

THESE DE DOCTORAT DE

ONIRIS

ECOLE DOCTORALE N° 600

Ecole doctorale Ecologie, Géosciences, Agronomie et Alimentation

Spécialité : Statistiques/Modélisation en écologie, géosciences, agronomie et alimentation

Par

Fabien LLOBELL

Classification de tableaux de données, applications en analyse sensorielle

Thèse présentée et soutenue à Nantes, le 15/10/2020

Unité de recherche : Statistique, Sensométrie & Chimiométrie - USC 1381 Oniris INRAE [StatSC]

Rapporteurs avant soutenance :

Robert SABATIER
Ndeye NIANG

Professeur, Université de Montpellier 1
Maître de Conférences, CNAM Paris

Composition du Jury :

Président :	Pascal SCHLICH	Directeur de recherches, INRAE Dijon
Examineurs :	Stéphanie BOUGEARD	Ingénieure en recherche statistique, ANSES
	Robert SABATIER	Professeur, Université de Montpellier 1
	Ndeye NIANG	Maître de Conférences, CNAM Paris
Dir. de thèse :	El Mostafa QANNARI	Professeur, Oniris
Co-dir. de thèse :	Evelyne VIGNEAU	Professeur, Oniris
Co-dir. de thèse :	Véronique CARIOU	Maître de Conférences, Oniris
Co-encadrant :	Amaury LABENNE	Docteur, Addinsoft

« *La seule chose qui puisse empêcher un
rêve d'aboutir, c'est la peur d'échouer.* »
Paulo Coelho

Remerciements

Je souhaite adresser un grand merci à Ndèye Niang et Robert Sabatier d’avoir honoré ma demande de rapporter sur ces travaux de thèse. Merci également à Stéphanie Bougeard et Pascal Schlich qui ont accepté de faire partie du jury de thèse.

Mes remerciements vont bien évidemment à mon directeur de thèse El Mostafa Qanari. Si cette thèse a pu être réalisée, c’est avant tout parce qu’il m’a accordé sa confiance, ce que je n’oublierai jamais. Mostafa, merci de m’avoir consacré autant de temps. Jamais à court d’idées, tu as toujours su me guider vers les véritables priorités, et tu as ainsi permis que de nombreux projets aboutissent.

Je ne peux également que remercier ma co-directrice de thèse Evelyne Vigneau, au vu de sa grande contribution à mes travaux. Evelyne, grâce à ta rigueur de travail, ainsi que ta détermination pour que les recherches soient les plus abouties possibles, tu as porté cette thèse dans le bon sens. Tu m’as toujours poussé à faire mieux, et sans toi certains travaux n’auraient pas été aussi poussés.

Un grand merci à ma co-directrice de thèse Véronique Cariou, qui a toujours répondu présente quand j’ai eu besoin d’elle. Véronique, par ton investissement et tes conseils très précieux, tu as souvent su me diriger dans la bonne direction. Grâce notamment à ta rigueur bibliographique, tu as eu un apport considérable dans ces travaux.

Je voudrais remercier toute l’unité de recherche d’Oniris StatSC pour son excellent accueil au sein de l’établissement. Un merci tout particulièrement à Philippe Courcoux et Mohamed Hanafi qui ont accepté de m’aider sur plusieurs sujets.

Mon encadrant dans l’entreprise Addinsoft, Amaury Labenne, m’aura tant appris au cours de cette thèse que je ne peux que lui dire un très grand merci. Amaury, même quand tu étais débordé de travail, tu n’as jamais refusé de m’aider. Toute la découverte de l’environnement de programmation du logiciel XLSTAT s’est faite petit à petit, car tu as su me former sur la longueur afin que je puisse continuer mon travail de recherche en parallèle. Ta gestion aussi souple que sérieuse aura été vraiment bénéfique.

Merci à Thierry Fahmy, CEO d’Addinsoft, pour sa confiance accordée. Malgré tout le travail que représente la gestion d’une entreprise, il n’a jamais cessé de s’intéresser

à mon bien-être et à l'avancée de mes travaux. De plus, il n'a pas hésité à rendre possible ma participation aux conférences, ce qui m'a fait progresser dans différents domaines.

Je souhaite remercier toute l'équipe d'Addinsoft de m'avoir bien intégré. L'ambiance dans l'entreprise est vraiment très sympathique, et les temps de repas du midi sont toujours très agréables. L'équipe Stat' m'a permis de venir tous les jours avec le sourire, et travailler dans de si bonnes conditions n'a pas de prix.

Je tiens à apporter ma reconnaissance à Lise Bellanger, Pascal Schlich, Christian Derquenne et Robin Genuer, membres de mon comité de suivi individuel, pour leurs pertinents conseils. Leur regard extérieur a apporté un vrai plus à mes travaux. Ils n'ont pas hésité à se déplacer, que ce soit à Nantes ou à Bordeaux, et leurs présences auront été un réel plaisir.

Merci à Davide Giacalone de l'université de Southern Denmark d'avoir accepté de collaborer avec moi ainsi que de m'avoir autorisé à utiliser ses nombreux jeux de données de type Check-All-That-Apply.

J'aimerais remercier l'ANRT pour l'attribution de la bourse CIFRE sans laquelle cette thèse n'aurait pas été possible. Au vu de la bonne entente et du respect mutuel entre Oniris et Addinsoft, notamment sur mon temps de travail, je n'ai pu que constater les effets bénéfiques d'une thèse CIFRE.

Je tiens à remercier les professeurs du Master de statistique de l'université de Nantes pour leur formation de qualité, qui a été mon socle dans l'entreprise de ces travaux.

J'adresse un grand merci à à toute ma famille et ma belle-famille pour leurs soutiens et leurs encouragements tout au long de ce doctorat. Je remercie en particulier mes parents, qui m'ont toujours soutenu durant l'ensemble de mes études, ainsi que mon frère et mes grands-parents, qui, malgré leur éloignement géographique, se sont continuellement intéressés à l'avancée de ma thèse.

Enfin, je pense que tous les sacrifices réalisés par ma compagne, Marie, tout au long de mes études, méritent toute ma gratitude. En effet, elle aura déménagé pas moins de quatre fois en l'espace de six ans pour me suivre à travers la France !

Table des matières

1	Introduction	17
1.1	Position du problème	17
1.2	L'analyse sensorielle et les données de tableaux multiples	18
1.3	État de l'art et intérêts	19
1.3.1	Analyse exploratoire	19
1.3.2	Classification de tableaux	21
1.4	Structure du manuscrit	24
2	Analyse exploratoire de tableaux multiples par la méthode STATIS	27
2.1	Introduction	27
2.2	La méthode STATIS	27
2.3	Application en évaluation sensorielle	29
2.4	Exemples d'application	31
2.4.1	Données de Projective mapping ou Napping	31
2.4.2	Données de tri libre	33
2.5	Conclusion	37
3	Classification de tableaux : la méthode CLUSTATIS	39
3.1	Introduction	39
3.2	Principe de CLUSTATIS	40
3.3	Classification en écartant les tableaux atypiques	43
3.3.1	Classe « $K + 1$ »	43
3.3.2	Choix du seuil minimal de similarité ρ	44
3.4	Exemples d'application en analyse sensorielle	46
3.4.1	Projective mapping ou Napping	46
3.4.2	Tri libre	55
3.5	Conclusion	56

4	Données Check-All-That-Appl	59
4.1	Introduction	59
4.2	Accord entre les sujets	60
4.2.1	Accord entre deux sujets	60
4.2.2	Accord global	61
4.2.3	Tests de consistance	61
4.3	Analyse exploratoire : la méthode CATATIS	62
4.4	La méthode CLUSCATA	64
4.5	Classification en écartant les sujets atypiques	66
4.6	Exemples d'application	67
4.6.1	Données « fraises »	67
4.6.2	Autres jeux de données	72
4.7	Conclusion	75
5	Choix du nombre de classes	77
5.1	Introduction	77
5.2	Procédure générale	78
5.3	Étape 1 : une classe ou plus d'une classe ?	78
5.4	Étape 2 : combien de classes ?	80
5.4.1	Stratégie séquentielle de tests	80
5.4.2	Détermination du nombre de classes à l'aide d'indices	81
5.5	Adaptation aux données CATA	82
5.6	Application à des données simulées et réelles	83
5.6.1	Simulations	83
5.6.2	Données réelles	90
5.7	Conclusion	92
6	Implémentation informatique	95
6.1	Introduction	95
6.2	Package <i>R</i> : ClustBlock	96
6.2.1	Présentation du package	96
6.2.2	Fonctions	96
6.2.3	Jeux de données	98
6.2.4	Exemples d'utilisation	99
6.3	XLSTAT	103
6.4	Conclusion	103

7 Discussion et conclusion	105
Annexes	
A Démonstrations	111
A.1 Méthode STATIS	111
A.2 Méthode CLUSTATIS	112
B Valorisation des travaux de thèse	115
B.1 Publications	115
B.2 Communications orales	117
B.3 Développements	118
B.3.1 Logiciel <i>R</i>	118
B.3.2 XLSTAT	118

Liste des tableaux

2.1	Données chocolats : Coefficients RV entre les deux premières composantes de STATIS et DISTATIS avec celles obtenues au moyen de méthodes basées sur l'AFC, l'ACM et la MDS.	36
2.2	Données « chocolats » : Rapports de corrélation moyens associés aux deux premiers axes des diverses méthodes.	37
3.1	Données « smoothies » : Taille, homogénéité de chaque classe et homogénéité globale provenant de l'algorithme hiérarchique de CLUSTATIS.	48
3.2	Données « smoothies » : Taille, homogénéité de chaque classe et homogénéité globale obtenues en utilisant CLUSTATIS sans et avec la classe « $K + 1$ ».	49
3.3	Coefficients RV entre les tableaux compromis des trois classes.	51
3.4	Données « yaourts » : Comparaison, sur la base des indices d'homogénéité, de Proclustrees et CLUSTATIS (sans classe « $K + 1$ »), et de SCR et CLUSTATIS (avec classe « $K + 1$ »).	52
3.5	Données « arômes » : Taille, homogénéité de chaque classe et homogénéité globale obtenues en utilisant CLUSTATIS avec et sans classe « $K + 1$ ».	55
4.1	Données « fraises » : Pseudo p -value du test de consistance appliqué à chacun des attributs.	68
4.2	Données « fraises » : Homogénéité de chacune des classes et homogénéité globale obtenues au moyen de CLUSCATA avec l'option « $K + 1$ ».	71
4.3	Description des jeux de données étudiés.	72
4.4	Principaux résultats des diverses stratégies d'analyse.	74
5.1	Plan de simulation de données de Projective mapping/Napping. Chaque cas est également testé avec des données équilibrées.	84
5.2	Pourcentage de résultats corrects des tests et indices pour les données de Projective mapping/Napping simulées pour une, deux et trois classes. Coefficient RV moyen intra-classes égal à 0.5. Pour deux classes, coefficient RV entre les tableaux de base égal à 0.3.	86

5.3	Affectations des produits P1, . . . , P14 aux groupes par le sujet support et par 10 sujets simulés avec un niveau de consistance $p = 0.7$.	87
5.4	Plan de simulation de données de tri libre. Chaque cas est également testé avec des données équilibrées.	88
5.5	Données de tri libre : Pourcentage de nombres corrects de classes lorsqu'une seule classe est simulée.	89
5.6	Plan de simulation de données CATA. Chaque cas est également testé avec des données équilibrées.	89
5.7	Données CATA : Pourcentages de nombres corrects de classes du Test 1 et des indices dans les différentes conditions définies par le Tableau 5.6, excepté le cas où le coefficient s moyen intra-classes est égal à 0.7.	91
5.8	Nombres de classes préconisés par le Test 2 et l'Indice H sur cinq jeux de données réels provenant d'épreuves de Projective mapping/Napping.	92
5.9	Résultats du Test 2 et de l'Indice H sur trois jeux de données réels provenant d'épreuves de tri libre.	92
5.10	Résultats du Test 1 et de l'Indice H sur neuf jeux de données réels provenant d'épreuves CATA.	93

Table des figures

1.1	Données de tableaux multiples.	19
2.1	Description de la méthode STATIS.	28
2.2	Données « smoothies » : Poids de chacun des sujets obtenu par STATIS.	32
2.3	Données « smoothies » : Représentation des produits sur la base des deux premières composantes du tableau compromis de STATIS.	33
2.4	Données « smoothies » : Représentation des produits sur la base des deux premières composantes calculées au moyen de l'APG et de l'AFM.	34
2.5	Données « smoothies » : Pourcentages d'inertie des tableaux de données originaux expliquée par les composantes successives dérivées de STATIS, de l'APG et de l'AFM.	35
2.6	Données « chocolats » : Représentation des marques de chocolat sur la base des deux premières composantes de STATIS (gauche) et DISTATIS (droite).	36
2.7	Données « chocolats » : Représentation graphique des proximités entre les méthodes de tri libre sur les deux premiers axes de l'analyse MDS.	37
3.1	Représentation schématique de trois classes de tableaux. Les tableaux non circonscrits sont considérés comme atypiques.	45
3.2	Données « smoothies » : Dendrogramme et variation du critère D lors de l'algorithme hiérarchique de CLUSTATIS.	47
3.3	Données « smoothies » : Coefficients RV entre les matrices des produits scalaires associées aux sujets S1 à S24 avec le tableau compromis de leur classe.	48
3.4	Données « smoothies » : Évolution de l'indice d'homogénéité globale (gauche) et du nombre de sujets dans la classe « $K + 1$ » (droite) en fonction du seuil ρ .	49
3.5	Données « smoothies » : Représentation des produits sur la base des deux premières composantes de STATIS dans chaque classe construite par CLUSTATIS avec l'option « $K + 1$ ».	50

3.6	Données « yaourts » : Dendrogramme et variation du critère D provenant de l'algorithme hiérarchique de CLUSTATIS.	52
3.7	Données « yaourts » : Représentation des produits sur la base des deux premières composantes de STATIS de chaque classe.	53
3.8	Exemple de données simulées pour la classe 1 (gauche) et la classe 2 (droite). Les chiffres 1 à 9 correspondent aux produits simulés. Les points noirs représentent les produits initiaux.	54
3.9	Données « arômes » : Dendrogramme et variation du critère D provenant de l'algorithme hiérarchique de CLUSTATIS.	56
3.10	Données « arômes » : Représentation des produits sur la base des deux premières composantes de STATIS associées à la classe 1 obtenue par l'algorithme de partitionnement de CLUSTATIS sans option « $K + 1$ » (gauche) et avec option « $K + 1$ » (droite).	57
4.1	Description de la méthode CATATIS.	63
4.2	Données « fraises » : Poids des sujets donnés par CATATIS.	69
4.3	Données « fraises » : Représentation du tableau compromis de CATATIS sur les deux premiers axes de l'AFC.	70
4.4	Données « fraises » : Ellipses de confiance à 95% obtenues par l'AFC sur le tableau de fréquences (gauche) et par l'AFC sur le tableau compromis obtenu par CATATIS (droite).	71
4.5	Données « fraises » : Dendrogramme obtenu par l'algorithme hiérarchique de CLUSCATA (gauche) et variation du critère d'agrégation au cours de cet algorithme (droite).	72
4.6	Données « fraises » : Représentation du tableau compromis de chacune des quatre classes sur les deux premiers axes de l'AFC.	73
5.1	Procédure générale de détermination du nombre de classes.	79
5.2	Évolution de l'indice d'homogénéité I en fonction du niveau de consistance p .	88
6.1	Architecture des fonctionnalités du package ClustBlock.	97
6.2	Représentation du tableau compromis obtenu par la commande <code>plot(st, col=rep("darkblue", 8), cex=1.3, axes=c(1,3))</code> .	100

Liste des algorithmes

1	Algorithme hiérarchique de CLUSTATIS.	41
2	Algorithme de partitionnement de CLUSTATIS.	42
3	Algorithme de partitionnement de CLUSTATIS avec option « $K + 1$ ».	44
4	Algorithme hiérarchique de CLUSCATA.	65
5	Algorithme de partitionnement de CLUSCATA.	66
6	Algorithme de partitionnement de CLUSCATA avec option « $K + 1$ ».	67

Chapitre 1

Introduction

1.1 Position du problème

Dans plusieurs domaines d'application, les données se présentent sous forme de tableaux multiples appelés aussi données multi-blocs. Par exemple, dans le domaine de la santé et de l'environnement, le praticien peut être amené à collecter des données relatives à l'exposition d'un ensemble d'individus à des agents chimiques ou biologiques, des données relatives aux effets associés à cette exposition, et, parallèlement, des données phénotypiques générées par des plateformes à haut débit. En chimiométrie, il arrive souvent que des produits ou des systèmes soient caractérisés par des données instrumentales issues de différents types d'analyses (spectroscopie infra-rouge, analyse d'images, mesures physico-chimiques, ...). En analyse sensorielle, domaine qui nous intéresse particulièrement, plusieurs procédures d'évaluation conduisent à la collecte de données qui peuvent être présentées sous forme de plusieurs tableaux. Certaines de ces procédures sont détaillées dans ce manuscrit.

La recherche dans le cadre de l'analyse de données structurées en tableaux multiples a été très active ces trente dernières années (De Roover *et al.*, 2012). Une pléthore de méthodes ont été proposées soit dans une visée exploratoire soit dans une visée prédictive. Dans la première famille de méthodes, nous pouvons citer l'Analyse Procrustéenne Généralisée (Gower, 1975), STATIS (Lavit *et al.*, 1994), l'Analyse Factorielle Multiple (Pagès, 2005), ComDim (Qannari *et al.*, 2000). Dans la deuxième famille, nous pouvons évoquer à titre d'exemple la régression PLS multi-blocs (MacGregor *et al.*, 1994), P-ComDim (Carriou *et al.*, 2018). Cependant, le problème de classification de tableaux multiples, qui fait l'objet de cette thèse, a été très peu abordé. Pourtant, ce problème devient de plus en plus étendu car les situations où l'utilisateur est conduit à collecter un nombre relativement

important de tableaux de données deviennent très fréquentes. Il est à noter que si des travaux ont été réalisés sur la classification des individus avec ce type de données (Niang et Ouattara, 2019), nous nous focalisons sur la classification des tableaux.

1.2 L’analyse sensorielle et les données de tableaux multiples

Plusieurs procédures d’évaluation sensorielle conduisent à des données de tableaux multiples. Généralement, chaque tableau de données est associé à un juge « naïf » ou entraîné, à un consommateur, *etc.*

Le profil conventionnel et le profil libre (Williams et Langron, 1984) sont les exemples les plus simples : des juges plus ou moins entraînés attribuent à chaque produit des notes d’intensité de différents descripteurs sensoriels (*e.g.* intensité de l’odeur, saveur sucrée, ...). La différence entre ces deux types de profils réside dans le choix des descripteurs. Pour le cas du profil conventionnel, la liste des descripteurs est la même pour tous les juges, tant et si bien que les données peuvent être présentées sous forme d’un ensemble de tableaux qui s’apparentent en lignes (produits) et en colonnes (descripteurs sensoriels) ; chaque tableau étant associé à un juge. Pour le cas du profil libre, comme son nom l’indique, la liste des descripteurs peut être différente d’un juge à un autre. Dans ce cas, les données peuvent être présentées sous forme d’un ensemble de tableaux s’apparentant en lignes mais pas en colonnes (Figure 1.1).

L’épreuve dite Projective mapping (Risvik *et al.*, 1994), appelée aussi Napping (Pagès, 2005), est une épreuve sensorielle consistant à demander aux sujets de positionner les produits sur une feuille de papier de type A4. L’instruction donnée est que, si deux produits sont perçus comme similaires, le sujet devrait les positionner proche l’un de l’autre. Les coordonnées en abscisse et ordonnée sur la feuille de papier des produits évalués sont récupérées pour former un tableau par sujet où les lignes réfèrent aux produits et les colonnes réfèrent aux coordonnées de ces produits. Des données de tableaux multiples sont ainsi obtenues.

L’épreuve de tri libre (ou Free Sorting, Faye *et al.*, 2004) est très utilisée. Dans cette épreuve, l’instruction donnée à chaque sujet est de former des groupes de produits sur la base de leurs similarités perçues. De fait, chaque sujet procure une partition des produits qui pourrait être exprimée sous forme d’un tableau disjonctif complet indiquant l’appar-

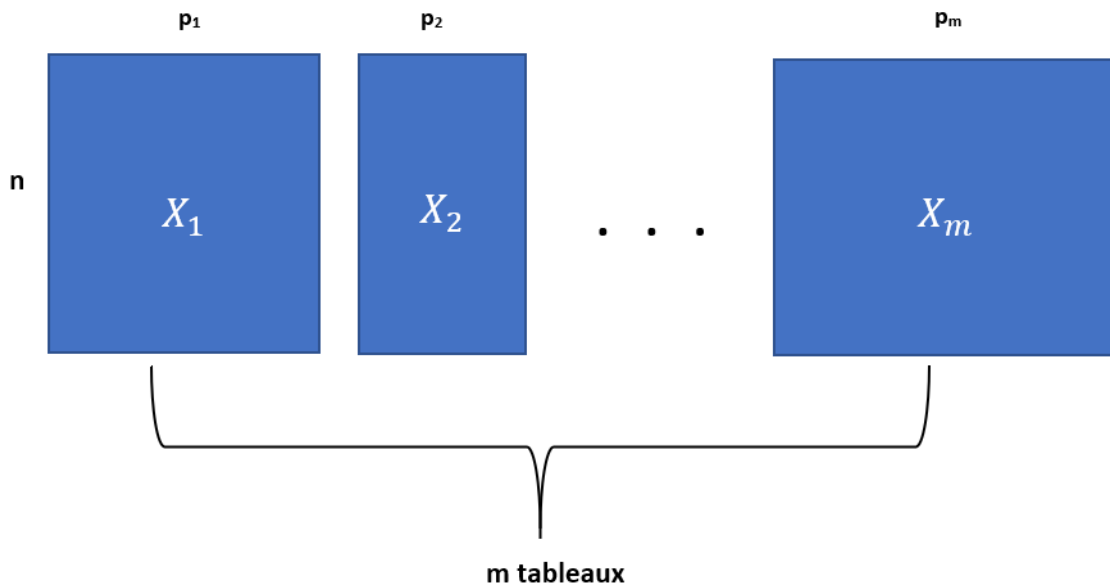


FIGURE 1.1 : Données de tableaux multiples.

tenance des produits aux différents groupes de la partition. En définitive, nous obtenons des données de tableaux multiples.

Une autre épreuve sensorielle largement discutée dans ce manuscrit se nomme Check-All-That-Apply (CATA, [Ares et Jaeger, 2015](#)). Il s'agit pour chaque sujet de cocher parmi une liste d'attributs ceux qui s'appliquent à chaque produit. Des données binaires, pouvant être structurées en tableaux multiples apparentés en lignes (produits) et en colonnes (attributs) sont ainsi obtenues : pour un produit donné, la valeur 1 indique qu'un attribut a été coché, et la valeur 0 qu'il n'a pas été coché. Pour en savoir plus sur toutes ces techniques d'analyse sensorielle, nous référons au livre de [Delarue et al. \(2014\)](#).

1.3 État de l'art et intérêts

1.3.1 Analyse exploratoire

En présence de données de tableaux multiples, un objectif majeur est d'explorer la structure des données et d'investiguer la typologie des individus (*i.e.* les lignes des tableaux). Dans ce cadre, STATIS ([Lavit et al., 1994](#)) est une méthode tout à fait adaptée, ce qui lui vaut sa popularité dans beaucoup de domaines, dont en particulier l'évaluation sensorielle ([Schlich, 1996](#); [Abdi et al., 2007](#)). Nous présentons dans le deuxième chapitre une nouvelle formulation de la méthode STATIS basée sur un nouveau critère d'optimisation et nous discutons son application en analyse sensorielle, notamment pour traiter

les données qualitatives issues d'épreuves telles que le tri libre. Sur la base d'une étude de cas, les résultats de la méthode STATIS sont comparés à ceux de méthodes alternatives : l'Analyse Factorielle Multiple (AFM, Pagès, 2005) et l'Analyse Procrustéenne Généralisée (APG, Gower, 1975). Fondamentalement, la méthode STATIS consiste à déterminer un tableau compromis sur la base duquel la typologie des individus est explorée. De manière astucieuse, STATIS limite l'impact des tableaux atypiques : les tableaux sont combinés en tenant compte de leurs accords mutuels. Ainsi, les tableaux atypiques sont pondérés par des poids relativement faibles. Cependant, cette stratégie s'avère insuffisante en présence de plusieurs classes de tableaux. Dans ce cas, l'idée qui semble pertinente est d'identifier les classes de tableaux et opérer la méthode STATIS à l'intérieur de chaque classe. Ceci est schématiquement l'objectif de la méthode CLUSTATIS, qui est développée dans le troisième chapitre.

Comme indiqué plus haut, dans une épreuve de tri libre, les sujets qui participent à l'épreuve sont chargés de partitionner les produits. Le nombre de groupes peut différer d'un sujet à l'autre. Plusieurs façons de coder les données sont proposées et conduisent à différentes méthodes d'analyse. Ces méthodes sont basées sur le Multidimensional Scaling (MDS, Faye *et al.*, 2004; Lawless *et al.*, 1995) ou sur l'analyse des correspondances simples et multiples (Cadoret *et al.*, 2009; Faye *et al.*, 2011; Qannari *et al.*, 2010; Cariou et Qannari, 2018). La méthode DISTATIS (Abdi *et al.*, 2007), à l'instar de l'approche présentée dans ce manuscrit, s'articule autour de la méthode STATIS. Pour cette raison, une étude de comparaison est effectuée entre les deux approches.

Pour le cas des données CATA, la méthode d'analyse usuelle est d'effectuer une Analyse Factorielle des Correspondances (AFC, Greenacre, 2017) sur le tableau de fréquences qui croise les produits et les attributs pour tout le panel de sujets (Meyners *et al.*, 2013). L'approche présentée ici, appelée CATATIS, utilise également l'AFC pour caractériser les produits. Cependant, l'AFC est effectuée sur un tableau compromis obtenu en combinant les tableaux individuels des sujets tout en tenant compte de leurs accords mutuels. En d'autres termes, les sujets qui ont un faible accord avec le reste du panel sont sous-pondérés afin de réduire leur impact sur les résultats de l'analyse. Il est clair que cette approche présente une similitude avec la méthode STATIS. La principale différence est que CATATIS est basée sur les données originales, alors que STATIS est basée sur les matrices des produits scalaires associées aux tableaux. Fondamentalement, nous nous appuyons sur le fait que les tableaux correspondants aux données des divers sujets se rapportent aux mêmes entités en termes de lignes (*i.e.*, les produits) et de colonnes (*i.e.*, les attributs).

Une comparaison est faite entre la méthode usuelle d'analyse des données CATA et CATATIS.

Pour l'épreuve CATA, comme pour les autres épreuves en analyse sensorielle, les utilisateurs sont intéressés par l'évaluation de l'accord entre les sujets. L'importance de cette évaluation a été soulignée par Worch et Piqueras-Fiszman (2015). Ces auteurs ont proposé d'utiliser des stratégies fondées sur l'Analyse Factorielle Multiple (AFM) et sur les tests de McNemar (Lachenbruch, 2014). Cependant, Meyners *et al.* (2016) ont remarqué que si le test de McNemar permet bien d'évaluer la significativité de la différence entre les produits, il est peu approprié pour explorer les accords entre les sujets. Nous proposons une procédure pour vérifier si un accord significatif entre les sujets existe, et ce de manière globale ou par attribut.

1.3.2 Classification de tableaux

Dans le cas de données de tableaux multiples, la classification de tableaux a déjà été abordée. Dans le cadre d'épreuves dites de profils sensoriels, Dahl et Næs (2004) ont proposé de segmenter un ensemble de tableaux (de nombre généralement peu élevé) afin de détecter des classes de juges, et, surtout, d'identifier des juges atypiques. Pour cela, ces auteurs ont préconisé une classification hiérarchique sur la base de la matrice des distances de Procrustes entre les tableaux. Pour la détermination de tableaux atypiques, l'hypothèse de travail est que ces tableaux devraient apparaître dans l'arbre hiérarchique en tant que branches excentrées. Nous pouvons citer également les deux articles publiés par Cariou et Wilderjans (Wilderjans et Cariou, 2016; Cariou et Wilderjans, 2018), qui décrivent une classification de tableaux ternaires (*i.e.*, apparentés par lignes et par colonnes) basée sur un modèle Parafac. Berget *et al.* (2019) ont opéré une segmentation de plusieurs tableaux dans le cas spécifique de l'épreuve Projective mapping/Napping via plusieurs stratégies. Nous reviendrons sur leurs approches en les comparant à la stratégie que nous proposons.

La question spécifique de la classification des sujets impliqués dans une épreuve de tri libre a été abordée par Lawless (1989), qui a suggéré de réaliser une classification ascendante hiérarchique sur les coordonnées factorielles obtenues grâce à une analyse MDS non métrique. D'autre part, Courcoux *et al.* (2014) ont calculé un critère global basé sur l'indice de Rand ajusté (Hubert et Arabie, 1985), qui est la version corrigée de l'indice de Rand (Rand, 1971). Ce critère permet d'évaluer l'accord entre les sujets avec une partition consensus des produits déterminée par l'algorithme. Par la suite, une stratégie de regroupement des sujets est mise en place. Elle combine une classification ascendante hiérarchique suivie d'une procédure de partitionnement. Dans un article plus récent, Ca-

riou et Qannari (2018) ont étudié l'accord entre les sujets impliqués dans une épreuve de tri libre en calculant une matrice de cooccurrences qui donne, pour chaque paire de sujets, le nombre de paires de produits que ces sujets ont placé dans le même groupe. Cette matrice de cooccurrences fait l'objet d'une AFC ainsi que d'une classification hiérarchique.

Nous proposons une approche de classification de tableaux multiples basée sur un critère d'optimisation, appelée CLUSTATIS. Le principe de cette méthode de classification est de déterminer simultanément les classes de tableaux et les compromis associés aux différentes classes de manière à ce que les tableaux de chaque classe soient fortement liés à leur compromis. Cette stratégie d'analyse permet également de procurer des indicateurs statistiques de nature à caractériser la pertinence de la classification. CLUSTATIS pourrait s'apparenter à la méthode de classification de variables appelée « Classification autour des Variables Latentes » (CLV ; Vigneau et Qannari, 2003). En effet, cette dernière méthode consiste à déterminer des classes de variables de manière à ce que les variables de chaque classe soient très liées à une composante latente associée à cette classe. CLUSTATIS consiste en un algorithme hiérarchique et un algorithme de partitionnement. Ces deux stratégies visent à optimiser le même critère. Elles peuvent être exécutées indépendamment ou en combinaison dans le but d'obtenir une solution encore meilleure que celle obtenue par l'une ou l'autre des deux stratégies. Dans le cadre de l'analyse sensorielle, il est à noter que CLUSTATIS peut être appliquée dans plusieurs situations, dont, en particulier, les données issues d'une épreuve de Projective mapping/Napping ou de tri libre. Des exemples d'application sont présentés dans ce manuscrit.

Nous nous sommes également intéressés à l'épreuve d'évaluation sensorielle appelée CATA. Comme nous l'avons déjà indiqué, cette méthode conduit à des tableaux de données binaires qui s'apparentent en lignes et en colonnes. Les sujets d'une épreuve CATA peuvent avoir des perceptions différentes des produits. Ainsi, la segmentation de ces sujets est d'un intérêt primordial. Cette question n'a pas été suffisamment abordée dans la littérature. Cependant, puisque les données CATA sont intrinsèquement binaires (un attribut s'applique ou ne s'applique pas à un produit), la première idée pouvant venir à l'esprit est d'utiliser l'une des nombreuses mesures de proximité dédiées pour ce type de données (Warrens *et al.*, 2008) afin d'évaluer la (dis)similarité entre les sujets. Ensuite, le regroupement des sujets pourrait se faire sur la base de la matrice de (dis)similarité ainsi obtenue. Nous proposons une stratégie de classification des sujets appelée CLUSCATA. Cette méthode d'analyse peut être vue comme une extension de CATATIS puisqu'elle s'appuie sur un critère étroitement lié à celui de cette dernière analyse. Le but de CLUS-

1.3 État de l'art et intérêts

CATA est de trouver des classes homogènes de sujets au sens où les tableaux associés aux sujets d'une classe donnée sont aussi proches que possible d'un tableau compromis de cette classe. Ce dernier tableau est calculé à l'aide de CATATIS. Dans un second temps, le tableau compromis de chaque classe de sujets est soumis à l'AFC afin de décrire les relations entre les produits et les attributs. En fait, CLUSCATA est une adaptation de la méthode CLUSTATIS pour le cas particulier des données CATA. La plus grande différence est que CLUSTATIS s'intéresse à la classification de plusieurs tableaux qui se rapportent aux mêmes individus (*i.e.*, lignes), mais pas nécessairement aux mêmes variables (*i.e.*, colonnes).

Il existe très souvent des tableaux atypiques ne se conformant à aucune classe déterminée par la méthode de classification. Il n'en demeure pas moins vrai que ces tableaux sont assignés à l'une ou l'autre des classes. Par conséquent, l'homogénéité de l'ensemble des classes peut être affectée et la performance de l'algorithme de classification peut se détériorer (Dave, 1991). Une stratégie d'identification des tableaux atypiques et leur mise à l'écart est discutée et appliquée à des données issues d'épreuves de Projective mapping/Napping et de tri libre. En fait, nous suivons une stratégie d'analyse similaire à celle décrite par Vigneau *et al.* (2016) dans le cadre de la classification de variables. En analyse sensorielle, les sujets atypiques que nous cherchons à identifier peuvent correspondre à des sujets à part, sujets sans évaluation claire, sujets indécis, *etc.* Ce genre de comportement a été discuté dans le contexte des études de préférences par Westad *et al.* (2004). La stratégie d'identification des tableaux atypiques est d'inclure dans le critère d'optimisation associé à CLUSTATIS une classe supplémentaire « $K + 1$ » en plus des K classes principales. L'objet de cette classe « $K + 1$ » est de contenir les tableaux atypiques. Dave (1991), qui a été le premier à formaliser le concept de la stratégie « $K + 1$ », nomme cette classe supplémentaire « groupe de bruit ». Une caractéristique notable de cette stratégie est qu'elle est simple et intuitive. Elle suppose, dans un premier temps, de déterminer un seuil qui reflète l'accord minimal qu'un tableau devrait avoir avec le compromis de sa classe. Tout tableau ayant un degré d'accord avec le compromis de sa classe inférieur à ce seuil est placé dans la classe « $K + 1$ ». Une contribution notable dans ce manuscrit est le choix automatique du seuil de similarité minimal. Une adaptation de cette procédure pour la classification des sujets dans une épreuve CATA avec la méthode CLUSCATA est également proposée.

De manière générale, le choix du nombre de classes est une question complexe (Hardy, 1996). Généralement, ce choix est basé sur des outils graphiques telle que la structure du

dendrogramme (Bellanger, 2015) ou les graphiques appelés « silhouettes » (Rousseeuw, 1987). À notre connaissance, il n'existe pas d'articles consacrés à cette question difficile dans les approches de classification de tableaux. Le choix peut être fait empiriquement par le praticien en examinant l'évolution du critère de regroupement. Une telle stratégie s'apparente à un scree-test et a été proposée dans le cadre de la classification de données de type trois voies (Cariou et Wilderjans, 2018). Par ailleurs, il existe des critères basés sur la variation d'inertie intra-classes (Kothari et Pitts, 1999; Gupta *et al.*, 2018), des approches de test d'hypothèses telle que la statistique Gap (Tibshirani *et al.*, 2001; Cariou *et al.*, 2009) ou des méthodes directes pour lesquelles le nombre de classes est intrinsèquement lié à la stratégie de classification (Gialampoukidis *et al.*, 2019). Dans ce manuscrit, l'objectif est d'adapter certaines de ces approches au contexte de la classification de tableaux à partir de critères d'inertie ou de tests d'hypothèses, puis de comparer leurs performances sur des données réelles ou simulées.

1.4 Structure du manuscrit

Dans le chapitre 2, dédié à l'analyse exploratoire de tableaux multiples, la méthode STATIS est présentée. Un pré-traitement adapté afin de pouvoir appliquer STATIS à des données de tri libre est ensuite introduit. Des exemples d'application avec des données de Projective mapping/Napping et de tri libre sont discutés. Les résultats de l'approche que nous préconisons sont comparés à ceux d'autres méthodes pour chaque type de données.

Le chapitre 3, central à ce manuscrit, introduit la méthode CLUSTATIS dédiée à la classification de tableaux. Une extension de cette méthode en ajoutant une classe supplémentaire « $K + 1$ » est également proposée. Cette dernière option permet de mettre de côté les tableaux atypiques (*i.e.*, les tableaux ne se conformant de manière claire à aucune des classes). Des exemples d'application sur des données de Projective mapping/-Napping ainsi que sur des données de tri libre sont présentés afin de montrer la pertinence de la méthode. Des comparaisons avec des méthodes existantes, basées sur des simulations et une étude de cas, sont également réalisées.

Dans le chapitre 4, le cas des données CATA est traité, partant de l'analyse exploratoire (avec la méthode CATATIS) jusqu'à la classification (avec la stratégie CLUSCATA), en passant par la mise à l'écart des sujets atypiques. Des tests de consistance des sujets sont également proposés. Ils permettent de tester si un accord significatif entre les sujets existe de manière globale et par attribut. Des exemples d'application sont présentés.

1.4 Structure du manuscrit

Le chapitre 5 apporte des préconisations sur le choix du nombre de classes pour les méthodes CLUSTATIS et CLUSCATA. Des tests et des indices sont proposés puis comparés sur la base de simulations et de données réelles. De plus, les simulations permettent de souligner davantage la pertinence de CLUSTATIS et CLUSCATA.

Le chapitre 6 est dédié aux développements informatiques réalisés durant la thèse afin de rendre les méthodes proposées accessibles à tous. Le package *R* (R Core Team, 2020) nommé ClustBlock (Llobell *et al.*, 2020) est présenté et des exemples d'utilisation sont fournis. De plus, des développements réalisés dans le logiciel XLSTAT (Addinsoft, 2020) sont également esquissés.

Le chapitre 7 est consacré aux discussions, conclusions et perspectives.

Chapitre 2

Analyse exploratoire de tableaux multiples par la méthode STATIS

2.1 Introduction

Comme indiqué dans le chapitre précédent, dans le cadre de données de tableaux multiples, un des objectifs est d'investiguer la typologie des individus. Le plus souvent, cette investigation se fait sur la base de cartes factorielles. Un autre but est de déterminer si les tableaux (sujets sans le cas de l'analyse sensorielle) présentent de fortes similitudes entre eux. Dans le cadre de la méthode STATIS, cette dernière étape s'appelle « analyse inter-structures ». Cette analyse consiste à explorer la similarité entre les tableaux de données.

La méthode STATIS est d'abord présentée, puis un pré-traitement afin d'analyser les données de tri libre par cette stratégie est introduit. Enfin, des exemples d'application en analyse sensorielle sont discutés, tout en comparant les résultats à des méthodes usuelles.

2.2 La méthode STATIS

Considérons m tableaux $\mathbf{X}_1, \dots, \mathbf{X}_m$ supposés être centrés. STATIS est basée sur les matrices des produits scalaires associées aux tableaux de données. Ces matrices sont déterminées comme suit : $\mathbf{W}_1 = \mathbf{X}_1 \mathbf{X}_1^\top, \dots, \mathbf{W}_m = \mathbf{X}_m \mathbf{X}_m^\top$. Il s'agit de m matrices symétriques de taille $n \times n$ qui reflètent la configuration spatiale des individus puisque la l^{e} ligne et la j^{e} colonne d'une matrice indique le produit scalaire entre les deux individus correspondants. Il est recommandé de standardiser les matrices $\mathbf{W}_i$ de manière à ce que leur norme soit égale à 1. Cette opération est réalisée en divisant chaque matrice $\mathbf{W}_i$ par

sa norme de Frobenius. Nous pouvons rappeler que la norme de Frobenius d'une matrice $\mathbf{A}$ est donné par $\|\mathbf{A}\| = \sqrt{\text{trace}(\mathbf{A}\mathbf{A}^\top)} = \sqrt{\sum_l \sum_j a_{lj}^2}$, où a_{lj} est la $(l, j)^e$ entrée de la matrice $\mathbf{A}$.

Le coefficient RV (Robert et Escoufier, 1976), très populaire dans beaucoup de domaines et en particulier en analyse sensorielle (El Ghaziri et Qannari, 2015; Næs et al., 2017), est au cœur de la méthode STATIS. Ce coefficient reflète la similarité entre deux tableaux, varie entre 0 et 1 en augmentant avec la similarité entre les deux tableaux considérés. Pour deux matrices des produit scalaires $\mathbf{W}_i$ et $\mathbf{W}_s$, $RV(\mathbf{W}_i, \mathbf{W}_s)$ est défini par $\frac{\text{trace}(\mathbf{W}_i \mathbf{W}_s)}{\sqrt{\text{trace}(\mathbf{W}_i \mathbf{W}_i)} \sqrt{\text{trace}(\mathbf{W}_s \mathbf{W}_s)}}$.

La méthode STATIS vise à déterminer un tableau compromis $\mathbf{W}$ comme étant une somme pondérée des matrices $\mathbf{W}_i$: $\mathbf{W} = \alpha_1 \mathbf{W}_1 + \dots + \alpha_m \mathbf{W}_m$. Comme indiqué sur la Figure 2.1, les pondérations $\alpha_1, \dots, \alpha_m$ sont les composantes du premier vecteur propre de la matrice des coefficients RV entre les différents tableaux. De fait, le coefficient α_i est relativement grand si $\mathbf{W}_i$ est en accord (coefficients RV élevés) avec les autres matrices. La décomposition spectrale de $\mathbf{W}$ permet d'écrire $\mathbf{W} = \mathbf{C}\mathbf{C}^\top$. La matrice $\mathbf{C}$ est la configuration compromis des tableaux de données $\mathbf{X}_1, \dots, \mathbf{X}_m$, et peut être utilisée pour représenter les individus sur une carte factorielle afin d'explorer leurs similarités.

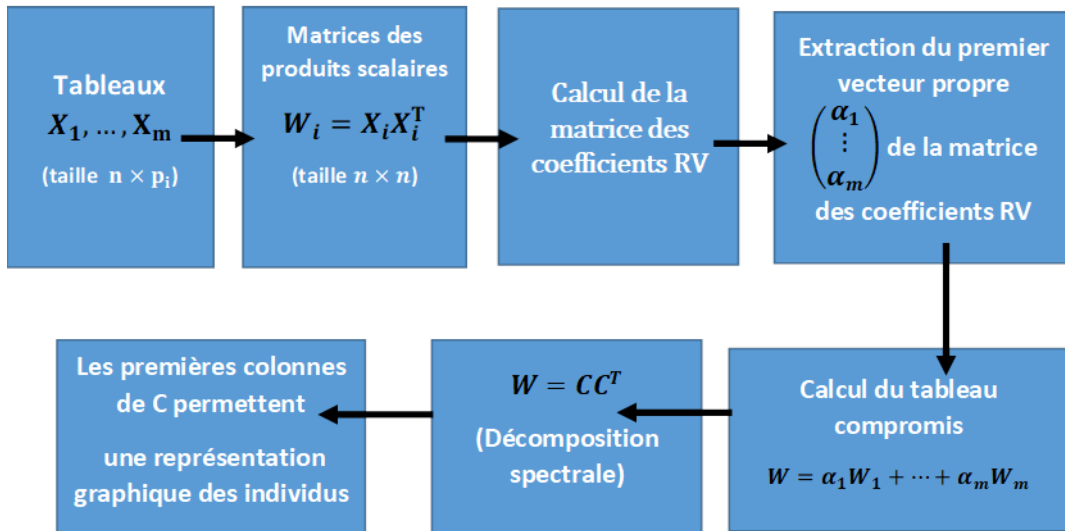


FIGURE 2.1 : Description de la méthode STATIS.

Formellement, la méthode STATIS cherche à trouver des poids positifs $\alpha_1, \dots, \alpha_m$ et un tableau compromis $\mathbf{W}$ de manière à minimiser le critère original suivant (la démonstration

2.3 Application en évaluation sensorielle

que ce critère revient bien à la méthode STATIS se trouve en Annexe [A](#) :

$$F = \sum_{i=1}^m \|\mathbf{W}_i - \alpha_i \mathbf{W}\|^2 \quad (2.1)$$

Ainsi, STATIS vise à minimiser la distance des matrices des produits scalaires au tableau compromis. Comme contrainte de détermination, nous supposons que $\sum_{i=1}^m \alpha_i^2 = 1$.

Des propriétés supplémentaires sont exposées ci-dessous. La plupart de ces propriétés sont déjà connues (voir par exemple [Robert et Escoufier, 1976](#); [Glaçon, 1981](#)). Ces propriétés sont démontrées dans l'Annexe [A.1](#). Nous notons par λ_1 la première valeur propre de la matrice des coefficients RV.

- (i) $\|W\|^2 = \sum_{i=1}^m RV^2(\mathbf{W}_i, \mathbf{W}) = \lambda_1$
- (ii) $\alpha_i = \frac{\text{trace}(\mathbf{W}_i \mathbf{W})}{\|\mathbf{W}\|^2} = \frac{RV(\mathbf{W}_i, \mathbf{W})}{\sqrt{\lambda_1}}$
- (iii) $\sum_{i=1}^m \|\mathbf{W}_i - \alpha_i \mathbf{W}\|^2 = m - \sum_{i=1}^m RV^2(\mathbf{W}_i, \mathbf{W}) = m - \lambda_1$
- (iv) $\sum_{i=1}^m \|\mathbf{W}_i\|^2 = m = \|\mathbf{W}\|^2 + \sum_{i=1}^m \|\mathbf{W}_i - \alpha_i \mathbf{W}\|^2 = \lambda_1 + \sum_{i=1}^m \|\mathbf{W}_i - \alpha_i \mathbf{W}\|^2$

Ces propriétés indiquent que la méthode STATIS recherche à maximiser la similarité (coefficient RV) entre les matrices $\mathbf{W}_i$ et le tableau compromis $\mathbf{W}$. La propriété [\(iv\)](#) établit que la variation totale mesurée par $\sum_{i=1}^m \|\mathbf{W}_i\|^2 = m$ peut être décomposée entre la variation expliquée par le tableau compromis (*i.e.*, $\|\mathbf{W}\|^2 = \lambda_1$) et la variation résiduelle (*i.e.*, $\sum_{i=1}^m \|\mathbf{W}_i - \alpha_i \mathbf{W}\|^2$). Ainsi, l'indice $I = \frac{\|\mathbf{W}\|^2}{\sum_{i=1}^m \|\mathbf{W}_i\|^2} = \frac{\lambda_1}{m}$ reflète la part de variation dans les différentes matrices $\mathbf{W}_i$ expliquée par le tableau compromis $\mathbf{W}$. Cet indice est compris entre $1/m$ et 1. Plus cet indice est élevé, plus l'accord entre les tableaux $\mathbf{X}_1, \dots, \mathbf{X}_m$ est grand.

2.3 Application en évaluation sensorielle

[Schlich \(1996\)](#) fut le premier à introduire la méthode STATIS en analyse sensorielle en soulignant notamment son intérêt par rapport à l'Analyse Procrustéenne Généralisée (APG, [Gower, 1975](#)), qui était alors la méthode phare dans le domaine. L'utilisation de STATIS en évaluation sensorielle a été initialement confinée au traitement statistique des profils sensoriels. Nous discutons de nouvelles applications de cette méthode en analyse sensorielle et nous comparons ses résultats sur la base d'études de cas avec des méthodes alternatives.

Les données de Projective mapping ou Napping

Comme cela est indiqué dans le chapitre précédent, les données issues d’une épreuve de « Projective mapping », dite aussi « Napping », peuvent être structurées sous forme d’un ensemble de tableaux de données. Chaque tableau est associé à un sujet participant à l’épreuve et consiste en deux colonnes correspondant aux coordonnées des produits sur la feuille de papier. Initialement introduite sous l’appellation « Projective mapping » (Risvik *et al.*, 1994), la méthode de traitement proposée fut l’APG. Par la suite, cette méthode d’évaluation sensorielle a été redécouverte par Pagès (2005) sous l’appellation « Napping ». La méthode de traitement de données qui fut préconisée est l’Analyse Factorielle Multiple (AFM). Une comparaison des deux méthodes d’analyse de données a été effectuée par Tomic *et al.* (2015). « Projective mapping » et « Napping » ont connu une grande popularité du fait de leur facilité d’utilisation et la disponibilité d’un bon environnement informatique, notamment grâce aux packages FactoMineR (Lê *et al.*, 2008) et SensoMineR (Lê et Husson, 2008) au sein du logiciel R (R Core Team, 2020). Dans la suite, nous appliquons la méthode STATIS sur les données issues d’une étude de de cas et nous comparons les résultats à ceux des méthodes APG et AFM.

Les données de tri libre

Nous rappelons que le tri libre est une épreuve de catégorisation consistant à demander à chaque sujet de former des groupes à partir des n produits qui lui sont présentés. L’instruction donnée aux sujets est que les produits qui sont perçus comme étant similaires devraient être dans un même groupe et les produits perçus comme étant différents devraient être dans différents groupes. Les seules restrictions sont de ne former qu’un seul groupe constitué par tous les produits, ou, à l’inverse, former autant de groupes que de produits à raison d’un produit par groupe. De fait, chaque sujet procure une partition des produits et, de ce point de vue, nous pouvons lui associer une variable qualitative opérant sur les produits et ayant pour modalités les groupes de la partition donnée par le sujet. Il n’est pas surprenant que l’Analyse des Correspondances Multiples (ACM) fut extensivement appliquée à ce type de données (Van der Kloot et Van Herk, 1991; Takane, 1981; Cadoret *et al.*, 2009; Qannari *et al.*, 2010). La stratégie DISTATIS (Abdi *et al.*, 2007) et la démarche d’analyse de données de tri libre proposée dans ce manuscrit partagent le fait que le codage disjonctif complet et la méthode STATIS sont utilisés. Fondamentalement, la principale différence est que, dans DISTATIS, les colonnes des tableaux disjonctifs complets ne sont pas normalisées comme nous le préconisons dans le présent document. En utilisant une normalisation courante en analyse des correspondances, nous adoptons, de

2.4 Exemples d'application

fait, la métrique du χ^2 , qui est bien adaptée aux données issues de variables qualitatives. Cependant, nous devons mettre au crédit de la méthode DISTATIS le fait qu'elle procure des outils et concepts intéressants pour l'étude des données de tri libre (Lahne *et al.*, 2018).

Supposons que le sujet i ($i = 1, \dots, m$) ait trié les n produits en p_i groupes. Nous désignons par $\mathbf{Y}_i$ ($n \times p_i$) le tableau disjonctif complet associé au sujet i . Chaque colonne de $\mathbf{Y}_i$ est une variable binaire qui est associée à un groupe spécifique des produits considérés. Elle indique pour chaque produit s'il appartient à ce groupe (auquel cas elle prend la valeur 1) ou non (auquel cas elle prend la valeur 0). Nous proposons de diviser chaque colonne de $\mathbf{Y}_i$ par la racine carrée de sa moyenne, qui représente la proportion de 1 dans cette colonne. Ce traitement est courant pour les données catégorielles, en particulier dans le cadre de l'analyse des correspondances (Saporta, 1976). Dans le cas des données de tri libre, il permet de ne pas favoriser un groupe comportant plus de produits que les autres. Dans ce qui suit, nous désignerons par $\mathbf{X}_i$ la matrice obtenue à partir de $\mathbf{Y}_i$ par normalisation des colonnes selon cette stratégie, suivie d'un centrage des colonnes. Par la suite, les matrices des produits scalaires $\mathbf{W}_i = \mathbf{X}_i \mathbf{X}_i^\top$ peuvent être calculées et divisées par leur norme, comme préconisé dans la section 2.2. Nous pouvons montrer que la norme de $\mathbf{W}_i$ est proportionnelle à $\sqrt{p_i - 1}$, où p_i est le nombre de groupes dans la partition des produits définie par le sujet i (Cazes *et al.*, 1976; Saporta, 1976). Cette procédure permet de tenir compte du fait que le nombre de groupes est différent d'un tableau à un autre. Il s'agit, en effet, d'un paramètre ayant un impact significatif sur l'analyse des données et qui est souvent négligé. Une fois ces transformations effectuées, la méthode STATIS peut être utilisée.

2.4 Exemples d'application

Afin d'illustrer l'application de STATIS en analyse sensorielle, deux exemples sont présentés. L'un concerne une épreuve de Projective mapping/Napping et l'autre des données de tri libre.

2.4.1 Données de Projective mapping ou Napping

Les données proviennent d'une expérimentation basée sur le procédé « Napping » réalisée par 24 étudiants d'Agrocampus-Rennes. Elles sont disponibles dans les packages *R* *SensoMineR* (Lê et Husson, 2008) et *ClustBlock* (Llobell *et al.*, 2020). À chaque étudiant, il a été demandé de positionner huit smoothies sur une feuille de papier.

Appliquée à ces données, la méthode STATIS procure des « poids » associés aux sujets qui reflètent leur accord avec le point de vue général du panel (Figure 2.2). Ainsi, les sujets S18 et S23 ont un poids plus grand que les sujets S2 et S4 (environ 3 fois plus). La première valeur propre λ_1 de la matrice des coefficients RV entre les tableaux associés aux sujets est égale à 10.19. Ainsi, l'indice d'homogénéité des sujets est de 42.4%, ce qui n'est pas très élevé.

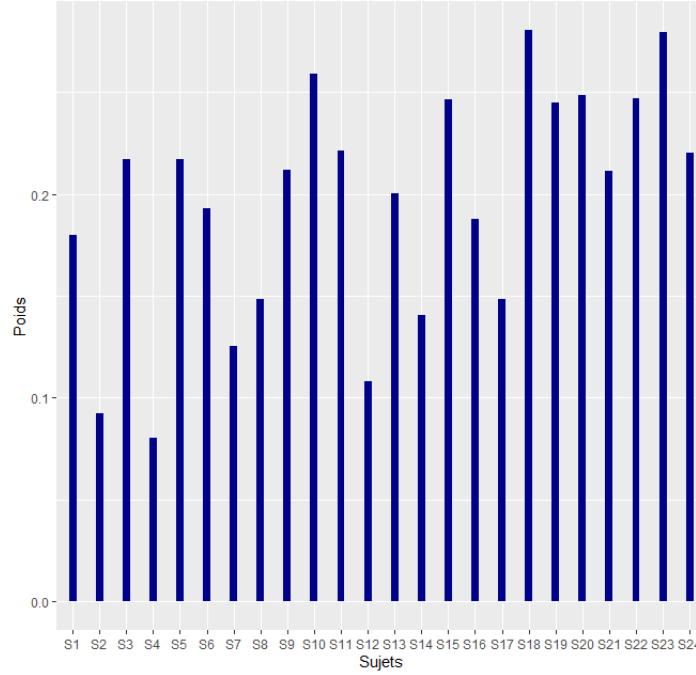


FIGURE 2.2 : Données « smoothies » : Poids de chacun des sujets obtenu par STATIS.

Afin d'avoir une représentation graphique des produits, la décomposition spectrale du tableau compromis, $\mathbf{W} = \mathbf{C}\mathbf{C}^\top$, est réalisée. La Figure 2.3 montre la représentation graphique des smoothies sur la base des deux premières composantes de STATIS, c'est à dire les deux premières colonnes de $\mathbf{C}$. Il est clair que les smoothies Casino_PBC et Innocent_PBC sont proches. Ils sont opposés aux smoothies Carrefour_MP et Immedia_SRB.

Afin d'établir une comparaison entre STATIS, l'APG et l'AFM, chacune des méthodes a été utilisée sur ce jeu de données. Sur la Figure 2.4, les représentations des produits sur la base des deux premières composantes calculées au moyen de l'APG et de l'AFM sont indiquées. Il est clair qu'il y a une grande similitude entre ces deux configurations et celle de STATIS (Figure 2.3). Ceci est confirmé par les valeurs élevées du RV entre les paires de configurations : $RV(\text{STATIS}, \text{APG}) = 0.99$, $RV(\text{STATIS}, \text{AFM}) = 0.97$, $RV(\text{APG}, \text{AFM}) = 0.96$. La Figure 2.5 souligne davantage les similitudes et les différences

2.4 Exemples d'application

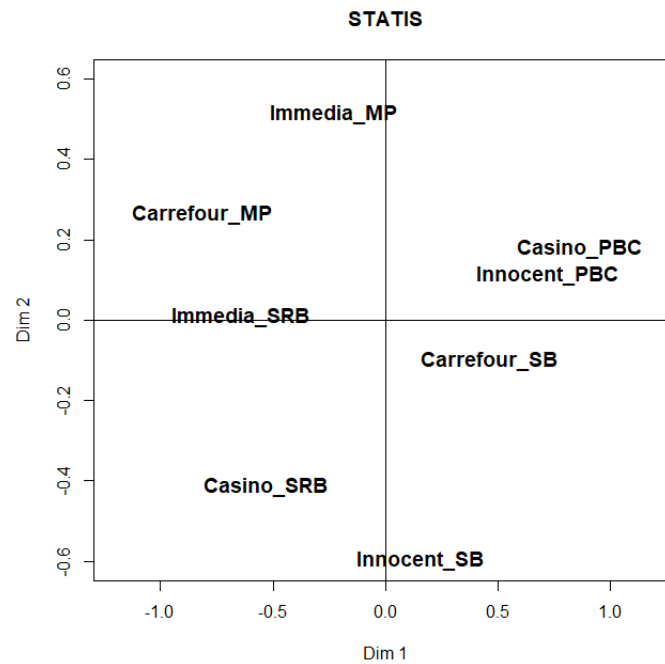


FIGURE 2.3 : Données « smoothies » : Représentation des produits sur la base des deux premières composantes du tableau compromis de STATIS.

entre les trois méthodes d'analyse. Dans cette figure, nous avons représenté, pour chacune des trois méthodes, l'inertie cumulée des tableaux de données originaux expliquée par les composantes dérivées de STATIS, de l'APG et de l'AFM. Comme tous les tableaux de données originaux sont bidimensionnels, la configuration moyenne du panel dérivée de l'APG est également bidimensionnelle et, par conséquent, seules deux composantes peuvent être déterminées. Ces deux composantes expliquent moins de 65% de la variation des tableaux originaux. Les deux premières composantes dérivées de STATIS et de l'AFM expliquent autant de variation que celles de l'APG. Cependant, avec les méthodes STATIS et AFM, d'autres composantes peuvent être calculées pour tenir compte des variations supplémentaires dans les tableaux. Au vu de la Figure 2.5, il est clair que ces deux dernières méthodes d'analyse présentent la même tendance en termes d'inertie restituée par les composantes successives.

2.4.2 Données de tri libre

Vingt-cinq sujets provenant de l'entreprise Product Perceptions Ltd. ont participé à une épreuve de tri libre. Il leur a été demandé de trier 14 marques de chocolat. Les données considérées ici sont extraites d'une expérience plus large détaillée dans Courcoux *et al.* (2012) et sont disponibles dans le package *R* ClustBlock (Llobell *et al.*, 2020). Les sujets

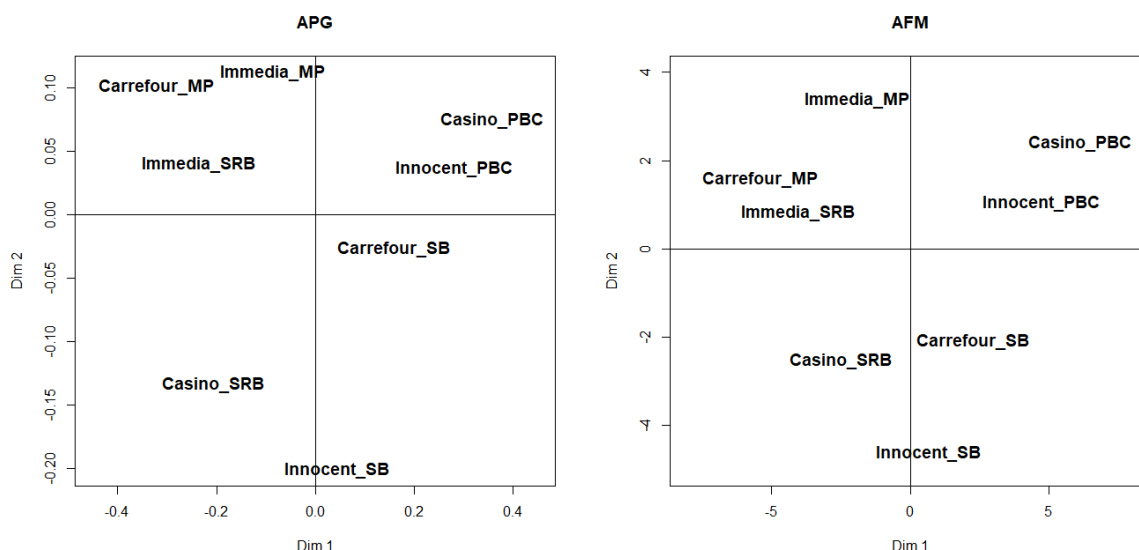


FIGURE 2.4 : Données « smoothies » : Représentation des produits sur la base des deux premières composantes calculées au moyen de l'APG et de l'AFM.

ont créé entre 5 et 10 groupes de produits.

Les données de chaque sujet ont été codées en tableau disjonctif complet, centrées et normalisées comme indiqué dans la section 2.3, puis soumises à la méthode STATIS. L'indice d'homogénéité I est égal à 63,9%, ce qui indique un accord convenable entre les sujets.

Nous avons comparé les résultats de la méthode STATIS à ceux d'autres méthodes dédiées à l'analyse de données de tri libre. Néanmoins, nous nous sommes surtout concentrés sur la méthode DISTATIS (Abdi *et al.*, 2007), puisque, comme expliqué précédemment, elle est étroitement liée à STATIS. Le côté gauche (respectivement droit) de la Figure 2.6 montre la position des produits sur la base des deux premières composantes obtenues au moyen de STATIS (respectivement DISTATIS). Il y a des similitudes évidentes entre ces deux configurations. Par exemple, les chocolats CDM et Galaxy, d'une part, et Tesco Value et JSValue, d'autre part, sont placés ensemble et chacune de ces paires de produits est très éloignée de l'autre paire. Cependant, il y a aussi des différences entre les deux configurations, avec notamment le chocolat Divine qui est plutôt extrême pour les deux premières composantes de STATIS alors qu'il occupe une position centrale dans la configuration de DISTATIS. Ces similitudes et ces différences sont reflétées par le coefficient RV entre ces deux configurations égal à 0.71, ce qui indique un accord assez élevé entre les résultats des deux méthodes d'analyse. En fait, les différences entre ces deux stratégies découlent de la normalisation des colonnes des tableaux disjonctifs complets (*i.e.*, des

2.4 Exemples d'application

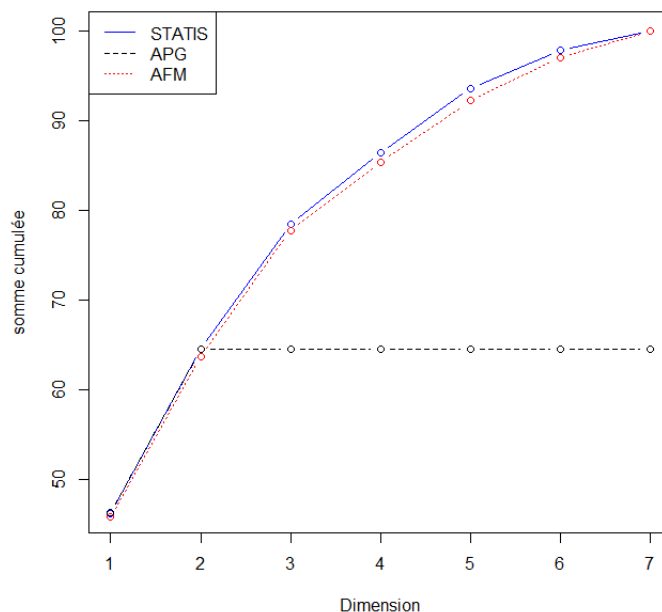


FIGURE 2.5 : Données « smoothies » : Pourcentages d’inertie des tableaux de données originaux expliquée par les composantes successives dérivées de STATIS, de l’APG et de l’AFM.

tailles des groupes) que nous avons préconisé dans la section 2.3. Si nous utilisons les tableaux disjonctifs complets non normalisés, le coefficient RV entre les configurations bidimensionnelles obtenues au moyen de STATIS et DISTATIS est égal à 0.96, ce qui indique une forte concordance entre les deux méthodes. Il est à noter que nous avons également comparé les deux premières composantes de STATIS avec les configurations obtenues à partir de méthodes basées sur l’Analyse des Correspondances Multiples (ACM, Cadoret *et al.*, 2009; Qannari *et al.*, 2010), sur l’analyse des correspondances simples (AFC, Carriou et Qannari, 2018) et sur le Multidimensional Scaling (MDS, Lawless *et al.*, 1995; Faye *et al.*, 2004). Tous ces résultats sont détaillés dans le Tableau 2.1, qui indique les coefficients RV entre les configurations bidimensionnelles dérivées des différentes méthodes. Afin de mieux mettre en évidence les similarités entre les méthodes, nous avons transformé les coefficients RV consignés dans le Tableau 2.1 en indices de distance ($d = 1 - RV$), puis nous avons effectué une analyse MDS sur la matrice de distances ainsi obtenue. La Figure 2.7 montre la position des différentes méthodes d’analyse sur la base des deux premiers axes de l’analyse MDS (stress=0.09). De toute évidence, DISTATIS semble se distinguer des autres méthodes, car il s’oppose le long du premier axe à toutes les autres méthodes. Le deuxième axe distingue la méthode d’analyse basée sur la MDS des méthodes STATIS et ACM qui sont très proches. La méthode AFC occupe une position centrale.

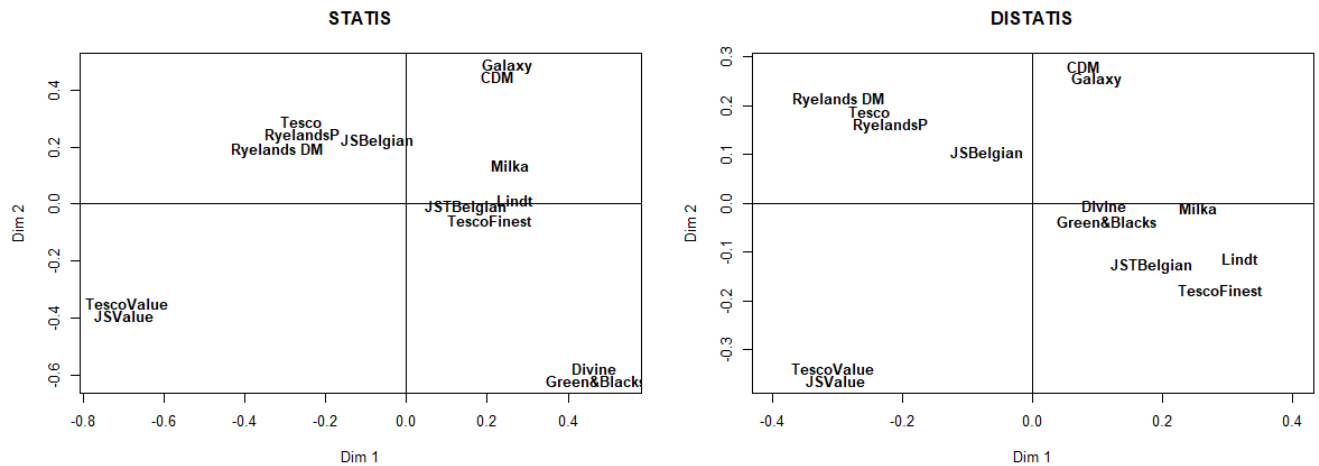


FIGURE 2.6 : Données « chocolats » : Représentation des marques de chocolat sur la base des deux premières composantes de STATIS (gauche) et DISTATIS (droite).

	STATIS	DISTATIS	AFC	ACM	MDS
STATIS	1.00	0.71	0.94	0.99	0.82
DISTATIS		1.00	0.79	0.69	0.69
AFC			1.00	0.92	0.94
ACM				1.00	0.83
MDS					1.00

Tableau 2.1 : Données chocolats : Coefficients RV entre les deux premières composantes de STATIS et DISTATIS avec celles obtenues au moyen de méthodes basées sur l'AFC, l'ACM et la MDS.

Afin de comparer les méthodes d'analyse de données de tri libre sur une base objective, nous avons confronté les deux premiers axes issus de chaque méthode aux données originales, à savoir les partitions des produits réalisées par les sujets. De manière plus précise, nous avons calculé pour chaque axe et chaque sujet, le rapport de corrélation qui est le rapport de la variance inter-classes et de la variance totale de classes. Par la suite, nous avons calculé la moyenne de cet indicateur pour tous les sujets. Il est évident que la comparaison des méthodes sur cette base est à l'avantage de l'ACM, puisque c'est justement l'indicateur qu'elle cherche à maximiser pour la détermination des axes. Il n'en reste pas moins vrai que c'est un indicateur pertinent pour les données considérées. Le Tableau 2.2 donne les rapports de corrélations moyens associés aux deux premiers axes pour les diverses méthodes. Il apparaît que la démarche basée sur STATIS a exactement la même performance que celle basée sur l'ACM. La méthode basée sur l'AFC a une performance très proche de l'ACM et STATIS, suivie de très près par l'approche basée sur MDS. La stratégie DISTATIS présente une performance relativement plus faible.

2.5 Conclusion

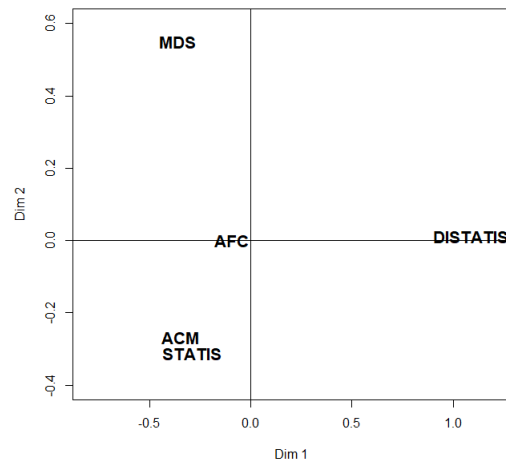


FIGURE 2.7 : Données « chocolats » : Représentation graphique des proximités entre les méthodes de tri libre sur les deux premiers axes de l'analyse MDS.

Pour toutes ces raisons, nous préconisons l'utilisation de STATIS pour l'analyse des données de tri libre, car, tout en ayant une performance similaire à celle de l'ACM, elle procure des outils supplémentaires pour mieux appréhender les différences entre sujets (analyse inter-structures).

	Axe 1	Axe 2
STATIS	0.915	0.814
DISTATIS	0.865	0.783
AFC	0.913	0.805
ACM	0.915	0.814
MDS	0.910	0.725

Tableau 2.2 : Données « chocolats » : Rapports de corrélation moyens associés aux deux premiers axes des diverses méthodes.

2.5 Conclusion

Nous avons proposé un critère de minimisation original pour la méthode STATIS et rappelé certaines propriétés essentielles. Par la suite, nous nous sommes intéressés à l'application de cette méthode d'analyse à deux épreuves sensorielles : l'épreuve de Projective mapping/Napping et l'épreuve de tri libre.

Dans le cadre d'une épreuve de Projective mapping/Napping, les données sont composées de tableaux de deux colonnes correspondant aux coordonnées sur les feuilles de

papier des sujets. L'application de STATIS sur ce type de données a montré que certains sujets pouvaient être atypiques, mais que STATIS permettait de se concentrer sur les sujets en accord avec le point de vue général du panel. Nous avons également réalisé une comparaison des résultats avec les méthodes courantes d'analyse sur ce type de données, à savoir l'Analyse Procrustéenne Généralisée (APG) et l'Analyse Factorielle Multiple (AFM). Cette comparaison basée sur une étude de cas a indiqué de fortes similarités entre les résultats des différentes stratégies.

Concernant les données de tri libre, nous avons proposé un pré-traitement avant d'appliquer la méthode STATIS sur ce type de données. Nous avons également proposé une standardisation qui permet de tenir compte du nombre de groupes formés par les sujets. Une comparaison avec les méthodes d'analyse existantes a permis de conclure à des résultats très similaires, excepté pour la stratégie DISTATIS et dans une moindre mesure la méthode basée sur l'analyse MDS.

Chapitre 3

Classification de tableaux : la méthode CLUSTATIS

3.1 Introduction

Nous avons déjà souligné l'intérêt de la classification de tableaux de données. La méthode CLUSTATIS est dédiée à cet effet. Elle s'applique à un ensemble de tableaux portant sur les mêmes individus mais pas nécessairement sur les mêmes variables. Elle est fortement liée à la méthode STATIS, car elle consiste à identifier des classes homogènes de tableaux, et, à l'intérieur de chaque classe, elle détermine un tableau compromis selon la démarche préconisée par STATIS. Elle peut être aussi vue comme une extension de la méthode de classification de variables CLV (Vigneau et Qannari, 2003) au cas de tableaux multiples.

Dans le cadre d'une classification d'entités, il arrive que certaines entités atypiques ne se conforment à aucune des classes identifiées. Il conviendrait alors de trouver ces entités et de les mettre à l'écart. En suivant une démarche préconisée par Dave (1991) et Vigneau *et al.* (2016), nous proposons une adaptation de la méthode CLUSTATIS de manière à mettre de côté les tableaux atypiques.

Les stratégies d'analyse sont illustrées sur la base de données provenant d'épreuves sensorielles de Projective mapping/Napping et de tri libre. Une comparaison avec des méthodes existantes (Berget *et al.*, 2019) pour segmenter les sujets lors d'une épreuve de Projective mapping/Napping a également lieu via des simulations et une étude de cas.

3.2 Principe de CLUSTATIS

La méthode CLUSTATIS vise à minimiser un critère reflétant la recherche de classes homogènes de tableaux de données. Plus précisément, les tableaux de chaque classe doivent être fortement liés à une configuration latente déterminée par la méthode STATIS. Formellement, les tableaux $\mathbf{X}_1, \dots, \mathbf{X}_m$ sont centrés, puis les matrices des produits scalaires $\mathbf{W}_1 = \mathbf{X}_1 \mathbf{X}_1^\top, \dots, \mathbf{W}_m = \mathbf{X}_m \mathbf{X}_m^\top$ sont calculées et standardisées par la norme de Frobenius. Enfin, K classes de tableaux $G_1, \dots, G_K$ sont déterminées de manière à minimiser le critère suivant :

$$D = \sum_{k=1}^K \sum_{i \in G_k} \|\mathbf{W}_i - \alpha_i \mathbf{W}^{(k)}\|^2 \quad (3.1)$$

où α_i ($i \in G_k$) sont des scalaires à déterminer sous la contrainte $\sum_{i \in G_k} \alpha_i^2 = 1$, et, pour $k = 1, \dots, K$, $\mathbf{W}^{(k)}$ est le tableau compromis de la classe G_k . Il est clair que lorsqu'il n'y a qu'une seule classe de tableaux (*i.e.*, $K = 1$), le critère F qui définit la méthode STATIS (équation 2.1) est retrouvé.

Le problème de minimisation D peut être résolu par deux stratégies complémentaires. La première est une classification hiérarchique des tableaux. La seconde approche est un algorithme de partitionnement semblable à l'algorithme des K -means (Pena et al., 1999).

L'algorithme hiérarchique suit une stratégie ascendante. Il démarre par la situation où chaque tableau de données forme une classe. Naturellement, dans ce cas, $D = 0$. À chaque étape, deux tableaux sont agrégés, ou, plus généralement au fur et à mesure que l'algorithme évolue, deux classes de tableaux sont fusionnées jusqu'à ce que tous les tableaux soient agrégés en une seule classe. Il découle de la propriété (iii) de la méthode STATIS (section 2.2) qu'à l'étape courante de l'algorithme hiérarchique, le critère D est égal à $D = m - \sum_{k=1}^K \lambda_1^{(k)}$, où $\lambda_1^{(k)}$ est la plus grande valeur propre de la matrice qui contient les coefficients RV entre les tableaux de la classe G_k . À chaque étape de l'algorithme, lorsque deux classes que nous désignons par A et B sont agrégées, la variation du critère D est égale à $\Delta = D_{A \cup B} - D_{A+B} = \lambda_1^{(A)} + \lambda_1^{(B)} - \lambda_1^{(A \cup B)}$, où $\lambda_1^{(A)}$, $\lambda_1^{(B)}$ et $\lambda_1^{(A \cup B)}$ sont respectivement les plus grandes valeurs propres des matrices composées des coefficients RV entre les tableaux des classes A , B et $A \cup B$. Nous pouvons montrer que cette variation est positive (voir Annexe A.2). Cela signifie que la fusion de deux classes A et B entraîne inéluctablement une détérioration de l'homogénéité au sein des classes, c'est à dire une augmentation du critère D . La stratégie d'agrégation vise à agréger les deux classes A et B qui entraînent la plus petite augmentation du critère D (*i.e.* la plus

3.2 Principe de CLUSTATIS

petite variation $\lambda_1^{(A)} + \lambda_1^{(B)} - \lambda_1^{(A \cup B)}$). L'Algorithme 1 résume la stratégie de classification ascendante hiérarchique.

Algorithme 1 Algorithme hiérarchique de CLUSTATIS.

Initialisation : Chaque tableau forme une classe. $D = 0$.

tant que il y a plus d'une classe **faire**

 Fusion des deux classes conduisant à la plus faible augmentation Δ de D .

fin tant que

Tous les tableaux sont dans la même classe.

Il est recommandé d'examiner l'augmentation Δ du critère D à chaque étape de l'algorithme hiérarchique, puisqu'il reflète l'augmentation de l'hétérogénéité à l'intérieur des classes au cours de l'évolution de l'algorithme. Une grande variation de ce critère indique que nous essayons de fusionner deux classes hétérogènes, ce qui doit être considéré comme un signal pour arrêter l'algorithme. En pratique, l'augmentation du critère D se reflète dans l'arbre hiérarchique (ou dendrogramme), puisqu'elle correspond à la hauteur des branches qui relient deux classes. Alternativement, les variations entre les étapes successives peuvent être représentées à l'aide de diagrammes en bâtons, indiquant leur évolution au fur et à mesure que le nombre de classes diminue. Ainsi, l'algorithme hiérarchique peut être très utile pour déterminer graphiquement le nombre de classes.

Le problème de classification basé sur le critère D peut également être résolu au moyen d'un algorithme de partitionnement similaire à l'algorithme des K -means. Au cours de cet algorithme, les tableaux de données peuvent changer de classe, de manière à minimiser le critère D . Cet algorithme suppose que le nombre de classes, K , est connu à l'avance, et qu'une partition initiale est donnée. L'Algorithme 2 détaille cette approche.

Dans la pratique, il est recommandé d'exécuter les deux algorithmes pour obtenir une meilleure solution que lorsque chaque algorithme est appliqué seul. Premièrement, la classification hiérarchique peut être effectuée, permettant d'indiquer un nombre approprié de classes en examinant l'évolution du critère d'agrégation D au cours du processus de regroupement. Deuxièmement, les tableaux sont soumis à l'algorithme de partitionnement en utilisant comme partition initiale la solution obtenue en coupant l'arbre hiérarchique au niveau indiqué (*i.e.*, avec le nombre de classes sélectionné K). En permettant aux tableaux de changer de classe lors de l'exécution de l'algorithme de partitionnement, la solution obtenue par l'algorithme hiérarchique est susceptible d'être améliorée puisque l'on tente de minimiser davantage le critère D . Cette dernière étape est appelée « consolidation » (Sa-

Algorithme 2 Algorithme de partitionnement de CLUSTATIS.

Initialisation : Donner une partition initiale en K classes.

tant que au moins un tableau change de classe **faire**

pour $k = 1, \dots, K$ **faire**

 Application de STATIS dans chaque classe de tableaux pour calculer les tableaux compromis $\mathbf{W}^{(k)}$.

fin pour

De nouvelles classes de tableaux sont formées en déplaçant chaque tableau, $\mathbf{X}_i$, vers la classe G_k pour laquelle la quantité $\|\mathbf{W}_i - \alpha_i \mathbf{W}^{(k)}\|$ est la plus faible, ou, de façon équivalente, la quantité $RV(\mathbf{W}_i, \mathbf{W}^{(k)})$ est la plus élevée. Cette équivalence découle du fait que $\|\mathbf{W}_i - \alpha_i \mathbf{W}^{(k)}\|^2 = 1 - RV^2(\mathbf{W}_i, \mathbf{W}^{(k)})$ (propriété (iii) de STATIS).

fin tant que

(porta, 2006). Il est à noter que lorsque le nombre de tableaux est relativement grand (de l'ordre de quelques centaines), l'algorithme hiérarchique peut demander un temps d'exécution assez élevé. Comme alternative, nous préconisons une solution « multi-start ». Cela consiste à exécuter l'algorithme de partitionnement plusieurs fois en considérant à chaque fois une partition initiale choisie au hasard. *In fine*, c'est la solution qui correspond à la plus petite valeur du critère D qui est retenue.

Une importante propriété est la décomposition suivante :

$$\begin{aligned}
 \sum_{i=1}^m \|\mathbf{W}_i\|^2 &= m = \sum_{i=1}^m \sum_{i \in G_k} \|\mathbf{W}_i - \alpha_i \mathbf{W}^{(k)}\|^2 + \sum_{k=1}^K \|\mathbf{W}^{(k)}\|^2 \\
 &= \sum_{i=1}^m \sum_{i \in G_k} \|\mathbf{W}_i - \alpha_i \mathbf{W}^{(k)}\|^2 + \sum_{k=1}^K \sum_{i \in G_k} RV^2(\mathbf{W}_i, \mathbf{W}^{(k)}) \\
 &= \sum_{i=1}^m \sum_{i \in G_k} \|\mathbf{W}_i - \alpha_i \mathbf{W}^{(k)}\|^2 + \sum_{k=1}^K \lambda_1^{(k)} \\
 &= D + H
 \end{aligned} \tag{3.2}$$

avec $H = \sum_{k=1}^K \lambda_1^{(k)}$.

Ces égalités, qui découlent des propriétés de STATIS (section 2.2), indiquent que la variation totale présente dans les tableaux originaux se décompose en la variation expliquée par le modèle (ou variance inter-classes), H , et la variation résiduelle (ou variance intra-classes), D . Ainsi, il apparaît qu'en minimisant D , nous cherchons à maximiser la variation inter-classes, H .

Plusieurs indices associés à la solution finale sont d'un intérêt primordial. En premier

3.3 Classification en écartant les tableaux atypiques

lieu, nous considérons pour chaque classe G_k ($k = 1, \dots, K$), l'indice $I_k = \frac{\lambda_1^{(k)}}{m_k}$, où $\lambda_1^{(k)}$ est la plus grande valeur propre de la matrice des coefficients RV entre les tableaux de la classe G_k et m_k est le nombre de tableaux dans cette classe. Cet indice varie entre $1/m_k$ et 1 et reflète l'homogénéité de G_k . Un indice global pour évaluer la qualité de la partition des tableaux obtenue par l'approche de classification est donné par la moyenne pondérée des indices I_k : $I = \sum_{k=1}^K \frac{m_k I_k}{m} = \sum_{k=1}^K \frac{\lambda_1^{(k)}}{m}$. Cet indice peut être interprété comme le pourcentage de variation des tableaux originaux expliqué par les tableaux compromis des différentes classes. Dans la classe G_k , nous pouvons calculer pour chaque tableau $\mathbf{X}_i$ ($i \in G_k$), le coefficient RV entre $\mathbf{W}_i$ et $\mathbf{W}^{(k)}$. Ce coefficient reflète la proximité de chaque tableau avec le tableau compromis de la classe qui lui est associée. Alternativement, il est possible de considérer les poids α_i ($i = 1, \dots, m$) qui reflètent la même idée. Enfin, pour évaluer dans quelle mesure les différentes classes sont proches les unes des autres, nous pouvons calculer les coefficients RV entre les tableaux compromis $\mathbf{W}^{(k)}$ associés aux différentes classes.

3.3 Classification en écartant les tableaux atypiques

3.3.1 Classe « $K + 1$ »

En classification, il arrive fréquemment que parmi les entités que nous cherchons à segmenter, certaines ne se rapportent à aucune des classes déterminées. En identifiant ces entités atypiques puis en les mettant de côté, la classification est susceptible d'être de meilleure qualité. Le principe de la stratégie dite « classe $K + 1$ » ou « classe de bruit » est de définir, à côté des K classes principales, une classe supplémentaire dans laquelle toutes les entités qui ne se conforment à aucune des classes principales sont affectées. La stratégie décrite ci-après est une adaptation de celle proposée par [Dave \(1991\)](#) pour la classification d'individus et de celle de [Vigneau et al. \(2016\)](#) pour la classification de variables.

La première étape consiste à choisir une valeur seuil ρ entre 0 et 1. En dessous de ρ , la similarité entre deux tableaux de données, mesurée par le coefficient RV, est considérée comme insignifiante. Dans un second temps, l'objectif est de maximiser le critère suivant :

$$H_\rho = \sum_{k=1}^K \sum_{i=1}^m (\delta_i^k RV^2(\mathbf{W}_i, \mathbf{W}^{(k)}) + \delta_i^{K+1} \rho^2) \quad (3.3)$$

où δ_i^k est le symbole de Kronecker, égal à 1 si le tableau $\mathbf{X}_i$ appartient à la classe G_k et à 0 sinon, sous la contrainte $\sum_{k=1}^{K+1} \delta_i^k = 1$. Cette contrainte implique qu'un tableau

appartient à une et une seule classe, y compris la classe « $K + 1$ ». La justification de ce critère est simple : chaque tableau, $\mathbf{X}_i$, est affecté à une classe G_k pour laquelle la valeur $RV(\mathbf{W}_i, \mathbf{W}^{(k)})$ est la plus grande. Cependant, si cette dernière quantité est plus petite que ρ , alors $\mathbf{X}_i$ est affecté à la classe « $K + 1$ ».

Pour résoudre ce problème d'optimisation, nous proposons un nouvel algorithme de partitionnement (Algorithme 3). Cet algorithme est similaire à l'Algorithme 2 excepté pour la phase d'affectation. Il nécessite de fournir une partition initiale ainsi qu'un paramètre ρ . Pour le choix de la partition initiale, les mêmes considérations que celles discutées dans la section précédente s'appliquent. Le choix du paramètre ρ sera discuté plus en détail ci-après.

Algorithme 3 Algorithme de partitionnement de CLUSTATIS avec option « $K + 1$ ».

Initialisation : Donner une partition initiale (idéalement à partir de la classification hiérarchique de CLUSTATIS) et un seuil minimal de similarité ρ .

tant que au moins un tableau change de classe **faire**

pour $k = 1, \dots, K$ **faire**

 Application de STATIS dans chaque classe pour calculer les tableaux compromis $\mathbf{W}^{(k)}$.

fin pour

 Affectation de chaque tableau à la classe G_k pour laquelle le coefficient $RV(\mathbf{W}_i, \mathbf{W}^{(k)})$ est le maximum, ou à la classe « $K + 1$ » si $RV(\mathbf{W}_i, \mathbf{W}^{(k)}) < \rho$.

fin tant que

3.3.2 Choix du seuil minimal de similarité ρ

Nous proposons une procédure pour sélectionner une valeur seuil appropriée, ρ . Nous considérons comme un tableau atypique un tableau aberrant au sens où il est éloigné de toutes les classes, ou un tableau qui se place entre deux ou plusieurs classes. Ces cas sont représentés sur la Figure 3.1, qui donne une représentation schématique de tableaux divisés en trois classes. Chaque point représente un tableau de données. Les trois cercles délimitent clairement trois classes et les points à l'extérieur de ces cercles représentent des tableaux atypiques. Plus précisément, les cercles délimitent la distance maximale au centre (*i.e.* la similarité minimale avec le tableau compromis) nécessaire pour appartenir à la classe considérée.

Puisque le paramètre ρ représente un indice de similarité mesuré par le coefficient RV entre un tableau de données et un tableau compromis associé à une classe de tableaux, il doit être compris entre 0 et 1. Pour $\rho = 0$, la classe « $K + 1$ » est vide, et pour $\rho = 1$,

3.3 Classification en écartant les tableaux atypiques

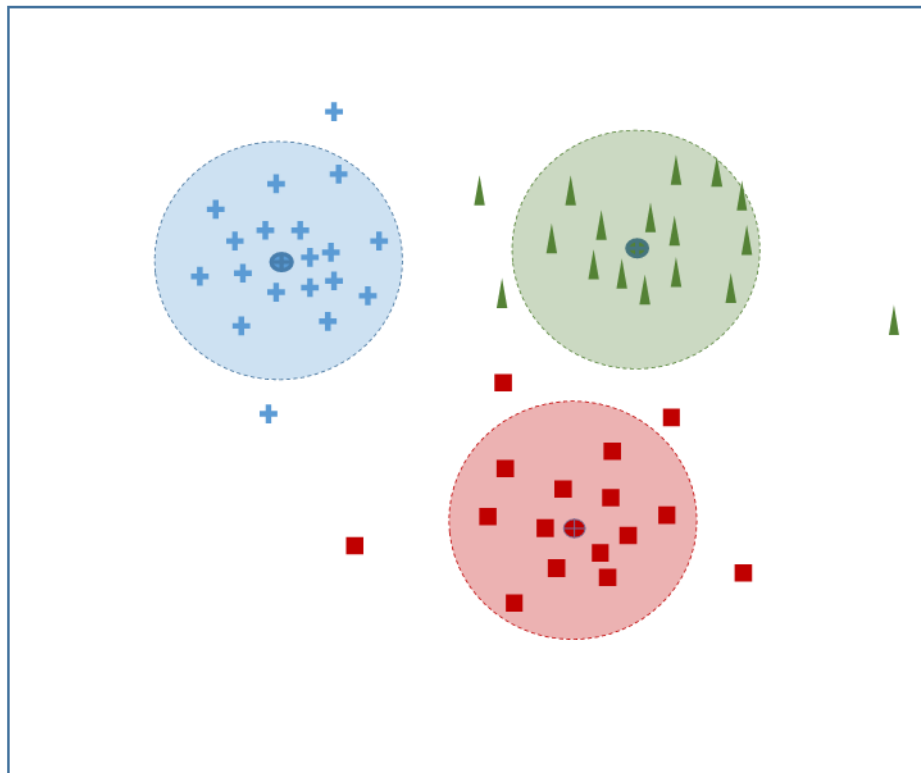


FIGURE 3.1 : Représentation schématique de trois classes de tableaux. Les tableaux non circonscrits sont considérés comme atypiques.

tous les tableaux sont placés dans la classe « $K + 1$ ». [Dave \(1991\)](#), qui a été le premier à introduire le concept de classe « de bruit » pour la classification d'individus, a suggéré une stratégie pour le choix de ρ . Il s'agit, dans un premier temps, de calculer la moyenne des distances au carré entre chaque individu et chaque centre des différentes classes. Ensuite, une moyenne de toutes ces distances est calculée et multipliée par un autre paramètre (hyperparamètre). Cependant, peu d'indications sont données concernant cet hyperparamètre. [Vigneau et al. \(2016\)](#), qui ont préconisé d'utiliser le concept de la classe « $K + 1$ » dans le contexte de la classification de variables, ont proposé pour la sélection du seuil de considérer plusieurs valeurs de ρ entre 0 et 1, ce qui dans leur contexte s'apparente à un coefficient de corrélation, et d'examiner l'évolution de l'homogénéité intra-classes en fonction de ρ . Parallèlement, l'évolution du nombre de variables affectées à la classe « $K + 1$ » est étudiée. De ce point de vue, une valeur de ρ appropriée devrait correspondre à une homogénéité satisfaisante sans pour autant mettre de côté une grande proportion de variables. Cette stratégie pourrait être adaptée sans difficulté à la méthode CLUSTATIS.

Nous proposons de calculer automatiquement une valeur appropriée du paramètre

ρ dans le cadre de CLUSTATIS. Dans un premier temps, l'algorithme hiérarchique de CLUSTATIS doit être utilisé pour déterminer une partition des tableaux. Ceci suppose en particulier le choix du nombre K de classes. La quantité ρ est alors définie comme la moyenne des coefficients RV entre chaque tableau de données et les deux tableaux compromis les plus proches, qui correspondent aux deux plus grandes valeurs des coefficients RV entre la matrice des produits scalaires $\mathbf{W}_i$ associée au tableau $\mathbf{X}_i$ et les compromis des classes. Formellement, la valeur ρ que nous proposons est la suivante :

$$\rho = \frac{\sum_{i=1}^m (RV(\mathbf{W}_i, \mathbf{W}^{(k_i)}) + RV(\mathbf{W}_i, \mathbf{W}^{(k_{i'})}))}{2m} \quad (3.4)$$

où $\mathbf{W}^{(k_i)}$ est le tableau compromis de classe le plus proche du i^e tableau, et $\mathbf{W}^{(k_{i'})}$ le deuxième tableau compromis de classe le plus proche du i^e tableau.

Cette proposition est guidée par le principe de trouver un seuil ρ en rapport avec les données considérées. Plus la similarité des tableaux avec le tableau compromis de leur classe est élevée, plus le seuil ρ est grand. De surcroît, plus les tableaux sont proches des classes voisines, plus la valeur ρ est élevée. La première propriété implique que pour les classes compactes, nous devrions choisir une valeur seuil ρ relativement grande. La deuxième propriété implique que si les différentes classes sont relativement proches les uns des autres, alors la valeur seuil devrait également être élevée afin de délimiter des frontières claires entre les classes.

3.4 Exemples d'application en analyse sensorielle

3.4.1 Projective mapping ou Napping

Dans cette partie, l'accent est placé sur l'application de CLUSTATIS dans le cas de données de Projective mapping/Napping. Dans un premier temps, un exemple concret est détaillé. Dans la seconde étape, dédiée à la comparaison avec d'autres méthodes, une étude de cas réelle puis des simulations sont présentées.

Exemple d'application

Cet exemple concerne les mêmes données que dans l'exemple d'application de STATIS (section 2.4.1). Huit smoothies ont été évalués par 24 sujets notés S1, ..., S24. Par conséquent, il y a 24 tableaux de données, chacun étant associé à un sujet et se composant des coordonnées X-Y des smoothies sur la feuille de papier. Nous avons constaté dans la

3.4 Exemples d'application en analyse sensorielle

section 2.4.1 que l'indice d'homogénéité trouvé en appliquant la méthode STATIS (*i.e.* avec une seule classe) est $I = 42.4\%$. Une segmentation des sujets participant à l'épreuve pourrait améliorer cette homogénéité.

Le dendrogramme associé à l'algorithme hiérarchique ainsi que la variation Δ du critère D de l'analyse CLUSTATIS sont représentés sur la Figure 3.2. Ces deux manières de présenter la variation du critère d'agrégation permettent de déterminer visuellement le nombre de classes. Il s'avère que lors du passage d'une partition en trois classes à une partition en deux classes, la variation Δ du critère D est relativement grande. Cela implique que la fusion de deux des trois classes donne une classe hétérogène. Par conséquent, nous retenons la partition en trois classes. Le Tableau 3.1 donne des indicateurs concernant la partition ainsi obtenue. En particulier, nous pouvons noter que l'homogénéité globale est égale à 58.3%, ce qui représente une nette amélioration par rapport à la situation initiale où tous les tableaux étaient dans une même classe. Par la suite, deux stratégies d'analyse s'offrent à nous : utiliser l'algorithme de partitionnement sans classe « $K + 1$ » (Algorithme 2) ou l'algorithme incluant la classe « $K + 1$ » (Algorithme 3). Ces deux possibilités sont appliquées et comparées ci-après.

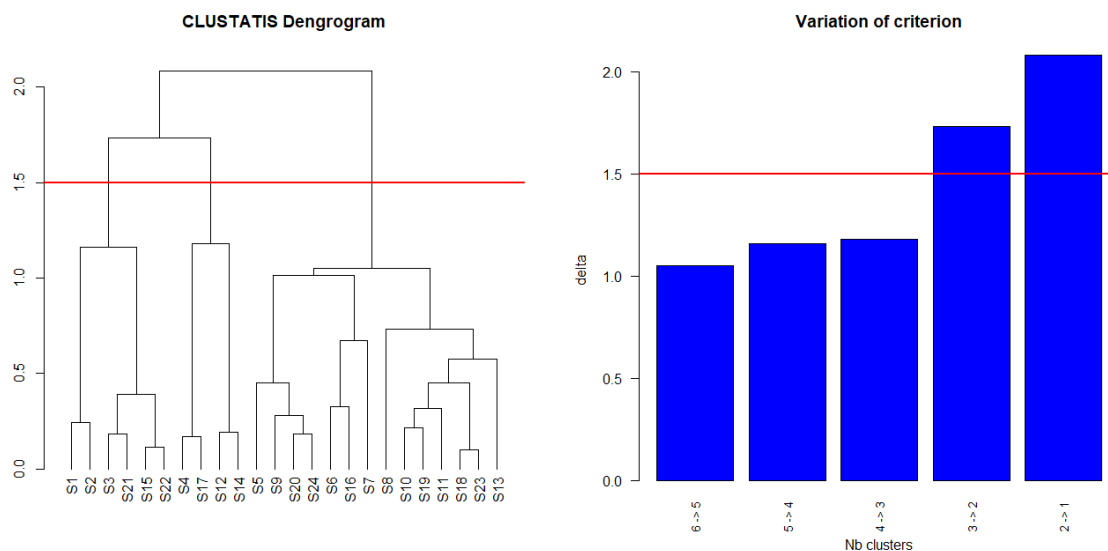


FIGURE 3.2 : Données « smoothies » : Dendrogramme et variation du critère D lors de l'algorithme hiérarchique de CLUSTATIS.

Suite à la consolidation par l'algorithme de partitionnement sans option « $K + 1$ », un seul sujet a changé de classe, et l'indice d'homogénéité global a été légèrement amélioré (+0.6%). Nous avons calculé le coefficient RV entre la matrice des produits scalaires as-

Classe	Homogénéité	Nb sujets
1	65.1%	6
2	61.5%	4
3	54.5%	14
Homogénéité globale	58.3%	24

Tableau 3.1 : Données « smoothies » : Taille, homogénéité de chaque classe et homogénéité globale provenant de l'algorithme hiérarchique de CLUSTATIS.

sociée à chaque sujet et le tableau compromis de la classe à laquelle le sujet considéré appartient (Figure 3.3). Nous constatons que, dans l'ensemble, ces coefficients sont relativement importants. Cependant, quelques rares coefficients sont relativement faibles. C'est le cas du sujet S2 dans la classe 1 et des sujets S7 et S8 dans la classe 3. Ces constatations mettent en évidence l'existence de sujets atypiques qui ne se conforment pas à leur classe.

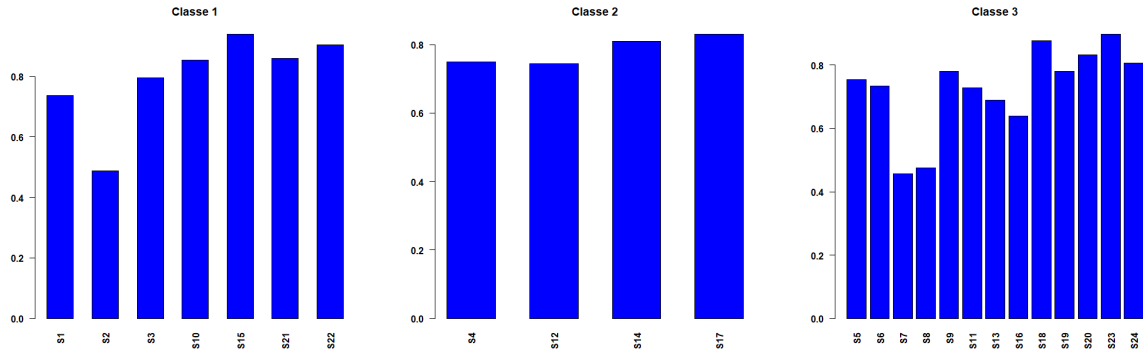


FIGURE 3.3 : Données « smoothies » : Coefficients RV entre les matrices des produits scalaires associées aux sujets S1 à S24 avec le tableau compromis de leur classe.

À présent, nous discutons les résultats de l'algorithme de partitionnement en ajoutant une classe « $K + 1$ ». Afin de choisir le seuil minimal de similarité, ρ , la Figure 3.4 montre l'évolution de l'indice d'homogénéité globale et du nombre de sujets qui ont été affectés à la classe « $K + 1$ » en fonction du seuil ρ . Un seuil entre 0.5 et 0.6 est indiqué puisqu'il permet d'écarter seulement trois sujets tout en gagnant environ 6% d'homogénéité globale. En considérant $\rho = 0.65$, un autre sujet est mis de côté mais le gain en termes d'homogénéité globale est négligeable. Avec $\rho = 0.7$, l'homogénéité globale augmente d'environ 4%, mais cela provoque de la mise à l'écart de 7 sujets au total. Ceci peut être considéré comme trop élevé au regard du bénéfice obtenu. Le paramètre ρ calculé en utilisant l'équation 3.4 est égal à 0.6, ce qui semble cohérent avec le constat graphique. Les trois sujets qui ont été écartés sont le sujet S2 de la classe 1 et les sujets S7 et S8 de la classe 3. Ce sont exactement les sujets qui ont été mis en évidence comme ayant de faibles similarités avec

3.4 Exemples d'application en analyse sensorielle

le tableau compromis de leur classe sur la Figure 3.3.

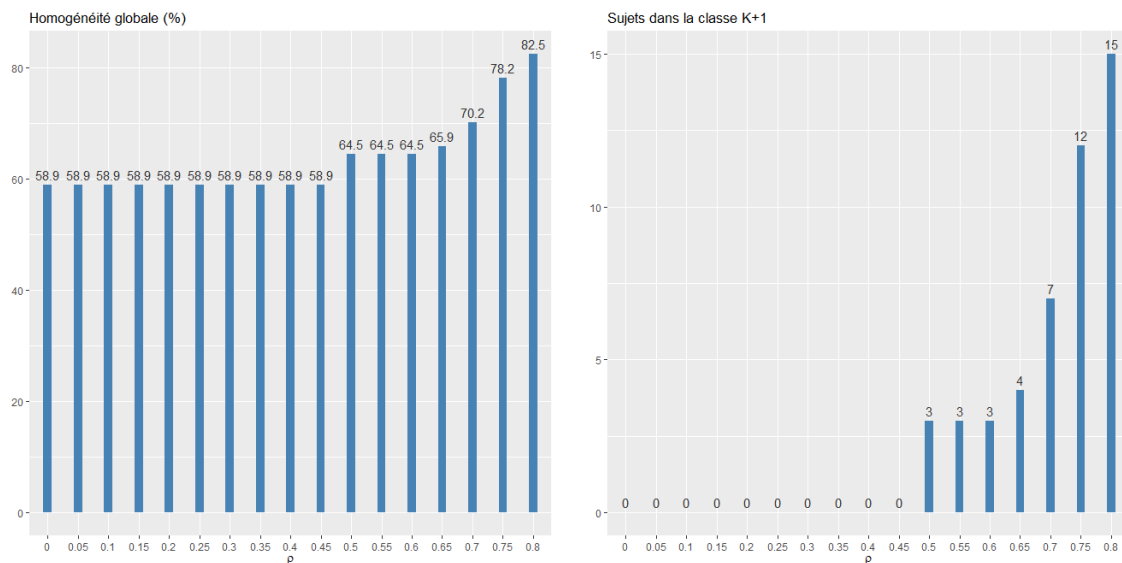


FIGURE 3.4 : Données « smoothies » : Évolution de l'indice d'homogénéité globale (gauche) et du nombre de sujets dans la classe « $K + 1$ » (droite) en fonction du seuil ρ .

Le Tableau 3.2 indique la taille de chacune des trois classes et leur homogénéité. Il apparaît que la classe 1 est très homogène. La classe 2 n'a pas été affectée par la procédure « $K + 1$ », et la classe 3 a une meilleure homogénéité du fait que les deux sujets atypiques S7 et S8 ont été écartés. L'homogénéité globale, moyenne pondérée de l'homogénéité des classes, a par conséquent augmenté de 5.6%. Ces résultats montrent l'importance de la classe « $K + 1$ », qui permet, en écartant les tableaux atypiques, d'améliorer l'homogénéité des classes.

Classe	Sans classe « $K + 1$ »		Avec classe « $K + 1$ »	
	Homogénéité	Nb sujets	Homogénéité	Nb sujets
1	65.4%	7	73.2%	6
2	61.5%	4	61.5%	4
3	54.5%	13	60.9%	11
Homogénéité globale	58.9%	24	64.5%	21

Tableau 3.2 : Données « smoothies » : Taille, homogénéité de chaque classe et homogénéité globale obtenues en utilisant CLUSTATIS sans et avec la classe « $K + 1$ ».

La Figure 3.5 représente les huit smoothies sur la base des deux premières composantes principales du tableau compromis de chaque classe obtenue à l'issue de la procédure

« $K + 1$ ». Il est clair qu'il y a des différences entre ces trois configurations, notamment concernant le smoothie Carrefour_SB. Ainsi, chaque classe de consommateurs débouche sur une perception différente des produits et ces graphiques permettent de mettre en évidence les différences.

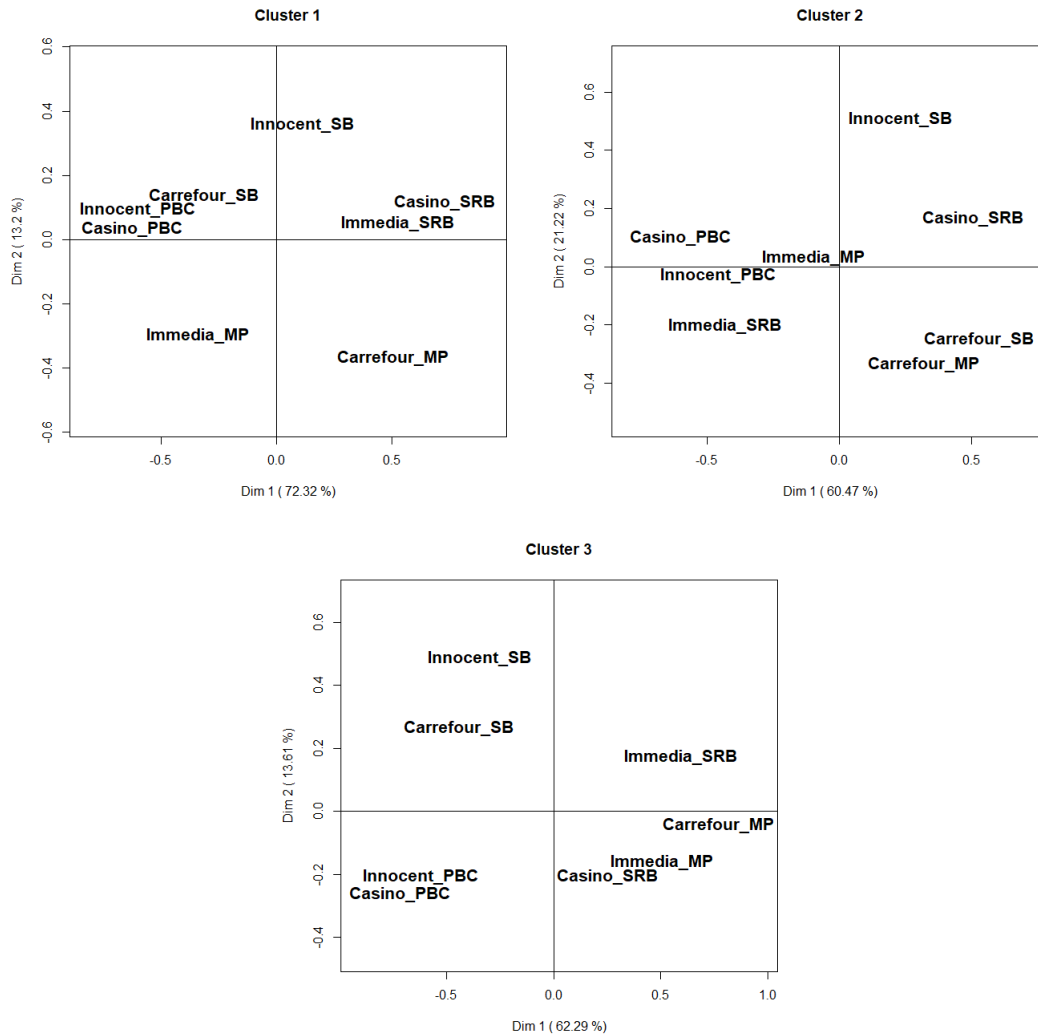


FIGURE 3.5 : Données « smoothies » : Représentation des produits sur la base des deux premières composantes de STATIS dans chaque classe construite par CLUSTATIS avec l'option « $K + 1$ ».

Les coefficients RV entre les tableaux compromis associés aux trois classes sont consignés dans le Tableau 3.3. Ils permettent d'évaluer la proximité des classes les unes par rapport aux autres. Dans le cas présent, les classes 1 et 2 apparaissent comme relativement différentes.

Classe	1	2	3
1	1.00	0.37	0.65
2		1.00	0.35
3			1.00

Tableau 3.3 : Coefficients RV entre les tableaux compromis des trois classes.

Comparaisons avec d'autres méthodes

L'objectif de l'étude de comparaison est de confirmer l'efficacité de CLUSTATIS et de montrer la pertinence du choix automatique du seuil ρ pour mettre de côté les tableaux atypiques, tout en comparant avec des méthodes existantes. Ainsi, des comparaisons avec les résultats des méthodes « Sequential Clusterwise Rotations » (SCR) et « Proclustrees » (Berget *et al.*, 2019) sont réalisées. La méthode SCR consiste à identifier dans un premier temps un groupe central parmi tous les sujets, puis à le mettre de côté en tant que première classe et à répéter la procédure jusqu'à ce qu'un nombre maximum de classes soit identifié. À chaque étape de la séquence, un critère de type APG est utilisé pour l'identification du groupe central. La stratégie Proclustrees consiste à construire un arbre hiérarchique sur la base de la distance de Procrustes (Gower, 1975). Cette approche avait été introduite par Dahl et Næs (2004) pour la détection de juges atypiques dans le cadre de profils sensoriels.

Nous commençons par réaliser une comparaison des méthodes sur la base d'une étude de cas, puis, dans un second temps, grâce à des simulations. Ces deux étapes de comparaison se rapportent à des données de Projective mapping/Napping.

Données réelles

L'objectif de cette partie est de comparer CLUSTATIS avec les stratégies SCR et Proclustrees sur une étude de cas. À cette fin, nous utilisons les mêmes données que celles qui ont été utilisées pour illustrer ces deux dernières stratégies (Berget *et al.*, 2019). Elles concernent une épreuve de Projective mapping/Napping impliquant 100 sujets à qui l'on a demandé d'évaluer 12 emballages de yaourts aux myrtilles commercialisés en Norvège notés P1, ..., P12. Pour plus de détails sur ces données, nous nous référons à trois articles (Næs *et al.*, 2017; Varela *et al.*, 2017; Berget *et al.*, 2019). Le dendrogramme obtenu au moyen de CLUSTATIS (Figure 3.6) suggère de choisir deux classes de sujets. Cette conclusion est conforme à celle qui a été tirée à partir du dendrogramme obtenu au moyen de Proclustrees (Berget *et al.*, 2019).

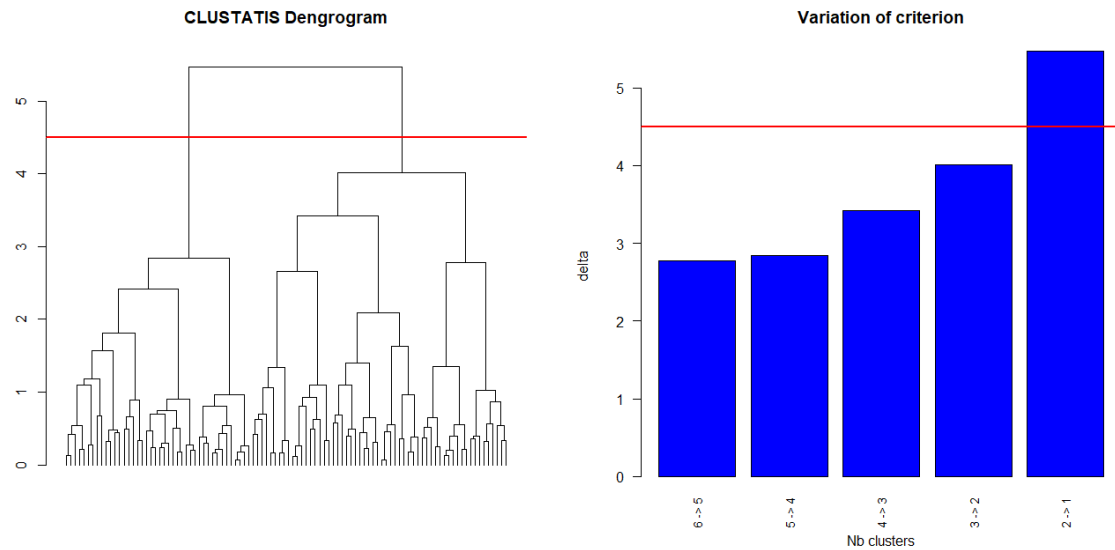


FIGURE 3.6 : Données « yaourts » : Dendrogramme et variation du critère D provenant de l'algorithme hiérarchique de CLUSTATIS.

Le Tableau 3.4 présente les indices d'homogénéité des classes. Dans ce tableau, deux situations sont considérées, à savoir la situation où l'option concernant la classe « $K + 1$ » n'est pas activée (CLUSTATIS sans option « $K + 1$ » et Proclustrees, puisque cette méthode ne propose pas de mettre de côté les tableaux atypiques) et la situation où elle est activée (CLUSTATIS avec option « $K + 1$ » et SCR, puisque cette stratégie met de côté les tableaux atypiques). Dans tous les cas, il est clair que l'homogénéité des classes est plus élevée avec CLUSTATIS qu'avec les deux autres méthodes, ce qui indique une meilleure performance de classification. La seule exception à cette conclusion générale concerne la classe 2 définie par Proclustrees qui présente une homogénéité légèrement supérieure à celle obtenue par CLUSTATIS.

	Classe 1		Classe 2		Homogénéité globale	
<i>Sans classe « $K + 1$ »</i>						
Méthode d'analyse	Proclustrees	CLUSTATIS	Proclustrees	CLUSTATIS	Proclustrees	CLUSTATIS
(Taille de la classe)	(47)	(43)	(53)	(57)	(100)	(100)
Indice d'homogénéité	25.7	40.4	24.5	23.3	25.1	30.7
<i>Avec classe « $K + 1$ »</i>						
Méthode d'analyse	SCR	CLUSTATIS	SCR	CLUSTATIS	SCR	CLUSTATIS
(Taille de la classe)	(42)	(37)	(26)	(26)	(68)	(63)
Indice d'homogénéité	39.0	44.6	27.1	35.5	34.5	40.8

Tableau 3.4 : Données « yaourts » : Comparaison, sur la base des indices d'homogénéité, de Proclustrees et CLUSTATIS (sans classe « $K + 1$ »), et de SCR et CLUSTATIS (avec classe « $K + 1$ »).

3.4 Exemples d'application en analyse sensorielle

La Figure 3.7 représente les produits sur la base des deux premières composantes de STATIS, qui est exécutée sur les tableaux des deux classes construites par CLUSTATIS avec l'option « $K + 1$ ». Il apparaît que les sujets des deux classes ont une perception différente de ces produits puisque les deux configurations semblent être dissemblables. Par exemple, les produits P5 et P9 sont opposés dans la classe 1 et placés ensemble dans la classe 2. Il convient également de noter que la configuration associée à la classe 1 est similaire à la classe 1 obtenue au moyen de Proclustrees (voir Berget *et al.*, 2019). Le coefficient RV entre ces deux configurations est égal à 0.80. La configuration associée à la classe 2 présente une assez bonne similitude avec celle associée à la classe 2 obtenue au moyen de Proclustrees (RV=0.76, voir Berget *et al.*, 2019). La principale différence entre ces deux dernières configurations concerne les produits P4 et P8 qui figurent à proximité des produits P2 et P6 dans CLUSTATIS (Figure 3.7, droite), alors que dans Proclustrees, ces produits occupent une position intermédiaire entre P9 et P5, d'une part, et P1, P10 et P11, d'autre part.

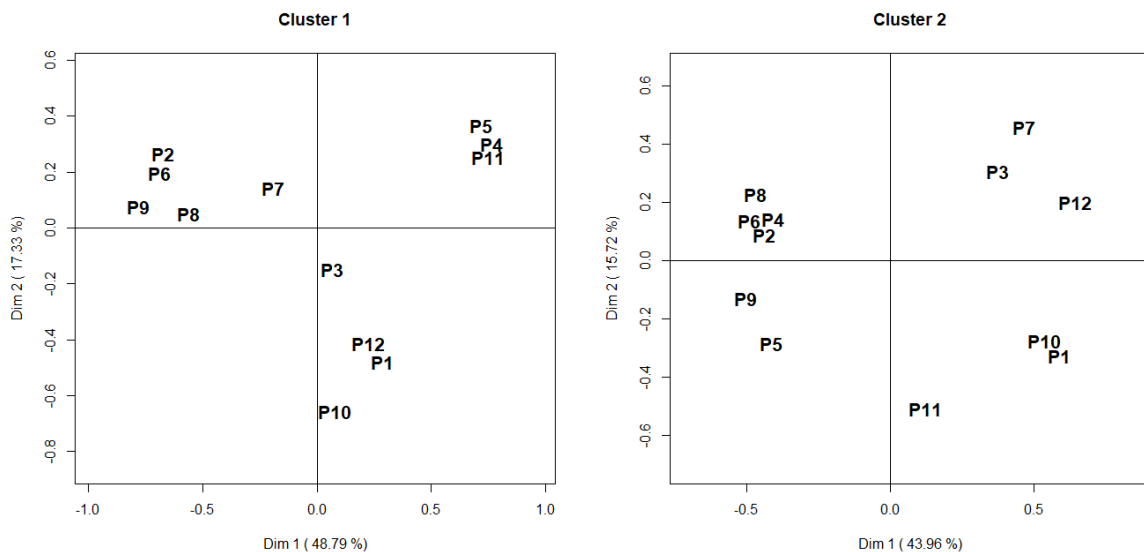


FIGURE 3.7 : Données « yaourts » : Représentation des produits sur la base des deux premières composantes de STATIS de chaque classe.

Simulations

Afin d'aller plus loin dans la comparaison des approches de classification, des simulations ont été réalisées. Nous avons suivi strictement le cadre de simulation proposé par Berget *et al.* (2019). Des données de Projective mapping/Napping ont été simulées en considérant le cas de 100 sujets et de 9 produits. Les sujets sont supposés appartenir à

trois classes nommées 1, 2 et 3. Les données ont été générées de manière à ce que les classes 1 et 2 aient une perception différente des produits, et que la classe 3 contienne des sujets qui devraient faire partie de la classe « $K + 1$ », puisqu'ils disposent les produits plus ou moins aléatoirement, et, par conséquent, représentent le bruit (Berget *et al.*, 2019). Les configurations de base (coordonnées des produits initiaux) à partir desquelles les classes 1 et 2 ont été simulées sont représentées sur la Figure 3.8. Les simulations ont été réalisées en considérant différents niveaux de variabilité suivant une distribution normale et en faisant varier les proportions des sujets dans les trois classes. Plus précisément, dans les classes 1 et 2, les coordonnées du produit l placé par le sujet i dans la classe k est généré à partir d'une loi normale $\mathcal{N}(x_{lk}, v_k \mathbf{S}_0)$, où x_{lk} représente le produit initial (Figure 3.8) et $\mathbf{S}_0 = \begin{bmatrix} 1 & 0 \\ 0 & 2/3 \end{bmatrix}$. Pour les variations faible (L), v_k vaut 1000, et pour les variations fortes (H), v_k est fixé à 2000. La classe 3 est simulée à partir de la loi uniforme $\mathcal{U}([-300, 300], [-200, 200])$. Pour plus de détails, nous renvoyons à l'article de Berget *et al.* (2019).

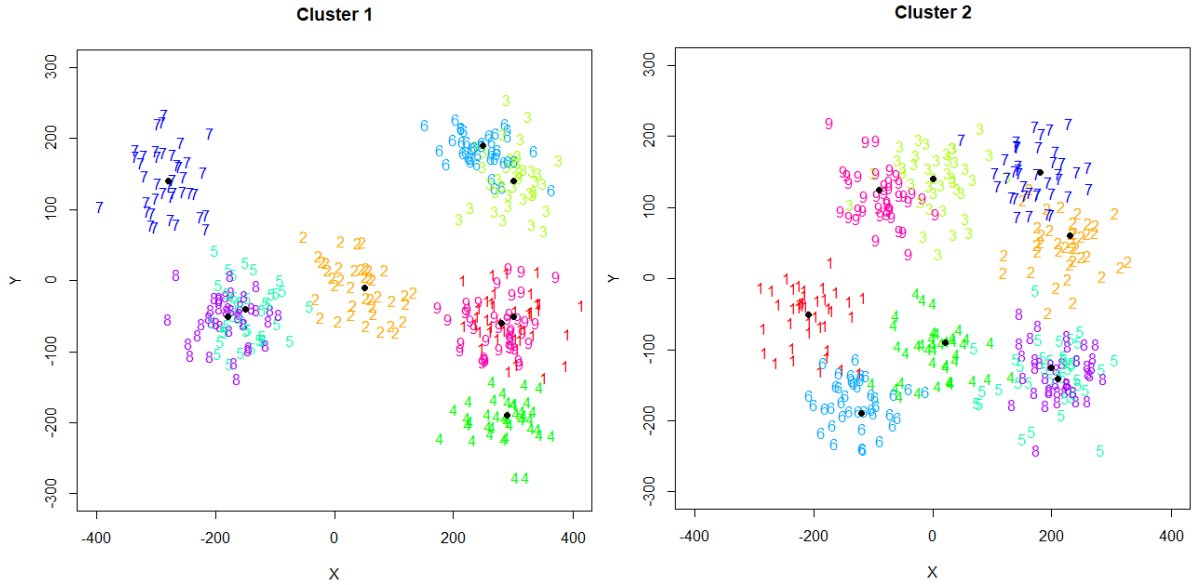


FIGURE 3.8 : Exemple de données simulées pour la classe 1 (gauche) et la classe 2 (droite). Les chiffres 1 à 9 correspondent aux produits simulés. Les points noirs représentent les produits initiaux.

Pour chaque situation simulée, nous avons exécuté CLUSTATIS avec l'option « $K+1$ ». Le paramètre ρ a été calculé à l'aide de l'équation 3.4. Pour toutes les données simulées, nous avons obtenu 100% d'identification correcte. Cela signifie que toutes les classes, y compris la classe « $K + 1$ » (classe 3), ont été parfaitement retrouvées. Ce qui n'est pas le cas de Proclustrees ni de SCR, même si leur taux d'erreur est inférieur à 2% (Berget *et al.*,

(2019). Il est à noter qu'une étude de simulation concernant CLUSTATIS est également présentée dans le chapitre 5 en liaison avec le problème de détermination du nombre de classes.

3.4.2 Tri libre

Les données utilisées pour illustrer l'application de la méthode CLUSTATIS sur des données de tri libre sont disponibles dans le package *R* nommé FreeSortR (Courcoux, 2017). Elles concernent 16 arômes soumis à une épreuve de catégorisation selon la procédure de tri libre par 31 étudiants d'Oniris à Nantes. Les seize arômes, qui sont connus pour être perceptibles dans les vins blancs ou rouges, ont été sélectionnés dans le coffret « Le nez du vin » (Edition Jean Lenoir, 2006). Nous pouvons citer quelques exemples d'arômes : citron, pamplemousse, miel, beurre, pain grillé, noix grillée, *etc.* Les arômes, codés par un nombre aléatoire à trois chiffres, ont été présentés simultanément aux sujets dans des flacons de verre selon une disposition aléatoire. Chaque sujet a formé des groupes avec les 16 arômes en fonction des similitudes et des différences perçues.

Après avoir opéré le pré-traitement indiqué dans la section 2.3, la méthode CLUSTATIS a été appliquée. En traçant le dendrogramme et la variation du critère au cours de l'algorithme hiérarchique (Figure 3.9), il apparaît que cinq classes peuvent être considérées. Le seuil ρ , sélectionné avec l'équation 3.4, est égal à 0.66. Avec ce seuil, quatre sujets sur 31 ont été placés dans la classe « $K + 1$ ». Le Tableau 3.5 indique l'homogénéité de chacune des classes avant et après la mise à l'écart des sujets atypiques. En particulier, nous pouvons remarquer que la mise à l'écart des quatre sujets a entraîné une augmentation de l'indice d'homogénéité globale qui est passé de 57.0% à 60.8%. Ces indices sont nettement supérieurs à l'homogénéité du panel entier (43.9%).

Classe	Sans classe « $K + 1$ »		Avec classe « $K + 1$ »	
	Homogénéité	Nb sujets	Homogénéité	Nb sujets
1	54.3%	6	58.4%	5
2	48.7%	5	55.5%	4
3	58.0%	4	58.0%	4
4	56.8%	13	60.8%	11
5	75.7%	3	75.7%	3
Homogénéité globale	57.0%	31	60.8%	27

Tableau 3.5 : Données « arômes » : Taille, homogénéité de chaque classe et homogénéité globale obtenues en utilisant CLUSTATIS avec et sans classe « $K + 1$ ».

La Figure 3.10 montre la configuration des arômes associée à la première classe de

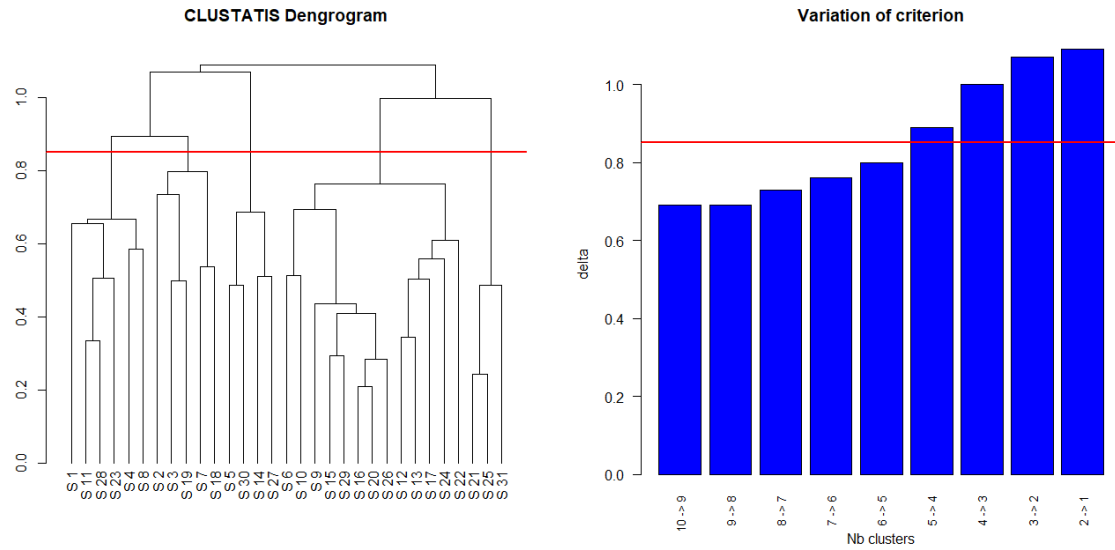


FIGURE 3.9 : Données « arômes » : Dendrogramme et variation du critère D provenant de l’algorithme hiérarchique de CLUSTATIS.

sujets avec et sans l’option « $K + 1$ ». La principale différence entre ces deux configurations concerne la position de l’arôme « beurre ». Ce changement peut s’expliquer par le fait que le sujet placé dans la classe « $K + 1$ » a formé un groupe de produits contenant l’arôme « beurre » et l’arôme « pain grillé », alors qu’aucun autre sujet de cette classe n’avait placé ces deux arômes ensemble. Par conséquent, lorsque ce sujet a été mis de côté, les deux arômes se sont éloignés l’un de l’autre sur la représentation graphique.

3.5 Conclusion

La méthode CLUSTATIS, dédiée à la classification de tableaux de données, vise à minimiser un critère qui est une extension de celui de la méthode STATIS dans une perspective de classification. Ainsi, chaque classe est construite autour d’un tableau latent déterminé par STATIS. Deux algorithmes de classification qui se complètent mutuellement ont été proposés. L’algorithme hiérarchique permet de déterminer graphiquement le nombre de classes en scrutant l’évolution du critère de CLUSTATIS. L’algorithme de partitionnement permet de consolider la partition issue de la classification hiérarchique.

Une extension de CLUSTATIS afin de détecter les tableaux atypiques a été proposée. Elle consiste en une modification de l’algorithme de partitionnement, en fixant un seuil minimal de similarité entre chacun des tableaux de données avec le tableau latent de sa classe. La détermination de ce seuil est importante dans ce type de procédures. Nous avons

3.5 Conclusion

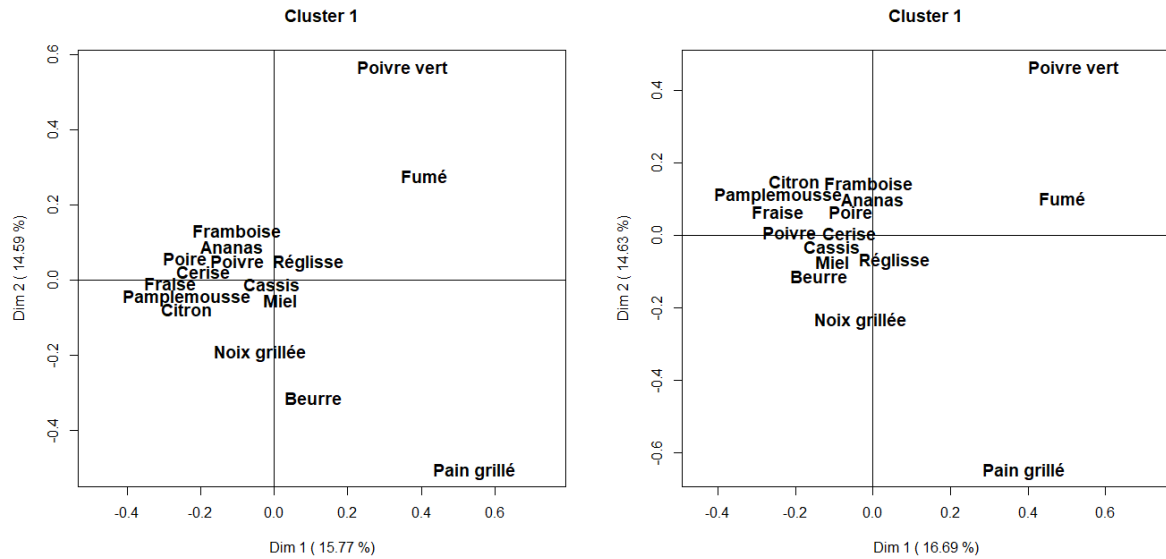


FIGURE 3.10 : Données « arômes » : Représentation des produits sur la base des deux premières composantes de STATIS associées à la classe 1 obtenue par l'algorithme de partitionnement de CLUSTATIS sans option « $K + 1$ » (gauche) et avec option « $K + 1$ » (droite).

proposé de déterminer de manière automatique le seuil en tenant compte de la compacité des classes ainsi que de leur proximité.

Un exemple d'application sur des données de Projective mapping/Napping concernant des smoothies a été présenté. Il a permis de montrer que des classes de sujets ayant des perceptions différentes des produits peuvent être identifiées. Cependant, des sujets atypiques ont été détectés, à la fois sur la base des coefficients RV entre les tableaux de données et le compromis de la classe à laquelle ils appartiennent, et automatiquement par CLUSTATIS avec l'option de la classe « $K + 1$ ». L'importance de cette option a ainsi été démontrée, notamment en termes de gain d'homogénéité.

Afin de comparer l'application de CLUSTATIS sur des données de Projective mapping/Napping avec d'autres stratégies existantes, une étude de cas et des simulations ont été réalisées. L'étude de cas a montré que l'homogénéité des classes construites était clairement meilleure avec CLUSTATIS qu'avec les deux autres stratégies étudiées. Les simulations ont donné des résultats très satisfaisant pour CLUSTATIS, en comparaison de ceux des méthodes SCR et Proclustrees.

La stratégie CLUSTATIS a également été appliquée à des données issues d'une épreuve de tri libre, afin de procéder à une classification des sujets. Nous avons pu remarquer que

l'influence d'un seul sujet atypique pouvait être forte sur les résultats de l'analyse. Ainsi, il paraît essentiel de procéder à la détection de ces sujets atypiques et de les mettre à l'écart de l'analyse.

Chapitre 4

Données Check-All-That-Apply (CATA)

4.1 Introduction

Les épreuves d'évaluation sensorielle impliquant directement les consommateurs sont de plus en plus populaires. Parmi ces épreuves, nous nous sommes déjà intéressés à l'évaluation de produits par catégorisation (tri libre) et à l'épreuve nommée Projective mapping/Napping. Une autre épreuve d'évaluation qui connaît une popularité grandissante est l'épreuve appelée Check-All-That-Apply (CATA). En effet, cette épreuve est facile à mettre en œuvre : les sujets évaluent chacun des produits en cochant les attributs qui s'appliquent pour le décrire. Ainsi, nous obtenons un tableau de données $\mathbf{X}_i$ par sujet, ces tableaux étant apparentés par produits (lignes) et par attributs (colonnes). Nous avons ainsi affaire à un cube de données binaires $n \times p \times m$, avec n produits, p attributs et m sujets (Varela et Ares, 2014). La $(l, j)^e$ entrée du tableau $\mathbf{X}_i$ indique si le sujet considéré a coché l'attribut figurant dans la j^e colonne pour le produit de la l^e ligne ($\mathbf{X}_{i,l,j} = 1$) ou pas ($\mathbf{X}_{i,l,j} = 0$).

Dans un premier temps, nous discutons des indices d'accord entre deux sujets ainsi que pour l'ensemble du panel. Des tests de consistance, afin de déterminer si un accord significatif existe entre les sujets de manière globale et par attribut, sont ensuite présentés. Dans un second temps, nous proposons la méthode CATATIS, qui consiste en une adaptation de STATIS pour l'analyse exploratoire de ce type de données. La stratégie CLUSCATA, qui permet de segmenter les sujets provenant d'une épreuve CATA, est ensuite introduite. Cette méthode offre également la possibilité de détecter les sujets atypiques. Enfin, des exemples d'application sont discutés.

4.2 Accord entre les sujets

4.2.1 Accord entre deux sujets

La matrice associée au sujet i est une matrice binaire qui indique si chaque attribut (en colonnes) a été coché ou non pour les produits (en lignes). Elle a pour norme $\sqrt{n_i}$, où n_i est le nombre total de fois que le sujet i a coché les attributs pour l'ensemble des n produits. Nous proposons de standardiser la matrice associée au sujet i en la divisant par sa norme. Ce type de standardisation permet de mettre les sujets sur un pied d'égalité en tenant compte du fait que certains sujets ont tendance à cocher davantage d'attributs que d'autres. Dans la suite, les tableaux standardisés seront notés $\mathbf{X}_i$ ($i = 1, \dots, m$).

Soit $\mathbf{X}_j$ un tableau supposé standardisé associé au sujet j . Le produit scalaire entre $\mathbf{X}_i$ et $\mathbf{X}_j$ est un indice de similarité entre ces deux matrices. Il est donné par $s(\mathbf{X}_i, \mathbf{X}_j) = \langle \mathbf{X}_i, \mathbf{X}_j \rangle = \text{trace}(\mathbf{X}_i \mathbf{X}_j^\top) = \frac{n_{ij}}{\sqrt{n_i} \sqrt{n_j}}$, où n_{ij} est le nombre d'attributs cochés pour les mêmes produits par les deux sujets i et j . Ce coefficient de similarité est connu sous le nom d'indice d'Ochiai (Ochiai, 1957), et, dans le domaine du traitement de l'information, sous le nom de cosinus de Salton (Salton et McGill, 1983). De toute évidence, le coefficient $s(\mathbf{X}_i, \mathbf{X}_j)$ se situe entre 0 et 1. Il est égal à 0 lorsque $n_{ij} = 0$ (désaccord complet) et il est égal à 1 lorsque $\mathbf{X}_i = \mathbf{X}_j$ (accord parfait). Il est clair que $s(\mathbf{X}_i, \mathbf{X}_j)$ ne considère que les accords positifs (*i.e.*, les accords sur les attributs cochés) entre $\mathbf{X}_i$ et $\mathbf{X}_j$ et ne tient pas compte des accords négatifs (c'est-à-dire les accords sur les attributs non cochés). Cette caractéristique est partagée par plusieurs indices de similarité pour ce type de données (Warrens *et al.*, 2008). Parmi ces indices, le plus connu est certainement celui de Jaccard (Jaccard, 1901) dont l'expression est donnée par $J(\mathbf{X}_i, \mathbf{X}_j) = \frac{n_{ij}}{n_i + n_j - n_{ij}}$. Hamers *et al.* (1989) ont fait une comparaison des indices d'Ochiai et de Jaccard et ont conclu que ces deux indices sont très liés.

Comme indiqué ci-dessus, l'indice d'Ochiai entre deux sujets i et j apparaît comme le produit scalaire entre leurs tableaux $\mathbf{X}_i$ et $\mathbf{X}_j$. Cela permet également d'évaluer leur proximité sur la base de leur distance euclidienne associée à ce produit scalaire :

$$d(\mathbf{X}_i, \mathbf{X}_j) = \|\mathbf{X}_i - \mathbf{X}_j\| = \sqrt{\text{trace}((\mathbf{X}_i - \mathbf{X}_j)(\mathbf{X}_i - \mathbf{X}_j)^\top)} = \sqrt{2(1 - s(\mathbf{X}_i, \mathbf{X}_j))} \quad (4.1)$$

Nous retrouvons une mesure de dissimilarité entre les données binaires connue sous le nom de distance de Hellinger (Wijaya *et al.*, 2016). Cette distance est généralement utilisée comme mesure de ressemblance dans les études écologiques (Conde et Domínguez, 2018).

4.2 Accord entre les sujets

Il est à noter qu'il s'agit ici de la distance de Hellinger entre deux matrices binaires (standardisées), à ne pas confondre avec la distance de Hellinger utilisée pour l'analyse des tableaux de données de type CATA comme alternative à la distance du χ^2 (Meyners et Castura, 2014; Vidal *et al.*, 2015).

4.2.2 Accord global

Nous considérons les données d'une épreuve CATA dans laquelle m sujets ont évalué n produits en utilisant p attributs, ces données étant codées selon la stratégie décrite ci-dessus. Nous désignons par $\mathbf{X}_i$ le tableau standardisé associé au sujet i . La matrice $m \times m$ contenant les coefficients de similarité d'Ochiai entre les sujets est désignée par $\mathbf{S}$. Les éléments diagonaux de $\mathbf{S}$ sont égaux à 1 et les éléments extra-diagonaux sont compris entre 0 et 1. Il s'ensuit que la plus grande valeur propre de la matrice $\mathbf{S}$ que nous désignons par λ_1 varie de 1 à m . Un indice global d'homogénéité entre les sujets est donné par $I = \lambda_1/m$. Cet indice varie entre $1/m$ et 1. Plus ce coefficient est grand, plus l'accord entre les sujets est élevé.

Désignons par $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)^\top$ le vecteur propre de $\mathbf{S}$ associé à la plus grande valeur propre λ_1 . Puisque toutes les entrées de $\mathbf{S}$ sont non négatives, toutes les valeurs du vecteur $\boldsymbol{\alpha}$ sont de même signe, qui peut être choisi comme positif (Meyer, 2000). La i^{e} entrée du vecteur $\boldsymbol{\alpha}$ reflète le niveau d'accord du sujet i avec le point de vue global du panel. Ces coefficients permettent donc de classer les sujets en fonction de leur accord avec reste du panel. De toute évidence, leur interprétation dépend de la valeur de l'indice d'accord global I . En particulier, les petites valeurs indiquent des sujets atypiques. La méthode CATATIS, présentée dans la section 4.3, utilise les composantes du vecteur $\boldsymbol{\alpha}$ afin de calculer un tableau compromis.

4.2.3 Tests de consistance

Nous proposons d'évaluer la significativité de l'accord entre les sujets sur la base de l'indice d'accord global I . En d'autres termes, il s'agit d'évaluer l'hypothèse $H_0 : I = \frac{1}{m}$ (absence totale d'accord entre les sujets) contre l'hypothèse $H_1 : I > \frac{1}{m}$. Le test d'hypothèses que nous proposons peut être effectué à l'échelle de chaque attribut ou à l'échelle globale. Si à l'échelle globale, l'hypothèse H_0 n'est pas rejetée, cela signifie que les évaluations des sujets sont inconsistantes. Si à l'échelle d'un attribut, l'hypothèse H_0 n'est pas rejetée, cela signifie que cet attribut a été mal compris par les sujets ou a été interprété différemment d'un sujet à un autre.

Afin de mettre en œuvre le test d’hypothèses présenté ci-dessus, nous proposons d’effectuer un test de permutations consistant dans un premier temps à permuter aléatoirement les lignes de chaque tableau $\mathbf{X}_i$. Ces permutations conduisent aux tableaux $\mathbf{X}_i^*$ ($i = 1, \dots, m$). Ensuite, l’accord entre les tableaux $\mathbf{X}_i^*$ est évalué au moyen de $I^* = \frac{\lambda_1^*}{m}$, où λ_1^* est la plus grande valeur propre de la matrice $\mathbf{S}^* = (s(\mathbf{X}_i^*, \mathbf{X}_j^*))_{i,j=1,\dots,m}$. Cette procédure de permutations est répétée un grand nombre de fois (*e.g.*, 1000 fois). La distribution des valeurs I^* ainsi obtenues représente la distribution de la statistique de test sous H_0 . Cela signifie que nous pouvons calculer une pseudo p -value, donnée par la proportion des valeurs simulées I^* supérieures ou égales à la valeur observée I . Par conséquent, l’hypothèse H_0 sera rejetée si cette pseudo p -value est inférieure à un seuil de significativité choisi par le praticien (*e.g.*, 5%). Il est clair que ce test suit la même démarche que celle préconisée pour tester la significativité du coefficient RV (Kazi-Aoual *et al.*, 1995; Josse *et al.*, 2008).

Pour le cas d’un attribut, le test Q de Cochran (Cochran, 1950; Meyners *et al.*, 2013) permet de tester la significativité de la différence entre les produits. Ce test est conceptuellement basé sur un principe de test de permutations même si une distribution de probabilité théorique est associée à la statistique de test proposée par Cochran. Nous pensons que le test proposé ici répond à une problématique différente de celle du test de Cochran, mais qu’un lien existe entre ces deux problématiques. En effet, si à l’issue du test que nous proposons, nous concluons qu’il y a une absence de cohérence entre les sujets (*i.e.*, hypothèse H_0) alors il y a une forte probabilité que les produits ne soient pas discriminés. En revanche, si le test préconise le rejet de H_0 , il est possible que les produits ne soient pas discriminés car, naturellement, les sujets pourraient être tout à fait d’accord sur le fait que les produits ne sont pas différents.

4.3 Analyse exploratoire : la méthode CATATIS

La pratique usuelle pour l’analyse exploratoire des données CATA consiste à considérer les tableaux de données binaires indiquant pour chaque produit si les sujets ont coché les attributs ou non. Ces tableaux sont moyennés selon les sujets; ce qui débouche sur un tableau de fréquences croisant les produits et les attributs. Chaque entrée de ce tableau indique la proportion de sujets qui ont coché un attribut pour un produit donné. Ce tableau est ensuite soumis à l’Analyse Factorielle des Correspondances (AFC) afin de représenter les relations entre les produits et les attributs (Jaeger *et al.*, 2015; Meyners *et al.*, 2013). Nous proposons un raffinement de cette stratégie d’analyse dans laquelle, à l’instar de

4.3 Analyse exploratoire : la méthode CATATIS

STATIS, une pondération des sujets est introduite. Ces pondérations tiennent compte de l'accord des sujets avec le point de vue global. Cela signifie que les sujets qui diffèrent des autres dans leur évaluation des produits sont sous-pondérés. En conséquence, le résultat est plus robuste que celui obtenu par la stratégie usuelle d'analyse où tous les sujets sont pondérés de manière égale. Cette méthode suit un schéma similaire à celui de la stratégie STATIS présentée dans la section 2.2. De ce point de vue, CATATIS est une adaptation de STATIS au cas des données CATA.

Notons $\mathbf{X}_i$ le tableau standardisé associé au sujet i , et $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)^\top$ le premier vecteur propre de la matrice de similarité $\mathbf{S}$ composée des indices d'Ochiai entre les sujets pris deux à deux. Ce vecteur détermine la pondération des sujets dans CATATIS. Ainsi, comme décrit dans la section 4.2.2, plus l'accord d'un sujet avec le reste du panel est élevé, plus son poids sera élevé. Les sujets atypiques sont ainsi sous-pondérés et n'influent que très peu la construction du tableau compromis $\mathbf{C} = \alpha_1 \mathbf{X}_1 + \dots + \alpha_m \mathbf{X}_m$. Enfin, ce tableau compromis est soumis à une AFC. La Figure 4.1 résume les différentes étapes de CATATIS.

Formellement, la méthode CATATIS consiste à minimiser la quantité suivante :

$$L = \sum_{i=1}^m \|\mathbf{X}_i - \alpha_i \mathbf{C}\|^2 \quad (4.2)$$

avec comme contrainte de détermination $\sum_{i=1}^m \alpha_i^2 = 1$.

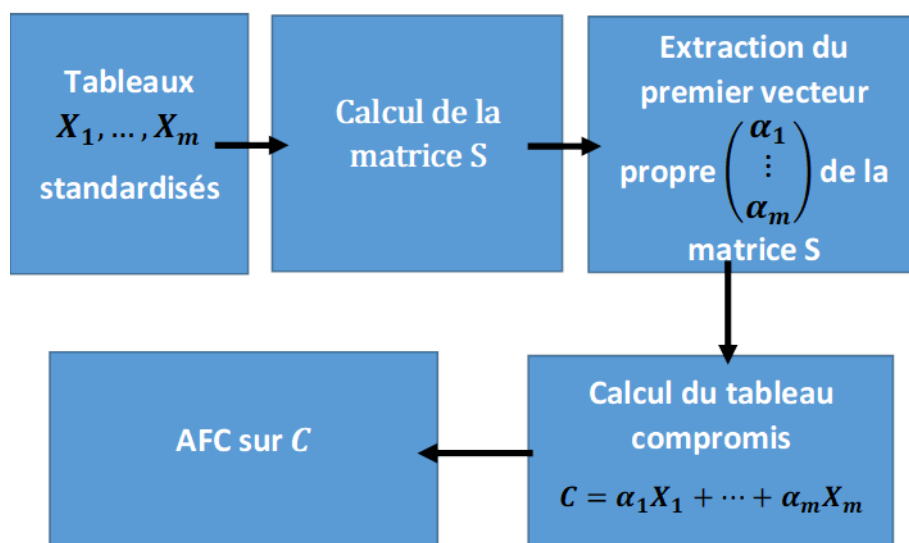


FIGURE 4.1 : Description de la méthode CATATIS.

De manière similaire à STATIS, nous pouvons citer les propriétés suivantes, dont les démonstrations suivent exactement le même raisonnement que celles de l'Annexe [A.1](#) :

- $\|\mathbf{C}\|^2 = \sum_{i=1}^m s^2(\mathbf{X}_i, \mathbf{C}) = \lambda_1$
- $\alpha_i = \frac{\text{trace}(\mathbf{X}_i \mathbf{C})}{\|\mathbf{C}\|^2} = \frac{s(\mathbf{X}_i, \mathbf{C})}{\sqrt{\lambda_1}}$
- $\sum_{i=1}^m \|\mathbf{X}_i - \alpha_i \mathbf{C}\|^2 = m - \sum_{i=1}^m s^2(\mathbf{X}_i, \mathbf{C}) = m - \lambda_1$
- $\sum_{i=1}^m \|\mathbf{X}_i\|^2 = m = \|\mathbf{C}\|^2 + \sum_{i=1}^m \|\mathbf{X}_i - \alpha_i \mathbf{C}\|^2 = \lambda_1 + \sum_{i=1}^m \|\mathbf{X}_i - \alpha_i \mathbf{C}\|^2$

où λ_1 est la première valeur propre de la matrice $\mathbf{S}$. Par conséquent, minimiser L revient à maximiser $\|\mathbf{C}\|^2 = \sum_{i=1}^m s^2(\mathbf{X}_i, \mathbf{C})$. Enfin, ces propriétés indiquent que le pourcentage de la variation expliquée par le modèle CATATIS est donné par $I = \frac{\|\mathbf{C}\|^2}{\sum_{i=1}^m \|\mathbf{X}_i\|^2} = \frac{\lambda_1}{m}$, qui a été présenté comme un indice d'homogénéité dans la section [4.2.2](#).

4.4 La méthode CLUSCATA

Comme les sujets impliqués dans une épreuve CATA ne suivent généralement pas de formation spécifique, ils peuvent différer dans leur évaluation des produits. L'objectif de la méthode CLUSCATA est d'explorer l'existence de classes homogènes de sujets et de les identifier. Cette méthode est étroitement liée à CATATIS. En effet, elle s'appuie sur un critère de minimisation qui est une extension du critère L de CATATIS (équation [4.2](#)) au cas de plus d'un segment de sujets. CLUSCATA suit la même stratégie générale d'analyse que la méthode CLV ([Vigneau et Qannari, 2003](#)) et CLUSTATIS (chapitre [3](#)). Plus précisément, CLUSCATA vise à déterminer des classes homogènes de sujets de manière à ce que les sujets dans chaque classe soient aussi proches que possible d'un tableau compromis déterminé au moyen de la méthode CATATIS.

Formellement, CLUSCATA a pour objectif de minimiser le critère suivant :

$$Q = \sum_{k=1}^K \sum_{i \in G_k} \|\mathbf{X}_i - \alpha_i \mathbf{C}^{(k)}\|^2 \quad (4.3)$$

où G_k est la k^{e} classe et $\mathbf{C}^{(k)}$ est le tableau compromis de la classe G_k . Enfin, α_i ($i \in G_k$) sont des scalaires à déterminer sous la contrainte $\sum_{i \in G_k} \alpha_i^2 = 1$. Naturellement, dans le cas où il n'y a qu'une seule classe, le même critère que CATATIS est retrouvé.

Le critère d'optimisation ci-dessus présente une grande similitude avec celui de CLUSTATIS. Toutefois, CLUSTATIS utilise les matrices des produits scalaires associées aux tableaux de données à regrouper, alors que CLUSCATA opère directement sur les tableaux

4.4 La méthode CLUSCATA

de données eux-mêmes. Cela permet de tenir compte du fait que les données CATA associées aux divers sujets se rapportent non seulement aux mêmes produits (lignes), mais aussi aux mêmes attributs (colonnes).

Pour résoudre le problème d'optimisation de CLUSCATA, nous proposons deux stratégies de classification : un algorithme hiérarchique (Algorithme 4) et un algorithme de partitionnement (Algorithme 5). Les deux stratégies visent à minimiser le critère Q (équation 4.3) et peuvent être exécutés l'un après l'autre. Comme pour CLUSTATIS, l'algorithme hiérarchique peut indiquer un choix du nombre de classes, K . Il consiste en une stratégie ascendante où, à chaque étape, deux classes de tableaux sont fusionnées. En suivant exactement le même raisonnement que pour CLUSTATIS, nous pouvons montrer que l'indice d'agrégation de deux classes A et B se traduit par une augmentation du critère Q égale à $\Delta = Q_{A \cup B} - Q_{A+B} = \lambda_1^{(A)} + \lambda_1^{(B)} - \lambda_1^{(A \cup B)}$ où $\lambda_1^{(A)}$, $\lambda_1^{(B)}$ et $\lambda_1^{(A \cup B)}$ sont respectivement les plus grandes valeurs propres des matrices de similarité $\mathbf{S}$ des classes A , B et $A \cup B$. Ainsi, la stratégie d'agrégation consiste à fusionner, à chaque étape, les classes A et B pour lesquelles la variation Δ est la plus petite. Par la suite, l'algorithme de partitionnement est opéré en considérant la partition en K classes obtenue au moyen de l'algorithme hiérarchique comme point de départ, afin de tenter d'améliorer la solution obtenue par cet algorithme. Cependant, utiliser l'algorithme de partitionnement avec plusieurs partitions aléatoires et retenir, *in fine*, la solution qui minimise le plus Q est également possible (stratégie multi-start).

Algorithme 4 Algorithme hiérarchique de CLUSCATA.

Initialisation : Chaque tableau associé à un sujet forme une classe. $Q = 0$.

tant que il y a plus d'une classe **faire**

Regroupement des deux classes A et B conduisant à la plus faible augmentation Δ de Q .

fin tant que

Tous les tableaux sont dans la même classe.

Il est à noter que nous avons la décomposition suivante :

$$\sum_{i=1}^m \|\mathbf{X}_i\|^2 = m = \sum_{k=1}^K \sum_{i \in G_k} \|\mathbf{X}_i - \alpha_i \mathbf{C}^{(k)}\|^2 + \sum_{k=1}^K \|\mathbf{C}^{(k)}\|^2 \quad (4.4)$$

Cette équation indique que la variation totale dans les tableaux de données $\mathbf{X}_i$ est égale à la somme de la variation expliquée par le modèle de classification en K classes (ou variation inter-classes) : $Z = \sum_{k=1}^K \|\mathbf{C}^{(k)}\|^2$, et de la variation inexpliquée par le modèle (ou variation intra-classes) : $Q = \sum_{k=1}^K \sum_{i \in G_k} \|\mathbf{X}_i - \alpha_i \mathbf{C}^{(k)}\|^2$. Ainsi, minimiser Q revient

Algorithme 5 Algorithme de partitionnement de CLUSCATA.

Initialisation : Donner une partition initiale en K classes.

tant que au moins un tableau change de classe **faire**

pour $k = 1, \dots, K$ **faire**

 Application de CATATIS pour calculer les tableaux compromis $\mathbf{C}^{(k)}$.

fin pour

 De nouvelles classes de tableaux sont formées en déplaçant chaque tableau, $\mathbf{X}_i$, vers la classe G_k pour laquelle la quantité $\|\mathbf{X}_i - \alpha_i \mathbf{C}^{(k)}\|$ est la plus faible ou, de façon équivalente, la quantité $s(\mathbf{X}_i, \mathbf{C}^{(k)})$ est la plus élevée.

fin tant que

à maximiser la quantité suivante :

$$Z = \sum_{k=1}^K \|\mathbf{C}^{(k)}\|^2 = \sum_{k=1}^K \sum_{i \in G_k} s^2(\mathbf{X}_i, \mathbf{C}^{(k)}) = \sum_{k=1}^K \lambda_1^{(k)} \quad (4.5)$$

où $\lambda_1^{(k)}$ est la plus grande valeur propre de la matrice $\mathbf{S}$ composée des coefficients s entre les sujets de la classe G_k . Ces égalités découlent des propriétés de CATATIS énoncées dans la section 4.3. Enfin, il résulte des équations 4.4 et 4.5 que le pourcentage de variation expliquée par CLUSCATA est égal à $I = \sum_{k=1}^K \frac{\lambda_1^{(k)}}{m}$. Cet indice représente l'homogénéité globale des classes, et peut être décomposé par classe : $I = \sum_{k=1}^K \frac{m_k I_k}{m}$, où $I_k = \frac{\lambda_1^{(k)}}{m_k}$ est l'homogénéité de la classe G_k , et m_k est le nombre de sujets dans G_k .

4.5 Classification en écartant les sujets atypiques

Il arrive fréquemment que des sujets aient un avis qui diverge de celui de chacune des classes obtenues suite à une classification. Les mettre de côté permet d'augmenter l'homogénéité de chaque classe, et donc l'homogénéité globale. En suivant l'idée de Dave (1991), nous introduisons, en plus des K classes principales, une classe « $K + 1$ » qui contient les sujets jugés atypiques. Cette stratégie nécessite la sélection d'une valeur seuil, ρ , comprise entre 0 et 1, en dessous de laquelle la similarité entre un tableau associé à un sujet et le tableau compromis d'une classe est considérée comme trop faible. Naturellement, l'algorithme de partitionnement cherche à placer un sujet dans la classe pour laquelle cette similarité est la plus grande, mais, si cette similarité est inférieure à la valeur seuil ρ , le sujet est affecté à la classe « $K + 1$ ». Formellement, nous cherchons à maximiser le critère suivant :

$$Z_\rho = \sum_{i=1}^m \sum_{k=1}^K (\delta_{ik} s^2(\mathbf{X}_i, \mathbf{C}^{(k)}) + \delta_{i(K+1)} \rho^2) \quad (4.6)$$

4.6 Exemples d'application

où δ_{ik} est le symbole de Kronecker, égal à 1 si le sujet $\mathbf{X}_i$ appartient à la classe G_k et 0 sinon, avec la contrainte $\sum_{k=1}^{K+1} \delta_{ik} = 1$, stipulant qu'un sujet appartient à une seule et unique classe, y compris à la classe « $K + 1$ ». Pour résoudre ce problème de maximisation, l'Algorithme 6 peut être exécuté.

Algorithme 6 Algorithme de partitionnement de CLUSCATA avec option « $K + 1$ ».

Initialisation : Donner une partition initiale (idéalement à partir de la classification hiérarchique de CLUSCATA) et le seuil minimal de similarité ρ .

tant que au moins un sujet change de classe **faire**

pour $k = 1, \dots, K$ **faire**

 Application de CATATIS pour calculer les tableaux compromis $\mathbf{C}^{(k)}$.

fin pour

 Affectation de chaque tableau à la classe G_k pour laquelle le coefficient $s(\mathbf{X}_i, \mathbf{C}^{(k)})$ est le maximum ou à la classe « $K + 1$ » si $s(\mathbf{X}_i, \mathbf{C}^{(k)}) < \rho$.

fin tant que

Une grande similitude est encore une fois à noter avec la stratégie CLUSTATIS. Cette similitude est corroborée par le calcul du seuil ρ automatique, qui est réalisé avec le même raisonnement que pour CLUSTATIS :

$$\rho = \frac{\sum_{i=1}^m (s(\mathbf{X}_i, \mathbf{C}^{(k_i)}) + s(\mathbf{X}_i, \mathbf{C}^{(k_{i'})}))}{2m} \quad (4.7)$$

où $\mathbf{C}^{(k_i)}$ est le tableau compromis de classe le plus proche du tableau $\mathbf{X}_i$ et $\mathbf{C}^{(k_{i'})}$ est le second tableau compromis le plus proche. La justification de ce choix est de définir une valeur intermédiaire qui délimite des frontières entre les classes, afin que les sujets trop éloignés ou situés entre plusieurs classes soient affectés à la classe « $K + 1$ ».

4.6 Exemples d'application

4.6.1 Données « fraises »

Les données utilisées pour illustrer les stratégies CATATIS et CLUSCATA se rapportent à une épreuve CATA concernant six variétés de fraises qui ont été évaluées par 114 consommateurs au moyen de 16 attributs. Les sujets ont été recrutés dans un supermarché de Montevideo, en Uruguay, alors qu'ils achetaient des fruits et légumes. Ces données sont extraites d'une étude de cas décrite par Ares et Jaeger (2013). Deux sujets de l'étude initiale ont été exclus puisque certaines de leurs données étaient manquantes.

Tests de consistance

En considérant un seuil de 5%, les tests de consistance ont donné un accord général significatif de manière globale (pseudo p -value < 0.001), ainsi que pour tous les attributs pris séparément, exceptés les attributs « saveur de fraise » et « sec ». Pour réaliser ces tests, 1000 permutations ont été considérées. Les résultats sont donnés dans le Tableau 4.1.

Attribut	Pseudo p -value
sucré	< 0.001
acide	< 0.001
ferme	< 0.001
dur	< 0.001
mou	< 0.001
juteux	0.003
forme irrégulière	0.007
forme régulière	< 0.001
petit	< 0.001
grand	< 0.001
odeur de fraise	0.028
saveur de fraise	0.844
savoureux	< 0.001
couleur rouge	0.005
sans goût	0.005
sec	0.191

Tableau 4.1 : Données « fraises » : Pseudo p -value du test de consistance appliqué à chacun des attributs.

Application de CATATIS

À partir de la matrice $\mathbf{S}$ contenant les indices de similarité s entre les sujets pris deux à deux, l'indice d'homogénéité globale I a été calculé. Il est égal à 37.3%, ce qui indique un accord assez faible entre les sujets. Nous avons également calculé les poids des sujets α_i ($i = 1, \dots, 114$), qui reflètent l'accord de chaque sujet avec le point de vue global du panel. Ces indices sont représentés sur la Figure 4.2. Il s'avère que plusieurs sujets semblent avoir des poids relativement faibles, dont un en particulier qui a été doté d'un poids très faible. Un examen des données de ce sujet a montré qu'il ou elle a coché seulement six attributs pour tous les produits. En comparaison, le nombre moyen d'attributs cochés par les sujets est de 28, tous produits confondus. De plus, le sujet en question n'était que très peu en accord avec les autres consommateurs sur la faible quantité d'attributs qu'il a coché.

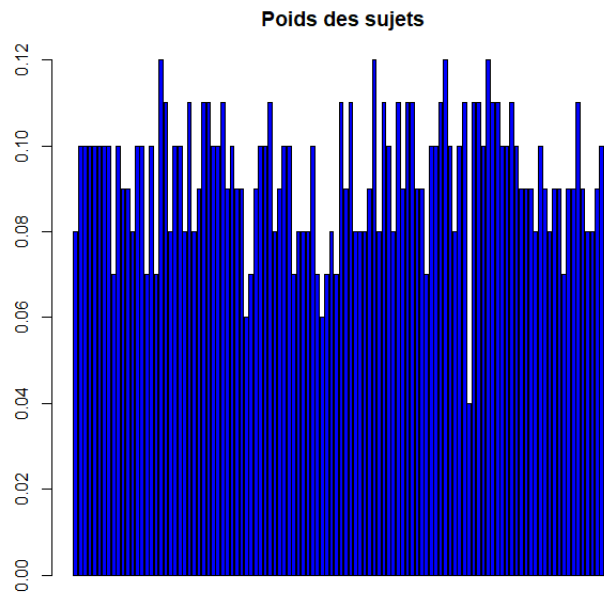


FIGURE 4.2 : Données « fraises » : Poids des sujets donnés par CATATIS.

Le tableau compromis obtenu avec l'application de CATATIS sur ces données a été soumis à l'AFC (Figure 4.3). La fraise Yvahé est considérée comme petite et irrégulière, alors que la fraise K31.5 est perçue comme amère. L'axe 1 est déterminé par une opposition fraise molle / fraise dure alors que l'axe 2 est construit par une opposition fraise / grande fraise.

À titre de comparaison, nous avons également soumis le tableau des fréquences à l'AFC, comme le préconise l'approche courante (Meyners *et al.*, 2013). Le résultat est quasiment identique à la Figure 4.3 (Figure non montrée), la valeur du coefficient RV entre les deux configurations étant presque égale à 1. Afin de mettre en évidence les avantages de CATATIS par rapport à l'approche habituelle, nous avons procédé à une permutation des données de 20 sujets (sur 114) concernant 4 attributs (sur 16). Cela signifie que les données concernant ces 20 sujets pour les 4 attributs sélectionnés sont maintenant totalement incohérentes. Sur les tableaux ainsi obtenus, nous avons effectué une AFC sur le tableau compromis obtenu par CATATIS et sur le tableau de fréquences (approche courante). Cette procédure (*i.e.*, permutation et AFC) a été répétée 100 fois. Sur la Figure 4.4, les ellipses de confiance à 95% autour des produits sont représentées. Il s'avère que les ellipses associées à CATATIS sont plus petites et se chevauchent moins que celles associées à l'approche courante. Cela signifie que la pondération des différents sujets confère une plus grande stabilité à la configuration finale des produits.

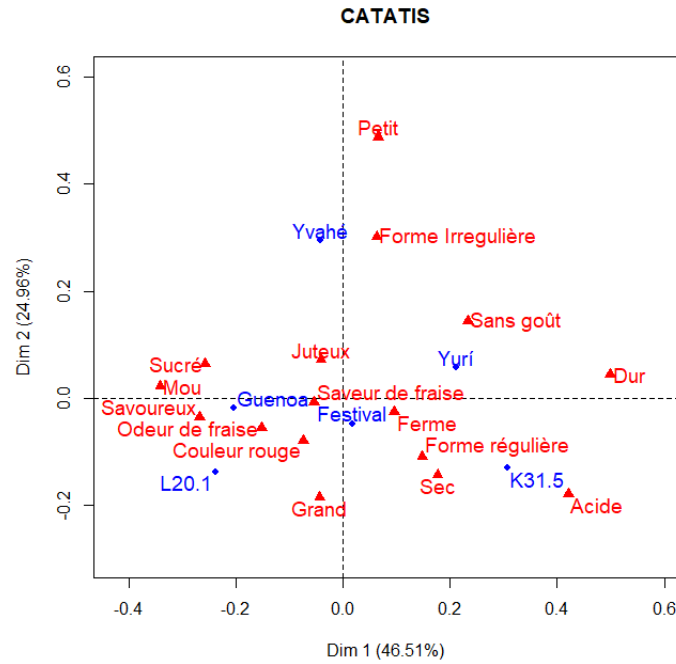


FIGURE 4.3 : Données « fraises » : Représentation du tableau compromis de CATATIS sur les deux premiers axes de l'AFC.

Application de CLUSCATA

La méthode CLUSCATA a été appliquée sur les données « fraises » afin d'effectuer une segmentation des consommateurs. Le dendrogramme, qui résulte de l'algorithme hiérarchique, est représenté sur la Figure 4.5. Cette figure montre également la variation Δ du critère d'agrégation lorsque deux classes sont fusionnées. Cette variation doit être aussi petite que possible puisqu'elle reflète l'augmentation de l'hétérogénéité intra-classes résultante du regroupement de ces deux classes. Dans le cas présent, cette variation suggère de considérer quatre classes de sujets, puisqu'elle indique un saut net lors du passage d'une partition en quatre classes à une partition en trois classes.

Dans un second temps, nous avons appliqué l'algorithme de partitionnement avec l'option « $K + 1$ » en considérant la partition en quatre classes obtenue au moyen de l'algorithme hiérarchique comme point de départ. Le Tableau 4.2 indique l'homogénéité de chacune des quatre classes constituées, ainsi que l'homogénéité globale ($I = 49.4\%$), qui est bien supérieure à celle obtenue sans classification ($I = 37.3\%$). Il apparaît également que les sujets atypiques sont très hétérogènes entre eux ($I_{K+1} = 27.5\%$). La mise à l'écart de ces sujets conduit à une augmentation de l'homogénéité globale puisque I passe de 42.4% (si l'on utilise CLUSCATA sans classe « $K + 1$ ») à 49.4%.

4.6 Exemples d'application

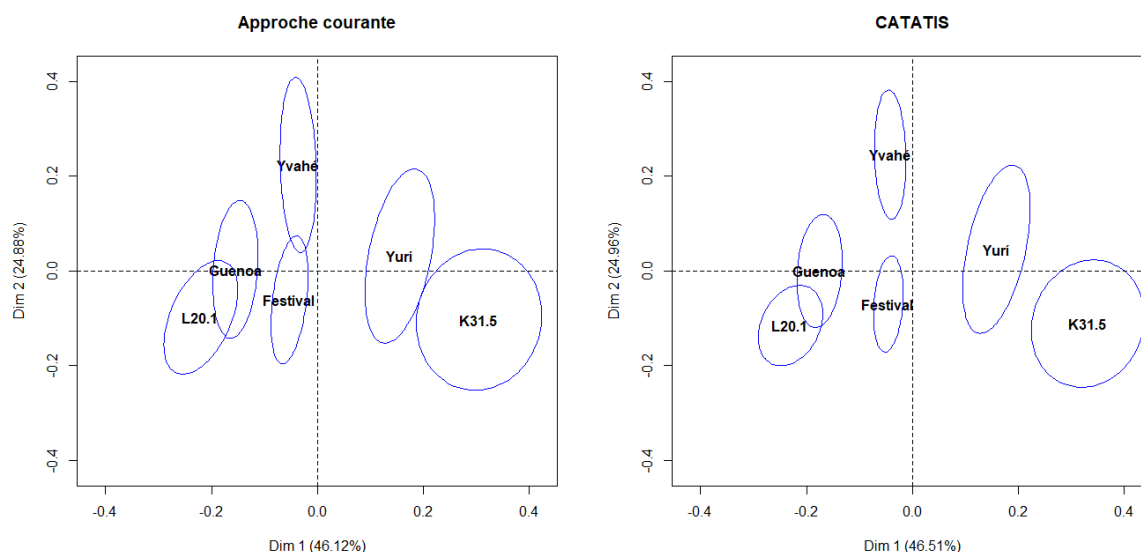


FIGURE 4.4 : Données « fraises » : Ellipses de confiance à 95% obtenues par l'AFC sur le tableau de fréquences (gauche) et par l'AFC sur le tableau compromis obtenu par CATATIS (droite).

Classe	Homogénéité (%)	Nb sujets
1	49.4	24
2	49.4	28
3	49.7	16
4	49.1	8
« $K + 1$ »	27.5	38
Homogénéité globale	49.4	76

Tableau 4.2 : Données « fraises » : Homogénéité de chacune des classes et homogénéité globale obtenues au moyen de CLUSCATA avec l'option « $K + 1$ ».

En plus de la partition des sujets, CLUSCATA procure, pour chaque classe de sujets, un tableau compromis qui croise les produits et les attributs. Chacun de ces tableaux a été soumis à l'AFC (Figure 4.6). Nous pouvons remarquer que des différences existent entre ces représentations graphiques, mettant en évidence des différences de perception des produits par les quatre classes de sujets. La classe 4 se distingue clairement des autres en ne proposant pas d'opposition petite fraise / grande fraise sur le second axe. La fraise L20.1 est perçue comme dure dans la classe 1 alors qu'elle est plutôt indiquée comme molle dans les autres classes. Quant aux sujets de la classe 2, ils semblent associer la variété de fraise Yuri à « sans goût », ce qui est une perception contraire aux autres classes.

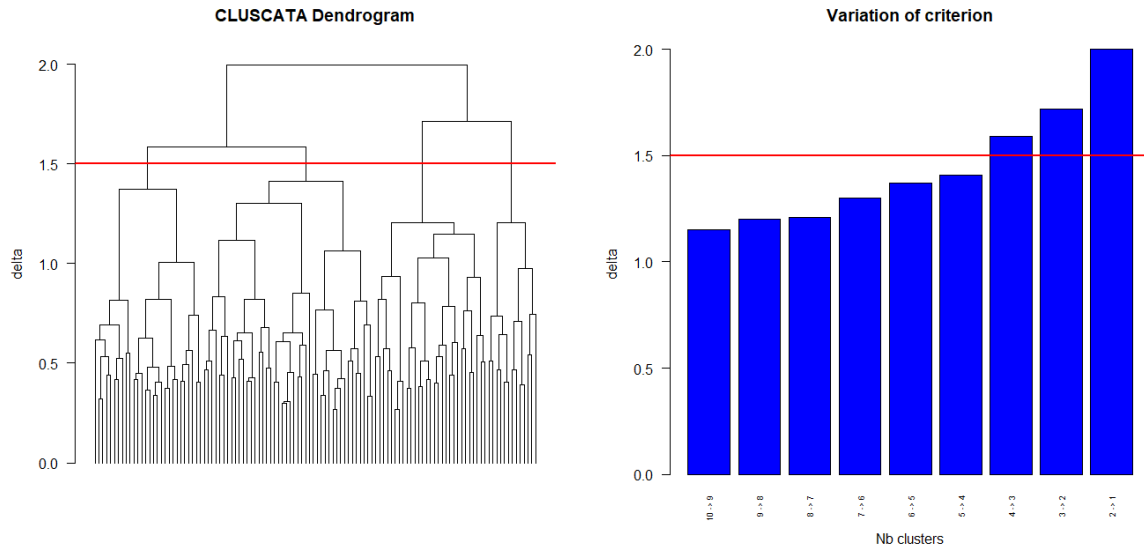


FIGURE 4.5 : Données « fraises » : Dendrogramme obtenu par l'algorithme hiérarchique de CLUSCATA (gauche) et variation du critère d'agrégation au cours de cet algorithme (droite).

4.6.2 Autres jeux de données

Afin d'illustrer la stratégie d'analyse sur une base de données plus grande, nous avons examiné huit autres études de cas provenant de publications antérieures sur lesquelles nous avons appliqué l'approche générale proposée dans ce chapitre. L'objectif est également de donner des ordres de grandeur concernant les indices d'homogénéité, le nombre d'attributs pour lesquels les sujets n'étaient pas consistants, le gain en termes d'homogénéité lorsque nous exécutons CLUSCATA avec et sans l'option « $K + 1$ », le nombre de sujets qui sont affectés à la classe « $K + 1$ », *etc.* Les caractéristiques des études de cas et les références bibliographiques sont indiquées dans le Tableau 4.3. Il est à noter que le nombre de sujets est compris entre 73 et 154, que celui de produits varie de 4 à 12, et, enfin, que le nombre d'attributs se situe entre 10 et 38.

Jeu de données	Produit	# sujets	# produits	# attributs	Référence
D1	Bière	154	6	27	Giacalone <i>et al.</i> (2013)
D2	Bière	73	8	38	Reinbach <i>et al.</i> (2014)
D3	Bière	97	9	15	Giacalone <i>et al.</i> (2015)
D4	Bière	87	4	10	Jaeger <i>et al.</i> (2013)
D5	Vin	112	12	17	Giacalone et Jaeger (2016)
D6	Café	83	6	30	Giacalone <i>et al.</i> (2019)
D7	Pain de seigle	134	6	14	Giacalone (2018)
D8	Bière	76	9	15	Giacalone <i>et al.</i> (2015)

Tableau 4.3 : Description des jeux de données étudiés.

4.6 Exemples d'application

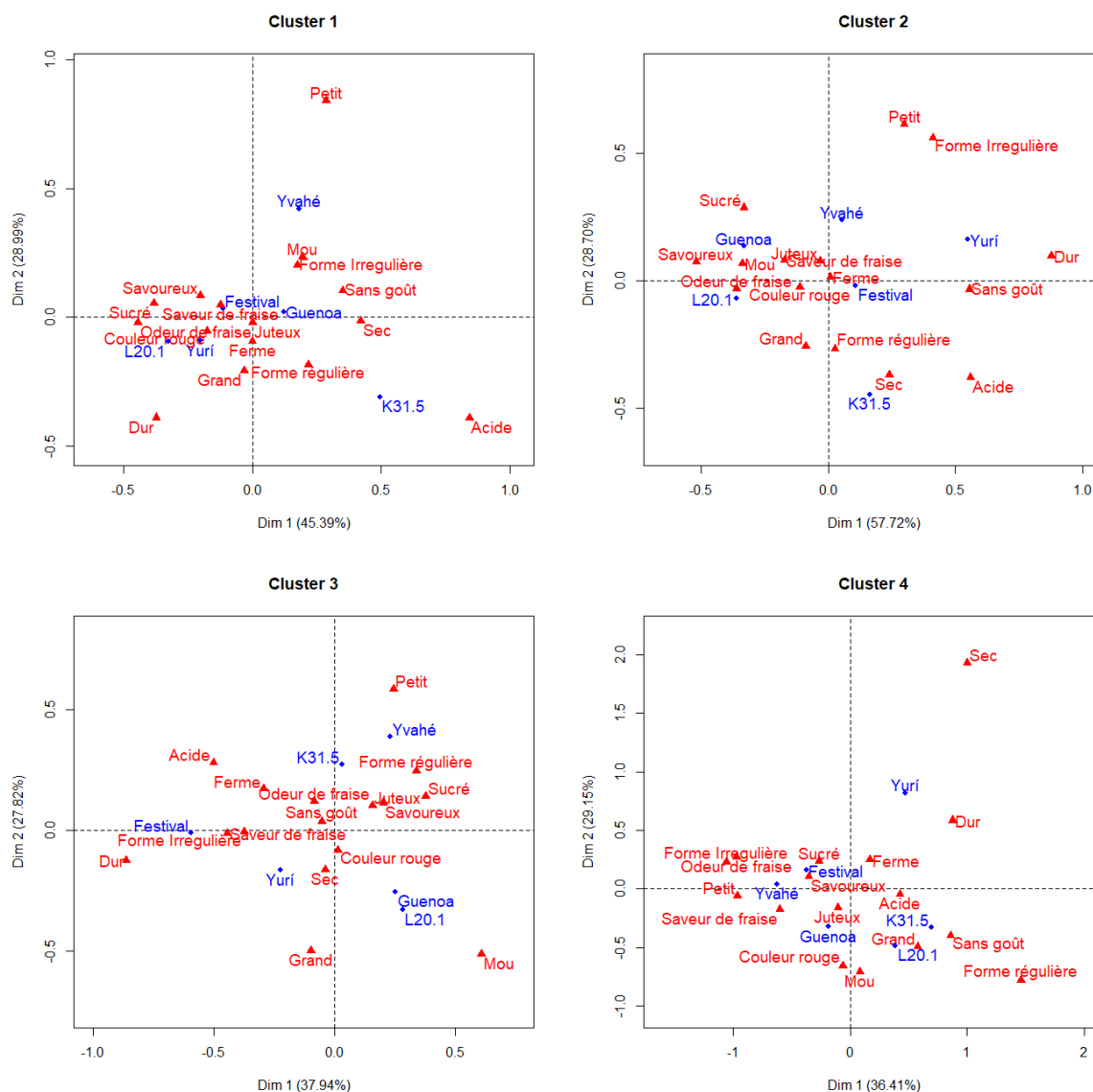


FIGURE 4.6 : Données « fraises » : Représentation du tableau compromis de chacune des quatre classes sur les deux premiers axes de l'AFC.

Dans le Tableau 4.4, les principaux résultats concernant les différentes étapes de la stratégie d'analyse sont exposés. Les indices d'homogénéité se situent entre 21.3% pour l'étude de cas D1 à 70.7% pour les données D4. Cependant, si nous excluons ces deux cas extrêmes, l'homogénéité des autres études de cas est moyenne ou faible. Les tests de permutations appliqués à chaque étude de cas mènent à la conclusion que l'hypothèse nulle stipulant que les sujets sont en complet désaccord est rejetée.

Les tests de permutations concernant la consistance des sujets pour chaque attribut pris séparément ont mis en évidence que dans les études de cas D3, D4, D5, D7 et D8, très peu d'attributs ont révélé un problème de consistance. Cependant, plus de 10 attributs

étaient impliqués dans un problème de consistance pour les études de cas D1, D2 et D6 (jusqu'à 40% des attributs testés). Pour ces études de cas, le nombre initial d'attributs est relativement important et l'accord entre les sujets évalué par l'indice d'homogénéité est faible. Parallèlement au test de permutations appliqué aux divers attributs, nous avons exécuté également le test Q de Cochran pour évaluer leur capacité de discrimination. Les conclusions tirées des deux tests (*i.e.*, les tests Q de Cochran et de permutations) sont les mêmes pour presque tous les 166 attributs présents dans toutes les études de cas. Il y a eu trois exceptions : un attribut a été évalué comme non discriminant mais consistant, et deux attributs ont été jugés inconsistants mais discriminants.

Comme attendu, l'homogénéité globale a augmenté après avoir utilisé CLUSCATA, et cette augmentation est encore plus marquée après la mise à l'écart des sujets jugés atypiques. En moyenne, un tiers des sujets ont été placés dans la classe « $K + 1$ ». Le seuil ρ est logiquement très en lien avec l'homogénéité de CLUSCATA sans classe « $K + 1$ », puisque son calcul est basé sur l'homogénéité des différentes classes.

Il est à noter que le nombre de classes a été déterminé visuellement dans chaque cas. Cependant, le jeu de données D4 a une homogénéité très élevée. Ce qui implique qu'une seule classe aurait pu être considérée. Ainsi, une classification des sujets n'est pas toujours nécessaire, et peut impliquer que des sujets soient considérés à tort comme atypiques. Afin de contrer ce genre de problèmes, un test d'hypothèses pour déterminer s'il y a plus d'une classe est présenté dans le chapitre 5.

Jeu de données	Homogénéité	# attributs inconsistants (%)	# classes	Hom. suite à CLUSCATA (sans classe « $K + 1$ »)	Seuil ρ	# sujets atypiques (%)	Hom. suite à CLUSCATA (avec classe « $K + 1$ »)
D1	21.3%	10 (37.0%)	3	24.8%	0.45	53 (34.4%)	30.5%
D2	38.2%	11 (28.9%)	3	42.7%	0.61	20 (27.4%)	48.2%
D3	45.5%	0 (0%)	4	51.8%	0.67	33 (34.0%)	60.1%
D4	70.7%	1 (10%)	3	74.7%	0.83	22 (29.9%)	79.8%
D5	48.5%	1 (5.9%)	2	51.1%	0.66	39 (34.8%)	62.8%
D6	31.8%	12 (40.0%)	3	35.8%	0.55	22 (26.5%)	40.7%
D7	31.7%	3 (21.4%)	3	37.3%	0.55	43 (32.1%)	44.1%
D8	46.8%	0 (0%)	4	54.0%	0.68	18 (23.7%)	61.7%

Tableau 4.4 : Principaux résultats des diverses stratégies d'analyse.

4.7 Conclusion

Ce chapitre aborde quatre sujets importants pour l'analyse de données de type CATA : (i) indices de similarité entre les sujets, (ii) significativité de l'accord entre les sujets globalement et par attribut (tests de consistance), (iii) calcul d'un tableau compromis du panel (CATATIS), (iv) segmentation des sujets impliqués dans l'épreuve CATA (CLUSCATA). Ces quatre questions sont étroitement liées, puisque le test de significativité de l'accord entre les sujets, qui est basé sur les indices de similarité entre ces derniers, indique dans le cas de non-significativité que les données ne peuvent pas être exploitées. Dans le cas contraire, les poids attribués aux sujets sont dérivés des indices de similarité et l'approche de classification est basée sur l'idée que les tableaux de données associés aux sujets dans chaque classe doivent être aussi proche que possible du tableau compromis de la classe obtenu au moyen de CATATIS.

Un des avantages de la méthode CATATIS par rapport à la stratégie courante est d'être plus stable et moins sensible aux sujets atypiques. Un autre avantage est de fournir, en plus du tableau compromis, des indices utiles concernant l'accord de chaque sujet avec le point de vue général et concernant l'accord global du panel. S'il s'avère que ce dernier indice est petit, alors une segmentation du panel est conseillée. L'approche de classification proposée (CLUSCATA) est directement liée à la méthode CATATIS, comme cela est souligné par les critères d'optimisation qui définissent ces deux méthodes. En plus de proposer une segmentation des sujets, CLUSCATA fournit des outils pour aider à choisir le nombre de classes et évaluer la pertinence de la solution finale. Enfin, CLUSCATA dispose d'une option permettant d'écarter les sujets atypiques.

La pertinence de la stratégie globale a été démontrée sur la base de plusieurs études de cas. Les résultats ont exhibé des situations où les sujets avaient des perceptions très différentes et des cas présentant un degré d'accord relativement élevé. Le regroupement des sujets permet une amélioration de l'accord global. En ajoutant une classe « $K + 1$ », cette amélioration est encore plus perceptible.

Il est clair que, suivant le nombre de classes, un sujet peut être atypique ou non. Prenons l'exemple d'une classe homogène dans laquelle un sujet est central. Si nous séparons cette classe en deux, le sujet risque alors d'être à la frontière entre les deux classes constituées, et d'être ainsi considéré comme atypique. Par conséquent, le choix du nombre de classes revêt une importance capitale dans la méthode d'analyse. Le chapitre

5 est dédié à ce problème et propose des stratégies pour faire ce choix.

Les études de cas ont également mis en évidence un problème important lié à l'épreuve CATA : le nombre d'attributs cochés tous produits confondus peut être très différent d'un sujet à l'autre. Nous avons proposé une normalisation qui consiste à diviser chaque tableau de données binaires par la racine carrée du nombre total d'attributs cochés pour tous les produits. Néanmoins, il reste que les sujets qui ont tendance à cocher plus d'attributs auront souvent une plus grande contribution que ceux qui cochent les attributs avec parcimonie. Un moyen pour contrer ce problème serait de demander aux sujets de cocher, pour chaque produit, les attributs qui semblent les plus probants, et, en tout état de cause, de ne pas en cocher plus qu'un nombre fixé a priori. Une autre question concerne le nombre d'attributs total qui, lorsqu'il est relativement grand, peut causer la fatigue des sujets. Ce problème est encore plus aigu lorsque le nombre de produits à évaluer est lui-même élevé. En effet, nous pouvons remarquer, à partir des études de cas discutées ci-dessus, qu'en général l'accord entre les sujets est plutôt faible lorsque le nombre d'attributs est grand.

Chapitre 5

Choix du nombre de classes

5.1 Introduction

En classification, le choix du nombre de classes est une décision importante. C'est pourquoi ce problème a fait couler beaucoup d'encre. De nombreux indices et des tests d'hypothèses pour guider l'utilisateur dans ce choix ont été proposés (Milligan et Cooper, 1985; Charrad *et al.*, 2014; Tibshirani *et al.*, 2001). En particulier, les indices de Calinski-Harabasz (Caliński et Harabasz, 1974) et d'Hartigan (Hartigan, 1975) ont retenu notre attention. Nous proposons de les adapter au cadre de la classification de tableaux multiples pour CLUSTATIS et CLUSCATA. Nous proposons également des stratégies de choix du nombre de classes basées sur des tests d'hypothèses.

Avant d'étudier le choix du nombre de classes, il conviendrait de se poser préalablement une question importante : existe-t-il des classes, ou est-ce que l'ensemble des tableaux est suffisamment homogène et ne se prête pas à une classification ? Très souvent, cette question est éludée, ce qui conduit parfois à des décisions erronées. C'est pourquoi, dans un premier temps, nous proposons des tests basés sur des permutations qui permettent d'évaluer s'il y a plus d'une classe de tableaux. Dans le cas d'une réponse affirmative, nous présentons diverses stratégies de détermination du nombre de classes.

La procédure générale est tout d'abord introduite, puis des tests d'hypothèses ainsi que des indices adaptés à CLUSTATIS sont proposés. Nous mentionnerons brièvement comment ces approches peuvent être adaptées au cas de CLUSCATA. Les différentes propositions sont comparées sur la base de données simulées et de données réelles. Ces études seront également une nouvelle occasion pour confirmer l'efficacité de CLUSTATIS et de CLUSCATA.

5.2 Procédure générale

Comme nous l'avons indiqué, avant de procéder à une classification, il convient de vérifier s'il existe une structure des données en plusieurs classes. En d'autres termes, dans le contexte des tableaux multiples, la classification n'est pertinente que si les tableaux ne forment pas un ensemble homogène. Ainsi, la première étape que nous proposons est un test d'hypothèses pour vérifier s'il y a une seule classe (hypothèse $H_0^{(1)}$) ou s'il y a plus d'une classe (hypothèse $H_1^{(1)}$).

Sous réserve du rejet de l'hypothèse nulle, la seconde étape consiste à déterminer le nombre de classes approprié. En plus de déterminer ce nombre visuellement sur la base du dendrogramme, divers indices proposés dans le cadre de la classification d'individus (Everitt *et al.*, 2011) peuvent être adaptés au contexte de CLUSTATIS.

Une autre stratégie pour la sélection du nombre de classes est basée sur des tests d'hypothèses. Elle est mise en œuvre sous forme d'une procédure séquentielle, consistant à tester, à une étape K , l'hypothèse $H_0^{(K)}$ stipulant qu'il y a K classes, contre l'hypothèse $H_1^{(K)}$, indiquant que le nombre de classes est supérieur à K . Si $H_0^{(K)}$ est rejetée, la procédure est la même en augmentant K de 1, et ainsi de suite (Figure 5.1).

5.3 Étape 1 : une classe ou plus d'une classe ?

Nous proposons un test d'hypothèses basé sur des permutations afin de tester l'hypothèse $H_0^{(1)}$: il y a une classe, contre l'hypothèse $H_1^{(1)}$: il y a plus d'une classe. Sous l'hypothèse $H_0^{(1)}$, l'analyse inter-structures de STATIS indique que tous les tableaux sont fortement liés les uns aux autres et nettement reliés à leur compromis. De manière plus précise, elle indique que la matrice de similarité composée des coefficients RV entre les tableaux est de rang égal à 1. De manière équivalente, la seconde valeur propre de cette matrice ne devrait pas être significativement élevée.

Nous nous inspirons d'une démarche proposée par Sahmer *et al.* (2006), qui, pour tester l'unidimensionnalité d'un tableau de données, se sont intéressés à la seconde valeur propre de la matrice des corrélations. Pour le cas de CLUSTATIS, nous considérons la seconde valeur propre de la matrice des coefficients RV entre les tableaux de données. Si nous désignons cette valeur propre par λ_2 , il s'agit de tester l'hypothèse $H_0^{(1)}$: λ_2 n'est pas significativement élevée, contre $H_1^{(1)}$: λ_2 est significativement élevée. Pour cela, nous

5.3 Étape 1 : une classe ou plus d'une classe ?

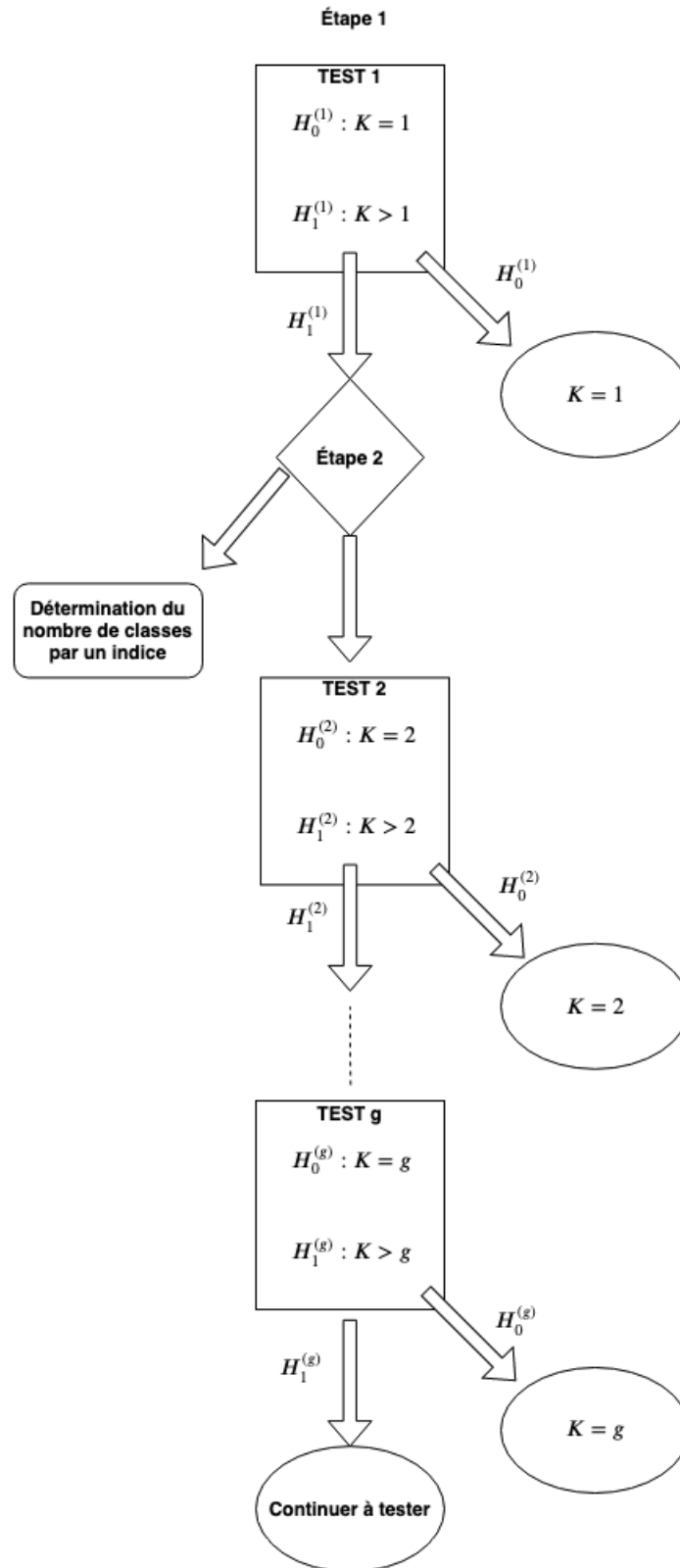


FIGURE 5.1 : Procédure générale de détermination du nombre de classes.

proposons deux procédures de permutations. Chaque procédure conduit à une série de valeurs propres simulées λ_2^* et la décision pour départager les hypothèses $H_0^{(1)}$ et $H_1^{(1)}$

est prise en comparant la valeur réellement observée aux valeurs simulées. De manière plus précise, l'hypothèse $H_0^{(1)}$ est rejetée si la proportion des valeurs simulées qui sont supérieures à la valeur observée, λ_2 , est supérieure à un seuil de significativité α (e.g., $\alpha = 0.05$) fixé par l'utilisateur.

Test 1

La première procédure de permutation consiste à permuter indépendamment les lignes de chaque colonne de chaque tableau. Sur les tableaux ainsi simulés, nous calculons la matrice des coefficients RV, et, par la suite, nous déterminons la deuxième valeur propre de cette matrice.

Test 2

Dans cette seconde procédure, nous proposons d'effectuer les permutations sur les matrices des produits scalaires associées aux tableaux, puisque les coefficients RV sont calculés sur la base de ces matrices. Plus précisément, les éléments de la diagonale supérieure de chaque matrice des produits scalaires sont permutés. Par la suite, les éléments sont dupliqués par symétrie dans la diagonale inférieure. Les coefficients RV entre ces matrices sont calculés, et, comme précédemment, la seconde valeur propre de la matrice constituée de ces coefficients est récupérée. La procédure de permutation proposée ici s'apparente d'une certaine manière à celle préconisée afin de tester la nullité d'un coefficient de corrélation. En effet, il est bien connu (Glaçon, 1981) que le coefficient RV entre deux matrices $\mathbf{W}_i$ et $\mathbf{W}_j$ est égal à $\frac{\mathbf{w}_i^T \mathbf{w}_j}{\sqrt{\mathbf{w}_i^T \mathbf{w}_i} \sqrt{\mathbf{w}_j^T \mathbf{w}_j}}$, où $\mathbf{w}_i$ (respectivement $\mathbf{w}_j$) est le vecteur obtenu en empilant les colonnes de la matrice $\mathbf{W}_i$ (respectivement $\mathbf{W}_j$) les unes sur les autres.

5.4 Étape 2 : combien de classes ?

Dans le cas du rejet de l'hypothèse $H_0^{(1)}$, la question du nombre de classes K se pose. Comme indiqué dans la Figure 5.1, deux possibilités sont envisageables : continuer avec une stratégie séquentielle de tests, ou utiliser un indice donnant directement le nombre de classes approprié.

5.4.1 Stratégie séquentielle de tests

Supposons que nous mettons en balance l'hypothèse $H_0^{(2)} : K = 2$ et l'hypothèse $H_1^{(2)} : K > 2$. Nous proposons d'effectuer la classification des tableaux en deux classes

5.4 Étape 2 : combien de classes ?

grâce à CLUSTATIS. Par la suite, à l'intérieur de chaque classe, nous pouvons mettre en œuvre le test basé sur la significativité de la deuxième valeur propre de la matrice des coefficients RV. En d'autres termes, il s'agit d'évaluer si la matrice des coefficients RV entre les tableaux d'une classe est de rang 1, ce qui revient à dire que la classe considérée est homogène. Pour éviter l'inflation de l'erreur de première espèce α , la correction de Bonferroni (Weisstein, 2004) est utilisée. Plus précisément, un seuil égal à $\alpha/2$ pour chacun des deux tests (associés aux deux classes) est considéré. Cette stratégie séquentielle peut-être réalisée avec les deux manières de permuter définies précédemment (*i.e.*, conduisant aux Test 1 et Test 2).

Dans le cas du rejet de l'hypothèse $H_0^{(2)}$, la procédure de test est réitérée en considérant les hypothèses $H_0^{(3)} : K = 3$ contre $H_1^{(3)} : K > 3$, et ainsi de suite.

5.4.2 Détermination du nombre de classes à l'aide d'indices

Comme nous l'avons indiqué ci-dessus, pour déterminer le nombre de classes approprié, il existe des stratégies basées sur l'évolution d'indices statistiques en fonction du nombre de classes. Pour une partition en K classes, ces indices sont basés sur la variation inter-classes et la variation intra-classes. Nous proposons d'adapter deux indices utilisés dans le cadre de la classification d'individus au cas de CLUSTATIS. Pour cela, nous nous basons sur la relation que nous avons démontrée dans le chapitre 3 :

$$\sum_{i=1}^m \|\mathbf{W}_i\|^2 = \sum_{i=1}^m \sum_{i \in G_k} \|\mathbf{W}_i - \alpha_i \mathbf{W}^{(k)}\|^2 + \sum_{k=1}^K \|\mathbf{W}^{(k)}\|^2.$$

Nous avons alors interprété cette relation en indiquant que la variation totale ($\sum_{i=1}^m \|\mathbf{W}_i\|^2$) est décomposée en une variation inter-classes ($H = \sum_{k=1}^K \|\mathbf{W}^{(k)}\|^2 = \sum_{k=1}^K \lambda_1^{(k)}$) et une variation intra-classes ($D = \sum_{i=1}^m \sum_{i \in G_k} \|\mathbf{W}_i - \alpha_i \mathbf{W}^{(k)}\|^2$).

Indice CH

Dans le cadre de la classification de n individus en K classes, sur la base d'un tableau de données quantitatives, l'indice de Calinski-Harabasz est donné par $\frac{(BGSS/K-1)}{(WGSS/n-K)}$, où K est le nombre de classes, $BGSS$ est la somme des carrés inter-classes et $WGSS$ est la somme des carrés intra-classes (Desgraupes, 2013).

L'adaptation de cet indice à la classification de m tableaux conduit à :

$$CH(K) = \frac{H/(K-1)}{D/(m-K)} \quad (5.1)$$

Pratiquement, cet indice est calculé pour les valeurs successives de K , et la valeur de K

retenue est celle qui correspond à la valeur maximale de l'indice (Charrad *et al.*, 2014).

Indice H

L'indice d'Hartigan (Hartigan, 1975) est basé sur le rapport entre la somme des carrés intra-classes calculée avec deux nombres consécutifs de classes. Il est utilisé pour le choix du nombre de classes dans le cadre de la classification de n individus; étant défini de la façon suivante : $(\frac{WGSS_K}{WGSS_{K+1}} - 1)(n - K - 1)$, où $WGSS_K$ (respectivement $WGSS_{K+1}$) est la somme des carrés intra-classes obtenue avec une partition en K (respectivement $K + 1$) classes.

Nous proposons d'adapter l'indice d'Hartigan pour le cas de tableaux multiples comme suit :

$$H(K) = (\frac{D_K}{D_{K+1}} - 1)(m - K - 1) \quad (5.2)$$

où D_K (respectivement, D_{K+1}) correspond à la variation intra-classes pour K (respectivement, $K + 1$) classes et m est le nombre de tableaux. La valeur de K qui correspond à la plus grande variation $H(K - 1) - H(K)$ est retenue comme étant le nombre de classes (Milligan et Cooper, 1985).

5.5 Adaptation aux données CATA

Nous rappelons que, dans le cadre de données CATA, les tableaux de données (sujets) sont apparentés à la fois par lignes (produits) et par colonnes (attributs). Pour déterminer le nombre de classes de tableaux, la procédure générale décrite dans ce chapitre reste la même. Cette section vise à expliciter les adaptations des diverses stratégies au cas des données CATA.

Afin de déterminer si une classe de tableaux (sujets) existe, le test basé sur la significativité de la seconde valeur propre de la matrice de similarité peut être utilisé. Cependant, au lieu de considérer la matrice RV, c'est la matrice S composée des coefficients d'Ochiai entre les tableaux qui est étudiée. De plus, étant donné que nous avons recommandé de travailler directement sur les tableaux de données eux-mêmes, la permutation sur les matrices de produits scalaires n'est pas adaptée. Ainsi, seul le Test 1 est défini, puisqu'il utilise une permutation sur les données brutes.

Dans le cas du rejet de l'hypothèse d'une seule classe, la stratégie séquentielle de tests

5.6 Application à des données simulées et réelles

décrite ci-dessus peut être utilisée. Pour cela, les classes définies par CLUSCATA sont considérées.

Étant donné la propriété $\sum_{i=1}^m \|\mathbf{X}_i\|^2 = m = \sum_{k=1}^K \sum_{i \in G_k} \|\mathbf{X}_i - \alpha_i \mathbf{C}^{(k)}\|^2 + \sum_{k=1}^K \|\mathbf{C}^{(k)}\|^2 = Q + Z$ (équation 4.4), l'adaptation des indices de Calinski-Harabasz et d'Hartigan peuvent être réalisées exactement de la même manière que pour CLUSTATIS. Il suffit d'utiliser le critère $Q = \sum_{k=1}^K \sum_{i \in G_k} \|\mathbf{X}_i - \alpha_i \mathbf{C}^{(k)}\|^2$ de CLUSCATA, qui représente la variation intra-classes, et $Z = \sum_{k=1}^K \|\mathbf{C}^{(k)}\|^2$, définie comme la variation inter-classes. Nous obtenons donc :

$$CH(K) = \frac{Z/(K-1)}{Q/(m-K)} \text{ et } H(K) = (\frac{Q_K}{Q_{K+1}} - 1)(m - K - 1).$$

5.6 Application à des données simulées et réelles

5.6.1 Simulations

Afin de comparer les différentes stratégies détaillées dans la section précédente, des simulations ont été réalisées. Ces simulations sont basées sur des configurations conformes aux trois épreuves sensorielles discutées dans ce manuscrit : des données de Projective mapping/Napping, de tri libre et CATA ont été simulées. Dans chaque cas, des données peu bruitées, moyennement bruitées et très bruitées ont été générées. De plus, nous avons fait varier le nombre de classes, le nombre de tableaux, la proximité entre les tableaux d'une même classe ou de différentes classes, et l'équilibre des effectifs entre les classes. Ces simulations permettent également d'évaluer l'efficacité des méthodes CLUSTATIS et CLUSCATA par le biais de l'indice de Rand entre la partition simulée et la partition trouvée par la méthode de classification.

Projective mapping/Napping

Plan de simulation

Les simulations ont été réalisées en prenant comme base, pour chaque classe, les coordonnées X-Y des produits données par un sujet provenant d'une épreuve réelle de Projective mapping/Napping sur huit smoothies (données présentées dans la section 2.4.1). Sur ces coordonnées, du bruit est ajouté en considérant une loi normale centrée avec un écart-type plus ou moins élevé. La proximité (évaluée par le coefficient RV) des sujets qui forment le noyau de chaque classe (tableaux de base) est un paramètre important pour contrôler la difficulté à trouver le nombre de classes. De plus, le coefficient RV moyen entre

les tableaux d'une classe (plus la variance de la distribution normale est élevée, plus le coefficient RV moyen intra-classes est faible) régule également la complexité à déterminer le nombre de classes correct. Enfin, le nombre total de tableaux (sujets) et l'équilibre des classes en termes d'effectifs complètent les différents éléments variants. Le Tableau 5.1 recense les paramètres utilisés afin de simuler différentes configurations de données. Toutes les combinaisons de paramètres sont testées. Dans chaque cas, 100 simulations ont été réalisées et le pourcentage de nombres corrects de classes indiqués a été évalué. De plus, la moyenne de l'indice de Rand entre la partition donnée par CLUSTATIS et la partition réelle simulée a été calculée.

	1 classe	2 classes	3 classes
Nombre de tableaux	30, 60, 100	30, 60, 100	60, 100 (99 si équilibrées)
RV moyen intra-classes	0.5, 0.7, 0.9	0.5, 0.7, 0.9	0.5, 0.7, 0.9
RV entre les tableaux de base		0.1, 0.3	Fixé : 0.1, 0.2 et 0.3
Taille de chaque classe si déséquilibrées		10/20, 15/45, 25/75	25/25/10, 30/15/15, 40/40/20, 50/25/25

Tableau 5.1 : Plan de simulation de données de Projective mapping/Napping. Chaque cas est également testé avec des données équilibrées.

Résultats

La première étape consiste à ne construire qu'une seule classe, puis à appliquer les deux tests proposés. Avec un coefficient RV moyen intra-classes égal à 0.9 et 0.7, les deux tests préconisent toujours une seule classe (100% de réussite), ce qui montre que, dans des conditions relativement tranchées, les deux tests conduisent à de bons résultats. Les pourcentages de réussite avec un coefficient RV moyen intra-classes égal à 0.5 sont donnés dans la première partie du Tableau 5.2. Le Test 2 est incontestablement le plus performant.

Lorsque deux classes sont générées, les résultats des tests et indices sont à nouveau toujours corrects (100%) lorsque le coefficient RV moyen intra-classes est égal à 0.7 ou 0.9, quels que soient les autres paramètres du Tableau 5.1. Ainsi, seuls les résultats avec le coefficient RV moyen intra-classes égal à 0.5 sont présentés dans la deuxième partie du Tableau 5.2. L'accent est mis sur le cas correspondant à un coefficient RV entre les tableaux qui servent de base de simulation égal à 0.3. Cependant, les résultats sont similaires pour une valeur de 0.1. Deux situations sont considérées : les effectifs des classes équilibrés et déséquilibrés. Le Test 2, l'Indice CH et l'Indice H sont plus performants que le Test 1 dans les deux situations.

5.6 Application à des données simulées et réelles

La dernière partie du Tableau 5.2 concerne le cas où il existe trois classes de tableaux. Deux types de données non équilibrées en termes d'effectifs sont utilisés dans ce contexte : le cas où il y a une grande classe et deux petites classes, et celui comprenant une petite classe et deux grandes classes (détails dans le Tableau 5.1). Ici encore, les résultats sont toujours corrects pour les tests et les indices avec un coefficient RV moyen intra-classes égal à 0.9. Ils sont également excellents avec un coefficient RV moyen intra-classes égal à 0.7, sauf pour l'Indice CH ($< 15\%$ dans tous les cas), ce qui nous amène à ne montrer que les résultats pour un coefficient RV moyen intra-classes égal à 0.5 dans le Tableau 5.2. Les deux tests donnent de bons résultats dans ce cas. Concernant les indices, une différence nette est à noter. L'Indice CH ne permet pas de retrouver le bon nombre de classes, alors que l'Indice H semble performant, excepté dans le cas de classes déséquilibrées comprenant une grande classe et deux petites classes.

Pour un coefficient RV moyen intra-classes égal à 0.7 ou 0.9, la moyenne de l'indice de Rand entre la partition réelle et la partition trouvée est égale à 1 pour tous les cas, ce qui confirme la pertinence de la méthode. Dans le cas de données fortement bruitées, ses performances restent très correctes (Rand moyen ≥ 0.91). Un autre aspect intéressant est que CLUSTATIS ne semble pas être affectée par les effectifs de classes déséquilibrés, puisque la moyenne des indices de Rand entre la partition réelle et la partition trouvée pour les données équilibrées est pratiquement égale à celle obtenue pour les données déséquilibrées.

Tri libre

Plan de simulation

Pour simuler des données de tri libre pour une classe de sujets, nous commençons, tout d'abord, par fixer un niveau de consistance, p , qui sera précisé ci-après. Plus ce niveau de consistance sera élevé, plus la classe de sujets sera homogène (voir plus bas). Ensuite, nous considérons les données d'un sujet pris comme configuration support. Supposons que le sujet ait effectué une partition des produits en trois groupes, que nous désignons par g_1 , g_2 et g_3 , respectivement. La simulation des données d'un sujet consiste à affecter des produits à des groupes, formant une partition de ces produits. Considérons le cas d'un produit qui, pour le sujet support, était affecté au groupe g_1 . Ce produit sera alors affecté au même groupe avec une probabilité p . Afin de ne pas limiter tous les sujets simulés à avoir une partition en trois classes comme le sujet support, le produit considéré sera affecté aux groupes g_2 , g_3 et à un nouveau groupe g_4 avec la probabilité $\frac{1-p}{3}$. De

1 classe					
	Nb tableaux	Test 1	Test 2	Indice CH	Indice H
	30	91	100	-	-
	60	76	100	-	-
	100	36	96	-	-
2 classes					
	Nb tableaux	Test 1	Test 2	Indice CH	Indice H
Équilibrées	30	97	100	100	100
	60	85	100	100	100
	100	82	100	100	100
Déséquilibrées	30	97	93	100	97
	60	93	98	100	98
	100	79	99	100	100
3 classes					
	Nb tableaux	Test 1	Test 2	Indice CH	Indice H
Équilibrées	60	97	99	0	94
	100	85	98	0	100
Déséquilibrées 2 grandes 1 petite	60	88	74	0	3
	100	92	97	0	31
Déséquilibrées 1 grande 2 petites	60	99	100	0	80
	95	90	99	0	93

Tableau 5.2 : Pourcentage de résultats corrects des tests et indices pour les données de Projective mapping/Napping simulées pour une, deux et trois classes. Coefficient RV moyen intra-classes égal à 0.5. Pour deux classes, coefficient RV entre les tableaux de base égal à 0.3.

plus, si au cours de la simulation, un produit se trouve effectivement affecté au groupe g_4 , alors nous donnons la possibilité d'avoir un cinquième groupe. De fait, la probabilité d'affectation d'un produit placé dans le groupe g_1 par le sujet support aux groupes g_2 , g_3 , g_4 ou g_5 devient $\frac{1-p}{4}$. Cependant, nous limitons la simulation à un maximum de deux groupes supplémentaires. A contrario, nous pouvons remarquer qu'avec cette procédure, il est possible de simuler des partitions avec moins de trois classes. Ceci correspond à la situation où aucun produit n'a été affecté à l'une ou l'autre des classes initiales.

Nous illustrons cette démarche avec un exemple concret. Supposons que le sujet support ait affecté 14 produits à six groupes, comme indiqué dans le Tableau 5.3. Nous considérons un niveau de consistance $p = 0.7$. Dans le Tableau 5.3, nous donnons l'exemple de dix sujets simulés selon la procédure décrite. Le nombre de groupes de ces sujets simulés varie entre 5 (Simulations 6, 7 et 9) et 8 (Simulation 5).

Naturellement, si le niveau de consistance, p , diminue, alors l'indice d'homogénéité diminue. Pour en avoir une idée plus précise, nous avons considéré la même étude de

5.6 Application à des données simulées et réelles

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14
Sujet support	1	2	3	4	1	1	4	5	6	5	5	4	6	1
Simulation 1	1	2	3	7	1	7	2	5	6	5	5	4	6	1
Simulation 2	4	2	6	1	1	4	7	5	6	5	5	4	6	1
Simulation 3	6	2	3	4	1	1	5	5	4	5	3	4	5	1
Simulation 4	4	2	3	2	1	1	3	5	6	6	5	4	6	1
Simulation 5	1	2	3	4	7	1	4	5	6	5	5	8	6	7
Simulation 6	1	5	3	4	1	6	6	5	4	1	5	3	6	1
Simulation 7	4	2	6	4	1	1	4	5	6	6	4	2	6	1
Simulation 8	5	2	3	4	1	2	4	5	6	5	7	4	1	3
Simulation 9	1	2	3	4	3	1	4	5	1	5	5	4	4	1
Simulation 10	1	2	3	4	1	1	4	5	7	5	5	4	1	1

Tableau 5.3 : Affectations des produits P1, ..., P14 aux groupes par le sujet support et par 10 sujets simulés avec un niveau de consistance $p = 0.7$.

simulation reflétée par le Tableau 5.3 en considérant 50 sujets et nous avons fait varier le niveau de consistance entre 0.05 et 0.95. La Figure 5.2 montre l'évolution de l'indice d'homogénéité en fonction du niveau de consistance p . Nous pouvons remarquer que l'indice d'homogénéité reste faible tant que le niveau de consistance ne dépasse pas la valeur 0.55. Ce constat est logique car un niveau de consistance inférieur à 0.5 signifie que, pour le sujet simulé, un produit donné a une probabilité plus grande d'être affecté dans un autre groupe que celui du sujet support. À partir de la valeur $p = 0.55$, la courbe forme un coude en passant d'une évolution sous forme d'un plateau à une évolution indiquant une forte croissance. Naturellement, plus p se rapproche de 1, plus l'indice d'homogénéité se rapproche également de 1. Nous pouvons noter que cette procédure originale de simulation s'inspire d'un modèle basé sur le concept de classes latentes pour les données de tri libre (Wiley, 1967).

À présent, nous pouvons définir le plan de simulation des données de tri libre. Le Tableau 5.4 recense les différents cas simulés, chaque paramètre étant utilisé avec tous les autres. Chaque simulation est réalisée 100 fois.

Résultats

Le Tableau 5.5 indique les pourcentages de décisions correctes de chacun des tests d'hypothèses lorsqu'une seule classe est simulée. Si le Test 1 procure de bons résultats, le Test 2 est encore plus performant. Concernant ces tests, les conclusions sont exactement les mêmes pour des données simulées avec deux ou trois classes. Au vu de la redondance des résultats avec ceux de Projective mapping/Napping, nous ne les détaillerons pas.

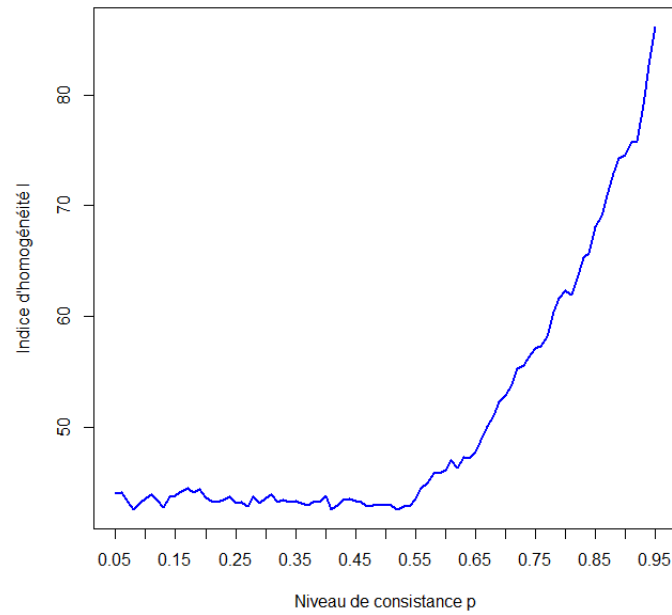


FIGURE 5.2 : Évolution de l'indice d'homogénéité I en fonction du niveau de consistance p .

	1 classe	2 classes	3 classes
Nombre de tableaux	30, 60, 100	30, 60, 100	60, 100 (99 si équilibrées)
RV moyen intra-classes	0.6, 0.7, 0.8	0.6, 0.7, 0.8	0.6, 0.7, 0.8
RV entre les tableaux de base		0.35, 0.51	Fixé : 0.35, 0.44 et 0.51
Taille de chaque classe si déséquilibrées		10/20, 15/45, 25/75	25/25/10, 30/15/15, 40/40/20, 50/25/25

Tableau 5.4 : Plan de simulation de données de tri libre. Chaque cas est également testé avec des données équilibrées.

Cependant, une remarque importante est que l'indice de Rand moyen entre les partitions simulées et les partitions retrouvées par CLUSTATIS est au minimum égal à 0.97, ce qui illustre une fois de plus l'efficacité de l'application de CLUSTATIS sur des données de tri libre.

CATA

Plan de simulation

Afin de correspondre au mieux à la spécificité des données CATA, entre 60 et 140 tableaux binaires sont considérés à chaque fois. Comme pour les épreuves précédentes, un

5.6 Application à des données simulées et réelles

RV moyen intra-classes = 0.8		
Nb tableaux	Test 1	Test 2
30	100	100
60	100	100
100	100	100
RV moyen intra-classes = 0.7		
Nb tableaux	Test 1	Test 2
30	100	100
60	100	100
100	98	100
RV moyen intra-classes = 0.6		
Nb tableaux	Test 1	Test 2
30	100	100
60	93	100
100	56	100

Tableau 5.5 : Données de tri libre : Pourcentage de nombres corrects de classes lorsqu'une seule classe est simulée.

sujet différent pour chaque classe est considéré comme base de simulation. Ces sujets proviennent d'une épreuve CATA sur des fraises (données présentées dans la section 4.6.1). Les tableaux binaires ainsi obtenus sont « bruités » en tirant, pour chaque valeur, un 0 ou un 1 avec une probabilité donnée de tirer la même valeur que celle du tableau de base. Plus cette probabilité est élevée, plus le tableau simulé devrait ressembler au tableau initial. Le concept est en fait très proche de celui que nous avons détaillé pour les données de tri libre, à l'exception que seules deux valeurs sont possibles : 0 et 1. Les paramètres du plan de simulation sont détaillés dans le Tableau 5.6. Nous rappelons que, puisque des données CATA sont considérées, le Test 2 n'est pas utilisable. Toutes les simulations sont réalisées 100 fois.

	1 classe	2 classes	3 classes
Nombre de tableaux	60, 100, 140	60, 100, 140	60, 100 (99 si équilibrées), 140 (138 si équilibrées)
s moyen intra-classes	0.5, 0.6, 0.7	0.5, 0.6, 0.7	0.5, 0.6, 0.7
s entre les tableaux de base		Fixé : 0.47	Fixé : 0.47, 0.44 et 0.43
Taille de chaque classe si déséquilibrées		15/45, 25/75, 100/40	25/25/10, 30/15/15, 40/40/20, 50/25/25, 60/60/20, 75/35/35

Tableau 5.6 : Plan de simulation de données CATA. Chaque cas est également testé avec des données équilibrées.

Résultats

Le nombre de classes préconisé par les tests et indices est toujours correct (100%) lorsque le coefficient s moyen intra-classes est égal à 0.7, excepté pour l'Indice CH, qui n'est une nouvelle fois pas très performant lorsque nous considérons une structure en trois classes de tableaux. Ainsi, le Tableau 5.7 n'indique que les résultats des cas considérant un coefficient s moyen intra-classes égal à 0.6 ou 0.5. Il apparaît que, dans le cas de trois classes, d'effectifs déséquilibrés et de données très bruitées, l'Indice H présente une mauvaise performance. Le Test 1 se démarque des indices en présentant tout le temps un pourcentage de réussite égal à 100, excepté une fois (94).

Ces simulations permettent également d'évaluer les performances de CLUSCATA, grâce à l'indice de Rand moyen entre la partition simulée et la partition trouvée par CLUSCATA. Étant donné que cet indice n'est jamais plus faible que 0.96, la pertinence de la méthode CLUSCATA est une nouvelle fois vérifiée. De plus, CLUSCATA ne montre aucune baisse de performances dans le cadre d'effectifs de classes déséquilibrés.

Conclusion générale de l'étude basée sur des simulations

Dans le cadre de données de Projective mapping/Napping et de tri libre, le Test 2 s'est démarqué par ses bonnes performances, à la fois pour déterminer si une classification est nécessaire, et pour évaluer le nombre de classes, le cas échéant. L'Indice CH a été moins performant pour retrouver les structures en trois classes. En revanche, l'Indice H a procuré de bons résultats. Il s'est cependant montré moins performant lors d'une structure en trois classes comprenant une classe de faible effectif et deux classes de grand effectif. L'étude des données CATA a montré les excellentes performances du Test 1, et a donné les mêmes conclusions pour les indices. Il est néanmoins à noter que si un des tests est le plus performant à chaque fois, ces tests nécessitent des permutations, et par conséquent un temps de calcul pouvant s'avérer élevé. En revanche, l'Indice H est immédiat à déterminer, et peut ainsi être une alternative. Nous proposons donc de retenir le Test 2 et l'Indice H dans le cadre de données multi-tableaux quantitatives, et le Test 1 et l'Indice H pour le cas des épreuves CATA.

5.6.2 Données réelles

Afin d'avoir un aperçu des indications des différentes stratégies de sélection du nombre de classes qui ont été retenues suite aux études de simulation, nous les avons mis en œuvre

5.6 Application à des données simulées et réelles

1 classe							
		<i>s moyen intra-classes = 0.6</i>			<i>s moyen intra-classes = 0.5</i>		
	Nb tableaux	Test 1	Indice CH	Indice H	Test 1	Indice CH	Indice H
	60	100	-	-	100	-	-
	100	100	-	-	100	-	-
	140	100	-	-	100	-	-
2 classes							
		<i>s moyen intra-classes = 0.6</i>			<i>s moyen intra-classes = 0.5</i>		
	Nb tableaux	Test 1	Indice CH	Indice H	Test 1	Indice CH	Indice H
Équilibrées	60	100	100	100	100	100	100
	100	100	100	100	100	100	100
	140	100	100	100	100	100	100
Déséquilibrées	60	100	100	100	100	100	100
	100	100	100	100	100	100	100
	140	100	100	100	100	100	100
3 classes							
		<i>s moyen intra-classes = 0.6</i>			<i>s moyen intra-classes = 0.5</i>		
	Nb tableaux	Test 1	Indice CH	Indice H	Test 1	Indice CH	Indice H
Équilibrées	60	100	0	100	100	0	93
	100	100	0	100	100	0	99
	140	100	0	100	100	0	99
Déséquilibrées 2 grandes 1 petite	60	100	0	88	94	0	12
	100	100	0	100	100	0	55
	140	100	0	51	100	0	2
Déséquilibrées 1 grande 2 petites	60	100	0	100	100	0	25
	100	100	0	100	100	0	40
	140	100	0	100	100	0	35

Tableau 5.7 : Données CATA : Pourcentages de nombres corrects de classes du Test 1 et des indices dans les différentes conditions définies par le Tableau 5.6, excepté le cas où le coefficient *s* moyen intra-classes est égal à 0.7.

sur des jeux de données réels. Nous avons exécuté le Test 2 et l'Indice H sur cinq jeux de données provenant d'épreuves de Projective mapping/Napping (Tableau 5.8) et sur trois jeux de données de tri libre (Tableau 5.9). Par la suite, neuf épreuves CATA ont été soumises au Test 1 et à l'Indice H (Tableau 5.10). Dans chaque tableau sont indiqués les caractéristiques des jeux de données, l'homogénéité des sujets sans classification et le nombre de classes préconisé par chaque stratégie. Ces données comportent des nombres différents de lignes (produits) et de tableaux (sujets), voire d'attributs dans le cas des données CATA. De plus, l'homogénéité des sujets sans classification est variable d'un jeu de données à l'autre.

Dans le cadre de données de Projective mapping/Napping, les données utilisées proviennent d'Agrocampus, excepté le second jeu de données qui provient de Norvège (Berget et al., 2019). Le Tableau 5.8 indique que, sans regroupement, les jeux de données 1 et 4 sont les plus homogènes. Il n'est donc pas surprenant de constater que, pour ces données, le Test 2 préconise de ne retenir qu'une classe de tableaux. Nous rappelons que l'Indice H

#	Nb produits	Nb sujets	Homogénéité	Test 2	Indice H
1	8	24	42.4%	1	3
2	12	100	24.3%	11	2
3	12	97	26.0%	6	2
4	8	17	59.1%	1	4
5	8	17	28.9%	2	2

Tableau 5.8 : Nombres de classes préconisés par le Test 2 et l'Indice H sur cinq jeux de données réels provenant d'épreuves de Projective mapping/Napping.

#	Nb produits	Nb sujets	Homogénéité	Test 2	Indice H
1	14	25	63.9%	1	2
2	16	31	43.9%	1	5
3	12	30	40.3%	1	2

Tableau 5.9 : Résultats du Test 2 et de l'Indice H sur trois jeux de données réels provenant d'épreuves de tri libre.

ne peut pas indiquer une seule classe, et proposera donc toujours une classification. Les cas des jeux de données 2 et 3 sont particuliers. En effet, dans chaque cas, l'homogénéité très faible et le nombre de produits élevé montrent que ces épreuves se sont avérées très difficiles pour les sujets. Ainsi, beaucoup de classes sont envisagées par le Test 2. L'Indice H n'indique pas de retenir un nombre relativement grand de classes. Cette étude montre que les tests ont tendance à proposer plus de classes que l'Indice H.

Pour le cas du tri libre, les deux jeux de données utilisés dans le chapitre 2 et chapitre 3 sont de nouveau considérés. Le dernier provient d'Agrocampus. Toutes ces épreuves ont débouché sur une homogénéité de plus de 40%, et le Test 2 a proposé de ne considérer qu'une seule classe dans chaque cas.

Les neuf jeux de données CATA considérés sont les mêmes que ceux utilisés dans le chapitre 4. Le Test 1 indique parfois un nombre de classes relativement important, au contraire de l'Indice H.

5.7 Conclusion

Les tests et indices proposés peuvent être très utiles pour aider l'utilisateur dans son choix du nombre de classes. La procédure de simulation a permis de comparer ces différentes approches. Le Test 2 et l'Indice H se sont révélés respectivement le test et l'indice les plus efficaces dans le cadre de données de Projective mapping/Napping et de tri libre, avec une meilleure performance du Test 2 que de l'Indice H. Le Test 1 est le

5.7 Conclusion

#	Nb produits	Nb attributs	Nb sujets	Homogénéité	Test 1	Indice H
1	6	27	154	21.3%	5	3
2	8	38	73	38.2%	2	3
3	9	15	97	45.5%	4	2
4	4	10	87	70.7%	1	3
5	12	17	112	48.5%	1	2
6	6	17	83	31.8%	6	2
7	6	14	134	31.7%	6	2
8	9	15	76	46.8%	8	2
9	6	16	114	37.3%	1	2

Tableau 5.10 : Résultats du Test 1 et de l'Indice H sur neuf jeux de données réels provenant d'épreuves CATA.

plus performant sur des données issues d'épreuves CATA. Les jeux de données réels ont indiqué que les tests ont tendance à préconiser un nombre de classe plus élevé que l'Indice H. Une remarque importante est que l'Indice H a l'avantage d'être immédiat en termes de temps de calcul, contrairement aux tests qui nécessitent des permutations. Suite à ce constat, il peut être choisi comme alternative à la procédure séquentielle de tests. Ainsi, il est possible de déterminer si une classification est nécessaire via le test sélectionné, puis d'utiliser l'Indice H afin de préconiser le nombre de classes.

Les méthodes CLUSTATIS et CLUSCATA contiennent une option « $K + 1$ » permettant de détecter les tableaux atypiques. Nous avons remarqué dans le chapitre 4 que, puisqu'un tableau atypique est un tableau qui n'appartient à aucun modèle de classe, il est clair que pour deux nombres différents de classes, un tableau peut être considéré comme atypique ou non. Le fait de travailler simultanément sur les deux aspects, en considérant les classes issues de la classification avec l'option « $K + 1$ », pourrait être examiné.

À travers ces simulations, la pertinence des méthodes CLUSTATIS et CLUSCATA a été une nouvelle fois confirmée, même dans des situations de classes avec des effectifs déséquilibrés.

Il est utile de remarquer que les études de simulations concernant les données de type tri libre et CATA présentent un certain nombre de difficultés que nous avons cherché à contourner. Une perspective intéressante serait d'approfondir cet aspect en intégrant, en particulier, d'autres modèles de simulation.

Chapitre 6

Implémentation informatique

6.1 Introduction

R (R Core Team, 2020) est un environnement pour la programmation principalement dédié à la statistique et aux représentations graphiques. C'est un logiciel libre distribué selon les termes de la licence dite GPL. Afin de rendre les outils que nous avons développés disponibles et à la portée de tous les utilisateurs, nous avons conçu un package *R* nommé ClustBlock, dédié à l'analyse exploratoire et à la classification de tableaux multiples.

Dans le cadre de l'analyse de tableaux multiples, d'autres packages sont disponibles. Pour l'analyse exploratoire, le package *ade4* (Thioulouse *et al.*, 2007) permet d'exécuter la méthode STATIS. Le package *FactoMineR* (Lê *et al.*, 2008) rends les méthodes Analyse Factorielle Multiple (AFM) et Analyse Procrustéenne Généralisée (APG) disponibles. Pour analyser des données de tri libre, le package *FreeSortR* (Courcoux, 2017) se base sur des indices de Rand et de Rand ajusté entre les partitions des sujets. À notre connaissance, ClustBlock est le seul package permettant d'analyser directement des données CATA. De plus, d'après nos recherches, aucun autre package *R* n'existe pour la classification de tableaux de données.

Il est à noter que ClustBlock est dépendant de certaines fonctions du package *FactoMineR* (Lê *et al.*, 2008). De plus, il suggère d'installer *ClustVarLV* (Vigneau *et al.*, 2020), qui est un package contenant la méthode CLV (Vigneau et Qannari, 2003) qui permet de segmenter les variables avec des algorithmes similaires à CLUSTATIS.

XLSTAT (Addinsoft, 2020) est un logiciel de statistique très simple d'utilisation. En effet, il suffit de sélectionner des données dans Excel, de choisir une méthode statistique

appropriée et de « lancer » l’analyse après avoir renseigné d’éventuels paramètres dans une boîte de dialogue. S’affichent alors les résultats de l’analyse dans une nouvelle feuille Excel. Ce logiciel est très pratique car il ne demande pas de savoir coder, et tient compte du fait que les données des utilisateurs sont généralement stockées dans Excel. De plus, il est possible de charger les données via d’autres types de fichiers. Des méthodes introduites dans ce manuscrit ont été implémentées dans XLSTAT.

Dans ce chapitre, les fonctionnalités pour utiliser les différentes méthodes introduites précédemment sont présentées via le package *R* ClustBlock. Des exemples d’application sont donnés. D’autre part, une présentation du logiciel XLSTAT est réalisée.

6.2 Package *R* : ClustBlock

6.2.1 Présentation du package

Le package ClustBlock est composé majoritairement de fonctions d’analyse et de classification de tableaux de données. Les deux fonctions de base sont *statis* (méthode STATIS) et *clustatis* (méthode CLUSTATIS), dédiées aux données quantitatives de tableaux multiples. De plus, des fonctions spécifiques pour l’analyse sensorielle comme *catatis*, *cluscata* ou encore le pré-traitement des données de tri libre pour pouvoir exécuter les stratégies STATIS et CLUSTATIS sont également incluses. Naturellement, de nombreuses fonctions utiles (*plot*, *summary*, *etc.*) font partie du package, tout comme des jeux de données pouvant servir de test. La Figure 6.1 résume l’architecture de ce package.

6.2.2 Fonctions

La liste suivante recense les fonctions de ClustBlock, triées par catégories :

Analyse exploratoire

- *statis* : méthode STATIS.
- *catatis* : méthode CATATIS.
- *statis_FreeSort* : méthode STATIS sur des données de tri libre.
- *consistency_cata* : test de permutations afin d’identifier les attributs pour lesquels les sujets ne sont pas consistants dans une épreuve CATA.
- *consistency_cata_panel* : test de permutations afin de déterminer si le panel est consistant dans une épreuve CATA.

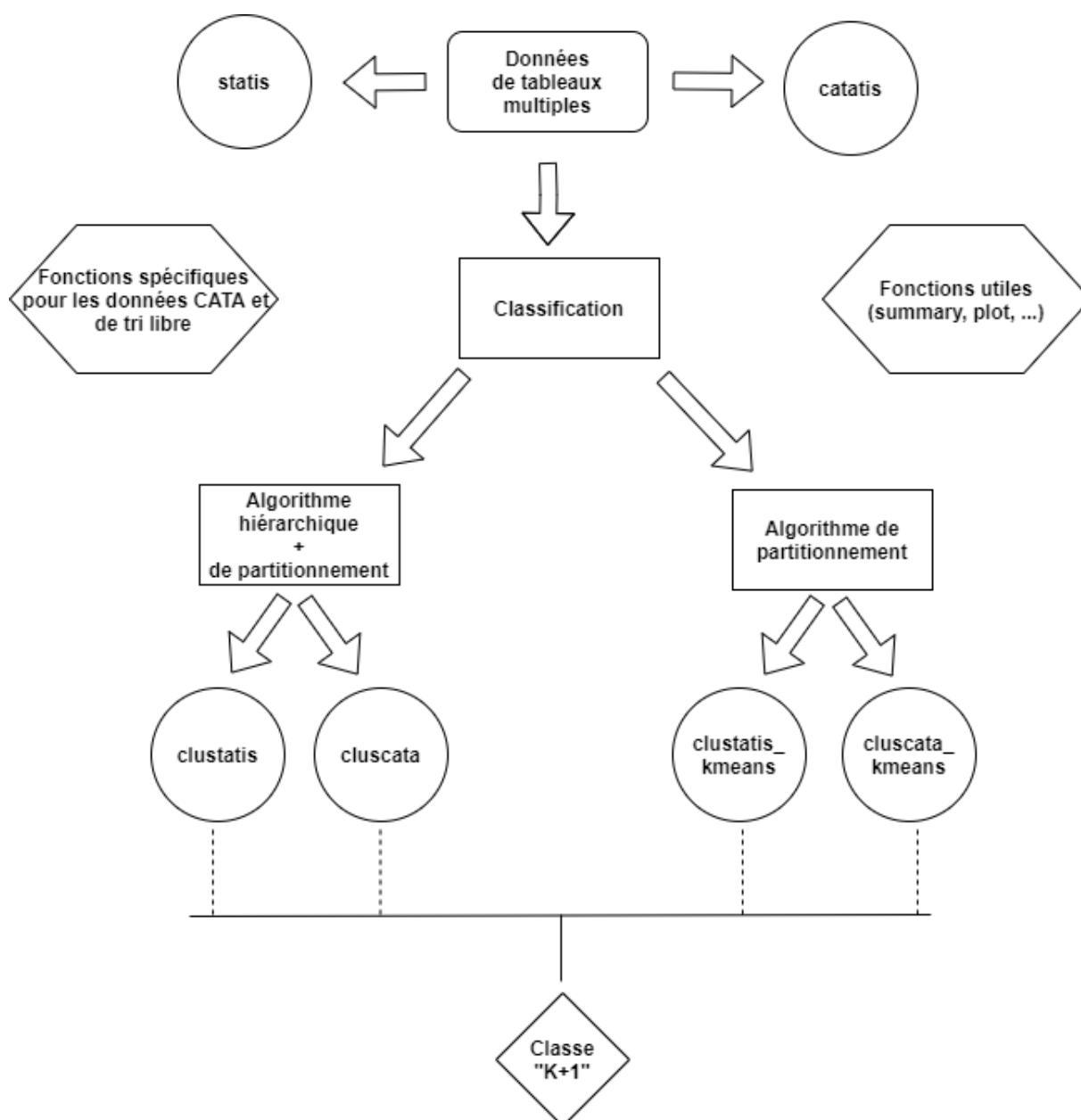


FIGURE 6.1 : Architecture des fonctionnalités du package ClustBlock.

Classification

- *clustatis* : méthode CLUSTATIS : algorithme hiérarchique consolidé par l'algorithme de partitionnement, option classe « $K + 1$ » disponible, aide au choix du nombre de classes.
- *clustatis_kmeans* : algorithme de partitionnement de CLUSTATIS (possibilité de multi-start ou de donner une partition initiale), option classe « $K + 1$ » disponible.
- *cluscata* : méthode CLUSCATA : algorithme hiérarchique consolidé par l'algorithme de partitionnement, option classe « $K + 1$ » disponible, aide au choix du nombre de

classes.

- *cluscata_kmeans* : algorithme de partitionnement de CLUSCATA (possibilité de multi-start ou de donner une partition initiale), option classe « $K + 1$ » disponible.
- *clustatis_FreeSort* : méthode CLUSTATIS sur des données de tri libre : algorithme hiérarchique consolidé par l'algorithme de partitionnement, option classe « $K + 1$ » disponible, aide au choix du nombre de classes.
- *clustatis_FreeSort_kmeans* : algorithme de partitionnement de CLUSTATIS sur des données de tri libre (possibilité de multi-start ou de donner une partition initiale), option classe « $K + 1$ » disponible.

Fonctions utiles

- *summary* : fonction pour afficher les principaux résultats à la suite de toutes les fonctions d'analyse et de classification.
- *plot* : fonction pour afficher les principaux graphiques à la suite de toutes les fonctions d'analyse et de classification.
- *print* : fonction pour afficher la méthode qui a été utilisée.
- *change_cata_format* : fonction permettant de changer de format de données CATA si le format initial n'est pas celui attendu par *catatis* ou *cluscata*. Utilisable avec les deux autres formats de données usuels, qui correspondent aux tableaux CATA placés les uns en dessous des autres, organisés par sujet ou par produit.
- *preprocess_FreeSort* : fonction qui permet de réaliser le pré-traitement des données de tri libre afin d'utiliser STATIS et CLUSTATIS. Notamment utilisée au sein de *statis_FreeSort*, *clustatis_FreeSort* et *clustatis_FreeSort_kmeans*.

6.2.3 Jeux de données

Afin de fournir des exemples aux utilisateurs, des jeux de données sont disponibles dans le package :

- *smoo* : Données de Projective mapping/Napping sur des smoothies, qui ont été présentées dans la section [2.4.1](#).
- *straw* : Jeu de données provenant d'une épreuve CATA sur des fraises, décrit dans la section [4.6.1](#).
- *choc* : Données de tri libre sur des chocolats, présentées dans la section [2.4.2](#).

6.2.4 Exemples d'utilisation

Des exemples d'utilisation des méthodes STATIS et CLUSTATIS sur les données de Projective mapping/Napping incluses dans le package sont présentés dans cette section.

Commandes de base

Pour installer le package et le lancer, il faut utiliser les commandes suivantes :

```
> install.packages("ClustBlock")
> library(ClustBlock)
```

Pour chaque fonction, il existe une aide avec des exemples d'utilisation. Pour la visualiser, il suffit de rentrer « ? » suivi du nom de la fonction. Par exemple, pour avoir l'aide de *clustatis*, il faut rentrer l'instruction :

```
> ?clustatis
```

Les données concernant le Projective mapping/Napping sur les smoothies peuvent être téléchargées et visualisées de la manière suivante :

```
> data(smoo)
> View(smoo)
```

STATIS

Pour appliquer la fonction *statis*, il faut rentrer le code suivant :

```
> st=statis(Data=smoo, Blocks=rep(2,24))
```

Les données sont constituées de 48 colonnes correspondant aux tableaux de données contigus des 24 sujets, chacun des 24 sujets ayant 2 colonnes correspondant aux coordonnées X et Y récupérées sur les feuilles de papier. C'est pourquoi dans les fonctions *statis* et *clustatis*, il faut renseigner en paramètre `Blocks=rep(2,24)`.

Suite à l'utilisation de la fonction *statis*, le graphique représentant les individus (les smoothies ici) sur les deux premiers axes et le diagramme en bâtons des poids des tableaux (les sujets ici) donnés par la méthode STATIS apparaissent. Cependant, le graphique des individus peut être contrôlé via la fonction *plot*, en changeant par exemple la paire d'axes que nous désirons visualiser, ou la couleur des individus. Par exemple, si nous voulons représenter les individus sur les axes 1 et 3, en bleu foncé et un peu plus gros que le réglage par défaut, nous pouvons rentrer :

```
> plot(st, col=rep("darkblue", 8), cex=1.3, axes=c(1,3))
```

Le résultat de cette commande est donné sur la Figure 6.2.

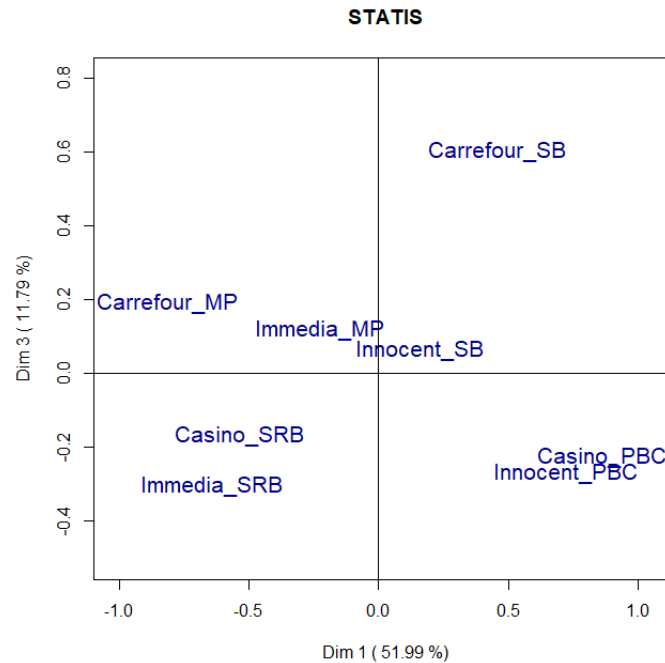


FIGURE 6.2 : Représentation du tableau compromis obtenu par la commande `plot(st, col=rep("darkblue", 8), cex=1.3, axes=c(1,3))`.

Il est possible et même vivement conseillé d’afficher les résultats les plus importants de l’analyse STATIS (homogénéité, valeurs numériques des poids, valeurs propres avec les pourcentages d’inertie des axes). Pour cela il faut utiliser la fonction `summary` :

```
> summary(st)

$homogeneity
[1] 42.4

$weights
B-1    B-2    B-3    B-4    B-5    B-6    B-7    B-8    B-9    B-10   B-11
0.18016 0.09257 0.21710 0.08030 0.21725 0.19287 0.12514 0.14862 0.21206 0.25926 0.22141
B-12   B-13   B-14   B-15   B-16   B-17   B-18   B-19   B-20   B-21   B-22
0.10826 0.20029 0.14059 0.24648 0.18754 0.14848 0.28062 0.24486 0.24846 0.21113 0.24686
B-23   B-24
0.27941 0.22010

$eigenvalues
Dim 1 Dim 2 Dim 3 Dim 4 Dim 5 Dim 6 Dim 7
2.921 0.924 0.662 0.429 0.333 0.237 0.113
```

```
$inertia
Dim 1 Dim 2 Dim 3 Dim 4 Dim 5 Dim 6 Dim 7
51.99 16.44 11.79 7.63 5.93 4.22 2.01
```

CLUSTATIS

La fonction *clustatis* contient les résultats de la classification hiérarchique consolidée par l'algorithme de partitionnement de CLUSTATIS pour chaque nombre de classes envisagé (de 1 jusqu'à un nombre de classes maximum contrôlé par le paramètre « *gpmax* »). Cet algorithme peut être utilisé soit sans classe « $K + 1$ », soit avec une classe « $K + 1$ ». Cette dernière option est gérée par le paramètre « *Noise_cluster* ». Ainsi, pour exécuter la méthode CLUSTATIS avec l'option « $K + 1$ », il faut rentrer l'instruction :

```
> cl=clustatis(Data= smoo, Blocks= rep(2,24), Noise_cluster = TRUE)
```

Afin d'aider l'utilisateur à faire son choix du nombre de classes, le dendrogramme et un diagramme en bâtons représentant l'évolution du critère d'agrégation (équation 3.1) s'affichent alors. De plus, un nombre recommandé sur la base de l'Indice H (voir section 5.4.2) est indiqué :

```
Recommended number of clusters = 3
```

Il est à noter que la décision du test pour savoir s'il y a une classe ou plus (Test 2, voir section 5.3) est disponible dans les résultats (via la fonction *summary*, voir ci-dessous).

Une fois le choix du nombre de classes effectué, il est conseillé d'appliquer les fonctions *plot* et *summary* afin de visualiser les principaux résultats de l'analyse. Pour cela, il faut renseigner le paramètre « *ngroups* » (s'il n'est pas renseigné, le nombre de classes recommandé sera choisi). Comme pour *statis*, la fonction *plot* possède diverses options pour la représentation graphique des individus dans chaque classe (pour l'exemple des smoothies, les résultats sont donnés sur la Figure 3.5). La fonction *summary* indique les compositions des classes, les indices d'homogénéité, le poids de chaque tableau au sein de sa classe, le seuil ρ qui a été calculé puis utilisé pour construire la classe « $K + 1$ », et, enfin, les résultats du test déterminant si une seule classe peut être considérée. Les noms des tableaux sont par défaut « B-1, B-2, ... » mais peuvent être changés par le paramètre « *NameBlocks* » dans *statis* et *clustatis*.

```

> summary(cl, ngroups=3)
$groups
$groups$`Cluster 1`
[1] "B-1" "B-3" "B-10" "B-15" "B-21" "B-22"

$groups$`Cluster 2`
[1] "B-4" "B-12" "B-14" "B-17"

$groups$`Cluster 3`
[1] "B-5" "B-6" "B-9" "B-11" "B-13" "B-16" "B-18" "B-19" "B-20" "B-23" "B-24"

$groups$`Noise cluster (K+1)`
[1] "B-2" "B-7" "B-8"

$homogeneity

```

	homogeneity (%)	nb_blocks
Cluster 1	73.2	6
Cluster 2	61.5	4
Cluster 3	60.9	11
Noise cluster	45.3	3
Overall	64.5	21
One group	42.4	24

```

$weights
$weights$`Cluster 1`
B-1      B-3      B-10     B-15     B-21     B-22
0.3241664 0.3964542 0.4154155 0.4395361 0.4228519 0.4395162

$weights$`Cluster 2`
B-4      B-12     B-14     B-17
0.4775003 0.4750231 0.5159443 0.5292901

$weights$`Cluster 3`
B-5      B-6      B-9      B-11     B-13     B-16     B-18
0.2945563 0.2778527 0.3115948 0.2720356 0.2644849 0.2458681 0.3448218
B-19     B-20     B-23     B-24
0.2968135 0.3283228 0.3506106 0.3111469

$rho
[1] 0.5966896

$test_one_cluster
$test_one_cluster$decision
[1] "Only one cluster can be considered"

$test_one_cluster$pvalue
[1] 1

```

Il est à noter que l'algorithme de partitionnement de CLUSTATIS avec plusieurs partitions aléatoires (stratégie multi-start) est exécutable grâce à la fonction *clustatis_kmeans*, en indiquant le nombre de classes voulu dans le paramètre « clust ». Il est également

6.3 XLSTAT

possible de donner une partition de départ dans le paramètre « clust » et de régler le seuil ρ grâce au paramètre « rho ».

```
> cl_km=clustatis_kmeans(Data=smoo,Blocks=rep(2,24), clust=3, nstart=100)
> partition
[1] 2 1 1 2 3 3 2 3 1 2 2 3 1 3 1 1 3 3 1 1 2 1 2 1
> cl_km2=clustatis_kmeans(Data=smoo,Blocks=rep(2,24), clust=partition, rho=0.5)
```

6.3 XLSTAT

Le travail de thèse ayant été réalisé dans le cadre d'une bourse CIFRE en collaboration avec l'entreprise Addinsoft, détentrice du logiciel XLSTAT (Addinsoft, 2020), des développements ont été réalisés dans ce logiciel. Ces derniers sont directement en rapport avec les travaux présentés dans ce manuscrit.

XLSTAT est un logiciel de statistique à la fois simple d'utilisation et très puissant. Il permet d'analyser, visualiser et modéliser les données tout en produisant des rapports sous Microsoft Excel, exportables vers d'autres formats. Il comporte plus de 220 fonctionnalités statistiques allant de la statistique descriptive au « machine learning ». Actuellement, plus de 100 000 utilisateurs sont recensés, situés dans plus de 120 pays. XLSTAT est compatible avec toutes les versions d'Excel depuis 2003 jusqu'à 2016.

L'offre proposée par XLSTAT pour les utilisateurs en analyse sensorielle est très variée. Elle englobe, par exemple, les cartographies des préférences, l'analyse de panel, l'Analyse Procrustéenne Généralisée (APG), l'analyse de données provenant d'une épreuve de Dominance Temporelle des Sensations, les tests de discrimination sensorielle, ...

Suite aux travaux qui ont été accomplis dans le cadre de cette thèse, nous avons enrichi le logiciel XLSTAT en implémentant les méthodes STATIS, CATATIS, CLUSTATIS, CLUSCATA et d'analyse de données de tri libre.

6.4 Conclusion

Les développements réalisés dans le cadre de deux logiciels permettent aux utilisateurs d'appliquer les méthodes introduites dans ce manuscrit. À ce stade, le package *R* ClustBlock est plus complet, puisque l'on y trouve quasiment l'ensemble des méthodes

présentées. Cependant, utiliser *R* nécessite une certaine familiarité avec l’environnement de programmation, ce qui n’est pas le cas de tous les utilisateurs. En revanche, utiliser XLSTAT est à la portée de tout le monde, y compris les utilisateurs novices.

D’un point de vue conceptuel, la difficulté de réalisation des programmes sous XLSTAT est sans commune mesure avec celle dans l’environnement *R*. En effet, l’utilisation de différents langages de programmation sont nécessaires, la conception de la boîte de dialogue doit être très précise, plusieurs langues doivent être disponibles, un tutoriel est créé, des tests très poussés sont réalisés, *etc.*

Voici quelques « chiffres » afin de mesurer l’impact de ces développements : ClustBlock est sorti le 6 mars 2019, et exactement un an après sa sortie, le package était téléchargé 5233 fois. Quatre mises à jour ont été effectuées. Neuf mois après sa sortie, les clients du logiciel XLSTAT avaient utilisé la méthode CATATIS environ 5000 fois.

Chapitre 7

Discussion et conclusion

Les données de tableaux multiples sont de plus en plus fréquentes dans divers domaines d'application. Si les méthodes d'analyse exploratoire ont fait couler beaucoup d'encre, le problème de classification de tableaux n'a été que très peu exploré, alors qu'un réel besoin existe. Ce besoin est notamment très présent dans le cadre de l'analyse sensorielle. En effet, des épreuves dites rapides ne nécessitant pas d'entraînement des sujets sont actuellement en vogue, et, très souvent, conduisent à l'obtention de tableaux multiples à raison d'un tableau par sujet. Considérant les différences de perception au sein d'un panel, il est souvent recommandé d'effectuer une segmentation des sujets.

Trois épreuves sensorielles ont particulièrement retenu notre attention : les données de Projective mapping/Napping, de tri libre et Check-All-That-Apply (CATA). Nous avons montré que la méthode STATIS pouvait être utilisée avec efficacité directement sur les données de Projective mapping/Napping, et, suite à un pré-traitement adapté, sur les données issues d'une épreuve de tri libre. Le cas spécifique des données CATA a également été exploré, puisqu'un raffinement de la méthode usuelle d'analyse a été proposé. Cette méthode, que nous avons nommé CATATIS, consiste en une adaptation de la méthode STATIS aux données CATA.

La méthode de classification de tableaux CLUSTATIS a été proposée. Cette approche, basée sur un critère d'optimisation, consiste à exécuter un algorithme hiérarchique ou de partitionnement. Si chacun de ces algorithmes peut être utilisé indépendamment, une complémentarité entre ces deux stratégies existe. En effet, la solution donnée par l'algorithme hiérarchique peut servir comme point de départ à l'algorithme de partitionnement (stratégie de consolidation). La méthode STATIS est centrale à la stratégie CLUSTATIS, aussi bien pour la détermination des classes de tableaux que pour la représentation

des résultats. Plusieurs indices permettant d'évaluer la qualité de la partition sont fournis.

Il est fréquent que certaines entités parmi celles que nous voulons segmenter ne se conforment à aucune des classes établies. Chacune de ces entités sera tout de même placée dans une des classes, et, par conséquent, risque de détériorer son homogénéité. Détecter et mettre de côté ces entités permet de faire des conclusions plus pertinentes. Dans le cadre de tableaux multiples, une option dite « $K + 1$ » consistant à ajouter une classe supplémentaire contenant les tableaux atypiques a été proposée. Cette option est basée sur un principe simple : si un tableau a un indice de similarité inférieur à une valeur seuil ρ , alors il est placé dans la classe « $K + 1$ ». Une autre contribution importante est la détermination automatique du seuil de similarité ρ . Cette dernière repose sur les proximités des tableaux aussi bien à l'intérieur des classes qu'entre les classes. Dans le cadre de l'analyse sensorielle, un tableau placé dans la classe « $K + 1$ » peut correspondre à un sujet qui n'a pas compris le sens de l'épreuve, qui a fait des erreurs, ou qui tout simplement se démarque des autres dans sa perception des produits.

Le cas spécifique des données CATA a été très largement abordé. Tout d'abord, comme énoncé plus haut, par la proposition de la méthode d'analyse exploratoire CATATIS. Un test dit de consistance a été introduit, permettant de vérifier si le degré d'accord entre les sujets de l'épreuve obtenu est significatif. Ce test peut être réalisé de manière globale ou par attribut. Une stratégie de classification des sujets, nommée CLUSCATA, a également été proposée. Cette méthode est une adaptation de la méthode CLUSTATIS au cas des données CATA et suit la même stratégie globale. Une option « $K + 1$ » permettant de détecter les sujets atypiques est disponible. L'application des diverses stratégies sur des données provenant d'épreuves CATA a permis de mesurer l'impact du nombre d'attributs sur la consistance des sujets. Un trop grand nombre d'attributs a tendance à provoquer la fatigue des sujets, et peut, par conséquent, conduire à des données de mauvaise qualité.

Le choix du nombre de classes revêt une importance primordiale dans toute classification. Ainsi, nous avons proposé des tests de permutations ainsi que des adaptations d'indices existant pour la classification d'individus. De plus, les tests permettent d'évaluer si une classification de tableaux est nécessaire. Des simulations ont permis de comparer les tests et indices proposés. À travers ces simulations, l'efficacité de CLUSTATIS et de CLUSCATA a pu être une fois de plus confirmée. Nous recommandons l'utilisation d'un test de permutations afin de vérifier si une classification est nécessaire. Le cas échéant, la détermination du nombre de classes peut être réalisée via une procédure séquentielle

de tests ou une adaptation de l'indice d'Hartigan au cas des données de tableaux multiples. Une remarque importante est que si la procédure séquentielle a procuré de meilleurs résultats sur les simulations, elle a tendance à préconiser plus de classes que l'indice proposé et qu'elle est beaucoup plus « gourmande » en temps de calcul.

Il est à noter que si la stratégie CLUSTATIS a été utilisée avec succès sur des données sensorielles, qui représentent le cadre de cette thèse, il serait intéressant de l'appliquer dans d'autres domaines.

CLUSTATIS traite le cas tableaux multiples quantitatifs, et CLUSCATA celui de tableaux binaires apparentés en lignes et colonnes. Cependant, divers autres types de données de tableaux multiples méritent encore réflexion. Nous pouvons citer le cas des tableaux multiples contenant des variables qualitatives ou mixtes.

Une étude sensorielle spécifique mérite de l'attention : il s'agit de données CATA, où, en supplément, une variable hédonique est procurée par chaque sujet. Cette dernière évalue dans quelle mesure le sujet considéré apprécie les différents produits. Il s'avère que dans cette épreuve, les sujets diffèrent à la fois par leurs données CATA et par leurs appréciations. Une segmentation peut ainsi être nécessaire. Ce type de problématique a fait l'objet d'une présentation à la conférence Sensometrics en octobre 2020. Une adaptation de la méthode CLUSCATA a notamment été introduite. Cependant, ces travaux mériteraient d'être plus approfondis.

À partir d'un tableau de données quantitatives, nous pourrions imaginer que chaque variable forme un tableau et, par la suite, appliquer les méthodes de classification de tableaux préconisées dans cette thèse. De fait, une classification de variables quantitatives peut être ainsi réalisée. Des comparaisons avec les méthodes existantes sont ainsi nécessaires. Des études à ce propos sont entreprises actuellement.

Comme nous l'avons expliqué, les données de tri libre se rapportent, de fait, à des variables qualitatives, qui sont ensuite transformées en tableaux multiples par un pré-traitement adapté. Ainsi, si la méthode CLUSTATIS a été appliquée aux données de tri libre, elle pourrait l'être également afin de procéder à une classification des variables qualitatives. Des travaux sont en cours à ce sujet.

Pour terminer, une analyse exploratoire et une classification de variables mixtes peuvent

également être envisagées. Pour cela, nous pouvons considérer les variables quantitatives comme des tableaux à une seule variable, et utiliser les codages disjonctifs complets associés aux variables qualitatives, auxquels nous appliquons un traitement similaire à celui que nous avons été préconisé pour les données de tri libre. Cette perspective, très prometteuse, sera explorée dans un futur proche.

Annexes

Annexe A

Démonstrations

A.1 Méthode STATIS

Dans cette annexe, plusieurs propriétés sont démontrées. Les tableaux de données ont été centrés et normalisés de façon à avoir $\|\mathbf{W}_i\| = 1$. Par conséquent, nous avons $RV(\mathbf{W}_i, \mathbf{W}_j) = \text{trace}(\mathbf{W}_i \mathbf{W}_j)$. Il faut souligner que plusieurs propriétés sont déjà connues (Robert et Escoufier, 1976; Glaçon, 1981).

Considérons le problème de la minimisation de la quantité F avec la contrainte

$\sum_{i=1}^m \alpha_i^2 = 1$. Nous avons :

$$F = \sum_{i=1}^m \|\mathbf{W}_i - \alpha_i \mathbf{W}\|^2 = \sum_{i=1}^m \|\mathbf{W}_i\|^2 - 2 \sum_{i=1}^m \alpha_i \langle \mathbf{W}_i, \mathbf{W} \rangle + \sum_{i=1}^m \alpha_i^2 \|\mathbf{W}\|^2 = m - 2 \sum_{i=1}^m \alpha_i \langle \mathbf{W}_i, \mathbf{W} \rangle + \|\mathbf{W}\|^2.$$

L'expression lagrangienne associée au problème de minimisation s'écrit comme suit :

$L(\alpha_i, \mathbf{W}, \mu) = m - 2 \sum_{i=1}^m \alpha_i \langle \mathbf{W}_i, \mathbf{W} \rangle + \|\mathbf{W}\|^2 + \mu(\sum_{i=1}^m \alpha_i^2 - 1)$ où μ est le multiplicateur de Lagrange. En fixant la dérivée partielle de L par rapport à $\mathbf{W}$ à 0, nous obtenons :

$$\frac{\partial L}{\partial \mathbf{W}} = 0 \Leftrightarrow -2 \sum_{i=1}^m \alpha_i \mathbf{W}_i + 2\mathbf{W} = 0 \Leftrightarrow \mathbf{W} = \sum_{i=1}^m \alpha_i \mathbf{W}_i.$$

En remplaçant, dans l'expression de F , $\sum_{i=1}^m \alpha_i \mathbf{W}_i$ par $\mathbf{W}$, nous avons :

$$F = m - 2 \sum_{i=1}^m \alpha_i \langle \mathbf{W}_i, \mathbf{W} \rangle + \|\mathbf{W}\|^2 = m - 2\|\mathbf{W}\|^2 + \|\mathbf{W}\|^2 = m - \|\mathbf{W}\|^2.$$

Si à la place, nous remplaçons $\mathbf{W}$ par $\sum_{i=1}^m \alpha_i \mathbf{W}_i$, nous obtenons :

$F = m - \sum_{l=1}^m \sum_{i=1}^m \alpha_i \alpha_l \langle \mathbf{W}_i, \mathbf{W}_l \rangle = m - \sum_{l=1}^m \sum_{i=1}^m \alpha_i \alpha_l RV(\mathbf{W}_i, \mathbf{W}_l) = m - \boldsymbol{\alpha}^\top \mathbf{R} \boldsymbol{\alpha}$ où $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)^\top$ et $\mathbf{R}$ est la matrice qui contient les coefficients RV entre les matrices des produits scalaires. Il découle que la minimisation de F par rapport à $\boldsymbol{\alpha}$ revient à la

maximisation de la quantité $\boldsymbol{\alpha}^\top \mathbf{R} \boldsymbol{\alpha}$ sous la contrainte $\|\boldsymbol{\alpha}\| = 1$. La solution de ce problème est donnée par le vecteur propre de $\mathbf{R}$ associé à la plus grande valeur propre, λ_1 . Ceci montre que nous obtenons la même solution que la méthode STATIS.

Nous avons donc montré que $\|\mathbf{W}\|^2 = \boldsymbol{\alpha}^\top \mathbf{R} \boldsymbol{\alpha} = \lambda_1 \boldsymbol{\alpha}^\top \boldsymbol{\alpha} = \lambda_1$ (propriété (i)).

De l'expression $\mathbf{R} \boldsymbol{\alpha} = \lambda_1 \boldsymbol{\alpha}$, il découle que pour chaque i ,

$$\sum_{l=1}^m RV(\mathbf{W}_i, \mathbf{W}_l) \alpha_l = \lambda_1 \alpha_i \Leftrightarrow \sum_{l=1}^m \langle \mathbf{W}_i, \mathbf{W}_l \rangle \alpha_l = \lambda_1 \alpha_i \Leftrightarrow \langle \mathbf{W}_i, \mathbf{W} \rangle = \lambda_1 \alpha_i \Leftrightarrow \alpha_i = \frac{\langle \mathbf{W}_i, \mathbf{W} \rangle}{\lambda_1} = \frac{RV(\mathbf{W}_i, \mathbf{W})}{\|\mathbf{W}\|} = \frac{RV(\mathbf{W}_i, \mathbf{W})}{\sqrt{\lambda_1}} \text{ (propriété (ii))}.$$

Puisque $\|\mathbf{W}\|^2 = \lambda_1$, $F = m - \|\mathbf{W}\|^2 = m - \lambda_1$ (propriété (iii)) et $m = F + \lambda_1$ (propriété (iv)).

A.2 Méthode CLUSTATIS

Dans cette annexe, nous montrons que pour deux classes de tableaux que nous désignons par A et B , $\lambda_1^{(A \cup B)} \leq \lambda_1^{(A)} + \lambda_1^{(B)}$, où $\lambda_1^{(A)}$, $\lambda_1^{(B)}$ et $\lambda_1^{(A \cup B)}$ sont respectivement les plus grandes valeurs propres des matrices composées des coefficients RV entre les tableaux des classes A , B et $A \cup B$.

Comme remarque préliminaire, nous pouvons noter que $RV(\mathbf{W}_i, \mathbf{W}_j) = \mathbf{v}_i^\top \mathbf{v}_j$ où $\mathbf{v}_i = \text{Vec}(\mathbf{W}_i)$ et $\mathbf{v}_j = \text{Vec}(\mathbf{W}_j)$. Le vecteur $\mathbf{v}_i$ (respectivement, $\mathbf{v}_j$) est obtenu en mettant les colonnes de $\mathbf{W}_i$ (respectivement, $\mathbf{W}_j$) les unes en dessous des autres de façon à former un vecteur unique.

Notons $\mathbf{R}_{A \cup B}$ la matrice des coefficients RV entre les tableaux dans $A \cup B$ pris deux à deux. Nous avons $\mathbf{R}_{A \cup B} = \mathbf{V}_{A \cup B}^\top \mathbf{V}_{A \cup B}$ où $\mathbf{V}_{A \cup B}$ contient comme colonnes les $\mathbf{v}_i$ (*i.e.*, les $\mathbf{W}_i$ vectorisés) associés aux tableaux dans $A \cup B$. De la même façon, nous notons $\mathbf{V}_A$ (respectivement, $\mathbf{V}_B$) la matrice obtenue par vectorisation des matrices des produits scalaires associées aux tableaux dans A (respectivement, B). La première valeur propre de $\mathbf{V}_{A \cup B}^\top \mathbf{V}_{A \cup B}$ est égale à celle de $\mathbf{V}_{A \cup B} \mathbf{V}_{A \cup B}^\top$. Nous avons ainsi :

$$\begin{aligned}
\lambda_1^{(A \cup B)} &= \max_{\alpha, \|\alpha\|=1} \{\alpha^\top V_{A \cup B} V_{A \cup B}^\top \alpha\} \\
&= \max_{\alpha, \|\alpha\|=1} \{\alpha^\top (V_A V_A^\top + V_B V_B^\top) \alpha\} \\
&= \max_{\alpha, \|\alpha\|=1} \{\alpha^\top V_A V_A^\top \alpha + \alpha^\top V_B V_B^\top \alpha\} \\
&\leq \max_{\alpha, \|\alpha\|=1} \{\alpha^\top V_A V_A^\top \alpha\} + \max_{\alpha, \|\alpha\|=1} \{\alpha^\top V_B V_B^\top \alpha\} \\
&= \lambda_1^{(A)} + \lambda_1^{(B)}
\end{aligned}$$

Annexe B

Valorisation des travaux de thèse

B.1 Publications

- Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2020). Analysis and clustering of multiblock datasets by means of the STATIS and CLUSTATIS methods. Application to sensometrics. *Food Quality and Preference*, 79, 103520.

Abstract : *The STATIS method has been successfully applied to the analysis of sensory profiling data and other kinds data in sensometrics. We discuss its use and benefits and compare its outcomes to alternative methods for the analysis of multiblock data arising in situations such as projective mapping and free sorting experiments. More importantly, a method of clustering a collection of datasets measured on the same individuals, called CLUSTATIS, is introduced. It is based on the optimization of a criterion and consists in a hierarchical cluster analysis and a partitioning algorithm akin to the K-means algorithm. The procedure of analysis can be seen as an extension of the cluster analysis of variables around latent components (Vigneau et Qannari, 2003) to the case of blocks of variables. Alongside the determination of the clusters, a latent configuration is determined by the STATIS method. The interest of CLUSTATIS in sensometrics is discussed and illustrated on the basis of two case studies pertaining to the projective mapping also called Napping and the free sorting tasks, respectively.*

- Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2019). A new approach for the analysis of data and the clustering of subjects in a CATA experiment. *Food Quality and Preference*, 72, 31-39.

Abstract : *A new approach for the analysis of the data and the clustering of the subjects in a Check All That Apply (CATA) experiment is outlined. It encompasses indices to assess the agreements among the subjects. These indices are taken into account for the analysis of the CATA data and the segmentation of the subjects. The analysis of the CATA data bears some similarities to the STATIS method and the cluster analysis of the subjects follows the same pattern as a clustering approach, called CLV, for the cluster analysis of variables, and a clustering approach, called CLUSTATIS, for clustering datasets.*

- Llobell, F., Vigneau, E., Qannari, E. M. (2019). Clustering datasets by means of CLUSTATIS with identification of atypical datasets. Application to sensometrics. *Food Quality and Preference*, 75, 97-104.

Abstract : *A method of clustering a collection of datasets measured on the same individuals, called CLUSTATIS, was introduced and applied to the segmentation of the subjects participating in a projective mapping experiment or a free sorting task. A refinement of this method of clustering is proposed. It consists in segmenting the subjects while discarding those subjects who can be considered as atypical because they do not fit to the pattern of any cluster. This strategy of analysis requires the determination of a threshold parameter that delineates the boundary between the main clusters and the noise cluster that contains the atypical subjects. An appropriate selection of this parameter is proposed. The general strategy of analysis is illustrated on the basis of a simulation study and data from projective mapping and free sorting experiments.*

- Llobell, F., Giacalone, D., Labenne, A., Qannari, E.M. (2019). Assessment of the agreement and cluster analysis of the respondents in a CATA experiment. *Food Quality and Preference*, 77, 184-190.

Abstract : *Statistical tools to assess the agreement of respondents in a Check-All-That-Apply (CATA) experiment are discussed. An overall index of agreement is introduced and a hypothesis test to assess the significance of this index is outlined. A similar investigation at the level of each attribute is undertaken. The permutation test proposed in this latter situation can be compared to Cochran's Q test. We also propose to cluster the respondents while setting aside those respondents that are*

deemed to be atypical. This strategy of clustering stands a refinement of the cluster analysis called *CLUSCATA* that was introduced in a previous paper. The various strategies of analysis are illustrated by means of real case studies.

- Llobell, F. & Qannari, E.M. (2020). **CLUSTATIS : Cluster analysis of blocks of variables.** *Electronic Journal of Applied Statistical Analysis*, in Press.

Abstract : *The STATIS method is one of many strategies of analysis devoted to the unsupervised analysis of multiblock data. A new optimization criterion to define this method of analysis is introduced and an extension to the cluster analysis of several blocks of variables is discussed. This consists in a hierarchical cluster analysis and a partitioning algorithm akin to the K-means algorithm. Moreover, in order to improve the cluster analysis outcomes, an additional cluster called noise cluster which contains atypical blocks of variables is introduced. The general strategy of analysis is illustrated by means of two case studies.*

B.2 Communications orales

- Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2018). CLUSTATIS : a cluster analysis of multiblock datasets. Application to sensometrics. *AgroStat*, Marseille, France.
- Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2018). Cluster analysis of multiblock datasets. Application to Projective mapping/Napping and free sorting task. *Sensometrics*, Montevideo, Uruguay.
- Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2018). Consumer segmentation in a check-all-that-apply task. *SenseAsia*, Kuala Lumpur, Malaisie.
- Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2018). CLUSTATIS : a cluster analysis of multiblock datasets. *STDMA*, Crète, Grèce.
- Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2018). Consumer segmentation in a Check All That Apply experiment. *Rencontres de la Société Francophone de Classification*, Paris, France.
- Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2019). A cluster analysis of multiblock datasets. *ASMDA*, Florence, Italie.
- Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2019). Clust-Block : a package for clustering datasets. *UseR !*, Toulouse, France.

- Llobell, F., Cariou, V., Vigneau, E., Giacalone, D., Labenne, A., Qannari, E.M. (2019). Segmentation of the subjects in a CATA experiment while setting aside atypical subjects. *Pangborn*, Édimbourg, Écosse.
- Llobell, F., Cariou, V., Vigneau, E., Labenne, A., Qannari, E. M. (2020). Determination of the number of clusters of subjects in Projective Mapping, Free Sorting and CATA experiments. *Sensometrics*, Stavanger, Norvège.
- Vigneau, E., Cariou, V., Giacalone, D., Berget, I., Llobell, F. (2020). Combining hedonic information and CATA description for consumer segmentation : new methodological proposals and comparison. *Sensometrics*, Stavanger, Norvège.

B.3 Développements

B.3.1 Logiciel *R*

- Package *R* ClustBlock

B.3.2 XLSTAT

- STATIS
- CATATIS
- Analyse de données de tri libre
- CLUSTATIS
- CLUSCATA

Bibliographie

- ABDI, H., VALENTIN, D., CHOLLET, S. et CHREA, C. (2007). Analyzing assessors and products in sorting tasks : DISTATIS, theory and applications. *Food quality and preference*, 18(4):627–640.
- ADDINSOFT (2020). XLSTAT : Logiciel de statistique et d’analyse de données pour Microsoft Excel.
- ARES, G. et JAEGER, S. (2015). Check-all-that-apply (CATA) questions with consumers in practice : experimental considerations and impact on outcome. *In Rapid sensory profiling techniques*, pages 227–245. Elsevier.
- ARES, G. et JAEGER, S. R. (2013). Check-all-that-apply questions : Influence of attribute order on sensory product characterization. *Food Quality and Preference*, 28(1):141–153.
- BELLANGER, L. (2015). Contributions à l’exploration de données environnementales, écologiques, médicales et archéologiques. Statistiques [math.ST]. Université de Nantes.
- BERGET, I., VARELA, P. et NÆS, T. (2019). Segmentation in projective mapping. *Food Quality and Preference*, 71:8–20.
- CADORET, M., LÊ, S. et PAGÈS, J. (2009). A factorial approach for sorting task data (FAST). *Food quality and preference*, 20(6):410–417.
- CALIŃSKI, T. et HARABASZ, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.
- CARIOU, V. et QANNARI, E. (2018). Statistical treatment of free sorting data by means of correspondence and cluster analyses. *Food quality and preference*, 68:1–11.
- CARIOU, V., QANNARI, E. M., RUTLEDGE, D. N. et VIGNEAU, E. (2018). ComDim : From multiblock data analysis to path modeling. *Food Quality and Preference*, 67:27–34.

- CARIOU, V., VERDUN, S., DIAZ, E., QANNARI, E. M. et VIGNEAU, E. (2009). Comparison of three hypothesis testing approaches for the selection of the appropriate number of clusters of variables. *Advances in data analysis and classification*, 3(3):227.
- CARIOU, V. et WILDERJANS, T. F. (2018). Consumer segmentation in multi-attribute product evaluation by means of non-negatively constrained CLV3W. *Food Quality and Preference*, 67:18–26.
- CAZES, P., BONNEFOUS, S., BAUMERDER, A. *et al.* (1976). Description coh rente des variables qualitatives prises globalement et de leurs modalit s. *Statistique et analyse des donn es*, 1(2):48–62.
- CHARRAD, M., GHAZZALI, N., BOITEAU, V., NIKNAFS, A. et CHARRAD, M. M. (2014). Package ‘nbclust’. *Journal of statistical software*, 61:1–36.
- COCHRAN, W. G. (1950). The comparison of percentages in matched samples. *Biometrika*, 37(3/4):256–266.
- CONDE, A. et DOM NGUEZ, J. (2018). Scaling the chord and Hellinger distances in the range $[0, 1]$: An option to consider. *Journal of Asia-Pacific Biodiversity*, 11(1):161–166.
- COURCOUX, P. (2017). *FreeSortR : Free Sorting Data Analysis*. R package version 1.3.
- COURCOUX, P., FAYE, P. et QANNARI, E. (2014). Determination of the consensus partition and cluster analysis of subjects in a free sorting task experiment. *Food quality and preference*, 32:107–112.
- COURCOUX, P., QANNARI, E., TAYLOR, Y., BUCK, D. et GREENHOFF, K. (2012). Taxonomic free sorting. *Food Quality and Preference*, 23(1):30–35.
- DAHL, T. et N ES, T. (2004). Outlier and group detection in sensory panels using hierarchical cluster analysis with the Procrustes distance. *Food Quality and Preference*, 15(3):195–208.
- DAVE, R. N. (1991). Characterization and detection of noise in clustering. *Pattern Recognition Letters*, 12(11):657–664.
- DE ROOVER, K., CEULEMANS, E. et TIMMERMAN, M. E. (2012). How to perform multiblock component analysis in practice. *Behavior Research Methods*, 44(1):41–56.
- DELARUE, J., LAWLOR, B. et ROGEAUX, M. (2014). *Rapid sensory profiling techniques : Applications in new product development and consumer research*. Elsevier.

BIBLIOGRAPHIE

- DESGRAUPES, B. (2013). Clustering indices. *University of Paris Ouest-Lab Modal'X*, 1:34.
- EL GHAZIRI, A. et QANNARI, E. M. (2015). Measures of association between two datasets ; Application to sensory data. *Food Quality and Preference*, 40:116–124.
- EVERITT, B. S., LANDAU, S., LEESE, M. et STAHL, D. (2011). Cluster analysis. –John Wiley & Sons. *Ltd., New York*, page 330.
- FAYE, P., BRÉMAUD, D., DAUBIN, M. D., COURCOUX, P., GIBOREAU, A. et NICOD, H. (2004). Perceptive free sorting and verbalization tasks with naive subjects : an alternative to descriptive mappings. *Food quality and preference*, 15(7-8):781–791.
- FAYE, P., COURCOUX, P., QANNARI, M. et GIBOREAU, A. (2011). Méthodes de traitement statistique des données issues d’une épreuve de tri libre. *Revue Modulad*, 43:1–24.
- GIACALONE, D. (2018). Product performance optimization. *In Methods in Consumer Research, Volume 1*, pages 159–185. Elsevier.
- GIACALONE, D., BREDIE, W. L. et FRØST, M. B. (2013). “All-In-One Test”(Ai1) : A rapid and easily applicable approach to consumer product testing. *Food Quality and Preference*, 27(2):108–119.
- GIACALONE, D., DEGN, T. K., YANG, N., LIU, C., FISK, I. et MÜNCHOW, M. (2019). Common roasting defects in coffee : Aroma composition, sensory characterization and consumer perception. *Food quality and preference*, 71:463–474.
- GIACALONE, D., FRØST, M. B., BREDIE, W. L., PINEAU, B., HUNTER, D. C., PAISLEY, A. G., BERESFORD, M. K. et JAEGER, S. R. (2015). Situational appropriateness of beer is influenced by product familiarity. *Food Quality and Preference*, 39:16–27.
- GIACALONE, D. et JAEGER, S. R. (2016). Better the devil you know ? How product familiarity affects usage versatility of foods and beverages. *Journal of Economic Psychology*, 55:120–138.
- GIALAMPOUKIDIS, I., VROCHIDIS, S., KOMPATSIARIS, I. et ANTONIOU, I. (2019). Probabilistic density-based estimation of the number of clusters using the DBSCAN-martingale process. *Pattern Recognition Letters*, 123:23–30.
- GLAÇON, F. (1981). *Analyse conjointe de plusieurs matrices de données : Comparaison de différentes méthodes*. Thèse de doctorat, Université scientifique et médicale de Grenoble.

- GOWER, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, 40(1):33–51.
- GREENACRE, M. (2017). *Correspondence analysis in practice*. Chapman and Hall/CRC.
- GUPTA, A., DATTA, S. et DAS, S. (2018). Fast automatic estimation of the number of clusters from the minimum inter-center distance for k-means clustering. *Pattern Recognition Letters*, 116:72–79.
- HAMERS, L. *et al.* (1989). Similarity measures in scientometric research : The Jaccard index versus Salton’s cosine formula. *Information Processing and Management*, 25(3): 315–18.
- HARDY, A. (1996). On the number of clusters. *Computational Statistics & Data Analysis*, 23(1):83–96.
- HARTIGAN, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.
- HUBERT, L. et ARABIE, P. (1985). Comparing Partitions. *Journal of classification*, 2(1):193–218.
- JACCARD, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.
- JAEGER, S. R., BERESFORD, M. K., PAISLEY, A. G., ANTÚNEZ, L., VIDAL, L., CADENA, R. S., GIMÉNEZ, A. et ARES, G. (2015). Check-all-that-apply (CATA) questions for sensory product characterization by consumers : Investigations into the number of terms used in CATA questions. *Food Quality and Preference*, 42:154–164.
- JAEGER, S. R., GIACALONE, D., ROIGARD, C. M., PINEAU, B., VIDAL, L., GIMÉNEZ, A., FRØST, M. B. et ARES, G. (2013). Investigation of bias of hedonic scores when co-eliciting product attribute information using CATA questions. *Food quality and preference*, 30(2):242–249.
- JOSSE, J., PAGÈS, J. et HUSSON, F. (2008). Testing the significance of the RV coefficient. *Computational Statistics & Data Analysis*, 53(1):82–91.
- KAZI-AOUAL, F., HITIER, S., SABATIER, R. et LEBRETON, J.-D. (1995). Refined approximations to permutation tests for multivariate inference. *Computational statistics & data analysis*, 20(6):643–656.
- KOTHARI, R. et PITTS, D. (1999). On finding the number of clusters. *Pattern Recognition Letters*, 20(4):405–416.

BIBLIOGRAPHIE

- LACHENBRUCH, P. A. (2014). McNemar test. *Wiley StatsRef : Statistics Reference Online*.
- LAHNE, J., ABDI, H. et HEYMANN, H. (2018). Rapid sensory profiles with DISTATIS and Barycentric Text Projection : An example with amari, bitter herbal liqueurs. *Food quality and preference*, 66:36–43.
- LAVIT, C., ESCOUFIER, Y., SABATIER, R. et TRAISSAC, P. (1994). The act (statis method). *Computational Statistics & Data Analysis*, 18(1):97–119.
- LAWLESS, H. T. (1989). Exploration of fragrance categories and ambiguous odors using multidimensional scaling and cluster analysis. *Chemical Senses*, 14(3):349–360.
- LAWLESS, H. T., SHENG, N. et KNOOPS, S. S. (1995). Multidimensional scaling of sorting data applied to cheese perception. *Food Quality and Preference*, 6(2):91–98.
- LÊ, S. et HUSSON, F. (2008). Sensominer : A package for sensory data analysis. *Journal of Sensory Studies*, 23(1):14–25.
- LÊ, S., JOSSE, J., HUSSON, F. *et al.* (2008). FactoMineR : an R package for multivariate analysis. *Journal of statistical software*, 25(1):1–18.
- LLOBELL, F., VIGNEAU, E., CARIOU, V. et QANNARI, E. M. (2020). *ClustBlock : Clustering of Datasets*. R package version 2.2.0.
- MACGREGOR, J. F., JAECKLE, C., KIPARISSIDES, C. et KOUTOUDI, M. (1994). Process monitoring and diagnosis by multiblock PLS methods. *AIChE Journal*, 40(5):826–838.
- MEYER, C. D. (2000). *Matrix analysis and applied linear algebra*, volume 71. Siam.
- MEYNER, M. et CASTURA, J. C. (2014). Check-all-that-apply questions. In *Novel techniques in sensory characterization and consumer profiling*, pages 284–319. CRC Press.
- MEYNER, M., CASTURA, J. C. et CARR, B. T. (2013). Existing and new approaches for the analysis of CATA data. *Food Quality and Preference*, 30(2):309–319.
- MEYNER, M., CASTURA, J. C. et WORCH, T. (2016). Statistical evaluation of panel repeatability in Check-All-That-Apply questions. *Food quality and preference*, 49:197–204.
- MILLIGAN, G. W. et COOPER, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

- NÆS, T., BERGET, I., LILAND, K. H., ARES, G. et VARELA, P. (2017). Estimating and interpreting more than two consensus components in projective mapping : INDSCAL vs. multiple factor analysis (MFA). *Food quality and preference*, 58:45–60.
- NIANG, N. et OUATTARA, M. (2019). Weighted consensus clustering for multiblock data. *In SFC 2019*.
- OCHIAI, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. *Bulletin of Japanese Society of Scientific Fisheries*, 22:526–530.
- PAGÈS, J. (2005). Collection and analysis of perceived product inter-distances using multiple factor analysis : Application to the study of 10 white wines from the loire valley. *Food quality and preference*, 16(7):642–649.
- PENA, J. M., LOZANO, J. A. et LARRANAGA, P. (1999). An empirical comparison of four initialization methods for the k-means algorithm. *Pattern recognition letters*, 20(10):1027–1040.
- QANNARI, E., CARIOU, V., TEILLET, E. et SCHLICH, P. (2010). SORT-CC : A procedure for the statistical treatment of free sorting data. *Food quality and preference*, 21(3):302–308.
- QANNARI, E. M., WAKELING, I., COURCOUX, P. et MACFIE, H. J. (2000). Defining the underlying sensory dimensions. *Food Quality and Preference*, 11(1-2):151–154.
- R CORE TEAM (2020). R : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>.
- RAND, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.
- REINBACH, H. C., GIACALONE, D., RIBEIRO, L. M., BREDIE, W. L. et FRØST, M. B. (2014). Comparison of three sensory profiling methods based on consumer perception : CATA, CATA with intensity and Napping®. *Food Quality and Preference*, 32:160–166.
- RISVIK, E., MCEWAN, J. A., COLWILL, J. S., ROGERS, R. et LYON, D. H. (1994). Projective mapping : A tool for sensory analysis and consumer research. *Food quality and preference*, 5(4):263–269.

BIBLIOGRAPHIE

- ROBERT, P. et ESCOUFIER, Y. (1976). A unifying tool for linear multivariate statistical methods : the RV-coefficient. *Journal of the Royal Statistical Society : Series C (Applied Statistics)*, 25(3):257–265.
- ROUSSEEUW, P. J. (1987). Silhouettes : a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.
- SAHMER, K., HANAFLI, M. et EL QANNARI, M. (2006). Assessing unidimensionality within PLS path modeling framework. *In From data and information analysis to knowledge engineering*, pages 222–229. Springer.
- SALTON, G. et MCGILL, M. J. (1983). *Introduction to modern information retrieval*. mcgraw-hill.
- SAPORTA, G. (1976). Quelques applications des opérateurs d’escoufier au traitement des variables qualitatives. *Statistique et analyse des données*, 1(1):38–46.
- SAPORTA, G. (2006). *Probabilités, analyse des données et statistique*. Editions Technip.
- SCHLICH, P. (1996). Defining and validating assessor compromises about product distances and attribute correlations. *In Data handling in science and technology*, volume 16, pages 259–306. Elsevier.
- TAKANE, Y. (1981). MDSORT : A special-purpose multidimensional scaling program for sorting data. *Behavior Research Methods & Instrumentation*, 13(5):698–698.
- THIOULOUSE, J., DRAY, S. *et al.* (2007). Interactive multivariate data analysis in R with the ade4 and ade4TkGUI packages. *Journal of Statistical Software*, 22(5):1–14.
- TIBSHIRANI, R., WALTHER, G. et HASTIE, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 63(2):411–423.
- TOMIC, O., BERGET, I. et NÆS, T. (2015). A comparison of generalised procrustes analysis and multiple factor analysis for projective mapping data. *Food quality and preference*, 43:34–46.
- Van der KLOOT, W. A. et VAN HERK, H. (1991). Multidimensional scaling of sorting data : A comparison of three procedures. *Multivariate Behavioral Research*, 26(4):563–581.

- VARELA, P., ANTÚNEZ, L., BERGET, I., OLIVEIRA, D., CHRISTENSEN, K., VIDAL, L., NAES, T. et ARES, G. (2017). Influence of consumers' cognitive style on results from projective mapping. *Food Research International*, 99:693–701.
- VARELA, P. et ARES, G. (2014). *Novel techniques in sensory characterization and consumer profiling*. CRC Press.
- VIDAL, L., TÁRREGA, A., ANTÚNEZ, L., ARES, G. et JAEGER, S. R. (2015). Comparison of correspondence analysis based on Hellinger and chi-square distances to obtain sensory spaces from check-all-that-apply (CATA) questions. *Food Quality and Preference*, 43: 106–112.
- VIGNEAU, E., CHEN, M. et CARIOU, V. (2020). *ClustVarLV : Clustering of Variables Around Latent Variables*. R package version 2.0.1.
- VIGNEAU, E. et QANNARI, E. (2003). Clustering of variables around latent components. *Communications in Statistics-Simulation and Computation*, 32(4):1131–1150.
- VIGNEAU, E., QANNARI, E., NAVEZ, B. et COTTET, V. (2016). Segmentation of consumers in preference studies while setting aside atypical or irrelevant consumers. *Food quality and preference*, 47:54–63.
- WARRENS, M. J. *et al.* (2008). *Similarity coefficients for binary data : properties of coefficients, coefficient matrices, multi-way metrics and multivariate coefficients*. Thèse de doctorat, University of Groningen.
- WEISSTEIN, E. W. (2004). *Bonferroni correction*. Wolfram Research, Inc.
- WESTAD, F., HERSLETH, M. et LEA, P. (2004). Strategies for consumer segmentation with applications on preference data. *Food Quality and Preference*, 15(7-8):681–687.
- WIJAYA, S. H., AFENDI, F. M., BATUBARA, I., DARUSMAN, L. K., ALTAF-UL-AMIN, M. et KANAYA, S. (2016). Finding an appropriate equation to measure similarity between binary vectors : case studies on Indonesian and Japanese herbal medicines. *BMC bioinformatics*, 17(1):520.
- WILDERJANS, T. F. et CARIOU, V. (2016). CLV3W : A clustering around latent variables approach to detect panel disagreement in three-way conventional sensory profiling data. *Food Quality and Preference*, 47:45–53.
- WILEY, D. E. (1967). Latent partition analysis. *Psychometrika*, 32(2):183–193.

BIBLIOGRAPHIE

- WILLIAMS, A. A. et LANGRON, S. P. (1984). The use of free-choice profiling for the evaluation of commercial ports. *Journal of the Science of Food and Agriculture*, 35(5): 558–568.
- WORCH, T. et PIQUERAS-FISZMAN, B. (2015). Contributions to assess the reproducibility and the agreement of respondents in CATA tasks. *Food Quality and Preference*, 40:137–146.

Titre : Classification de tableaux de données, applications en analyse sensorielle

Mots clés : Tableaux multiples, Classification, Analyse sensorielle

Résumé : Les données structurées sous forme de tableaux se rapportant aux mêmes individus sont de plus en plus fréquentes dans plusieurs secteurs d'application. C'est en particulier le cas en évaluation sensorielle où plusieurs épreuves conduisent à l'obtention de tableaux multiples ; chaque tableau étant rapporté à un sujet (juge, consommateur, ...). L'analyse exploratoire de ce type de données a suscité un vif intérêt durant les trente dernières années. Cependant, la classification de tableaux multiples n'a été que très peu abordée alors que le besoin pour ce type de données est important.

Dans ce contexte, une méthode appelée CLUSTATIS permettant de segmenter les tableaux de données est proposée. Au cœur de cette approche se trouve la méthode STATIS, qui est une stratégie d'analyse exploratoire de tableaux multiples.

Plusieurs extensions de la méthode de classification CLUSTATIS sont présentées. En

particulier, le cas des données issues d'une épreuve dite « Check-All-That-Apply » (CATA) est considéré. Une méthode de classification ad-hoc, nommée CLUSCATA, est discutée.

Afin d'améliorer l'homogénéité des classes issues aussi bien de CLUSTATIS que de CLUSCATA, une option consistant à rajouter une classe supplémentaire, appelée « K+1 », est introduite. Cette classe additionnelle a pour vocation de collecter les tableaux de données identifiés comme atypiques.

Le choix du nombre de classes est abordé, et des solutions sont proposées. Des applications dans le cadre de l'évaluation sensorielle ainsi que des études de simulation permettent de souligner la pertinence de l'approche de classification.

Des implémentations dans le logiciel XLSTAT et dans l'environnement R sont présentées.

Title: Clustering of datasets, applications to sensory analysis

Keywords: Multiblock datasets, Cluster analysis, Sensory analysis

Abstract: Multiblock datasets are more and more frequent in several areas of application. This is particularly the case in sensory evaluation where several tests lead to multiblock datasets, each dataset being related to a subject (judge, consumer, ...). The statistical analysis of this type of data has raised an increasing interest over the last thirty years. However, the clustering of multiblock datasets has received little attention, even though there is an important need for this type of data.

In this context, a method called CLUSTATIS devoted to the cluster analysis of datasets is proposed. At the heart of this approach is the STATIS method, which is a multiblock datasets analysis strategy.

Several extensions of the CLUSTATIS clustering method are presented. In

particular, the case of data from the so-called "Check-All-That-Apply" (CATA) task is considered. An ad-hoc clustering method called CLUSCATA is discussed.

In order to improve the homogeneity of clusters from both CLUSTATIS and CLUSCATA, an option to add an additional cluster, called "K+1", is introduced. The purpose of this additional cluster is to collect datasets identified as atypical.

The choice of the number of clusters is discussed, and solutions are proposed. Applications in sensory analysis as well as simulation studies highlight the relevance of the clustering approach.

Implementations in the XLSTAT software and in the R environment are presented.