la maladie, bien que très importante, n'est pas forcément l'aspect qui reflète le plus les différentes facettes de la qualité de vie [16]. L'incapacité physique n'est qu'une sous-dimension de la qualité de vie d'un patient, et la présence d'une sclérose implique une réduction de la qualité de vie de ce dernier [130]. Le plus souvent, la qualité de vie est mesurée par le biais de questionnaires auto-rapportés [157], qui peuvent être séparés en deux types : les questionnaires génériques (par exemple le SF-36 [168]), qui sont généralement utilisés pour comparer la qualité de vie entre différentes populations, et les questionnaires spécifiques qui se concentrent sur les problèmes de santé spécifiques à la pathologie, et qui sont plus sensibles pour détecter et quantifier de faibles variations [133] (par exemple le MusiQoL [155] ou le MSQOL-51 [164]). Etant donné que les questionnaires génériques ne mesurent pas exactement les mêmes aspects de la qualité de vie que les questionnaires spécifiques, il peut être intéressant d'utiliser une combinaison des deux, de manière à optimiser les avantages de chaque instrument, et ainsi obtenir une mesure complète de la qualité de vie [101]. Cependant dans la recherche et la pratique cliniques, il n'est pas pratique d'utiliser plusieurs mesures, puisque le temps est limité, et que les questionnaires, souvent longs à compléter, peuvent alourdir la tâche à la fois des patients et des cliniciens. De plus, certains auteurs ont mis en évidence la nécessité d'utiliser des questionnaires aussi courts que possible, étant donné les difficultés spécifiques rencontrées chez les patients atteints de sclérose en plaques (troubles cognitifs, fatigue, dépression) [16].

Les questionnaires adaptatifs offrent de nombreux avantages, permettant de surpasser ces problèmes [55, 59]. Basés sur une banque d'items développée à partir de méthodes de la théorie de réponse de l'item, les questionnaires adaptatifs permettent de sélectionner des items dans la banque qui seront les plus appropriés pour un patient donné. Le questionnaire adaptatif est un algorithme

qui administre un certain item en fonction des réponses précédentes de l'individu, et il permet de ne prendre en compte que les items les plus pertinents, réduisant le temps de remplissage tout en maintenant un niveau satisfaisant de précision du questionnaire [170, 143, 89]. Plusieurs questionnaires adaptatifs ont déjà été développés dans le champs de la sclérose en plaques, ils visent à mesurer différents aspects de la qualité de vie (souvent liée aux symptômes tels que la dépression, l'anxiété, la douleur, la fatigue et le fonctionnement physique). Ce sont souvent des initiatives issues du projet américain d'envergure *Patient Reported Outcomes Measurement Information System* (PROMIS) [46, 18, 154]. On peut également citer le projet *neurology quality-of-life measurement initiative* (Neuro-QoL), issu de PROMIS, qui vise à développer de grandes banques d'items mesurant la qualité de vie pour différentes maladies neurologiques [36, 37], notamment la sclérose en plaques. Certaines limites ont cependant été notées vis-à-vis de ces mesures existantes, qui ne sont pas disponibles en plusieurs langages (à part l'anglais et l'espagnol pour le projet Neuro-QoL). De plus le développement des banques d'items (et des questionnaires adaptatifs qui en sont issus) ne prennent pas en compte des combinaisons de questionnaires spécifiques et génériques.

Le but de cette étude préliminaire est de développer et calibrer une banque d'items, disponible en plusieurs langues, permettant de mesurer la qualité de vie liée à la santé mentale, en considérant la combinaison entre un questionnaire générique (le SF-36, [168]) et un questionnaire spécifique (le MusiQoL, [155]), dans le but de disposer d'une base pour le développement futur d'un questionnaire adaptatif.

4.1.2 Méthodes

4.1.2.1 Conception de l'étude

Cette étude est une seconde analyse d'une étude internationale, multicentrique et transversale [155]. Les patients ont été recrutés entre le mois de Janvier 2004 et le mois de Février 2005 dans des départements de neurologie de 15 pays : Argentine, Canada, France, Allemagne, Grèce, Israël, Italie, Liban, Norvège, Russie, Afrique du Sud, Espagne, Turquie, Royaume Uni et les Etats Unis.

4.1.2.2 Population

Les critères d'inclusion sont les suivants : diagnostic de sclérose en plaques selon les critères de McDonald [117], patients hospitalisés ou ambulatoires, âge supérieur à 18 ans, consentement écrit du patient. Les critères de non-inclusion sont les suivants : diagnostic neurologique autre que la sclérose en plaques, démence, grave rechute en cours et incapacité de remplir le questionnaire sans assistance.

4.1.2.3 Récolte des données

Les données suivantes sont considérées dans cette étude :

Données sociodémographiques Le genre, l'âge, le niveau scolaire, le statut marital, l'activité professionnelle et la zone géographique. Les 15 pays sont agrégés en 6 zones géographiques (1 : Afrique du Sud ; 2 : Allemagne, France, Grèce, Italie, Norvège, Espagne, Royaume Uni ; 3 : Argentine ; 4 : Etats-Unis, Canada ; 5 : Israël, Liban, Turquie ; 6 : Russie).

Données cliniques Le sous-type de sclérose [112], la durée de la maladie, l'invalidité due à la sclérose mesurée avec l'échelle d'invalidité étendue EDSS [99], la sévérité des symptômes mesurée avec une échelle contenant 14 items (perte de sensation tactile, perte de sensation de position, mouvements involontaires du corps, vibrations dans les bras et les mains, faiblesse des membres, impossibilité d'avaler, mouvements d'yeux involontaires, problèmes visuels, difficulté de concentration, fatigue, incontinence urinaire, incontinence intestinale).

Données de qualité de vie La qualité de vie est mesurée avec les questionnaires MusiQoL [155] et SF-36 [103]. Ces deux questionnaires sont décrits dans la section 1.1.2.

4.1.2.4 Développement de la banque d'items

Nous suivons une procédure standardisée adaptée à l'analyse psychométrique des items d'une banque d'items [143, 148, 38, 52]. Cette procédure est divisée en trois étapes : démarche conceptuelle et sélection d'items, calibration et validation clinique.

Démarche conceptuelle : sélection d'items mesurant la qualité de vie liée à la santé mentale Un ensemble initial de 67 items est construit en agrégeant les items du MusiQoL (31 items) et les items du SF-36 (36 items). Les items mesurant des concepts relatifs à la santé mentale sont sélectionnés comme items candidats par deux experts du domaine. Au final, 8 items du MusiQoL sont conservés, représentant trois dimensions liées à la santé mentale : bien-être psychologique (4 items), rejet (2 items) et coping (2 items). De plus, 14 items du SF-36 sont conservés, représentant les dimensions ayant une pondération supérieure à 0,2 dans le calcul du score composite mental [102, 179] : fonctionnement social (2 items), santé mentale (5 items), vitalité (4 items) et limitations du à l'état émotionnel (3 items). Finalement, un ensemble de 22 items, mesurant les concepts de la qualité de vie liée à la santé mentale globale, est utilisé dans l'étape suivante de calibration.

Etape de calibration

1. Analyse descriptive et asymétrie : Les proportions de données manquantes ainsi qu'une mesure de l'asymétrie des distribution des réponses sont calculés pour chaque item sélectionné. Nous calculons également les corrélations inter-items avec les coefficients de corrélation non-paramétriques de Spearman. A cette étape, un item peut être supprimé si leur taux de données manquantes est trop important ($> 70\%$), s'il présente une asymétrie extrême (plus de 95 % de taux de réponse pour une des modalités ou valeur absolue du coefficient d'asymétrie supérieure à 4), ou si son coefficient de corrélation avec un autre item est trop élevé (> 0.7 , l'item de la paire le moins pertinent cliniquement est supprimé).
2. Unidimensionnalité et indépendance locale : L'unidimensionnalité des 22 items est vérifiée en utilisant plusieurs indicateurs. Tout d'abord, nous analysons la courbe de l'alpha de Cronbach [94] (une stricte monotonie doit être observée). Nous utilisons également le rapport des valeurs propres, issu d'une analyse en composantes principales [6] (un ratio supérieur à 3 est attendu), et un modèle d'analyse factorielle confirmatoire avec le logiciel MPlus [129]. Comme indicateurs de bon ajustement de ce dernier modèle, nous nous basons sur les valeurs de la racine de l'erreur quadratique moyenne d'approximation (noté RMSEA, valeur inférieure à 0.07 attendue) et de l'indice d'ajustement comparatif (noté CFI, valeur supérieure à 0.95 attendue), en fonction de seuils définis dans la littérature [90, 159]. L'indépendance locale est vérifiée en analysant les corrélations résiduelles inter-items. Si une paire d'items présente une corrélation résiduelle supérieur à 0.3, alors l'item ayant les plus fortes corrélations résiduelles avec les autres items de la banque est supprimé [22, 21].
3. Estimation des paramètres d'items et qualité d'ajustement : Les paramètres des items sont estimés en considérant le modèle de crédit partiel [116] avec la méthode de maximisation de la vraisemblance conditionnelle [9], implémentée dans le package R *eRm* [114]. Le modèle de crédit partiel permet de considérer ayant des nombres différents de modalités et différents seuils de difficulté. Soit θ un niveau de trait latent (ici, de qualité de vie liée à la santé mentale), on note $P_{ix}(\theta)$ la probabilité qu'un patient ayant un niveau de trait latent θ réponde la modalité x à l'item i , définie par :

$$P_{ix}(\theta) = \frac{e^{\sum_{k=0}^x (\theta - \beta_{ik})}}{\sum_{r=0}^{m_i} e^{\sum_{k=0}^r (\theta - \beta_{ik})}}, \quad (4.1)$$

où β_{ik} sont les paramètres des modalités d'items (ou seuils de difficulté) et m_i est le nombre de modalités de réponse de l'item i .

Dans un premier temps, nous étudions les courbes caractéristiques des items pour vérifier que chaque modalité de réponse d'un item a une probabilité maximale d'être sélectionné sur un certain intervalle des valeurs de

trait latent. Lorsque deux modalités de réponse ne sont pas visuellement discriminantes ou paraissent désordonnées, nous décidons d'agréger ces deux modalités et de ré-estimer le modèle avec l'item ainsi recodé. Cette approche nous permet d'assurer une autre hypothèse fondamentale de la théorie de réponse à l'item, à savoir la monotonie. A chaque fois qu'un item est recodé, nous calculons les critères d'information d'Akaike [7] (noté AIC) et de Bayes [153] (noté BIC) du modèle nouvellement estimé. Le recodage de l'item correspondant est donc validé si une amélioration de l'ajustement est vérifié au regard de ces deux indices.

Nous étudions ensuite la qualité d'ajustement de chaque item en calculant la statistique INFIT [104], permettant de détecter de potentiels items aberrants. L'INFIT permet de détecter des profils de réponses incohérents. Un intervalle de valeurs acceptables est défini par $[0.7, 1.3]$ [172].

Enfin, les valeurs de traits latents sont estimées par maximisation de la vraisemblance, fournissant un score pour chaque patient. Ce score reflète le niveau de qualité de vie liée à la santé mentale d'un patient, déterminé à partir de son profil de réponses et les paramètres d'items du modèle. La distribution du score est observée via un histogramme. L'information totale de la banque d'items est également calculé à partir des paramètres du modèle et du trait latent estimé [151], afin de mesurer la précision de la mesure en fonction des valeurs du trait latent.

4. Comportement différentiel des items :

Nous effectuons des tests de comportement différentiel d'items afin d'évaluer de potentiels biais d'items dus à l'appartenance des patients à certains groupes définis par les variables suivantes : classe d'âge (≤ 40 ans versus > 40 ans), genre (homme versus femme), situation familiale (célibataire versus en couple), niveau scolaire ($\leq$ BAC versus $>$ BAC), zone géographique (6 zones géographiques définies plus haut) et sous-type de sclérose (récurrente-rémittente, progressive primaire, progressive secondaire et syndrome clinique isolé). Dans cette étude le fonctionnement différentiel des items est mesuré en utilisant une approche itérative basé sur un modèle réponse à l'item logistique ordinal, implémentée dans le package R *lordif* [43]. Cette méthode utilise le modèle de réponse graduée [151] pour estimer le niveau de trait latent de chaque individu. Soit X_i la variable aléatoire représentant la réponse ordinaire d'un item i et soit $x_i \in \{0, 1, \dots, m - 1\}$. Pour chaque item, on considère les deux modèles emboîtés suivants :

$$\begin{aligned} \text{Modèle A : } \text{logit}(P(X_i \geq k)) &= \alpha_k + \beta_1 \theta_{IRT}, \\ \text{Modèle B : } \text{logit}(P(X_i \geq k)) &= \alpha_k + \beta_1 \theta_{IRT} + \beta_2 G + \beta_3 (\theta_{IRT} \times G), \end{aligned} \quad (4.2)$$

où $P(X_i \geq k)$ est la probabilité que la réponse à l'item, notée x_i , soit une

valeur supérieure ou égale à k , α_k est la constante des modèles, θ_{IRT} est le trait latent mesuré par la banque d'items, G est une variable groupe, β_1 (respectivement β_2) est le coefficient de θ_{IRT} (respectivement G), et $\theta_{IRT} \times G$ représente l'interaction entre le trait latent et la variable groupe, avec le coefficient β_3 .

Ces deux modèles sont comparés en calculant la statistique Khi-deux du rapport de vraisemblance et la mesure pseudo- R^2 entre les deux modèles emboîtés. Le test du rapport de vraisemblance permet de détecter un comportement différentiel global, avec un risque d'erreur $\alpha = 0.01$. En cas de significativité, on utilise le critère de Zumbo pour évaluer la magnitude du comportement différentiel d'item. Cette magnitude est jugée négligeable si $\Delta R^2 < 0.13$, modérée si $0.13 \leq \Delta R^2 \leq 0.26$ et élevé si $\Delta R^2 > 0.26$ [182].

5. Validité externe de la banque d'items :

Le score de qualité de vie liée à la santé mentale estimé, issu du modèle qui a permis de calibrer la banque d'items, est exprimé sur une échelle de 0 à 100 (0 étant le score le plus faible et 100 le score le plus élevé). Afin d'explorer la validité convergente, on étudie les relations entre le score de la banque d'items et les scores issus des dimensions du MusiQoL et du SF-36. L'hypothèse sous-jacente est que le score de la banque d'items devrait être plus corrélé avec les scores de dimensions correspondant aux items qui ont été utilisés pour constituer la banque qu'avec les scores des autres dimensions. La validité discriminante de la banque d'items est étudiée en comparant les moyennes du score de la banque en fonction de différents critères sociodémographiques (âge, genre, niveau scolaire, statut familial et activité professionnelle) et cliniques (score EDSS, échelle de symptômes, durée de la maladie et sous-type de sclérose), en utilisant des test t de Student, des ANOVA et des corrélations de Pearson.

L'intégralité des analyses statistiques sont entreprises avec les logiciels suivants : IBM PASW SPSS 17.0 [158], R version 2.15.2 [160], et MPlus [129].

4.1.3 Résultats

Un total de 1992 patients atteints de sclérose en plaques ont été inclus dans cette étude internationale. Les patients sont originaires de 15 pays différents : Argentine (27 patients), Canada (77 patients), France (179 patients), Allemagne (209 patients), Grèce (92 patients), Israël (66 patients), Italie (379 patients), Liban (20 patients), Norvège (104 patients), Russie (201 patients), Afrique du Sud (53 patients), Espagne (224 patients), Turquie (228 patients), Royaume-Uni (36 patients) et Etats-Unis (97 patients). La moyenne d'âge est 42.2 ans (écart-type = 11.9), 578 patients (29.5%) sont des hommes, 601 (36.8%) sont sans-emploi, 592 (35.2%) ont un haut niveau scolaire et 372 (21.7%) sont célibataires. Près de trois quarts (70.4%) ont une forme de sclérose récurrente-rémittente. La du-

rée moyenne de la maladie est de 11.1 ans (écart-type = 8.8) et le score EDSS médian vaut 3.0 (écart interquartile = 3.5).

4.1.3.1 Analyse descriptive et asymétrie

Pour les 22 items contenus dans la banque d'items, les taux de données manquantes varient de 1.6% à 7.6%. Les effets plancher et plafond varient de 4.0% à 50.2% et de 2.1% à 15.0% respectivement. Chaque item a un coefficient d'asymétrie acceptable (variant de -1.16 à 0.40) et un taux de réponse inférieur à 95% pour chaque modalité. Les coefficients de corrélation inter-items varient de 0.21 à 0.68 (avec des p-valeurs inférieures à 0.001), indiquant une faible redondance. A l'issue de cette analyse, nous décidons de conserver les 22 items dans la banque. Toutes les caractéristiques décrites ci-dessus sont présentées à la Table 4.1.

4.1.3.2 Unidimensionalité et indépendance locale

Nous observons une stricte monotonie de la courbe de l'alpha de Cronbach, comme le montre la figure 4.1. De plus, le rapport de valeurs propres est élevé (5.50), les indicateurs d'ajustement du modèle issus de l'analyse factorielle confirmatoire sont satisfaisants (RMSEA = 0.07, CFI = 0.95). Aucun des coefficients de corrélation résiduelle ne dépasse la valeur 0.3. Ces indices confirment l'unidimensionnalité et l'indépendance locale des 22 items contenus dans la banque. A l'issue de cette analyse, nous décidons de conserver tous les items.

Item	Questionnaire	Question	Effet plancher (%)	Effet plafond (%)	Données manquantes (%)	Asymétrie
1	MusiQoL	Felt nervous or irritated by a few things or situations ?	7.5%	9.4%	1.6%	0.11
2	MusiQoL	Felt anxious ?	6.4%	18.1%	2.6%	-0.16
3	MusiQoL	Felt depressed or gloomy ?	5.7%	17.9%	2.1%	-0.17
4	MusiQoL	Felt like crying ?	7.7%	20.0%	2.1%	-0.19
5	SF-36	Have you felt downhearted and blue ?	2.3%	18.8%	2.2%	-0.55
6	SF-36	Have you felt so down in the dumps that nothing could cheer you up ?	2.1%	36.8%	1.7%	-0.97
7	SF-36	Have you felt calm and peaceful ?	5.6%	8.2%	2.0%	-0.01
8	SF-36	Did you feel full of life ?	8.0%	9.0%	1.8%	0.09
9	SF-36	Did you feel worn out ?	4.7%	16.0%	2.1%	-0.24
10	SF-36	Did you have a lot of energy ?	15.0%	5.4%	2.4%	0.40
11	SF-36	Did you feel tired ?	10.5%	4.0%	1.6%	-0.07
12	SF-36	Have you been a happy person ?	5.8%	11.2%	2.2%	-0.09
13	SF-36	To what extent have your emotional problems interfered with your normal social activities ?	3.2%	39.2%	1.7%	-0.67
14	SF-36	How much of the time has emotional problems interfered with your social activities ?	5.2%	27.6%	3.3%	-0.37
15	SF-36	Have you been a very nervous person ?	4.7%	17.6%	1.9%	-0.50
16	MusiQoL	Been embarrassed when in public ?	4.1%	41.9%	3.9%	-0.80
17	MusiQoL	Felt bitter ?	7.8%	33.2%	3.5%	-0.54
18	SF-36	Accomplished less than you would like ?	46.4%	49.0%	4.6%	-0.05
19	SF-36	Cut down the amount of time you spent on work or other activities ?	38.1%	57.2%	4.8%	-0.41
20	SF-36	Didn't do work or other activities as carefully as usual ?	40.7%	54.6%	4.7%	-0.29
21	MusiQoL	Been upset by the stares of other people ?	4.1%	50.2%	7.6%	-1.16
22	MusiQoL	Felt that your situation is unfair ?	13.2%	28.6%	3.2%	-0.31

Table 4.1 – Analyse descriptive des items de la banque.

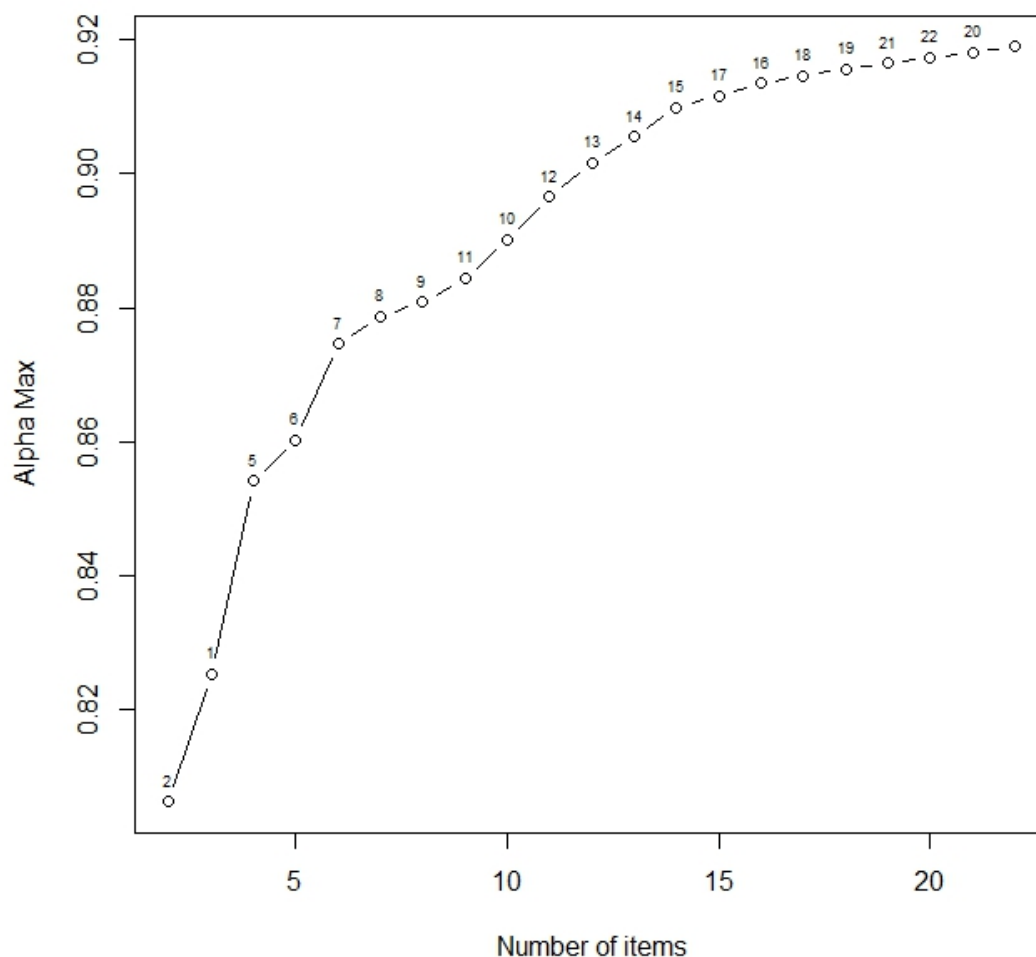


Figure 4.1 – Courbe de l'alpha de Cronbach.

4.1.3.3 Estimation des paramètres d'items et qualité d'ajustement

Sur les 22 items contenus dans la banque, 8 d'entre eux sont recodés après analyse des courbes caractéristiques des items. La différence des valeurs du critère AIC (respectivement BIC) entre le modèle final et le modèle initial est de -8468.78 (respectivement -8519.20), indiquant une amélioration globale de la qualité d'ajustement du modèle. Les courbes caractéristiques des items pour le modèle final sont fournis dans la figure 4.2. A l'issue de cette étape de recodage on observe que les seuils de difficulté sont bien distribués dans l'ordre croissant. Les détails de cette étape sont disponibles dans les tables 4.2 et 4.3.

Item	Modalités initiales	Modalités agrégées	Différence AIC	Différence BIC
8	{1,2,3,4,5,6}	{3,4}	-737.66	-742.97
10	{1,2,3,4,5,6}	{3,4}	-1416.97	-1427.61
12	{1,2,3,4,5,6}	{3,4}	-2197.61	-2213.57
14	{1,2,3,4,5,6}	{3,4}	-3481.02	-3507.62
16	{1,2,3,4,5}	{1,2} et {3,4}	-4583.37	-4620.6
17	{1,2,3,4,5}	{1,2} et {3,4}	-5806.22	-5854.09
21	{1,2,3,4,5}	{1,2} et {3,4}	-6786.85	-6845.36
22	{1,2,3,4,5}	{1,2} et {3,4}	-8353.32	-8422.47

Table 4.2 – Etape de recodage et qualité d'ajustement du modèle. Les différences des valeurs des critères d'information d'Akaike (AIC) et Bayes (BIC) sont calculées après chaque recodage.

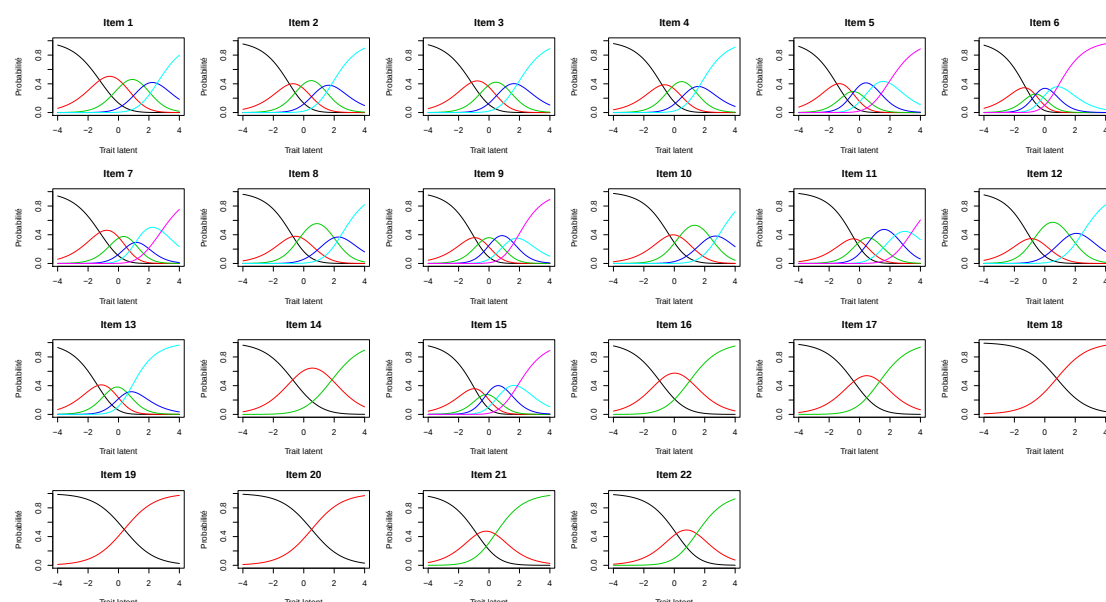


Figure 4.2 – Courbes caractéristiques des 22 items de la banque.

La figure 4.3 montre la distribution empirique du trait latent ainsi que la courbe de l'information de la banque d'items en fonction des différentes valeurs du trait latent. La courbe du trait latent obtenu est en forme de cloche, ne paraît pas symétrique, et son mode avoisine la valeur 1. La courbe d'information de la banque d'items observe un pic au voisinage de $\theta = 0.13$, et 55.8% de l'information totale est observée sur l'intervalle de valeurs de traits latents $[0, 4]$.

Item	Seuil 1	Seuil 2	Seuil 3	Seuil 4	Seuil 5	INFIT
1	-1.24	0.29	1.65	2.53	-	0.96
2	-0.93	-0.19	1.21	1.78	-	0.94
3	-1.16	-0.09	1.08	1.85	-	0.76
4	-0.80	-0.16	1.19	1.63	-	1.03
5	-1.49	-0.53	-0.35	0.91	1.91	0.77
6	-1.24	-0.66	-0.57	0.33	0.84	0.78
7	-1.23	0.04	1.01	1.12	2.83	0.95
8	-0.71	-0.23	1.98	2.37	-	0.93
9	-0.95	-0.45	0.35	1.42	1.82	1.09
10	-0.31	0.34	2.36	2.87	-	1.02
11	-0.34	0.04	0.78	2.32	3.40	1.09
12	-0.91	-0.72	1.62	2.41	-	1.03
13	-1.37	-0.50	0.55	0.66	-	1.04
14	-0.72	1.85	-	-	-	0.90
15	-0.94	-0.29	-0.16	1.08	1.86	1.27
16	-0.97	1.02	-	-	-	0.99
17	-0.38	1.32	-	-	-	0.94
18	0.78	-	-	-	-	0.84
19	0.37	-	-	-	-	0.87
20	0.52	-	-	-	-	0.91
21	-0.78	0.40	-	-	-	1.05
22	0.14	1.46	-	-	-	1.10

Table 4.3 – Seuils de difficulté des items et ajustement des items après recodage.

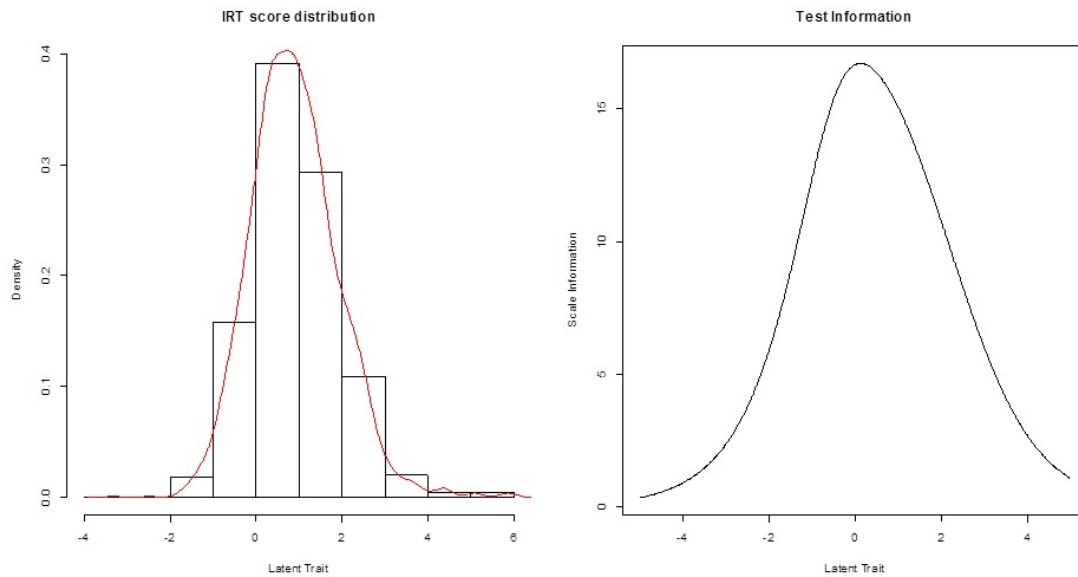


Figure 4.3 – Distribution du score issu de la banque d'items et courbe d'information de la banque.

4.1.3.4 Comportement différentiel des items

Sur les 132 tests de comportement différentiel d'items réalisés (c'est-à-dire 22 items et 6 facteurs de confusion), 50 montrent un comportement différentiel global (voir Table 4.4). Au regard de la classification définie par Zumbo, aucun item ne présente un comportement différentiel de magnitude modérée ou élevée. Quelques items présentent cependant un comportement différentiel de magnitude négligeable : 3 pour le genre (items 2, 3 et 4), 8 pour l'âge (items 6, 8, 10, 11, 14, et 18 à 20), 4 pour le niveau scolaire (18 à 21), 2 pour le statut familial (items 12 et 13), 15 items pour le sous-type de sclérose (items 2 à 4, 6 à 11, 13 à 15, 18, 19, et 21) et 18 items pour la zone géographique (1 à 13, 15 à 17, 20, et 22). Etant donné que les magnitudes sont toutes négligeables, les 22 items sont conservés dans la banque suite à cette analyse.

4.1.3.5 Validité externe de la banque d'items

Les résultats de validité externe sont présentés à la table 4.5. La moyenne du score de qualité de vie liée à la santé mentale (issu de la banque d'items) vaut 67.6 (l'écart-type vaut 18.1). L'âge n'est pas corrélé avec le score. Les scores sont significativement plus élevés pour les hommes, pour les patients ayant un niveau scolaire élevé, ainsi que chez les actifs. Aucune différence ne semble apparaître vis-à-vis de la situation familiale. Le score issu de la banque est négativement

Item	Genre		Age		Niveau scolaire		Situation familiale		Sous-type de sclérose		Zone géographique	
	p-valeur	ΔR^2	p-valeur	ΔR^2	p-valeur	ΔR^2	p-valeur	ΔR^2	p-valeur	ΔR^2	p-valeur	ΔR^2
1	0.08	-	0.59	-	0.41	-	0.36	-	0.02	-	0.00	0.01
2	0.00	0.01	0.33	-	0.08	-	0.62	-	0.00	0.01	0.00	0.01
3	0.00	0.00	0.26	-	0.71	-	0.11	-	0.00	0.00	0.00	0.01
4	0.00	0.04	0.35	-	0.04	-	0.29	-	0.00	0.01	0.01	0.00
5	0.26	-	0.51	-	0.27	-	0.43	-	0.05	-	0.00	0.01
6	0.81	-	0.00	0.00	0.22	-	0.77	-	0.00	0.01	0.00	0.02
7	0.82	-	0.95	-	0.50	-	0.86	-	0.00	-	0.00	0.01
8	0.17	-	0.00	0.01	0.19	-	0.36	-	0.00	0.01	0.00	0.03
9	0.26	-	0.05	-	0.04	-	0.73	-	0.00	0.01	0.00	0.04
10	0.34	-	0.00	0.02	0.42	-	0.05	-	0.00	0.02	0.00	0.02
11	0.27	-	0.00	0.01	0.32	-	0.25	-	0.00	0.01	0.00	0.01
12	0.20	-	0.02	-	0.38	-	0.01	0.00	0.18	-	0.00	0.03
13	0.39	-	0.13	-	0.52	-	0.00	0.00	0.00	0.00	0.00	0.01
14	0.34	-	0.00	0.01	0.11	-	0.06	-	0.00	0.01	0.14	-
15	0.85	-	0.06	-	0.63	-	0.07	-	0.00	0.01	0.00	0.01
16	0.38	-	0.06	-	0.54	-	0.60	-	0.13	-	0.00	0.01
17	0.92	-	0.77	-	0.27	-	0.81	-	0.50	-	0.00	0.01
18	0.58	-	0.00	0.02	0.00	0.01	0.28	-	0.00	0.01	0.11	-
19	0.56	-	0.00	0.01	0.00	0.01	0.33	-	0.00	0.02	0.24	-
20	0.31	-	0.00	0.01	0.01	0.01	0.63	-	0.04	-	0.00	0.02
21	0.60	-	0.07	-	0.01	0.01	0.34	-	0.00	0.01	0.15	-
22	0.85	-	0.56	-	0.48	-	0.95	-	0.24	-	0.00	0.01

Table 4.4 – Résultats relatifs au comportement différentiel des items. Les valeurs en gras correspondent à des p-valeurs < 0.01 . La mesure ΔR^2 représente la magnitude du comportement différentiel d'items.

		R	M (SD)	p-value
Données sociodémographiques				
Age		-0.07	-	< 0.01
Genre	Homme	-	69.61 (17.01)	< 0.01
	Femme	-	66.80 (18.45)	
Niveau scolaire	Elevé	-	71.37 (16.90)	< 0.01
	Bas	-	64.99 (18.34)	
Situation familiale	En couple	-	68.44 (18.73)	0.26
	Seul	-	67.05(17.93)	
Activité professionnelle	Actif	-	70.03 (17.33)	< 0.01
	Sans emploi	-	63.29 (18.88)	
Données cliniques				
Score EDSS		-0.18	-	< 0.01
Echelle de symptômes		-0.44	-	< 0.01
Durée de la maladie		0.00	-	0.96
Sous-type de sclérose	RR	-	68.51 (18.11)	< 0.01
	PP	-	68.42 (17.29)	
	SP	-	63.65 (17.60)	
	CIS	-	76.33 (18.96)	
Données de qualité de vie				
MusiQoL	ADL	0.56	-	< 0.01
	PWB	0.81	-	< 0.01
	RFr	0.23	-	< 0.01
	SPT	0.50	-	< 0.01
	Rfa	0.27	-	< 0.01
	RHCS	0.22	-	< 0.01
	SSL	0.36	-	< 0.01
	COP	0.57	-	< 0.01
	REJ	0.54	-	< 0.01
	Index	0.80	-	< 0.01
SF-36	PF	0.37	-	< 0.01
	SF	0.67	-	< 0.01
	RP	0.52	-	< 0.01
	RE	0.62	-	< 0.01
	MH	0.89	-	< 0.01
	VT	0.80	-	< 0.01
	BP	0.47	-	< 0.01
	GH	0.57	-	< 0.01
	PCS	0.33	-	< 0.01
	MCS	0.85	-	< 0.01

Table 4.5 – Comparaison des scores de qualité de vie liée à la santé mentale en fonction de données sociodémographiques, clinique, et de qualité de vie.

corrélé avec les indices cliniques (c'est-à-dire le score EDSS et l'échelle de symptômes), ces corrélations étant modérément élevées ($-0.44 \leq R \leq 0.00$). Aucune corrélation ne semble apparaître avec la durée de la maladie. Le score présente des différences significatives entre les différents groupes de sous-type de sclérose, les plus hauts scores étant observés pour les patients ayant un syndrome cliniquement isolé, et les plus faibles scores pour les patients atteints de sclérose progressive secondaire. Enfin le score issu de la banque est fortement corrélé avec les scores des dimensions issues du SF-36 et du MusiQoL qui mesurent des concepts proches de notre banque d'items.

4.1.4 Conclusion

Les travaux présentés dans cette section représentent la première proposition de développement et de calibration d'une banque d'items mesurant la qualité de vie pour des patients atteints de sclérose en plaques, qui peut être utilisée dans différents pays et dans des études internationales. La banque d'items obtenue démontre de bonnes propriétés psychométriques ainsi qu'une validité cliniques satisfaisante.

Bien qu'une large part de la recherche soit aujourd'hui concentrée sur le développement de banques d'items et questionnaires adaptatifs (par exemple les initiatives PROMIS et Neuro-QoL), la banque d'items décrite dans cette section peut jouer un rôle dans la pratique clinique et la recherche sur la sclérose en plaques. En effet notre banque d'items est basée sur deux questionnaires qui sont largement utilisés de nos jours (c'est-à-dire le MusiQoL et le SF-36). Ces deux questionnaires ont été développés et validés simultanément dans de nombreux pays, la banque d'items résultante peut donc être appliquée internationalement.

La première application future de cette banque d'items consiste à développer un questionnaire adaptatif. La fin de ce chapitre présentera le développement d'un questionnaire adaptatif basé sur cette banque d'items (voir section 4.3). Il pourrait aussi être possible de compléter la banque avec d'autres dimensions de la qualité de vie (par exemple des dimensions physiques ou sociales) afin de préserver la structure multidimensionnelle de la qualité de vie [1], et ainsi développer un questionnaire adaptatif multidimensionnel. La section suivante s'intéresse à ce type de questionnaires.

4.2 Questionnaires adaptatifs multidimensionnels basés sur la théorie de réponse à l'item

Dans cette section, nous nous intéressons aux questionnaires adaptatifs informatisés multidimensionnels basés sur la théorie de réponse à l'item. Un questionnaire adaptatif est un outil puissant qui permet d'améliorer la précision d'une mesure (un score de qualité de vie par exemple) tout en réduisant le nombre total d'items requis dans les questionnaires utilisés. Dans ce type de questionnaires, un ensemble d'items adapté à l'individu est administré, construit successivement en prenant en compte l'information contenue dans les réponses aux items précédents. Les questionnaires adaptatifs sont initialement basés sur la théorie de réponse à l'item unidimensionnelle, ils sont utilisés pour choisir les items qui doivent être administrés lors de la passation d'un questionnaire. Aujourd'hui, il est possible de développer des questionnaires adaptatifs multidimensionnels, permettant d'améliorer la précision de la mesure et de réduire le nombre d'items nécessaires pour mesurer des traits latents multiples (différentes dimensions de la qualité de vie par exemple).

4.2.1 Introduction

L'application des questionnaires adaptatifs dans la pratique clinique est une approche qui permet de réduire la longueur d'un questionnaire. Contrairement aux questionnaires fixes (questionnaires "papier-crayon", ou tout questionnaire administrant des items de façon séquentielle), les questionnaires adaptatifs permettent de sélectionner des items optimaux. Les méthodes basées sur la théorie

Cette section présente le développement de deux questionnaires adaptatifs multidimensionnels, basés sur la théorie de réponse à l'item multidimensionnelle. La sous-section 4.2.2 présente le développement du MusiQoL-MCAT, qui est une version adaptative du questionnaire MusiQoL, la section 4.2.3 présente le développement du SQoL-MCAT, qui est la version adaptative du questionnaire SQoL à 41 items.

4.2.2 Le MusiQoL-CAT : Développement d'un questionnaire adaptatif multidimensionnel mesurant la qualité de vie chez les patients atteints de sclérose en plaques

Ici, nous présentons brièvement le développement du questionnaire adaptatif multidimensionnel MusiQoL-MCAT. Ce questionnaire adaptatif a été développé à partir des items du questionnaire MusiQoL.

Les données de 1992 patients utilisées ont déjà été utilisées dans des travaux précédents (voir la section 4.1 par exemple). Le développement du MusiQoL-MCAT est basé sur l'ajustement d'un modèle de réponse à l'item multidimension-

nelle et sur des simulations de questionnaire adaptatif. L'algorithme du questionnaire adaptatif est basé sur la procédure de sélection d'items de Kullback-Leibler. Nous analysons différents scénarios basés sur un nombre fixe d'items administrés. La précision du score obtenu est mesurée en utilisant les corrélations (r) entre les scores initiaux issus du modèle de réponse à l'item et les scores issus des questionnaires adaptatifs. On utilise également l'erreur type de mesure (notée SEM) ainsi que la racine de l'erreur quadratique (notée $RMSE$).

Le modèle multidimensionnel de réponse graduée est utilisé pour estimer les paramètres d'items ainsi que des scores multidimensionnels pour chaque patient. Parmi les différentes simulations de questionnaires adaptatifs, on définit finalement la version du MusiQoL-MCAT à 16 items comme étant celles obtenant les niveaux de précision les plus satisfaisants, selon certaines valeurs seuils définies au préalable ($r \geq 0.9$, $SEM \leq 0.55$, $RMSE \leq 0.3$). La validité externe du MusiQoL-MCAT est également satisfaisante.

Le MusiQoL-MCAT présente des propriétés psychométriques satisfaisantes et peut s'adapter individuellement à la qualité de vie de chaque patient. Ce nouvel instrument permet de réduire le fardeau ressenti par les patients atteints de sclérose en plaques lors du remplissage d'un questionnaire de qualité de vie et est mieux adapté pour l'utilisation dans la pratique clinique. Les détails du développement du MusiQoL-MCAT sont présentés dans un article [122].

4.2.3 Le SQoL-CAT : Développement d'un questionnaire adaptatif multidimensionnel mesurant la qualité de vie chez les patients atteints de schizophrénie

Dans cette section, nous présentons les résultats d'une étude ayant pour but de développer un questionnaire adaptatif multidimensionnel mesurant la qualité de vie spécifique aux patients atteints de schizophrénie, basé sur la théorie de réponse à l'item. La première partie introduit le contexte de cette étude. Les sections suivantes présentent la méthodologie utilisée et les principaux résultats obtenus.

4.2.3.1 Introduction

Malgré le besoin grandissant de mesurer la qualité de vie chez les patients atteints de schizophrénie, les outils de mesure de la qualité de vie ne sont pas encore utilisés et implémentés couramment dans la pratique clinique [13]. Parmi les différentes raisons expliquant ce manque d'implémentation [27], une des limites exprimées par les cliniciens concerne le fardeau administratif associé au manque d'efficacité de l'évaluation de la qualité de vie dans les hôpitaux et centres de santé [125, 81, 80]. La plupart des questionnaires de qualité de vie sont des questionnaires papier-crayon, ce qui rend très difficile pour les cliniciens

l'obtention de scores de qualité de vie en temps réel. De plus, ces questionnaires sont souvent trop longs, le nombre d'items qu'ils contiennent est fixé, cela rend l'analyse des patients fastidieuse et fait également apparaître des problèmes dus aux données manquantes potentielles [75].

De nos jours, il est donc essentiel de pouvoir fournir des nouvelles mesures de la qualité de vie, combinant à la fois les nouvelles technologies (en informatique) et la théorie moderne de la mesure, afin de limiter autant que possible les différentes limites administratives et le potentiel fardeau ressenti par les patients lors du remplissage de questionnaires. De nombreuses études ont été menées dans ce sens ces dernières années, la plus significative étant la fondation d'un programme de recherche américain soutenu par le Institut National de Santé (National Institutes of Health, NIH), appelé PROMIS (pour Patient-Reported Outcomes Measurement Information System) [38, 39]. Les méthodes basées sur les modèles multidimensionnels de la théorie de réponse à l'item ainsi que le développement de questionnaires adaptatifs multidimensionnels peuvent être utilisés pour s'affranchir des différentes limites rencontrées avec l'utilisation des questionnaires traditionnels [55, 59].

Les questionnaires adaptatifs multidimensionnels permettent de mesurer la qualité au travers de plusieurs scores de dimensions, estimés simultanément. Ils permettent de n'administrer que les items les plus pertinents pour un individu donné, impliquant une réduction du nombre d'items administrés et du temps de remplissage, tout en assurant un niveau de précision de la mesure satisfaisant [170, 143, 89]. Plusieurs développements de questionnaires adaptatifs multidimensionnels ont été récemment entrepris, pour mesurer la gravité de différents problèmes de santé (par exemple symptomatologie, fatigue, fonctionnement physique et émotionnel) pour différentes maladies chroniques (par exemple la santé mentale chez les enfants ou encore le cancer) [67, 136, 78], mais à notre connaissance aucun outil n'est encore disponible à ce jour pour mesurer la qualité de vie chez les patients atteints de schizophrénie.

Dans la suite de la section, nous présentons le développement d'un questionnaire adaptatif multidimensionnel pour les patients atteints de schizophrénie, le SQoL-MCAT. Nous nous appuyons sur un questionnaire existant, le SQoL à 41 items, qui est (et a été) largement utilisé dans de nombreuses applications en prenant en compte différentes populations et différents paramètres (par exemple des patients présentant des troubles cognitifs, des patients sans domicile fixe, des patients hospitalisés ou encore dans des études d'imagerie cérébrale).

4.2.3.2 Méthodes

Le questionnaire SQoL à 41 items Le questionnaire sur lequel est basée cette étude est la version du SQoL à 41 items. C'est un questionnaire de qualité de vie spécifique à la schizophrénie, auto-rapporté et multidimensionnel. Le contenu de ce questionnaire est décrit à la section 1.1.2. Le contenu des items est présentée

Item	Items en français	Items en anglais
	Actuellement,	At the present time,
1	j'ai confiance en la vie	I'm confident in life
2	je me bats pour réussir dans la vie	I fight to succeed in my life
3	je fais des projets professionnels et/ou personnels pour l'avenir	I'm able to plan for my professional or personal future
4	je réalise mes projets professionnels et/ou personnels	I'm able to achieve my professional or personal projects
5	j'ai confiance en moi	I feel self-confident
6	je suis heureux(se)	I'm happy
7	je suis bien dans ma tête	I feel in a good mood, I'm at ease with myself
8	je suis épanoui(e)	I feel in full bloom
9	je suis libre de prendre des décisions	I feel free to take decisions
10	je suis libre d'agir	I feel free to act
11	j'ai une vie active	I have an active life
12	je fais des efforts pour travailler	I make efforts to work
13	je peux sortir (cinéma, promenade, restaurant ...)	I'm able to go out (cinema, walking, restaurant ...)
14	je réalise mes projets familiaux, sentimentaux	I'm able to achieve my family and sentimental projects
15	je suis en bonne forme physique	I'm in good physical shape
16	je suis plein(e) d'énergie	I'm full of energy
17	je fais du sport, j'ai des activités physiques	I do sports, I practise physical activities
18	j'ai une vie stable, équilibrée	I have a stable well balanced life style
19	je peux parler à ma famille	I'm able to talk with my family
20	je suis aidé(e) par ma famille	I'm helped, supported by my family
21	je suis compris(e) par ma famille	I'm understood by my family
22	je suis autonome, indépendant (e) par rapport à ma famille	I'm self-sufficient, independent of my family
23	je vois ma famille	I see, meet my family
24	je suis écouté(e) par ma famille	My family pays attention to me
25	je vois, j'invite mes amis (proches)	I see, invite my friends or my relatives
26	je peux me confier à quelqu'un	I'm able to confide in someone
27	je suis aidé(e) par mes amis (proches)	I'm helped, supported by my friends or my relatives
28	je suis compris(e) par mes amis (proches)	I'm understood by my friends or my relatives
29	j'ai des amis	I have friends
30	j'ai une vie sentimentale satisfaisante	I'm satisfied with my love life
31	je suis à l'aise en public	I feel comfortable when in public
32	j'ai peur de l'avenir	I fear for my future
33	je me sens inutile	I feel useless
34	je suis angoissé(e)	I feel anxious
35	je suis seul(e)	I feel lonely
36	j'ai des difficultés à me concentrer, à réfléchir	I have difficulty concentrating, thinking straight
37	je m'ennuie	I get bored
38	je suis coupé(e) du monde extérieur	I feel myself cut-off from the outside world
39	je crains d'accomplir des formalités administratives	I fear accomplishing administrative procedures
40	j'ai du mal à exprimer ce que je ressens	I have difficulty expressing my feelings
41	j'ai des difficultés à m'intéresser aux choses qui m'entourent	I have difficulty paying attention things to my surroundings

Table 4.6 – Contenu des 41 items du SQoL en français et en anglais.

dans la table 4.6.

Conception de l'étude et population La base de données utilisée dans cette étude provient de l'agrégation de 4 études qui ont précédemment été menées par les membres du groupe de travail SQoL (composé de professionnels de la santé publique, de psychiatres, de psychologues et de statisticiens), dans lesquelles le SQoL a été utilisé pour mesurer la qualité de vie des patients. La base de données contient les informations de 517 patients, recrutés dans les hôpitaux psychiatriques de différentes villes de France (Marseille, Lyon et Toulon). Les critères d'inclusion sont les suivants : un diagnostic de schizophrénie selon la 4ème édition du manuel diagnostique et statistique des troubles mentaux (DSM-IV-TR) [11], âge supérieur à 18 ans, un consentement écrit du patient et le français comme langue maternelle. Les critères de non-inclusion sont les suivants : un diagnostic autre que la schizophrénie, maladie organique ou retard mental.

Récolte des données En plus des données relatives au SQoL, les données ont été récoltées pour cette étude :

- Données sociodémographiques : âge (en années), genre (homme ou femme), niveau scolaire (scolarisé moins de 12 ans ou scolarisé 12 ans ou plus) et

statut du patient (hospitalisé ou ambulatoire).

- Données cliniques : durée de la maladie (en années), sévérité des symptômes psychotiques basée sur l'échelle de symptômes positifs et négatifs (PANSS), comprenant trois facteurs (positif, négatif et psychopathologie générale) [98, 100], et dépression, mesurée par l'échelle de dépression de Calgary pour la schizophrénie (CDSS) [2].
- Données de qualité de vie : La qualité de vie est également mesurée en utilisant le questionnaire générique SF-36, décrit dans la section 1.1.2.

Développement du questionnaire adaptatif SQoL-MCAT Le développement du SQoL-MCAT est divisé en trois étapes distinctes : une analyse des 41 items basée sur la théorie de réponse à l'item multidimensionnelle, des simulations de questionnaires adaptatifs suivies d'analyse de la précision et l'analyse de la validité externe du questionnaire adaptatif.

1. Analyse basée sur la théorie de réponse à l'item multidimensionnelle : Dans un premier temps nous calculons le taux de données manquantes pour chaque item. La structure à 8 dimensions définie par les concepteurs du SQoL est tout d'abord validée par une analyse factorielle confirmatoire, en utilisant un estimateur des moindres carrés pondérés (adaptés aux données ordinales), via le logiciel MPlus [129], afin de vérifier la validité de construit des 41 items. La qualité d'ajustement du modèle est mesurée en utilisant l'erreur quadratique moyenne d'approximation ($RMSEA < 0.07$ attendu) ainsi que l'indice d'ajustement comparatif ($CFI > 0.95$ attendu) [159]. Nous ajustons ensuite un modèle de réponse graduée multidimensionnel [151, 152] à nos données. C'est un modèle "inter-item", c'est-à-dire que la réponse d'un item n'influe qu'une seule dimension, cependant on modélise les corrélations entre les différentes dimensions. Ce modèle est choisi après avoir testé un autre modèle multidimensionnel, le modèle de crédit partiel généralisé multidimensionnel [178]. Le modèle de réponse graduée multidimensionnel est finalement retenu au regard des critères d'information d'Akaike [7] (AIC) et de Bayes [153] (BIC).

Les paramètres des items sont estimés avec la méthode Metropolis-Hastings Robbins-Monro (MH-RM) [33], implémentée dans le package R *mirt* [40]. L'algorithme MH-RM calcule la vraisemblance des données complètes, imputées de façon stochastique, en considérant la population distribuée selon une loi normale multivariée. Son utilisation est préférée à celle de l'algorithme espérance-maximisation [25] lorsque le nombre de dimensions est supérieure à 3. La qualité d'ajustement des items est mesurée en utilisant la statistique $S - \chi^2$ [97], qui est adaptée aux items multidimensionnels. Les items obtenant une p-valeur inférieure à 0.05 sont considérés comme mal ajustés et peuvent être retirés de l'étude le cas échéant. Les scores des dimensions sont obtenus en utilisant une estimation bayésienne maximum

a posteriori (MAP) [55], en fonction des paramètres d'items et les réponses observées aux items. L'information moyenne de chaque item ainsi que l'information globale sont également calculées. En théorie de réponse à l'item l'information d'un item dépend des paramètres du modèle ajusté (c'est-à-dire les paramètres de difficulté et de discrimination) et est une fonction du trait latent. Un item possédant une plus grande quantité d'information est plus discriminant et fournit une plus faible erreur de mesure. L'information globale (ou information du test) est la somme de l'information de chaque item. La fiabilité de chaque dimension est aussi mesurée en utilisant les estimateurs de fiabilité marginales empiriques [139] : nous utilisons ici les mêmes seuils qu'en théorie classique, c'est-à-dire qu'un coefficient de fiabilité supérieur ou égal à 0.7 est jugé raisonnable [166].

Le modèle de réponse graduée multidimensionnel est composé de deux modèles logistiques multidimensionnel à deux paramètres, il est défini de la façon suivante :

$$P(X_{ij} = k | \theta_i, \alpha_j, \beta_{jk}) = P(X_{ij} \geq k | \theta_i, \alpha_j, \beta_{jk}) - P(X_{ij} \geq k+1 | \theta_i, \alpha_j, \beta_{j,k+1}) \quad (4.3)$$

où,

$$P(X_{ij} \geq k | \theta_i, \alpha_j, \beta_{jk}) = \frac{1}{1 + e^{-\alpha_j(\theta_i - \beta_{jk} \mathbf{1})}} \quad (4.4)$$

où i désigne l'individu i , j désigne l'item j , X_{ij} désigne la réponse ordinale qui peut prendre la valeur $k \in \{1, \dots, K\}$, α_j est le paramètre de discrimination, θ_i et le vecteur de traits latents et β_{jk} est le k -ième paramètre de seuil de difficulté .

Le comportement différentiel des items a été mesuré pour prendre en compte le potentiel biais dû à l'appartenance aux groupes suivants : genre (homme et femme), présence de symptômes paranoïdes (schizophrénie paranoïde et autre) et niveau de conscience de la maladie (aussi appelée insight, basé sur l'échelle SUMD). La méthode utilisée ici est identique à celle déjà utilisée précédemment, elle est décrite à la section 4.1.2.

2. Simulations de questionnaires adaptatifs et analyse de la précision : Maintenant que le modèle multidimensionnel est calibré (c'est-à-dire ajusté à nos données), il est possible de réaliser des simulations de questionnaires adaptatifs à partir de ce modèle. L'idée est de proposer plusieurs scénarios, c'est-à-dire définir plusieurs paramètres pour mettre en place un algorithme adaptatif d'administration d'items. Nous utilisons une méthode de simulation basée sur les données observées. Nous utilisons donc les réponses aux 41 items du SQoL contenus dans notre base de données pour simuler l'administration des items, comme cela est indiqué dans une étude

récente [8]. Pour chaque patient ayant complété les 41 items du SQoL, on simule une administration adaptative d'un sous-ensemble d'items. L'algorithme repose sur une procédure de sélection d'items basée sur l'information de Kullback-Liebert [126]. Comme critère initial, nous choisissons comme premier item l'item ayant la quantité moyenne d'information la plus élevée. La sélection de l'item suivant dépend des réponses à l'item précédent, qui sont issues des données observées. Après chaque réponse à un item, le trait latent de l'individu est ré-estimé en utilisant l'estimation bayésienne MAP. Le critère d'arrêt est basé sur un niveau de précision que doit atteindre l'individu, basé sur l'erreur standard de mesure (notée SEM). Un intervalle d'acceptabilité est défini par $[0.33, 0.55]$, correspondant à des coefficients de fiabilité compris entre 0.7 et 0.9 [83]. Le questionnaire adaptatif est donc simulé en fixant trois valeurs seuils de SEM (0.33, 0.44 et 0.55), c'est-à-dire le critère est vérifié lorsque le seuil est atteint pour chaque dimension. Pour ces trois simulations, les scores issus du questionnaire adaptatif sont calculés, ainsi que la précision de la mesure. La précision est mesurée en utilisant les coefficients de corrélation entre les scores issus du questionnaire adaptatif et les scores issus de l'ensemble initial de 41 items (des coefficients supérieurs à 0.9 sont attendus pour chaque dimension) ainsi que la racine de l'erreur quadratique (notée RMSE). Plus les valeurs du RMSE sont proches de 0, meilleure est la précision de la mesure. Le RMSE permet d'évaluer l'écart entre les scores des 41 items et les scores du questionnaire adaptatif. Les deux scores étant sur la même métrique et standardisés, des valeurs inférieures ou égales à 0.3 reflètent un excellent niveau de précision [44]. Finalement, la meilleure version du SQoL-MCAT est sélectionnée comme étant celle qui minimise le nombre d'items administrés tout en obtenant un niveau satisfaisant de précision.

3. Validité externe du SQoL-MCAT : Afin de mesurer la validité des scores issus du SQoL-MCAT, nous évaluons la validité convergente et divergente. Pour la validité convergente, nous utilisons les corrélations de Pearson pour évaluer la relation entre les scores du SQoL-MCAT et ceux du SF-36, l'hypothèse sous-jacente étant que les scores du SQoL-MCAT (c'est-à-dire les scores obtenus par estimation MAP) devraient être plus corrélés avec les scores des dimensions du SF-36 mesurant des aspects similaires à ceux du SQoL-MCAT. La validité divergente est déterminée en explorant les relations entre les scores du SQoL-MCAT et différentes caractéristiques socio-démographiques (âge, genre, niveau scolaire, statut du patient) et cliniques (durée de la maladie, scores PANSS et score CDSS), en utilisant des tests t de Student et des corrélations de Pearson. En accord avec des études précédentes, on s'attend à ce que les scores du SQoL-MCAT ne diffèrent en fonction des données sociodémographiques, qu'ils soient plus élevés chez les patients ambulatoires, ne dépendent pas de la durée de la maladie et qu'enfin, ils ne soient pas négativement corrélés avec la sévérité de la ma-

Données sociodémographiques		
Age (M $\pm$ SD)	Années	36.5 $\pm$ 10.9
Genre (N [%])	Homme	362 [70%]
	Femme	154 [30%]
Niveau scolaire (N [%])	< 12 ans	162 [79%]
	$\geq$ 12 ans	43 [21%]
Statut du patient (N [%])	Hospitalisé	131 [64%]
	Ambulatoire	74 [36%]
Données cliniques		
Durée de la maladie	Années	13.8 $\pm$ 9.3
Score PANSS (M $\pm$ SD)	Score total	69.6 $\pm$ 18.4
	Score positif	15.7 $\pm$ 6.1
	Score négatif	19.2 $\pm$ 7.0
	Score psychopathologie générale	35.8 $\pm$ 9.7
Score CDSS (M $\pm$ SD)	Score total	3.1 $\pm$ 3.6
Données de qualité de vie		
SF-36 (M $\pm$ SD)	PF	77.8 $\pm$ 23.3
	SF	52.8 $\pm$ 28.8
	RP	42.4 $\pm$ 36.1
	RE	40.9 $\pm$ 41.0
	MH	56.1 $\pm$ 20.3
	VT	48.1 $\pm$ 19.6
	BP	63.7 $\pm$ 25.9
	GH	55.2 $\pm$ 21.7
	PCS	46.4 $\pm$ 8.2
	MCS	37.0 $\pm$ 11.1

Table 4.7 – Caractéristiques descriptives de l'échantillon d'étude.

ladié.

4.2.3.3 Résultats

L'échantillon d'étude comprend 571 patients atteints de schizophrénie. La moyenne d'âge est de 36.5 ans (l'écart-type vaut 10.8), et 29.8% des patients sont des femmes. la durée moyenne de la maladie est de 13.8 ans (l'écart-type vaut 9.3). Les patients ont une sévérité des symptômes modérée, avec un score de PANSS total de 69.6 (l'écart-type vaut 18.4). Toutes ces caractéristiques sont présentées dans la table 4.7.

Analyse basée sur la théorie de réponse à l'item multidimensionnelle La structure à 8 facteurs testée dans le modèle d'analyse factorielle confirmatoire montre des indices de qualité d'ajustement satisfaisants ($RMSE = 0.05$, $CFI =$

Item	Dimension	DM (%)	α	β_1	β_2	β_3	β_4	PIT	$S - \chi^2$ (p-valeur)
1	SE	2.71%	2.136	-1.64	-0.92	-0.13	0.92	2.39%	0.20
2	RE	4.26%	1.731	-1.65	-0.84	-0.30	0.62	1.62%	0.03
3	RE	4.84%	1.58	-1.41	-0.71	-0.18	0.71	1.39%	0.91
4	RE	5.22%	1.814	-1.19	-0.44	0.31	1.26	1.82%	0.85
5	SE	1.93%	2.2	-1.65	-0.89	-0.04	1.02	2.53%	<0.01
6	SE	2.71%	2.367	-1.67	-0.92	-0.03	1.15	2.87%	0.09
7	SE	2.13%	2.826	-1.92	-1.01	-0.12	1.10	3.89%	0.24
8	SE	3.87%	2.654	-1.57	-0.70	0.28	1.44	3.55%	0.02
9	AU	2.90%	3.045	-2.15	-1.21	-0.42	1.35	4.22%	0.22
10	AU	2.32%	3.376	-2.54	-1.56	-0.56	1.50	4.82%	0.32
11	RE	4.06%	1.988	-1.10	-0.37	0.41	1.41	2.15%	0.30
12	RE	8.51%	1.609	-1.41	-0.77	-0.13	0.84	1.44%	0.67
13	AU	3.09%	1.137	-1.08	-0.56	-0.16	0.78	0.74%	0.89
14	SL	6.19%	2.314	-1.04	-0.28	0.43	1.47	2.83%	0.53
15	PhW	2.13%	2.319	-1.79	-1.07	-0.06	1.27	2.73%	0.07
16	PhW	4.26%	2.553	-1.54	-0.72	0.21	1.44	3.32%	0.15
17	PhW	6.19%	1.254	-0.68	-0.10	0.36	1.08	0.90%	0.65
18	PhW	3.48%	1.832	-1.39	-0.83	-0.08	1.25	1.79%	0.04
19	Rfa	4.45%	2.151	-1.58	-0.99	-0.43	0.91	2.33%	0.22
20	Rfa	6.00%	2.9	-1.92	-1.30	-0.65	0.91	3.78%	0.06
21	Rfa	7.35%	3.136	-1.96	-1.14	-0.31	1.27	4.53%	0.27
22	AU	5.61%	1.177	-1.07	-0.62	-0.09	0.77	0.80%	0.25
23	Rfa	4.06%	1.706	-1.39	-0.97	-0.52	0.76	1.50%	0.12
24	Rfa	6.00%	3.378	-2.01	-1.30	-0.49	1.28	4.92%	0.26
25	RFr	4.64%	2.12	-0.96	-0.33	0.25	1.36	2.37%	0.22
26	RFr	2.90%	1.535	-1.18	-0.68	-0.21	0.97	1.29%	0.17
27	RFr	7.54%	2.775	-1.43	-0.49	0.06	1.65	3.72%	0.17
28	RFr	7.35%	2.588	-1.56	-0.74	-0.04	1.65	3.28%	0.60
29	RFr	6.77%	2.575	-1.23	-0.47	0.10	1.50	3.30%	0.12
30	SL	7.54%	2.314	-0.64	-0.01	0.67	1.53	2.71%	0.94
31	SE	2.71%	1.506	-0.84	-0.35	0.16	1.17	1.26%	0.82
32	PsW	3.29%	1.841	-1.13	-0.56	0.07	0.62	1.82%	0.14
33	PsW	5.61%	2.359	-1.68	-1.02	-0.47	0.12	2.54%	0.23
34	PsW	5.03%	2.109	-1.63	-0.91	-0.41	0.23	2.16%	0.23
35	PsW	5.03%	1.615	-1.10	-0.57	-0.16	0.48	1.41%	0.85
36	PsW	3.87%	1.841	-1.32	-0.59	-0.03	0.52	1.80%	0.30
37	PsW	4.84%	1.8	-1.39	-0.76	-0.27	0.33	1.68%	0.36
38	PsW	4.64%	2.11	-1.46	-0.85	-0.35	0.20	2.14%	0.76
39	PsW	5.80%	1.442	-1.08	-0.59	-0.18	0.35	1.13%	0.19
40	PsW	3.87%	1.983	-1.43	-0.76	-0.25	0.47	2.03%	0.88
41	PsW	4.45%	2.268	-1.66	-1.02	-0.42	0.28	2.47%	0.46

Table 4.8 – Caractéristiques des items. SE : self-esteem ; RE : resilience ; AU : autonomy ; SL : sentimental life ; PhW : physical well-being ; Rfa : relationships with family ; RFr : relationships with friends ; PsW : psychological well-being. α : paramètre de discrimination associé à la dimension de l’item. $\beta_1, \beta_2, \beta_3, \beta_4$: paramètres de seuils de difficulté. DM : Données manquantes. PIT : Proportion d’information totale.

0.95). Les caractéristiques des items (c’est-à-dire les taux de données manquantes, les paramètres d’items, l’information des items et l’indice d’ajustement multidimensionnel) sont présentées dans la table 4.8, et la distribution du trait latent pour chaque dimension est illustrée dans la figure 4.4. La matrice de corrélations des 8 traits latents est présentée dans la table 4.9.

	PsW	SE	RFa	RFr	RE	PhW	AU	SL
PsW	1	0.7	0.3	0.38	0.59	0.63	0.55	0.58
SE	0.7	1	0.3	0.5	0.75	0.77	0.63	0.73
Rfa	0.3	0.3	1	0.37	0.24	0.26	0.24	0.31
RFr	0.38	0.5	0.37	1	0.52	0.5	0.44	0.54
RE	0.59	0.75	0.24	0.52	1	0.75	0.68	0.69
PhW	0.63	0.77	0.26	0.5	0.75	1	0.63	0.69
AU	0.55	0.63	0.24	0.44	0.68	0.63	1	0.57
SL	0.58	0.73	0.31	0.54	0.69	0.69	0.57	1

Table 4.9 – Matrice de corrélations des huit dimensions du SQoL.

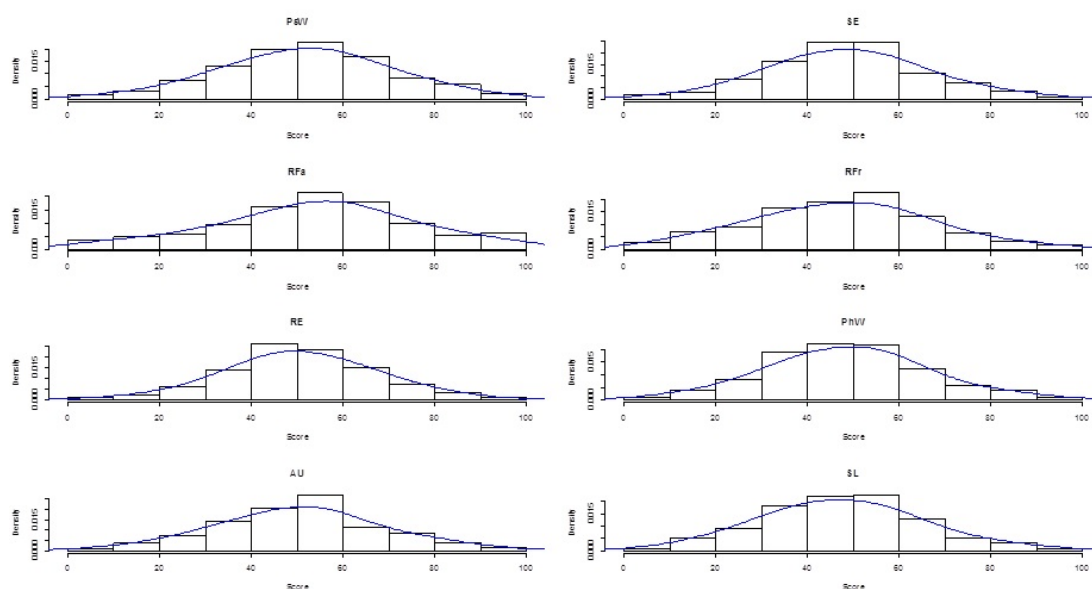


Figure 4.4 – Distribution des scores issus des 41 items du SQoL, obtenus par estimation bayésienne maximum a posteriori.

L'item 24 issu de la dimension relations avec la famille ("Je suis écouté(e) par ma famille") est l'item le plus informatif parmi les 41 items, et l'item 13 issue de la dimension autonomie est l'item le moins informatif ("Je peux sortir (cinéma, promenade, restaurant..."). La fiabilité est élevée pour chaque dimension (variant entre 0.80 et 0.92).

La qualité d'ajustement des items est satisfaisante au regard de l'indice $S - \chi^2$ adapté aux items multidimensionnels. Cependant, certains pourraient être supprimés de l'étude. Nous décidons de conserver les items mal ajustés, étant donné leur haute proportion d'information globale ou encore leur qualité d'ajustement unidimensionnel dans leur dimension (mesuré par l'indice INFIT), indiquant une importance de ces items lors de l'administration adaptative.

Parmi les 123 tests de comportement différentiel d'items (41 items et 3 facteurs de confusion), 6 items montrent un comportement différentiel global (voir table 4.10). La magnitude du comportement différentiel de ces items reste cependant négligeable, au regard du critère de Zumbo, tous les items sont donc supposés invariant en fonction des différents facteurs de confusion.

Simulations de questionnaires adaptatifs et analyse de la précision Les simulations de questionnaires adaptatifs basées sur les réponses observées sont effectuées en utilisant les réponses des 348 patients ayant répondu complètement aux 41 items du SQoL. Les indicateurs de précision de chaque scénario sont présentés dans la table 4.11.

Pour chaque simulation, on considère la précision comme la corrélation entre les scores issus du questionnaire adaptatif et les scores basés sur l'ensemble des 41 items. Les huit dimensions ont une précision satisfaisante, avec des coefficients de corrélations supérieurs à 0.9. Même pour le modèle moins restrictif (c'est-à-dire $SEM < 0.55$), la précision au regard de cet indice est satisfaisante, avec des corrélations supérieures à 0.94.

La précision est améliorée lorsque les simulations sont réalisées avec une valeur seuil de SEM plus faible. Cependant, les valeurs du RMSE sont également satisfaisantes pour les trois scénarios, avec des valeurs de RMSE inférieures à 0.3 (à part pour la dimension résilience pour le cas $SEM < 0.55$, dans lequel la valeur du RMSE dépasse légèrement ce seuil).

Le nombre moyen d'items administrés est de 25 (l'écart-type vaut 5) pour le modèle basé sur le niveau de précision $SEM < 0.55$, il est de 35 (l'écart-type vaut 6) pour le modèle basé sur le niveau de précision $SEM < 0.44$ et il est de 41 (d'écart-type 0) pour le modèle basé sur le niveau de précision $SEM < 0.33$. Pour ce dernier, les 41 items du SQoL sont administrés à chaque patient, ce scénario correspond donc à la version séquentielle initiale du SQoL. Dans chaque scénario, tous les patients se voient cependant administrer au moins un item de chaque dimension.

Le modèle basé sur le niveau de précision $SEM < 0.55$ est finalement défini comme étant le plu satisfaisant parmi les 3 scénarios testés, ce modèle est celui qui administre le moins d'items, tout en conservant des niveaux satisfaisants de précision de la mesure. Le taux d'exposition de chaque item est présenté à la figure 4.5. Tous les items sont administrés au moins une fois, parmi lesquels 18 items sont administrés plus de 9 fois sur 10 (items 1, 5-11, 14-16, 20, 21, 24, 27-29, 41).

Item	Genre		Présence de symptômes paranoïdes			Niveau d'insight	
	p-valeur	ΔR^2	p-valeur		ΔR^2	p-valeur	ΔR^2
1	0.94	0.00	0.40		0.00	0.99	0.00
2	0.66	0.00	0.35		0.00	0.21	0.00
3	0.37	0.00	0.62		0.00	0.00	0.04
4	0.35	0.00	0.23		0.00	0.69	0.00
5	0.06	0.00	0.17		0.00	0.15	0.01
6	0.09	0.00	0.10		0.00	0.07	0.01
7	0.36	0.00	0.26		0.00	0.60	0.00
8	0.28	0.00	0.95		0.00	0.20	0.00
9	0.90	0.00	0.13		0.00	0.77	0.00
10	0.70	0.00	0.48		0.00	0.35	0.00
11	0.06	0.00	0.32		0.00	0.39	0.00
12	0.66	0.00	0.56		0.00	0.21	0.00
13	0.77	0.00	0.40		0.00	0.63	0.00
14	0.77	0.00	0.90		0.00	0.77	0.00
15	0.22	0.00	0.56		0.00	0.77	0.00
16	0.36	0.00	0.33		0.00	0.35	0.00
17	0.07	0.00	0.55		0.00	0.63	0.00
18	0.01	0.00	0.78		0.00	0.90	0.00
19	0.74	0.00	0.69		0.00	0.84	0.00
20	0.68	0.00	0.74		0.00	0.38	0.00
21	0.60	0.00	0.41		0.00	0.58	0.00
22	0.88	0.00	0.41		0.00	0.90	0.00
23	0.49	0.00	0.38		0.00	0.44	0.00
24	0.00	0.01	0.61		0.00	0.55	0.00
25	0.95	0.00	0.33		0.00	0.76	0.00
26	0.50	0.00	0.24		0.00	0.00	0.02
27	0.98	0.00	0.00		0.01	0.58	0.00
28	0.14	0.00	0.42		0.00	0.45	0.00
29	0.10	0.00	0.13		0.00	0.11	0.01
30	0.88	0.00	0.70		0.00	0.35	0.00
31	0.19	0.00	0.20		0.00	0.42	0.00
32	0.06	0.00	0.66		0.00	0.24	0.00
33	0.00	0.00	0.42		0.00	0.46	0.00
34	0.40	0.00	0.11		0.00	0.31	0.00
35	0.19	0.00	0.94		0.00	0.34	0.00
36	0.15	0.00	0.28		0.00	0.58	0.00
37	0.43	0.00	0.53		0.00	0.10	0.01
38	0.11	0.00	0.13		0.00	0.15	0.01
39	0.54	0.00	0.09		0.00	0.77	0.00
40	0.59	0.00	0.30		0.00	0.87	0.00
41	0.26	0.00	0.74		0.00	0.07	0.01

Table 4.10 – Comportement différentiel des items en fonction du genre (homme et femme), de la présence de symptômes paranoïdes (schizophrénie paranoïde et autres) et de l'insight (conscient et inconscient).

Niveau de précision	Indicateur	PsW	SE	Rfa	RFr	RE	PhW	AU	SL
SEM < 0.33	Score moyen	51.56	48.66	53.45	47.56	50.68	48.37	51.01	46.94
	Précision	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	RMSE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	SEM	0.30	0.28	0.32	0.34	0.37	0.37	0.37	0.43
Nombre d'items moyen		10.0	6.0	5.0	5.0	5.0	4.0	4.0	2.0
SEM < 0.44	Score moyen	51.45	48.73	53.52	47.63	50.87	48.41	51.12	46.96
	Précision	0.99	1.00	1.00	1.00	0.99	1.00	1.00	1.00
	RMSE	0.12	0.05	0.05	0.06	0.15	0.09	0.09	0.02
	SEM	0.32	0.29	0.33	0.35	0.40	0.38	0.38	0.43
Nombre d'items moyen		8.3	5.5	4.6	4.5	3.8	3.2	2.9	2.0
SEM < 0.55	Score moyen	51.72	48.68	53.59	47.53	50.84	48.39	51.05	46.92
	Précision	0.95	1.00	0.99	0.99	0.94	0.98	0.99	0.99
	RMSE	0.27	0.09	0.13	0.12	0.32	0.17	0.13	0.10
	SEM	0.39	0.30	0.34	0.36	0.48	0.40	0.39	0.44
Nombre d'items moyen		4.2	5.1	3.8	3.9	1.6	2.3	2.1	1.9

Table 4.11 – Score moyen, indicateurs de précision et nombre d'items moyen administrés pour chaque niveau de précision testé.

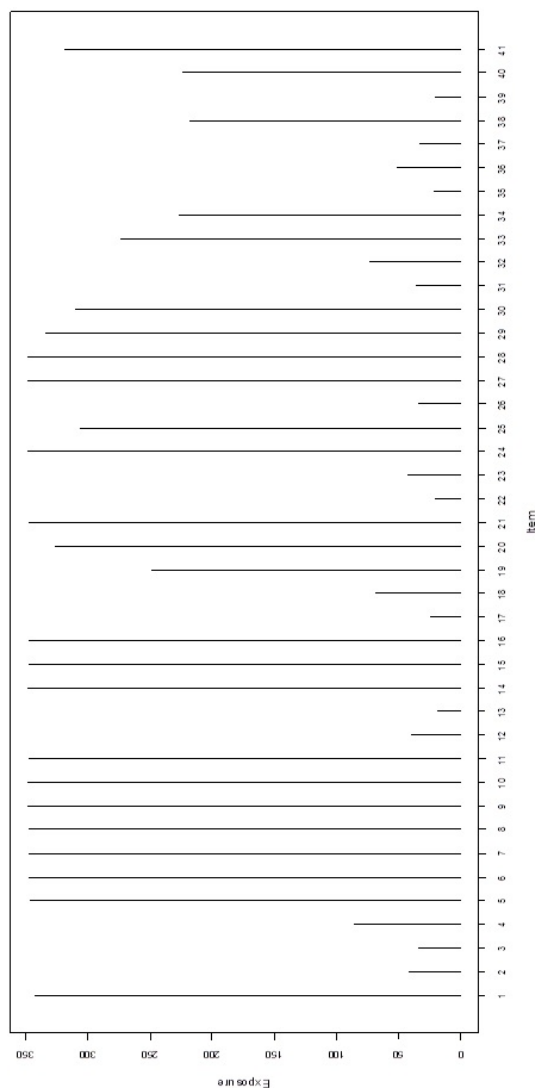


Figure 4.5 – Taux d'exposition des items lors des simulations, avec $SEM < 0.55$.

Validité externe du SQoL-MCAT Les validités convergente et divergente sont mesurées pour la version du SQoL-MCAT sélectionné à l'étape précédente (c'est-à-dire pour un niveau de précision $SEM < 0.55$), les résultats de validité sont présentés dans la table 4.12.

Les scores des dimensions issus du SQoL-MCAT sont plus corrélés avec les scores des dimensions explorant des aspects similaires qu'avec des scores mesurant des aspects plus éloignés. Les scores des dimensions "mentales et psychologiques" du SQoL-CAT (bien-être psychologique et estime de soi) sont fortement corrélés avec la dimension santé mentale et le score composite mental du SF-

	P _{SW}	SE	R _{Fa}	R _{Ft}	RE	P _{HW}	AU	SL
SF-36								
PF	0.38**	0.34**	0.10	0.21*	0.43**	0.51**	0.30**	0.34**
SF	0.50**	0.39**	0.09	0.21*	0.34**	0.38**	0.39**	0.39**
RP	0.37**	0.30**	0.03	0.17	0.32**	0.36**	0.36**	0.32**
RE	0.40**	0.33**	-0.02	0.29**	0.34**	0.32**	0.37**	0.43**
MH	0.65**	0.62**	0.17	0.35**	0.53**	0.53**	0.51**	0.55**
VT	0.48**	0.49**	0.12	0.25**	0.41**	0.60**	0.39**	0.39**
BP	0.33**	0.19*	0.05	0.14	0.22*	0.29**	0.28**	0.18*
GH	0.46**	0.43**	0.08	0.18	0.39**	0.46**	0.37**	0.41**
PCS	0.26**	0.17	0.04	0.15	0.26**	0.38**	0.23*	0.17
MCS	0.65**	0.61**	0.13	0.39**	0.49**	0.52**	0.50**	0.58**
Age	0.05	0.07	-0.19**	-0.12*	0.02	-0.01	0.03	0.02
Genre								
Homme	51.8±16.8	48.7±16.4	54.4±20.7	48.3±18.4	50.8±15.2	48.7±16.9	51.2±17.1	45.9±16.4
Femme	51.9±17.7	48.8±16.8	51.3±23.4	45.7±20.3	51.2±15.1	47.8±17.1	50.9±18.2	49.2±17.4
p-valeur	0.96	0.94	0.24	0.26	0.78	0.65	0.90	0.11
Niveau scolaire								
< 12 years	50.2±18.2	47.8±16.0	50.1±22.8	45.1±19.5	49.7±14.2	46.7±16.5	48.2±18.1	46.4±17.0
≥ 12 years	57.8±12.6	53.8±15.3	55.9±16.7	51.7±16.4	54.0±15.5	54.8±13.5	53.6±18.9	51.4±15.6
p-valeur	0.02	0.11	0.18	0.10	0.23	0.02	0.23	0.19
Statut du patient								
Hospitalisé	48.6±17.2	47.2±15.4	50.9±23.0	43.7±20.3	48.8±14.5	47.0±16.4	45.8±18.4	45.8±16.6
Ambulatoire	57.5±16.5	52.4±16.7	52.0±19.5	51.1±15.7	53.8±14.1	50.6±15.8	55.6±16.6	50.4±16.8
p-valeur	0.01	0.09	0.78	0.05	0.08	0.26	0.01	0.16
Durée de la maladie	0.04	0.09	-0.06	-0.07	0.03	0.01	0.00	0.05
p-valeur	0.59	0.25	0.44	0.40	0.67	0.91	0.99	0.51
PANSS								
Score total	-0.08	-0.08	-0.11	-0.07	-0.10	-0.10	-0.10	-0.09
Score positif	-0.16**	-0.17**	-0.06	-0.18**	-0.18**	-0.11	-0.14*	-0.12*
Score négatif	-0.24**	-0.23**	-0.17**	-0.13*	-0.23**	-0.19**	-0.19**	-0.21**
Score psychopathologie générale	-0.25**	-0.23**	-0.14*	-0.18**	-0.24**	-0.19**	-0.18**	-0.22**
Score CDSS	-0.31**	-0.38**	-0.14	-0.17*	-0.34**	-0.30**	-0.22**	-0.34**

Table 4.12 – Validité externe du SQoL-MCAT. * : p-valeur < 0.05. ** : p-valeur < 0.01

36. Le score de la dimension "physique" du SQoL-MCAT (bien-être physique) est fortement corrélé avec les dimensions fonctionnement physique et vitalité du SF-36. Les scores des dimensions "sociales" du SQoL-MCAT (relations avec la famille, relations avec les amis et vie sentimentale) sont modérément corrélés avec la dimension fonctionnement social du SF-36. L'autonomie et la résilience n'étant pas mesurées par le SF-36, les scores correspondant issus du SQoL-MCAT ne sont pas fortement corrélés, à part avec la dimension santé mentale.

Les corrélations entre les scores issus du SQoL-MCAT et l'âge ne sont pas significatives, à part pour les scores de dimensions "sociales" (relations avec la famille et les amis). Les tests de comparaison par genre et par niveau scolaire ne sont pas non plus significatifs, ce qui confirme les hypothèses attendues. Les scores issus du SQoL-MCAT ne sont pas corrélés avec la durée de la maladie. Concernant les corrélations entre les scores du SQoL-MCAT et les caractéristiques cliniques (scores PANSS et CDSS), on peut en conclure que des hauts niveaux de qualité de vie sont globalement associés à des faibles niveaux de sévérité.

4.2.3.4 Conclusion

Les mesures de la qualité de vie permettent de fournir aux cliniciens des informations concernant la santé générale des patients, mais ces mesures ne sont de nos jours pas encore bien implémentées en psychiatrie. Cependant, il est simple d'obtenir des données de qualité de vie en temps réel en utilisant les nouvelles technologies [81]. Le développement d'algorithmes d'administration adaptative d'items, associé aux nouvelles technologies informatiques, peut permettre d'améliorer l'utilisation des mesures de la qualité de vie dans la prise de décision clinique. Contrairement aux mesures de la qualité de vie classiques (questionnaires "papier-crayon"), les questionnaires adaptatifs multidimensionnels sélectionnent les items qui sont les plus pertinents pour un patient, et les scores de toutes les dimensions considérées sont calculés et mis à jour après chaque réponse à un item. Ces scores sont exprimés sur une métrique standardisée, ce qui permet de comparer des patients qui n'ont pas reçu le même sous-ensemble d'items. Le SQoL-MCAT est le premier questionnaire adaptatif multidimensionnel spécifique aux patients atteints de schizophrénie, et il permet d'administrer moins d'items (en moins de temps) que la version séquentielle initiale du questionnaire sur lequel il est basé.

Le SQoL-MCAT possède des propriétés de précision satisfaisantes. Tous les scores des dimensions ont des niveaux de corrélation élevés (coefficients supérieurs à 0.9) avec les scores issus de la banque d'items (qui contient les 41 items du SQoL), et les valeurs de RMSE obtenues sont satisfaisantes (c'est-à-dire < 0.3), excepté pour une dimension (résilience). De plus l'analyse de la validité externe du SQoL-MCAT est cohérente au regard de la validité du questionnaire initial.

Dans la méthode utilisée, nous avons choisi d'utiliser la méthode d'estima-

tion bayésienne MAP pour obtenir des scores initiaux pour les huit traits latents considérés, pour recalculer le score après chaque réponse à un item dans le questionnaire adaptatif et pour l'estimation finale des scores. Dans la littérature, il existe deux approches différentes pour estimer des traits latents : estimation par maximum de vraisemblance et estimation bayésienne, cette dernière comprenant les méthodes maximum a posteriori (MAP) et espéré a priori (EAP). Bien que ce choix méthodologique puisse paraître discutable dans le sens où les scores MAP peuvent faire apparaître certains biais [156], une étude précédente [176] montre que les scores MAP apportent plus de précision que des scores obtenus par maximum de vraisemblance et obtiennent des performances similaires (voire supérieures) aux scores EAP. De plus, au regard de récents travaux [40], l'utilisation de scores EAP pour des modèles de réponse à l'item considérant plus de trois dimensions est fortement déconseillée car cela peut résulter en une estimation plus lente et moins précise des scores. Etant donnée la structure multidimensionnelle (avec 8 dimensions) du SQoL, nous avons décidé d'utiliser l'estimation MAP dans ce cas.

Cette section a permis de démontrer qu'un questionnaire adaptatif peut être un outil de mesure efficace comparé aux mesures classiques, impliquant une augmentation de la précision, et en évitant les questions les moins pertinentes. Cependant, un important fondement du développement de questionnaires adaptatifs est la calibration de banques d'items, qui contiennent un grand nombre d'items mesurant un même trait latent. Le développement de banques d'items est un processus important avant de pouvoir proposer une nouvelle mesure adaptative. Cependant, cela nécessite beaucoup de ressource (matérielles et temporelles), et certains problèmes sont encore aujourd'hui non résolus : est-il possible d'associer des items basés sur des domaines théoriques et conceptuels multiples dans une même banque d'items ? Peut-on associer des questionnaires génériques et spécifiques ? Peut-on associer des items prenant en compte différents points de vue (patient, aidant, soignant) dans une même banque ?

La nature multidimensionnelle de la qualité de vie requiert le développement de plusieurs banques d'items unidimensionnelles, qui peuvent ensuite être liées par la calibration d'un modèle de réponse multidimensionnel. Dans le développement du SQoL-MCAT nous avons fait le choix de développer notre mesure adaptative à partir des 41 items du questionnaire SQoL, étant données les barrières logistiques que représente le développement de banques d'items unidimensionnelles. Bien que le nombre d'items à notre disposition soit relativement faible comparé à d'autres études (américaines notamment), nous montrons que le développement de questionnaires adaptatifs multidimensionnels basés sur des questionnaires classiques peut être une alternative astucieuse pour pallier manque de ressources financières et temporelles.

Le SQoL-MCAT peut s'adapter aux caractéristiques individuelles d'un patient et est significativement plus court que la version initiale du questionnaire sur lequel il est basé. Il permet d'améliorer la mesure de la qualité de vie dans la pratique

clinique, puisqu'il réduit le fardeau ressenti par les patients lors du remplissage des questions et permet aux professionnels de la santé d'obtenir des données de qualité de vie dans un intervalle de temps raisonnable.

4.3 Questionnaires adaptatifs basés sur les arbres de décision binaires

Nous proposons dans cette section une alternative aux approches basées sur la théorie de réponse à l'item pour le développement de questionnaires adaptatifs informatisés. Nous articulerons la suite de cette section autour des applications aux mesures de la qualité de vie liée à la santé. Cette approche alternative est basée sur les arbres de décision binaires, dont les capacités de prédiction peuvent être utilisées pour développer un algorithme d'administration adaptative d'items. Une comparaison avec l'approche classique basée sur la théorie de réponse à l'item est entreprise, en fournissant des indicateurs de précision des mesures obtenues.

4.3.1 Introduction

De nos jours, les ressources temporelles et matérielles nécessaires pour la récolte et l'analyse des données de qualité de vie issues de questionnaires sont de réelles contraintes pour les cliniciens et les fournisseurs de santé. Les questionnaires utilisés dans la pratique clinique doivent être le moins restrictif et pénible possible pour les patients, étant données les difficultés rencontrées chez certaines populations cliniques. Il est donc indispensable de fournir aux cliniciens des formes courtes de questionnaires pour la mesure de la qualité de vie. Cependant, les formes courtes des questionnaires utilisées contiennent un nombre fixe d'items, et cela peut présenter certaines limites, notamment des risques de perte d'information, ou une diminution de la précision de la mesure. Les items disponibles peuvent être inadaptés à certains patients, ces derniers pouvant ressentir un manque d'intérêt et décider d'arrêter le remplissage du questionnaire. Les méthodes basées sur la théorie de réponse à l'item peuvent cependant être considérées pour résoudre ces problèmes (voir la section 4.2 par exemple). Les modèles paramétriques de la théorie de réponse à l'item sont communément utilisés pour le développement de banques d'items unidimensionnelles, et sont la base du développement de questionnaire adaptatifs [59, 170].

Un questionnaire adaptatif permet d'administrer les items qui apportent le plus d'information pour un individu donné, en améliorant à la fois le temps de remplissage et la précision de la mesure [170, 143, 89]. Cependant, dans la pratique, il est possible que certaines mesures ou certaines échelles ne soient pas adaptées pour une analyse de données basée sur la théorie de réponse à l'item. Le développement de questionnaires adaptatifs basés sur la théorie de réponse à l'item (voir section 4.2) est très coûteux, étant données les ressources nécessaires pour la calibration de banques d'items unidimensionnelles, même si l'étape de calibration peut être simplifiée en considérant un faible nombre d'items et ainsi éviter des étapes trop longues et fastidieuses de sélection d'items, ou encore en considérant

les modèles de réponse multidimensionnels, qui permettent de s'affranchir des hypothèses fondamentales de la théorie de réponse à l'item. De plus, au regard de l'interprétation des résultats issus de la théorie de réponse à l'item, les algorithmes de sélection d'items disponibles dans la littérature sont assez complexes et ne permettent pas au clinicien d'avoir une vision d'ensemble des étapes du questionnaire adaptatif.

Une manière naturelle et intuitive de développer un questionnaire adaptatif sur la base d'un ensemble d'items serait d'utiliser les arbres de décision binaires. Les arbres de décision binaires, également appelés arbres de régression et de classification (notés CART) [29]. CART est une méthode de partitionnement récursive des données qui permet de prédire la valeur d'une variable à expliquer Y , qui peut être continue (dans le cas de la régression) ou catégorielle (dans le cas de la classification). Les arbres de décision en régression sont utilisés en pratique pour obtenir une prédiction de Y , notée $\hat{Y}$, à partir d'une suite de divisions binaires basées sur les p variables explicatives $X_1, \dots, X_p$. Cette méthode est très simple à implémenter, elle ne considère aucune hypothèse, elle est non-paramétrique, et l'arbre résultant est simple à interpréter.

Yan [173, 174, 175] a déjà proposé d'appliquer les arbres de décision en régression pour le développement de questionnaires adaptatifs. Dans ces travaux, il montre que l'approche non-paramétrique basée sur les arbres de décision permet d'obtenir de meilleures performances qu'en utilisant l'approche "classique", basée sur la théorie de réponse à l'item, notamment lorsque les hypothèses fondamentales ne peuvent pas être vérifiées de manière satisfaisante. Certains auteurs ont également étudié la performance de questionnaires adaptatifs basés sur des arbres de décision, en se comparant à des questionnaires adaptatifs basés sur la théorie de réponse à l'item [162, 145]. A notre connaissance, la seule application de CART en médecine est le développement d'un outil diagnostic adaptatif pour les patients atteints de dépression [72]. L'auteur de cette étude a d'ailleurs publié récemment un état de l'art des différentes méthodes permettant de développer des questionnaires et outils de diagnostic adaptatifs en santé mentale [71]. Aucune étude n'a encore été réalisée en santé publique pour comparer l'approche basée sur la théorie de réponse à l'item avec l'approche basée sur les arbres de décision.

Le but de cette section est donc de proposer une alternative à la théorie de réponse à l'item pour le développement de questionnaires adaptatifs mesurant la qualité de vie. Cette approche repose sur CART, dont les capacités de prédiction sont utilisés pour définir un nouvel algorithme d'administration adaptative d'items. Une comparaison à l'approche "classique", basée sur la théorie de réponse à l'item est proposée, au regard des différentes propriétés de précision obtenues. Cette application est concentrée sur les 22 items de la banque d'items développée au laboratoire de santé publique de Marseille, mesurant la qualité de vie liée à la santé mentale. La section 4.3.2 présente la méthodologie employée, c'est-à-dire la conception de l'étude et la population, le contenu de la banque

d'items et le développement des questionnaires adaptatifs. La section 4.3.3 présente les principaux résultats de cette étude.

4.3.2 Méthodes

La méthodologie présentée dans cette section consiste à développer un questionnaire adaptatif informatisé à partir d'un ensemble d'items mesurant un trait latent commun (la qualité de vie liée à la santé mentale), en considérant deux approches différentes : la théorie de réponse à l'item et les arbres de décision. La théorie de réponse à l'item est actuellement l'approche référence pour le développement de questionnaires adaptatifs [105], mais elle peut ne pas être adaptée à toutes les situations dans la pratique clinique.

4.3.2.1 Approche basée sur les arbres de décision binaires

Nous nous concentrons ici sur l'approche basée sur les arbres de décision binaires en régression. Un arbre de régression est construit en utilisant un processus itératif, dans lequel des règles de divisions binaires sont définies à partir de valeurs seuils qui peuvent être prises par les variables d'un jeu de données. Pour chaque variable $X_{.j}$ d'un jeu de données, une règle de division binaire à la forme suivante :

$$x_{.j} < a, \quad (4.5)$$

où $a \in \mathbb{R}$ est une valeur seuil prise par la variable j , permettant de diviser le jeu de données en deux sous-ensembles, le jeu de données étant assigné au noeud racine de l'arbre, noté t_0 . Les deux sous-groupes d'observations ainsi obtenus sont alors assignés aux deux noeuds enfants de la racine, notés t_g et t_d . Parmi toutes les divisions possibles d'un noeud (c'est-à-dire pour $j \in \{1, \dots, p\}$ et pour tout $a \in \mathbb{R}$, la meilleure division est celle qui minimise la somme de l'hétérogénéité intra-classe au sein des noeuds enfants. Une fois que la meilleure division est définie et que le noeud est séparé en deux noeuds, le même processus est appliqué aux noeuds enfants ainsi obtenu. Ce processus est répété jusqu'à l'obtention de noeuds terminaux (appelés feuilles) d'une certaine taille (c'est-à-dire contenant un minimum d'observations) ou si un critère de minimisation de l'hétérogénéité intra-classe est vérifié. Chaque feuille de l'arbre se voit attribuer une valeur (la moyenne, ou encore le mode des observations), qui peut être utilisée pour attribuer de nouveaux individus aux feuilles, et prédire ainsi la valeur de la variable à expliquer pour ces nouveaux individus. Une étape d'élagage de l'arbre, permettant de regrouper certaines paires de feuilles, peut être ensuite envisagée par validation croisée en utilisant un critère de dissimilarité.

Un arbre de décision peut être représenté graphiquement, il est facilement interprétable étant donné sa structure binaire. Dans le contexte du développement

de questionnaires adaptatifs, chaque variable prise en compte dans la division d'un noeud de l'arbre peut être vue comme un item issu de la banque d'items, par exemple, la variable définissant la division du noeud racine de l'arbre correspond au premier item administré au patient. En fonction de la réponse fournie par le patient à un item, et étant donné la valeur du seuil définissant la division du noeud, l'item suivant est sélectionné, correspondant à la variable prise en compte dans la division du noeud enfant de gauche si la division binaire est satisfaite, du noeud enfant de droite sinon. L'ensemble d'items administré correspond donc aux noeuds reliant la feuille dans laquelle le patient est assigné à la racine de l'arbre. Un score peut ainsi être prédit pour ce patient, comme étant la moyenne (des score) des observations contenues dans la feuille de l'arbre. La figure 4.6 donne un exemple de structure d'arbre obtenue à partir de la banque d'items utilisée dans la suite de cette section. Dans cet exemple, une contrainte au niveau de la taille des noeuds est fixée, ces derniers ne pouvant pas être divisés s'ils contiennent moins de 100 observations. La méthode CART qui est utilisée ici comprend une multitude de paramètres et de détails techniques qui sont décrits dans les travaux de Léo Breiman [29].

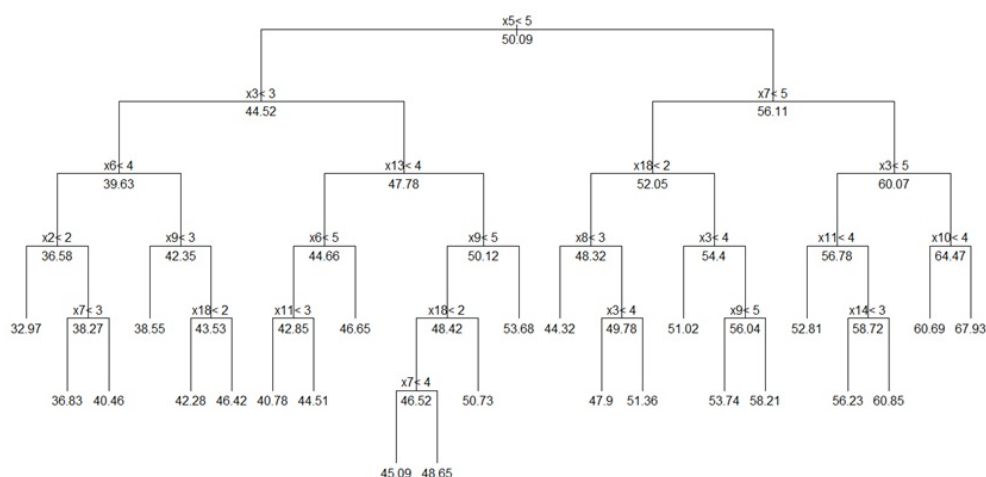


Figure 4.6 – Structure de l'arbre obtenu avec la méthode CART, avec $minsplitleft = 100$.

4.3.2.2 Conception de l'étude et population

Nous utilisons les mêmes données utilisées pour la validation de la version internationale du questionnaire MusiQoL et le développement de la banque d'items issue du questionnaire MusiQoL mesurant la qualité de vie liée à la santé men-

tale. Une description du recrutement des patients ainsi que les critères d'inclusion de l'étude sont fournis dans les sections 4.1.2.1 et 4.1.2.2 respectivement.

4.3.2.3 La banque d'items mesurant la qualité de vie liée à la santé mentale

Les items sur lesquels sera basé le questionnaire adaptatif présenté dans la suite de cette section sont ceux issus de la banque d'items mesurant la qualité de vie liée à la santé mentale. La banque d'items contient 22 items qui sont initialement issus d'un questionnaire générique (le SF-36) et d'un questionnaire spécifique (le MusiQoL), son processus de développement est décrit en détails dans la section 4.1.2.

4.3.2.4 Développement du questionnaire adaptatif

Approche "classique" : théorie de réponse à l'item Dans un premier temps, nous utilisons l'approche "classique" basée sur la théorie de réponse à l'item. Un questionnaire adaptatif développé en utilisant ce type de modèles est basé sur une banque d'items calibrée (c'est-à-dire dont les paramètres d'items ont été estimés grâce à un modèle paramétrique). La première étape d'un questionnaire adaptatif consiste à assigner un score initial $\hat{\theta}$ à l'individu (dans la suite on fixera $\hat{\theta} = 0$). Le questionnaire sélectionne ensuite l'item dont l'information est maximale pour ce niveau de trait latent (pour cela on utilise les fonctions d'information des items). L'estimateur du trait latent $\hat{\theta}$ est alors recalculé en prenant en compte la réponse de l'individu à l'item et les paramètres de l'item en question, en utilisant une estimation bayésienne MAP. Un nouvel item est ensuite sélectionné de la même façon que l'item précédent. Ce processus peut être stoppé lorsqu'un nombre d'items fixé a été administré au patient, ou lorsqu'un certain niveau de précision est atteint, on utilise ici l'erreur standard de mesure (SEM). Dans nos simulations nous utilisons les mêmes valeurs seuils de SEM définies précédemment (0.33, 0.44 et 0.55, voir section 4.2.3). L'estimateur final du trait latent $\hat{\theta}$ est alors attribué au patient, il son score de qualité de vie liée à la santé mentale. Nous appliquons cette méthode en utilisant package R *mirtCAT*.

Approche alternative : arbres de régression Nous utilisons maintenant l'approche basée sur les arbres de régression, en utilisant la méthode CART. La variable à expliquer Y représente ici le score issu de la banque d'items mesurant la qualité de vie liée à la santé mentale (voir section 4.1). Les variables ordinales explicatives correspondent aux 22 items contenus dans cette banque. Pour construire l'arbre de régression, le seul paramètre que nous fixons est appelé *minsplit*, qui permet de contrôler le nombre minimum d'observations que doit contenir un noeud pour pouvoir être divisé. Quatre valeurs de *minsplit* sont testés (100, 50, 10 et 5), des valeurs élevées de *minsplit* impliquant des arbres moins

profonds. Pour prédire le score d'un patient, on administre les items correspondant les variables impliquées dans chaque noeud de l'arbre, de la racine, jusqu'à la feuille correspondante (différente selon les patients). Par exemple dans la figure 4.6, si un patient choisit la 5ème modalité pour les items 5, 7, 3 et 10, alors il est assigné à la feuille se situant à l'extrême droite de la figure, et son score de qualité de vie liée à la santé mentale est prédit et vaut 67.93. Le score final de l'individu (une fois que tous les items sont administrés) correspond à la valeur prédite de Y , notée $\hat{Y}$. Nous appliquons cette méthode en utilisant le package R *rpart*.

Simulations de questionnaires adaptatifs Finalement, une fois que nos deux algorithmes d'administration adaptative sont définis, nous effectuons des simulations de questionnaires adaptatifs pour chacune des approches. Les réponses des patients utilisés pour calibrer la banque d'items sont maintenant utilisées pour simuler les deux processus adaptatifs. Pour chaque cas, on calcule le score moyen obtenu (et son écart-type), les scores minimum et maximum, ainsi que le nombre moyen (et son écart-type) d'items administré par l'algorithme. Afin de quantifier la performance et comparer les deux approches, on calcule également les indices de précision utilisés précédemment (c'est à dire les corrélations inter-scores et le RMSE, voir 4.2.3).

4.3.3 Résultats

La figure 4.7 illustre le questionnaire adaptatif basé sur la théorie de réponse à l'item, utilisé dans cette étude. La table 4.13. Le meilleur scénario de questionnaire adaptatif obtenu avec la théorie de réponse à l'item est celui dont le niveau de précision est le plus faible. Avec une valeur de $SEM < 0.33$, cela revient à dire que le score obtenu est fiable à plus de 90%. Le score issu de ce questionnaire adaptatif est fortement corrélé avec le score issu de la banque d'items ($R = 0.96$). De plus, la valeur du RMSE est de 0.22, ce qui se traduit par un excellent niveau de précision de la mesure. En plus de ces résultats très satisfaisants, ce questionnaire administre moins de la moitié des items initiaux (9 en moyenne) pour obtenir ce score. Les deux autres questionnaires de ce type ($SEM < 0.44$ et $SEM < 0.55$) sont un peu moins satisfaisants. Le premier ($SEM < 0.44$) obtient une précision acceptable au regard du coefficient de corrélation ($R = 0.90$) mais le RMSE est en-dessous du seuil attendu. Le second ($SEM < 0.55$), malgré son très faible nombre d'items administré (2 en moyenne) ne parvient pas à obtenir des résultats de précision satisfaisants.

Concernant les questionnaires adaptatifs basés sur les arbres de régression, on peut voir que pour une valeur de $minsplit = 100$, malgré un faible nombre d'items administrés, les résultats au regard de la précision ne sont pas satisfaisants. Le deuxième questionnaire adaptatif ($minsize = 50$) montre des résultats de précision satisfaisants ($R = 0.93$, $RMSE = 0.28$), mais ne dépasse

pas la performance du questionnaire adaptatif basé sur la théorie de réponse à l'item ($SEM < 0.33$). En revanche, les deux derniers scénarios ($minsplit = 10$ et $minsplit = 5$) nous permettent d'observer que ces questionnaires adaptatifs ont une précision satisfaisante. Le premier ($minsplit = 10$) administre autant d'items (9 en moyenne) que le questionnaire adaptatif basé sur la théorie de réponse à l'item ($SEM < 0.33$). Il est cependant plus performant en termes de précision ($R = 0.98$ et $RMSE = 0.16$). Nous décidons de ne pas nous intéresser au questionnaire adaptatif ($minsplit = 5$), étant donné qu'il administre plus de la moitié des items initiaux (9 en moyenne, avec un écart-type élevé) pour un gain en précision négligeable comparé au précédent.

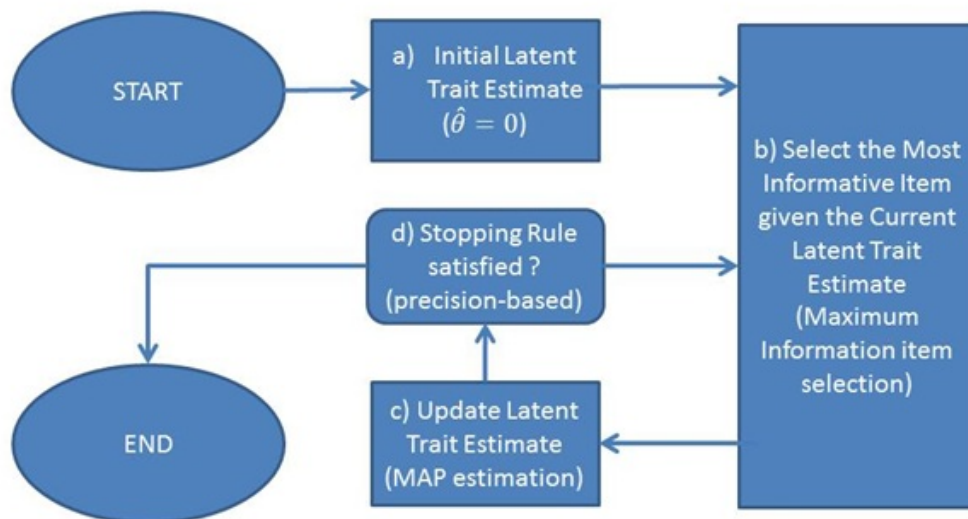


Figure 4.7 – Questionnaire adaptatif basé sur la théorie de réponse à l'item.

4.3.4 Conclusion

Ces résultats préliminaires confirment l'intérêt d'utiliser des approches alternatives pour le développement de questionnaires adaptatifs, alors que la méthode la plus utilisée de nos jours repose sur la théorie de réponse à l'item, qui implique des modèles complexes et une interprétabilité difficile des résultats. L'approche

	Théorie de réponse à l'item			Arbres de régression			
	$SEM < 0.33$	$SEM < 0.44$	$SEM < 0.55$	$ms = 100$	$ms = 50$	$ms = 10$	$ms = 5$
Score moyen	50	49.77	49.68	50.09	50.09	50.09	50.09
Ecart-type score	8.64	8.37	7.86	8.26	8.52	8.94	9.05
Score minimum	15.77	15.77	27.19	32.97	32.97	23.4	18.46
Score maximum	79.56	79.56	73.62	67.93	73.23	78.38	78.83
Nombre d'items	9 (3)	5 (2)	2 (1)	5 (1)	6 (1)	9 (2)	14 (11)
Précision	0.96	0.9	0.83	0.91	0.93	0.98	0.99
RMSE	0.22	0.35	0.45	0.33	0.28	0.16	0.09

Table 4.13 – Comparaisons des deux approches : théorie de réponse à l'item et arbres de régression. ms désigne le paramètre *minsize*.

par les arbres de régression nous permet de construire simplement un algorithme d'administration adaptative, qui peut être facilement implémenté dans la pratique clinique. En effet, la succession des divisions binaires de l'arbre revient à évaluer après chaque réponse d'un individu si cette réponse est en-dessous ou au-dessus d'une certaine valeur seuil, puis d'assigner l'individu au sous-groupe correspondant. Le score final est obtenu en considérant le score moyen des individus dans ce sous-groupe.

En général, les arbres de décision présentent de nombreux avantages. Ils sont simples à interpréter, étant donnée leur structure simple. Ils peuvent être représentés graphiquement, ce qui permet, dans le cadre du questionnaire adaptatif, d'énumérer l'ensemble des chemins possibles (c'est-à-dire l'ensemble des sous-groupes d'items pouvant être administrés) d'un patient. Ils peuvent également être utilisés avec des variables explicatives qualitatives, sans utiliser de transformation particulière, cela reviendrait donc à créer un questionnaire adaptatif pour items nominaux. Cependant, un des défauts majeurs des arbres de régression vient de leur manque de robustesse. En effet, les arbres de régression sont connus pour être instables lorsque l'on effectue de petites perturbations sur les données. Cependant, il est possible de contrôler ce défaut, en mettant en place des méthodes d'aggrégation d'arbres (les plus connues étant le "bagging", les forêts aléatoires, et le "boosting"), permettant de construire des modèles de prédiction plus performants, mais ceci implique forcément une perte d'interprétabilité.

Des travaux sont actuellement en cours pour prendre en compte ces défauts et tester plusieurs méthodes d'aggrégation pour améliorer encore les résultats obtenus dans cette étude, une analyse de la stabilité du score obtenu avec un questionnaire adaptatif basé sur un arbre de décision reste à entreprendre.

L'utilisation des arbres de décision binaires est une manière naturelle et intuitive qui peut être envisagée pour le développement de questionnaires adaptatifs. Nous avons présenté un nouvel algorithme d'administration adaptative d'items, basé sur la méthode CART. Ce dernier surpasse l'approche "classique" basée sur l'IRT lorsque l'on contrôle un des paramètres de cette méthode. Les simulations de questionnaires adaptatifs, l'analyse de leur précision, et les comparaisons des deux approches nous ont permis de définir un algorithme optimal pour mesurer la qualité de vie liée à la santé mentale des patients atteints de sclérose en plaques.

Conclusion et perspectives

Dans cette thèse, nous nous sommes intéressés aux méthodes d'apprentissage non supervisé, souvent utilisées pour résoudre des problèmes de sélection de variables. Nous avons également présenté plusieurs contributions statistiques appliquées à la santé publique, plus particulièrement dans la recherche sur les questionnaires adaptatifs.

Dans un premier temps, nous avons étudié une méthode récente de classification non supervisée basée sur les arbres de décision binaires. Quelques applications sur des données de qualité de vie ont également été entreprises et l'une d'elles a été décrite en détails dans ce chapitre. Ces résultats empiriques nous permettent de confirmer l'intérêt de cette méthode pour obtenir une partition d'un échantillon, de manière à obtenir des sous-groupes d'individus tout pouvant interpréter la partition obtenue. Nous avons proposé une extension de cette méthode au cas des données nominales. Par le biais de modèles de simulation de données, nous comparons ces méthodes avec d'autres méthodes de classification non supervisée. Cette extension démontre des performances satisfaisantes et cela nous a encouragé pour mettre en place d'autres applications de cette méthode dans d'autres champs de recherche clinique (en imagerie cérébrale ou encore en cancérologie).

Ensuite, nous avons utilisé la méthode présentée précédemment pour définir un score d'importance des variables. Nous avons montré qu'il est possible de mesurer l'importance des variables d'un jeu de données de façon non-supervisée, de la même manière que dans les arbres de décision supervisés. Cette méthode consiste à calculer l'apport de chaque variable vis-à-vis d'un certain critère d'hétérogénéité. Cela nous a permis en pratique d'obtenir des classements pour les variables d'un jeu de données et ainsi identifier les variables les plus pertinentes. Par le biais de simulations, nous montrons que cette approche obtient d'excellentes performances pour détecter les variables importantes dans un jeu de données, et inversement détecter les variables non pertinentes (ce que l'on appelle le bruit). Cette nouvelle méthode de mesure de l'importance des variables est adaptée aux données continues et qualitatives, et permet d'entreprendre des démarches de sélection de variables.

Enfin, les différentes applications présentées dans la thèse permettent de mettre en évidence l'importance des mesures dites "modernes" de la qualité de vie. En effet, par le biais de l'utilisation de banques d'items, il est possible de développer des questionnaires adaptatifs, ayant pour but d'alléger le nombre d'items administré actuellement par les questionnaires traditionnels dits "papier-crayon". Une banque d'items mesurant la qualité de vie liée à la santé mentale a donc été développée en respectant les standards imposés par la théorie de réponse à l'item unidimensionnelle. De plus, nous nous sommes également intéressés au cas des

questionnaires adaptatifs multidimensionnels, permettant de mesurer plusieurs traits latents simultanément, et ainsi se conformer à la structure multidimensionnelle de la qualité de vie. Deux nouveaux questionnaires adaptatifs multidimensionnels, mesurant la qualité de vie spécifique des patients atteints de sclérose en plaque (le MusiQoL-MCAT) et de schizophrénie (le SQoL-MCAT), ont été développés en collaboration avec le laboratoire de santé publique de Marseille. Enfin, nous avons proposé une alternative non-paramétrique pour le développement de questionnaires adaptatifs. Cette nouvelle méthode, basée sur les arbres de décision binaires, permet de s'affranchir des méthodes paramétriques de la théorie de réponse à l'item. Des résultats préliminaires issus de simulations nous permettent de confirmer que cette méthode peut être une sérieuse concurrente de la méthode standard.

Perspectives

Concernant la méthode de classification non supervisée et son extension aux données nominales (la méthode CUBT), des travaux supplémentaires sont à effectuer pour définir un critère de division binaires adaptés au cas de données mélangées. Ce critère mixte devra s'adapter au cas où le jeu de données contient à la fois des variables continues et des variables qualitatives. Il sera aussi envisagé d'adapter CUBT au cas des données longitudinales. En effet il pourrait être intéressant d'obtenir une partition d'un ensemble d'observations représentant des séries temporelles. Pour cela, une nouvelle mesure de dissimilarité (de type "dynamic time warping" par exemple), adaptée aux données longitudinales doit être implémentée au niveau des divisions binaires et des étapes d'élagage de l'arbre.

Au sujet de la sélection de variables, maintenant que le score d'importance basé sur CUBT a été défini, jugé satisfaisant vis-à-vis des résultats issus des simulations, il est envisagé de mettre en place un nouvel algorithme de sélection de variables, basés sur ce score. Dans le cadre de l'apprentissage non supervisé, la sélection de variables est beaucoup moins aisée qu'en apprentissage supervisé. Notre future méthode sera comparée à des méthodes issues de l'état de l'art. De plus nous chercherons à effectivement améliorer la qualité d'un modèle de prédiction en ne sélectionnant qu'un sous-ensemble de variables les plus importantes. L'utilisation des divisions concurrentes, utilisés pour définir le score d'importance des variables, peut aussi être envisagée pour développer une nouvelle méthode de traitement des données manquantes.

Enfin, nous envisageons plusieurs perspectives concernant les applications des méthodes statistiques utilisées, sur des données de qualité de vie. En effet, les résultats préliminaires relatifs au développement de questionnaires adaptatifs basé sur les arbres de décision binaires méritent quelques approfondissements. Cette nouvelle méthode a déjà été proposée dans des études sur d'autres domaines

(en sciences éducationnelles par exemple) mais à notre connaissance, aucune ne s'est concentrée sur le principal défaut des arbres de décision, à savoir leur instabilité. En effet la structure d'un arbre peut être grandement modifiée lorsque l'on modifie quelque peu l'échantillon d'apprentissage. Cette contrainte est à prendre en compte avant de pouvoir proposer un questionnaire adaptatif utilisable en pratique clinique. Pour cela, nous chercherons à évaluer la stabilité des scores obtenus par ce type de questionnaires adaptatifs. Afin de contrôler la potentielle instabilité, nous envisagerons des méthodes d'agrégation, permettant de stabiliser les estimateurs obtenus et améliorer leur précision. Un travail plus approfondi, en testant différentes méthodes d'agrégation ("bagging", forêts aléatoires, "boosting") est à envisager.

En conclusion, cette thèse a permis de proposer de nouvelles méthodes pour l'analyse des données de qualité de vie et le développement de questionnaires adaptatifs, tout en se penchant sur des aspects plus techniques des méthodes statistiques utilisées dans ce cadre.

Bibliographie

- [1] N. K. AARONSON, S. AHMEDZAI, B. BERGMAN et al. « The European Organization for Research and Treatment of Cancer QLQ-C30 : a quality-of-life instrument for use in international clinical trials in oncology ». In : *Journal of the National Cancer Institute* 85.5 (3 mar. 1993), p. 365–376. ISSN : 0027-8874.
- [2] D. ADDINGTON, J. ADDINGTON et B. SCHISSEL. « A depression rating scale for schizophrenics ». In : *Schizophrenia Research* 3.4 (août 1990), p. 247–251. ISSN : 0920-9964.
- [3] Food {and} Drug ADMINISTRATION. « Guidance for industry : patient-reported outcome measures : use in medical product development to support labeling claims. » In : *Health and Quality of Life Outcomes* 4 (11 oct. 2006). 00000, p. 79. ISSN : 1477-7525. DOI : 10.1186/1477-7525-4-79. URL : <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1629006/> (visité le 21/05/2013).
- [4] European Medicines AGENCY. « Reflection paper on the regulatory guidance for the use of HRQoL measures in the evaluation of medicinal products ». In : <http://www.ema.europa.eu/ema/pages> (2004).
- [5] Alan AGRESTI. *Categorical Data Analysis*. Google-Books-ID : hpEzw4T0sPUC. John Wiley & Sons, 14 avr. 2003. 736 p. ISBN : 978-0-471-45876-0.
- [6] Seung C. AHN et Alex R. HORENSTEIN. « Eigenvalue Ratio Test for the Number of Factors ». In : *Econometrica* 81.3 (2013). 00154, p. 1203–1227. ISSN : 1468-0262. DOI : 10.3982/ECTA8968. URL : <http://onlinelibrary.wiley.com/doi/10.3982/ECTA8968/abstract> (visité le 28/11/2013).
- [7] Hirotogu AKAIKE. « Information Theory and an Extension of the Maximum Likelihood Principle ». In : *Selected Papers of Hirotugu Akaike*. Sous la dir. d'Emanuel PARZEN, Kunio TANABE et Genshiro KITAGAWA. Springer Series in Statistics. 00000. Springer New York, 1^{er} jan. 1998, p. 199–213. ISBN : 978-1-4612-7248-9 978-1-4612-1694-0. URL : http://link.springer.com/chapter/10.1007/978-1-4612-1694-0_15 (visité le 28/11/2013).
- [8] Milena D ANATCHKOVA, Renee N SARIS-BAGLAMA, Mark KOSINSKI et al. « Development and preliminary testing of a computerized adaptive assessment of chronic pain ». In : *The journal of pain : official journal of the American Pain Society* 10.9 (sept. 2009), p. 932–943. ISSN : 1528-8447. DOI : 10.1016/j.jpain.2009.03.007.

- [9] John ANDERSEN. « Asymptotic Properties of Conditional Maximum Likelihood Estimators ». In : *Journal of the Royal Statistical Society B.32* (1970), p. 283–301.
- [10] David ANDRICH. « A rating formulation for ordered response categories ». In : *Psychometrika* 43.4 (déc. 1978), p. 561–573. ISSN : 0033-3123, 1860-0980. DOI : 10.1007/BF02293814. URL : <http://link.springer.com/article/10.1007%2FBF02293814?LI=true> (visité le 09/11/2012).
- [11] American Psychiatric ASSOCIATION. « Diagnostic and statistical manual of mental disorders (4th ed., text rev.) » 00000. Washington, DC, USA, 2000.
- [12] P. AUQUIER, M. C. SIMEONI, C. SAPIN et al. « Development and validation of a patient-based health-related quality of life questionnaire in schizophrenia : the S-QoL ». In : *Schizophrenia Research* 63.1 (1^{er} sept. 2003), p. 137–149. ISSN : 0920-9964.
- [13] A. George AWAD et Lakshmi N. P. VORUGANTI. « Measuring quality of life in patients with schizophrenia : an update ». In : *PharmacoEconomics* 30.3 (mar. 2012). 00000, p. 183–195. ISSN : 1179-2027. DOI : 10.2165/11594470-000000000-00000.
- [14] Frank B. BAKER et Seock-Ho KIM. *Item Response Theory : Parameter Estimation Techniques, Second Edition*. 2^e éd. New York : CRC Press, 20 juil. 2004. 528 p. ISBN : 978-0-8247-5825-7.
- [15] K BAUMSTARCK, J PELLETIER, H BUTZKUEVEN et al. « Health-related quality of life as an independent predictor of long-term disability for patients with relapsing-remitting multiple sclerosis ». In : *European journal of neurology : the official journal of the European Federation of Neurological Societies* 20.6 (juin 2013). 00015, 907–914, e78–79. ISSN : 1468-1331. DOI : 10.1111/ene.12087.
- [16] Karine BAUMSTARCK, Laurent BOYER, Mohamed BOUCEKINE et al. « Measuring the quality of life in patients with multiple sclerosis in clinical practice : a necessary challenge ». In : *Multiple sclerosis international* 2013 (2013), p. 524894. ISSN : 2090-2654. DOI : 10.1155/2013/524894.
- [17] K BAUMSTARCK-BARRAU, J PELLETIER, M-C SIMEONI et al. « [French validation of the Multiple Sclerosis International Quality of Life Questionnaire] ». In : *Revue neurologique* 167.6 (juil. 2011), p. 511–521. ISSN : 0035-3787. DOI : 10.1016/j.neurol.2010.10.008.
- [18] Heather BECKER, Alexa STUIFBERGEN, HwaYoung LEE et al. « Reliability and Validity of PROMIS Cognitive Abilities and Cognitive Concerns Scales Among People with Multiple Sclerosis ». In : *International Journal of MS Care* 16.1 (2014). 00006, p. 1–8. ISSN : 1537-2073. DOI :

- 10.7224/1537-2073.2012-047. URL : <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3967698/> (visité le 02/07/2014).
- [19] Mikhail BELKIN et Partha NIYOGI. « Laplacian Eigenmaps for Dimensionality Reduction and Data Representation ». In : *Neural Comput.* 15.6 (juin 2003). 04884, p. 1373–1396. ISSN : 0899-7667. DOI : 10.1162/089976603321780317. URL : <http://dx.doi.org/10.1162/089976603321780317> (visité le 29/08/2016).
- [20] Adi BEN-ISRAEL et Cem IYIGUN. « Probabilistic D-Clustering ». In : *Journal of Classification* 25.1 (18 juin 2008). 00065, p. 5. ISSN : 0176-4268, 1432-1343. DOI : 10.1007/s00357-008-9002-z. URL : <http://link.springer.com/article/10.1007/s00357-008-9002-z> (visité le 15/08/2016).
- [21] Jakob B. BJORNER, Mark KOSINSKI et John E. Ware JR. « Calibration of an item pool for assessing the burden of headaches : An application of item response theory to the Headache Impact Test (HIT™) ». In : *Quality of Life Research* 12.8 (1^{er} déc. 2003), p. 913–933. ISSN : 0962-9343, 1573-2649. DOI : 10.1023/A:1026163113446. URL : <http://link.springer.com/article/10.1023/A:1026163113446> (visité le 05/06/2013).
- [22] Jakob Bue BJORNER, Chih-Hung CHANG, David THISSEN et al. « Developing tailored instruments : item banking and computerized adaptive assessment ». In : *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 16 Suppl 1 (2007). 00114, p. 95–108. ISSN : 0962-9343. DOI : 10.1007/s11136-007-9168-6.
- [23] Hendrik BLOCKEEL et Luc DE RAEDT. « Top-down induction of first-order logical decision trees ». In : *Artificial Intelligence* 101.1 (1^{er} mai 1998), p. 285–297. ISSN : 0004-3702. DOI : 10.1016/S0004-3702(98)00034-4. URL : <http://www.sciencedirect.com/science/article/pii/S0004370298000344> (visité le 27/07/2016).
- [24] R. Darrell BOCK. « Estimating item parameters and latent ability when responses are scored in two or more nominal categories ». In : *Psychometrika* 37.1 (1^{er} mar. 1972). 01094, p. 29–51. ISSN : 0033-3123, 1860-0980. DOI : 10.1007/BF02291411. URL : <http://link.springer.com/article/10.1007/BF02291411> (visité le 18/05/2015).
- [25] R. Darrell BOCK et Murray AITKIN. « Marginal maximum likelihood estimation of item parameters : Application of an EM algorithm ». In : *Psychometrika* 46.4 (1^{er} déc. 1981). 01987, p. 443–459. ISSN : 0033-3123, 1860-0980. DOI : 10.1007/BF02293801. URL : <http://link.springer.com/article/10.1007/BF02293801> (visité le 15/01/2013).

- [26] Anne£4340 BOOMSMA et T. A. B. £4340 SNIJDERS. *Essays on Item Response Theory*. 00000. Springer-Verlag, 2001. 464 p. ISBN : 978-0-387-95147-8.
- [27] Laurent BOYER, Karine BAUMSTARCK, Mohamed BOUCEKINE et al. « Measuring quality of life in patients with schizophrenia :an overview ». In : *Expert review of pharmacoeconomics & outcomes research* 13.3 (juin 2013), p. 343–349. ISSN : 1744-8379. DOI : 10.1586/erp.13.15.
- [28] Laurent BOYER, Marie-Claude SIMEONI, Anderson LOUNDOU et al. « The development of the S-QoL 18 : a shortened quality of life questionnaire for patients with schizophrenia ». In : *Schizophrenia research* 121.1 (août 2010), p. 241–250. ISSN : 1573-2509. DOI : 10.1016/j.schres.2010.05.019.
- [29] Leo BREIMAN. *Classification and regression trees*. Wadsworth International Group, 1984. 376 p. ISBN : 978-0-534-98053-5.
- [30] Leo BREIMAN. « Heuristics of instability and stabilization in model selection ». In : *The Annals of Statistics* 24.6 (déc. 1996). 00993, p. 2350–2383. ISSN : 0090-5364, 2168-8966. DOI : 10.1214/aos/1032181158. URL : <http://projecteuclid.org/euclid.aos/1032181158> (visité le 30/08/2016).
- [31] Leo BREIMAN. « Looking inside the black box. » In : *Wald Lecture 2, Berkeley University* (2001). 00023.
- [32] Leo BREIMAN. « Random Forests ». In : *Machine Learning* 45.1 (oct. 2001). 00000, p. 5–32. ISSN : 0885-6125, 1573-0565. DOI : 10.1023/A:1010933404324. URL : <http://link.springer.com/article/10.1023/A%3A1010933404324> (visité le 30/03/2016).
- [33] Li CAI. « Metropolis-Hastings Robbins-Monro Algorithm for Confirmatory Item Factor Analysis ». In : *Journal of Educational and Behavioral Statistics* 35.3 (1^{er} juin 2010), p. 307–335. ISSN : 1076-9986, 1935-1054. DOI : 10.3102/1076998609353115. URL : <http://jeb.sagepub.com/content/35/3/307> (visité le 15/01/2015).
- [34] D. T. CAMPBELL et D. W. FISKE. « Convergent and discriminant validation by the multitrait-multimethod matrix ». In : *Psychological Bulletin* 56.2 (mar. 1959). 15763, p. 81–105. ISSN : 0033-2909.
- [35] R. G. CAREY et J. H. SEIBERT. « A patient survey system to measure quality improvement : questionnaire reliability and validity ». In : *Medical Care* 31.9 (sept. 1993). 00264, p. 834–845. ISSN : 0025-7079.
- [36] D. CELLA, J.-S. LAI, C. J. NOWINSKI et al. « Neuro-QOL : brief measures of health-related quality of life for clinical research in neurology ». In : *Neurology* 78.23 (5 juin 2012), p. 1860–1867. ISSN : 1526-632X. DOI : 10.1212/WNL.0b013e318258f744.

- [37] David CELLA, Cindy NOWINSKI, Amy PETERMAN et al. « The neurology quality-of-life measurement initiative ». In : *Archives of Physical Medicine and Rehabilitation* 92.10 (oct. 2011), S28–36. ISSN : 1532-821X. DOI : 10.1016/j.apmr.2011.01.025.
- [38] David CELLA, William RILEY, Arthur STONE et al. « The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks : 2005-2008 ». In : *Journal of Clinical Epidemiology* 63.11 (nov. 2010), p. 1179–1194. ISSN : 1878-5921. DOI : 10.1016/j.jclinepi.2010.04.011.
- [39] David CELLA, Susan YOUNT, Nan ROTHROCK et al. « The Patient-Reported Outcomes Measurement Information System (PROMIS) : progress of an NIH Roadmap cooperative group during its first two years ». In : *Medical Care* 45.5 (mai 2007), S3–S11. ISSN : 0025-7079. DOI : 10.1097/01.mlr.0000258615.42478.55.
- [40] R. Philip CHALMERS. « mirt : A Multidimensional Item Response Theory Package for the R Environment ». In : *JSS Journal of Statistical Software* 48 (2012).
- [41] Marie CHAVENT, Yves LECHEVALLIER et Olivier BRIANT. « DIVCLUS-T : A monothetic divisive hierarchical clustering method ». In : *Computational Statistics & Data Analysis* 52.2 (15 oct. 2007), p. 687–701. ISSN : 0167-9473. DOI : 10.1016/j.csda.2007.03.013. URL : <http://www.sciencedirect.com/science/article/pii/S0167947307001181> (visité le 25/07/2016).
- [42] Xiaojun CHEN, Xiaofei XU, J. Z. HUANG et al. « TW-k-means : Automated two-level variable weighting clustering algorithm for multiview data ». In : *IEEE Transactions on Knowledge and Data Engineering* 25.4 (2013). 00038, p. 932–944. ISSN : 1041-4347. DOI : 10.1109/TKDE.2011.262.
- [43] Seung W. CHOI, Laura E. GIBBONS et Paul K. CRANE. « lordif : An R Package for Detecting Differential Item Functioning Using Iterative Hybrid Ordinal Logistic Regression/Item Response Theory and Monte Carlo Simulations ». In : *Journal of statistical software* 39.8 (1^{er} mar. 2011), p. 1–30. ISSN : 1548-7660. URL : <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3093114/> (visité le 22/10/2013).
- [44] Seung W. CHOI, Steven P. REISE, Paul A. PILKONIS et al. « Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms ». In : *Quality of Life Research* 19.1 (1^{er} fév. 2010), p. 125–136. ISSN : 0962-9343, 1573-2649. DOI : 10.1007/s11136-009-9560-5. URL : <http://link.springer.com/article/10.1007/s11136-009-9560-5> (visité le 15/01/2013).

- [45] Seung W. CHOI et Richard J. SWARTZ. « Comparison of CAT Item Selection Criteria for Polytomous Items ». In : *Applied Psychological Measurement* 33.6 (9 jan. 2009). 00053, p. 419–440. ISSN : 0146-6216, 1552-3497. DOI : 10.1177/0146621608327801. URL : <http://apm.sagepub.com/content/33/6/419> (visité le 07/02/2013).
- [46] Karon F. COOK, Alyssa M. BAMER, Toni S. RODDEY et al. « A PROMIS fatigue short form for use by individuals who have multiple sclerosis ». In : *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation* 21.6 (août 2012), p. 1021–1030. ISSN : 1573-2649. DOI : 10.1007/s11136-011-0011-8.
- [47] James E. CORTER et Mark A. GLUCK. « Explaining basic categories : Feature predictability and information. » In : *Psychological Bulletin* 111.2 (1992), p. 291–303. ISSN : 0033-2909. DOI : 10.1037/0033-2909.111.2.291. URL : <http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.111.2.291> (visité le 04/08/2016).
- [48] Linda CROCKER et James ALGINA. *Introduction to Classical and Modern Test Theory*. 04842. Mason, Ohio : Wadsworth Pub Co, 9 nov. 2006. 527 p. ISBN : 978-0-495-39591-1.
- [49] Lee J. CRONBACH. « Coefficient alpha and the internal structure of tests ». In : *Psychometrika* 16.3 (). 30047, p. 297–334. ISSN : 0033-3123, 1860-0980. DOI : 10.1007/BF02310555. URL : <http://link.springer.com/article/10.1007/BF02310555> (visité le 02/09/2016).
- [50] A. P. DEMPSTER, N. M. LAIRD et D. B. RUBIN. « Maximum likelihood from incomplete data via the EM algorithm ». In : *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B* 39.1 (1977). 45646, p. 1–38.
- [51] Michel C. DESMARAIS et Xiaoming PU. « Computer Adaptive Testing : Comparison of a Probabilistic Network Approach with Item Response Theory ». In : *User Modeling 2005*. Sous la dir. de Liliana ARDISSONO, Paul BRNA et Antonija MITROVIC. Lecture Notes in Computer Science 3538. Springer Berlin Heidelberg, 1^{er} jan. 2005, p. 392–396. ISBN : 978-3-540-27885-6 978-3-540-31878-1. URL : http://link.springer.com/chapter/10.1007/11527886_51 (visité le 07/02/2013).
- [52] Darren A. DEWALT, Nan ROTHROCK, Susan YOUNT et al. « Evaluation of Item Candidates : The PROMIS Qualitative Item Review ». In : *Medical care* 45.5 (mai 2007), S12–S21. ISSN : 0025-7079. DOI : 10.1097/01.mlr.0000254567.79743.e2. URL : <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2810630/> (visité le 18/12/2012).
- [53] R A DEYO et D L PATRICK. « Barriers to the use of health status measures in clinical investigation, patient care, and policy research ». In : *Medical care* 27.3 (mar. 1989), S254–268. ISSN : 0025-7079.

- [54] J. C. DUNN. « A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters ». In : *Journal of Cybernetics* 3.3 (1^{er} jan. 1973), p. 32–57. ISSN : 0022-0280. DOI : 10.1080/01969727308546046. URL : <http://dx.doi.org/10.1080/01969727308546046> (visité le 18/06/2015).
- [55] Susan E. EMBRETSON et Steven P. REISE. *Item Response Theory for Psychologists*. 1 edition. Mahwah, N.J : Psychology Press, 3 mai 2000. 384 p. ISBN : 978-0-8058-2819-1.
- [56] M. ERHART, C. HAGQUIST, P. AUQUIER et al. « A comparison of Rasch item-fit and Cronbach's alpha item reduction analysis for the development of a Quality of Life scale for children and adolescents ». In : *Child : Care, Health and Development* 36.4 (juil. 2010). 00020, p. 473–484. ISSN : 1365-2214. DOI : 10.1111/j.1365-2214.2009.00998.x.
- [57] Martin ESTER, Hans-peter KRIEGEL, Jörg S et al. « A density-based algorithm for discovering clusters in large spatial databases with noise ». In : 00000. AAAI Press, 1996, p. 226–231.
- [58] P. M. FAYERS et D. J. HAND. « Factor analysis, causal indicators and quality of life ». In : *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation* 6.2 (mar. 1997). 00000, p. 139–150. ISSN : 0962-9343.
- [59] Peter FAYERS et David MACHIN. *Quality of Life : The Assessment, Analysis and Interpretation of Patient-reported Outcomes*. 2^e éd. Wiley, 9 avr. 2007. 566 p. ISBN : 0-470-02450-X.
- [60] Douglas H. FISHER. « Knowledge Acquisition Via Incremental Conceptual Clustering ». In : *Machine Learning* 2.2 (1^{er} sept. 1987). 02405, p. 139–172. ISSN : 0885-6125, 1573-0565. DOI : 10.1023/A:1022852608280. URL : <http://link.springer.com/article/10.1023/A%3A1022852608280> (visité le 18/06/2015).
- [61] R. A. FISHER. « The Use of Multiple Measurements in Taxonomic Problems ». In : *Annals of Eugenics* 7.2 (1^{er} sept. 1936), p. 179–188. ISSN : 2050-1439. DOI : 10.1111/j.1469-1809.1936.tb02137.x. URL : <http://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.1936.tb02137.x/abstract> (visité le 15/08/2016).
- [62] K. FLOREK, J. ŁUKASZEWICZ, J. PERKAL et al. « Sur la liaison et la division des points d'un ensemble fini ». In : *Colloquium Mathematicae* 2.3 (1951). 00276, p. 282–285. ISSN : 0010-1354. URL : <https://eudml.org/doc/209969> (visité le 26/08/2016).
- [63] M F FOLSTEIN, L N ROBINS et J E HELZER. « The Mini-Mental State Examination ». In : *Archives of general psychiatry* 40.7 (juil. 1983), p. 812. ISSN : 0003-990X.

- [64] Ricardo FRAIMAN, Badih GHATTAS et Marcela SVARC. « Interpretable clustering using unsupervised binary trees ». In : *Advances in Data Analysis and Classification* 7.2 (29 mar. 2013). 00009, p. 125–145. ISSN : 1862-5347, 1862-5355. DOI : 10.1007/s11634-013-0129-3. URL : <http://link.springer.com/article/10.1007/s11634-013-0129-3> (visité le 15/08/2016).
- [65] Chris FRALEY et Adrian E. RAFTERY. « Model-Based Clustering, Discriminant Analysis, and Density Estimation ». In : *Journal of the American Statistical Association* 97.458 (2002). 02429, p. 611–631. ISSN : 0162-1459. URL : <http://www.jstor.org/stable/3085676> (visité le 15/08/2016).
- [66] Luis Angel GARCÍA-ESCUADERO, Alfonso GORDALIZA, Carlos MATRÁN et al. « A review of robust clustering methods ». In : *Advances in Data Analysis and Classification* 4.2 (18 juin 2010). 00066, p. 89–109. ISSN : 1862-5347, 1862-5355. DOI : 10.1007/s11634-010-0064-5. URL : <http://link.springer.com/article/10.1007/s11634-010-0064-5> (visité le 15/08/2016).
- [67] William GARDNER, Kelly J. KELLEHER et Kathleen A. PAJER. « Multidimensional adaptive testing for mental health problems in primary care ». In : *Medical Care* 40.9 (sept. 2002), p. 812–823. ISSN : 0025-7079. DOI : 10.1097/01.MLR.0000025436.30093.77.
- [68] Badih GHATTAS. « Importance des variables dans les methodes CART. » In : *Modulad* 24 (1999). 00000, p. 29–39.
- [69] Badih GHATTAS. « Agrégation d'arbres de classification ». In : *Revue de Statistique Appliquée* 48.2 (2000). 00020, p. 85–98. URL : http://www.numdam.org/numdam-bin/item?id=RSA_2000__48_2_85_0 (visité le 30/08/2016).
- [70] Badih GHATTAS et Pierre MICHEL. « Clustering ordinal data using binary decision trees. » In : *Proceedings of COMPSTAT 2014* 21st International Conference on Computational Statistics, Geneva, Switzerland. (2014).
- [71] Robert D. GIBBONS, David J. WEISS, Ellen FRANK et al. « Computerized Adaptive Diagnosis and Testing of Mental Health Disorders ». In : *Annual Review of Clinical Psychology* 12.1 (2016). 00000, p. 83–104. DOI : 10.1146/annurev-clinpsy-021815-093634. URL : <http://dx.doi.org/10.1146/annurev-clinpsy-021815-093634> (visité le 26/05/2016).
- [72] Robert D GIBBONS, David J WEISS, Paul A PILKONIS et al. « Development of a computerized adaptive test for depression ». In : *Archives of general psychiatry* 69.11 (nov. 2012). 00046, p. 1104–1112. ISSN : 1538-3636. DOI : 10.1001/archgenpsychiatry.2012.14.

- [73] Mark A. GLUCK. « Information, uncertainty and the utility of categories ». In : *Proc. of the Seventh Annual Conf. on Cognitive Science Society*. Lawrence Erlbaum, 1985, p. 283–287. URL : <http://ci.nii.ac.jp/naid/10004577544/> (visité le 04/08/2016).
- [74] W E GOLDEN. « Health status measurement. Implementation strategies ». In : *Medical care* 30.5 (mai 1992), MS187–195, discussion MS196–209. ISSN : 0025-7079.
- [75] Joanne GREENHALGH, Andrew F LONG et Rob FLYNN. « The use of patient reported outcome measures in routine clinical practice : lack of impact or lack of theory ? » In : *Social science & medicine* (1982) 60.4 (fév. 2005), p. 833–843. ISSN : 0277-9536. DOI : 10.1016/j.socscimed.2004.06.022.
- [76] Jolie J GUTTELING, Jan J V BUSSCHBACH, Robert A de MAN et al. « Logistic feasibility of health related quality of life measurement in clinical practice : results of a prospective study in a large population of chronic liver patients ». In : *Health and quality of life outcomes* 6 (2008). 00000, p. 97. ISSN : 1477-7525. DOI : 10.1186/1477-7525-6-97.
- [77] Isabelle GUYON, Jason WESTON, Stephen BARNHILL et al. « Gene Selection for Cancer Classification using Support Vector Machines ». In : *Machine Learning* 46.1 (2002). 05038, p. 389–422. ISSN : 0885-6125, 1573-0565. DOI : 10.1023/A:1012487302797. URL : <http://link.springer.com/article/10.1023/A:1012487302797> (visité le 15/08/2016).
- [78] Stephen M. HALEY, Pengsheng NI, Helene M. DUMAS et al. « Measuring global physical health in children with cerebral palsy : illustration of a multidimensional bi-factor model and computerized adaptive testing ». In : *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation* 18.3 (avr. 2009), p. 359–370. ISSN : 0962-9343. DOI : 10.1007/s11136-009-9447-5.
- [79] Mark HALL, Eibe FRANK, Geoffrey HOLMES et al. « The WEKA Data Mining Software : An Update ». In : *SIGKDD Explor. Newsl.* 11.1 (nov. 2009). 11779, p. 10–18. ISSN : 1931-0145. DOI : 10.1145/1656274.1656278. URL : <http://doi.acm.org/10.1145/1656274.1656278> (visité le 15/08/2016).
- [80] Michele Y HALYARD, Marlene H FROST et Amylou DUECK. « Integrating QOL assessments for clinical and research purposes ». In : *Current problems in cancer* 30.6 (déc. 2006), p. 319–330. ISSN : 0147-0272. DOI : 10.1016/j.currproblcancer.2006.08.009.
- [81] Michele Y HALYARD, Marlene H FROST, Amylou DUECK et al. « Is the use of QOL data really any different than other medical testing ? » In : *Current problems in cancer* 30.6 (déc. 2006), p. 261–271. ISSN : 0147-0272. DOI : 10.1016/j.currproblcancer.2006.08.004.

- [82] Ronald K. HAMBLETON, Hariharan SWAMINATHAN et H. Jane ROGERS. *Fundamentals of Item Response Theory*. SAGE, 23 juil. 1991. 192 p. ISBN : 978-0-8039-3647-8.
- [83] Leo M. HARVILL. « Standard Error of Measurement ». In : *Educational Measurement : Issues and Practice* 10.2 (1^{er} juin 1991), p. 33–41. ISSN : 1745-3992. DOI : 10.1111/j.1745-3992.1991.tb00195.x. URL : <http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3992.1991.tb00195.x/abstract> (visité le 09/04/2015).
- [84] Trevor HASTIE, Robert TIBSHIRANI et Jerome FRIEDMAN. *The Elements of Statistical Learning*. Springer Series in Statistics. 00885. New York, NY : Springer New York, 2009. ISBN : 978-0-387-84857-0 978-0-387-84858-7. URL : <http://link.springer.com/10.1007/978-0-387-84858-7> (visité le 15/08/2016).
- [85] Ron D. HAYS, Honghu LIU, Karen SPRITZER et al. « Item response theory analyses of physical functioning items in the medical outcomes study ». In : *Medical Care* 45.5 (mai 2007), S32–38. ISSN : 0025-7079. DOI : 10.1097/01.mlr.0000246649.43232.82.
- [86] Christian HENNIG. « A Method for Visual Cluster Validation ». In : *Classification — the Ubiquitous Challenge*. Sous la dir. de Professor Dr Claus WEIHS et Professor Dr Wolfgang GAUL. Studies in Classification, Data Analysis, and Knowledge Organization. 00010 DOI : 10.1007/3-540-28084-7_15. Springer Berlin Heidelberg, 2005, p. 153–160. ISBN : 978-3-540-25677-9 978-3-540-28084-2. URL : http://link.springer.com/chapter/10.1007/3-540-28084-7_15 (visité le 02/09/2016).
- [87] Christian HENNIG. « Clustering strategy and method selection ». In : *arXiv :1503.02059 [stat]* (6 mar. 2015). 00000. arXiv : 1503.02059. URL : <http://arxiv.org/abs/1503.02059> (visité le 02/09/2016).
- [88] I. J. HIGGINSON et A. J. CARR. « Measuring quality of life : Using quality of life measures in the clinical setting ». In : *BMJ (Clinical research ed.)* 322.7297 (26 mai 2001). 00000, p. 1297–1300. ISSN : 0959-8138.
- [89] Cheryl D. HILL, Michael C. EDWARDS, David THISSEN et al. « Practical issues in the application of item response theory : a demonstration using items from the pediatric quality of life inventory (PedsQL) 4.0 generic core scales ». In : *Medical Care* 45.5 (mai 2007), S39–47. ISSN : 0025-7079. DOI : 10.1097/01.mlr.0000259879.05499.eb.
- [90] D. HOOPER, J. COUGHLAN et M. MULLEN. « Structural equation modelling : guidelines for determining model fit ». In : *Electronic Journal of Buisness Research Methods* 6.1 (2008), p. 53–60.

- [91] Zhexue HUANG. « Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values ». In : *Data Mining and Knowledge Discovery* 2.3 (), p. 283–304. ISSN : 1384-5810, 1573-756X. DOI : 10.1023/A:1009769707641. URL : <http://link.springer.com/article/10.1023/A:1009769707641> (visité le 15/08/2016).
- [92] Lawrence HUBERT et Phipps ARABIE. « Comparing partitions ». In : *Journal of Classification* 2.1 (1^{er} déc. 1985), p. 193–218. ISSN : 0176-4268, 1432-1343. DOI : 10.1007/BF01908075. URL : <http://link.springer.com/article/10.1007/BF01908075> (visité le 02/08/2013).
- [93] Mallasandra V. JAGANNATHA REDDY et Balli KAVITHA. « Clustering the Mixed Numerical and Categorical Dataset using Similarity Weight and Filter Method ». In : (2012).
- [94] Zhezhen JIN et Mounir MESBAH. « Unidimensionality, Agreement and Concordance Probability ». In : *Statistical Models and Methods for Reliability and Survival Analysis*. Sous la dir. de Vincent COUALLIER, Léo GERVILLE-RÉACHE, Catherine HUBER-CAROL et al. 00001. John Wiley & Sons, Inc., 2013, p. 1–19. ISBN : 978-1-118-82680-5. URL : <http://onlinelibrary.wiley.com/doi/10.1002/9781118826805.ch1/summary> (visité le 02/09/2016).
- [95] Matthew S. JOHNSON. « Marginal Maximum Likelihood Estimation of Item Response Models in R ». In : *Journal of Statistical Software* 20.10 (2007), p. 1–24. ISSN : 1548-7660. URL : <http://www.jstatsoft.org/v20/i10>.
- [96] Karl G. JÖRESKOG. « Structural Equation Modeling with Ordinal Variables ». In : *Lecture Notes-Monograph Series* 24 (1994). 00026, p. 297–310. ISSN : 0749-2170. URL : <http://www.jstor.org/stable/4355811> (visité le 02/09/2016).
- [97] Taehoon KANG et Troy T. CHEN. « Performance of the Generalized S-X2 Item Fit Index for Polytomous IRT Models ». In : *Journal of Educational Measurement* 45.4 (2008), p. 391–406. ISSN : 1745-3984. DOI : 10.1111/j.1745-3984.2008.00071.x. URL : <http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.2008.00071.x/abstract> (visité le 16/01/2013).
- [98] Stanley R. KAY, Abraham FISZBEIN et Lewis A. OPLER. « The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia ». In : *Schizophrenia Bulletin* 13.2 (1^{er} jan. 1987), p. 261–276. ISSN : 0586-7614, 1745-1701. DOI : 10.1093/schbul/13.2.261. URL : <http://schizophreniabulletin.oxfordjournals.org/content/13/2/261> (visité le 18/03/2016).
- [99] J F KURTZE. « On the evaluation of disability in multiple sclerosis ». In : *Neurology* 11 (août 1961), p. 686–694. ISSN : 0028-3878.

- [100] C. LANÇON, G. REINE, P. M. LLORCA et al. « Validity and reliability of the French-language version of the Positive and Negative Syndrome Scale (PANSS) ». In : *Acta Psychiatrica Scandinavica* 100.3 (sept. 1999), p. 237–243. ISSN : 0001-690X.
- [101] K. P. LEONG, S. C. L. YEAK, A. S. M. SAURAJEN et al. « Why generic and disease-specific quality-of-life instruments should be used together for the evaluation of patients with persistent allergic rhinitis ». In : *Clinical and Experimental Allergy : Journal of the British Society for Allergy and Clinical Immunology* 35.3 (mar. 2005), p. 288–298. ISSN : 0954-7894. DOI : 10.1111/j.1365-2222.2005.02201.x.
- [102] A. LEPLÈGE, E. ECOSSE, J. POUCHOT et al. *MOS SF36 Questionnaire. Manual and Guidelines for Scores' Interpretation*. Vernouillet : Estem 156. Paris : Editions Estem, 2001. ISBN : 2-84371-118-5.
- [103] A. LEPLÈGE, E. ECOSSE, A. VERDIER et al. « The French SF-36 Health Survey : translation, cultural adaptation and preliminary psychometric evaluation ». In : *Journal of clinical epidemiology* 51.11 (nov. 1998), p. 1013–1023. ISSN : 0895-4356.
- [104] LINACRE J. M. et WRIGHT B. D. « (Dichotomous Mean-square) Chi-square fit statistics. » In : *Rasch Measurement Transactions* 8.2 (1994). 00000, p. 360.
- [105] Wim J. van der LINDEN et Cees A. W. GLAS, éd. *Elements of Adaptive Testing*. 2010 edition. New York : Springer, 16 fév. 2010. 438 p. ISBN : 978-0-387-85459-5.
- [106] Wim J. van der LINDEN et Ronald K. HAMBLETON. *Handbook of Modern Item Response Theory*. 01537. Springer, 15 nov. 1996. 530 p. ISBN : 978-0-387-94661-0.
- [107] Drew A. LINZER et Lewis B. JEFFREY. « poLCA : An R Package for Polytomous Variable Latent Class Analysis ». In : *Journal of Statistical Software* (2011), p. 1–29.
- [108] Bing LIU, Yiyuan XIA et Philip S. YU. « Clustering Through Decision Tree Construction ». In : *Proceedings of the Ninth International Conference on Information and Knowledge Management*. CIKM '00. New York, NY, USA : ACM, 2000, p. 20–29. ISBN : 978-1-58113-320-2. DOI : 10.1145/354756.354775. URL : <http://doi.acm.org/10.1145/354756.354775> (visité le 27/07/2016).
- [109] Huan LIU et Lei YU. « Toward integrating feature selection algorithms for classification and clustering ». In : *IEEE Transactions on Knowledge and Data Engineering* 17.4 (avr. 2005), p. 491–502. ISSN : 1041-4347. DOI : 10.1109/TKDE.2005.66.

- [110] Frederic M. LORD. « Maximum Likelihood and Bayesian Parameter Estimation in Item Response Theory ». In : *Journal of Educational Measurement* 23.2 (1986), p. 157–162. ISSN : 1745-3984. DOI : 10.1111/j.1745-3984.1986.tb00241.x. URL : <http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.1986.tb00241.x/abstract> (visit  le 07/02/2013).
- [111] Frederic M. LORD et Melvin R. NOVICK. *Statistical Theories of Mental Test Scores*. 07418. Charlotte, N.C. : Information Age Publishing, 18 avr. 2008. 592 p. ISBN : 978-1-59311-934-8.
- [112] F D LUBLIN et S C REINGOLD. « Defining the clinical course of multiple sclerosis : results of an international survey. National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis ». In : *Neurology* 46.4 (avr. 1996), p. 907–911. ISSN : 0028-3878.
- [113] James B. MACQUEEN. « Some methods for classification and analysis of multivariate observations ». In : *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics*. The Regents of the University of California, 1967. URL : <http://projecteuclid.org/euclid.bsmsp/1200512992> (visit  le 25/07/2016).
- [114] Patrick MAIR et Reinhold HATZINGER. *Extended Rasch Modeling : The eRm Package for the Application of IRT Models in R*. 00000. 2007. URL : <http://epub.wu.ac.at/332/> (visit  le 18/04/2014).
- [115] Trupti M. Kodinariya Dr Prashant R. MAKWANA. « Partitioning Clustering algorithms for handling numerical and categorical data : a review ». In : *arXiv :1311.7219 [cs]* (28 nov. 2013). arXiv : 1311.7219. URL : <http://arxiv.org/abs/1311.7219> (visit  le 18/08/2016).
- [116] Geoff N. MASTERS. « A rasch model for partial credit scoring ». In : *Psychometrika* 47.2 (1^{er} juin 1982). 02599, p. 149–174. ISSN : 0033-3123, 1860-0980. DOI : 10.1007/BF02296272. URL : <http://link.springer.com/article/10.1007/BF02296272> (visit  le 17/01/2013).
- [117] W. I. McDONALD, A. COMPSTON, G. EDAN et al. « Recommended diagnostic criteria for multiple sclerosis : guidelines from the International Panel on the diagnosis of multiple sclerosis ». In : *Annals of Neurology* 50.1 (juil. 2001), p. 121–127. ISSN : 0364-5134.
- [118] Ryszard S. MICHALSKI. « Learning by being told and learning from examples : an experimental comparison of the two methodes of knowledge acquisition in the context of developing an expert system for soybean disease diagnosis ». In : *International Journal of Policy Analysis and Information Systems* (1980). 00000, p. 125–161.

- [119] Pierre MICHEL, Karine BAUMSTARCK, Pascal AUQUIER et al. « Psychometric properties of the abbreviated version of the Scale to Assess Unawareness in Mental Disorder in schizophrenia ». In : *BMC Psychiatry* 13 (2013). 00007, p. 229. ISSN : 1471-244X. DOI : 10.1186/1471-244X-13-229. URL : <http://dx.doi.org/10.1186/1471-244X-13-229> (visité le 02/09/2016).
- [120] Pierre MICHEL, Karine BAUMSTARCK, Pascal AUQUIER et al. « How to interpret multidimensional quality of life questionnaires for patients with schizophrenia ? » In : *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation* (9 avr. 2015). 00000. ISSN : 1573-2649. DOI : 10.1007/s11136-015-0982-y.
- [121] Pierre MICHEL, Karine BAUMSTARCK, Laurent BOYER et al. « Defining Quality of Life Levels to Enhance Clinical Interpretation in Multiple Sclerosis : Application of a Novel Clustering Method ». In : *Medical Care* (15 mar. 2014). ISSN : 1537-1948. DOI : 10.1097/MLR.000000000000117.
- [122] Pierre MICHEL, Karine BAUMSTARCK, Badih GHATTAS et al. « A Multidimensional Computerized Adaptive Short-Form Quality of Life Questionnaire Developed and Validated for Multiple Sclerosis. The MusiQoL-MCAT ». In : *Medicine* (2016).
- [123] Deborah M MILLER et Rebecca ALLEN. « Quality of life in multiple sclerosis : determinants, measurement, and use in clinical practice ». In : *Current neurology and neuroscience reports* 10.5 (sept. 2010), p. 397–406. ISSN : 1534-6293. DOI : 10.1007/s11910-010-0132-4.
- [124] Alex J MITCHELL, Julián BENITO-LEÓN, José-Manuel Morales GONZÁLEZ et al. « Quality of life and its assessment in multiple sclerosis : integrating physical and psychological components of wellbeing ». In : *Lancet neurology* 4.9 (sept. 2005), p. 556–566. ISSN : 1474-4422. DOI : 10.1016/S1474-4422(05)70166-6.
- [125] J. MORRIS, D. PEREZ et B. MCNOE. « The use of quality of life data in clinical practice ». In : *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation* 7.1 (jan. 1998), p. 85–91. ISSN : 0962-9343.
- [126] Joris MULDER et Wim J. van der LINDEN. « Multidimensional Adaptive Testing with Kullback–Leibler Information Item Selection ». In : *Elements of Adaptive Testing*. Sous la dir. de Wim J. van der LINDEN et Cees A. W. GLAS. Statistics for Social and Behavioral Sciences. Springer New York, 2009, p. 77–101. ISBN : 978-0-387-85459-5 978-0-387-85461-8. URL : http://link.springer.com/chapter/10.1007/978-0-387-85461-8_4 (visité le 28/01/2015).

- [127] Eiji MURAKI. « A Generalized Partial Credit Model : Application of an EM Algorithm ». In : *Applied Psychological Measurement* 16.2 (6 jan. 1992), p. 159–176. ISSN : 0146-6216, 1552-3497. DOI : 10.1177/014662169201600206. URL : <http://apm.sagepub.com/content/16/2/159> (visité le 15/01/2015).
- [128] Fionn MURTAGH. *Multidimensional Clustering Algorithms*. 1 edition. 00410. Vienna : Physica, 1^{er} jan. 1985. ISBN : 978-3-7051-0008-4.
- [129] L.K MUTHÉN et B.O MUTHÉN. « MPlus User's Guide. Seventh Edition ». In : *Muthén & Muthén* (2012).
- [130] J Gareth NOBLE, Lisa A OSBORNE, Kerina H JONES et al. « Commentary on 'disability outcome measures in multiple sclerosis clinical trials' ». In : *Multiple sclerosis* 18.12 (déc. 2012), p. 1718–1720. ISSN : 1477-0970. DOI : 10.1177/1352458512457847.
- [131] Jum C. NUNNALLY et Ira BERNSTEIN. *Psychometric Theory*. 3rd Revised edition. 85689. New York : McGraw Hill Higher Education, 1^{er} nov. 1993. 736 p. ISBN : 978-0-07-047849-7.
- [132] Christos H. PAPADIMITRIOU et Kenneth STEIGLITZ. *Combinatorial Optimization : Algorithms and Complexity*. 00000. Upper Saddle River, NJ, USA : Prentice-Hall, Inc., 1982. ISBN : 978-0-13-152462-0.
- [133] D. L. PATRICK et R. A. DEYO. « Generic and disease-specific measures in assessing health status and quality of life ». In : *Medical Care* 27.3 (mar. 1989), S217–232. ISSN : 0025-7079.
- [134] Daniel PEÑA et Francisco J. PRIETO. « Cluster Identification Using Projections ». In : *Journal of the American Statistical Association* 96.456 (2001), p. 1433–1445. ISSN : 0162-1459. URL : <http://www.jstor.org/stable/3085911> (visité le 15/08/2016).
- [135] Daniel PEÑA et Francisco J. PRIETO. « A Projection Method for Robust Estimation and Clustering in Large Data Sets ». In : *Data Analysis, Classification and the Forward Search*. Sous la dir. de Prof Sergio ZANI, Prof Andrea CERIOLI, Prof Marco RIANI et al. Studies in Classification, Data Analysis, and Knowledge Organization. DOI : 10.1007/3-540-35978-8_24. Springer Berlin Heidelberg, 2006, p. 209–216. ISBN : 978-3-540-35977-7 978-3-540-35978-4. URL : http://link.springer.com/chapter/10.1007/3-540-35978-8_24 (visité le 15/08/2016).
- [136] Morten Aa PETERSEN, Mogens GROENVOLD, Neil AARONSON et al. « Multidimensional Computerized Adaptive Testing of the EORTC QLQ-C30 : Basic Developments and Evaluations ». In : *Quality of Life Research* 15.3 (1^{er} avr. 2006). 00000, p. 315–329. ISSN : 0962-9343, 1573-2649. DOI : 10.1007/s11136-005-3214-z. URL : <http://link.springer.com/article/10.1007/s11136-005-3214-z> (visité le 26/11/2012).

- [137] J. Ross QUINLAN. *C4.5 : Programs for Machine Learning*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 1993. ISBN : 978-1-55860-238-0.
- [138] Luc De RAEDT et Hendrik BLOCKEEL. « Using logical decision trees for clustering ». In : *Inductive Logic Programming*. Sous la dir. de Nada LAVRAČ et Sašo DŽEROSKI. Lecture Notes in Computer Science 1297. DOI : 10.1007/3540635149_41. Springer Berlin Heidelberg, 17 sept. 1997, p. 133–140. ISBN : 978-3-540-63514-7 978-3-540-69587-5. URL : http://link.springer.com/chapter/10.1007/3540635149_41 (visité le 27/07/2016).
- [139] Nambury S. RAJU, Larry R. PRICE, T. C. OSHIMA et al. « Standardized Conditional SEM : A Case for Conditional Reliability ». In : *Applied Psychological Measurement* 31.3 (5 jan. 2007). 00037, p. 169–180. ISSN : 0146-6216, 1552-3497. DOI : 10.1177/0146621606291569. URL : <http://apm.sagepub.com/content/31/3/169> (visité le 21/09/2016).
- [140] Alain RAKOTOMAMONJY. « Variable Selection Using Svm Based Criteria ». In : *J. Mach. Learn. Res.* 3 (mar. 2003), p. 1357–1370. ISSN : 1532-4435. URL : <http://dl.acm.org/citation.cfm?id=944919.944977> (visité le 18/08/2016).
- [141] Georg RASCH. *Probabilistic Models for Some Intelligence and Attainment Tests*. New edition. 07671. Chicago : University of Chicago Press, déc. 1980. 199 p. ISBN : 978-0-226-70554-5.
- [142] Tenko RAYKOV. « Evaluation of convergent and discriminant validity with multitrait-multimethod correlations ». In : *British Journal of Mathematical and Statistical Psychology* 64.1 (1^{er} fév. 2011). 00030, p. 38–52. ISSN : 2044-8317. DOI : 10.1348/000711009X478616. URL : <http://onlinelibrary.wiley.com/doi/10.1348/000711009X478616/abstract> (visité le 02/09/2016).
- [143] Bryce B. REEVE, Ron D. HAYS, Jakob B. BJORNER et al. « Psychometric evaluation and calibration of health-related quality of life item banks : plans for the Patient-Reported Outcomes Measurement Information System (PROMIS) ». In : *Medical Care* 45.5 (mai 2007), S22–31. ISSN : 0025-7079. DOI : 10.1097/01.mlr.0000250483.85507.04.
- [144] Steven P. REISE et Dennis A. REVICKI. *Handbook of Item Response Theory Modeling : Applications to Typical Performance Assessment*. Routledge, 20 nov. 2014. 484 p. ISBN : 978-1-317-56570-3.
- [145] Barth RILEY, Rodney FUNK, Michael DENNIS et al. *The Use of Decision Trees for Adaptive Item Selection and Score Estimation*. 00000. 2011.

- [146] Dimitris RIZOPOULOS. « ltm : An R Package for Latent Variable Modeling and Item Response Analysis ». In : *Journal of Statistical Software* 17.5 (2006).
- [147] Lior ROKACH. « A survey of Clustering Algorithms ». In : *Data Mining and Knowledge Discovery Handbook*. Sous la dir. d'Oded MAIMON et Lior ROKACH. 00066 DOI : 10.1007/978-0-387-09823-4_14. Springer US, 2009, p. 269–298. ISBN : 978-0-387-09822-7 978-0-387-09823-4. URL : http://link.springer.com/chapter/10.1007/978-0-387-09823-4_14 (visité le 27/07/2016).
- [148] M ROSE, J B BJORNER, J BECKER et al. « Evaluation of a preliminary physical function item bank supported the expected advantages of the Patient-Reported Outcomes Measurement Information System (PROMIS) ». In : *Journal of clinical epidemiology* 61.1 (jan. 2008), p. 17–33. ISSN : 0895-4356. DOI : 10.1016/j.jclinepi.2006.06.025.
- [149] William W. ROZEBOOM. « Review of Statistical Theories of Mental Test Scores ». In : *American Educational Research Journal* 6.1 (1969). Avec la coll. de Frederic M. LORD, Melvin R. NOVICK et Allan BIRNBAUM. 00000, p. 112–116. ISSN : 0002-8312. DOI : 10.2307/1162101. URL : <http://www.jstor.org/stable/1162101> (visité le 02/09/2016).
- [150] Richard A RUDICK et Deborah M MILLER. « Health-related quality of life in multiple sclerosis : current evidence, measurement and effects of disease severity and treatment ». In : *CNS drugs* 22.10 (2008). 00000, p. 827–839. ISSN : 1172-7047.
- [151] Fumiko SAMEJIMA. *Estimation of latent ability using a response pattern of graded scores*. Psychometric Society, 1969. 106 p.
- [152] Fumiko SAMEJIMA. « Normal ogive model on the continuous response level in the multidimensional latent space ». In : *Psychometrika* 39.1 (1^{er} mar. 1974), p. 111–121. ISSN : 0033-3123, 1860-0980. DOI : 10.1007/BF02291580. URL : <http://link.springer.com/article/10.1007/BF02291580> (visité le 05/12/2012).
- [153] Gideon SCHWARZ. « Estimating the Dimension of a Model ». In : *The Annals of Statistics* 6.2 (mar. 1978). Mathematical Reviews number (MathSciNet) MR468014, Zentralblatt MATH identifier 0379.62005, p. 461–464. ISSN : 0090-5364, 2168-8966. DOI : 10.1214/aos/1176344136. URL : <http://projecteuclid.org/euclid.aos/1176344136> (visité le 17/04/2014).
- [154] Angela SENDERS, Douglas HANES, Dennis BOURDETTE et al. « Reducing survey burden : feasibility and validity of PROMIS measures in multiple sclerosis ». In : *Multiple Sclerosis (Houndmills, Basingstoke, England)* 20.8 (8 jan. 2014), p. 1102–1111. ISSN : 1477-0970. DOI : 10.1177/1352458513517279.

- [155] M. C. SIMEONI, P. AUQUIER, O. FERNANDEZ et al. « Validation of the Multiple Sclerosis International Quality of Life questionnaire ». In : *Multiple Sclerosis* 14.2 (3 jan. 2008), p. 219–230. ISSN : 1352-4585, 1477-0970. DOI : 10.1177/1352458507080733. URL : <http://msj.sagepub.com/content/14/2/219> (visité le 03/05/2013).
- [156] Niels SMITS. « On the effect of adding clinical samples to validation studies of patient-reported outcome item banks : a simulation study ». In : *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation* 25.7 (juil. 2016). 00000, p. 1635–1644. ISSN : 1573-2649. DOI : 10.1007/s11136-015-1199-9.
- [157] Alessandra SOLARI. « Role of health-related quality of life measures in the routine care of people with multiple sclerosis ». In : *Health and Quality of Life Outcomes* 3.1 (18 mar. 2005), p. 16. ISSN : 1477-7525. DOI : 10.1186/1477-7525-3-16. URL : <http://www.hqlo.com/content/3/1/16/abstract> (visité le 02/07/2014).
- [158] « SPSS Inc. Released 2008. SPSS Statistics for Windows, Version 17.0. Chicago : SPSS Inc. » In : (2008).
- [159] James H. STEIGER. « Understanding the limitations of global fit assessment in structural equation modeling ». In : *Personality and Individual Differences* 42.5 (mai 2007), p. 893–898. ISSN : 0191-8869. DOI : 10.1016/j.paid.2006.09.017. URL : <http://www.sciencedirect.com/science/article/pii/S0191886906003825> (visité le 28/11/2013).
- [160] R Core TEAM. « R : A Language and Environment for Statistical Computing ». In : *R Foundation for Statistical Computing* (2012). 49156. ISSN : 3-900051-07-0. URL : <http://www.R-project.org/>.
- [161] David THISSEN et Howard WAINER, édés. *Test Scoring*. 1 édition. Mahwah, N.J : Routledge, 1^{er} mai 2001. 440 p. ISBN : 978-0-8058-3766-7.
- [162] M. UENO et P. SONGMUANG. « Computerized Adaptive Testing Based on Decision Tree ». In : *2010 IEEE 10th International Conference on Advanced Learning Technologies (ICALT)*. 2010 IEEE 10th International Conference on Advanced Learning Technologies (ICALT). Juil. 2010, p. 191–193. DOI : 10.1109/ICALT.2010.58.
- [163] Jeroen K. VERMUNT et Jay MAGIDSON. « Latent class cluster analysis ». In : *Applied latent class analysis*. 00000. Cambridge University Press, 2002, p. 89–106. URL : <http://books.google.com/books?hl=en&lr=&id=-0xrbRao0SsC&oi=fnd&pg=PA89&dq=info:aJajmECH63MJ:scholar.google.com&ots=Ou3A9f1cnN&sig=hCPiVD1UUA3-kM9enxzUIf9C624> (visité le 27/07/2016).

- [164] B. G. VICKREY, R. D. HAYS, R. HAROONI et al. « A health-related quality of life measure for multiple sclerosis ». In : *Quality of Life Research : An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation* 4.3 (juin 1995), p. 187–206. ISSN : 0962-9343.
- [165] Guenther WALTHER. « Optimal and fast detection of spatial clusters with scan statistics ». In : *The Annals of Statistics* 38.2 (avr. 2010), p. 1010–1033. ISSN : 0090-5364. DOI : 10.1214/09-AOS732. arXiv : 1002.4770. URL : <http://arxiv.org/abs/1002.4770> (visité le 15/08/2016).
- [166] Tianyou WANG et And OTHERS. « Conditional Standard Errors, Reliability and Decision Consistency of Performance Levels Using Polytomous IRT. » In : (avr. 1996). URL : <http://eric.ed.gov/?id=ED401323> (visité le 11/02/2015).
- [167] J. E. WARE et C. D. SHERBOURNE. « The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection ». In : *Medical Care* 30.6 (juin 1992), p. 473–483. ISSN : 0025-7079.
- [168] John E. WARE, Kristin K. SNOW, Mark KOSINSKI et al. *SF-36 health survey : manual and interpretation guide*. The Health Institute, New England Medical Center, 1993. 300 p.
- [169] Claus WEIHS, Uwe LIGGES, Karsten LUEBKE et al. « klaR Analyzing German Business Cycles ». In : *Data Analysis and Decision Support*. Sous la dir. de Prof Dr Daniel BAIER, Prof Dr Reinhold DECKER et Prof Dr Dr Lars SCHMIDT-THIEME. Studies in Classification, Data Analysis, and Knowledge Organization. DOI : 10.1007/3-540-28397-8_36. Springer Berlin Heidelberg, 2005, p. 335–343. ISBN : 978-3-540-26007-3 978-3-540-28397-3. URL : http://link.springer.com/chapter/10.1007/3-540-28397-8_36 (visité le 15/08/2016).
- [170] David J. WEISS. « Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education ». In : *Measurement and Evaluation in Counseling and Development* 37.2 (1^{er} juil. 2004). 00093, p. 70. ISSN : 0748-1756. (Visité le 16/02/2016).
- [171] Jason WESTON, André ELISSEEFF, Bernhard SCHÖLKOPF et al. « Use of the Zero Norm with Linear Models and Kernel Methods ». In : *J. Mach. Learn. Res.* 3 (mar. 2003). 00630, p. 1439–1461. ISSN : 1532-4435. URL : <http://dl.acm.org/citation.cfm?id=944919.944982> (visité le 30/08/2016).
- [172] B. D. WRIGHT et J. M. LINACRE. « Reasonable mean-square fit values. » In : *Rasch Measurement Transactions* 8 (1994), p. 370.
- [173] Duanli YAN, Charles LEWIS et Martha STOCKING. « Adaptive Testing without IRT. » In : (avr. 1998). URL : <http://eric.ed.gov/?id=ED422359> (visité le 18/11/2015).

- [174] Duanli YAN, Charles LEWIS et Martha STOCKING. *Adaptive Testing Without Irt in the Presence of Multidimensionality*. Educational Testing Service, 2002. 27 p.
- [175] Duanli YAN, Charles LEWIS et Martha STOCKING. « Adaptive Testing With Regression Trees in the Presence of Multidimensionality ». In : *Journal of Educational and Behavioral Statistics* 29.3 (21 sept. 2004). 00000, p. 293–316. ISSN : 1076-9986, 1935-1054. DOI : 10.3102/10769986029003293. URL : <http://jeb.sagepub.com/content/29/3/293> (visité le 18/11/2015).
- [176] Lihua YAO. « Multidimensional CAT Item Selection Methods for Domain Scores and Composite Scores With Item Exposure Control and Content Constraints ». In : *Journal of Educational Measurement* 51.1 (1^{er} mar. 2014), p. 18–38. ISSN : 1745-3984. DOI : 10.1111/jedm.12032. URL : <http://onlinelibrary.wiley.com/doi/10.1111/jedm.12032/abstract> (visité le 02/10/2014).
- [177] Lihua YAO et Keith A. BOUGHTON. « A Multidimensional Item Response Modeling Approach for Improving Subscale Proficiency Estimation and Classification ». In : *Applied Psychological Measurement* 31.2 (3 jan. 2007), p. 83–105. ISSN : 0146-6216, 1552-3497. DOI : 10.1177/0146621606291559. URL : <http://apm.sagepub.com/content/31/2/83> (visité le 13/11/2012).
- [178] Lihua YAO et Richard D. SCHWARZ. « A Multidimensional Partial Credit Model With Associated Item and Test Statistics : An Application to Mixed-Format Tests ». In : *Applied Psychological Measurement* 30.6 (11 jan. 2006), p. 469–492. ISSN : 0146-6216, 1552-3497. DOI : 10.1177/0146621605284537. URL : <http://apm.sagepub.com/content/30/6/469> (visité le 13/11/2012).
- [179] Yi ZHENG, Chih-Hung CHANG et Hua-Hua CHANG. « Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement ». In : *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 22.3 (avr. 2013). 00000, p. 491–499. ISSN : 1573-2649. DOI : 10.1007/s11136-012-0179-6.
- [180] Linling ZHU, Linsong MIAO et Daoqiang ZHANG. « Iterative Laplacian Score for Feature Selection ». In : *Pattern Recognition*. Sous la dir. de Cheng-Lin LIU, Changshui ZHANG et Liang WANG. Communications in Computer and Information Science 321. 00004 DOI : 10.1007/978-3-642-33506-8_11. Springer Berlin Heidelberg, 24 sept. 2012, p. 80–87. ISBN : 978-3-642-33505-1 978-3-642-33506-8. URL : http://link.springer.com/chapter/10.1007/978-3-642-33506-8_11 (visité le 29/08/2016).

- [181] Albrecht ZIMMERMANN et Luc De RAEDT. « Cluster-grouping : from subgroup discovery to clustering ». In : *Machine Learning* 77.1 (16 juin 2009), p. 125–159. ISSN : 0885-6125, 1573-0565. DOI : 10.1007/s10994-009-5121-y. URL : <http://link.springer.com/article/10.1007/s10994-009-5121-y> (visité le 18/06/2015).
- [182] B.D. ZUMBO. *A Handbook on the Theory and Methods of Differential Item Functioning (DIF) : Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, Directorate of Human Resources Research et Evaluation, Department of National Defense, 1999.