

# CLASSIFICATION DE PROFILS BINAIRES VIA LA MODÉLISATION DE LA STRUCTURE D'ACCORD/DÉSACCORD

Margot Brard <sup>1</sup> & Sébastien Lê <sup>2</sup> & David Causeur <sup>3</sup>

<sup>1</sup> *mbrard@centreculinaire.com*

<sup>2</sup> *sebastien.le@agrocampus-ouest.fr*

<sup>3</sup> *david.causeur@agrocampus-ouest.fr*

*Centre Culinaire Contemporain, 8 rue Jules Maillard-de-la Gournerie, Rennes & AGROCAMPUS OUEST (IRMAR, UMR CNRS 6625), 65 rue de Saint-Brieuc, Rennes*

**Résumé.** L'analyse de la perception par les consommateurs de concepts complexes associés à des produits fait l'objet d'une attention grandissante dans le domaine du marketing. La complexité d'un concept fait notamment référence à sa multidimensionnalité, l'évaluation subjective s'appuyant sur un ensemble de critères dont l'importance relative est propre à chaque consommateur. À l'échelle d'une population de consommateurs, cette complexité se traduit par une diversité de perceptions, que les études marketing visent à analyser. Inspirée d'approches utilisées dans le domaine de la psychologie pour étudier la subjectivité humaine, la QC-methodology permet d'étudier la diversité de perceptions d'un concept complexe donné au sein d'un ensemble de stimuli. Cette méthodologie fournit des données dans lesquelles chaque sujet peut être assimilé à un profil d'évaluations binaires. Dans ce travail, nous présentons une méthode de segmentation de ces profils basée sur un modèle spécifique de l'accord/désaccord entre sujets. En particulier, dans le cadre général que nous proposons, une séquence de tests de rapport de vraisemblance visant à déterminer le nombre de clusters de désaccord au sein d'un panel de sujets est présentée. La méthode est corroborée en évaluant ses performances dans des situations caricaturales de désaccord. La méthode est également appliquée à un jeu de données obtenu dans le cadre d'une étude réalisée au Centre Culinaire Contemporain (Rennes) auprès de 72 sujets pour caractériser le concept de « bon geste » au sein d'un ensemble de 39 stimuli vidéo représentant des gestes culinaires.

**Mots-clés.** Marketing, Classification non-supervisée, Accord inter-sujets, Variables binaires multivariées.

**Abstract.** Analysing consumers' perception of complex concepts through a set of products is becoming of the main interest in the marketing field. The complexity of a concept refers to its multidimensionality, the subjective assessment of this concept being based on various criteria for which the relative importance is specific to each consumer. At the level of a population of consumers, this complexity results in a diversity of perceptions, which marketing studies aim to analyze. Drawing its inspiration from approaches used in

the psychology field to study human subjectivity, the QC-methodology aims to study the diversity of perceptions of a given complex concept through a set of stimuli. This methodology provides data in which each subject can be assimilated to a profile of binary evaluations. In this research, we present a method allowing the segmentation of these profiles that is based on a specific model of agreement/disagreement between subjects. In particular, in the general framework that we propose, a strategy based on likelihood ratio tests aiming at determining the number of disagreed clusters within the panel of subjects is presented. The method is corroborated by evaluating its performances when being applied in caricatural situations of agreement/disagreement. The method is also applied to a dataset obtained during a study conducted at the Centre Culinaire Contemporain (Rennes, France) with 72 subjects who characterized the concept of « good gesture » through a set of 39 video stimuli representing culinary gestures.

**Keywords.** Marketing, Unsupervised clustering, Inter-rater agreement, Multivariate binary variables.

## 1 Contexte et problématique

L'analyse de la perception par les consommateurs de concepts complexes associés à des produits, par exemple leur caractère innovant, fait l'objet d'une attention grandissante dans le domaine du marketing. La complexité d'un concept fait notamment référence à sa multidimensionnalité : l'évaluation subjective consistant à déterminer qu'un produit est innovant ou non s'appuie ainsi sur un ensemble de critères, dont l'importance relative est propre à chaque consommateur. À l'échelle d'une population de consommateurs, cette complexité se traduit aussi par une diversité de perceptions, que les études marketing visent à analyser.

Inspirée d'approches utilisées dans le domaine de la psychologie pour étudier la subjectivité humaine, la QC-methodology permet d'étudier la diversité de perceptions d'un concept complexe donné au sein d'un ensemble de stimuli (i.e., généralement un ensemble de produits) [1]. En pratique,  $J$  sujets évaluent la représentation du concept à travers un ensemble de  $I$  stimuli. Pour le  $i$ -ème stimulus, l'évaluation du  $j$ -ème sujet est une variable binaire  $Y_{ij}$ , qui vaut 1 si le stimulus a été évalué comme représentatif du concept par le sujet, et 0 si le stimulus a été évalué comme non-représentatif du concept par le sujet. L'enjeu au cœur du travail présenté ici est de présenter une méthode de segmentation du panel de sujets dans ce genre de situations. Plus précisément, l'enjeu est de déterminer si les  $J$  sujets ont perçu le concept de manière consensuelle, ou, dans le cas contraire, de déterminer si des groupes homogènes de sujets (*clusters*) peuvent être identifiés. D'un point de vue méthodologique, cette problématique relève de la classification non-supervisée de profils multivariés binaires, dans laquelle la distance entre profils traduit de manière spécifique un désaccord entre sujets. Dans la suite, une mesure de distance spécifique de

notre problématique et un algorithme de classification ascendante hiérarchique associé sont présentés.

Lorsque les évaluations sur un même ensemble de stimuli sont binaires, différents indicateurs d'accord entre sujets, le plus intuitif étant le pourcentage de jugements identiques, sont habituellement utilisés pour quantifier le degré d'accord qu'il existe entre les évaluations fournies par deux sujets [2,3]. Cependant, de telles approches n'exploitent pas la structure multivariée de l'évaluation fournie par un sujet. Pour pallier ce problème, nous présentons une solution alternative à partir d'un modèle logistique appliqué aux évaluations fournies par les sujets [4]. En pratique, nous proposons de définir le désaccord entre deux sujets comme la distance Euclidienne entre leurs profils de résidus de la déviance d'un modèle d'accord parfait ajusté sur la globalité du panel. La matrice de distances ainsi obtenue, que l'on appellera par la suite *matrice de désaccord entre sujets*, est ensuite soumise à un algorithme de classification ascendante hiérarchique pour obtenir un dendrogramme représentant la structure de désaccord au sein du panel. À l'instar d'approches récentes visant à offrir des outils statistiques d'estimation du nombre de clusters [5,6,7], l'approche par un modèle logistique nous permet de compléter chaque niveau de la hiérarchie dans le dendrogramme par un test de l'existence d'une sous-structure en clusters.

Notre méthode est évaluée par des simulations dans différents scénarios caricaturaux de désaccord entre sujets et appliquée à des données issues d'une étude marketing menée au Centre Culinaire Contemporain, plateforme de recherche et d'ingénierie culinaire (Rennes).

## 2 Méthode : classification des sujets via la modélisation de la structure d'accord/désaccord

Une situation d'accord parfait au sein d'un panel de  $J$  sujets évaluant  $I$  stimuli correspond à l'absence d'effet de l'interaction stimulus  $\times$  sujet sur la probabilité  $\pi_{ij}$  que le stimulus  $i$  soit jugé représentatif du concept par le sujet  $j$ . Un modèle possible d'accord parfait est donc le modèle logistique suivant :

$$\text{logit}\pi_{ij} = \mu + \alpha_i + \beta_j, \tag{1}$$

où  $\alpha_i$  représente l'effet du stimulus  $i$  et  $\beta_j$  l'effet du sujet  $j$ . Notons que les paramètres  $(\alpha_i)_{i=1,\dots,I}$  s'interprètent comme des scores de représentativité du concept pour chaque stimulus : on parle de *scores de propension*. De la même manière, les paramètres  $(\beta_j)_{j=1,\dots,J}$  s'interprètent comme des scores caractérisant la tendance de chaque sujet à considérer qu'un stimulus, de manière générale, est représentatif du concept.

L'éventuelle structure de désaccord entre sujets au sein du panel peut être vue comme un écart entre les observations et leurs valeurs ajustées par le modèle d'accord parfait

ci-dessus, cet écart étant mesurable par les résidus de la déviance  $\hat{\varepsilon}_{ij}$ . Ces résidus sont distribués asymptotiquement selon une loi normale d'espérance nulle et d'écart-type 1, ce qui justifie l'utilisation de la distance Euclidienne  $d_{jj'} = \sqrt{\sum_{i=1}^I (\hat{\varepsilon}_{ij} - \hat{\varepsilon}_{ij'})^2}$  pour quantifier le désaccord qu'il existe entre les sujets  $j$  et  $j'$ . Dans la suite, on appelle  $d_{jj'}$  l'*indice de désaccord* entre les sujets  $j$  et  $j'$ , et  $D$  la *matrice de désaccord* de terme générique  $d_{jj'}$ .

Un algorithme de classification ascendante hiérarchique est ensuite appliqué à la matrice de désaccord entre sujets  $D$  obtenue précédemment. Parmi les critères d'agglomération disponibles, le critère de Ward est le seul qui soit basé sur un critère classique de somme de carrés, menant ainsi à l'obtention de clusters minimisant la dispersion intra-clusters à chaque niveau de la fusion [8]. Ce critère d'agglomération est utilisé pour la construction du dendrogramme représentant la structure de désaccord existant au sein du panel de  $J$  sujets.

Le cadre général du modèle logistique permet de construire des tests statistiques de l'homogénéité des évaluations au sein des clusters obtenus. Ainsi, sous l'hypothèse nulle d'un accord de l'ensemble des sujets sur tous les stimuli, les vecteurs  $\hat{\varepsilon}_j = (\hat{\varepsilon}_{1j}, \dots, \hat{\varepsilon}_{Ij})'$  de résidus de la déviance par sujet sont indépendants. Dans le cas contraire d'une structure en deux clusters, une dépendance entre les vecteurs  $\hat{\varepsilon}_j$  est générée par le modèle logistique à effets emboîtés suivant :

$$\text{logit}\pi_{ijk} = \mu + \alpha_i + \beta_{j|k} + \delta_k + (\alpha\delta)_{ik}, \quad (2)$$

où  $\pi_{ijk}$  représente la probabilité que le stimulus  $i$  soit évalué représentatif du concept par le sujet  $j$  du cluster  $k$ ,  $k = 1$  ou  $2$ ,  $\beta_{j|k}$  représente l'effet du sujet  $j$  (l'effet sujet étant emboîté dans l'effet cluster),  $\delta_k$  représente l'effet spécifique du cluster  $k$  et  $(\alpha\delta)_{ik}$  mesure le désaccord entre les clusters de sujets.

Une statistique naturelle de comparaison du modèle nul (1) d'accord parfait avec le modèle (2) de désaccord à deux clusters est la statistique du rapport de vraisemblance. Cependant, la distribution sous l'hypothèse nulle de cette statistique doit tenir compte de la variabilité liée à l'identification des deux clusters. Celle-ci est donc approchée par simulation de données sous l'hypothèse d'accord parfait entre sujets et par calcul, sur chaque tableau de données généré, de la statistique du rapport de vraisemblance obtenue après extraction de deux clusters.

Cette stratégie de test est étendue de manière séquentielle aux différents niveaux de la structure hiérarchique du dendrogramme pour tester l'existence d'une structure en  $K$  clusters contre une structure en  $K + 1$  clusters ( $K \geq 2$ ).

### 3 Évaluation de la méthode

On présente dans un premier temps le résultat de l'application de la méthode présentée ci-dessus dans deux scénarios caricaturaux de désaccord entre sujets, illustrant des

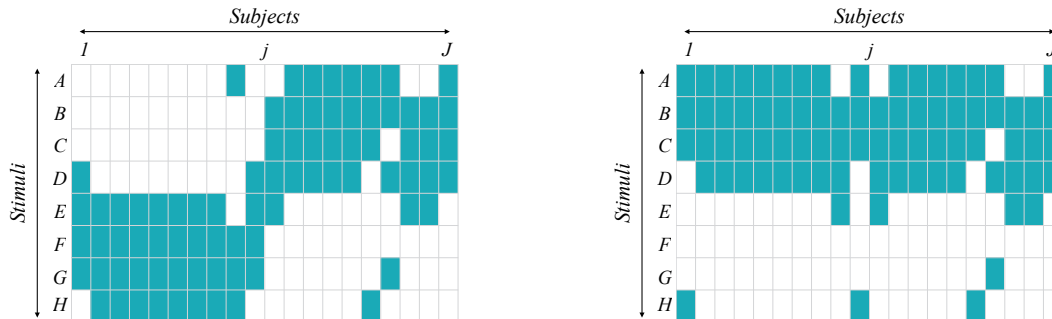


FIGURE 1 – Scénario 1 (à gauche) et scénario 2 (à droite) - une case bleue correspond à une évaluation « représentatif du concept » et une case blanche à une évaluation « non-représentatif du concept »

résultats plus complets basés sur des simulations intensives non-présentés ici.

Deux situations de désaccord impliquant 20 sujets ayant évalué 8 stimuli sont ainsi présentées sur la Figure 1. Un premier scénario fait état d’une structure forte de désaccord entre les sujets, les sujets numérotés de 1 à 10 ayant une perception du concept à travers l’ensemble de stimuli opposée à celle des sujets numérotés de 11 à 20. Le deuxième scénario est, au contraire, emblématique d’une structure forte d’accord entre les sujets.

Les résultats de l’application de la méthode que nous proposons aux deux scénarios sont présentés sous la forme d’un dendrogramme sur la Figure 2. En particulier, les p-values sur lesquelles s’appuient le choix de la structure en clusters semblent conformes aux attentes, en matière de contrôle de l’erreur de type-I (confirmé par les simulations).

Dans le cadre d’une étude réalisée au Centre Culinaire Contemporain lors de laquelle la QC-methodology a été implémentée auprès de 72 sujets pour caractériser le concept de « bon geste » au sein d’un ensemble de 39 stimuli vidéo représentant des gestes culinaires, on montrera la contribution de notre méthode pour explorer la diversité des perceptions de ce concept de « bon geste » au sein du panel.

Une comparaison avec des méthodes de classification basées sur des critères plus classiques d’écart entre distributions multivariées binaires sera présentée. Par ailleurs, la flexibilité de la méthode sera également illustrée, notamment à travers l’introduction de covariables décrivant les sujets ou les stimuli, permettant de décrire les clusters de sujets obtenus.

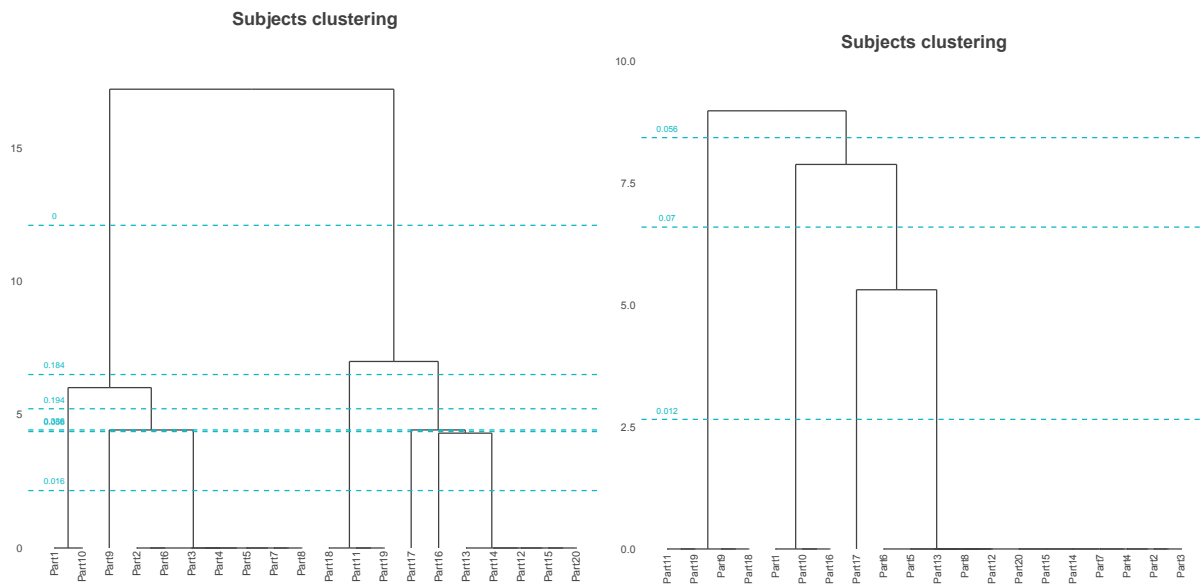


FIGURE 2 – Dendrogrammes obtenus dans le scénario 1 (à gauche) et dans le scénario 2 (à droite)

## Bibliographie

- [1] Brard, M., & Lê, S. (2017). *The QC-methodology, a categorical version of the Q-methodology : Application to the characterization of a complex concept through a set of stimuli*. Soumis le 04/10/2016 dans Food Quality and Preference.
- [2] Fleiss, J., Levin, B., & Paik, M. (2003). *Statistical methods for rates and proportions*. Troisième édition, John Wiley & Sons : Hoboken, New Jersey.
- [3] Warrens, M. (2008). *Similarity coefficients for binary data*. PhD thesis, Leiden University.
- [4] Agresti, A. (2007). *An introduction to categorical data analysis*. Seconde édition, John Wiley & Sons : Hoboken, New Jersey.
- [5] Park, P.J., Manjourides, J., Bonetti, M. & Pagano, M. (2009). *A permutation test for determining significance of clusters with applications to spatial and gene expression data*. Computational Statistics & Data Analysis, 53(12), 4290-4300.
- [6] Bruzzese, D., & Vistocco, D. (2015). *DESPOTA : DEndrogram Slicing through a PermutatiOn Test Approach*. Journal of Classification, 32(2), 285-304.
- [7] Greenacre, M. (2012). *A simple permutation test for clusteredness*. Disponible à l'adresse <http://hdl.handle.net/10230/19856>
- [8] Murtagh, F., & Legendre, P. (2014). *Ward's hierarchical agglomerative clustering method : Which algorithms implement Ward's criterion ?*. Journal of Classification, 31, 274-295.