Analyse et modélisation de la Dominance Temporelle des Sensations à l'aide de processus stochastiques

Guillaume Lecuelle

▶ **To cite this version:**

Guillaume Lecuelle. Analyse et modélisation de la Dominance Temporelle des Sensations à l'aide de processus stochastiques. Informatique. Université Bourgogne Franche-Comté, 2019. Français. NNT : 2019UBFCK031 . tel-02314887

HAL Id: tel-02314887

https://theses.hal.science/tel-02314887v1

Submitted on 14 Oct 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE BOURGOGNE-FRANCHE-COMTE

Centre des Sciences du Goût et de l'Alimentation
(UMR 6265 CNRS - UMR 1324 INRA - UBFC - AgroSup Dijon)

Ecole doctorale Environnements Santé n°554

THESE DE DOCTORAT

Pour obtenir le grade de Docteur de l'Université de Bourgogne-Franche-Comté

Discipline : Sciences de l'Alimentation

Analyse et modélisation de la Dominance Temporelle des Sensations à l'aide de processus stochastiques

Thèse présentée et soutenue à Dijon, le 1er octobre 2019 par

Guillaume LECUELLE

Composition du Jury :

M. BARBU Vlad	Maître de conférences HDR, Université de Rouen	Rapporteur
Mme SAINT-EVE Anne	Maître de conférences HDR, AgroParisTech	Rapporteur
M. POMMERET Denys	Professeur, Université Aix-Marseille	Examinateur
M. QANNARI El Mostafa	Professeur, ONIRIS	Examinateur
M. SCHLICH Pascal	Directeur de recherche, INRA	Directeur de thèse
M. CARDOT Hervé	Professeur, Université Bourgogne Franche-Comté	Codirecteur de thèse
M. VISALLI Michel	Ingénieur, INRA	Co-encadrant de thèse

AVANT-PROPOS

Cette thèse a pu être réalisée grâce à un cofinancement de l'INRA et de la région Bourgogne-Franche-Comté.

“Un pessimiste voit la difficulté dans chaque opportunité,
un optimiste voit l'opportunité dans chaque difficulté.”

Winston Churchill

Résumé

La Dominance Temporelle des Sensations (DTS) est une méthode d'analyse sensorielle qui mesure la perception temporelle d'un produit au cours de sa dégustation. Pour un panéliste, la DTS consiste à choisir parmi une liste de descripteurs lequel est dominant à chaque instant. Ce travail a pour but la modélisation des données DTS à l'aide de processus stochastiques et propose d'utiliser les processus semi-markoviens (PSM), une généralisation des chaînes de Markov qui permet de modéliser librement les durées de dominance. Le modèle obtenu peut être utilisé pour comparer des échantillons DTS en réalisant un rapport de vraisemblance. Étant donné que les probabilités de transition entre les descripteurs peuvent dépendre du temps, nous proposons d'utiliser des modèles différents par période et nous proposons un algorithme pour déterminer le nombre et les frontières de ces périodes de manière optimale. Le modèle est représenté sous forme d'un graphe montrant les transitions entre descripteurs les plus observées. Finalement, ce travail introduit les modèles de mélange de processus semi-markoviens afin de segmenter le panel en fonction des différences de perception interindividuelles.
Les méthodes développées sont appliquées à des jeux de données DTS variés : chocolats, fromages frais et Goudas. Les résultats montrent que la modélisation par un PSM apporte de nouvelles informations sur la perception temporelle, en particulier sur la variabilité de perception au sein d'un panel, alors que les méthodes classiques se focalisent sur une vision moyenne de la perception du panel. De plus, à notre connaissance, ce travail est le premier à proposer l'identification d'un modèle de mélange de processus semi-markoviens.

Mots-clés : *Analyse sensorielle ; Dominance Temporelle des Sensations (DTS) ; Processus semi-markoviens ; Modèles de mélange*

Abstract

Temporal Dominance of Sensations (TDS) is a technique to measure temporal perception of food product during tasting. For a panelist, it consists in choosing in a list of attributes which one is dominant at any time. This work aims to model TDS data with a stochastic process and proposes to use semi-Markov processes (SMP), a generalization of Markov chains which allows dominance durations to be modeled by any type of distribution. The model can then be used to compare TDS samples based on likelihood ratio. Because probabilities of transition from one attribute to another one can also depend on time, we propose to model TDS by period and we propose a method to select optimally the number of periods and the frontiers between periods. Graphs built upon the stochastic pattern can be plotted to represent main chronological transitions between attributes. Finally, this work introduces new statistical models based on finite mixtures of semi-Markov processes in order to derive consumer segmentation based on individual differences in temporal perception of a product.
The methods are applied to various TDS datasets: chocolates, fresh cheeses and Gouda cheeses. Results show that SMP modeling gives new information about temporal perception compared to classical methods. It particularly emphasizes the existence of several perceptions for a same product in a panel, whereas classical methods only provide a mean panel overview. Furthermore, as far as we know, this work is the first one that considers mixtures of semi-Markov processes.

Keywords: *Sensory analysis; Temporal Dominance of Sensations (TDS); Semi-Markov processes; Mixture models*

REMERCIEMENTS

Les premières personnes que je souhaite remercier sont bien sûr Pascal et Hervé pour m'avoir fait confiance et donné l'opportunité de vivre cette formidable aventure.

Merci Pascal pour avoir partagé toute ton expérience du monde de la recherche et pour ton optimisme sans faille. Tu auras su me guider à travers les obstacles de la thèse tout en me perdant, un beau matin, aux alentours de la Villa Clythia !

Hervé, je te remercie pour ta disponibilité et pour m'avoir appris la rigueur essentielle à la recherche en mathématiques. Coécrire un article avec toi aura vraiment été un plaisir.

Un grand merci également à Michel, encadrant au jour le jour, pour toutes nos discussions autour de l'analyse sensorielle qui m'auront éclairé sur ce domaine et m'auront grandement fait avancer dans ma thèse. Merci aussi pour toutes les autres discussions qui n'auront pas fait avancer ma thèse mais l'auront rendu agréable.

Merci à Célestin Kokonendji et El Mostafa Qannari d'avoir constitué mon comité de suivi de thèse et de m'avoir aidé à faire le point sur l'avancement de mes travaux et les suites à leur donner.

Je souhaite remercier Vlad Barbu et Anne Saint-Eve d'avoir accepté de rapporter sur ce travail de thèse. Merci également à Denys Pommeret et El Mostafa Qannari de faire partie de mon jury de thèse.

Je remercie tous les membres de la plateforme Chemosens, actuels et passés, en particulier Caroline et Arnaud qui m'ont transmis leur expérience de thésard. Un remerciement particulier aussi à Christine pour sa bienveillance et sa bonne humeur. Merci à ceux avec qui j'ai partagé des pauses parfois très animées ainsi que ceux que j'ai croisé régulièrement dans les couloirs.

Un grand merci à toute l'équipe du secrétariat du CSGA, et en particulier à Christine Chabert, qui a su résoudre toutes mes difficultés administratives.

Je tiens également à remercier mes anciens collègues de Michelin, et en particulier Benoît Gandar, qui m'ont convaincu de faire une thèse, ainsi que les enseignants du master « Modélisation statistique » de Besançon qui m'y ont, je pense, bien préparé.

Merci à tous mes amis qui m'ont donné le sourire et encouragé tout au long de cette thèse : la Badass Foundation pour être toujours présents depuis maintenant 12 ans (je ne compte plus les fous rires), la team soirées jeux toujours au top et tous ceux avec qui j'ai partagé mes passions que ce soit sur les pistes d'athlétisme ou sur les circuits automobiles !

Merci à l'ensemble de ma famille et belle-famille et en particulier à mes parents qui constituent un socle essentiel à mon équilibre et m'ont toujours

soutenu dans mes choix ainsi qu'à mon frère, grand frère exemplaire qui a su me donner le goût du travail.

Pour finir, je remercie bien sûr Madeleine, première contributrice à mon bonheur, qui a accepté de me suivre et de s'éloigner (un peu) de notre magnifique Jura natal !

VALORISATION DES TRAVAUX DE THESE

Publications

- Lecuelle, G., Visalli, M., Cardot, H., & Schlich, P. (2018). Modeling Temporal Dominance of Sensations with semi-Markov chains. *Food Quality and Preference, 67, 59-66.*

- Cardot, H., Lecuelle, G., Schlich, P., & Visalli, M. (2019). Estimating finite mixtures of semi-Markov chains: an application to the segmentation of temporal sensory data. *Journal of the Royal Statistical Society Series C–Applied statistics. (In Press)*

Communications orales

- Lecuelle, G., Visalli, M., Cardot, H., & Schlich, P. (2016). Modeling Temporal Dominance of Sensations data with semi-Markov chains. *In Sensometrics, Brighton, July 26-29th.*

- Lecuelle, G., Visalli, M., Cardot, H., & Schlich, P. (2016). A method based on semi-Markov chains for segmenting a consumer TDS panel. *In Eurosense, Dijon, September 11-14th.*

- Lecuelle, G., Visalli, M., Cardot, H., & Schlich, P. (2017). Une approche stochastique pour la modélisation de données temporelles sensorielles. *Aux Rencontres des Jeunes Statisticiens, Porquerolles, du lundi 3 au vendredi 7 avril.*

- Lecuelle, G., Visalli, M., Cardot, H., & Schlich, P. (2017). Une approche stochastique pour la modélisation de données temporelles sensorielles. *Au Forum des Jeunes Chercheurs, Dijon, le 15 juin.*

- Lecuelle, G., Visalli, M., Cardot, H., & Schlich, P. (2017). Modélisation stochastique : la DTS se déchaîne ! *A la journée des doctorants, Centre des Sciences du Goût et de l'alimentation, Dijon, le 22 juin.*

- Lecuelle, G., Visalli, M., Cardot, H., & Schlich, P. (2018). Modeling Temporal Dominance of Sensations data with semi-Markov chains. *In AgroStat, Marseille, March 14-16th*.

- Lecuelle, G., Visalli, M., Cardot, H., & Schlich, P. (2018). Modeling TDS data and segmenting consumers thanks to a mixture of semi-Markov processes. *In Sensometrics, Montevideo, April 9-12th*.

- Lecuelle, G., Visalli, M., Cardot, H., & Schlich, P. (2018). Understanding consumer segmentation in product perception thanks to semi-Markov chains modeling of TDS data. *In Eurosense, Verona, September 2-5th.*

- Cardot, H., Lecuelle, G., Visalli, M., & Schlich, P. (2018). Estimating finite mixtures of semi-Markov chains: an application to the segmentation of temporal sensory data. *In Stodep workshop, Saint-Etienne-du-Rouvray, October 3-5h.*

- Frascolla, C., Lecuelle, G., Visalli, M., Cardot, H., & Schlich, P. (2019). Comparison of qualitative trajectories with semi-Markovian chains: an application in sensory analysis. *Aux 51èmes Journées de Statistique, Vandoeuvre-lès-Nancy, du 3 au 7 juin.*

Communication affichée

- Lecuelle, G., Visalli, M., Cardot, H., & Schlich, P. (2017). Analysis of TDS data with a stochastic approach. *In Pangborn, Providence, USA, August 20-24*.

Encadrement de stagiaires

- Rihab Boubakri (2017)
  Stage de fin d'études en vue de l'obtention du Diplôme National d'Ingénieur en Statistique et Analyse de l'Information. Ecole Supérieure de la Statistique et de l'Analyse de l'Information, Université de Carthage.
  Développement statistique en analyse sensorielle : Application des chaînes de Markov à la Dominance Temporelle des Sensations.

- Cindy Frascolla (2018)
  Stage M2 Master Modélisation Statistique.
  Laboratoire de mathématiques de Besançon, Université de Franche-Comté.
  Etude de la performance des panels en analyse sensorielle temporelle.

TABLE DES MATIERES

Préambule

L'analyse sensorielle propose un ensemble de méthodes permettant d'objectiver les sensations perçues par des êtres humains. Les avancées dans le domaine peuvent provenir de nouvelles méthodes de mesure mais aussi de nouvelles méthodes d'analyse statistique des données. Recueillir des jeux de données coûte cher, aussi leur exploitation se doit d'être optimale.

Dans ce cadre, cette thèse, qui a eu lieu au sein de la plateforme Chemosens du Centre des Sciences du Goût et de l'Alimentation (CSGA) de Dijon sous la direction de Pascal Schlich (INRA) et Hervé Cardot (Institut de Mathématiques de Bourgogne) et l'encadrement de Michel Visalli (INRA), propose une nouvelle approche pour la modélisation de données sensorielles temporelles et se situe au croisement entre analyse sensorielle et statistique. Cette thèse en sciences de l'alimentation a été réalisée dans l'école doctorale « Environnements – Santé » (E-S) de l'Université Bourgogne Franche-Comté (UBFC). Même si des concepts mathématiques sont présentés, son contenu est destiné à être compréhensible par les professionnels de l'analyse sensorielle.

Ce travail de thèse se focalise sur la Dominance Temporelle des Sensations (DTS). Cette méthode, développée par ChemoSens, est certainement aujourd'hui la méthode de référence pour l'analyse sensorielle temporelle. Bien que largement utilisée, de nombreuses questions subsistent autour de la compréhension de la perception temporelle du consommateur. L'objectif de cette thèse est donc de fournir des outils statistiques pertinents, adaptés à la complexité des données et prenant en compte les différences de perception pouvant exister parmi les consommateurs. Ces outils s'adressent donc avant tout aux professionnels de l'analyse sensorielle, qui pourront les utiliser pour mieux comprendre ou formuler leurs produits en fonction des cibles visées pour un produit donné. Ce travail de thèse ouvre également de nouvelles opportunités en recherche fondamentale pour mieux comprendre les mécanismes impliqués dans la perception. Enfin, il propose un nouvel outil statistique d'estimation dans des modèles complexes qui pourrait s'appliquer à d'autres situations que l'analyse sensorielle .

S’organisant en quatre grands chapitres, ce manuscrit présente dans une première partie l’analyse sensorielle et principalement la méthode de la Dominance Temporelle des Sensations (DTS), puis introduit les différents outils mathématiques sur lesquels les travaux présentés sont fondés. Dans le deuxième chapitre, la théorie et les algorithmes mis en œuvre pour la modélisation des données DTS sont décrits en se basant sur deux articles publiés et fournis en annexe. Dans le troisième chapitre, plusieurs applications permettent de comprendre les avantages de l’approche de modélisation proposée par rapport aux outils d’analyse classiques. Enfin, la méthodologie proposée ainsi que les résultats obtenus sont discutés dans le quatrième chapitre.

Chapitre 1.

Introduction

1.1 Analyse sensorielle

Définition

La norme française NF ISO 5492 définit l'analyse sensorielle comme « l'examen des propriétés organoleptiques d'un produit par les organes des sens ». L'humain est ainsi l'instrument de mesure et va, à l'aide de ces cinq sens, apporter une information sur un produit complémentaire à une mesure physico-chimique.

La mesure sensorielle est réalisée lors d'une étude par un panel de personnes convoquées par un « animateur de panel ». Après avoir présenté le protocole de l'étude, ce dernier fait goûter un ou plusieurs produits à chaque panéliste qui doit les évaluer selon un ou plusieurs critères définis dans le protocole.

Historique

L'analyse sensorielle est une discipline plutôt récente qui intervient lors de la phase de conception ou d'amélioration de produit. Elle a été développée dans les années 50 par l'armée américaine afin d'améliorer la qualité sensorielle des rations militaires (Peryam, 1954).

Cette approche a ensuite été utilisée par les industriels de l'agro-alimentaire pour faire face aux exigences de qualité apparues lors de la crise économique des années 70 où il ne suffisait plus de produire pour vendre. Depuis cette période, l'analyse sensorielle connait un essor académique et industriel avec des congrès (Pangborn et Eurosense) regroupant plusieurs centaines de personnes et un nombre croissant d'articles de recherche publiés sur ce sujet (voir Figure 1).

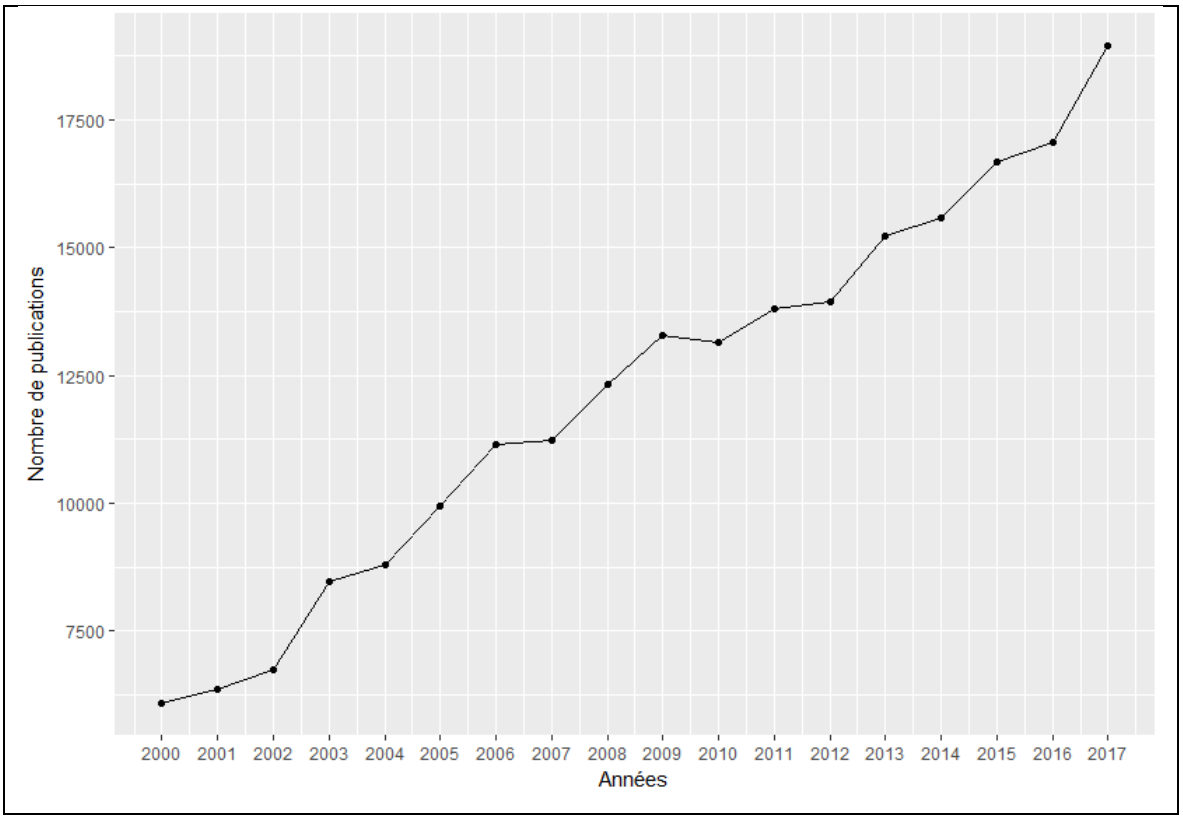


Figure 1 Nombre de résultats par année depuis 2000 pour la recherche des termes « Sensory Analysis » dans Science Direct.

Applications

L'analyse sensorielle est devenue un outil indispensable pour de nombreux industriels afin de répondre à des objectifs variés : développement de produits appréciés par le consommateur, modification de produits pour, par exemple, diminuer le coût de production ou limiter les teneurs en gras, en sucre ou en sel sans altérer la perception, contrôle qualité des matières premières, évolution du produit lors du stockage, etc. Le marketing fait aussi appel à l'analyse sensorielle pour obtenir une meilleure compréhension du marché et ainsi adapter la commercialisation et la communication aux spécificités du produit.

Développée pour les secteurs de la cosmétique et de l'agro-alimentaire, les applications de l'analyse sensorielle se sont depuis étendues à ceux de l'automobile, l'industrie textile, les matériels électroniques, l'ameublement, l'alimentation des animaux de compagnie, les matériaux, la publicité, le tabac, etc.

1.1.1 L'être humain comme outil de mesure

Actuellement, aucune machine ne peut encore remplacer le dégustateur du fait de la forte complexité des mécanismes sous-tendant la perception sensorielle humaine.

L'analyse sensorielle fait donc appel à un panel pour tester un ou plusieurs produits. Historiquement, le panel était composé d'experts entraînés et qui faisaient en général partie de l'entreprise. Néanmoins, ces dernières années l'intérêt pour les tests avec des consommateurs est grandissant, car ces tests permettent d'avoir une image plus réaliste du ressenti du client. Cet intérêt est également justifié pour des raisons de coût, aussi bien en terme de temps que d'argent, puisque la formation d'un panel d'experts requiert en général un long entraînement pour fournir des résultats répétables.

La perception se décompose en trois étapes : la stimulation des récepteurs sensoriels, la traduction de l'information physique ou chimique en un signal électrique par les récepteurs, la transmission au cerveau qui interprète ce signal en faisant appel à la mémoire et en déduit une perception.

Plusieurs biais à chacune des étapes expliquent que tout le monde n'ait pas la même perception d'un produit (Meilgaard, Civille, & Carr, 1999). Cette variabilité justifie de faire appel à un panel qui serait inutile si chaque individu avait exactement la même perception !

La perception peut être influencée par plusieurs facteurs. La mastication et la salive ont un rôle important et complexe qui peut en partie expliquer la variété de perception (Blissett, Hort, & Taylor, 2006). Les différences physiologiques telles que le nombre de capteurs sensoriels ou la taille de la cavité buccale jouent également un rôle important. Plusieurs facteurs psychologiques influencent aussi la perception. Par exemple, la perception d'un vin est fortement modifiée selon les informations données sur l'étiquette de la bouteille. Le manque de motivation ou une lassitude à la tâche peut aussi dégrader la qualité des résultats.

Il a été également montré que les fumeurs sont moins sensibles à certaines saveurs et aux odeurs (Ahlström, Berglund, Berglund, Engen, & Lindvall, 1987; Krut, Brontestewart, & Perrin, 1961) ; d'ailleurs, les fumeurs sont parfois exclus des études sensorielles, ou il leur est demandé de ne pas fumer dans l'heure précédent l'étude.

1.1.2 Méthodes statiques

L'analyse sensorielle comprend principalement trois types de méthodes permettant respectivement de savoir si des produits sont perçus différemment ou non, de décrire et quantifier ces différences et de mesurer l'appréciation d'un produit (Meilgaard et al., 1999).

1.1.2.1 Méthodes discriminatives

Les méthodes discriminatives consistent à évaluer l'existence ou non de différences entre produits. Ces méthodes très populaires sont généralement utilisées pour évaluer des produits avec des différences peu perceptibles. Dans l'industrie agroalimentaire ces méthodes permettent de tester si un changement de recette, tel qu'une diminution de la teneur en sucre ou en sel, a un impact sur la perception du produit. Plusieurs méthodes existent mais l'idée reste la même : après dégustation, les panélistes doivent identifier quels sont les produits identiques et quels produits sont différents. Les méthodes les plus couramment utilisées sont : le test triangulaire, le test de comparaison par paire, le test duo-trio, le test A/Non-A, l'épreuve p parmi n. Ces méthodes simples permettent de savoir si deux produits sont perçus différemment mais ne donnent aucune information sur ces différences.

1.1.2.2 Méthodes descriptives

Les méthodes descriptives visent à établir le profil sensoriel de chaque produit afin de pouvoir les comparer. Le profil conventionnel est la méthode de référence et prend son origine dans trois méthodes : Flavor Profile (Cairncross & Sjöström, 1950), Quantitative Descriptive Analysis (Stone, Sidel, Oliver, Woolsey, & Singleton, 1974) et Spectrum (Muñoz & Civille, 1992, 1998). Le profil d'un produit est construit en commençant par lister l'ensemble des descripteurs pouvant être associés au produit (sauf les termes

hédoniques), soit en utilisant la littérature, soit en faisant appel au panel avec par exemple la méthode Check All That Apply (Coombs, 1964). L'intensité de la perception de chaque descripteur est alors évaluée puis les valeurs obtenues pour l'ensemble du panel sont représentées graphiquement. Cette méthode requiert un entraînement pour que le panel soit capable de bien identifier les descripteurs et d'avoir une utilisation optimale de l'échelle de mesure de l'intensité.

1.1.2.3 Méthodes hédoniques

Les méthodes hédoniques permettent l'évaluation des préférences des consommateurs. Le panel est généralement constitué de consommateurs naïfs, c'est-à-dire n'ayant eu aucune pratique de l'analyse sensorielle, qui ne sont pas entraînés. Deux approches existent, la première consiste pour le panéliste à choisir le produit qu'il préfère ou à ordonner par ordre de préférence un certain nombre de produits et la deuxième consiste à donner une note d'appréciation hédonique, généralement sur une échelle de 1 à 9 (Jones, Peryam, & Thurstone, 1955), à chaque produit.

1.1.3 Méthodes temporelles

La perception d'un produit est un processus temporel. Cette assertion, qui peut paraître évidente, a été mise en évidence il y a une soixantaine d'année (Neilson, 1957). Prenons par exemple un chocolat contenant des grains de sel. Le dégustateur pourra successivement percevoir le côté croquant du chocolat, le cacao, le sel, l'amertume avant que le chocolat devienne fondant. De nombreux produits tels que le vin, les chewing-gums, les glaces ou même l'eau (Hort, Kemp, & Hollowood, 2017) ont des propriétés sensorielles qui évoluent au cours de la durée d'une prise de ces produits.

La prise de conscience de l'existence de cette temporalité et de l'intérêt de mesurer sa dynamique a conduit au développement de plusieurs méthodes dites temporelles. Dans un premier temps ces mesures ont été réalisées à temps discret, c'est-à-dire que la mesure a lieu à plusieurs instants de la dégustation. Ensuite, l'apparition d'outils informatiques a permis une mesure en continu réalisée pendant que les éléments sont perçus par le panéliste.

Le 7ème congrès *European Conference on Sensory and Consumer Research* (Dijon, France) a largement démontré l'importance prise par la temporalité en analyse sensorielle aussi bien pour le milieu industriel que le milieu académique avec de nombreuses présentations et posters portant sur cette thématique. Cette tendance a d'ailleurs été présente jusque dans le nom du congrès : « A sense of time ».

1.1.3.1 Temps Intensité

La méthode Temps-Intensité (Time-Intensity en anglais) (Lee & Pangborn, 1986) est une mesure de l'intensité de la perception d'un descripteur au cours du temps. Historiquement, cette méthode a connu ses prémices entre les années 30 et 60 (Holway & Hurvich, 1937; Jellinek, 1964; Sjöström, 1954) en demandant au panéliste de tracer une courbe correspondant à l'intensité des sensations perçues ou en notant l'intensité d'un descripteur toutes les secondes. Cependant, à cause des difficultés liées à l'acquisition des données, cette méthode n'a pris son essor que dans les années 80 grâce au développement de l'informatique. Pour le panéliste, la méthode consiste à déplacer un curseur sur une échelle d'intensité continue pendant une durée donnée. Cette méthode est la plus précise en termes de description sensorielle temporelle mais nécessite un long entraînement pour que les panélistes maîtrisent parfaitement le concept et pour limiter la variabilité interindividuelle. Cette nécessité d'entraînement qui se traduit par de nombreuses séances ainsi que le fait de n'évaluer qu'un descripteur par dégustation font que cette méthode est coûteuse aussi bien en termes d'argent que de temps.

1.1.3.2 Les premières méthodes multi-descripteurs

Les limites de la méthode Temps-Intensité ont conduit au développement de nouvelles méthodes visant à évaluer simultanément plusieurs descripteurs. La première méthode proposée a été le Dual-Attribute Time-Intensity (Duizer, Bloom, & Findlay, 1997). Cette méthode consistait à évaluer l'intensité de non plus un seul descripteur sur une échelle mais deux descripteurs simultanément en déplaçant un curseur dans un plan dont les deux dimensions correspondaient aux deux descripteurs. Cette méthode n'a, selon

la littérature, été utilisée qu'une fois pour une étude sur la tendreté et la jutosité de viandes de bœuf (Zimoch & Gullett, 1997). Cette méthode, limitée à deux descripteurs, est très difficile à réaliser pour le sujet et requiert ainsi un temps d'entraînement très long.

Le Temps-Intensité est aussi utilisé de façon discrète sous le nom de Temps-Intensité Discontinu (Clark & Lawless, 1994) et permet l'étude de plusieurs descripteurs en une seule dégustation en notant successivement les différents descripteurs.

La mesure de l'intensité est une tâche complexe et le développement de l'évaluation simultanée de tous les descripteurs s'est fait en se détachant de cette notion.

1.1.3.3 Dominance Temporelle des Sensations

La Dominance Temporelle des Sensations est une méthode apparue dans les années 2000 (Pineau, Cordelle, & Schlich, 2003; Pineau et al., 2009) qui offre la possibilité d'évaluer simultanément plusieurs descripteurs pendant la durée de la dégustation. Le panéliste est placé devant un écran sur lequel est affichée une liste d'une dizaine de descripteurs et va au cours de la dégustation sélectionner quel est selon lui le descripteur dominant à chaque instant (Figure 2). Bien qu'initialement il était demandé aux panélistes de donner une valeur d'intensité à chaque choix de descripteur dominant, cette idée a été progressivement abandonnée, permettant une évaluation simplifiée basée uniquement sur la dominance (Schlich, 2017). Ainsi, cette approche est devenue simple et ne nécessite pas ou très peu d'entraînement (Albert, Salvador, Schlich, & Fiszman, 2012) ce qui permet l'évaluation d'un produit en une seule séance, réduisant considérablement les coûts.

Figure 2 Écran DTS pour la dégustation de chocolats avec 10 descripteurs. Le panéliste doit cliquer sur « démarrer » au moment de la mise en bouche puis sélectionner consécutivement les descripteurs qui attirent le plus son attention avant de finalement cliquer sur « je ne perçois plus rien ».

Le concept de dominance est longtemps resté flou et nécessite encore d'être précisé, même si une définition relativement consensuelle s'est dégagée, décrivant un descripteur dominant comme étant « celui qui attire le plus l'attention à un moment donné ».

Le nombre minimum de sujets nécessaire pour réaliser une étude DTS dépend évidemment du type de produit évalué mais Pineau et al. (2012) suggèrent de faire appel à un panel composé d'au moins 16 sujets avec au moins une répétition c'est-à-dire deux dégustations du même produit. Ils suggèrent également de ne pas utiliser plus de 10 descripteurs.

Le choix de la liste de descripteurs utilisés est extrêmement important et peut se faire sur la base de connaissances antérieures, ou par des techniques classiques de génération de termes en profil sensoriel menées soit avec un petit groupe de sujets qui pourront ou pas faire partie du panel convoqué pour l'étude.

1.1.3.4 TCATA

La méthode Temporal Check All That Apply (Castura, Antunez, Gimenez, & Ares, 2016) consiste à sélectionner tous les descripteurs perçus à chaque instant. Le panéliste doit cocher et décocher les descripteurs au cours du temps selon l'apparition ou la disparition de leur perception des sensations associées à ces descripteurs. Cette méthode a été conçue afin de répondre à une limite de la DTS qui ne permet de citer qu'un seul descripteur à chaque instant même si plusieurs sensations sont présentes. La méthode TCATA est plus compliquée à utiliser que la DTS puisque décocher ce qui n'est plus ressenti se révèle assez compliqué, mais cette limite semble pouvoir être dépassée en faisant en sorte que les descripteurs sélectionnés soient automatiquement désélectionnés après un certain temps obligeant le panéliste à cliquer à nouveau sur ce descripteur s'il est encore présent (Ares et al., 2016).

1.1.3.5 Méthodologie temporelle hédonique

L'appréciation hédonique d'un produit, communément appelé « liking », peut facilement être mesurée par une note sur une échelle. Cette mesure est très utilisée, en revanche mesurer le liking de façon temporelle est une approche peu répandue. Le liking d'un produit a pourtant été reconnu comme pouvant évoluer au cours de la dégustation depuis les années 80 (Rozin, Ebert, & Schull, 1982) et peut être mesuré de façon continue comme un descripteur avec la méthode temps intensité (Lee & Pangborn, 1986; Taylor & Pangborn, 1990). Plusieurs études ont cherché à lier le liking temporel à la perception temporelle notamment avec les intensités de perception (Veldhuizen, Wuister, & Kroeze, 2006) et plus récemment avec les données DTS (Thomas et al., 2017; Thomas, Visalli, Cordelle, & Schlich, 2015).

1.1.4 Focus sur l'analyse des données de Dominance Temporelle des Sensations

Les données DTS consistent en l'enregistrement des temps des différents clics sur les descripteurs dominants.

A l'échelle du panel, les données DTS rassemblent l'ensemble des séquences, chaque séquence correspondant à une dégustation d'un produit par un panéliste (Figure 3). Plusieurs méthodes ont été proposées pour analyser ces données.

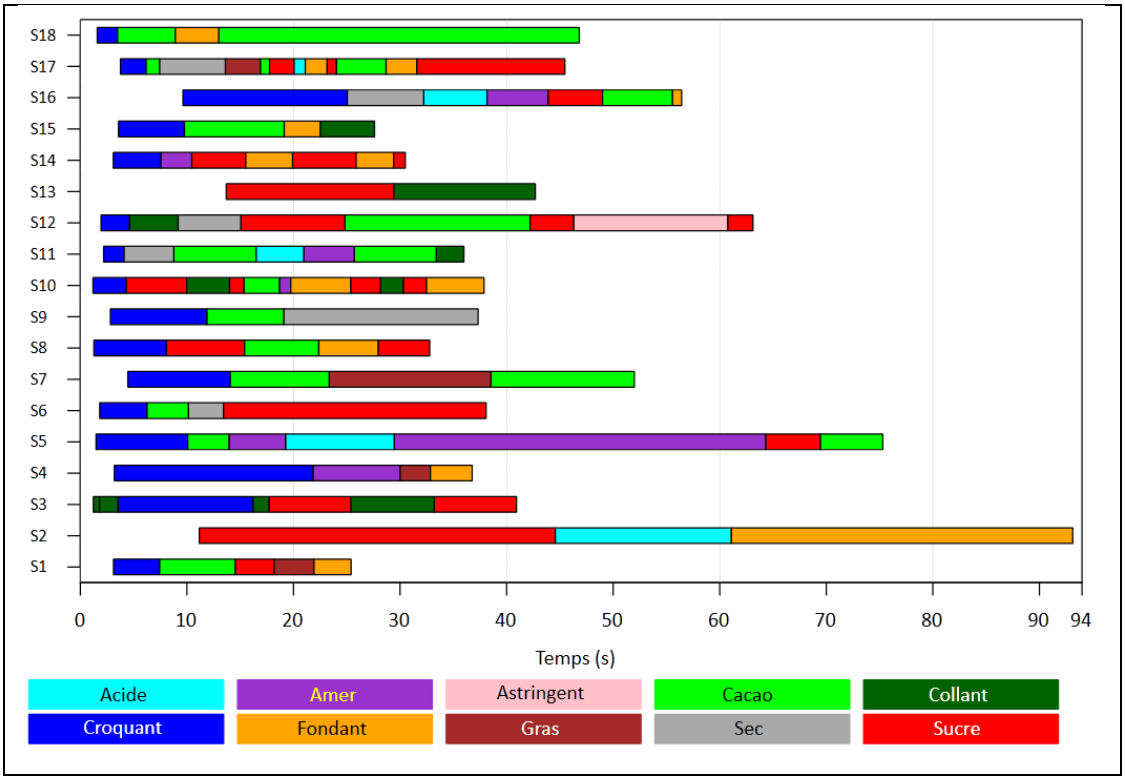


Figure 3 Bandplot d'un chocolat à 70% de cacao. Chaque bande représente une répétition de la dégustation de ce chocolat par un panéliste. Les panélistes sont ici codés de S1 à S18. Par exemple, le panéliste S1 a perçu dans l'ordre les descripteurs Croquant, Cacao, Sucre, Gras et Fondant.

Le comportement individuel des juges (temps de premier clic, nombre de descripteurs utilisés, nombre de clics, durée de dégustation) peut être étudié à l'aide de statistiques descriptives ou à l'aide de méthodes classiques telles que l'ANOVA. Il est ainsi possible de détecter des données aberrantes mais également d'obtenir des premières informations sur la temporalité du produit.

La méthode d'analyse la plus utilisée est la représentation graphique des données DTS par des courbes, appelées courbes DTS (Pineau et al., 2009), représentant le pourcentage de panélistes ayant choisi chaque descripteur à chaque instant (Figure 4). Pour construire ces courbes, le temps est discrétisé puis les taux de dominance sont calculés à chaque instant et une courbe, lissée en utilisant une base de splines, est tracée pour chaque descripteur.

Ces courbes sont complétées par deux niveaux : un niveau de hasard, égal à l'inverse du nombre de descripteurs, correspondant à la valeur observée si les panélistes choisissaient les descripteurs de manière aléatoire ; un niveau de significativité déterminé par un modèle binomial au-dessus duquel le choix de descripteur est considéré comme significativement supérieur au niveau de hasard avec un risque α généralement fixé à 10%.

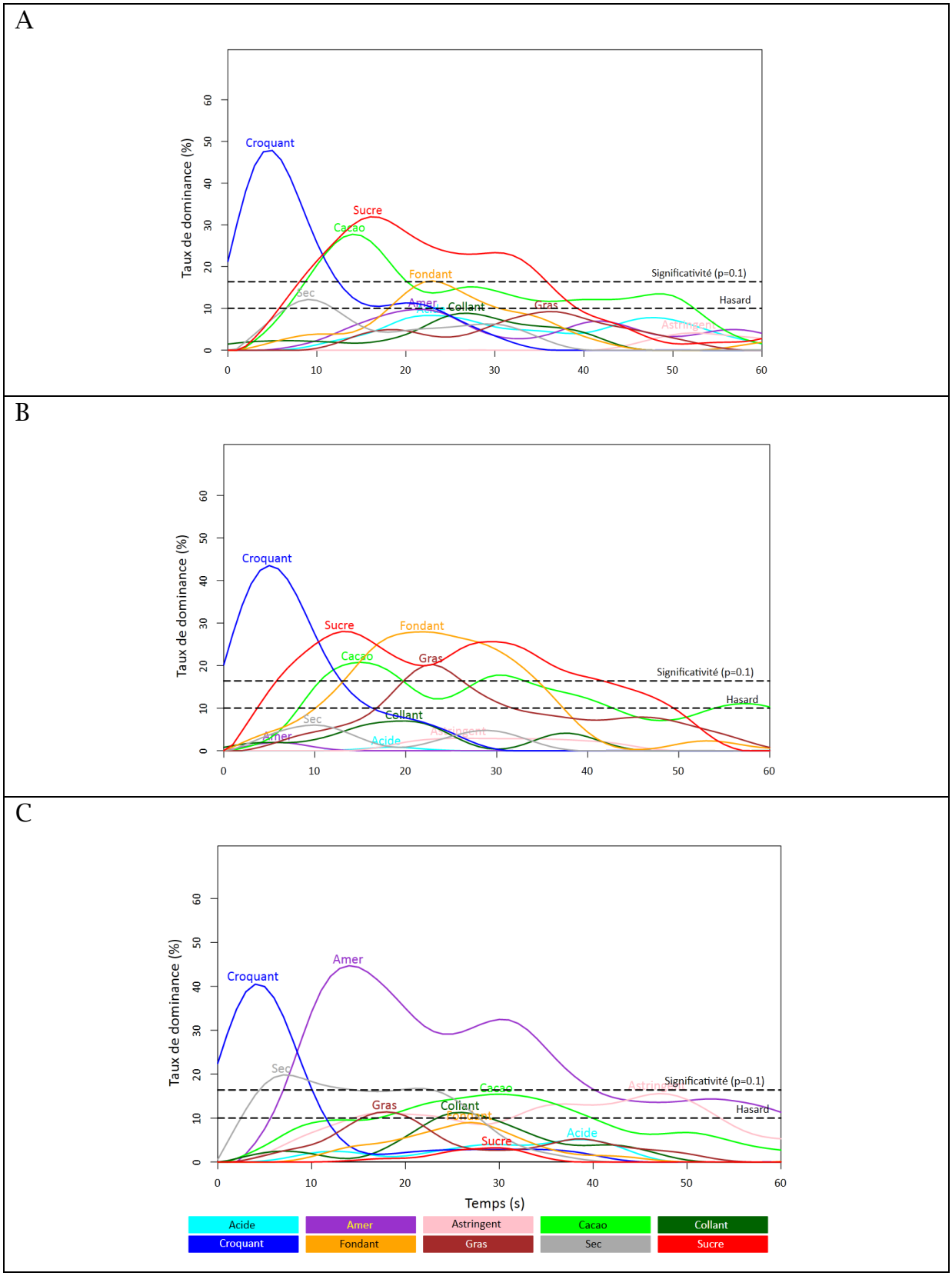


Figure 4 Courbes DTS pour 3 chocolats : A 70% de cacao, B 70% de cacao aussi mais plus sucré, C 90% de cacao. Chaque courbe représente le pourcentage de panélistes ayant choisi le descripteur correspondant à chaque instant. Pour les trois produits Croquant est significativement choisi comme dominant durant les 10 premières secondes. Ensuite les descripteurs Cacao et Sucre sont dominants pour les deux chocolats à 70% alors que le chocolat à 90% est perçu comme sec et amer. Les descripteurs Gras entre 20 et 25 secondes et Fondant entre 15 et 35 secondes ont également été significativement choisis pour le chocolat 70% doux.

Des différences existent généralement au sein d'un panel pour le temps avant le premier clic ainsi que pour la durée totale comme on le voit dans la Figure 3. Pour améliorer la représentation des données DTS, il est alors courant de standardiser les données en supprimant le temps avant le premier clic et en divisant la durée de chaque séquence par sa durée totale.

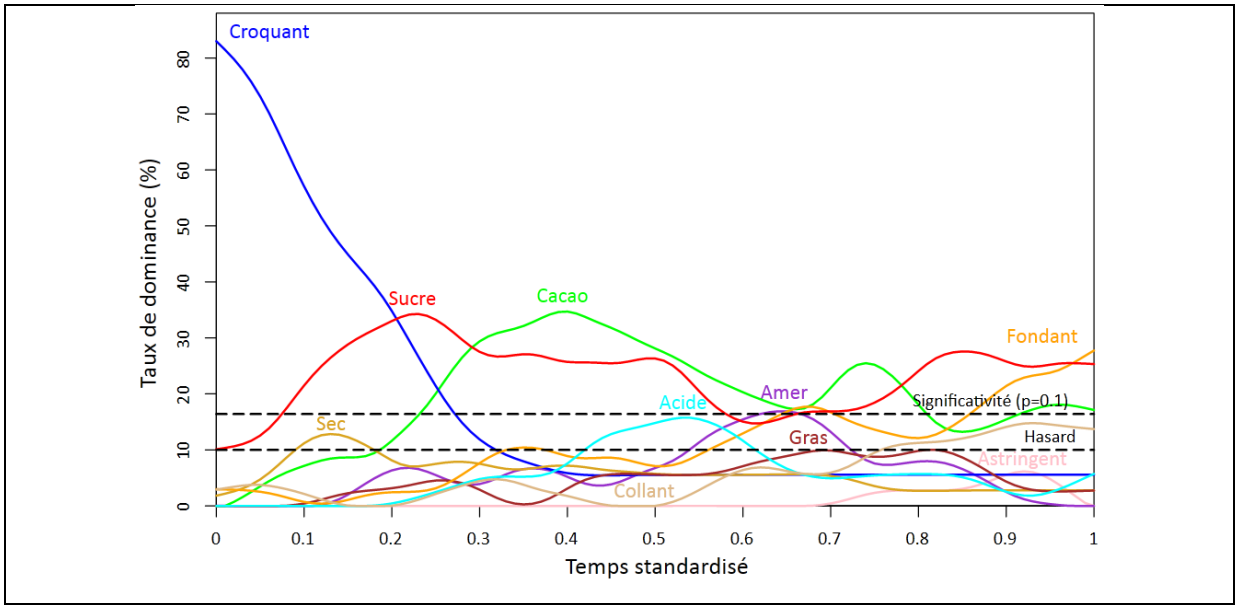


Figure 5 Courbes DTS standardisées pour un chocolat avec 70% de cacao. Ce chocolat est principalement perçu Croquant puis Sucré, Cacao et finalement Fondant.

Les courbes DTS réalisées avec les données standardisées proposent généralement une vision plus claire des résultats. Par exemple, pour le chocolat à 70%, les courbes standardisées (Figure 5) montrent clairement que Fondant est important à la fin de la dégustation ce qui était peu évident dans les courbes non-standardisées (Figure 4 A).

La comparaison des produits peut se faire de plusieurs manières. D'abord à l'aide de représentations graphiques en utilisant les courbes de différences entre deux produits (Figure 6). Les taux de dominance des deux produits comparés sont soustraits à chaque instant et seules les différences significativement non nulles (p=0.10) sont affichées.

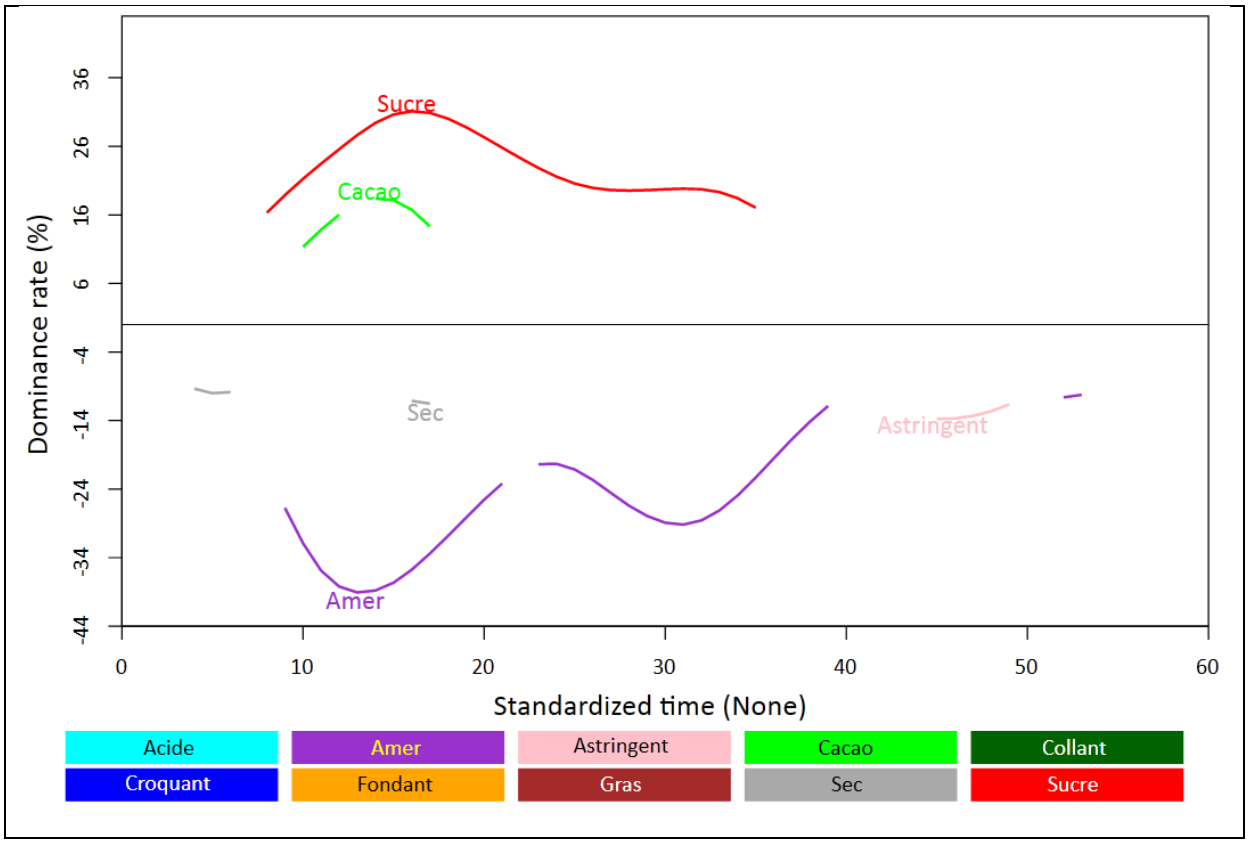


Figure 6 Courbes de différences entre un chocolat 70% et un chocolat 90%. Le chocolat à 90 % de cacao est perçu plus amer et moins sucré que le chocolat à 70% durant la période allant de 5 secondes après la mise en bouche à environ 40 secondes. Il est aussi perçu plus astringent entre 40 et 50 secondes alors que le cacao est plus choisi comme dominant pour le chocolat à 70% entre 10 et 15 secondes.

Une autre représentation graphique des données DTS appelée TDS-Band plot a été proposée (Monterymard, Visalli, & Schlich, 2010) où seul les descripteurs ayant un taux de dominance significatif sont affichés afin de pouvoir comparer visuellement tous les produits de l'étude à l'aide d'une seule représentation (Figure 7). Toutefois, celle-ci ne permettant plus de voir les importances relatives des taux de dominance, Galmarini et al. (2017) ont proposé une nouvelle représentation en bandplots comportant cette information.

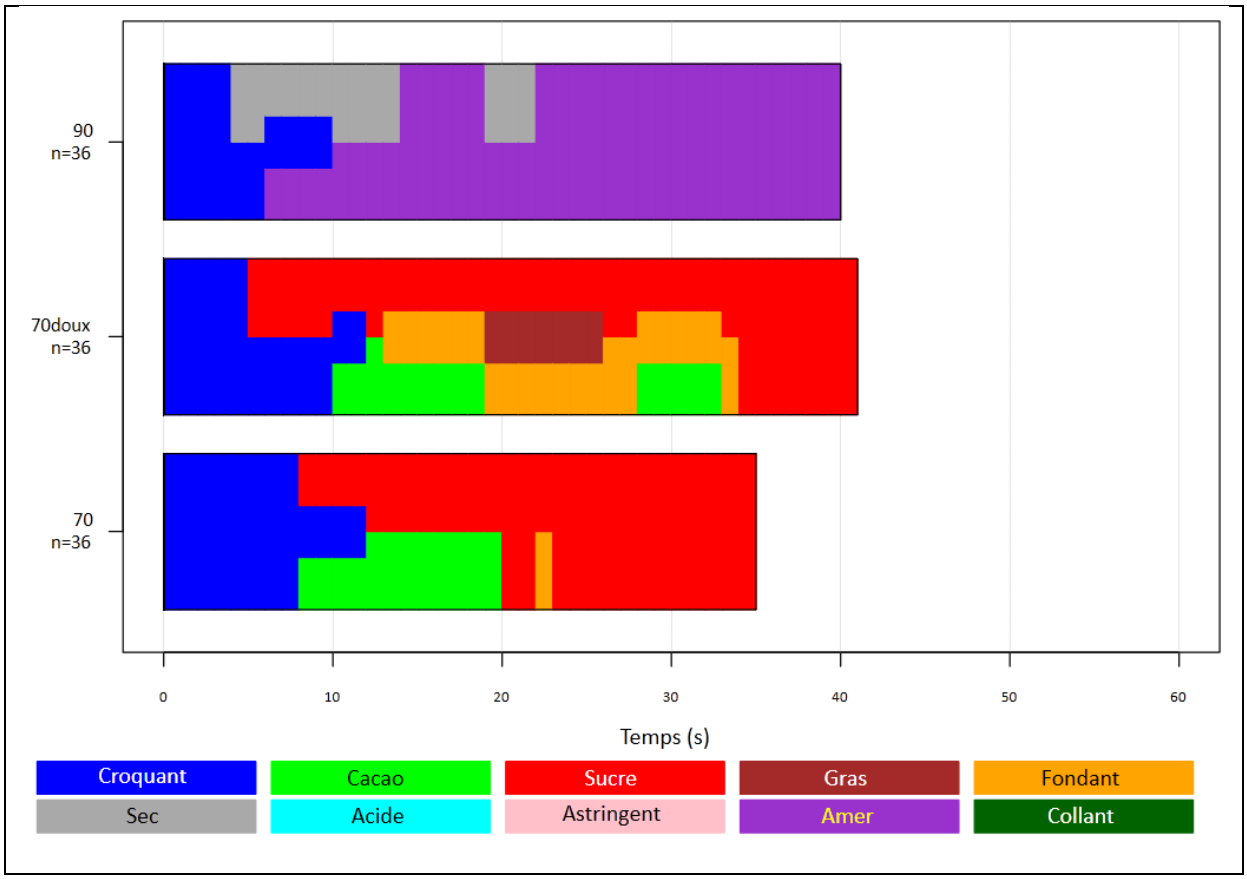


Figure 7 TDS Band-plots de chocolats 70%, 70% doux et 90%. Les trois chocolats sont d'abord perçus comme croquant durant 5 à 10 secondes. Ensuite le chocolat à 90% est perçu sec et amer alors que pour les deux autres chocolats les panélistes ont perçu le sucre et le cacao. Le chocolat à 70% doux est également perçu comme étant gras entre 20 et 25 secondes et fondant entre 15 et 35 secondes.

Ces approches graphiques sont simples à mettre en place mais ne fournissent qu'une information restreinte et pas de validation statistique de l'existence de différences entre produits basée sur l'ensemble des informations contenues dans les données DTS. Une autre approche proposée par Meyners & Pineau (2010) consiste à tester statistiquement à l'aide de tests de randomisations si deux produits sont différents à chaque instant discrétisé et pour chaque descripteur. Cette méthode est assez complexe à mettre en place notamment quant au choix des permutations à réaliser et est coûteuse en temps de calcul. Elle ne semble pas avoir été utilisée par d'autres auteurs.

Une autre approche pour analyser les données DTS consiste à calculer les durées de dominance, c'est-à-dire la durée pendant laquelle le descripteur a été dominant dans chaque séquence DTS (la durée étant de 0 si le descripteur n'est pas cité).

A

	FProd	70	70doux	90	FSubj	HSD	FProdSubj	RMSE
Amer	**24.23*** **	2.90(a)	0.14(a)	15.14(b)	1.75.	3.25	**2.89*** **	5.72
Sucre	**24.06*** **	9.10(b)	10.74(b)	0.41(a)	**2.66** **	2.34	**2.71*** **	4.13
Fondant	**9.64*** **	4.02(a)	6.95(b)	1.81(a)	**3.77*** **	2.23	1.61.	3.93
Astringent	**7.08** **	0.59(a)	0.70(a)	6.97(b)	0.95	1.75	**7.16*** **	3.08
Sec	**6.91** **	2.21(a)	1.14(a)	4.71(b)	**2.07** *	2.15	1.22	3.79
Acide	**3.54** *	2.74(b)	0.05(a)	1.14(ab)	1.03	2.15	1.31	3.78
Gras	2.85.	2.12	5.05	2.26	**2.44** *	1.87	**3.20*** **	3.28
Cacao	2.14	8.02	8.09	5.47	**9.64*** **	3.57	0.95	6.28
Croquant	2.13	6.23	5.53	4.29	**3.86*** **	1.78	**1.67** *	3.13
Collant	0.91	1.83	1.43	2.59	1.16	1.64	**1.66** *	2.88

Modèle: Sujet + Produit + Interaction (Sujet aléatoire).

Les moyennes des produits identifiées par les mêmes lettres ne sont pas significativement différentes.

(.) Significatif à 10%, (*) Significatif à 5%, (**) Significatif à 1%, (***) Significatif à 0.1%.

B

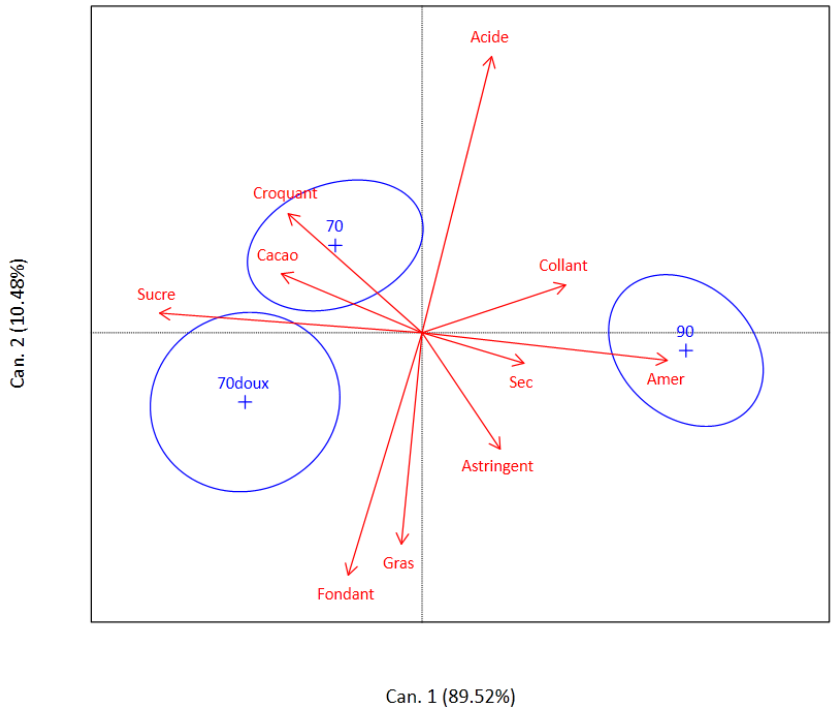


Figure 8 Analyse des durées de dominance à l'aide d'une ANOVA (A) et représentation des produits et des descripteurs dans le biplot de la CVA des durées de dominance (B) pour trois chocolats à 70%, 70% doux et 90% de cacao. Les descripteurs les plus discriminants sont Amer, Sucre, Fondant, Astringent et Sec. Le chocolat à 70% est caractérisé par les descripteurs Croquant, Cacao et Sucre. Le chocolat à 70% doux est caractérisé par les descripteurs Croquant, Cacao, Sucre, Fondant et Gras. Le chocolat à 90% est surtout associé aux descripteurs Amer, Collant, Sec et Astringent.

L'ensemble des méthodes utilisées pour analyser des intensités en profil sensoriel peut alors être appliqué à ces durées de dominances. Les effets

panéliste, produit et leur interaction peuvent être testés à l'aide d'une ANOVA (Figure 8 A) et une cartographie des produits basée sur l'Analyse en Composantes Principales (ACP) ou sur la Canonical Variate Analysis (CVA) (Peltier, Visalli, & Schlich, 2015) (Figure 8 B) peut être réalisée. Cette approche donne des résultats intéressants et facilement interprétables mais qui ne prennent plus en compte la temporalité des dominances. Le moment d'apparition des descripteurs est notamment occulté.

A l'aide d'un découpage en périodes et du calcul des durées de dominance par période, les méthodes précédentes peuvent aussi prendre en compte une temporalité, certes simplifiée mais probablement suffisante dans la plupart des cas. Le temps de la dégustation peut être séparé en trois périodes de mêmes durées qui pourront alors être analysées séparément (Dinnella, Masi, Naes, & Monteleone, 2013; Lepage et al., 2014; Thomas, van der Stelt, Prokop, Lawlor, & Schlich, 2016).

Une autre approche pour prendre en compte la temporalité et comparer plusieurs produits est l'ACP des trajectoires (Lenfant, Loret, Pineau, Hartmann, & Martin, 2009). La dégustation est découpée en sous-parties de même durée proportionnellement à la durée totale. Une ACP est alors réalisée sur les taux de dominance avec comme variable les descripteurs et comme individus les sous-parties pour l'ensemble des produits. Ces sous-parties sont ensuite projetées sur un biplot en se basant sur les deux premières composantes et pour chaque produit les différents points projetés sont reliés dans l'ordre chronologique de la dégustation (Figure 9). Le graphique obtenu permet de comparer les produits en tenant compte de la temporalité mais n'est pas toujours facile à interpréter.

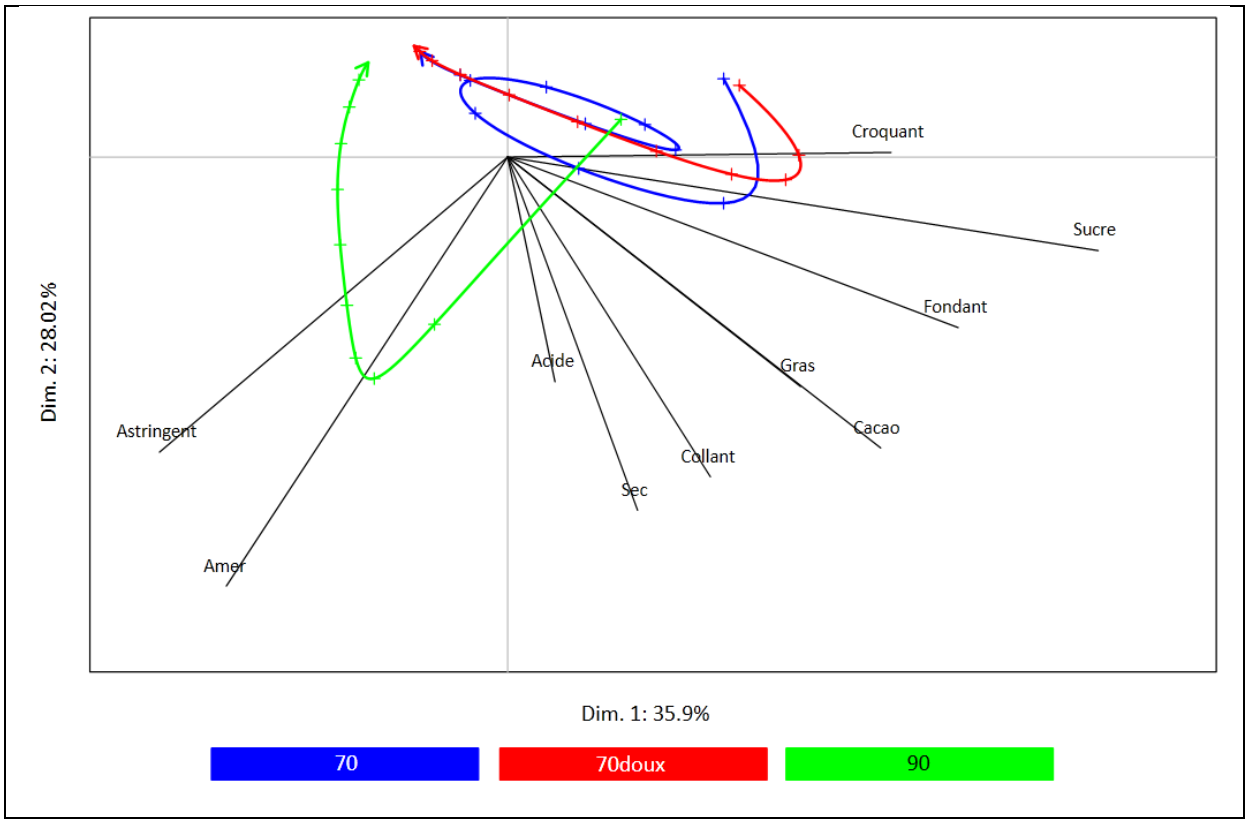


Figure 9 ACP des trajectoires avec 10 points pour trois chocolats à 70%, 70% doux et 90% de cacao. Tous sont d'abord croquants. Les chocolats avec 70% de cacao sont ensuite sucrés et fondants. Le chocolat à 90% est lui ensuite Amer avant d'être astringent.

Une évaluation des performances du panel et des panélistes a été proposée par Meyners (2011) se basant sur un calcul de distance entre matrices correspondant à différentes répétitions dont la significativité est évaluée à l'aide de tests de randomisation. Lepage et al. (2014) ont proposé une approche pour évaluer la capacité du panel et des panélistes à discriminer des produits se basant sur plusieurs indicateurs de performance résumés grâce à un arbre de décision. Ces approches s'avèrent compliquées à mettre en place, ne prennent pas en compte toute la complexité des données DTS et leur validité statistique est conditionnelle à un problème de multiplicité des risques de première espèce dus à la réalisation d'un grand nombre de tests statistiques.

Dans cette partie nous avons pu voir que plusieurs outils existent pour analyser les données DTS et nous avons pu en identifier les limites. Les méthodes existantes donnent une vision moyenne de la perception du panel mais ne permettent pas l'étude des différences interindividuelles de la perception temporelle. Ces méthodes regroupent un ensemble varié d'outils

statistiques qui manquent de cohérence et rendent une utilisation optimale difficile. Par ailleurs, plusieurs outils importants manquent encore à la DTS. Il est donc nécessaire de développer un ensemble de méthodes d'analyse se basant sur une approche unique et prenant en compte toute la richesse des données DTS. Ces méthodes devront entre autre permettre la visualisation des données, la comparaison de produits et la segmentation du panel selon la perception temporelle.

1.2 Concepts statistiques élémentaires

Cette partie présente deux concepts statistiques essentiels à la compréhension des travaux présentés par la suite : les variables aléatoires et la fonction de vraisemblance d'un échantillon ainsi que l'étude de son maximum.

1.2.1 Variables aléatoires

Définition

Une expérience est qualifiée d'aléatoire si son résultat ne peut être prévu à l'avance et si, répétée dans des conditions identiques, elle peut donner lieu à des résultats différents (Saporta, 2011). Une variable dont la valeur est déterminée en fonction du résultat d'une expérience aléatoire est appelée variable aléatoire et est généralement désignée par une des dernières lettres de l'alphabet (Dodge, 2007). D'un point de vue mathématique, c'est une application mesurable de l'ensemble fondamental Ω, composé de l'ensemble des évènements observables, vers un espace mesurable $(E, \mathcal{E})$ où E est un ensemble et $\mathcal{E}$ est une tribu sur E.

Une variable aléatoire est dite discrète si l'ensemble des valeurs qu'elle peut prendre est fini (dénombrable) et est dite continue si cet ensemble est un intervalle ou une union d'intervalles (non dénombrable).

Exemple de variable aléatoire discrète

La somme des valeurs de deux dés à 6 faces lancés simultanément est une variable aléatoire discrète définie sur l'ensemble $\Omega = \{(1,1); (1,2); (1,3); \dots; (6,6)\}$ et à valeur dans $E = \{2; 3; \dots; 11; 12\}$.

Exemple de variable aléatoire continue

Ce type de variable est souvent associé à une mesure. Par exemple la durée de dégustation d'un produit par un panéliste ou le poids d'une portion de fromage sont des variables aléatoires continues.

Loi de probabilité associée

Une loi de probabilité est un modèle qui vise à caractériser les fréquences d'observation des valeurs prises par une variable aléatoire. Dans le cas discret une loi de probabilité est définie par une fonction de probabilité qui associe à chaque évènement la probabilité de l'observer. Dans le cas continu les lois de probabilité sont définies par une fonction, appelée densité de probabilité et généralement notée f, positive ou nulle et intégrable, telle que la probabilité de l'intervalle $[a, b]$ est donnée par :

$$\int_a^b f(t)dt.$$

Les lois de probabilité peuvent également être caractérisées par leur fonction de répartition (ou fonction de distribution cumulative) notée F, telle que pour une variable aléatoire X de densité f_X, sa fonction de répartition F_X est définie pour tout nombre réel x par :

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt.$$

La valeur $F_X(x)$ est la probabilité d'observer une valeur inférieure ou égale à x.

Espérance et variance mathématique

L'espérance mathématique d'une variable aléatoire X, notée $E[X]$, est la moyenne pondérée des valeurs que la variable aléatoire peut prendre où les poids sont les probabilités avec lesquelles ces valeurs peuvent être prises. La variance d'une variable aléatoire mesure la dispersion de ces valeurs par rapport à l'espérance mathématique. Ainsi dans le cas discret, l'espérance mathématique de X est définie par :

$$E[X] = \sum_{i=1}^{n} p(x_i)\, x_i,$$

si X peut prendre n valeurs et où p est la fonction de probabilité de X. La variance de X est :

$$Var(X) = E[(X - E[X])^2]$$

$$= \sum_{i=1}^{n} p(x_i)\, x_i^2 - \left(\sum_{i=1}^{n} p(x_i)\, x_i \right)^2 .$$

Si la variable aléatoire X est continue alors son espérance mathématique est :

$$E[X] = \int_D x\, f_X(x)\, dx,$$

où D est l'intervalle sur lequel X prend ses valeurs et :

$$Var(X) = E[(X - E[X])^2]$$
$$= \int_D x^2\, f_X(x)\, dx - \left(\int_D x\, f_X(x)\, dx \right)^2 .$$

Quelques lois utiles pour la suite

Dans cette partie sont présentées trois lois usuelles de probabilité qui seront utilisées par la suite pour modéliser les durées de dominance. Pour plus de détails sur ces lois et une présentation plus complète des lois de probabilités usuelles, voir Saporta (2011).

Loi géométrique

Une épreuve de Bernoulli de paramètre p est une expérience aléatoire ayant pour probabilité de succès p et pour probabilité d'échec $1-p$. La loi géométrique de paramètre p est une loi discrète qui modélise le nombre d'épreuves de Bernoulli de paramètre p nécessaires pour obtenir un premier succès. La variable discrète X suit une loi géométrique si :

$$\Pr(X = k) = p(1-p)^{k-1}.$$

Par exemple le nombre de lancés d'un dé nécessaire pour obtenir un nombre pair suit une loi géométrique de paramètre $p = 1/2$ (Figure 10).

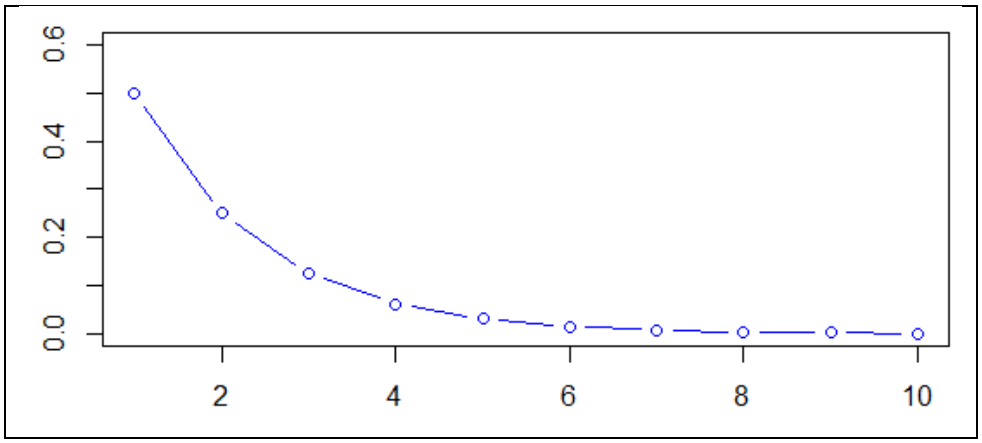


Figure 10 Distribution de la loi géométrique de paramètre p=1/2.

Loi exponentielle

La loi exponentielle est une loi continue qui est essentiellement utilisée pour modéliser la durée de vie d'un phénomène. Cette loi est beaucoup utilisée en fiabilité pour étudier l'état de fonctionnement d'un système. La loi exponentielle est dite sans mémoire ce qui signifie que la probabilité qu'une panne survienne à l'instant suivant, sachant que le système n'est pas en panne à l'instant présent, ne dépend pas du temps depuis lequel le système fonctionne. Une variable aléatoire X suit une loi exponentielle si sa densité (Figure 11) est donnée par :

$$f(t,\lambda) = \lambda e^{-\lambda t}.$$

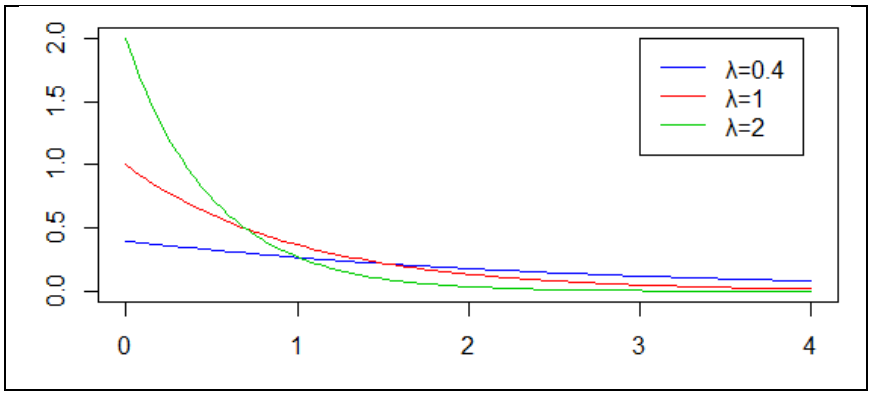


Figure 11 Densité de la loi exponentielle pour λ=0.4, λ=1 et λ=2.

Loi Gamma

La loi Gamma est une loi continue permettant la modélisation d'un grand nombre de phénomènes et est en particulier très utilisée pour modéliser des évènements évoluant au cours du temps. La loi Gamma est caractérisée par deux paramètres influençant respectivement sa forme et son intensité qui

permettent une grande flexibilité (Figure 12). Une variable à valeurs réelles positives suit une loi Gamma de paramètres a, λ si sa densité est donnée par :

$$f(t; a, \lambda) = \frac{t^{a-1}\lambda^{a} \exp(-\lambda t)}{\Gamma(a)},$$

où Γ désigne la fonction Gamma d'Euler.

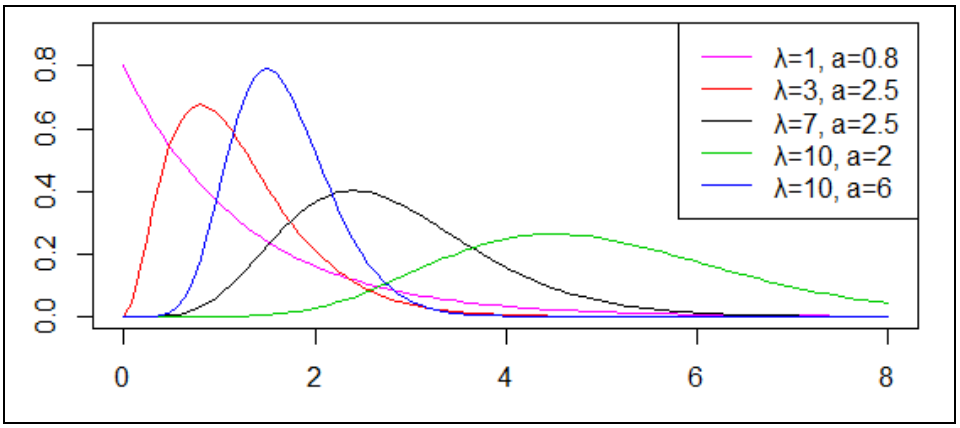


Figure 12 Densité de la loi Gamma selon différents paramètres.

1.2.2 La vraisemblance statistique

La modélisation d'une variable se fait généralement par l'estimation des paramètres propres à cette variable en se basant sur un échantillon d'observations. Par exemple, pour modéliser la taille des français à l'aide d'une loi normale, nous allons mesurer n français afin d'estimer l'espérance et la variance de cette variable, qui sont les deux paramètres d'une loi normale.

Plusieurs méthodes d'estimation existent. La plus utilisée est la méthode du maximum de vraisemblance. Les premières traces de cette méthode remontent au XVIII^{e} siècle dans les écrits de J.-H. Lambert et D. Bernoulli, mais la méthode est généralement attribuée à R. A. Fisher qui en 1912 l'introduisit sous le nom de « critère absolu » dans sa première publication statistique (Fisher, 1912). Il développera par la suite cette méthode et la nommera définitivement « maximum de vraisemblance » en 1922.

Le maximum de vraisemblance consiste à estimer les paramètres par les valeurs qui maximisent la probabilité d'obtenir l'échantillon observé. Il s'agit

donc de maximiser la fonction qui associe aux paramètres inconnus de la loi à ajuster la probabilité d'obtenir l'échantillon observé. Cette fonction est appelée fonction de vraisemblance. L'utilisation de la vraisemblance pour la modélisation est largement détaillée dans (Pawitan, 2013).

Définition de la fonction de vraisemblance

L'échantillon $x_1, x_2, \dots, x_n$ est supposé être la réalisation de n variables aléatoires indépendantes et identiquement distribuées ce qui signifie que les observations sont distribuées selon une même loi de probabilité dont nous appellerons la densité f ayant pour paramètre θ. La fonction de vraisemblance généralement notée L (pour likelihood) est le produit des probabilités d'observer chaque évènement en fonction des paramètres :

$$L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

L'estimateur du maximum de vraisemblance est alors obtenu en résolvant l'équation suivante qui vise à trouver le paramètre qui maximise la probabilité d'observer l'échantillon :

$$\frac{\partial}{\partial \theta} L(x_1, x_2, \dots, x_n; \theta) = 0.$$

Exemple

Nous souhaitons modéliser la taille des hommes en France à partir d'un échantillon composé de n individus par une loi normale $\mathcal{N}(\mu, \sigma^2)$ dont les paramètres doivent être estimés. La densité f de la loi normale est donnée par :

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

Ainsi la vraisemblance s'écrit :

$$L(x_1, x_2, \dots, x_n; \mu, \sigma^2) = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2).$$

Une première étape pour maximiser la vraisemblance consiste à passer au log ce qui permet de simplifier la maximisation en transformant les produits en sommes :

$$\begin{aligned}\log\big(L(x_1, x_2, \dots, x_n; \mu, \sigma^2)\big) &= \log\left(\prod_{i=1}^{n} f(x_i; \mu, \sigma^2)\right) \\ &= \sum_{i=1}^{n} \log\big(f(x_i; \mu, \sigma^2)\big) \\ &= \sum_{i=1}^{n} \left(-\log\big(\sigma\sqrt{2\pi}\big) - \frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\right).\end{aligned}$$

L'estimation $\hat{\mu}$ de μ est alors obtenue en résolvant l'équation suivante :

$$\frac{\partial}{\partial \mu} log\big(L(x_1, x_2, \dots, x_n; \mu, \sigma)\big) = 0.$$

$$\begin{aligned}\frac{\partial}{\partial \mu} log\big(L(x_1, x_2, \dots, x_n; \mu, \sigma)\big) &= \frac{\partial}{\partial \mu} \sum_{i=1}^{n} \left(-\log\big(\sigma\sqrt{2\pi}\big) - \frac{1}{2}\Big(\frac{x-\mu}{\sigma}\Big)^2\right) \\ &= \frac{\partial}{\partial \mu} \sum_{i=1}^{n} \left(-\frac{1}{2}\frac{x_i^2 + \mu^2 - 2x_i\mu}{\sigma^2}\right) \\ &= \sum_{i=1}^{n} \left(-\frac{\mu - x_i}{\sigma^2}\right)\end{aligned}$$

Il est alors facile de résoudre :

$$\sum_{i=1}^{n} \left(-\frac{\hat{\mu} - x_i}{\sigma^2}\right) = 0$$

$$-\frac{n\hat{\mu} - \sum_{i=1}^{n} x_i}{\sigma^2} = 0.$$

L'estimateur du maximum de vraisemblance est donc égal à la moyenne empirique : $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

L'estimateur $\widehat{\sigma^2}$ est déterminé de la même manière et est égal à la variance empirique : $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$.

La Figure 13 nous montre l'estimation obtenue par maximum de vraisemblance de la distribution de la taille des hommes pour une population simulée avec un échantillon de taille $n = 200$.

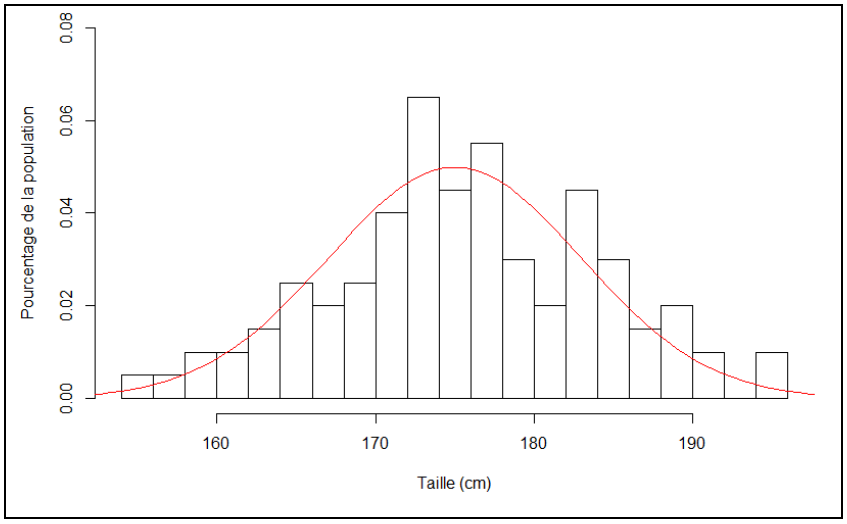


Figure 13 Histogramme et en rouge densité estimée pour $n = 200$ de la taille des hommes appartenant à une population simulée.

1.2.3 Simulation

Après avoir traduit un phénomène physique observé en modèle mathématique, il est possible d'utiliser ce modèle pour réaliser des simulations, c'est-à-dire générer de nouvelles données en cherchant à reconstituer le plus fidèlement possible le phénomène observé. La simulation de modèles stochastiques est connue sous le nom de méthode de Monte-Carlo en référence aux jeux de hasards pratiqués à Monte-Carlo.

Les techniques de simulation sont largement utilisées dans des domaines très divers tels que la simulation de vol en aéronautique, pour étudier la potentielle diffusion d'une maladie ou dans l'industrie automobile pour étudier les contraintes aérodynamiques au cours du développement d'une nouvelle carrosserie. La simulation, en remplaçant des expérimentations, permet de tester des hypothèses pour un coût minime.

La simulation est particulièrement utile pour étudier des enchaînements d'évènements aléatoires comme par exemple pour de la gestion de files d'attente. La simulation consiste alors à générer aléatoirement, selon un

tirage, la probabilité à chaque instant qu'une nouvelle commande arrive et qu'une commande soit honorée.

La simulation peut servir à contrôler la qualité du modèle en vérifiant que les données simulées sont semblables aux vraies données. Elle peut également être utilisée pour la réalisation d'un test statistique.

1.2.4 Tests statistiques

En statistique, un test ou test d'hypothèse est une procédure permettant de décider le rejet d'une hypothèse de départ H_0, appelée hypothèse nulle et considérée vraie *a priori*, ou son non-rejet en faveur de l'hypothèse alternative H_1. La statistique du test est une variable aléatoire qui va permettre de comparer l'échantillon aux données attendues si H_0 est vraie. Si la valeur observée de la statistique de test pour l'échantillon est trop peu probable selon la distribution de la statistique de test sous l'hypothèse H_0, alors on rejette l'hypothèse nulle au profit de l'hypothèse alternative, sinon on ne peut rejeter l'hypothèse nulle. Lorsque la distribution théorique de la statistique de test n'est pas connu, la simulation peut être utilisée pour estimer cette distribution en se basant sur les valeurs calculées sur des données simulées en respectant l'hypothèse H_0.

1.3 Les processus stochastiques

Historiquement, la théorie des probabilités a été développée pour la modélisation des jeux de hasard. Selon le probabiliste Émile Borel « le hasard n'est que le nom donné à notre ignorance et n'existerait pas pour un être omniscient ». En mathématiques, les probabilités ne visent pas à comprendre la nature profonde du hasard mais à permettre la modélisation de systèmes aléatoires ou en partie aléatoires afin de mieux les comprendre et éventuellement en prédire l'évolution. Il s'agit donc de prendre en compte l'effet de tout ce qui n'a pas été observé ou mesuré mais qui peut pourtant influencer le résultat de l'expérience ou du phénomène observé. Pour cela nous faisons appel à des modèles qui sont des abstractions des phénomènes, ou au moins d'une partie des phénomènes, construites en se basant sur les expérimentations ou les observations. La modélisation consiste à formaliser en langage mathématique un système afin de plus facilement pouvoir l'étudier et le comprendre. Les processus stochastiques (ou aléatoires) modélisent l'évolution temporelle de phénomènes aléatoires.

1.3.1 Chaînes de Markov

Le concept de chaîne de Markov a été proposé et développé par le mathématicien russe Andrej Andreevic Markov au début du XX^{ème} siècle. Les chaînes de Markov sont aujourd'hui largement utilisées dans de nombreux domaines : génétique, phylogénie, chimie, musique, contrôle qualité, réseaux de communication, finance, sciences sociales, …. Elles font l'objet d'une large littérature (Berchtold, 1998; Norris, 1997; Pardoux, 2007).

Définition

Une chaîne de Markov est une séquence de variables aléatoires, indexées par le temps, à valeurs dans un ensemble E appelé espace d'états et qui possède la propriété de Markov. Cette propriété signifie que l'état suivant pris par la chaîne ne dépend que de l'état présent et est indépendant du passé.

Propriété de Markov

Le processus stochastique $(X_n)_{n\in\mathbb{N}}$ à valeurs dans E respecte la propriété de Markov si pour tout $n \in \mathbb{N}$ et pour tout $i, j, i_0, \dots, i_{n-1} \in E^{n+2}$:

$$Pr(X_{n+1} = j | X_0 = i_0, X_1 = i_1, \dots, X_{n-1} = i_{n-1}, X_n = i) \; = \; Pr(X_{n+1} = j | \, X_n = i).$$

On peut parler de processus sans mémoire.

Quelques exemples d'application

Pour un réparateur d'ordinateurs, le nombre d'ordinateurs à réparer peut être modélisé à l'aide d'une chaîne de Markov en observant la fréquence d'arrivée d'un nouvel ordinateur en panne et le temps moyen pour réparer un ordinateur. Cet exemple peut être généralisé à de nombreux problèmes tels que la fiabilité d'une machine dans une usine ou la file d'attente à la caisse d'un magasin.

La navigation sur internet peut être modélisée à l'aide d'une chaîne de Markov en estimant les probabilités d'aller sur chaque page sachant sur quelle page l'internaute est actuellement.

La modélisation de la trajectoire réalisée par une grenouille sautant de nénuphar en nénuphar peut également être modélisée par une chaîne de Markov.

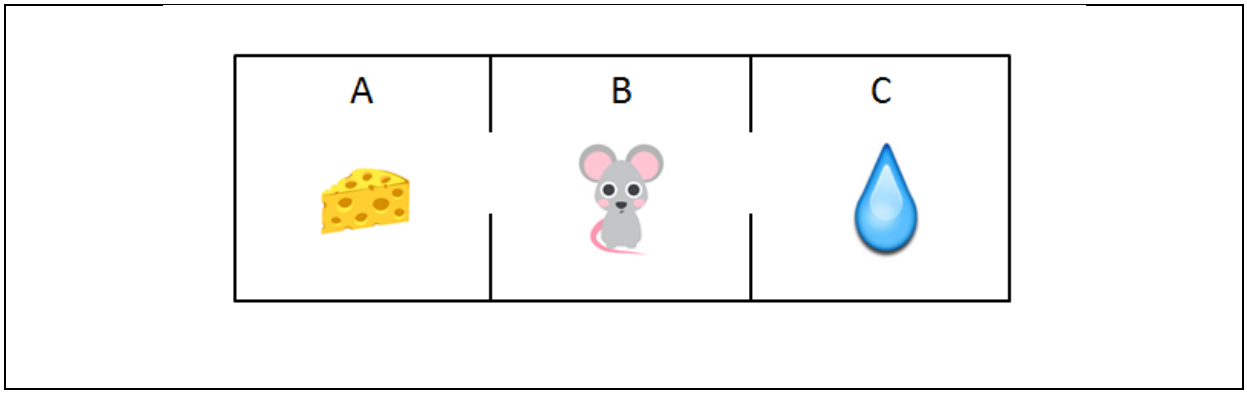


Figure 14 Exemple d'une cage séparée en trois compartiments avec à gauche la nourriture, à droite l'eau et entre les deux un compartiment vide.

La position d'une souris chaque minute dans une cage composée de trois compartiments (Figure 14) peut être modélisée par une chaîne de Markov $(X_n)_{n\in\mathbb{N}}$. Après trois séances de 5 minutes d'observation et en plaçant la

souris dans le compartiment B au départ, les trois séquences suivantes sont obtenues : BAABC, BBCBA, BABBA.

Homogénéité

Nous nous intéressons ici aux chaînes de Markov homogènes ce qui signifie que les probabilités de transition sont indépendantes du temps :

$$\Pr(X_{n+1} = i | X_n = j) = \Pr(X_{n+2} = i | X_{n+1} = j) = \Pr(X_1 = i | X_0 = j).$$

Par exemple, pour la souris, la probabilité d'être dans un des compartiments à l'instant suivant ne dépend que de sa position actuelle et est indépendante du moment auquel l'observation est faite.

Une chaîne de Markov est définie par une loi initiale donnant les probabilités d'occurrence du premier état de la chaîne et par une matrice de transition.

Matrice de transition (ou matrice stochastique)

Les probabilités de transition, c'est-à-dire de passer d'un état à un autre sont regroupées dans une matrice P, appelée matrice de transition. Pour deux états $i, j \in E$, P_{ij} est la probabilité de passer de l'état i à l'état j. La matrice P est dite markovienne ou stochastique et est caractérisée par les propriétés suivantes :

$$\forall\, i, j \in E, P_{ij} \geq 0;$$

$$\forall i \in E, \sum_{j \in I} P_{ij} = 1.$$

Les probabilités de transition sont estimées par le calcul des fréquences d'observation de chaque transition ce qui correspond à l'estimateur du maximum de vraisemblance qui sera présenté plus loin.

Exemple : la matrice de transition pour la position de la souris contient les probabilités pour la souris d'être dans chaque compartiment à l'instant suivant sachant sa position actuelle (Figure 15). Les probabilités sont estimées en se basant sur les transitions observées pour les 3 séquences présentées précédemment. Si la souris est dans le compartiment A alors à

l'instant d'après elle peut soit rester dans le compartiment A avec une probabilité égale à 1/3 soit aller dans le compartiment B avec une probabilité égale à 2/3.

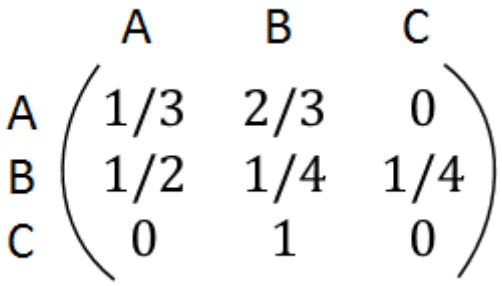


Figure 15 Matrice de transition pour la modélisation des déplacements d'une souris dans une cage composée de trois compartiments nommés A, B et C après avoir observé les trois parcours cités page 55.

Graphe de Markov

Les chaînes de Markov peuvent donner lieu à une représentation graphique qui permet de visualiser rapidement les relations entre les différents états. Selon Caumel (2015), « le graphe de transition de la chaîne discrète $(X_n)_{n\in\mathbb{N}}$ est le graphe orienté dont les sommets sont les états $(e_i)_i$ joints deux à deux par l'arc orienté e_i -> e_j, si et seulement si $p_{ij} > 0$ ».

Exemple

Le graphe de Markov pour l'exemple de la souris (Figure 16) permet de visualiser rapidement les probabilités de déplacement de la souris. Si la souris est dans le compartiment C de sa cage elle sera alors dans le compartiment B à l'instant suivant. Si la souris est dans le compartiment A alors à l'instant d'après il y a une chance sur trois qu'elle soit restée dans A et deux chances sur 3 qu'elle soit passée en B.

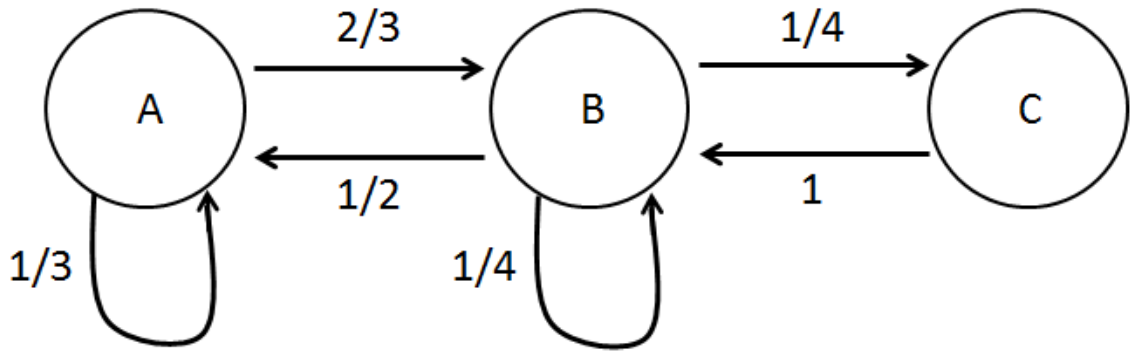


Figure 16 Graphe de Markov pour la modélisation des déplacements d'une souris dans une cage composé de trois compartiments nommés A, B et C après avoir observés les trois parcours cités page 55.

Temps de séjour

Le temps de séjour est le temps passé dans un état avant de passer à un autre état. A chaque instant deux évènements sont possibles, soit le système reste dans le même état soit le système change d'état. Le temps de séjour est distribué selon une loi géométrique qui modélise le temps nécessaire à l'obtention du premier succès en réalisant à chaque instant un tirage semblable à un lancer de pièce à pile ou face. Ici le premier succès est le passage à un autre état. Le nombre d'instants N_i passés dans l'état e_i avant que le système change d'état est donc modélisé comme suit :

$$\forall k \in \mathbb{N}, \Pr(N_i = k) = P(i,i)^k\big(1 - P(i,i)\big).$$

Processus de Markov à temps continu

Dans le cas où le système étudié évolue de manière continue dans le temps, il peut être modélisé avec une variante continue des chaînes de Markov appelée processus de Markov à temps continu principalement utilisée pour des problèmes de type file d'attente. Les temps de séjour dans chacun des états sont modélisés par une loi exponentielle qui a pour propriété l'absence de mémoire. Les probabilités de changement d'état ne dépendent que de l'état actuel et sont indépendantes du passé.

1.3.2 Chaînes semi-markoviennes et chaînes de renouvellement markovien

Les chaînes de Markov sont largement utilisées depuis des années et leur succès est sans doute largement dû à la simplicité de ce modèle et de la propriété de Markov. Néanmoins cette propriété impose des restrictions sur les temps de séjour qui doivent être distribués selon une loi géométrique dans le cas discret et selon une loi exponentielle dans le cas continu. La modélisation du temps de séjour par une loi géométrique ou par une loi exponentielle implique que les probabilités de rester dans le même état ou d'en changer sont constantes au cours du temps. En réalité cette propriété est rarement respectée. Par exemple, dans le cas de la modélisation du fonctionnement d'une machine qui va transiter entre les états opérationnel, en panne et en cours de réparation, la probabilité de panne est certainement faible quand la machine vient d'être réparée et va devenir de plus en plus élevée avec le temps. Au contraire il est possible par exemple que le temps de réparation soit en général très court, la probabilité de passer de en cours de réparation à opérationnel est alors élevée au début de la réparation et va être faible ensuite.

Les processus semi-markoviens sont une généralisation des processus markoviens permettant la modélisation des temps de séjour avec une loi quelconque tout en conservant la propriété de Markov mais uniquement pour les changements d'état. Les processus semi-markoviens ont été introduits dans les années 50 indépendamment par Levy (1954) et Smith (1955). Barbu & Limnios (2008) ont proposé une revue détaillée sur l'analyse théorique des chaînes semi-markoviennes à temps discret dont ce chapitre est en partie inspiré.

Le modèle

Une chaîne semi-markovienne $(Z_t)_{t\geq 0}$ décrit l'évolution au cours du temps d'un système à valeur dans un ensemble d'états E. L'idée sous-jacente des chaînes semi-markoviennes est de modéliser séparément les changements d'état avec une chaîne de Markov $(J_p)_{p=1,2,\dots}$ et le temps passé dans chaque

état successif $(X_p)_{p=1,2,\dots}$, c'est-à-dire le temps de séjour passé par le système dans l'état J_p . Le couple $(J_p, X_p)_{p=1,2,\dots}$ est appelé processus de renouvellement markovien (Pyke, 1961). Soit $N(t)$ la fonction qui à un instant t associe le nombre d'états successifs traversés par le système durant la période de temps $[0, t]$. L'évolution du système pendant une durée T peut ainsi être décrite par une séquence S de la manière suivante :

$$S = (J_1, X_1, \dots, J_{N(T)-1}, X_{N(T)-1}, J_{N(T)}, u_T),$$

où u_T est le temps de séjour censuré dans le dernier état. En effet la séquence a été interrompue à la durée T et le temps passé dans le dernier état aurait été potentiellement plus long sans cette interruption.

Le processus $Z_t = J_{N(t)}$ qui représente l'état du système à chaque instant constitue une chaine semi-markovienne homogène.

La fonction de vraisemblance pour une chaîne semi-markovienne s'écrit de la façon suivante :

$$L = \alpha_{J_1} \prod_{k=2}^{N(T)} P_{J_{k-1}J_k} f_{J_{k-1}J_k}(X_{k-1}) \bar{H}_{J_{N(T)}}(u_T),$$

où α_{J_1} est la probabilité initiale, c'est-à-dire la probabilité que le premier état pris par le système soit J_1, et $\bar{H}_{J_{N(T)}}(u_T)$ est la fonction qui donne la probabilité d'observer le dernier temps de séjour en prenant en compte la censure. $\bar{H}_{J_{N(T)}}(u_T)$ est négligeable lorsque T est suffisamment grand. La vraisemblance étant uniquement composée de produits, la log-vraisemblance est uniquement composée de sommes ce qui permet de maximiser indépendamment la vraisemblance des probabilités initiales, de la chaîne de Markov et des temps de séjour.

L'estimation des temps de séjour peut se faire soit par une approche non paramétrique, par exemple en affectant à chaque évènement une probabilité égale à sa fréquence d'apparition observée, soit par une approche

paramétrique, par exemple en estimant par maximum de vraisemblance une loi paramétrique telle que la loi gamma.

1.4 Les modèles de mélange

Les premières traces des modèles de mélange remontent à un mémoire de Siméon Denis Poisson (Poisson, 1837) dans lequel il modélise le résultat d'un procès comme une somme pondérée des différentes causes pouvant influencer le jugement. Les difficultés de diffusion des publications à l'époque ont fait que cet ouvrage n'a pas immédiatement été connu et que les mélanges de lois ont mis plusieurs décennies à être développés. D'autres grands noms de la statistique tels que Francis Galton et Karl Pearson ont contribué au développement des modèles de mélange. Un historique complet de cette méthode est donné dans (Droesbeke, Saporta, & Thomas-Agnan, 2013).

Dans le but de s'approcher toujours plus de la réalité, les modèles de mélange offrent la possibilité de modéliser des populations composées de G sous-populations ayant un comportement différent. Les modèles de mélange peuvent s'avérer très utiles dans de nombreux domaines tels que l'astronomie, la biologie, la bio-informatique, le traitement du signal, la chimie, l'économétrie, la robotique ou encore la biostatistique.

Les modèles de mélange, en plus d'être extrêmement utiles dans de nombreux domaines, soulèvent plusieurs problèmes théoriques, notamment l'identifiabilité des paramètres, ce qui a conduit à une vaste littérature.

Dans cette partie, nous définirons ce qu'est un modèle de mélange de lois puis nous introduirons l'algorithme EM, utilisé pour l'estimation des paramètres, avant de voir les différentes méthodes pouvant être utilisées pour sélectionner le meilleur modèle.

1.4.1 Définition des modèles de mélange

Exemple introductif

Supposons que nous souhaitons maintenant modéliser la taille d'une population sans tenir compte du sexe. Nous étudions pour cela un échantillon de $n = 400$ individus tirés au hasard dans la population. L'histogramme des tailles mesurées (Figure 17) suggère l'existence de deux

sous-populations et ne présente pas une distribution semblable à une loi normale. Pour modéliser la taille de la population nous allons donc faire appel à un mélange de deux lois normales. L'estimation réalisée à l'aide de l'algorithme EM, qui sera présenté plus loin, permet de retrouver les caractéristiques de la population des femmes pour une composante (densité en vert) et des hommes pour l'autre (densité en rouge). La somme pondéré des deux lois normales permet d'obtenir un modèle adapté aux données (densité en bleu).

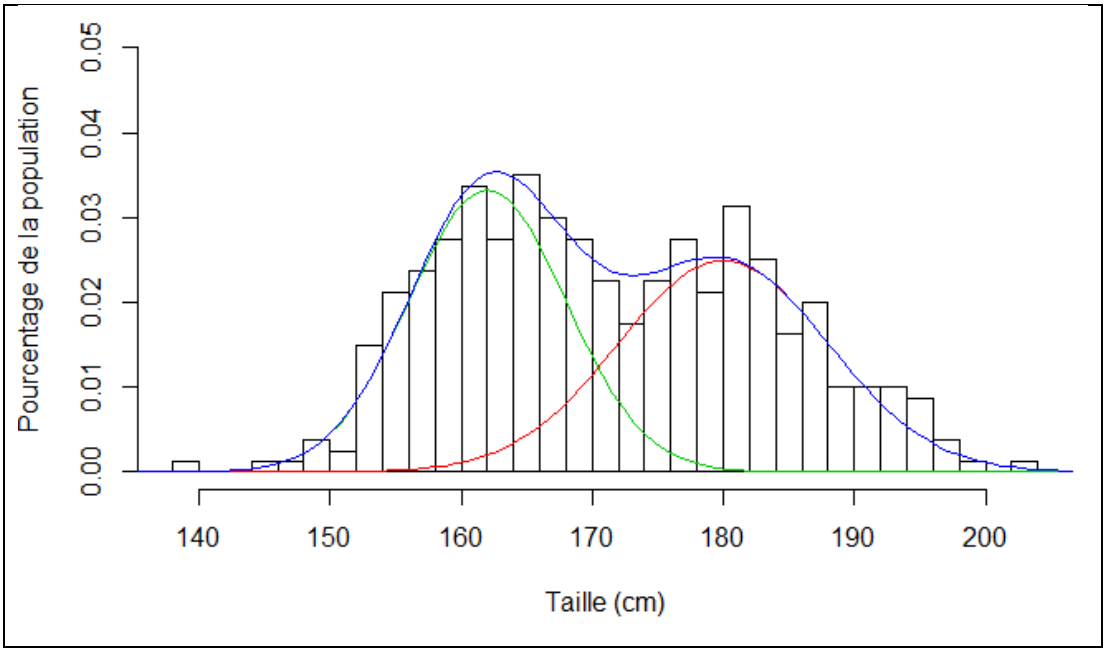


Figure 17 Histogramme représentant la répartition de la taille pour un échantillon simulé de 400 individus mélangeant hommes et femmes. La courbe rouge et la courbe verte représentent les densités des deux composantes du mélange et la courbe bleue représente la densité du modèle de mélange.

Définition

Une loi de mélange f est une combinaison linéaire de plusieurs lois de probabilité de densité respective $f_1, f_2, \dots, f_G$ pondérée par des poids $\pi_1, \pi_2, \dots, \pi_G$ tels que $0 < \pi_k \leq 1$ et $\sum_{g=1}^{G} \pi_g = 1$. La densité de f s'écrit alors :

$$f(x) = \sum_{g=1}^{G} \pi_g \, f_g(x).$$

Les π_g sont appelés les proportions du mélange et les f_g sont les composantes.

Les modèles de mélange sont généralement utilisés avec des distributions gaussiennes mais ils peuvent aussi bien être utilisés avec n'importe quel modèle paramétrique (voir par exemple (Banfield & Raftery, 1993; Frühwirth-Schnatter, 2006; McNicholas, 2016)).

1.4.2 L'algorithme EM

L'algorithme EM est apparu dans les années 60 et 70 à travers plusieurs publications mais est généralement attribué à Dempster, Laird, & Rubin (1977) qui ont réuni les principaux concepts. Cet algorithme permet de faire de l'estimation à l'aide du maximum de vraisemblance dans le cas de données incomplètes.

La maximisation de la fonction de vraisemblance d'un modèle de mélange est difficile car il faut estimer à la fois les paramètres des différentes composantes et les proportions du mélange. Une solution consiste à maximiser la vraisemblance du modèle complété qui consiste à attribuer à chaque observation une étiquette, qui est une variable non-observée (ou latente), l'associant à une composante. L'introduction du modèle complété permet l'utilisation de l'algorithme EM.

L'algorithme EM est un algorithme itératif où vont se succéder les étapes de calcul d'espérance de la variable latente, c'est-à-dire les probabilité t_{ig} pour chaque observation i d'avoir été générée par chaque composante g, et de maximisation de la vraisemblance complétée en tenant compte de la mise à jour des t_{ig} afin d'estimer au mieux les paramètres des composantes. L'algorithme EM se déroule comme suit :

- Initialisation de l'algorithme - Choix de la valeur initiale des paramètres $\theta^{(0)}$.
- Etape E - Calcul des probabilités conditionnelles à partir des estimations des paramètres $\theta^{(m-1)}$ et $\pi_j^{m-1}, j = 1, \dots, G$, obtenus à l'itération précédente :

$$t_{ig}^{(m)} = \frac{\pi_g^{(m-1)} L_g\left(x_i; \theta^{(m-1)}\right)}{\sum_{j=1}^{G} \pi_j^{(m-1)} L_j(x_i; \theta^{(m-1)})}.$$

La valeur $t_{ig}^{(m)}$ correspond à la probabilité que l'observation x_i provienne de la composante g.

- Etape M – Mise à jour de θ en maximisant la vraisemblance du modèle complet avec les nouvelles valeurs des t_{ig}. Les proportions du mélange estimées à l'étape m sont définies comme suit :

$$\pi_g^{(m)} = \frac{\sum_{i=1}^{n} t_{ig}^{(m)}}{n}.$$

L'estimation des paramètres des différentes composantes est faite par maximum de vraisemblance et dépend du modèle de mélange à estimer.

Initialisation de l'algorithme

L'initialisation est une étape cruciale pour s'assurer de la convergence vers le maximum global et pas vers un maximum local. Plusieurs stratégies peuvent être mises en place : chercher directement la meilleure initialisation, ou répéter un grand nombre de fois l'algorithme avec des initialisations différentes et choisir le résultat maximisant la vraisemblance.

Arrêt de l'algorithme

L'algorithme peut être arrêté une fois la convergence atteinte ou lorsque la différence de vraisemblance entre deux étapes successives est inférieure à un seuil fixé. L'algorithme peut également être arrêté après un nombre fixé d'itérations.

1.4.3 Les méthodes de choix du nombre de composantes

Le nombre de composantes n'étant pas toujours connu, il est nécessaire de faire appel à un outil statistique pour choisir un nombre de composantes optimal, c'est-à-dire un modèle de mélange s'ajustant bien aux données sans être trop complexe. Généralement, ce choix se fait en faisant appel à des critères d'information dont les deux plus utilisés sont AIC (Akaike Information Criterion) et BIC (Bayesian Information Criterion).

Le critère AIC (Akaike, 1974) est un compromis entre la vraisemblance du modèle et son nombre de paramètres. Le critère AIC est calculé de la manière suivante pour chaque nombre G de composantes envisagé :

$$AIC(G) = k - \ln\left(L\left(x_1, x_2, \dots, x_n; \hat{\theta}(G)\right)\right),$$

où $k = k(\theta(G))$ est le nombre de paramètres libres à estimer dans le cas où il y a G composantes.

Le nombre de composantes sélectionné est alors celui qui minimise le critère AIC.

Le critère BIC (Schwarz, 1978) est assez semblable au critère AIC à ceci près que le terme k est remplacé par $\frac{k}{2}\ln(n)$ ce qui lui confère de bonnes propriétés asymptotiques (Keribin, 2000).

Le critère BIC retrouve en général le bon nombre de composantes du modèle alors que AIC est connu pour choisir plus de composantes que nécessaire. D'autres critères ont été proposés tels que le critère AIC corrigé (AICc) conseillé lorsque le nombre de paramètres du modèle est grand comparé au nombre d'observations ou encore le critère ICL (Integrated Completed Likelihood) qui ajoute un terme d'entropie au critère BIC afin de sélectionner des composantes bien séparées.

1.4.4 Les modèles de mélange et la classification

Les modèles de mélange sont très largement utilisés pour réaliser de la classification automatique car l'idée de modéliser séparément les différentes sous-populations est assez intuitif et s'adapte à de très nombreuses situations. La classification se basant sur les modèles de mélange se fait principalement selon deux approches.

La première approche consiste à réaliser la classification par maximum a posteriori (Frühwirth-Schnatter, 2006). Après convergence de l'algorithme EM la classification se fait de manière assez naturelle en affectant chaque

individu à la composante à laquelle il maximise la probabilité d'appartenir t_{ik}.

La deuxième approche consiste à ajouter une étape de classification à l'algorithme EM qui devient alors l'algorithme CEM (Celeux & Govaert, 1992). L'étape C ajoutée avant l'étape M de l'algorithme convertie les t_{ik} en 1 si l'individu i a la plus grande probabilité d'appartenir au composante k et en 0 sinon. L'algorithme CEM converge en général très rapidement mais présente un risque élevé de converger vers un maximum local.

1.5 Objectifs et plan de cette thèse

Dans cette introduction, nous avons vu que l'analyse sensorielle consiste à mesurer la perception sensorielle, généralement pour des produits alimentaires. Les méthodes de mesure visent principalement à obtenir le profil sensoriel d'un ou plusieurs produits, à déterminer si deux produits sont différents, ou à connaître l'appréciation d'un produit par un groupe de consommateurs. Ces dernières années de nouvelles méthodes ont ajouté une dimension temporelle à ces mesures et sont désormais largement utilisées. Parmi ces méthodes, la DTS s'est démarquée, grâce à une utilisation simple ne nécessitant pas d'entraînement et permettant ainsi de faire appel directement au consommateur. Néanmoins, l'analyse des données DTS se limite soit à une analyse descriptive, soit à une analyse quantitative des durées de dominance qui ignore la séquentialité des sensations perçues. Les outils d'analyse existant fournissent des résultats intéressants mais souffrent de deux limites : les données DTS sont extrêmement riches mais cette richesse est en partie perdue à l'analyse ; les méthodes d'analyse des données DTS fournissent une vision moyenne du panel qui, étant donné la variabilité de perception au sein de la population, peut potentiellement ne correspondre à la perception de très peu de sujets, voire aucun.

La deuxième partie de cette introduction a présenté les concepts mathématiques pouvant être utilisés pour modéliser les données DTS. Les processus stochastiques, et notamment les processus markoviens et semi-markoviens, n'ont jamais été utilisés dans le domaine de l'analyse sensorielle malgré le virage pris depuis quelques années vers le temporel, explicitement modélisé par ces méthodes. De même, les modèles de mélange se révèlent utiles dans de très nombreuses situations où la population est en réalité composée de plusieurs sous-populations aux caractéristiques ou comportements différents, mais ne sont pas utilisés en analyse sensorielle. Non seulement les modèles de mélange fournissent une modélisation plus réaliste, mais ils permettent de plus de retrouver l'appartenance la plus vraisemblable des individus aux différentes sous-populations.

L'objectif de ce travail de thèse consiste à modéliser les données DTS en utilisant les processus stochastiques et les modèles de mélange, puis à partir de ce modèle, de développer des outils permettant de dépasser les limites évoqués ci-dessus et de répondre aux questions restant en suspens.

Ainsi, dans le chapitre 2, nous décrirons les différentes étapes de la modélisation des données DTS en justifiant chacun de nos choix. Nous commencerons par présenter la modélisation par une chaîne de Markov avant d'expliquer en quoi les chaînes semi-markoviennes sont plus adaptées à la modélisation des données DTS et nous décrirons l'estimation de ce modèle. Nous montrerons ensuite l'intérêt de découper en périodes la perception d'une prise de produit et comment déterminer automatiquement le nombre et la durée des périodes. Nous présenterons une approche pour tester l'existence de différence entre deux échantillons DTS. Finalement nous présenterons un modèle plus complexe tenant compte de l'existence de sous-populations avec des perceptions différentes et permettant de segmenter le panel en fonction des différences interindividuelles de perception temporelle.

Le chapitre 3 illustrera les méthodes présentées dans le chapitre 2 à travers des applications à 4 jeux de données issus d'études portant sur respectivement des chocolats Lindt Excellence, des chocolats Barry Callebaut, des fromages frais et des Goudas. Ce chapitre commencera par des exemples montrant l'adéquation du modèle proposé aux données DTS. L'interprétation des paramètres du modèle offre beaucoup d'informations mais le nombre de paramètres peut être élevé et leur lecture peut être complexe. Nous proposerons donc un moyen simple de visualiser les informations les plus importantes à l'aide d'un graphe. Nous verrons ensuite des exemples de découpage de la durée de dégustation en périodes. Plusieurs exemples de tests entre échantillons seront présentés à la fois entre produits différents et entre sous-échantillons d'un même panel pour tester l'existence de différences de perception temporelle entre les hommes et les femmes et entre des panélistes de pays différents. Finalement nous verrons des

exemples de segmentation des consommateurs selon leur perception temporelle pour un Gouda et pour un fromage frais.

Pour finir, ces résultats seront synthétisés puis discutés dans le chapitre 4, avec notamment un questionnement sur l'extension possible à d'autres méthodes de mesure temporelle et des recommandations sur l'emploi des nouveaux outils d'analyse proposés.

Chapitre 2.

Approche stochastique pour la modélisation de données DTS

Ce chapitre présente l'ensemble des méthodes statistiques utilisées dans cette thèse pour modéliser des données DTS à l'aide des modèles stochastiques. Dans un premier temps, nous introduirons les concepts et limites de la modélisation à l'aide d'une chaine de Markov. Ensuite, nous verrons comment améliorer la modélisation des données DTS avec un processus semi-markovien. Nous utiliserons alors ce modèle pour étudier l'existence de périodes temporelles lors de la dégustation d'un produit. Ces travaux ont fait l'objet d'une publication (Lecuelle, Visalli, Cardot, & Schlich, 2018) présentée en annexe p198. Nous utiliserons également le modèle pour tester si deux produits sont statistiquement perçus comme différents. Finalement, nous proposerons une modélisation des données DTS prenant en compte l'existence de segments au sein du panel avec des perceptions différentes (travaux présentés dans un article (Cardot, Lecuelle, Schlich, & Visalli, 2019) présent en annexe p207) et nous montrerons comment utiliser ce modèle pour identifier ces segments.

2.1 Modélisation par une chaîne de Markov

Un processus stochastique décrit l'évolution d'une variable aléatoire au cours du temps. Les données DTS sont constituées des différentes valeurs prises par la variable « descripteur dominant » au cours de la durée de la dégustation et le choix du descripteur dominant est entaché d'un aléa. Le phénomène étudié est donc par définition un processus stochastique. Suite à ce constat il est logique de vouloir faire appel à un modèle stochastique pour modéliser les données DTS mais les processus stochastiques constituent une large famille de méthodes et il va donc falloir trouver laquelle est la plus adaptée.

2.1.1 Le modèle

Une des méthodes les plus utilisées et simple à mettre en place est la modélisation par une chaîne de Markov homogène à temps discret. Cette approche a été appliquée pour la première fois aux données DTS par Franczak, Browne, McNicholas, Castura & Findlay (2015) mais a uniquement fait l'objet d'un poster et n'a pas été approfondie.

Pour modéliser les données DTS avec une chaîne de Markov homogène à temps discret il faut d'abord discrétiser le temps de dégustation, en prenant par exemple un pas de 1 seconde. L'idée est alors d'étudier les probabilités de choisir les différents descripteurs à l'instant $t+1$ en ne prenant en compte que le descripteur dominant à l'instant t. L'estimation du modèle consiste à estimer la matrice de transition P, c'est-à-dire l'ensemble des probabilités de transition p_{ij}, où p_{ij} est la probabilité de passer du descripteur i à l'instant t au descripteur j à l'instant $t+1$. Cette estimation se fait intuitivement en dénombrant le nombre d'observations de chaque transition puis en divisant par le nombre de transitions partant du même descripteur :

$$\hat{p}_{ij} = \frac{n_{ij}}{n_{i.}},$$

où n_{ij} est le nombre de transitions $i \to j$ observées et $n_{i.}$ est le nombre de transitions partant de l'état i observées.

Pour un jeu de données, il est possible de montrer que la modélisation par une chaîne de Markov est plus adaptée qu'une modélisation sans tenir compte du précédent descripteur qui consiste simplement à estimer les fréquences d'observation des descripteurs. Guttorp (1995) propose par exemple d'utiliser un test du Chi2 pour comparer les vraisemblances du jeu de données selon ces deux modèles.

L'analyse des données consiste à étudier le graphe de Markov associé à la matrice de transition. Il est alors possible d'observer quels changements de descripteurs sont les plus probables et éventuellement de rechercher un ou plusieurs chemins dans le graphe qui peuvent correspondre aux façons de percevoir le produit. Le niveau de probabilité de rester dans le même état apporte également une information sur le temps durant lequel les panélistes conservent ce descripteur comme dominant. Toutefois les données DTS proposent généralement un espace d'état composé d'une dizaine de descripteurs, ce qui conduit à un graphe complexe, chargé et donc difficilement lisible. Nous proposerons au chapitre 2 une amélioration de ce

graphe afin de le rendre à la fois plus simple à utiliser et plus riche en information.

2.1.2 Limites

Les chaînes de Markov constituent un outil intéressant pour l'analyse des données DTS mais la qualité d'ajustement du modèle peut être améliorée, en particulier pour les durées de dominance. Les temps de séjour sont en effet modélisés soit par une loi géométrique dans le cas discret, soit par une loi exponentielle dans le cas continu. Ces deux lois ont pour caractéristique d'être « sans mémoire » ce qui signifie qu'à chaque instant les probabilités de changer de descripteur ne dépendent pas de la durée depuis laquelle le descripteur est dominant. Cette caractéristique ne semble pas adaptée aux données DTS puisque, par exemple, la probabilité de cliquer sur un nouveau descripteur dominant est très faible juste après une seconde. Plusieurs raisons peuvent expliquer ce temps minimal observé entre deux clics : le temps physique nécessaire pour déplacer le pointeur de la souris sur un nouveau descripteur, le temps nécessaire à l'apparition d'une nouvelle sensation ou le temps de prise de décision.

Pour vérifier que les durées de dominance ne sont effectivement pas distribuées selon une loi exponentielle il est possible de réaliser un test statistique d'adéquation à cette loi tel qu'un test de Kolmogorov-Smirnov. Cette hypothèse a été vérifiée sur l'ensemble des jeux de données que nous avons eu l'occasion d'étudier.

2.2 Modélisation par un processus semi-markovien

Les processus semi-markoviens, contrairement aux processus de Markov, permettent une modélisation « libre » des temps de séjour dans les états successifs traversés. Les changements de descripteur dominant sont encore modélisés par une chaîne de Markov mais les durées durant lesquelles les descripteurs restent dominants sont modélisées séparément. Nous proposons de modéliser les durées de dominance en utilisant des lois binomiales négatives si le temps est discrétisé (Lecuelle et al., 2018) ou des lois Gamma dans le cas continu. Bien que la modélisation en temps discret puisse se justifier par le fait que les mesures sont discrètes, nous préférons modéliser les durées en temps continu pour mieux correspondre à la réalité physique de l'évènement.

2.2.1 Notations et estimation

Les changements de descripteurs sont modélisés par une chaîne de Markov homogène $(J_p)_{p\geq 1}$ à valeur dans l'espace d'état $S = \{1, \dots, D\}$ avec D le nombre de descripteurs. Cette chaîne de Markov est définie par la matrice de transition P qui prend comme valeurs l'ensemble des probabilités de passer d'un descripteur dominant à un autre : $P_{lj} = \Pr[J_{p+1} = j \mid J_p = l], l, j \in S$. Cette chaîne de Markov modélise uniquement les changements de descripteur, ce qui signifie que pour tout $j \in S, P_{jj} = 0$.

Le choix du premier descripteur dominant est modélisé par le vecteur des probabilités initiales $\alpha = (\alpha_1, \dots, \alpha_D)$, où pour tout $j \in S, \alpha_j = \Pr[J_1 = j]$.

On note $(X_p)_{p\geq 1}$ la séquence aléatoire constituée des durées de dominance successives, c'est-à-dire les temps de séjour dans chaque descripteur dominant. Ainsi pour tout $p \geq 1$, X_p est le temps de séjour passé dans l'état J_p et est donc à valeur dans $T = 1, 2, \dots$ si le temps est discrétisé et dans $T = \mathbb{R}_+$ sinon. Soit $\Phi_l(t) = \Pr[X_p \leq t \mid J_p = l]$ la fonction de répartition du temps de séjour étant donné l'état actuel du processus $(J_p)_{p\geq 1}$. Bien que classiquement la modélisation des temps de séjour dépende de l'état actuel et du suivant,

nous nous contentons ici d'une dépendance sur l'état actuel permettant de limiter le nombre de paramètres à estimer tout en conservant un modèle réaliste. En d'autres termes nous considérons que la distribution des temps de séjour est la même peu importe l'état suivant (ce qui nous semble légitime d'un point de vue sensoriel) :

$$Pr[X_p \leq t | J_p = l, J_{p+1} = j] = Pr[X_1 \leq t | J_1 = l].$$

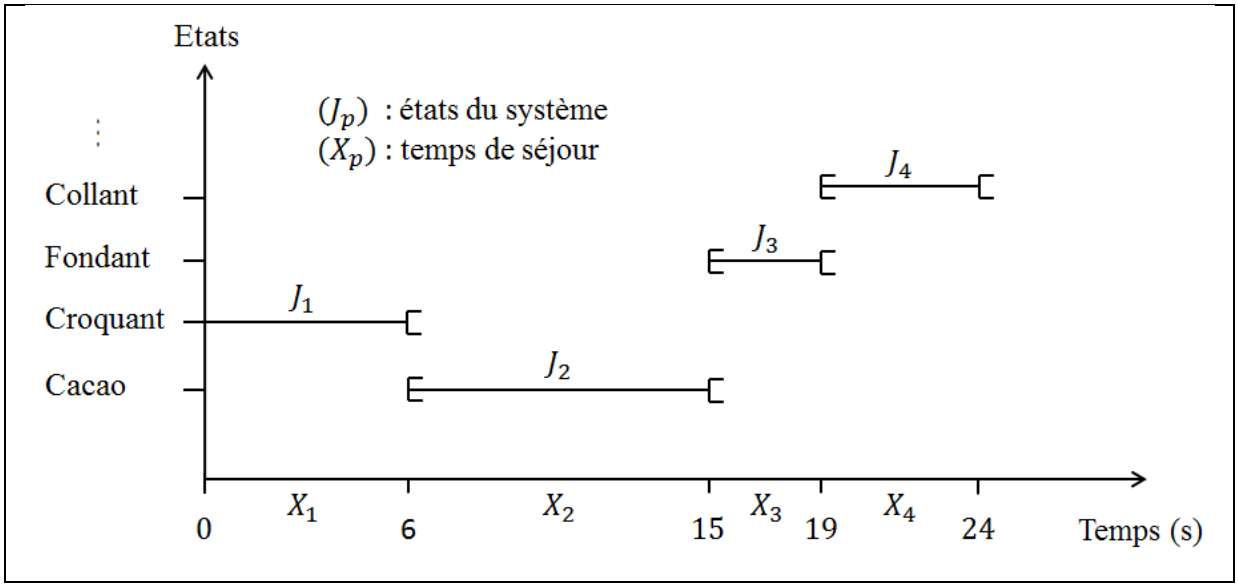


Figure 18 : Modélisation d'une séquence d'un chocolat avec un processus de renouvellement markovien $(J_p, X_p)_{p\geq 1}$. Les descripteurs dominants sont J_1 = Croquant, J_2 = Cacao, J_3 = Fondant et J_4 = Collant.

Le processus stochastique $(J_p, X_p)_{p\geq 1}$ satisfait la propriété de Markov, ce qui signifie que pour tout $t \in T$ et $l, j \in S$:

$$Pr[J_{p+1} = j, X_p \leq t \mid J_p = l, J_{p-1}, \dots, J_1, X_{p-1}, \dots, X_1] = Pr[J_{p+1} = j, X_p \leq t \mid J_p = l].$$

Cette propriété importante énonce que le choix de remplacer un descripteur dominant par un autre est indépendant du temps durant lequel le premier est resté dominant ; une fois encore, nous pensons que cela fait sens d'un point de vue sensoriel. Le processus $(J_p, X_p)_{p\geq 1}$ est appelé processus de renouvellement markovien et le processus stochastique donnant l'état du système à chaque instant $t \in T$ est appelé processus semi-markovien (Barbu &

Limnios, 2008). Un exemple de cette modélisation est donné dans la Figure 18 pour une séquence issue de la dégustation d'un chocolat par un panéliste.

Comme pour une simple chaine de Markov, la probabilité de transition P_{ij} est estimée de la manière suivante :

$$\hat{P}_{ij} = \frac{n_{ij}}{n_{i.}},$$

où n_{ij} est le nombre de transitions observées de l'état i à l'état j et $n_{i.}$ le nombre de transitions observées partant de l'état i.

Les temps de séjour dans les différents états sont modélisés par des lois Gamma. Ce choix se justifie par la nature continue des données, la grande flexibilité des lois Gamma et la simplicité d'estimation des paramètres de cette loi par la méthode des moments. La densité est définie selon deux paramètres $a > 0$ et $\lambda > 0$ comme suit :

$$f(t, a, \lambda) = \frac{t^{a-1}\lambda^a \exp(-\lambda t)}{\Gamma(a)}, t > 0,$$

où $\Gamma(a)$ est la fonction gamma. L'espérance et la variance de la loi Gamma de paramètres (a, λ) valent respectivement $\frac{a}{\lambda}$ et $\frac{a}{\lambda^2}$. Pour chaque descripteur d les durées sont modélisées par une loi Gamma ayant pour paramètres (a_j, λ_j) et dont la densité est notée $f_j(t) = f(t, a_j, \lambda_j), \forall\, t > 0$.

Les paramètres sont estimés par maximum de vraisemblance en faisant appel à un algorithme d'optimisation (par exemple la méthode de quasi-Newton) puisque l'un des deux paramètres ne peut pas être obtenu de manière explicite par le calcul. L'estimation peut également se faire par la méthode des moments qui est beaucoup plus simple mais l'estimateur est réputé moins bon pour les petits échantillons (Ye & Chen, 2017). L'estimation des paramètres par la méthode des moments est néanmoins utilisée pour initialiser l'algorithme d'optimisation utilisé pour obtenir le maximum de vraisemblance.

2.2.2 Calcul de la vraisemblance

Pour un produit, un jeu de données DTS est constitué de n séquences indépendantes $S_i,\ i = 1, \dots, n,$ correspondant chacune à une évaluation du produit par un panéliste et considérées comme n réalisations indépendantes d'un processus de renouvellement markovien. Chaque séquence S_i a été observée pendant une durée T_i au cours de laquelle le processus de renouvellement markovien a visité $N(T_i)$ descripteurs successifs et nous supposons que $N(T_i) \geq 2$. Ainsi une séquence S_i est définie comme suit :

$$S_i = \left(J_1^i, X_1^i, \dots, J_{N(T_i)-1}^i, X_{N(T_i)-1}^i, J_{N(T_i)}^i, X_{N(T_i)}^i\right), i = 1, \dots, n.$$

La vraisemblance $L(S_i, \theta)$ de la séquence S_i en fonction des paramètres $\theta = (\alpha, P, f_l, l \in S)$ du processus de renouvellement markovien est définie par :

$$L(S_i; \theta) = \alpha_{J_1^i} f_{J_1^i}\left(X_1^i\right) \prod_{k=2}^{N(T_i)} P_{J_{k-1}^i, J_k^i} f_{J_k^i}\left(X_k^i\right).$$

Grâce à l'indépendance des séquences S_i, la vraisemblance de l'ensemble du jeu de données est simplement donnée par :

$$L(S_1, \dots, S_n;\ \theta) = \prod_{i=1}^{n} L(S_i, \theta).$$

2.2.3 Simulation

A partir du modèle estimé, il est possible de simuler des nouvelles données. La simulation d'une réalisation d'une chaîne semi-markovienne est obtenue en tirant aléatoirement le premier descripteur selon les probabilités initiales du modèle à partir duquel les simulations sont réalisées, puis en tirant aléatoirement et successivement les descripteurs selon les probabilités de transitions et la durée passée dans chaque descripteur selon les lois estimées. Nous faisons le choix de contrôler la quantité d'information contenue dans les données simulées en fixant le nombre de transitions par séquence simulée au nombre moyen de transitions observées dans les séquences du jeu de données utilisé pour estimer le modèle servant aux simulations. Une autre approche est d'inclure dans le modèle les transitions

vers l'état absorbant « STOP », qui va mettre fin à la séquence dès qu'il sera atteint, et de continuer la simulation jusqu'à ce que cet état soit atteint. Dans ce cas, il y a un risque important d'obtenir des séquences soit très courtes soit très longues qui ne correspondent pas à ce qui peut être observé dans un vrai jeu de données. Si le but des simulations est de valider une hypothèse, il est préférable d'avoir le contrôle sur la quantité d'information simulée.

2.2.4 Limites

L'utilisation des processus semi-markoviens, en répondant à la problématique de la modélisation des durées de dominance, améliore sensiblement la qualité de la modélisation des données DTS, mais elle peut encore être améliorée. Jusqu'ici la modélisation est faite sous l'hypothèse que les probabilités de transition sont homogènes, c'est-à-dire indépendantes du temps. Cette hypothèse semble forte puisque lors de dégustations la probabilité d'aller vers certains descripteurs peut vraisemblablement être plus importante à certains moments. Par exemple, dans le cas de la dégustation d'un chocolat, il est souvent observé que les descripteurs de texture, tels que croquant et fondant ont des probabilités d'apparition respectivement plus élevées au début et à la fin de la dégustation.

Néanmoins, s'il se trouve que l'hypothèse d'une dépendance au temps des probabilités de transition est vraie, la quantité de données à disposition dans un jeu de données DTS classique ne permet pas de modéliser la DTS à l'aide d'un modèle hétérogène, c'est-à-dire d'évaluer les probabilités de transition à chaque instant de la dégustation. Pour prendre en compte les variations des probabilités de transition au cours d'une dégustation, une solution est de découper la durée de dégustation en plusieurs périodes consécutives.

2.3 Découpage en périodes temporelles

Le découpage des séquences DTS en périodes temporelles a déjà été proposé dans la littérature mais de manière assez réductrice en proposant systématiquement de découper la durée de la dégustation en 3 périodes de durée uniforme (Dinnella et al., 2013; Lepage et al., 2014; Thomas et al., 2016). Dans cette partie, nous proposons un découpage adapté aux données où le nombre de périodes et la position des frontières entre périodes vont être sélectionnés de manière à avoir des probabilités de transition les plus homogènes possible au sein de chaque période.

Dans un premier temps nous considérerons que le nombre de période est connu et nous proposerons une méthode pour positionner de manière optimale les frontières entre périodes temporelles. Dans un deuxième temps, nous proposerons un test statistique pour définir si un découpage en périodes temporelles est nécessaire et, si c'est le cas, quel est le nombre optimal de périodes.

2.3.1 Sélection optimale de la position des frontières entre périodes

La position des frontières entre périodes est calculée pour chaque produit. Cette position est déterminée comme un pourcentage de la durée de chaque séquence. Pour simplifier l'implémentation et la représentation des périodes, les frontières peuvent être déterminées après avoir standardisé à droite et à gauche les séquences. Chaque séquence commence alors à l'instant du premier clic et a une durée de 1. Le temps est alors discrétisé en 101 point de 0 à 1. La détermination de la position des frontières se fait en calculant pour chaque position possible des frontières la vraisemblance du modèle, en modélisant séparément chaque période par un processus semi-markovien, puis en sélectionnant la position maximisant cette vraisemblance. Soit F le nombre de frontières, ce qui signifie qu'il y a $F+1$ périodes, et soit n séquences observées $S_1, \dots, S_n$, alors la position optimale estimée pour les frontières $fr_1, \dots, fr_F$ est :

$$(\widehat{fr_1}, \dots, \widehat{fr_F}) = argmax_{(fr_1,\dots,fr_F)} \prod_{l=1}^{F+1} L(S_1^l, \dots, S_n^l; \hat{\theta}_l),$$

où S_i^l est la période l de la séquence i et $\hat{\theta}_l$ est l'ensemble des paramètres estimés du processus semi-markovien modélisant la période l.

Pour représenter graphiquement les périodes, la position des frontières peut être superposée sur les courbes DTS dans le cas où les données ont été préalablement standardisées avant la construction de ces courbes.

2.3.2 Sélection du nombre de périodes

L'algorithme proposé pour sélectionner le nombre de périodes nécessaire pour modéliser au mieux un produit est un algorithme de type « forward », c'est-à-dire que le nombre de périodes va être incrémenté jusqu'à ce qu'il soit montré statistiquement que l'ajout d'une période supplémentaire n'améliore pas la modélisation.

La première étape consiste à tester la nécessité de découper en deux périodes. La position de la frontière fr_1 est estimée pour le produit étudié. Il faut ensuite construire un test qui vise à montrer si deux matrices de transition sont significativement différentes ou non. Soit P^1 la matrice de transition estimée pour la première période et P^2 celle estimée pour la deuxième période alors :

$$H_0 : P^1 = P^2,$$

$$H_1 : P^1 \neq P^2.$$

Sous l'hypothèse H_0 un grand nombre de jeux de données sont simulés à partir du processus semi-markovien estimé sur le produit étudié sans découpage en périodes. Pour chaque jeu de données simulé, les données sont découpées en deux périodes en utilisant la frontière fr_1 déterminée précédemment, les matrices de transitions sont estimées pour ces deux périodes et la distance entre ces matrices est calculée de la manière suivante :

$$dist(P^1, P^2) = \sum_{i=1}^{D} \sum_{j=1}^{D} \left| P_{ij}^1 - P_{ij}^2 \right|.$$

La distribution de l'ensemble des distances calculées de la sorte offre ainsi une estimation de la distribution de la distance sous l'hypothèse H_0. Le test consiste alors à observer où se situe la distance calculée pour le vrai jeu de données par rapport à cette distribution. Les deux matrices de transition sont considérées comme significativement différentes si la p-value est inférieure à 5%, ce qui signifie que seul 5% des jeux de données simulés donnaient une distance entre matrices de transition supérieure à cette valeur.

Dans le cas où les matrices de transition ne sont pas significativement différentes, les probabilités de transition sont les mêmes tout au long de la dégustation, ce qui signifie que le produit est considéré comme n'ayant pas de périodes. En revanche si les matrices de transition sont significativement différentes, alors il y a au moins 2 périodes lors de la dégustation de ce produit. La prochaine étape est alors de tester si les matrices de transition sont différentes en découpant en 3 périodes. Les périodes sont comparées deux à deux dans l'ordre d'apparition : la première période est comparée à la deuxième, la deuxième à la troisième. Il est en effet inutile de comparer la première période à la troisième puisque si ces deux périodes ont des matrices de transitions identiques mais différentes de celle de la deuxième période alors il est tout de même nécessaire de découper en trois périodes.

Le nombre de périodes est incrémenté jusqu'au premier nombre de périodes s pour lequel deux périodes consécutives ont des matrices de transition non significativement différentes. Le nombre de périodes sélectionné pour le produit étudié est alors égal à $s - 1$.

2.4 Test de différence entre produits

Il est important en analyse sensorielle de définir un outil permettant de tester statistiquement si deux produits sont différents. Il s'agit d'une question complexe qui a été initiée dans ma thèse, puis continuée dans celle en cours de Cindy Frascolla. La démarche a fait l'objet d'une présentation au congrès annuel de la SFdS en 2019 (Frascolla, Lecuelle, Cardot, Schlich, & Visalli, 2019).

Soit deux produits testés respectivement par n_1 et n_2 juges générant les séquences $(S_1^1, \dots, S_{n_1}^1)$ et $(S_1^2, \dots, S_{n_2}^2)$. En utilisant la modélisation proposée précédemment, une solution pour tester si ces deux produits sont perçus significativement différemment est d'utiliser un rapport de vraisemblance entre un modèle unique pour les deux produits et un modèle différent pour chaque produit. L'idée est donc d'observer si l'amélioration de la vraisemblance en modélisant les deux produits séparément est statistiquement significative. Soit $\theta = (\alpha, P, (a_l, \lambda_l)_{l \in S})$ les paramètres du processus de renouvellement markovien estimés en modélisant les deux produits ensemble et $\theta^1 = (\alpha^1, P^1, (a_l^1, \lambda_l^1)_{l \in S})$, $\theta^2 = (\alpha^2, P^2, (a_l^2, \lambda_l^2)_{l \in S})$ les paramètres estimés respectivement pour la modélisation du produit 1 et du produit 2. L'hypothèse du test est alors définie comme suit :

$$H_0 : \theta_1 = \theta_2,$$

$$H_1 : \theta_1 \neq \theta_2.$$

La statistique de test est définie par le rapport de vraisemblance :

$$LR = \frac{\max_{\theta \in \Theta} \prod_{i=1}^{n_1} L(S_i^1; \theta) \, x \prod_{j=1}^{n_2} L(S_j^2; \theta)}{\max_{(\theta_1, \theta_2) \in \Theta x \Theta} \prod_{i=1}^{n_1} L(S_i^1; \theta_1) \prod_{j=1}^{n_2} L(S_j^2; \theta_2)},$$

où Θ est l'ensemble des valeurs que peuvent prendre les paramètres.

La loi décrivant la distribution de cette statistique de test n'est pas connue et nous proposons donc de l'estimer par la méthode de Monte-Carlo, c'est-à-dire en faisant appel à des simulations à partir du modèle aléatoire estimé, afin de pouvoir définir la zone de rejet du test. Frascolla et al. (2019)

proposent d'utiliser une loi du Chi-2. En effet, sous des hypothèses classiques (cf. van der Vaart (1998)) la statistique $-2\ln LR$ converge en loi sous H_0, lorsque le nombre de séquences tend vers l'infini, vers une loi du Chi-2 dont le nombre de degrés de liberté est égal au nombre de composantes de Θ. Ils ont également proposé de l'estimer par permutation.

L'estimation se fait en simulant un grand nombre de fois, à partir d'un même modèle estimé sur l'ensemble des 2 produits, 2 nouveaux jeux de données, considérés comme 2 produits différents et composés respectivement de n_1 et n_2 séquences. Puis il suffit de calculer pour chaque paire de produits simulée le rapport de vraisemblance. La distribution des valeurs ainsi obtenues est une estimation de la distribution de la statistique de test sous H_0.

2.5 Segmentation

Précédemment, les données DTS ont été modélisées à l'aide d'un processus de renouvellement markovien $(J_p, X_p)_{p\geq 1}$ ayant pour paramètres $\theta = (\alpha, P, \Phi_l, l \in S)$. Cette modélisation est réalisée sous l'hypothèse que tous les panélistes ont la même perception du produit étudié et que les données observées correspondent à cette perception entachée d'une erreur due à la complexité de la tâche. Cette vision est sans doute réductrice et il est probable que la variation au sein des réponses d'un panel puisse aussi s'expliquer par de réelles différences de perception (Prutkin et al., 2000). Il a été montré que de nombreux facteurs peuvent en effet influencer la perception tels que l'âge (Hutchings, Foster, Grigor, Bronlund, & Morgenstern, 2014; Schiffman & Graham, 2000) ou la sensibilité aux saveurs comme par exemple la perception du gras (Nachtsheim & Schlich, 2013; Pingel, Ostwald, Pau, Hummel, & Just, 2010; Schoumacker et al., 2017). Jaeger et al. (2017) ont listé les nombreux facteurs dont l'influence sur la perception a été montrée dans la littérature et ont souligné que bien souvent la réponse moyenne est une mauvaise représentation de la perception de la population.

Afin de prendre en compte l'hétérogénéité au sein d'un panel, la solution est la segmentation (Koster, 2009; Meiselman, 2013). Pour cela, une approche largement utilisée consiste à faire appel à la segmentation basée sur un modèle de mélange (G. J. McLachlan & Peel, 2000; Melnykov & Maitra, 2010). Les mélanges de chaînes de Markov sont utilisés dans différents domaines tels que la finance (Frydman, 2005), l'informatique (Song, Keromytis, & Stolfo, 2009), l'estimation du trafic routier (Lawlor & Rabbat, 2017) ou encore l'étude du marché du travail (Pamminger & Fruhwirth-Schnatter, 2010). A notre connaissance le mélange de processus semi-markoviens n'a en revanche jamais été proposé dans la littérature.

2.5.1 Notations

Soit G processus semi-markoviens indépendants à valeur dans un même espace d'états $S = \{1, \dots, D\}$, définis par un vecteur de probabilités initiales α^g, une matrice de transition P^g et des fonctions de répartition $\Phi_l^t(t), t \in T$. En

notant $\pi_g > 0$ la probabilité qu'un individu soit généré par la composante g, le modèle de mélange a pour loi :

$$\sum_{g=1}^{G} \pi_g Loi\left(\alpha^g, P^g, \Phi_l^g, l \in S\right).$$

Des informations théoriques sur ce modèle démontrant qu'un modèle de mélange de processus de renouvellement markovien est un processus de renouvellement markovien et montrant l'identifiabilité de ce modèle sous des hypothèses assez naturelles sont données dans Cardot, Lecuelle Schlich & Visalli (2019). L'identifiabilité est très importante pour les modèles de mélange (Frühwirth-Schnatter, 2006; Titterington, Smith, & Makov, 1985) puisqu'elle assure l'unicité des lois décrivant le mélange.

Les temps de séjour dans les différents états sont modélisés par des lois Gamma. La densité de la loi Gamma de paramètres $\left(a_{\mathrm{lg}}, \lambda_{lg}\right)$ modélisant les durées de dominance du descripteur l de la composante g est notée $f_l^g(t) = f\left(t, a_{\mathrm{lg}}, \lambda_{lg}\right), \forall\, t > 0$.

2.5.2 Estimation par maximum de vraisemblance

Nous nous intéressons à un échantillon composé de n panélistes qui ont réalisé chacun B dégustations indépendantes d'un même produit. Pour chaque panéliste i, avec $i = 1, \dots, n$, nous avons donc obtenu B séquences S_i^b, pour $b = 1, \dots, B$, de durée T_i^b et ayant un nombre d'états visités $N\left(T_i^b\right)$ supposé supérieur ou égal à 2. En reprenant les notations utilisées en 2.2.2 :

$$S_i^b = \left(J_1^{i,b}, X_1^{i,b}, \dots, J_{N\left(T_i^b\right)-1}^{i,b}, X_{N\left(T_i^b\right)-1}^{i,b}, J_{N\left(T_i^b\right)}^{i,b}, X_{N\left(T_i^b\right)}^{i,b}\right).$$

Nous supposons que les trajectoires $S_1^1, \dots, S_1^B, \dots, S_n^1, \dots, S_n^B$ ont été générées par un mélange de G processus semi-markoviens dont les paramètres sont à estimer. Nous supposons également dans un premier temps que le nombre de composantes du mélange G est connu.

Nous supposons comme dans 0 que les distributions des temps de séjour ne dépendent que de l'état actuel et sont indépendantes de l'état suivant.

La vraisemblance

La vraisemblance associée à un individu statistique i avec B répétitions indépendantes selon le processus de renouvellement markovien de paramètres $\theta_g = (\alpha^g, P^g, f_l^g, l \in S)$ est notée $L_g(S_i^1, \dots, S_i^B; \theta_g)$ et est définie comme suit :

$$L_g(S_i^1, \dots, S_i^B; \theta_g) = \prod_{b=1}^{B} L_g(S_i^b; \theta_g)$$
$$= \prod_{b=1}^{B} \left[\alpha^g_{J_1^{i,b}} f^g_{J_1^{i,b}}(X_1^{i,b}) \prod_{k=2}^{N(T_i)} P^g_{J_{k-1}^{i,b}, J_k^{i,b}} f^g_{J_k^{i,b}}(X_k^{i,b}) \right].$$

En considérant maintenant que la composante du mélange à laquelle appartient l'individu i n'est pas connue, la log-vraisemblance des nB séquences observées selon le modèle de mélange est alors :

$$\ln L(S_1^1, \dots, S_n^B; \theta) = \sum_{i=1}^{n} \ln \left(\sum_{g=1}^{G} \pi_g \prod_{b=1}^{B} L_g(S_i^b; \theta_g) \right),$$

où $\theta = (\pi, \theta_1, \dots, \theta_G)$ est l'ensemble des paramètres du modèle de mélange.

Une maximisation de la log-vraisemblance par rapport à θ est compliquée et les algorithmes classiques d'optimisation ne fonctionnent pas en général pour ce genre de problèmes (voir par exemple (Geoffrey J. McLachlan & Krishnan, 2008)). L'algorithme EM permet de réaliser facilement cette maximisation en décomposant la procédure d'optimisation en deux étapes simples.

EM pour un mélange semi-markovien

L'algorithme EM est une technique itérative d'optimisation utilisé dans le cas où une partie des données est considérée manquante. Nous allons donc d'abord « compléter » le modèle en ajoutant des variables décrivant l'appartenance des séquences aux composantes qui seront considérées comme données manquantes. Soit $Z_i, i = 1, \dots, n$ ces variables constituées d'un vecteur de longueur G dont les éléments notés Z_{ig} valent 1 si l'individu

appartient à la composante g et 0 sinon. Chaque vecteur Z_i est donc constitué de $G-1$ zéros et 1 un. L'ensemble des données et des variables $Z_i, i = 1, \dots, n$ est appelé données complétées. La log-vraisemblance des données complétées s'écrit de la manière suivante :

$$\ln L_C(S_1^1, \dots S_1^B, Z_1, \dots, S_n^1, \dots, S_n^B, Z_n; \theta) = \sum_{i=1}^{n}\sum_{g=1}^{G} Z_{ig} \ln\left(\pi_g \prod_{b=1}^{B} L_g\left(S_i^b; \theta_g\right)\right)$$

$$= \sum_{i=1}^{n}\sum_{g=1}^{G} Z_{ig} \ln \pi_g + \sum_{i=1}^{n}\sum_{g=1}^{G} Z_{ig} \sum_{b=1}^{B} \ln L_g\left(S_i^b; \theta_g\right).$$

Cette fonction de vraisemblance est plus simple à maximiser que la version non-complétée puisqu'elle n'est composée que de sommes et d'aucun produit.

L'algorithme EM se déroule alors de la manière suivante :

- Initialisation des paramètres θ par des valeurs $\theta^{(0)}$
- Étape E

L'étape E consiste à calculer l'espérance de la log-vraisemblance complétée avec les valeurs des séquences observées et les valeurs de paramètres obtenues à la précédente itération de l'algorithme. Soit :

$$Q\left(\theta, \theta^{(m-1)}\right) = E\left[\ln L_C(S_1, Z_1, \dots, S_n, Z_n; \theta) \mid S_1, \dots, S_n, \theta^{(m-1)}\right]$$

$$= \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{Z}_{ig}^{(m)} \ln \pi_g^{(m-1)} + \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{Z}_{ig}^{(m)} \sum_{b=1}^{B} \ln L_g\left(S_i^b; \theta_g^{(m-1)}\right),$$

avec $\hat{Z}_{ig}^{(m)} = E\left[Z_{ig} \mid S_1, \dots, S_n, \theta^{(m-1)}\right]$ la probabilité conditionnelle que la séquence S_i ait été générée par la composante g du modèle de mélange de paramètres $\theta^{(m-1)}$, où $\theta^{(m-1)}$ est la valeur des paramètres calculée à la précédente itération. Cette étape consiste donc à mettre à jour les valeurs des Z_i en fonction des nouvelles valeurs des paramètres obtenues à la précédente itération de l'algorithme. En utilisant le théorème de Bayes qui donne l'égalité suivante pour 2 évènements A et B :

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)},$$

alors la probabilité conditionnelle $\hat{Z}_{ig}^{(m)}$ s'écrit :

$$\begin{aligned}\hat{Z}_{ig}^{(m)} &= \Pr\left(Z_{ig} = 1 \mid S_i; \theta^{(m-1)}\right) \\ &= \frac{\Pr\left(S_i \mid Z_{ig} = 1; \theta^{(m-1)}\right)\Pr\left(Z_{ig} = 1; \theta^{(m-1)}\right)}{\Pr\left(S_i; \theta^{(m-1)}\right)} \\ &= \frac{\pi_g^{(m-1)} \prod_{b=1}^{B} L_g\left(S_i^b; \theta^{(m-1)}\right)}{\sum_{j=1}^{G} \pi_j^{(m-1)} \prod_{b=1}^{B} L_j\left(S_i^b; \theta^{(m-1)}\right)}.\end{aligned}$$

- Étape M

L'étape M consiste à mettre à jour les paramètres θ en fonction des nouvelles valeurs des $\hat{Z}_{ig}^{(m)}$ en maximisant $Q\left(\theta, \theta^{(m-1)}\right)$ selon θ.

Les différents paramètres sont estimés séparément grâce à la structure multiplicative de la vraisemblance. Le passage au log transforme cette multiplication en somme et seul le paramètre modélisé ne disparaît pas lors de la dérivation partielle.

L'estimation des proportions du mélange π s'obtient en maximisant la vraisemblance à laquelle est ajouté un terme pour prendre en compte la contrainte $\sum_{g=1}^{G} \pi_g = 1$. Cette maximisation est obtenue en cherchant la valeur pour laquelle la dérivée partielle est nulle (voir par exemple (Bierlaire, 2015) pour une revue détaillée des méthodes d'optimisation) :

$$\frac{\partial}{\partial \pi_g}\left[\sum_{i=1}^{n}\sum_{g=1}^{G} \hat{Z}_{ig}^{(m)} \ln\left(\pi_g\right) + \lambda\left(\sum_{g=1}^{G} \pi_g - 1\right)\right] = 0,$$

où λ, qui sert à pénaliser les solutions ne respectant pas la contrainte $\sum_{g=1}^{G} \pi_g = 1$, est le multiplicateur de Lagrange associé à cette contrainte.

On obtient :

$$\frac{\partial}{\partial \pi_g}\left[\sum_{i=1}^{n}\sum_{g=1}^{G}\hat{Z}_{ig}^{(m)}\ln(\pi_g)+\lambda\left(\sum_{g=1}^{G}\pi_g-1\right)\right]=\sum_{i=1}^{n}\frac{\hat{Z}_{ig}^{(m)}}{\pi_g}+\lambda,$$

et donc :

$$\pi_g=-\frac{\sum_{i=1}^{n}\hat{Z}_{ig}^{(m)}}{\lambda}.$$

En utilisant le fait que $\sum_{g=1}^{G}\pi_g=1$ on a alors :

$$\sum_{g=1}^{G}\left(-\frac{\sum_{i=1}^{n}\hat{Z}_{ig}^{(m)}}{\lambda}\right)=-\frac{\sum_{g=1}^{G}\sum_{i=1}^{n}\hat{Z}_{ig}^{(m)}}{\lambda}=-\frac{n}{\lambda}=1.$$

On en déduit que $\lambda=-n$ ce qui permet d'obtenir la solution classique $\pi_g^{(m)}=n^{-1}n_g^{(m)}$ avec $n_g^{(m)}=\sum_{i=1}^{n}\hat{Z}_{ig}^{(m)}$.

En faisant à nouveau appel à des multiplicateurs de Lagrange on obtient l'estimation des probabilités initiales :

$$\hat{\alpha}_j^{g(m)}=\frac{\sum_{i=1}^{n}\hat{Z}_{ig}^{(m)}\sum_{b=1}^{B}\mathbb{1}_{\left\{J_1^{i,b}=j\right\}}}{B\sum_{i=1}^{n}\hat{Z}_{ig}^{(m)}},$$

et l'estimation des probabilités de transition :

$$\hat{P}_{hj}^{g(m)}=\frac{\sum_{i=1}^{n}\hat{Z}_{ig}^{(m)}\sum_{b=1}^{B}n_{hj}^{ib}}{\sum_{l=1}^{D}\sum_{i=1}^{n}\hat{Z}_{ig}^{(m)}\sum_{b=1}^{B}n_{hl}^{ib}},$$

où n_{hj}^{ib} est le nombre de transitions $h\to j$ observées dans la séquence S_i^b.

Pour l'estimation des modèles de mélange de lois Gamma il est nécessaire d'ajouter un terme de pénalisation car la log vraisemblance n'est pas bornée (Chen, Li, & Tan, 2016). Intuitivement, si le ratio $a_{lg}/\lambda_{\mathrm{lg}}$, qui correspond au temps de séjour moyen dans l'état l pour la composante g, est gardé constant, quand a_{lg} tend vers l'infini alors $a_{lg}/\lambda_{\mathrm{lg}}^2$ tend vers 0 et la densité de la loi Gamma correspondante, se comportant comme une distribution de Dirac en $a_{lg}/\lambda_{\mathrm{lg}}$, ne sera pas bornée. En conséquence, pour éviter ce genre de

solution, il est préférable d'introduire une pénalisation à l'étape M qui empêche le paramètre a_{lg} de devenir trop grand. Ainsi, on ajoute une pénalité à la fonction Q similaire à celle proposée dans Chen et al. (2016) et définie comme suit :

$$Pen(a_{lg}, l \in S, g = 1, \dots, G) = -\frac{1}{\sqrt{\sum_{i=1}^{n}\sum_{b=1}^{B} N(T_i^b)}} \sum_{g=1}^{G}\sum_{l \in S}(a_{\mathrm{lg}} + ln\, a_{\mathrm{lg}}).$$

Il est à noter que cette pénalisation ne concerne pas le paramètre λ_{lg} et que son effet décroit lorsque la taille de l'échantillon et le nombre de transitions augmentent.

L'estimation des paramètres des distributions des temps de séjour se fait alors en maximisant :

$$\sum_{i=1}^{n}\sum_{g=1}^{G} \hat{Z}_{ig}^{(m)} \sum_{b=1}^{B}\sum_{k=1}^{N(T_i^b)} lnf_{j_k^{ib}}^{g}(X_k^{ib}) + Pen(a_{lg}, l \in S, g = 1, \dots, G).$$

Initialisation de l'algorithme

L'initialisation de l'algorithme, c'est-à-dire le choix des valeurs initiales des différents paramètres, est crucial pour assurer une convergence rapide et vers le résultat optimal. Galmarini, Visalli & Schlich (2017) ont montré que le temps passé dans chaque état constitue un indicateur intéressant pour étudier les données TDS. Chaque séquence est alors caractérisée par une variable à valeur réelle pour chaque état. A partir de ces données, une première segmentation est réalisée en utilisant l'algorithme des k-means d'Hartigan-Wong (Hartigan & Wong, 1979) et les paramètres du mélange $\theta^{(0)}$ sont estimés selon cette segmentation.

Sélection du nombre de composantes

Le nombre de composantes du mélange est sélectionné en utilisant le critère d'information BIC qui repose sur un compromis entre la qualité de la modélisation et la complexité du modèle :

$$BIC(G) = q\ln(nB) - 2\ln L\left(S_1^b, \dots, S_n^b, b = 1, \dots, B;\ \hat{\theta}(G)\right),$$

où $\hat{\theta}(G)$ est l'estimation des paramètres pour un mélange avec G composantes et $q = q\left(\theta(G)\right)$ est le nombre de paramètres à estimer. Dans le cas où les transitions vers l'état STOP ne sont pas modélisées, le nombre de paramètres est $q = G - 1 + G(D - 1 + D(D - 2) + Dd) = GD(D + d - 1) - 1$, avec d le nombre de paramètres de la loi utilisée pour modéliser les durées soit 2 lorsqu'il s'agit de la loi Gamma. Dans le cas où il y a un état absorbant, c'est-à-dire lorsque les transitions vers l'état STOP sont modélisées, et en supposant que l'état absorbant ne peut pas être un état initial, alors $q = G - 1 + G(D - 2 + (D - 1)(D - 2) + (D - 1)d)$.

D'autres critères d'information peuvent être utilisés tels que le critère d'information d'Akaike (AIC) qui est identique au BIC sauf que le terme $q\ln(nB)$ qui pénalise la complexité du modèle est remplacé par $2q$ ou encore le critère AIC corrigé, noté AIC_c dans lequel le terme $q\ln(nB)$ est remplacé par $2q + \frac{2q(q+1)}{nB-q-1}$.

Critère de segmentation

Une fois que l'algorithme a convergé, c'est-à-dire que la vraisemblance et les paramètres restent identique d'une itération à l'autre, la segmentation est réalisée en utilisant le critère de la probabilité maximum *a posteriori* (MAP) qui est défini comme suit : $MAP\left(\hat{Z}_{ih}\right) = 1$ si $g = argmax_h(\hat{Z}_{ih})$ et $MAP\left(\hat{Z}_{ih}\right) = 0$ sinon. Chaque panéliste est ainsi affecté dans le segment auquel il a la plus grande chance d'appartenir selon l'estimation du modèle.

2.5.3 Segmentation simultanée pour plusieurs produits

La méthode présentée précédemment propose de segmenter le panel pour un produit afin de distinguer les différentes façons de percevoir ce produit. Néanmoins les études utilisant la DTS sont généralement réalisées avec

plusieurs produits et il peut être intéressant de segmenter le panel en groupes ayant une perception similaire de l'ensemble des produits. Pour simplifier, nous considérons le cas où il n'y a pas de répétitions. Nous nous intéressons donc maintenant à n panélistes ayant testé une fois U produits. Pour chaque panéliste i, avec $i = 1, \dots, n$, nous avons donc obtenu U séquences S_i^u, pour $u = 1, \dots, U$, de durée T_i^u et ayant un nombre d'états visités $N(T_i^u)$ supposé supérieur ou égal à 2. En reprenant les notations utilisées précédemment on a :

$$S_i^u = \left(J_1^{i,u}, X_1^{i,u}, \dots, J_{N(T_i^u)-1}^{i,u}, X_{N(T_i^u)-1}^{i,u}, J_{N(T_i^u)}^{i,u}, X_{N(T_i^u)}^{i,u}\right).$$

Pour réaliser la segmentation sur l'ensemble des produits nous allons utiliser un modèle de mélange avec un processus semi-markovien par segment et par produit ayant pour paramètres $\theta_g^u = \left(\alpha^{gu}, P^{gu}, f_l^{gu}, l \in S\right)$. Il est à noter que le nombre de paramètres est multiplié par U par rapport à une modélisation par produit, mais la quantité de données utilisée pour l'estimation est elle aussi multipliée par ce même facteur.

La vraisemblance

La vraisemblance associée à un individu statistique i ayant testé indépendamment U produits selon le processus de renouvellement markovien de paramètres $\theta_g^u = \left(\alpha^{gu}, P^{gu}, f_l^{gu}, l \in S\right)$ est définie comme suit :

$$\begin{aligned} L_g\left(S_i^1, \dots, S_i^U; \theta_g^1, \dots, \theta_g^U\right) &= \prod_{u=1}^{U} L_g\left(S_i^u; \theta_g^u\right) \\ &= \prod_{u=1}^{U} \left[\alpha_{J_1^{i,u}}^{gu} f_{J_1^{i,u}}^{gu}\left(X_1^{i,u}\right) \prod_{k=2}^{N(T_i)} P_{J_{k-1}^{i,u}, J_k^{i,u}}^{gu} f_{J_k^{i,u}}^{gu}\left(X_k^{i,u}\right) \right]. \end{aligned}$$

En considérant maintenant que la composante du mélange à laquelle appartient l'individu i n'est pas connue, la log-vraisemblance des nU séquences observées selon le modèle de mélange est alors :

$$\ln L(S_1^1, \dots, S_n^U; \theta) = \sum_{i=1}^{n} ln\left(\sum_{g=1}^{G} \pi_g \prod_{u=1}^{U} L_g\left(S_i^u; \theta_g^u\right) \right),$$

où $\theta = (\pi, \theta_1^1, \dots, \theta_1^U, \dots, \theta_G^1, \dots, \theta_G^U)$ est l'ensemble des paramètres du modèle de mélange.

EM

La log-vraisemblance des données complétées s'écrit de la manière suivante :

$$\ln L_C(S_1^1, \dots S_1^U, Z_1, \dots, S_n^1, \dots, S_n^U, Z_n; \theta) = \sum_{i=1}^{n}\sum_{g=1}^{G} Z_{ig} \ln\left(\pi_g \prod_{u=1}^{U} L_g\left(S_i^u; \theta_g^u\right)\right)$$

$$= \sum_{i=1}^{n}\sum_{g=1}^{G} Z_{ig} \ln \pi_g + \sum_{i=1}^{n}\sum_{g=1}^{G} Z_{ig} \sum_{u=1}^{U} \ln L_g\left(S_i^u; \theta_g^u\right).$$

L'algorithme EM se déroule alors de la manière suivante :

- Initialisation des paramètres θ par des valeurs $\theta^{(0)}$
- Étape E

Soit :

$$Q\left(\theta, \theta^{(m-1)}\right) = E\left[\ln L_C(S_1, Z_1, \dots, S_n, Z_n; \theta) \mid S_1, \dots, S_n, \theta^{(m-1)}\right]$$

$$= \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{Z}_{ig}^{(m)} \ln \pi_g^{(m-1)} + \sum_{i=1}^{n}\sum_{g=1}^{G} \hat{Z}_{ig}^{(m)} \sum_{u=1}^{U} \ln L_g\left(S_i^u; \theta_g^{u(m-1)}\right).$$

La probabilité conditionnelle $\hat{Z}_{ig}^{(m)}$ s'écrit désormais :

$$\hat{Z}_{ig}^{(m)} = \frac{\pi_g^{(m-1)} \prod_{u=1}^{U} L_g\left(S_i^u; \theta_g^{u(m-1)}\right)}{\sum_{j=1}^{G} \pi_j^{(m-1)} \prod_{u=1}^{U} L_j\left(S_i^u; \theta_j^{u(m-1)}\right)}.$$

- Étape M

L'estimation des proportions π du mélange se fait de la même manière que précédemment.

L'estimation des probabilités initiales est obtenue comme suit :

$$\hat{\alpha}_j^{ug(m)} = \frac{\sum_{i=1}^{n} \hat{Z}_{ig}^{(m)} \mathbb{1}_{\left\{J_1^{i,u}=j\right\}}}{\sum_{i=1}^{n} \hat{Z}_{ig}^{(m)}},$$

L'estimation des probabilités de transition est désormais :

$$\hat{P}_{hj}^{ug(m)} = \frac{\sum_{i=1}^{n} \hat{Z}_{ig}^{(m)} n_{hj}^{iu}}{\sum_{l=1}^{D} \sum_{i=1}^{n} \hat{Z}_{ig}^{(m)} n_{hl}^{iu}},$$

où n_{hj}^{iu} est le nombre de transitions $h \to j$ observées dans la séquence S_i^u.

Le terme de pénalisation pour le mélange de lois Gamma vaut :

$$Pen\left(a_{lg}^{u}, l \in S, g = 1, \dots, G, u = 1, \dots, U\right) = -\frac{1}{\sqrt{\sum_{i=1}^{n} N(T_i^u)}} \sum_{g=1}^{G} \sum_{l \in S} (a_{\mathrm{lg}}^{u} + ln\, a_{\mathrm{lg}}^{u}).$$

L'estimation des paramètres des distributions des temps de séjour se fait alors en maximisant :

$$\sum_{i=1}^{n} \sum_{g=1}^{G} \hat{Z}_{ig}^{(m)} \sum_{k=1}^{N(T_i^u)} ln f_{j_k^{iu}}^{gu}\left(X_k^{iu}\right) + Pen\left(a_{lg}^{u}, l \in S, g = 1, \dots, G, u = 1, \dots, U\right).$$

2.6 Bilan

Ce chapitre présente la théorie de la modélisation des données DTS par un processus semi-markovien et les applications qui en découlent : découpage en périodes temporelles, test de différence entre produits et segmentation. Les méthodes présentées, contrairement aux méthodes existantes, permettent de prendre en compte toute la complexité des données DTS. Le cadre théorique des processus stochastiques apporte une réponse méthodologique cohérente aux différentes questions que soulève l'analyse des données DTS. Ce travail va même plus loin en proposant une méthode innovante non seulement en sensométrie mais plus largement en statistique puisque, à notre connaissance, le mélange de processus semi-markoviens n'avait encore jamais été proposé.

Dans le chapitre suivant nous allons présenter un outil graphique associé à ces méthodes et montrer des applications sur plusieurs jeux de données.

Chapitre 3.

Application à des études DTS

Les méthodes présentées dans le chapitre précédent ont été implémentées sous R (R Development Core Team, 2019) et les parties nécessitant le plus de calculs en C++ à l'aide de la librairie RCpp (Eddelbuettel & Balamuta, 2018). Pour illustrer ces méthodes, nous proposons d'utiliser ces nouveaux outils en les appliquant à des jeux de données DTS. Les 4 jeux de données utilisés sont issus d'études concernant des chocolats, des fromages frais et des Goudas.

Dans un premier temps, nous allons nous intéresser à l'adéquation du modèle proposé aux données DTS. Nous testerons l'hypothèse markovienne, puis nous vérifierons que les durées de dominance sont bien distribuées selon une loi Gamma et nous comparerons les données simulées aux données sur lesquelles le modèle a été estimé. Dans un deuxième temps nous présenterons le graphe DTS, une représentation graphique des matrices de transition basée sur le modèle semi-markovien, et nous comparerons ce graphe aux courbes DTS. Nous présenterons ensuite des résultats pour le découpage de la durée de dégustation en périodes avec la détermination du nombre de périodes et de la position optimale des frontières entre périodes. Par la suite, nous analyserons les résultats obtenus pour le test de différence entre produits et nous utiliserons également ce test pour étudier l'existence de différences de perception entre différentes populations constituant un panel. Finalement, nous nous intéresserons à la segmentation du panel basée sur la perception et son incidence sur les notes hédoniques moyennes observées dans les différents segments.

3.1 Présentation des jeux de données

3.1.1 Chocolats Lindt Excellence

Cette étude DTS sans répétition regroupe 4 chocolats de la gamme Lindt Excellence : Subtil, Intense, Puissant, Prodigieux. Le Subtil et l'Intense sont des chocolats avec 70% de cacao, le Puissant a 85% de cacao et le Prodigieux a 90% de cacao.

Ces chocolats ont été dégustés sans entraînement préalable par 106 consommateurs qui avaient à leur disposition 10 descripteurs : Acide, Amer, Astringent, Cacao, Collant, Croquant, Fondant, Gras, Sec, Sucré.

3.1.2 Chocolats Barry Callebaut

Ces chocolats ont été gracieusement offerts par l'entreprise Barry Callebaut afin de servir de produits lors de séances de dégustations réalisées à l'occasion de différents évènements tels que des cours ou des congrès.

L'étude concerne 5 chocolats dénommés 811NV (54% de cacao), Q65MAD (65% de cacao), Q68BRA (68% de cacao), R731EQU (73% de cacao), SAO THOME (70% de cacao).

184 panélistes sans entrainement préalable ont dégusté une fois chacun de ces 5 chocolats en utilisant la méthodologie DTS avec 12 descripteurs : Acide, Amer, Astringent, Boisé, Cacao, Collant, Floral, Fondant, Fruité, Gras, Sec, Sucré.

3.1.3 Fromages frais

Cette étude (Thomas et al., 2015) proposait d'une part de mesurer l'appréciation hédonique de manière temporelle, c'est-à-dire tout au long de la dégustation, et d'autre part la perception temporelle à l'aide de la DTS afin de pouvoir calculer les « Déterminants Temporels du Liking » (DTL ou TDL en anglais pour « Temporal Drivers of Liking ») qui déterminent le lien entre descripteurs et appréciation hédonique. Pour cette étude, 68 consommateurs français ont participé à 4 expérimentations en laboratoire. Durant la première session, ils ont goûté 6 fromages frais aromatisés pour lesquels ils devaient donner une note hédonique globale. Lors de la deuxième session, ils ont noté dynamiquement leur appréciation tout au long de la dégustation des 6 mêmes produits. La troisième session utilisait le protocole DTS et la quatrième consistait à évaluer ces produits 1 minute après les avoir goûtés. Nous nous intéresserons ici principalement aux résultats de la troisième session à laquelle 64 des panélistes ont participé. Les produits étaient 6 fromages frais aromatisés représentatifs du marché français nommés P1, P2,

P3, P4, P5 et P6. Le protocole DTS comprenait 8 descripteurs : Ail, Crème, Herbes fraîches, Herbes cuites, Acre, Poivre, Sel, Acide.

3.1.4 Goudas

Cette étude menée par l'European Sensory Network (ESN) visait à tester un nouveau protocole permettant de réaliser simultanément la DTS et l'évaluation hédonique temporelle d'un produit (Thomas et al., 2017). Elle a été réalisée dans 6 pays européens par : 117 allemands, 112 français, 105 polonais, 116 portugais, 100 hongrois et 117 britanniques. En tout 667 consommateurs ont testé 4 Goudas différents par leurs durées de maturation (4 semaines, 7 semaines ou 16 semaines) et leurs taux de matière grasse (normal = 48% ou faible en matière grasse = 30%), en utilisant le protocole DTS pour 3 bouchées successives avec 10 descripteurs : Mou, Fromage, Gras, Amer, Fondant, Crémeux/lacté, Dense/dur, Acide, Piquant, Salé. Le premier produit, un Gouda de 7 semaines avec un taux normal de matière grasse, a été utilisé uniquement comme échauffement et ne sera pas étudié ici. Les trois autres produits seront désignés dans la suite par les codes suivants : 16wk30 est un Gouda mature avec un faible taux de matière grasse, 4wk30 est un Gouda jeune avec un faible taux de matière grasse et 4wk48 est un Gouda jeune avec un taux normal de matière de grasse. Les 3 prises successives sont considérées dans la suite comme 3 répétitions indépendantes.

3.1.5 Utilisation des jeux de données

Nous avons fait le choix d'illustrer chaque problématique par une application sur au moins 2 jeux de données. La liste des jeux de données utilisées pour chaque problème est la suivante :

- Adéquation aux données DTS : Goudas, Chocolats Lindt Excellence ;
- Graphe DTS : Chocolats Barry-Callebaut, Chocolats Lindt Excellence ;
- Découpage en périodes : Chocolats Lindt Excellence, Fromages frais ;

- Test de différence : Chocolats Lindt Excellence, Fromages frais, Goudas ;
- Segmentation : Goudas, Fromages frais.

3.2 Adéquation du modèle aux données DTS

3.2.1 Hypothèse markovienne

Nous proposons ici d'utiliser un test statistique pour prouver qu'une chaîne de Markov est plus adaptée qu'une modélisation sans tenir compte du précédent descripteur. Ce test consiste à comparer à l'aide d'un test du rapport de vraisemblance le modèle markovien de paramètres $\mathbb{P}$ à un modèle de paramètres p qui estime simplement les fréquences d'observations de chacun des descripteurs indépendamment de l'état précédent (Guttorp, 1995). La statistique de test, qui est le log du rapport de vraisemblance $2\big(logL(\widehat{\mathbb{P}}) - logL(\hat{p})\big)$, est distribuée selon une loi du Khi2 avec un nombre de degrés de libertés égal à la différence de nombre de paramètres entre les deux modèles.

Tableau 1 Nombre de transitions observées et entre parenthèse nombre de transitions théorique qui devrait être observé si le choix du prochain descripteur dominant était indépendant du descripteur actuellement dominant. Ces résultats sont observés pour le Gouda 16wk30.

	Amer	**Fromage**	**Dense Dur**	**Gras**	**Fondant**	**Lacté Crémeux**	**Salé**	**Piquant**	**Acide**	**Tendre**
Amer		115 (106)	22 (28)	55 (55)	44 (69)	38 (45)	120 (105)	82 (81)	91 (68)	23 (33)
Fromage	112 (131)		55 (47)	77 (91)	105 (114)	108 (75)	204 (174)	141 (134)	85 (113)	51 (55)
Dense dur	151 (145)	204 (197)		124 (101)	119 (127)	48 (83)	202 (194)	158 (150)	122 (126)	59 (62)
Gras	64 (75)	91 (101)	26 (27)		67 (65)	43 (43)	100 (99)	89 (77)	63 (65)	40 (33)
Fondant	73 (89)	123 (121)	21 (32)	56 (62)		55 (51)	118 (119)	106 (92)	59 (77)	71 (38)
Lacté Crémeux	31 (48)	106 (65)	13 (17)	35 (34)	46 (42)		57 (64)	29 (50)	47 (42)	21 (20)
Salé	159 (119)	187 (162)	46 (43)	52 (83)	100 (105)	64 (68)		117 (123)	105 (104)	31 (51)
Piquant	111 (107)	149 (145)	61 (39)	69 (74)	88 (93)	46 (61)	126 (142)		105 (93)	44 (45)
Acide	96 (74)	86 (100)	31 (27)	38 (51)	49 (65)	45 (42)	121 (98)	81 (76)		17 (31)
Tendre	44 (72)	80 (98)	30 (26)	80 (50)	119 (63)	35 (41)	77 (96)	66 (74)	53 (63)	

Dans le Tableau 1, pour le Gouda 16wk30, nous observons des différences importantes entre le nombre de transitions observé et le nombre théorique sous l'hypothèse d'indépendance au descripteur dominant précédent. Nous pouvons entre autre constater que le choix de « Lacté crémeux » semble beaucoup plus probable si le descripteur actuellement dominant est « Fromage ». Pour ce produit, la statistique de test vaut 415,44. Le modèle markovien est paramétré par les probabilités de transition d'un descripteur à un autre. Il y a ici 10 descripteurs donc il y a $10 \times 10 = 100$ paramètres mais il est impossible dans ce modèle de rester dans le même état donc nous pouvons retirer 10 paramètres correspondant à la diagonale de la matrice de transitions. De plus, chaque ligne de la matrice de transition se sommant à 1, il suffit de connaître 8 des paramètres d'une ligne pour pouvoir en déduire les 9. Ainsi la chaîne de Markov a ici $10 \times 8 = 80$ paramètres (nous ne prenons pas en compte les probabilités initiales pour réaliser ce test). Le deuxième modèle est composé de 10 paramètres, correspondant à la probabilité d'observer chaque descripteur, mais ces 10 paramètres se sommant à 1, il n'est nécessaire d'en estimer que 9. Le rapport de vraisemblance est donc estimé par une loi du Khi2 à $80 - 9 = 71$ degrés de liberté. La statistique de test correspond à une p-value inférieure à 10^{-15}. Le test rejette l'hypothèse selon laquelle le choix de descripteur dominant est indépendant du descripteur actuellement dominant.

Nous avons observé pour l'ensemble des jeux de données DTS testés qu'une modélisation prenant en compte le précédent descripteur était toujours la plus adaptée.

3.2.2 Distribution des temps de séjour

Une hypothèse importante de notre modèle est que les durées de dominance ne sont pas distribuées selon une loi exponentielle.

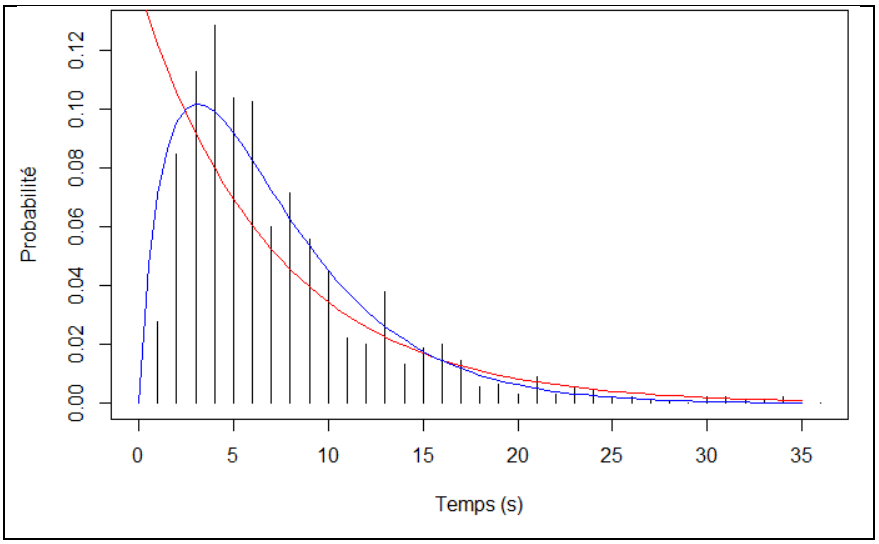


Figure 19 Distribution des durées de dominance pour le descripteur Amer et le produit 16wk30 avec en rouge l'estimation par une loi exponentielle et en bleu l'estimation par une loi Gamma.

L'exemple donné dans la Figure 19 est représentatif de la forme de distribution généralement observée pour les durées de dominance pour un descripteur d'une étude DTS. Nous observons que l'estimation par une loi Gamma est bien plus adaptée que l'estimation par une loi exponentielle notamment parce que la loi exponentielle surestime les probabilités d'observer des durées de dominance très courtes.

Tableau 2 P-values du test d'adéquation de Kolmogorov-Smirnov pour l'estimation des durées de dominance par une loi Gamma pour les différents descripteurs et les 3 Goudas de l'étude ESN.

	16wk30	4wk30	4wk48
Amer	0,022	0,046	0,244
Fromage	0,004	0,020	0,001
Dense dur	0,091	0,003	0,034
Gras	0,142	0,039	0,130
Fondant	0,084	0,066	0,008
Lacté crémeux	0,099	0,412	0,012
Salé	0,017	0,013	0,001
Piquant	0,001	0,017	0,004
Acide	0,045	0,045	0,420
Tendre	0,635	0,159	0,296

Nous vérifions cette hypothèse en réalisant un test de Kolmogorov-Smirnov pour vérifier si les durées de dominance sont distribuées selon une loi exponentielle ou une loi Gamma. Pour les 3 Goudas que nous prenons en exemple, le test de Kolmogorov-Smirnov donne des p-values inférieures à 10^{-15} pour la loi exponentielle quel que soit le descripteur et quel que soit le produit ce qui montre que la loi exponentielle n'est pas adaptée à ces données. Le Tableau 2 nous montre les p-values du test d'adéquation à une loi Gamma pour ces 3 Goudas et l'ensemble des descripteurs. Les p-values sont toutes supérieures à 10^{-3}, donc toujours meilleures que pour la loi exponentielle, et pour beaucoup l'hypothèse H_0 ne peut être rejetée en prenant un seuil α à 1% ce qui signifie que la modélisation par une loi Gamma est adaptée à ces données. Si certaines p-values sont tout de même assez faibles, il ne faut pas oublier que les descripteurs peuvent être très peu utilisés selon les produits ce qui complique l'estimation avec parfois un nombre faible d'observations.

3.2.3 Comparaison données simulées et données réelles

Afin de se faire une idée de la qualité de la modélisation par une chaîne semi-markovienne, nous réalisons des simulations de données à partir du modèle et comparons les courbes DTS des données simulées aux courbes DTS des données d'origine pour les chocolats Excellence.

La comparaison des courbes DTS calculées à partir des données d'origine et des données simulées (Figure 20) montre que les simulations sont proches mais que des petites différences existent avec notamment des pics de dominance qui se produisent à des moments différents de la dégustation. Par exemple, pour les chocolats Puissant et Prodigieux, le descripteur Astringent dépasse la significativité au milieu de la dégustation dans les données simulées mais à la fin pour les données d'origine. Ces décalages montrent l'intérêt de découper la durée de dégustation en périodes pour améliorer la qualité de la modélisation.

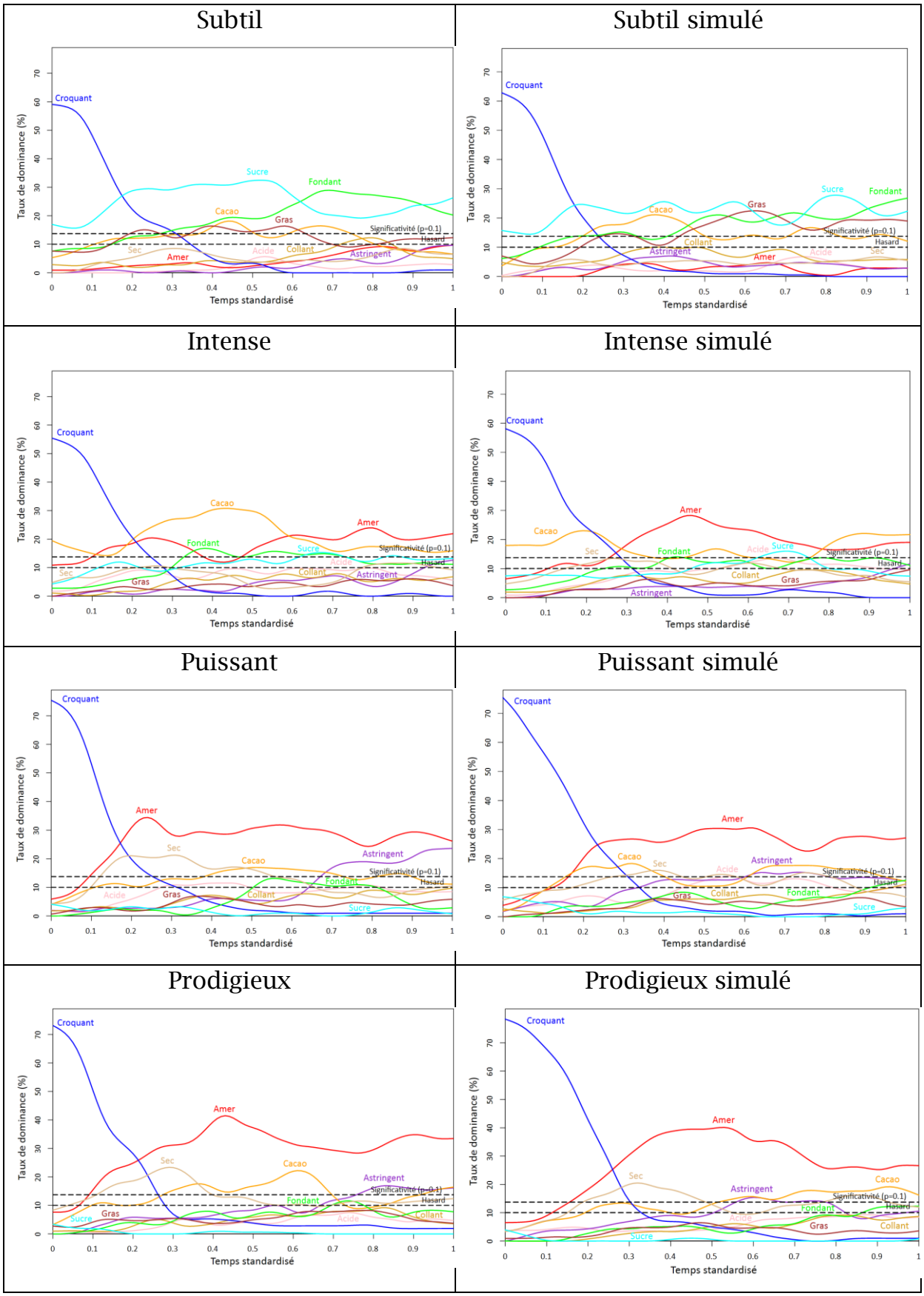


Figure 20 Courbes DTS standardisées des chocolats Excellence et des données simulées à partir des modèles estimés pour ces mêmes chocolats.

3.3 Graphe DTS

La chaîne de Markov modélisant les changements de descripteur est caractérisée par un grand nombre de paramètres composés des probabilités initiales et de la matrice de transition (exemple Tableau 3).

Tableau 3 Matrice de transitions estimée pour le chocolat 811NV et probabilité initiales (première ligne).

	Acide	Amer	Astringent	Boisé	Cacao	Collant	Floral	Fondant	Fruité	Gras	Sec	Sucré	STOP
START	0	0,033	0	0,043	0,207	0,027	0,022	0,049	0,103	0,043	0,179	0,293	0
Acide	0	0,167	0	0,167	0,500	0	0	0	0,167	0	0	0	0
Amer	0,062	0	0	0,062	0,281	0,219	0	0,031	0,062	0	0	0,188	0,094
Astringent	0,059	0	0	0,059	0,176	0	0,118	0	0,118	0	0,118	0,294	0,059
Boisé	0	0	0	0	0,250	0,042	0,042	0,104	0,042	0,125	0,042	0,229	0,125
Cacao	0	0,031	0,013	0,049	0	0,054	0,072	0,139	0,076	0,081	0,036	0,359	0,090
Collant	0	0,068	0,027	0,041	0,178	0	0	0,068	0,027	0,137	0	0,329	0,123
Floral	0	0,040	0	0	0,093	0	0	0,067	0,040	0,147	0,080	0,347	0,187
Fondant	0,013	0,013	0,013	0,040	0,192	0	0,073	0	0,106	0,066	0,013	0,331	0,139
Fruité	0	0	0,008	0,032	0,152	0,064	0,080	0,136	0	0,048	0,016	0,336	0,128
Gras	0	0	0,009	0,018	0,155	0,082	0,018	0,227	0,073	0	0,036	0,273	0,109
Sec	0,014	0,028	0,056	0,042	0,264	0,028	0,069	0,083	0,153	0,028	0	0,194	0,042
Sucré	0	0.018	0.012	0.020	0.158	0.082	0.067	0.137	0.123	0.114	0.038	0	0,231

Ce modèle est de fait difficile à interpréter, c'est pourquoi nous proposons une représentation graphique simplifiée pour en faciliter l'interprétation.

3.3.1 Présentation du graphe DTS

Pour représenter graphiquement le modèle nous proposons d'utiliser le graphe de Markov. Celui-ci étant peu lisible avec une dizaine d'états et le grand nombre de flèches représentant les transitions se chevauchant, nous proposons de le simplifier. Pour des raisons de lisibilité nous ne conservons dans le graphe que les informations les plus importantes en utilisant deux paramètres : le seuil de sélection des descripteurs *Dseuil* et le seuil de sélection des transitions *Tseuil*. Le seuil de sélection des descripteurs permet de définir quels descripteurs sont présents dans le graphe en ne sélectionnant que ceux ayant été utilisés par au moins un pourcentage *Dseuil* des panélistes. Nous utilisons 50% comme valeur par défaut pour ce seuil. Le seuil de sélection des transitions entre descripteurs permet de sélectionner quelles transitions sont affichées dans le graphe en fonction de leur

probabilité. La valeur par défaut que nous utilisons est 0,15. Les valeurs des 2 seuils peuvent être adaptées aux données, à leur complexité et également au degré de précision voulu du graphe, mais elles doivent être identiques pour tous les produits d'une étude pour qu'une comparaison des graphes reste possible. Nous enrichissons le graphe en ajoutant dans chaque bulle/descripteur le pourcentage de panélistes ayant utilisé le descripteur. Nous verrons plus loin comment le graphe peut être adapté dans le cas où la durée de dégustation est découpée en périodes.

L'implémentation sous R du graphe DTS a été réalisée par Rihab Boubakri lors de son stage de fin d'études (Boubakri, Lecuelle, Schlich, Visalli, & Ben Hassine, 2017). Pour permettre l'obtention d'un graphe le plus lisible possible, le graphe réalisé sous R est interactif de manière à ce que la disposition des différents descripteurs puisse être modifiée par l'utilisateur.

3.3.2 Exemple avec les chocolats BC

Nous allons observer les graphes DTS et les comparer aux courbes DTS pour les chocolats BC.

Tableau 4 Pourcentage de panélistes ayant utilisé chacun des descripteurs pour les 5 produits du jeu de données BC.

	811NV	Q65MAD	Q68BRA	R731EQU	SAO THOME
Acide	1,63	48,91	32,61	10,33	30,98
Amer	13,04	69,57	77,17	61,96	48,91
Astringent	8,70	32,61	37,50	27,72	26,63
Boisé	20,65	35,33	59,24	39,67	38,04
Cacao	69,57	71,74	70,11	76,09	80,98
Collant	29,89	30,98	33,15	25,00	39,13
Floral	29,89	13,04	15,22	15,22	14,67
Fondant	58,15	25,00	38,04	45,65	46,20
Fruité	43,48	36,96	19,57	19,02	23,37
Gras	42,93	34,24	30,98	32,61	42,93
Sec	29,89	37,50	38,59	42,39	30,98
Sucré	88,59	29,89	25,00	35,33	46,74

Dans un premier temps nous transformons les données DTS pour ne nous intéresser qu'au fait que les descripteurs soient utilisés ou non par les panélistes (Tableau 4) afin de déterminer le seuil de sélection des descripteurs.

En conservant la valeur par défaut $Dseuil = 50\%$, des informations importantes ne seront pas présentes dans le graphe telles que la présence du descripteur Acide uniquement pour le Q65MAD, du descripteur Fruité pour le 811NV ou Gras pour le 811NV et le SAO THOME. Afin de prendre en compte ces informations qui constituent des différences importantes entre les chocolats nous fixons le seuil *Dseuil* à 40%.

Etant donné le nombre de descripteurs plutôt élevé dans cette étude les probabilités de transitions sont mécaniquement plutôt faibles donc nous proposons également de modifier la valeur de *Tseuil* en la fixant à 12%.

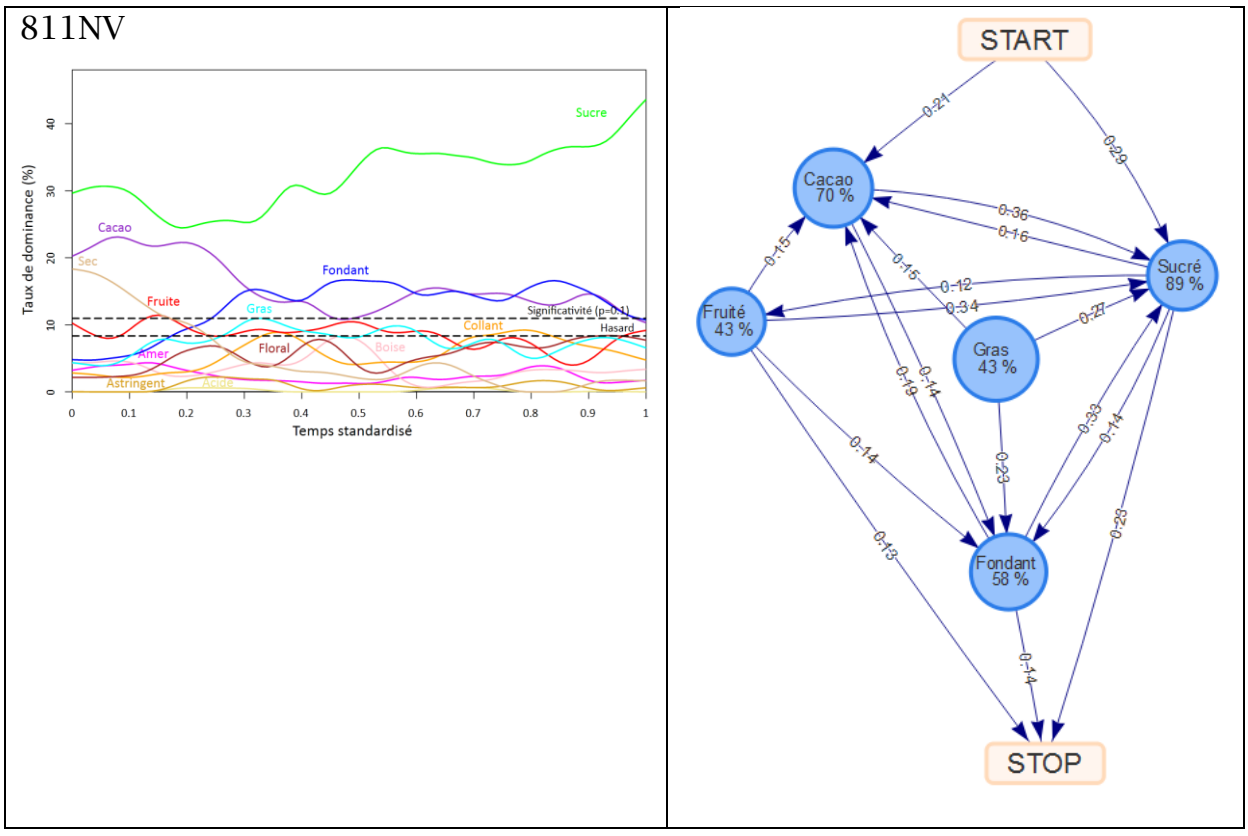


Figure 21 Courbes DTS et graphe DTS avec Dseuil=40% et Tseuil=12% du chocolat 811 NV.

Pour le 811NV (Figure 21), nous observons que les 3 descripteurs les plus importants dans les courbes, Sucré, Cacao et Fondant, sont présents dans le graphe et sont les plus utilisés par les panélistes. Le descripteur Sec qui est

significativement dominant au début de la dégustation dans les courbes est absent dans le graphe puisqu'il n'a été utilisé que par 29,89% des panélistes. Les descripteurs Gras et Fruité sont proches de la ligne de significativité sans la dépasser tout au long de la sélection mais sont présents dans le graphe. Ces 2 descripteurs sont utilisés par plus de 40% des panélistes pourtant ils ne semblent pas important dans les courbes car il n'y a pas d'accord sur le moment de citation.

Le graphe nous apprend que 29% des panélistes choisissent Sucré et 21% Cacao comme premier descripteur. Ensuite le graphe se lit de la façon suivante : par exemple, parmi les panélistes qui ont choisi Cacao, 36% ont ensuite choisi Sucré et 14% ont choisi Fondant. Il y a une boucle Cacao, Sucré et Fondant, c'est-à-dire des flèches aller-retour entre chaque duo de descripteurs parmi ces trois, montrant que ces trois descripteurs sont importants mais que les panélistes passent de l'un à l'autre sans logique temporelle. Après les deux descripteurs Gras et Fruité, les panélistes choisissent aussi majoritairement Cacao, Sucré ou Fondant. Dans tous les cas la probabilité la plus élevée est celle d'aller vers Sucré ce qui se traduit par une sur-dominance de ce descripteur dans les courbes DTS. Finalement la dégustation se termine principalement par Sucré, Fruité ou Fondant.

Il est intéressant d'observer qu'aucune transition ne va vers Gras dans le graphe et qu'il n'y a qu'une transition avec une probabilité faible qui va vers Fruité ce qui signifie que ces choix de descripteurs se font indépendamment du descripteur actuellement dominant. Ils ne sont donc pas inclus dans une séquence temporelle ce qui peut signifier que ces 2 descripteurs n'ont pas de temporalité pour ce produit, ce qui est exprimé par des courbes DTS plates et tout juste à la limite de la significativité.

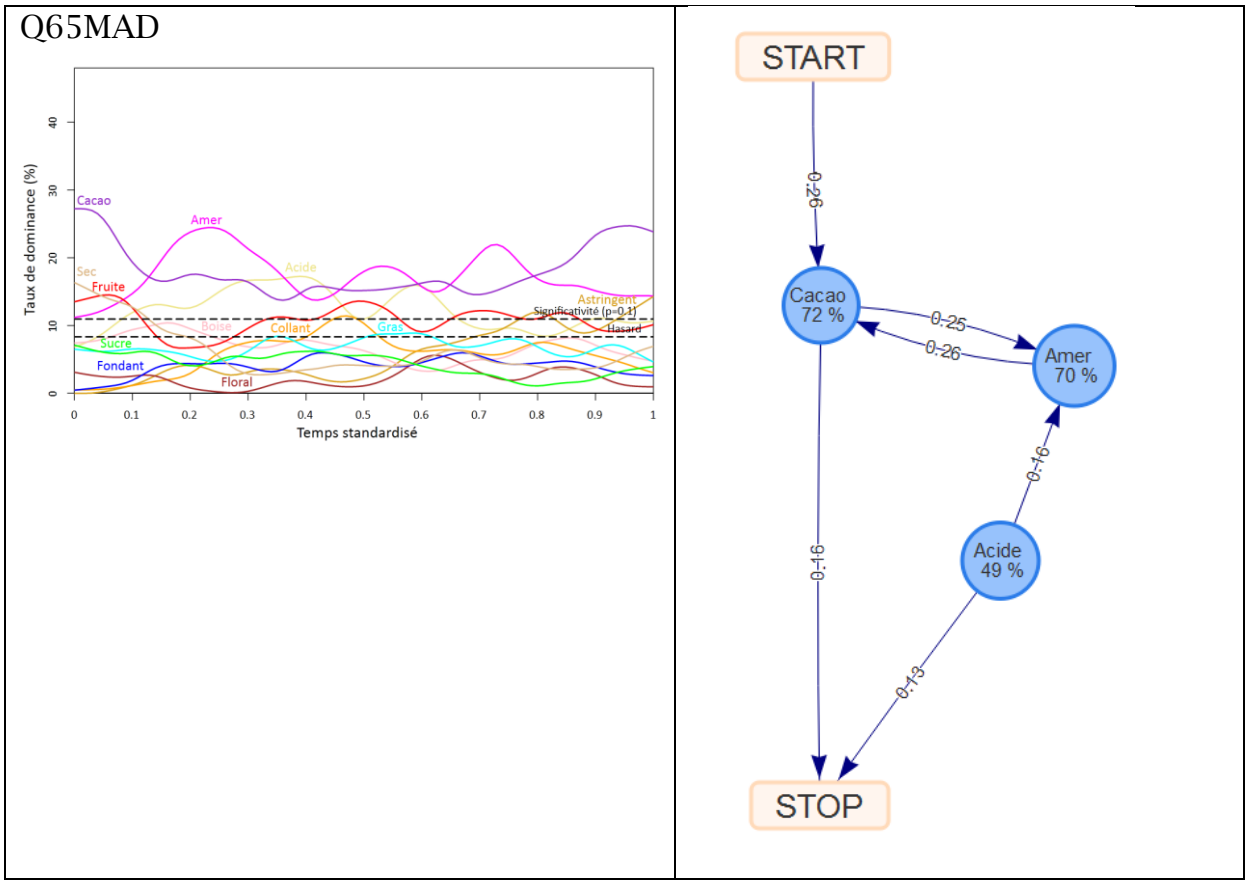


Figure 22 Courbes DTS et graphe DTS avec Dseuil=40% et Tseuil=12% du chocolat Q65MAD.

Pour le chocolat Q65MAD (Figure 22), 6 descripteurs dépassent la ligne de significativité : Cacao, Sec, Fruité, Amer, Acide et Astringent. Parmi ces 6 descripteurs seuls Cacao, Amer et Acide apparaissent dans le graphe. Les courbes de ces 3 descripteurs sont celles qui dépassent clairement le seuil de significativité alors que les courbes des 3 autres descripteurs n'atteignent la significativité que brièvement et ne la dépassent que de peu.

Le graphe nous permet de savoir que Cacao a été choisi comme premier descripteur par 26% des panélistes et qu'il est associé ensuite à Amer. Le descripteur Acide est également important pour ce produit mais son choix est indépendant du descripteur actuellement dominant. Parmi les panélistes qui ont choisi Acide comme dominant, la principale tendance est ensuite de choisir Amer. Finalement la perception de ce chocolat se termine soit par Cacao soit par Acide.

Pour ce chocolat, la lecture du graphe est plus simple que celle des courbes qui sont assez confuses.

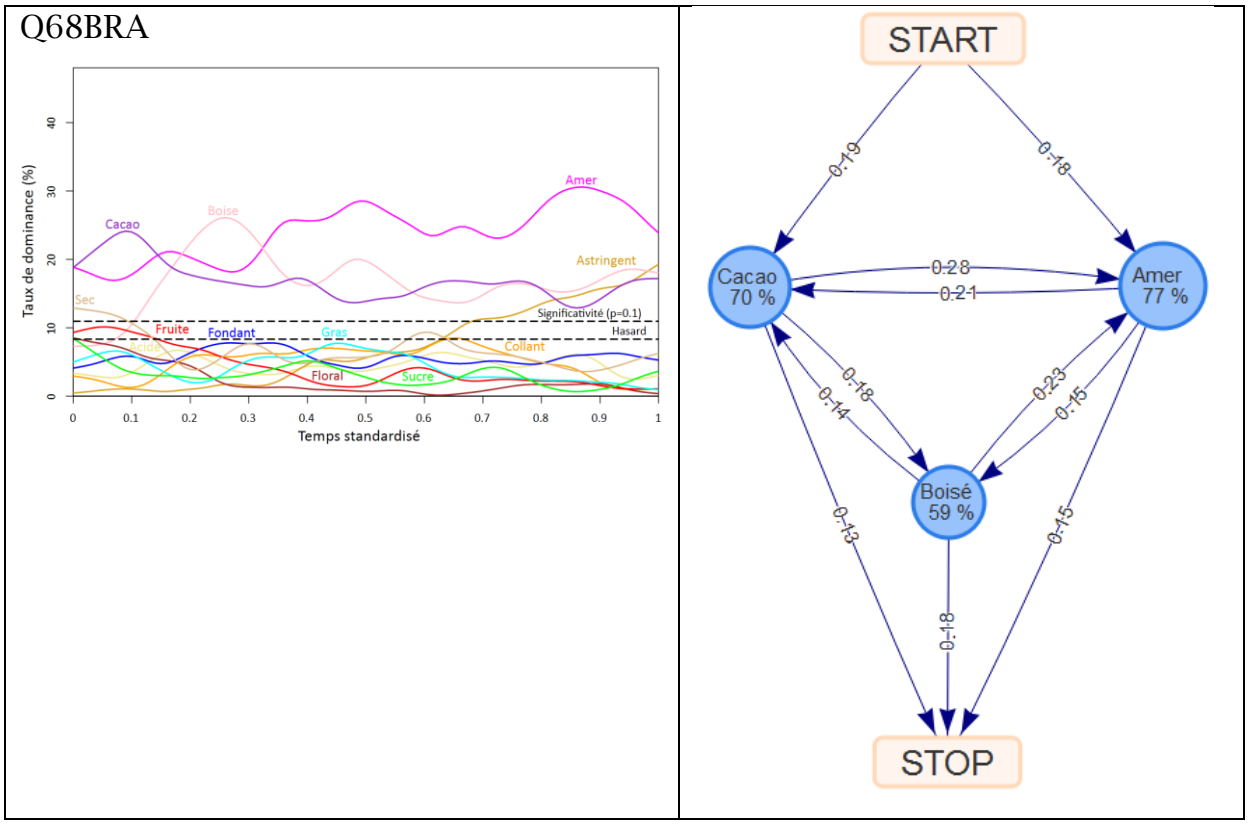


Figure 23 Courbes DTS et graphe DTS avec Dseuil=40% et Tseuil=12% du chocolat Q68BRA.

Les courbes DTS du chocolat Q68BRA (Figure 23) nous laissent penser qu'il y a une séquence temporelle Cacao, Boisé et Amer avec également Astringent qui fait son apparition à la fin.

En regardant le graphe DTS, les panélistes ont choisi de manière équiprobable soit Cacao soit Amer comme premier descripteur avant d'entrer dans une phase sans temporalité où les panélistes alternent entre Cacao, Amer et Boisé avec des probabilités tout de même plus élevées d'aller vers Amer. La perception se termine ensuite par l'un de ses 3 descripteurs.

Pour cet exemple, il nous semble que le graphe permet de souligner le seul véritable élément de temporalité dans la perception de ce produit à savoir que le boisé arrive après le Cacao et/ou l'Amer.

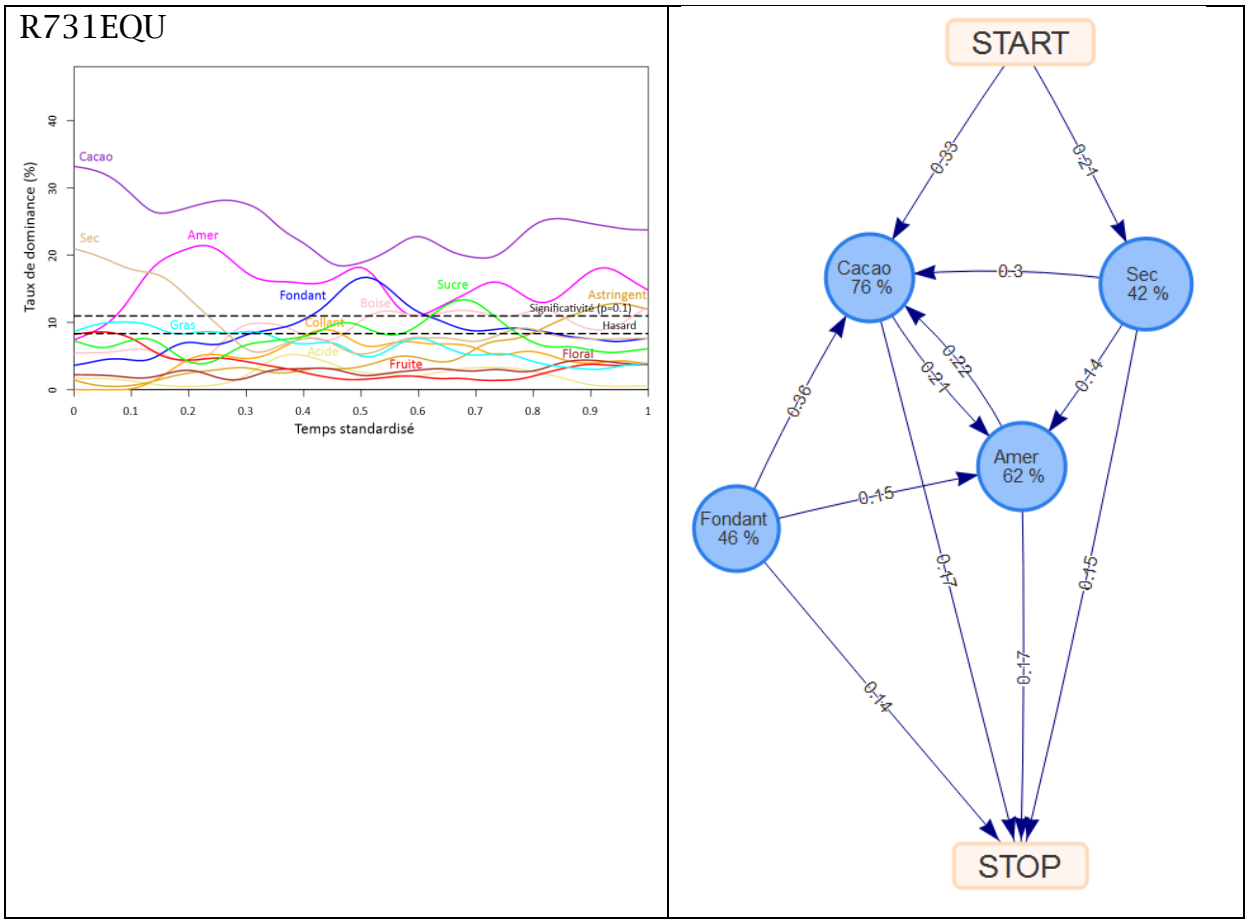


Figure 24 Courbes DTS et graphe DTS avec Dseuil=40% et Tseuil=12% du chocolat R731EQU.

Les courbes DTS du chocolat R731EQU (Figure 24) nous montrent que ce chocolat est perçu Cacao ou Sec puis Amer, Fondant, un peu Boisé et Sucré pour finir par Cacao ou Amer et légèrement Astringent. En regardant maintenant le graphe DTS, nous constatons que Boisé, Sucré et Astringent, qui sont très légèrement significatifs dans les courbes, n'apparaissent pas dans le graphe. Les panélistes ont choisis Cacao ou Sec comme premier descripteur. Il n'y a ensuite plus de transitions vers Sec ce qui indique que ce descripteur n'est présent qu'en début de dégustation. Le choix de Fondant est indépendant du descripteur dominant actuel. Le duo Cacao-Amer est associé et est au cœur de la perception. Finalement la perception peut se finir par chacun des 4 descripteurs présents dans le graphe.

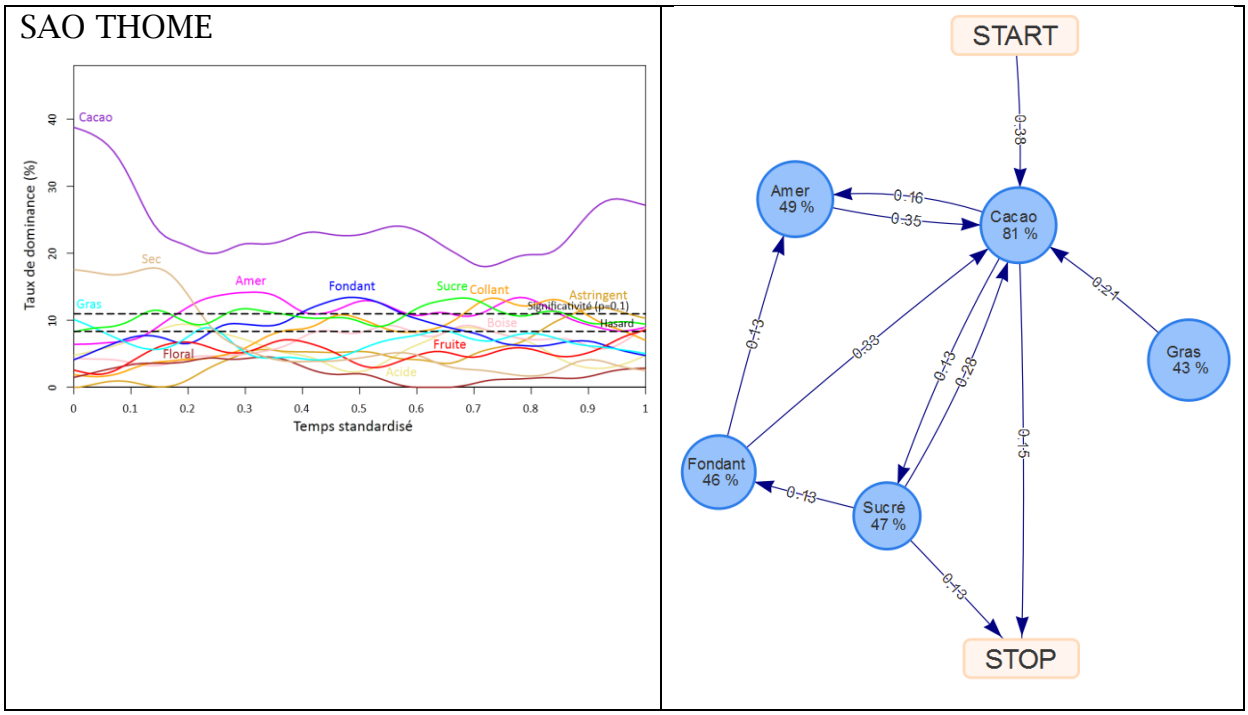


Figure 25 Courbes DTS et graphe DTS avec Dseuil=40% et Tseuil=12% du chocolat SAO THOME.

Les courbes DTS du chocolat SAO THOME (Figure 25) indiquent que les panélistes ont principalement trouvé que le Cacao était dominant. Pour les autres descripteurs les courbes sont plus confuses avec dans l'ordre mais à des taux de dominance restant assez faibles l'apparition des descripteurs suivant : Sec, Amer, Fondant, Sucré, Collant et Astringent. Le graphe DTS nous indique quant à lui que Cacao est effectivement le descripteur le plus important pour ce produit avec 81% des panélistes qui l'ont utilisé et 38% qui l'ont choisi comme premier descripteur dominant. De plus, les probabilités de transition les plus élevées sont toutes en direction de Cacao et tous les autres descripteurs ont une flèche vers Cacao. Les autres descripteurs présents qui ont tous été utilisés par entre 40 et 50% des panélistes sont Amer, Sucré, Fondant et Gras. Comme pour le 811NV il n'y a aucune flèche vers Gras, montrant que ce choix est indépendant de l'état actuel. Les duos Amer-Cacao et Sucré-Cacao ont des transitions dans les deux sens montrant que ces descripteurs sont liés mais que l'ordre n'a sans doute pas d'importance. Au contraire, la transition ne se fait que dans un sens pour la suite de descripteurs Sucré, Fondant puis Amer. Nous ne pensons pas que ces subtilités puissent être vues dans les courbes DTS.

3.3.3 Exemple avec les chocolats Excellence

Nous allons maintenant observer les courbes et les graphes DTS obtenus pour les chocolats Lindt Excellence.

Dans un premier temps, et comme précédemment, nous transformons les données DTS en données CATA (Tableau 5) pour déterminer le seuil de sélection des descripteurs.

Tableau 5 Pourcentage de panelistes ayant utilisés chacun des descripteurs pour les 4 chocolats Excellence.

	Subtil	Intense	Puissant	Prodigieux
Acide	12,26	38,68	44,34	23,58
Amer	23,58	68,87	83,02	85,85
Astringent	15,09	29,25	48,11	42,45
Cacao	57,55	85,85	70,75	63,21
Collant	28,30	33,96	37,74	30,19
Croquant	65,09	62,26	75,47	74,53
Fondant	77,36	62,26	37,74	41,51
Gras	50,00	26,42	27,36	29,25
Sec	17,92	37,74	54,72	57,55
Sucré	81,13	48,11	15,09	6,60

En observant les pourcentages de panélistes ayant utilisé chaque descripteur nous constatons qu'en conservant un seuil $Dseuil$ à 50% nous excluons le descripteur Astringent pour le Puissant et le Sucré pour l'Intense alors que ces deux descripteurs sont proches du seuil et importants pour caractériser ces chocolats. Nous décidons donc de baisser légèrement le seuil à 48% afin d'intégrer ces 2 descripteurs aux graphes. Pour ce jeu de données, qui utilise 10 descripteurs, la valeur par défaut $Tseuil = 0{,}15$ semble adaptée.

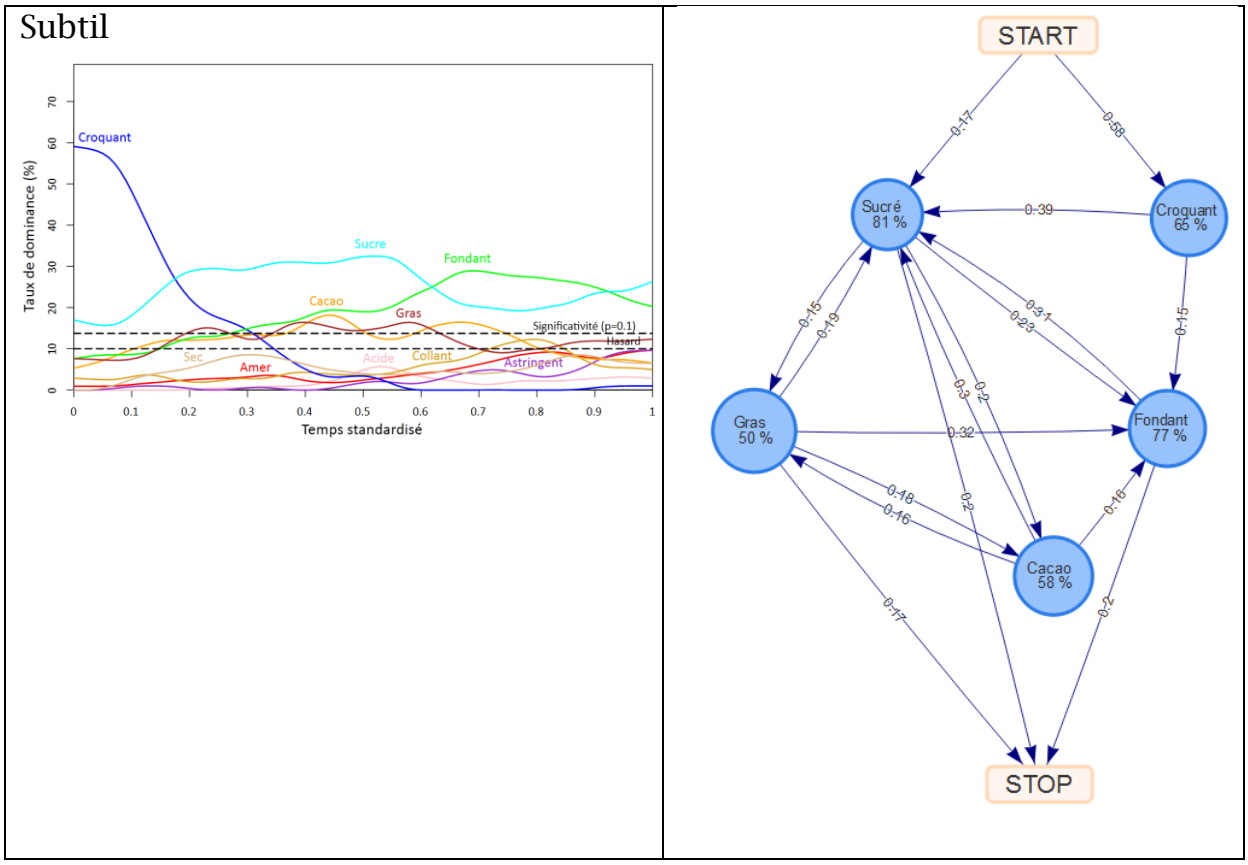


Figure 26 Courbes DTS et graphe DTS avec Dseuil=48% et Tseuil=15% du chocolat Lindt Subtil.

En observant les courbes DTS (Figure 26), nous pouvons être tentés de conclure que la perception de ce produit se limite à la séquence Croquant, Sucré, Fondant et éventuellement à nouveau Sucré. En regardant le graphe DTS et en suivant le chemin ayant les plus grandes probabilités de transition, nous retrouvons bien cette séquence mais nous constatons qu'il existe beaucoup d'autres chemins possibles dans ce graphe correspondant à différentes façons de percevoir ce produit.

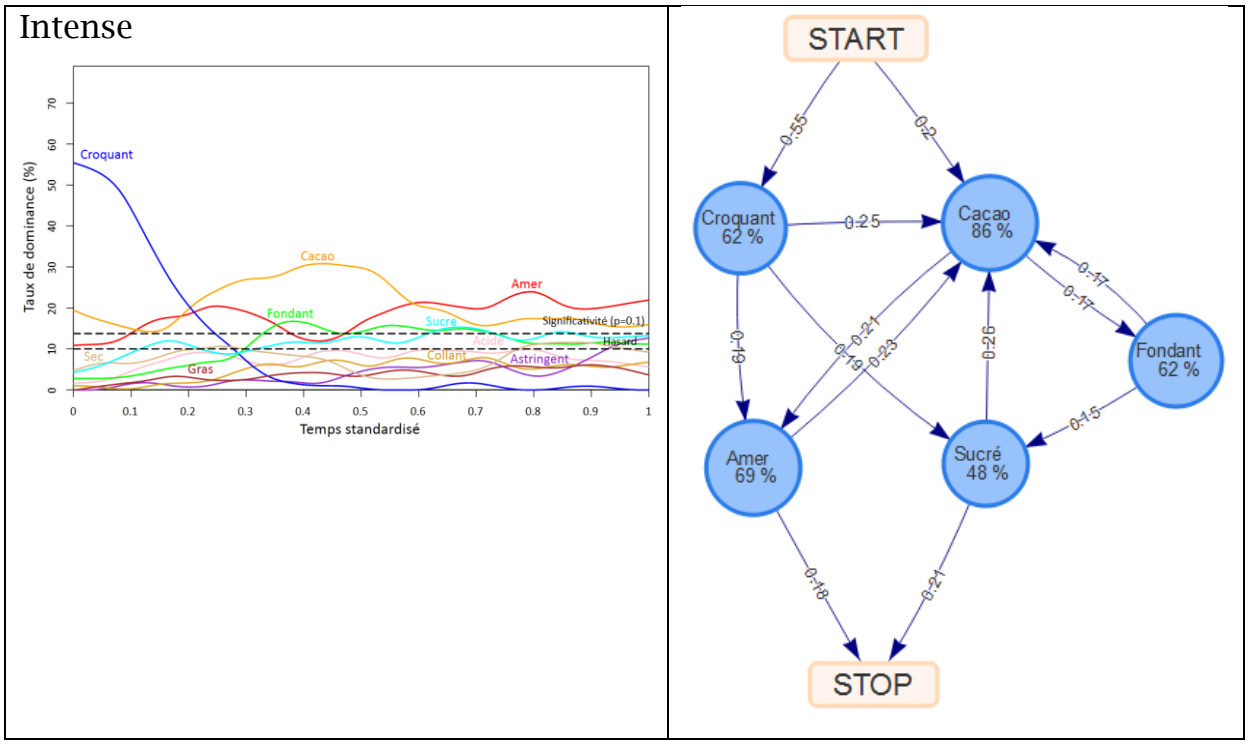


Figure 27 Courbes DTS et graphe DTS avec Dseuil=48% et Tseuil=15% du chocolat Lindt Intense.

Pour le chocolat Intense les descripteurs les plus importants sont Croquant, Cacao, Fondant, Sucré et Amer aussi bien selon les courbes que le graphe DTS (Figure 27). En observant le graphe, Croquant est utilisé uniquement en début de dégustation. Cacao est au cœur du graphe avec des flèches venant de tous les descripteurs et est suivi soit par Amer soit par Fondant. Amer est perçu après Croquant ou Cacao. Sucré est précédé par Croquant ou Fondant et suivi par Cacao. La probabilité que la perception se termine par Amer ou Sucré est supérieure à 0,15 alors qu'elle est inférieure à cette valeur pour les autres descripteurs du graphe.

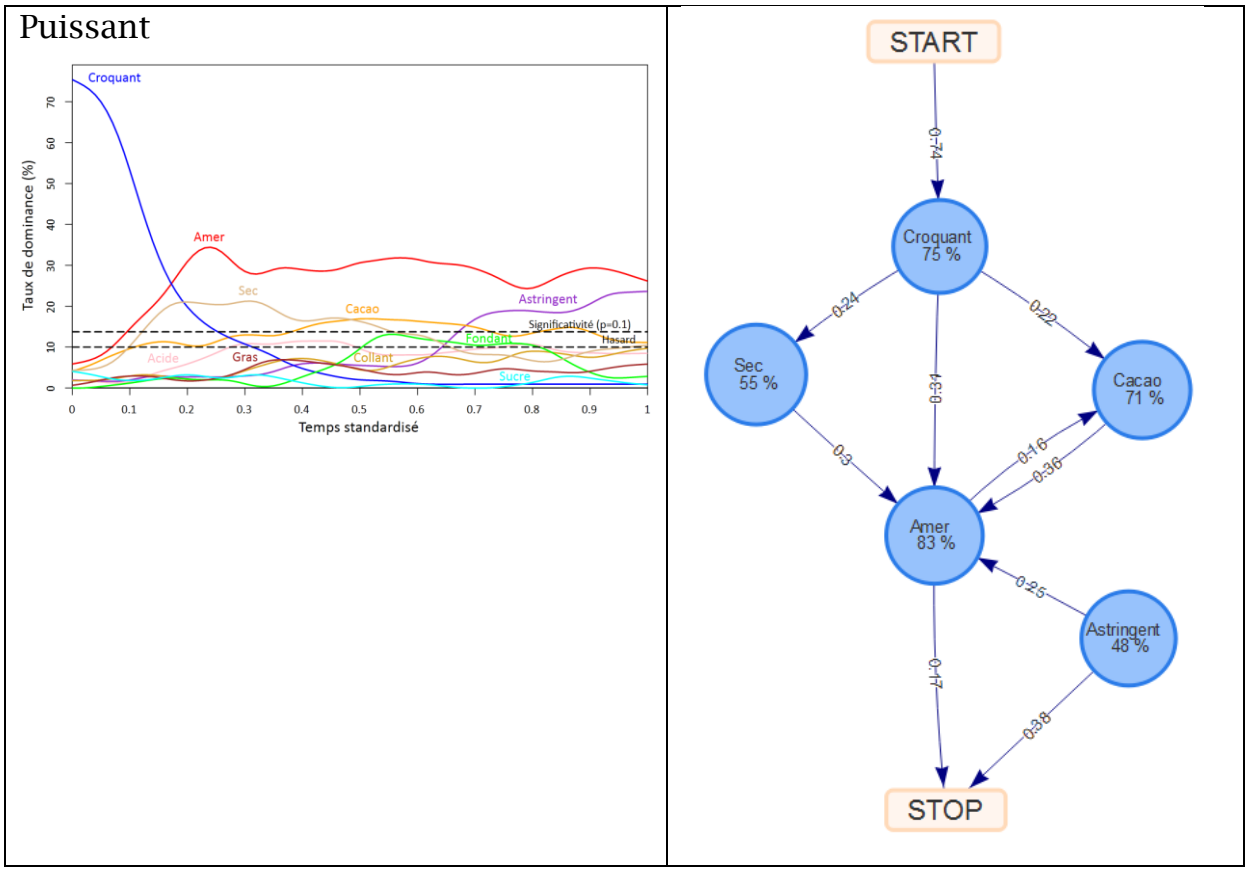


Figure 28 Courbes DTS et graphe DTS avec Dseuil=48% et Tseuil=15% du chocolat Lindt Puissant.

Selon le graphe DTS (Figure 28), le chocolat Puissant est perçu comme croquant par 71% des panélistes puis amer en passant éventuellement auparavant par Sec ou Cacao. Ce chocolat peut éventuellement être également perçu comme astringent en fin de dégustation. Nous remarquons également un lien entre les descripteurs Amer et Cacao. En regardant les courbes le descripteur Sec semble plus important que le Cacao pour ce produit alors que 71% des panelistes ont trouvé le cacao dominant alors que seulement 55% ont trouvé le descripteur Sec dominant. Cependant, les courbes nous informent que Sec est en général choisi avant Cacao, ceci est dû au fait que Cacao peut être choisi après Amer mais pas Sec, ce que le graphe montre et pas les courbes.

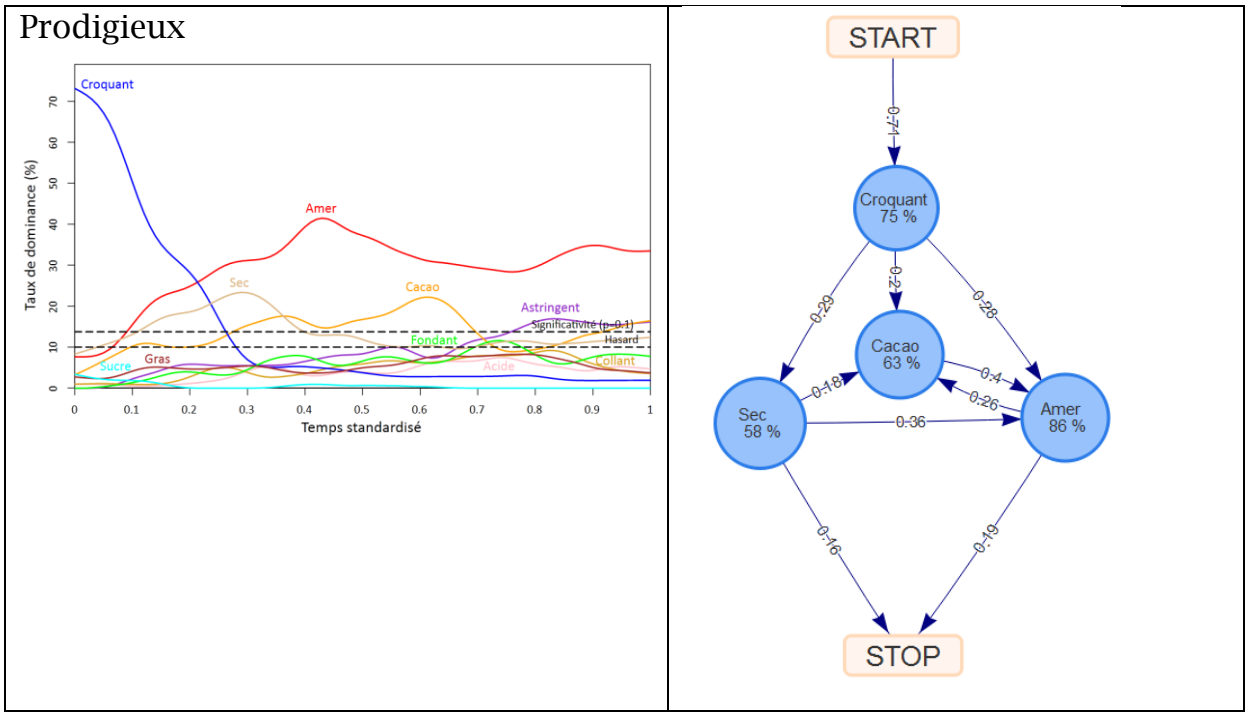


Figure 29 Courbes DTS et graphe DTS avec Dseuil=48% et Tseuil=15% du chocolat Lindt Prodigieux.

Le chocolat Prodigieux a, au regard des courbes et du graphe DTS (Figure 29), une perception très proche de celle du chocolat Puissant. Le descripteur Astringent est à un niveau plus bas dans les courbes et n'apparaît pas dans le graphe. Nous observons également qu'il y a 2 transitions supplémentaires dans le graphe depuis Sec vers Cacao et vers STOP mais elles ont des probabilités proches de *Tseuil* et ces différences sont peut être juste un effet de seuil.

3.4 Découpage en périodes

Dans cette partie nous allons, par produit, tester la nécessité de découper la durée de dégustation en périodes et, si c'est le cas, déterminer le nombre de périodes nécessaires et la position des frontières entre ces périodes. Chaque période est représentée par un graphe DTS avec $Tseuil$ conservé pour les probabilités initiales mais aussi pour les probabilités « finales », qui correspondent aux pourcentages de panélistes ayant choisi chacun des descripteurs comme dernier descripteur. Il est nécessaire d'ajouter cette information au graphe car le modèle ne prend pas en compte les transitions vers STOP pour le découpage de la durée de dégustation en périodes. Pour les transitions entre descripteurs, $Tseuil_{Decoupage}$ est fixé à $Tseuil + Tseuil \times \frac{N_p - 1}{N_p}$, où N_p est le nombre de périodes, de manière à ce qu'il soit plus exigeant quand le nombre de périodes est élevé puisque l'accord augmente lorsque le nombre de périodes augmente.

3.4.1 Chocolats Excellence

La première étape consiste à déterminer en combien de périodes la durée de dégustation doit être découpée.

Tableau 6 P-values des tests de différence entre matrices de transitions pour le découpage en périodes des chocolats Lindt Excellence.

	2 périodes	**3 périodes**		**4 périodes**		
	1ère vs 2ème	1ère vs 2ème	2ème vs 3ème	1ère vs 2ème	2ème vs 3ème	3ème vs 4ème
Subtil	0,008	0,015	0,002	0,353	0,075	0,134
Intense	0,013	$< 10^{-3}$	0,001	0,004	0,240	0,411
Puissant	0,018	0,331	0,028			
Prodigieux	0,302					

Pour les chocolats Excellence, selon le test statistique pour lequel la statistique de test est estimée en se basant sur 1000 simulations (voir 0), la durée de dégustation doit être découpée en 3 périodes pour le Subtil et l'Intense et en 2 périodes pour le Puissant en prenant un seuil $\alpha = 5\%$ (Tableau 6). Le chocolat Prodigieux ne nécessite pas de découpage temporel.

La perception du chocolat Subtil est découpée en 3 périodes avec des frontières à 55% et 80% de la durée de dégustation (Figure 30). En regardant les courbes il semblerait que la position des frontières soit principalement influencée par Sucré et Fondant. Dans la première période, les courbes de Sucré et Fondant sont légèrement croissantes puis dans la deuxième période la courbe de Sucré décroît rapidement alors que Fondant augmente rapidement. Dans la troisième partie la tendance s'inverse, la courbe de Sucré croît à nouveau alors que celle de Fondant décroît lentement. Les descripteurs Gras et Cacao sont présents dans les 2 premières périodes mais pas dans la troisième alors que Croquant n'est présent que dans la première période.

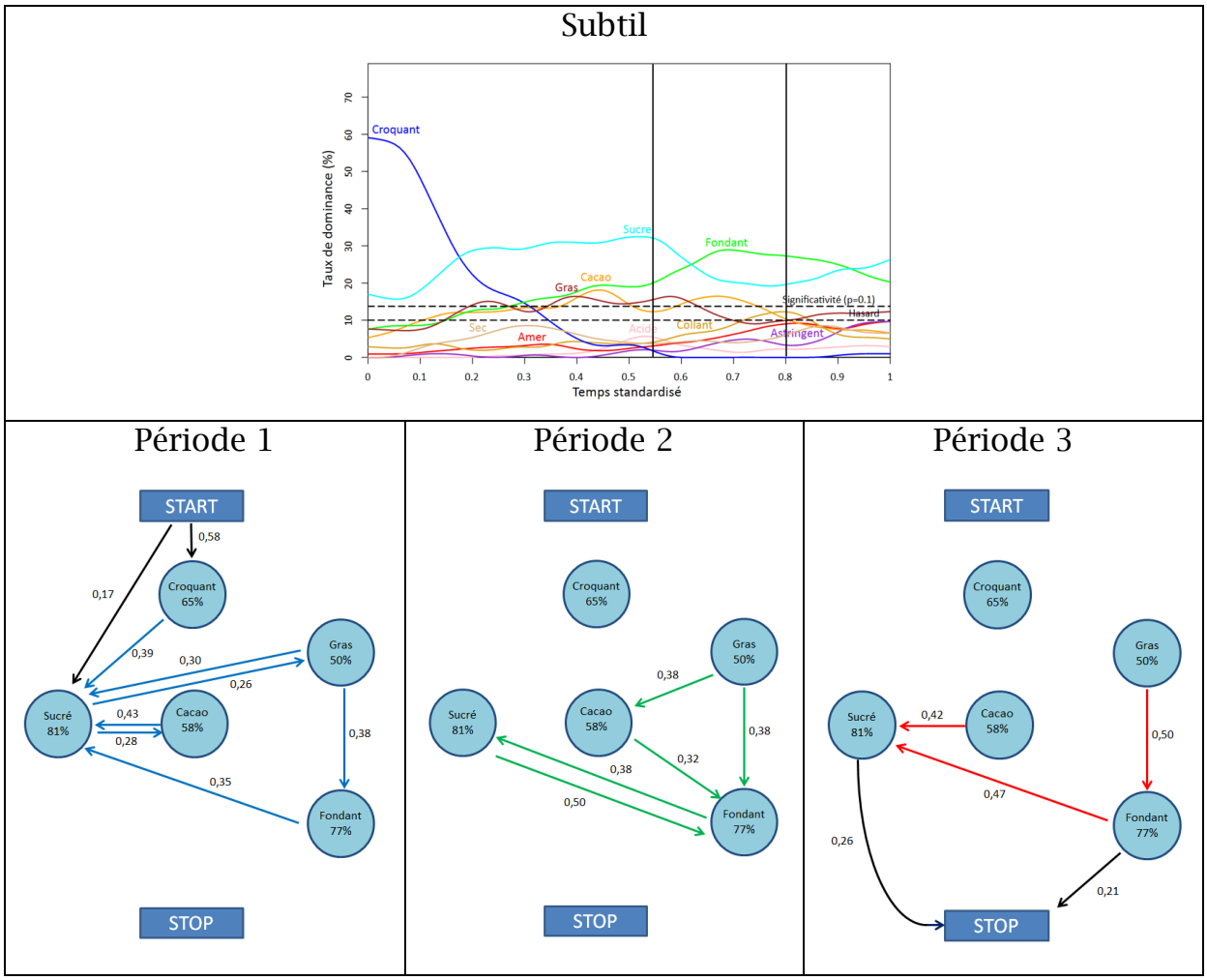


Figure 30 Courbes et graphe DTS avec découpage en 3 périodes du chocolat Subtil. Les flèches bleues correspondent aux transitions observées durant la première période, les vertes à celles observées durant la deuxième période et les rouges à celles observées durant la troisième période alors que les flèches noires correspondent aux probabilités initiales et finales.

En regardant le graphe DTS, il est d'abord intéressant de constater que les probabilités de transition changent d'une période à l'autre ce qui justifie le découpage de la durée de dégustation en périodes. La transition de Gras vers Fondant a par exemple une probabilité égale à 0,38 durant les deux premières périodes et à 0,50 durant la troisième période. Les panélistes choisissent principalement Sucré lorsque Cacao ou Fondant est dominant.

Les probabilités initiales nous apprennent que 58% des panélistes ont choisi Croquant comme premier descripteur et 17% ont choisi Sucré. Ceux qui ont choisi Croquant ont ensuite choisi majoritairement Sucré qui est suivi soit par Gras soit par Cacao. Après Gras, 30% des panélistes choisissent Sucré et 38% choisissent Fondant. Si Sucré est au cœur de la perception en première période, la majorité des transitions se font vers Fondant en deuxième période. Durant la troisième période, Gras est suivi par Fondant alors que Cacao et Fondant sont suivis avec des probabilités élevées par Sucré. Finalement la dégustation se finit par Sucré ou Fondant.

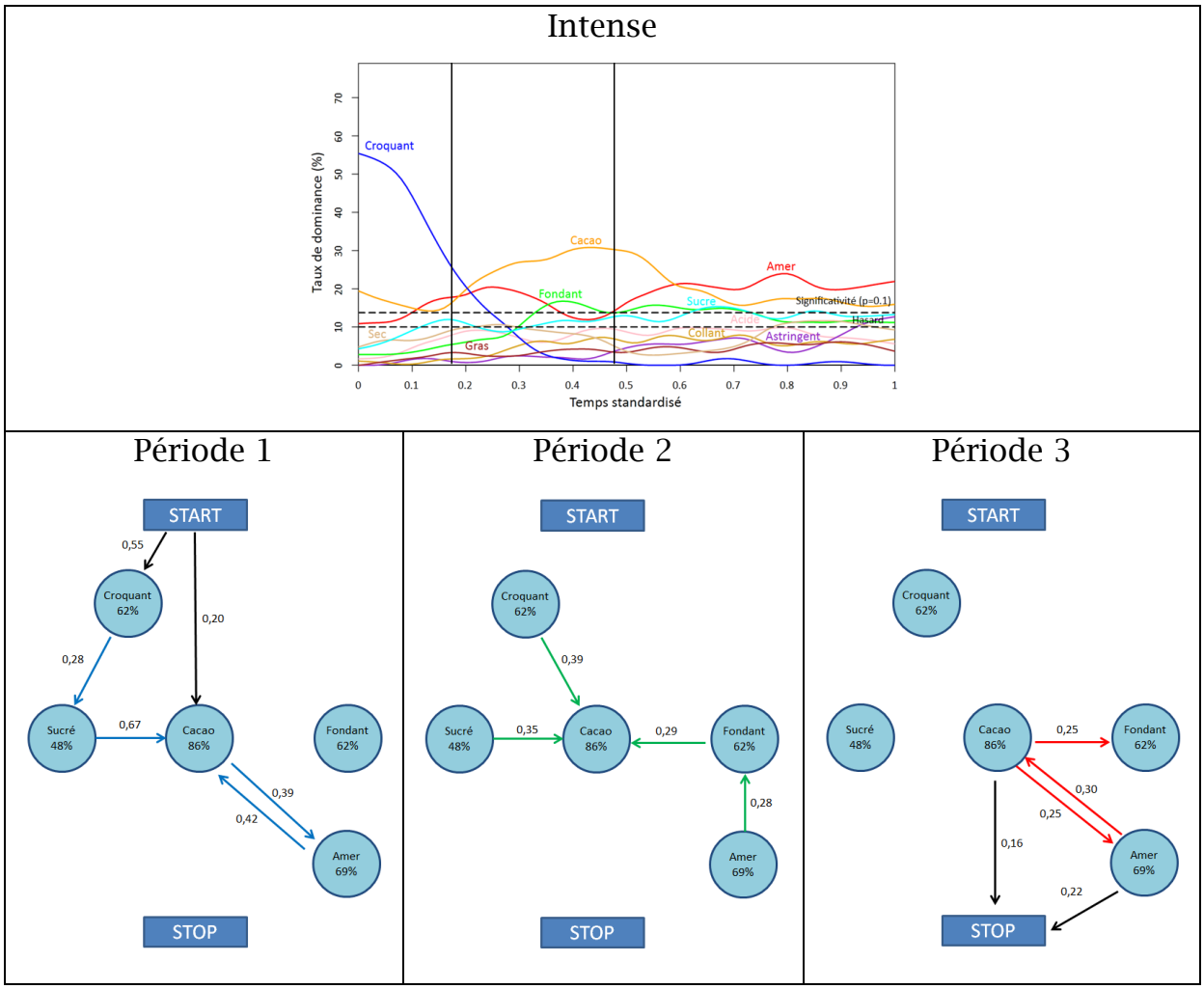


Figure 31 Courbes et graphe DTS avec découpage en 3 périodes du chocolat Intense. Les flèches bleues correspondent aux transitions observées durant la première période, les vertes à celles observées durant la deuxième période et les rouges à celles observées durant la troisième période alors que les flèches noires correspondent aux probabilités initiales et finales.

La perception du chocolat Intense est découpée en 3 périodes dont les frontières sont situées à 18% et 48% de la durée de dégustation (Figure 31). En regardant les courbes DTS, la première période est principalement caractérisée par Croquant, la deuxième par Cacao et la dernière par Amer.

Les panélistes ont choisi à 55% Croquant et à 20% Cacao comme premier descripteur. Durant la première période, ceux qui ont choisi Croquant choisissent ensuite Sucré puis Cacao. Cacao est associé à Amer. La deuxième période est caractérisée par Cacao vers lequel la majorité des flèches se dirigent. Les panélistes ayant choisi Amer comme dominant passent par Fondant avant de choisir Cacao. En troisième période nous observons à nouveau l'association Cacao-Amer. Cacao peut également être suivi par

Fondant. La dégustation se termine par la perception principalement d'Amer mais également de Cacao.

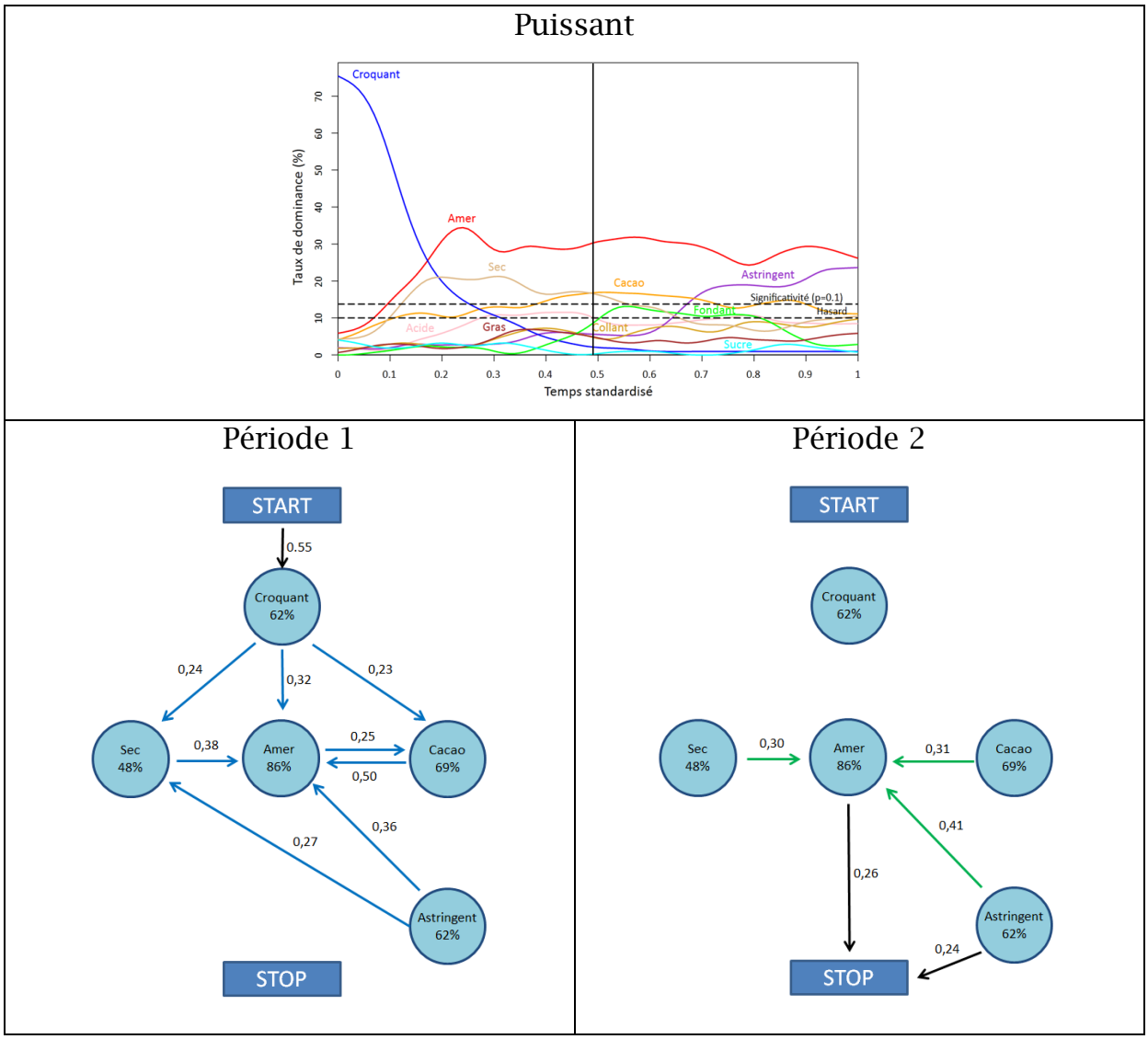


Figure 32 Courbes et graphe DTS avec découpage en 2 périodes du chocolat Puissant. Les flèches bleues correspondent aux transitions observées durant la première période et les vertes à celles observées durant la seconde période alors que les flèches noires correspondent aux probabilités initiales et finales.

La perception du chocolat Puissant est découpée en deux périodes dont la frontière est située à 49% de la durée totale de dégustation (Figure 32). En regardant les courbes, la première période est caractérisée par les descripteurs Croquant et Sec alors que dans la seconde ce sont les descripteurs Cacao, Astringent et dans une moindre mesure Fondant qui semblent importants. Le descripteur Amer est perçu quant à lui tout au long de la dégustation.

En regardant les graphes, nous apprenons que Croquant est choisi comme premier descripteur par 55% des panélistes et qu'il est suivi par Sec, Amer ou Cacao. Sec est choisi en première période mais pas en deuxième. La plupart des transitions se font vers Amer. Astringent est utilisé par 62% des panélistes pour ce produit mais il est choisi indépendamment du descripteur actuellement dominant. La perception se termine par Amer ou Astringent.

3.4.2 Fromages frais

Nous commençons par déterminer quel nombre de périodes est nécessaire pour chaque fromage frais en utilisant le test statistique pour lequel la statistique de test est estimée en se basant sur 1000 simulations.

Tableau 7 P-values des tests de différence entre matrices de transitions pour le découpage en périodes des chocolats Lindt Excellence.

	2 périodes	**3 périodes**		**4 périodes**		
	$1^{ère}$ vs $2^{ème}$	$1^{ère}$ vs $2^{ème}$	$2^{ème}$ vs $3^{ème}$	$1^{ère}$ vs $2^{ème}$	$2^{ème}$ vs $3^{ème}$	$3^{ème}$ vs $4^{ème}$
P1	0,021	0,011	0,036	0,007	0,059	0,009
P2	0,002	0,350	$< 10^{-3}$			
P3	0,060					
P4	0,017	0,256	0,133			
P5	0,090					
P6	0,003	0,067				

Avec un seuil $\alpha = 5\%$, le fromage frais P1 doit être découpé en 3 périodes (Tableau 7). Les fromages frais P2, P4 et P6 doivent être découpés en 2 périodes. Les fromages frais P3 et P5 ne nécessitent pas de découpage.

Tableau 8 Pourcentage de panélistes ayant utilisé chacun des descripteurs pour les 6 fromages frais.

	P1	P2	P3	P4	P5	P6
Herbes cuites	15,62	32,81	60,94	28,57	51,56	81,25
Crème	43,75	81,25	57,81	77,78	89,06	59,38
Herbes fraîches	48,44	37,50	35,94	53,97	46,88	37,50
Ail	76,56	56,25	45,31	61,90	59,38	59,38
Poivre	29,69	23,44	54,69	22,22	15,62	12,50
Âcre	26,56	23,44	23,44	14,29	12,50	15,62
Sel	76,56	65,62	60,94	63,49	45,31	45,31
Acide	40,62	29,69	25,00	28,57	28,12	28,12

En observant les pourcentages de panélistes ayant utilisé chaque descripteur (Tableau 8), nous constatons qu'en conservant un seuil $Dseuil$ à 50% nous excluons le descripteur Herbes fraîches pour P1 et P5. Pourtant, ce descripteur est proche du seuil, avec respectivement 48,44% et 46,88% des panélistes qui l'ont utilisé, et est important pour caractériser les fromages frais P1 et P5 par rapport aux fromages frais P2, P3 et P6. En revanche, nous considérons que l'Ail pour le fromage P3 et le Sel pour les fromages P5 et P6 qui ont été utilisés par 45% des panélistes peuvent ne pas apparaître dans les graphes de ces produits puisque ces descripteurs sont beaucoup moins souvent utilisés pour ces produits que pour les autres produits. Suite à ces constatations nous décidons donc de baisser légèrement le seuil à 46%. Pour ce jeu de données qui utilise 10 descripteurs la valeur par défaut $Tseuil = 0{,}15$ semble adaptée.

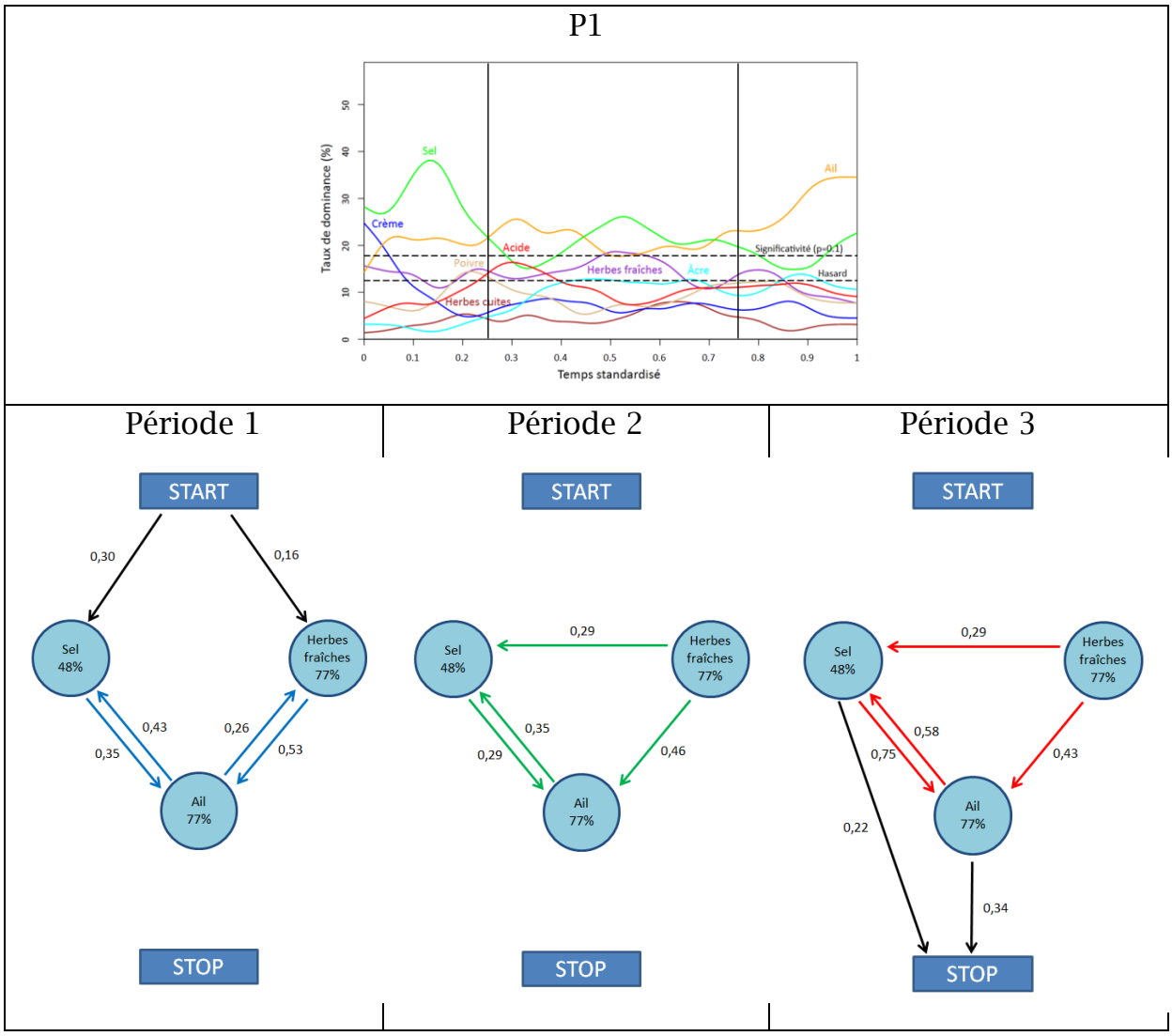


Figure 33 Courbes et graphe DTS avec découpage en 3 périodes du fromage frais P1. Les flèches bleues correspondent aux transitions observées durant la première période, les vertes à celles observées durant la deuxième période et les rouges à celles observées durant la troisième période alors que les flèches noires correspondent aux probabilités initiales et finales.

La perception du fromage frais P1 est découpée en 3 périodes dont les frontières sont situées à 25% et 76% de la durée totale de dégustation (Figure 33). En regardant les courbes DTS, la première période est caractérisée principalement par Sel et un peu par Crème et Ail. La deuxième période est plus confuse avec toujours Sel mais à un niveau plus bas, Ail et un peu Herbes fraîches et Acide alors que la troisième période est largement caractérisée par Ail.

En regardant les graphes DTS, 30% des panélistes ont choisi Sel comme premier descripteur et 16% ont choisi Herbes fraîches. Dans la première période, Sel est associé à Ail tout comme Herbes fraîches mais en revanche il

n'y a pas de transition entre Sel et Herbes fraîches. Dans la deuxième période, Sel et Ail sont encore associés. Il n'y a plus de transition vers Herbes fraîches qui est désormais suivi soit par Sel soit par Ail. La troisième période est similaire à la deuxième sauf que les transitions entre Sel et Ail ont des probabilités beaucoup plus élevées. Finalement la perception se termine par Ail pour 34% des panélistes et par Sel pour 22% des panélistes.

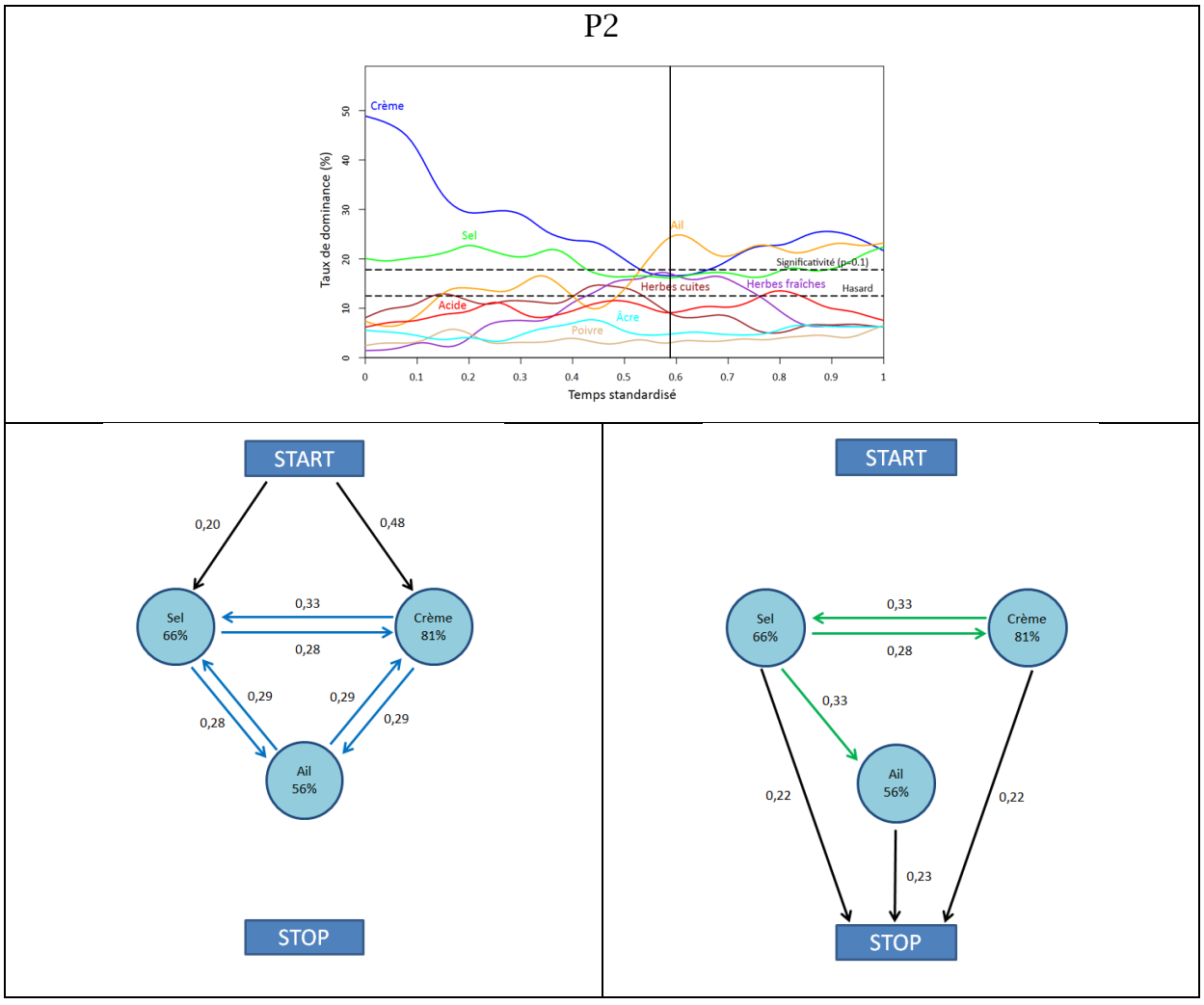


Figure 34 Courbes et graphe DTS avec découpage en 2 périodes du fromage frais P2. Les flèches bleues correspondent aux transitions observées durant la première période et les vertes à celles observées durant la deuxième période alors que les flèches noires correspondent aux probabilités initiales et finales.

La perception du fromage frais P2 est découpée en 2 périodes dont la frontière est située à 59% de la durée totale de dégustation (Figure 34). La première période est caractérisée par Crème et à un niveau plus faible Sel et la deuxième période se distingue par Ail en plus de Crème et Sel.

Les graphes DTS nous apprennent que 48% des panélistes ont choisi Crème comme premier descripteur et 20% ont choisi Sel. Durant la première période, les panélistes semblent choisir aléatoirement parmi Sel, Crème et Ail avec une même probabilité d'environ 30%. Pendant la deuxième période, il y a des transitions entre Sel et Crème et de Sel à Ail. La perception se termine par Sel, Crème ou Ail.

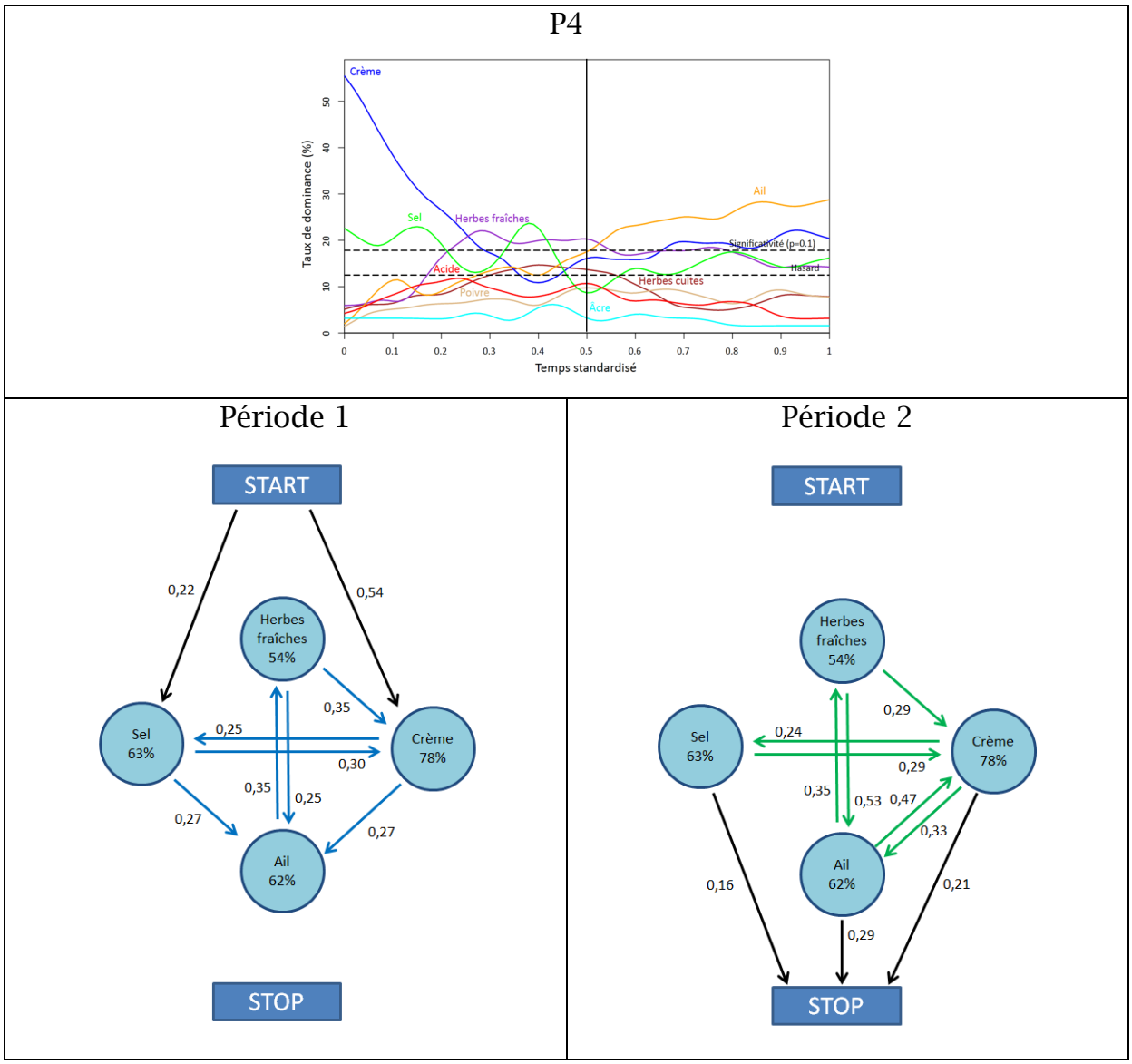


Figure 35 Courbes et graphe DTS avec découpage en 2 périodes du fromage frais P4. Les flèches bleues correspondent aux transitions observées durant la première période et les vertes à celles observées durant la deuxième période alors que les flèches noires correspondent aux probabilités initiales et finales.

La perception du fromage frais P4 est découpée en 2 périodes dont la frontière est située à 50% de la durée totale de dégustation (Figure 35). La

première période est caractérisée par Crème, Sel et Herbes fraîches alors que la deuxième se distingue par un niveau élevé pour Ail.

Pour ce produit, nous observons des associations entre Herbes fraîches et Ail et entre Sel et Crème ainsi que des transitions de Herbes fraîches à Crème et de Crème à Ail durant les deux périodes. Nous observons également des transitions de Sel à Ail durant la première période et de Ail à Crème durant la deuxième période. La probabilité d'aller de Herbes fraîches à Ail passe de 25% en première période à 53% en deuxième période. La perception de ce produit commence par Crème ou Sel et se termine par Sel, Ail ou Crème.

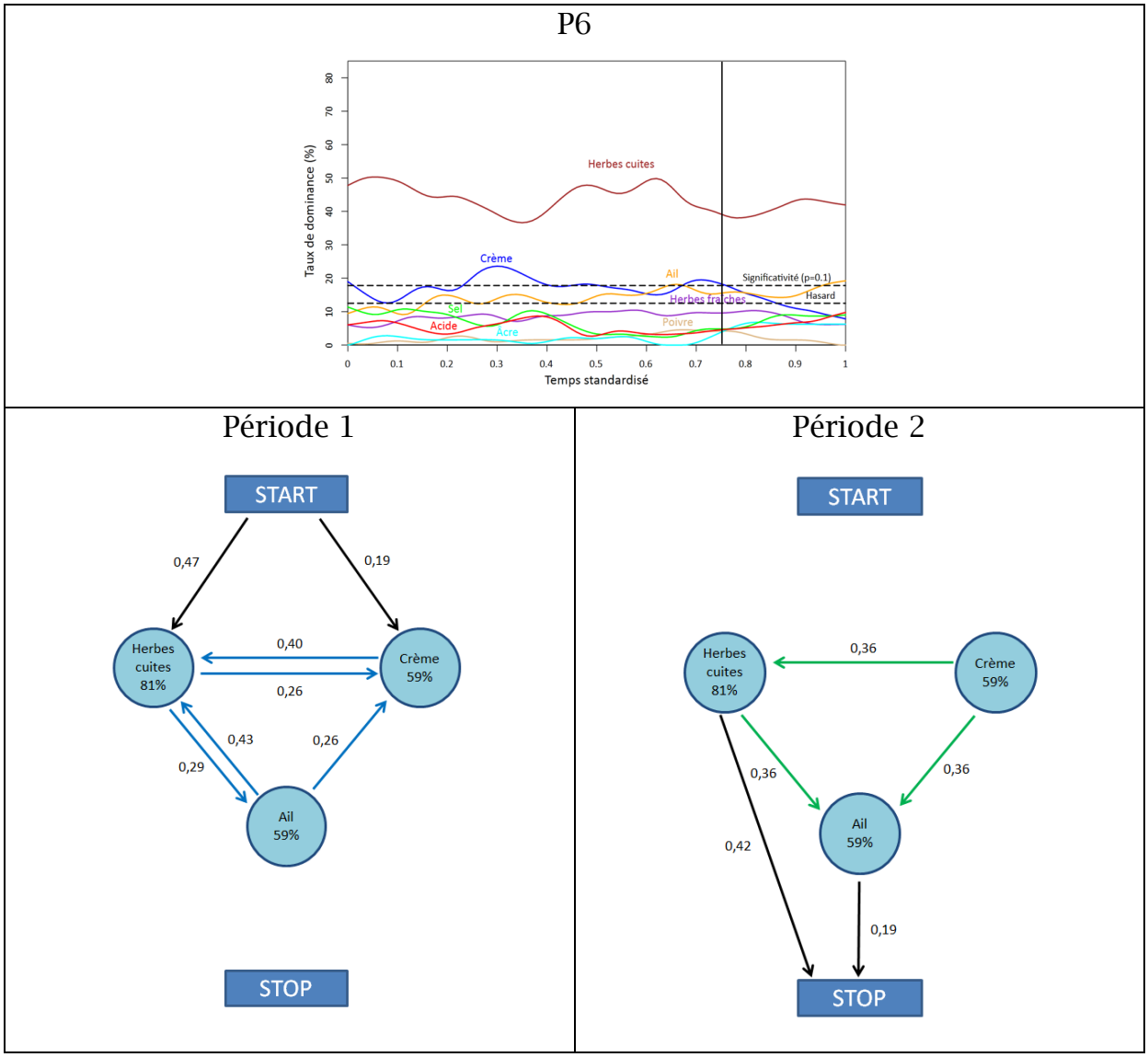


Figure 36 Courbes et graphe DTS avec découpage en 2 périodes du fromage frais P6. Les flèches bleues correspondent aux transitions observées durant la première période et les vertes à celles observées durant la deuxième période alors que les flèches noires correspondent aux probabilités initiales et finales.

La perception du fromage frais P6 est découpée en 2 périodes dont la frontière est située à 75% de la durée totale de dégustation (Figure 36). Les deux périodes sont assez similaires avec un niveau très élevé pour Herbes cuites. Dans la deuxième période il y a une diminution rapide du niveau du descripteur Crème.

Les graphes DTS nous indiquent que le fromage frais P6 est perçu comme Herbes cuites ou Crème puis durant la première période nous observons des transitions entre Herbes cuites et Ail et entre Herbes cuites et Crème ainsi que de Ail à Crème. Durant la deuxième période il n'y a plus que des transitions de Crème à Herbes cuites ainsi que de Herbes cuites et de Crème à Ail. Finalement la perception du fromage frais P6 se termine par Herbes cuites pour 46% des panélistes et par Ail pour 19% des panélistes.

3.5 Test de différence entre produits

Cette partie présente d'abord des résultats visant à observer le comportement du test puis les résultats du test de différence entre produits pour les fromages frais et les chocolats Excellence, du test de différence entre les hommes et les femmes puis entre les panélistes du Royaume-Unis et ceux des autres pays pour les Goudas.

3.5.1 Validation du test

Afin d'observer le comportement du test de différence dans des conditions contrôlées, nous proposons de séparer le panel de l'étude sur les fromages frais en deux par tirage aléatoire et d'utiliser le test de différence pour comparer les deux moitiés de panel pour un produit. De la même manière, nous réalisons la comparaison entre produits en tirant aléatoirement la moitié du panel pour chaque produit. Nous réalisons à chaque fois 100 tirages et nous observons la moyenne des p-values obtenues.

Tableau 9 Moyennes des p-values obtenues lors de 100 tirages aléatoires pour le test de différence entre les sous-ensembles des différents produits.

	P1	**P2**	**P3**	**P4**	**P5**	**P6**
P1	0,115	0,061	0,046	0,073	0,046	0,033
P2		0,162	0,050	0,199	0,071	0,035
P3			0,164	0,034	0,046	0,036
P4				0,120	0,137	0,034
P5					0,175	0,058
P6						0,133

Les résultats (Tableau 9) montrent que pour le test de différence entre deux échantillons correspondant à un même produit les p-values sont en moyennes supérieures à 10% ce qui signifie que l'hypothèse selon laquelle les échantillons sont issus d'un même modèle ne peut être rejetée. La proximité entre les produits P2 et P4 et entre les produits P4 et P5 est confirmée. Pour les autres comparaisons de produits les p-values sont autour de 5% malgré le fait que ces produits soient reconnus comme différents. La petite taille des échantillons ($n = 32$) peut peut-être expliquer qu'il soit difficile de conclure statistiquement à une différence.

3.5.2 Fromages frais

Tableau 10 P-values pour le test de différence, avec une statistique de test estimée à partir de 1000 simulations, entre les différents produits composants le jeu de données des fromages frais.

	P1	P2	P3	P4	P5	P6
P1		0,001	$< 10^{-3}$	0,003	$< 10^{-3}$	$< 10^{-3}$
P2			$< 10^{-3}$	0,351	0,001	$< 10^{-3}$
P3				$< 10^{-3}$	$< 10^{-3}$	$< 10^{-3}$
P4					0,057	$< 10^{-3}$
P5						$< 10^{-3}$

Pour les fromages frais (Tableau 10) nous observons que les modèles estimés pour toutes les paires de produits sont statistiquement différents sauf pour la paire de produits P2 et P4 pour laquelle l'hypothèse H_0 ne peut être rejetée et les produits P4 et P5 pour lesquels la p-value est légèrement supérieure à 5%.

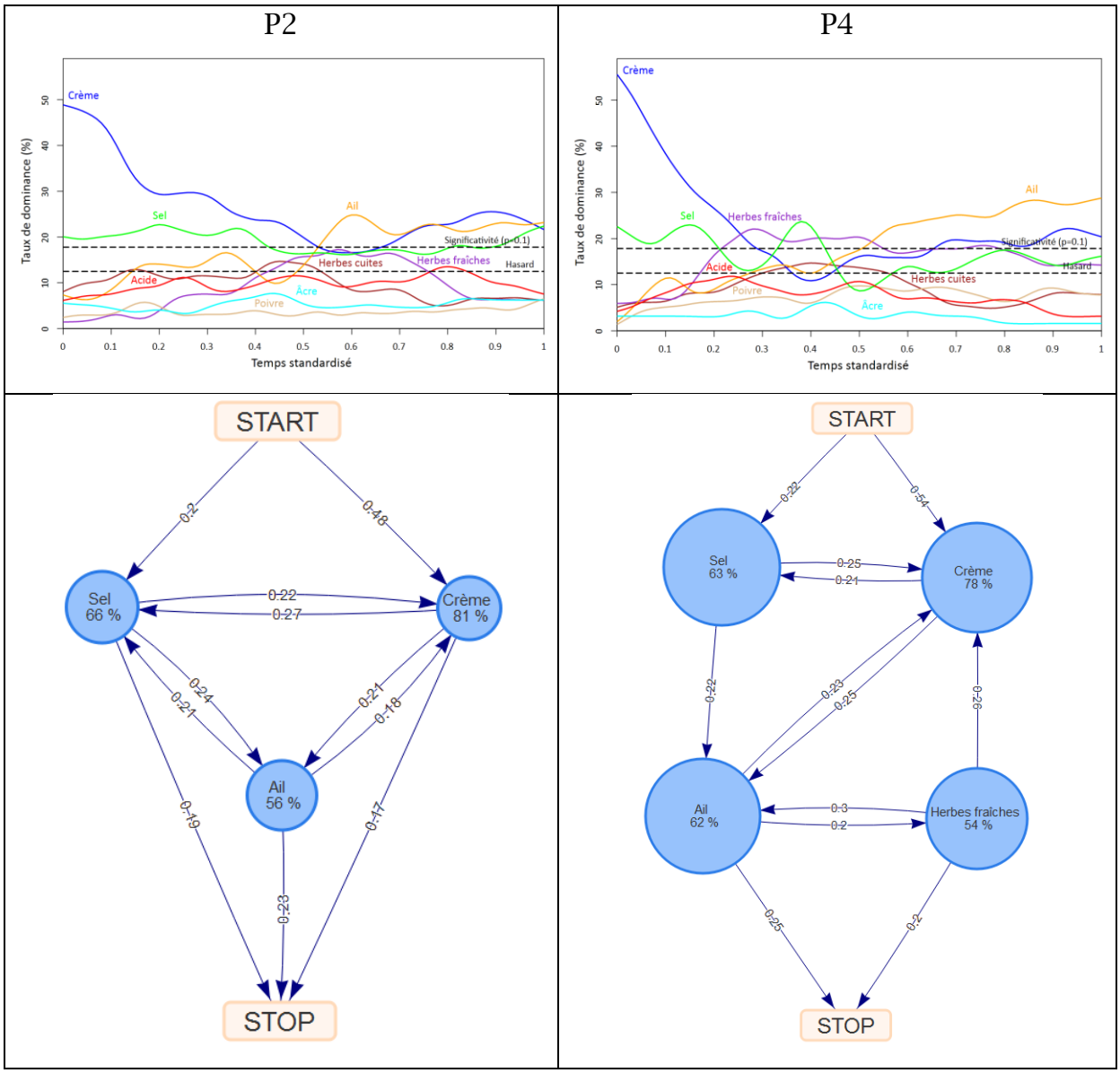


Figure 37 Graphes et courbes DTS des fromages frais P2 et P4.

Pour les produits P2 et P4, la différence de perception semble en effet très faible au regard des courbes DTS. Le niveau de la courbe du descripteur Crème est un peu plus élevé pour le P2, alors que le niveau d'Herbe fraîche est lui un peu plus élevé pour le P4, mais ce sont les seules différences. En ce qui concerne les graphes, le descripteur Herbes fraîches est présent dans le graphe du fromage P4 et pas dans celui du P2. Globalement les probabilités de transition sont assez similaires mais certaines transitions sont présentes dans le graphe du P2 mais pas dans celui du P4 : la transition de Ail vers Sel et les transitions de Sel et Crème vers STOP.

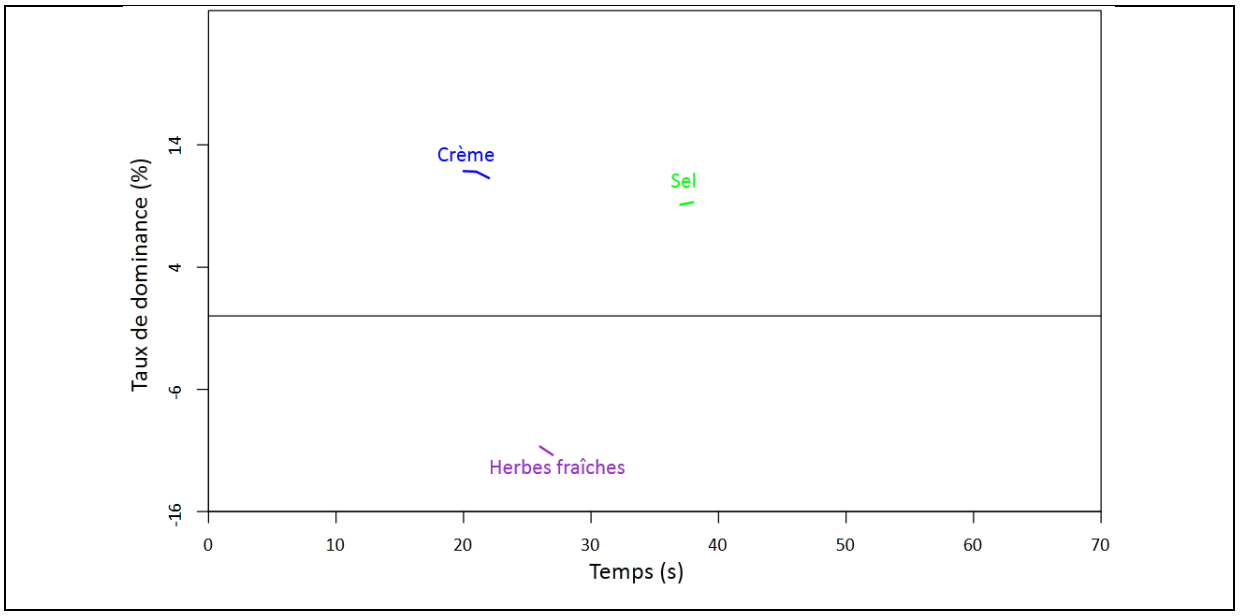


Figure 38 Courbes de différence entre le fromage frais P2 (en haut) et le fromage frais P4 (en bas). Risque alpha : 10%, durée minimale : 5%

Les courbes de différence entre le fromage frais P2 et le fromage frais P4 (Figure 38) nous montrent que les courbes DTS de ces produits sont en effet très proches avec uniquement quelques différences significatives mais sur des durées très courtes pour la crème, le sel et les herbes fraîches.

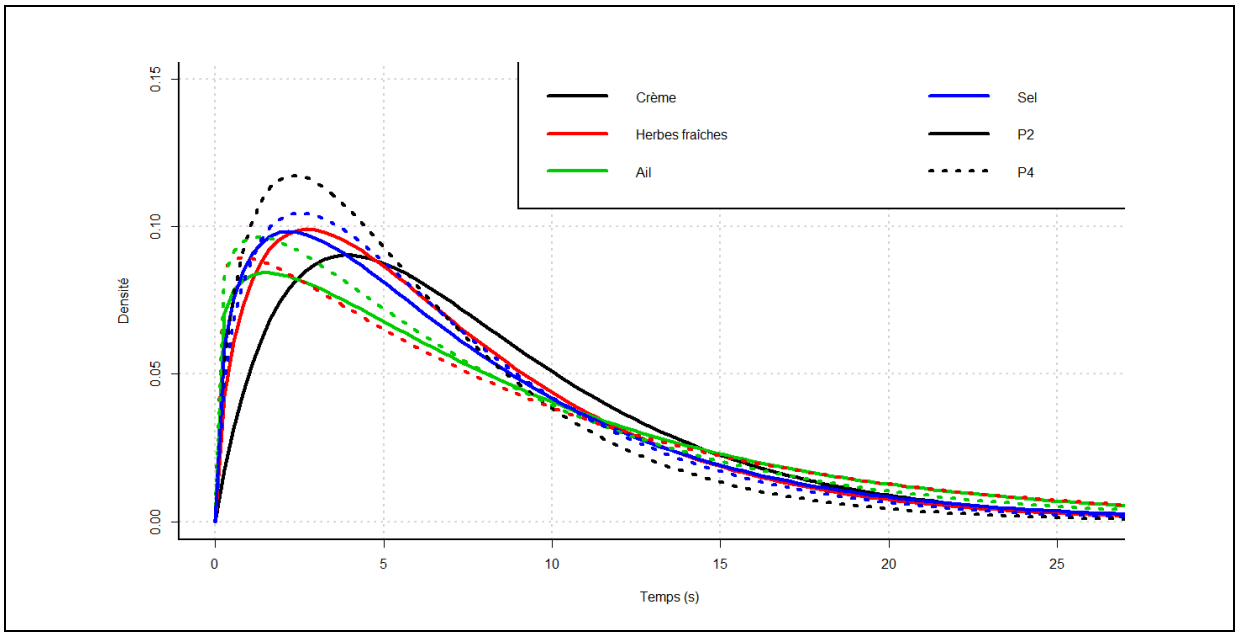


Figure 39 Lois Gamma estimées pour les descripteurs Crème, Herbes fraîche, Ail et Sel pour les fromages frais P2 et P4.

Il existe seulement de légères différences entre les durées de dominance observées pour P2 et celles pour P4 avec un peu plus de durées courtes pour P4 (Figure 39).

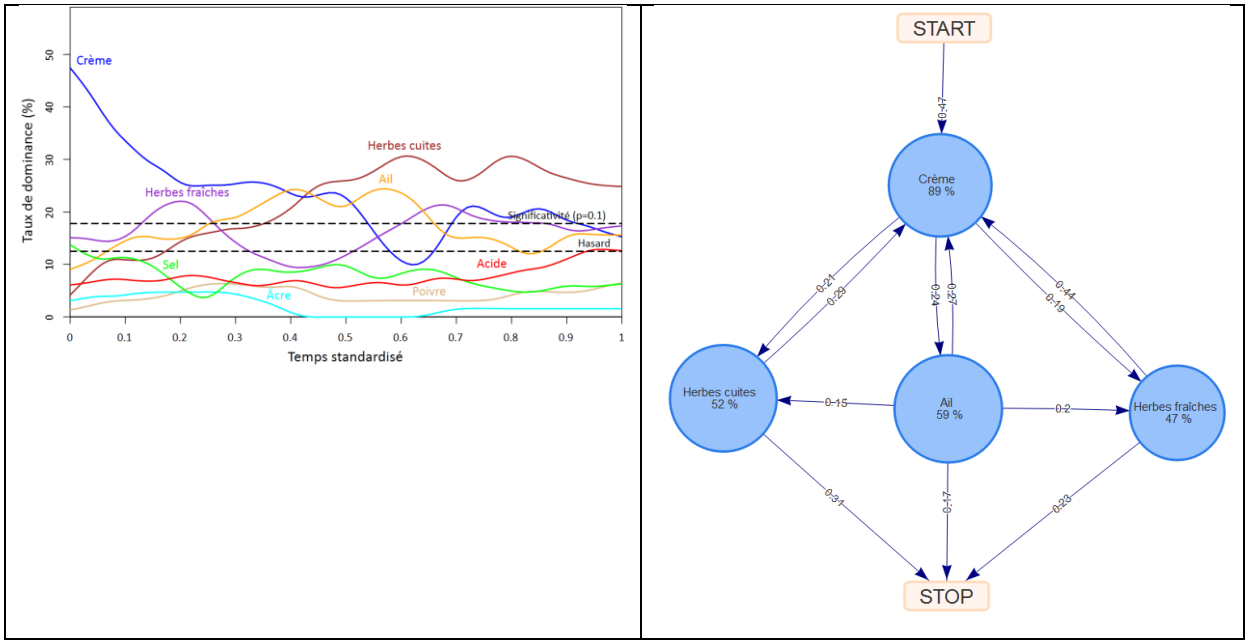


Figure 40 Graphes et courbes DTS du fromage frais P5.

Les courbes DTS du fromage P4 (Figure 37, côté gauche) et celles du fromage P5 (Figure 40) ont beaucoup de points communs, notamment pour Crème et Herbes fraîches, mais se distinguent par un nombre de citations beaucoup plus élevé d'Herbes cuites pour le P5 et un nombre de citations plus élevé de

Sel et Ail pour le P4. Pour les graphes, nous constatons que Sel est présent pour le P4 et pas pour le P5, alors que c'est l'inverse pour Herbes cuites. Pour le P5 seul Crème apparaît comme premier descripteur alors qu'il y a aussi Sel pour le P4.

3.5.3 Chocolats Excellence

Nous allons maintenant tester si les chocolats Excellence sont différents les uns des autres.

Tableau 11 P-values pour le test de différence, avec une statistique de test estimée à partir de 10000 simulations, entre les différents produits composants le jeu de données des chocolats Excellence.

	Subtil	**Intense**	**Puissant**	**Prodigieux**
Subtil		$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
Intense			$< 10^{-4}$	$< 10^{-4}$
Puissant				1.10^{-4}
Prodigieux				

Selon le test de différence, les produits sont tous perçus significativement différemment avec des p-values inférieures à 10^{-4} ou égales à 10^{-4} pour les chocolats Puissant et Prodigieux qui semblent pourtant proches au regard des courbes et des graphes DTS (Tableau 11). Nous avons d'ailleurs constaté que les courbes de différence entre ces deux produits ne montraient aucune différence sans standardisation et seulement de légères différences pour Amer et Astringent avec standardisation.

3.5.4 Goudas : Hommes-Femmes

Le test de différence, en plus d'être utilisé pour savoir si la perception de deux produits est significativement différente, peut permettre de tester si la perception d'un même produit diffère significativement au sein du panel entre, par exemple, les hommes et les femmes. C'est ce que nous allons tester pour l'étude sur les Goudas.

Tableau 12 P-values pour le test de différence entre les hommes et les femmes pour les Gouda composants le jeu de données ESN, avec une statistique de test estimée à partir de 10000 simulations,

	16wk30	4wk30	4wk48
P-value	0,0649	0,7538	0,2734

Pour les 3 goudas, le test ne permet pas de rejeter l'hypothèse selon laquelle les hommes et les femmes ont la même perception au seuil $\alpha = 5\%$ mais pour le Gouda le plus vieux (16wk30**)** la p-value est très proche de 5% montrant tout de même l'existence de légères différences.

Tableau 13 Pourcentage de panélistes ayant utilisé chacun des descripteurs selon le genre pour les 3 goudas du jeu de données ESN.

	16wk30		4wk30		4wk48	
	H	**F**	**H**	**F**	**H**	**F**
Amer	38,10	35,92	25,37	25,06	20,44	17,62
Fromage	46,15	49,61	50,37	54,01	50,73	53,63
Dense dur	68,86	59,95	52,21	52,97	19,34	21,50
Gras	29,67	25,84	34,93	31,01	38,69	38,86
Fondant	32,60	28,94	31,99	32,30	60,22	60,62
Lacté crémeux	21,25	23,26	36,03	35,92	48,18	48,45
Salé	46,15	51,16	34,93	36,69	32,48	40,16
Piquant	43,22	50,65	15,81	19,90	16,06	18,91
Acide	31,50	36,18	19,85	22,48	16,79	21,24
Tendre	24,54	27,91	42,28	45,99	66,06	67,88

En observant les pourcentages de panélistes ayant utilisé chaque descripteur (Tableau 13) nous avons décidé de fixer $Dseuil$ à 42% ce qui permet d'avoir les même descripteurs présents dans les graphes des hommes et des femmes et ainsi de pouvoir comparer les probabilités de transition. $Tseuil$ est conservé à 15%.

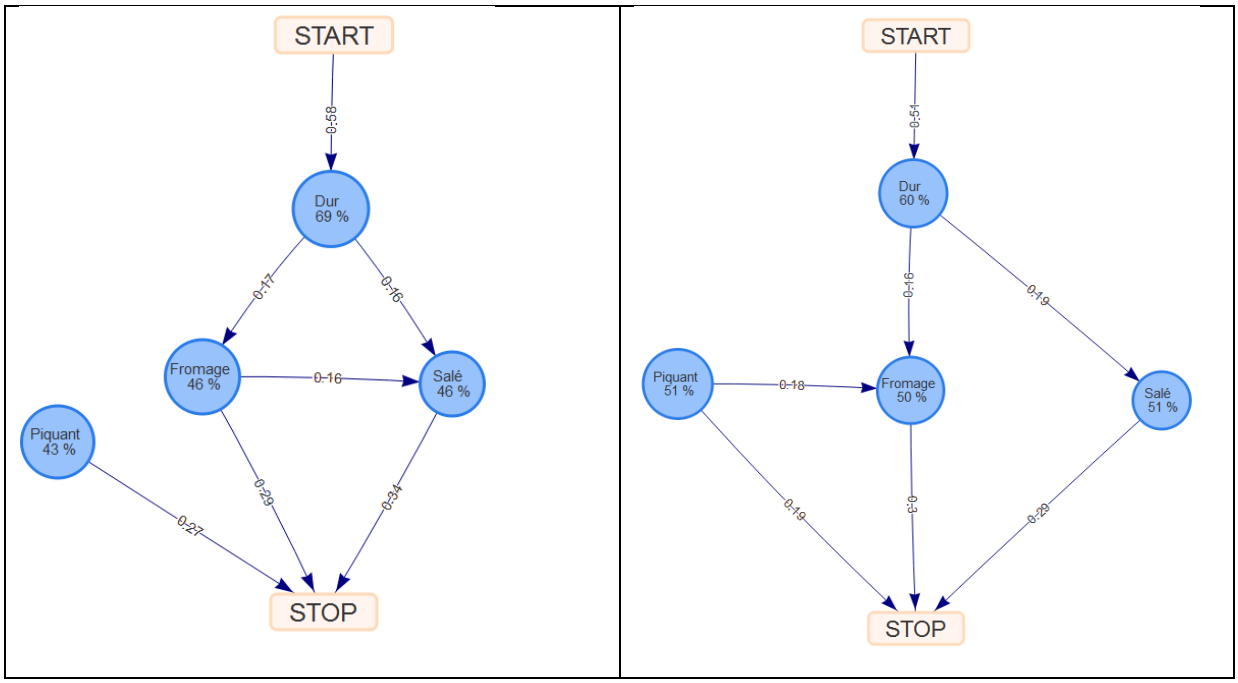


Figure 41 Graphes DTS avec Dseuil=42% et Tseuil=15% pour la perception du gouda 16wk30 pour les hommes à gauche et les femmes à droite.

En comparant les deux graphes DTS obtenus pour la perception du Gouda 16wk30 (Figure 41) nous constatons que les probabilités de transition sont légèrement différentes avec par exemple 58% de panélistes qui ont choisis Dur comme premier descripteur parmi les hommes et seulement 51% parmi les femmes. Nous remarquons également que la transition de Piquant à Fromage est présente pour les femmes mais par pour les hommes alors que la transition de Fromage à Salé est au contraire présente pour les hommes et pas pour les femmes mais dans les deux cas les probabilités sont proches du seuil. Il est intéressant aussi de noter que la proportion de panélistes ayant perçu ce Gouda piquant est plus élevée pour les femmes que pour les hommes. Ces petites différences expliquent que la p-value du test de différence soit certes supérieure à 5%, mais de peu pour ce produit en comparant la perception des hommes et celle des femmes.

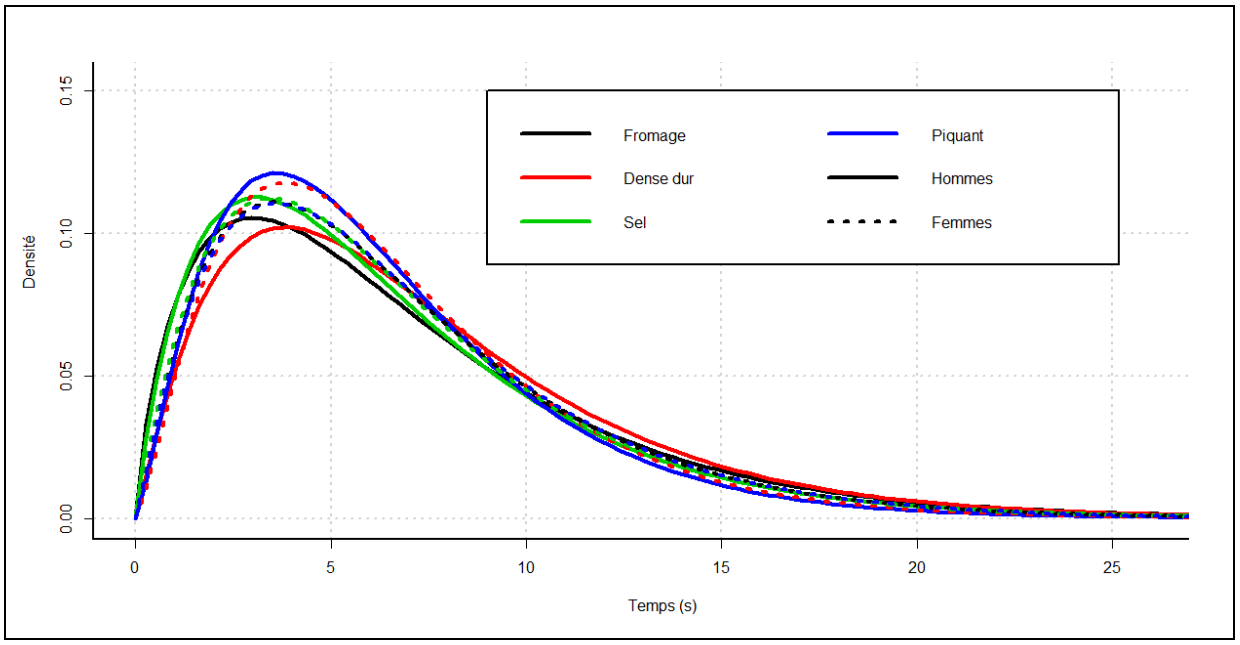


Figure 42 Lois Gamma estimées pour les descripteurs Fromage, Dense-dur, Sel et Piquant pour le Gouda 16wk30 pour les hommes et pour les femmes.

Les distributions estimées pour les durées de dominance observées lors de la dégustation du Gouda 4wk30 sont très proches entre les hommes et les femmes (Figure 42).

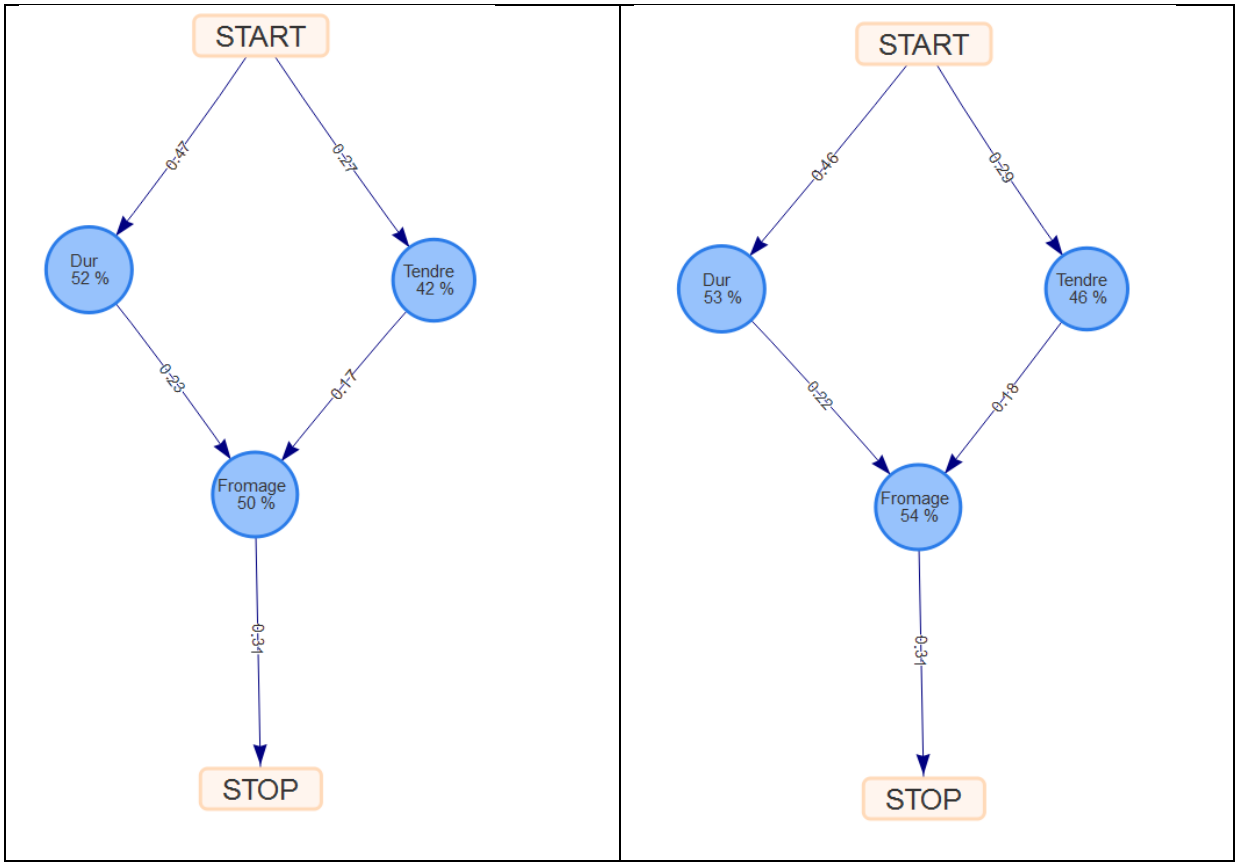


Figure 43 Graphes DTS avec Dseuil=42% et Tseuil=15% pour la perception du gouda 4wk30 pour les hommes à gauche et les femmes à droite.

Pour le Gouda 4wk30 les graphes ne montrent quasiment aucune différence (Figure 43) avec les mêmes descripteurs et les mêmes transitions avec des probabilités identiques à 2% prêts. Nous n'avons également pas observé de différences entre les distributions des durées de dominance. La p-value du test de différence pour ce produit en comparant la perception des hommes et celle des femmes est logiquement très élevée (0,7538).

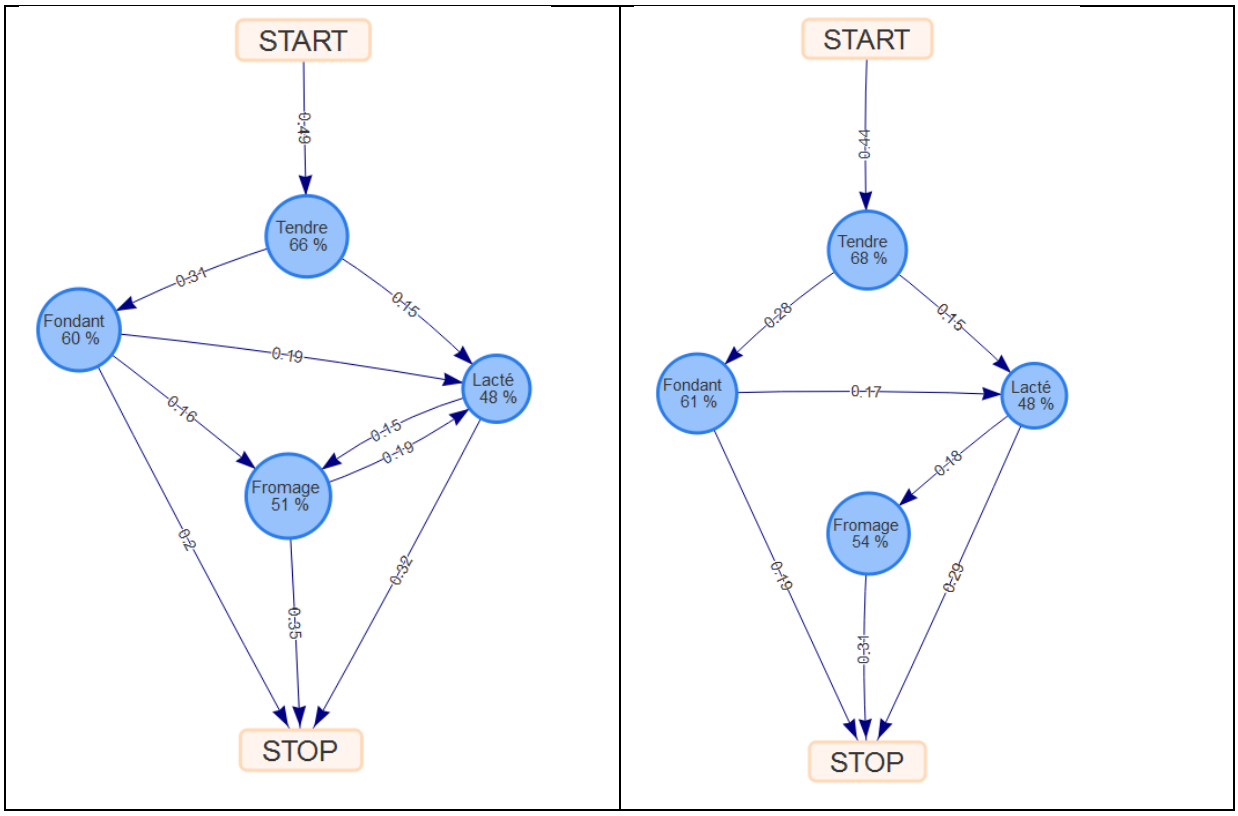


Figure 44 Graphes DTS avec Dseuil=42% et Tseuil=15% pour la perception du gouda 4wk48 pour les hommes à gauche et les femmes à droite.

Pour le Gouda 4wk48 les deux graphes sont très similaires (Figure 44). L'ensemble des valeurs sont proches, seules les transitions de Fondant à Fromage et de Fromage à Lacté différent dans le graphe puisqu'elles sont présentes pour les hommes et pas pour les femmes mais elles sont à la limite du seuil. Nous n'avons pas observé de différences entre les distributions des durées de dominance.

3.5.5 Gouda : Royaume-Uni contre les autres pays

Nous nous intéressons maintenant pour les mêmes Goudas à la comparaison du panel du Royaume-Uni par rapport aux panels des autres pays. En effet, Thomas et al. (2017) ont souligné que le panel du Royaume-Uni se distinguait

des autres par les préférences des Goudas et un niveau plus élevé de perception du gras pour les deux Goudas avec les durées de maturation les plus courtes.

Tableau 14 P-values pour le test de différence entre le panel du Royaume-Uni et les panels des autres pays pour les Goudas composants le jeu de données ESN, avec une statistique de test estimée à partir de 10000 simulations..

	16wk30	4wk30	4wk48
P-value	$< 10^{-4}$	$< 10^{-4}$	0,0851

Nous constatons que le test rejette l'hypothèse H_0 pour les goudas 16wk30 et 4wk30 selon laquelle le panel du Royaume-Uni a la même perception de ces goudas que les panels des autres pays. Pour le gouda 4wk48, l'hypothèse H_0 ne peut être rejetée au seuil $\alpha = 5\%$ mais la p-value est proche de ce seuil.

Tableau 15 Pourcentage de panélistes, pour le panel du Royaume-Uni et pour les panels des autres pays, ayant utilisés chacun des descripteurs pour les 3 goudas du jeu de données ESN.

	16wk30		4wk30		4wk48	
	RU	**Autres**	**RU**	**Autres**	**RU**	**Autres**
Amer	34,21	37,27	29,57	23,45	24,35	17,85
Fromage	63,16	45,09	49,57	53,10	49,57	53,01
Dense dur	55,26	65,45	34,78	56,75	13,04	22,04
Gras	27,19	27,45	58,26	27,37	53,04	36,07
Fondant	51,75	26,36	47,83	28,83	78,26	57,01
Lacté crémeux	27,19	21,27	42,61	34,67	63,48	44,99
Salé	50,00	49,09	40,87	34,85	36,52	37,34
Piquant	68,42	43,09	30,43	15,69	23,48	16,39
Acide	39,47	33,27	26,96	20,44	24,35	18,21
Tendre	38,60	23,82	60,87	40,69	84,35	63,57

En observant le pourcentage de panélistes ayant utilisé chacun des descripteurs (Tableau 15) nous constatons que pour le descripteur Fromage ou le descripteur Salé des valeurs sont légèrement en dessous du seuil de 50% mais doivent apparaître dans les graphes puisque les valeurs sont très

proches entre le panel du Royaume-Uni et les autres panels. Pour le Gouda 4wk30, le descripteur Fondant permet de différencier le panel du Royaume-Uni des autres mais il faut baisser le seuil à 47% pour qu'il apparaisse dans le graphe. A l'inverse, il est souhaitable que le Fromage n'apparaisse pas dans le graphe du Gouda 16wk30 pour les panels autres que celui du Royaume-Uni puisqu'il y a une différence importante. Pour cela il faut fixer un seuil supérieur à 46%. Nous choisissons donc de fixer le seuil de sélection des descripteurs *Dseuil* à 47%. Nous conservons *Tseuil* à 15%.

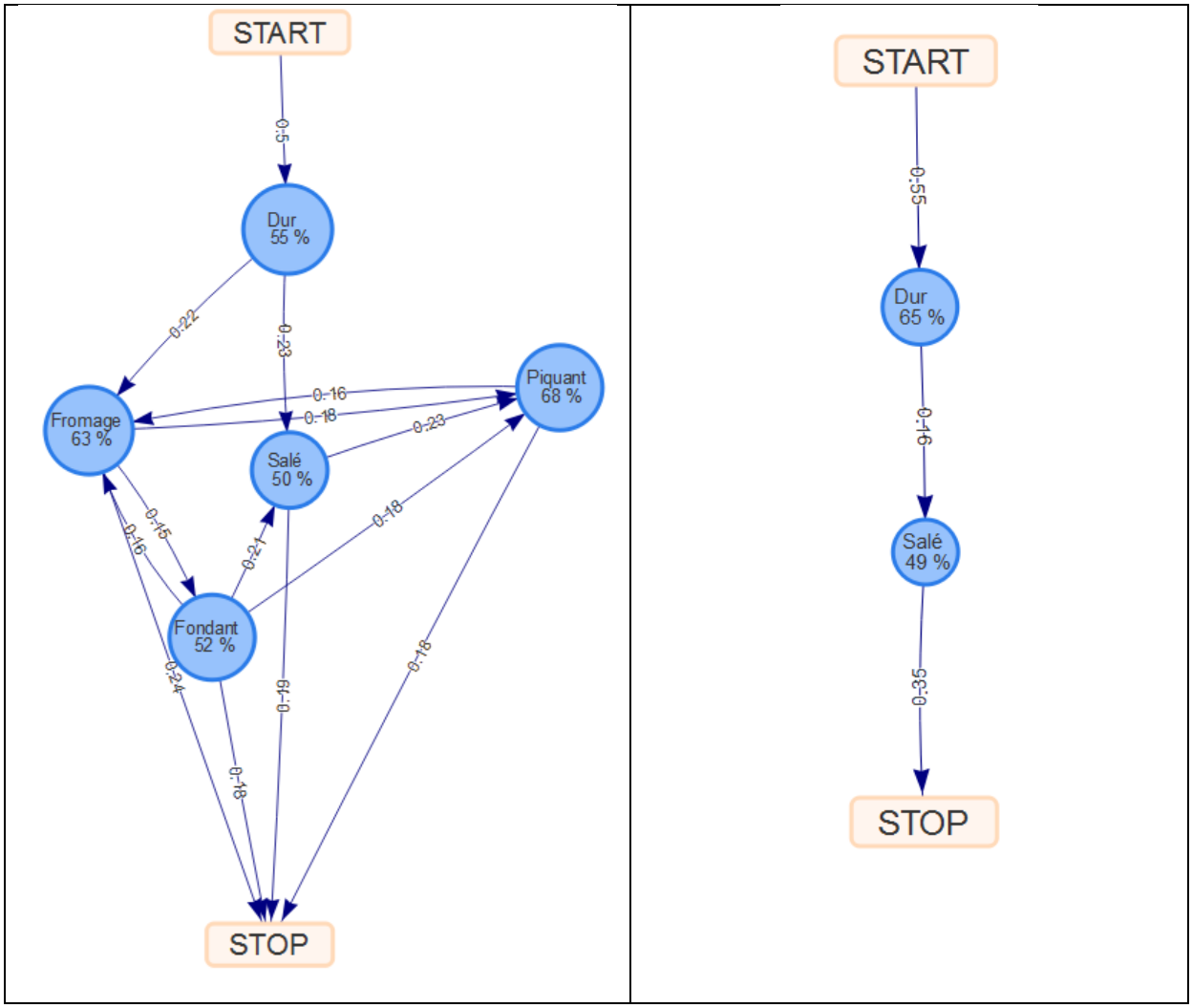


Figure 45 Graphes DTS avec Dseuil=47% et Tseuil=15% pour la perception du gouda 16wk30 pour le panel du Royaume-Unis à gauche et les autres panels à droite.

Pour la perception du Gouda 16wk30, le panel du Royaume-Uni se distingue nettement des autres panels par la présence, en plus de Dur et Salé, des descripteurs Piquant, Fromage et Fondant (Figure 45).

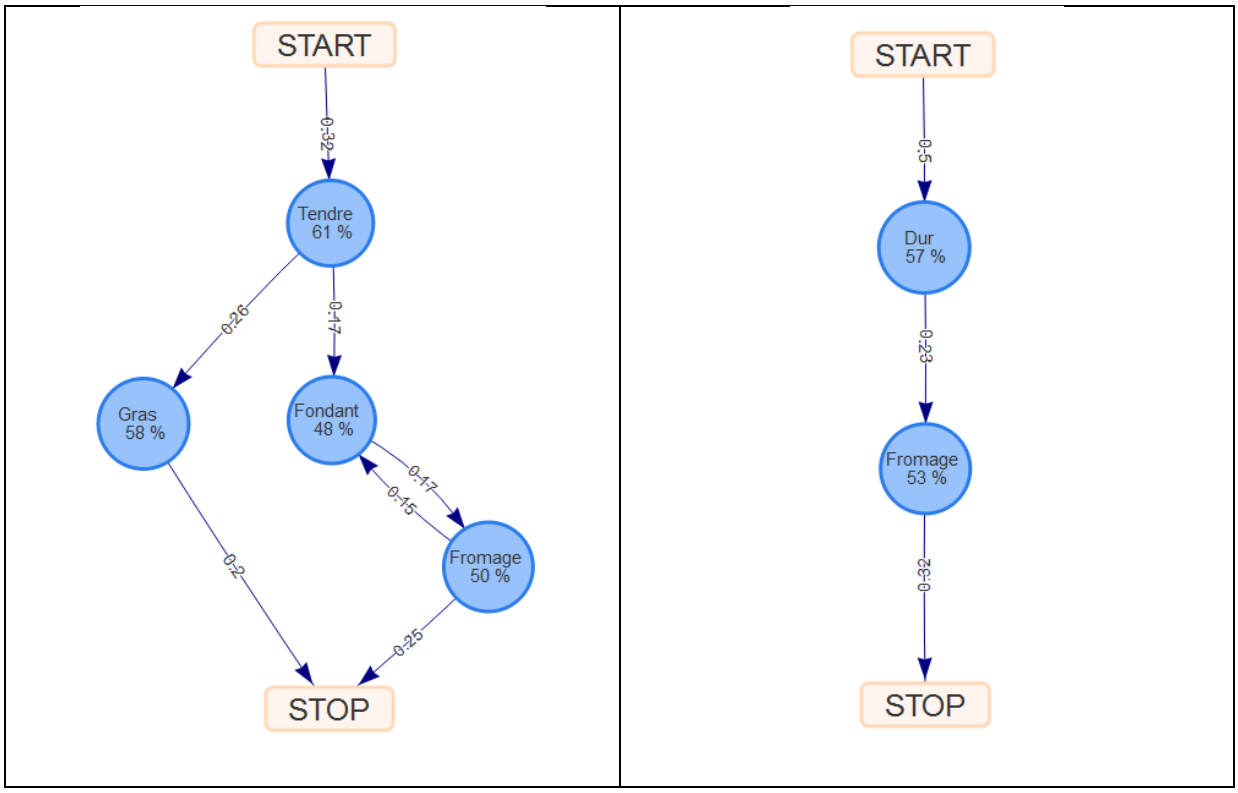


Figure 46 Graphes DTS avec Dseuil=47% et Tseuil=15% pour la perception du gouda 4wk30 pour le panel du Royaume-Unis à gauche et les autres panels à droite.

La différence de perception entre le panel du Royaume-Uni et les autres panels est encore plus importante pour le gouda 4wk30 avec une perception de Tendre, Fondant et Gras pour le panel du Royaume-Uni alors que les autres panels perçoivent ce Gouda comme dur (Figure 46).

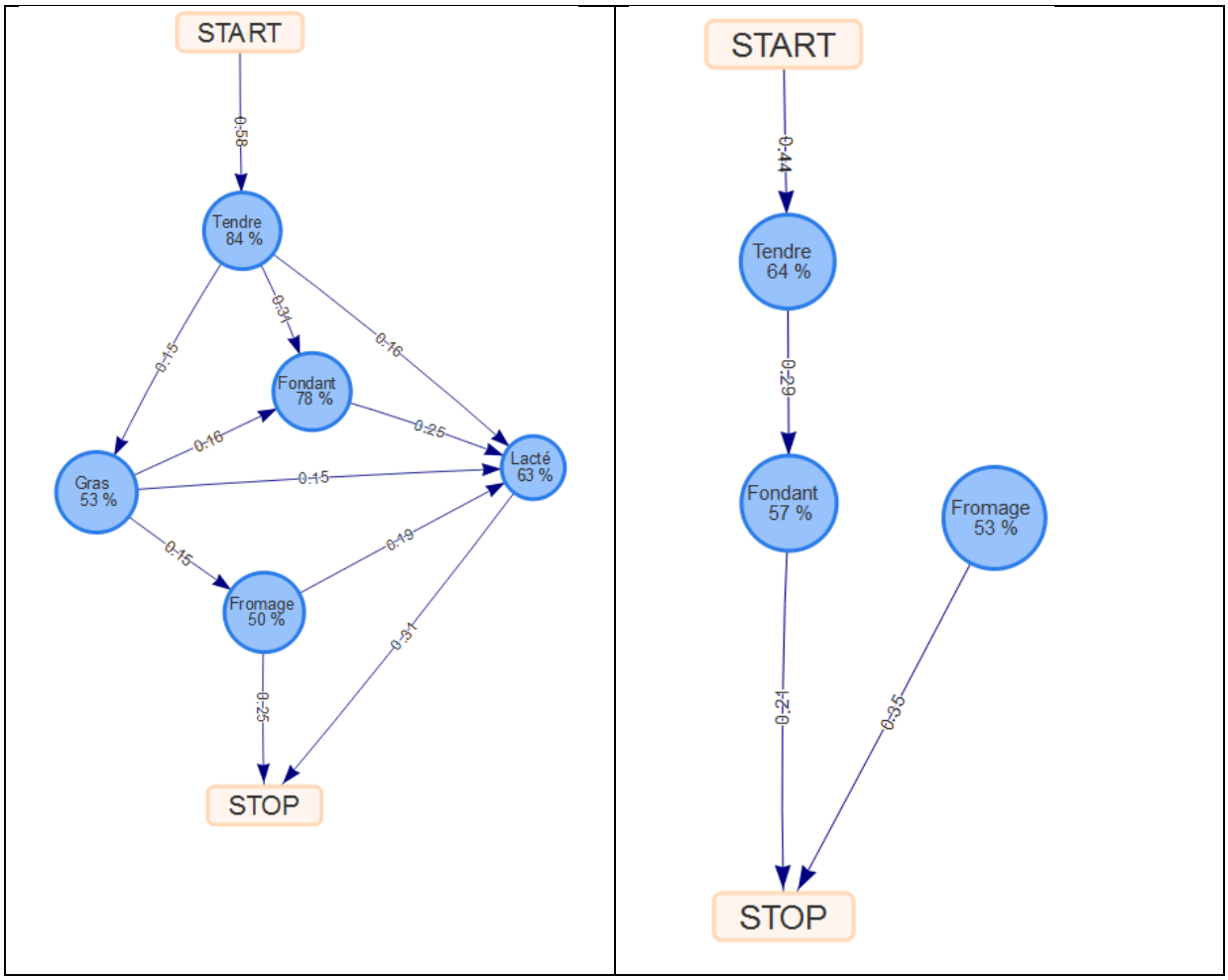


Figure 47 Graphes DTS avec Dseuil=47% et Tseuil=15% pour la perception du gouda 4wk48 pour le panel du Royaume-Uni à gauche et les autres panels à droite.

Pour le Gouda 4wk48, si nous retrouvons bien les descripteurs Tendre, Fondant et Fromage dans les deux graphes (Figure 47) les perceptions semblent tout de même assez différentes. La probabilité de choisir Tendre comme premier descripteur est plus élevée pour le panel du Royaume-Uni pour lequel nous remarquons également la présence de deux descripteurs supplémentaires : Gras et Lacté.

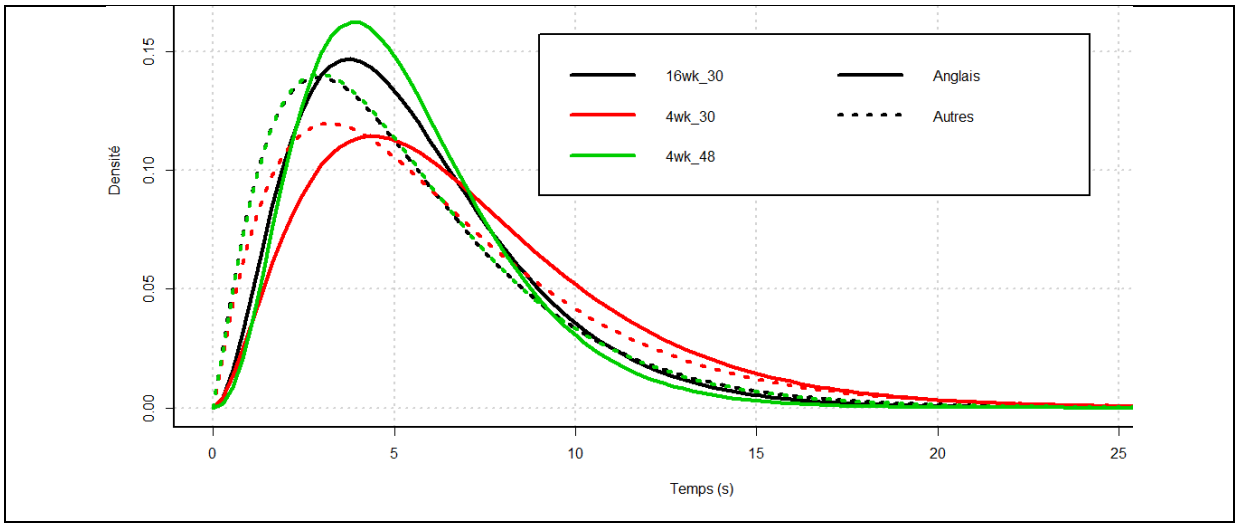


Figure 48 Lois Gamma estimées pour le descripteur Tendre pour les 3 Goudas pour le panel du Royaume-Uni et pour les panels des autres pays.

Nous observons que les durées de dominance du descripteur Tendre sont plus longues (distributions décalées vers la droite) pour les panélistes du Royaume-Uni (Figure 48). Nous avons observé le même phénomène pour les descripteurs Fondant et Fromage.

3.6 Segmentation

Dans cette partie nous allons voir deux applications de la segmentation basée sur un mélange de processus semi-markoviens présentée dans la partie 0.

3.6.1 Goudas

Nous présentons uniquement les résultats pour le Gouda 4wk30 (le plus jeune avec un taux de matière grasse faible) dont la perception par le panel est la plus complexe. Le nombre moyen de transitions observé pour ce produit est égal à 4,1. Le nombre d'itérations de l'algorithme EM est fixé à 400 ce qui est suffisant pour observer sa convergence.

Tableau 16 Valeurs prises par les critères d'information pour un nombre de segment G allant de 1 à 4 lors de la segmentation du Gouda.

	G=1	G=2	G=3	G=4
BIC	87525,93	86872,04	87058,21	87599,31
AIC	86859,73	85534,05	85048,43	84917,74
AIC_c	86874,99	85776,07	85794,99	86858,84

Le Tableau 16 nous montre que tous les critères d'information sélectionnent au moins deux segments ce qui confirme l'existence de différences de perception au sein du panel. Les critères BIC et AIC_c suggèrent de choisir 2 segments mais tous les deux prennent des valeurs très proches pour 2 et 3 segments. Nous proposons donc d'étudier la segmentation en 2 segments et celle en 3 segments. Le critère AIC suggère de sélectionner au moins 4 segments mais il est réputé pour être moins parcimonieux que BIC et AIC_c et a tendance à surestimer le nombre de segments. En segmentant le panel en 2, les segments obtenus sont composés respectivement de 398 et 267 panélistes. En segmentant le panel en 3, les segments obtenus sont cette fois composés respectivement de 242, 209 et 214 panélistes.

Tableau 17 Indice de Rand ajusté pour comparer la classification obtenue par la méthode des k-means à celle obtenue avec le modèle de mélange pour une segmentation de 2 à 5 segments pour le Gouda 4wk30.

Nombre de segments	**2**	**3**	**4**	**5**
Indice du rand ajusté	0.024	0.168	0.215	0.181

Nous proposons d'utiliser l'indice de Rand ajusté (Hubert & Arabie, 1985) pour comparer la segmentation des k-means et celle du modèle de mélange (Tableau 17). Cet indice prend comme valeur 0 dans le cas d'une partition totalement différente et 1 si les deux partitions sont identiques. Nous constatons que les valeurs observées sont faibles ce qui signifie que les classifications obtenues avec les k-means et celles obtenues avec le modèle de mélange sont différentes.

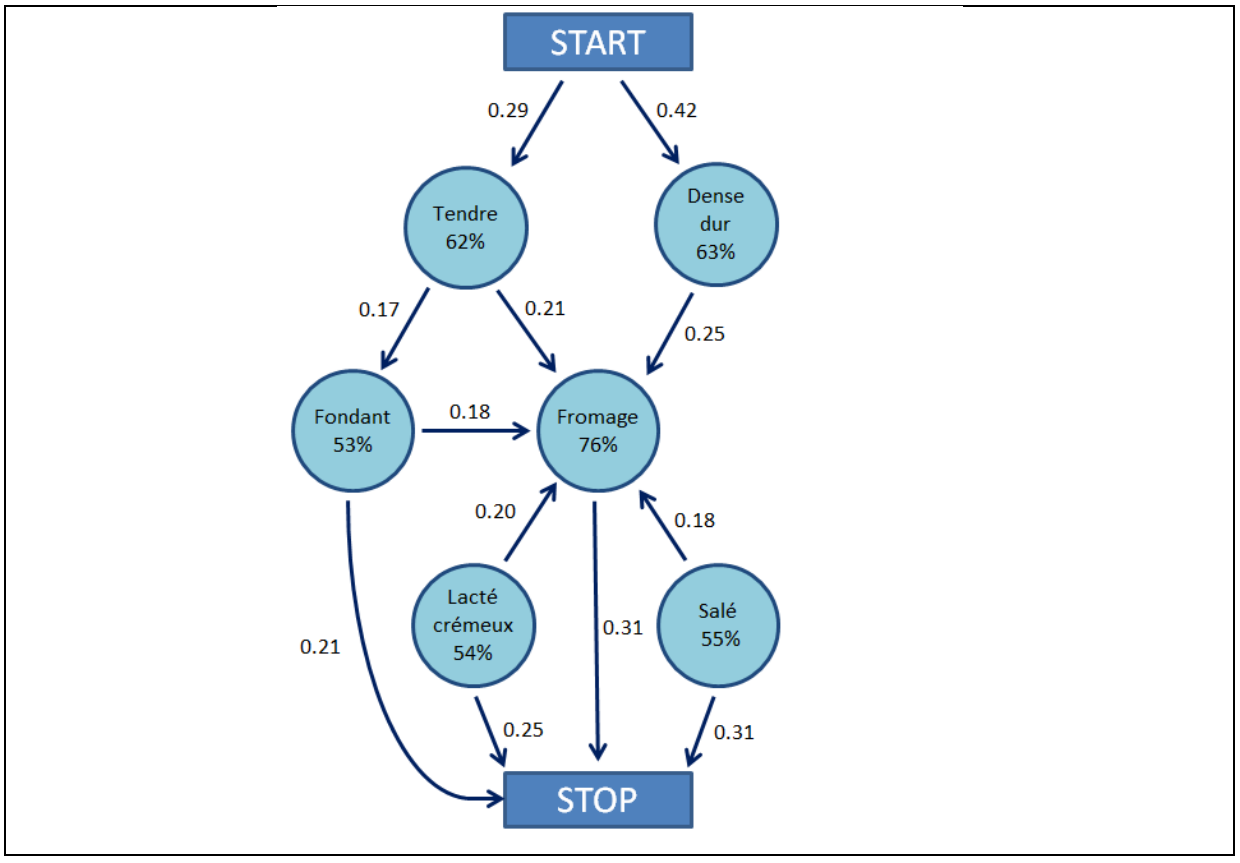


Figure 49 Graphe DTS du Gouda 4wk30.

Le graphe DTS du Gouda 4wk30 (Figure 49) montre qu'il existe certainement des différences de perception pour ce Gouda au sein du panel avec notamment des chemins dans le graphe qui s'opposent avec Dense dur d'un côté contre Tendre et Fondant de l'autre.

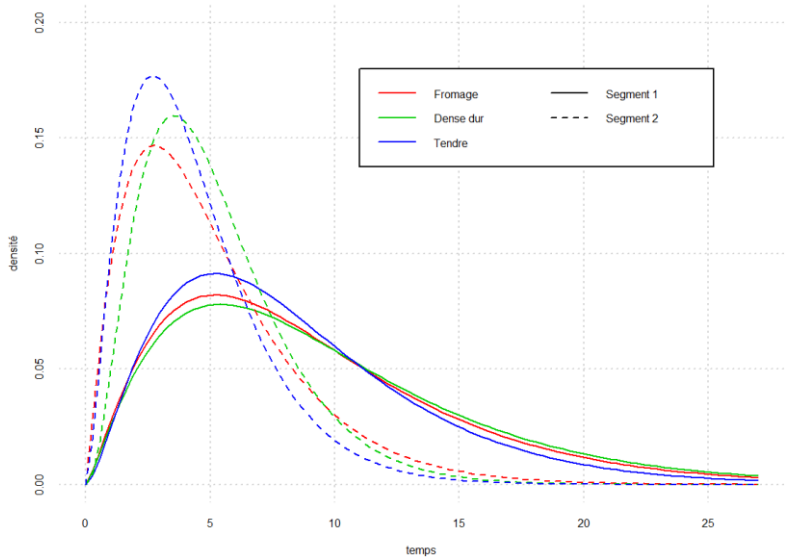


Figure 50 Temps de séjour estimés pour les descripteurs "Fromage", "Dense dur", et "Tendre" pour un mélange avec 2 composantes pour le Gouda 4wk30.

La Figure 50 présente les distributions Gamma estimées pour les descripteurs Fromage, Dense-dur et Tendre pour la segmentation en 2 segments du Gouda 4wk30. Nous constatons que pour chaque segment les distributions des différents descripteurs sont très similaires alors que les distributions sont très différentes entre segments. Le segment 1 est caractérisé par des durées de dominance plus longues que celles du segment 2.

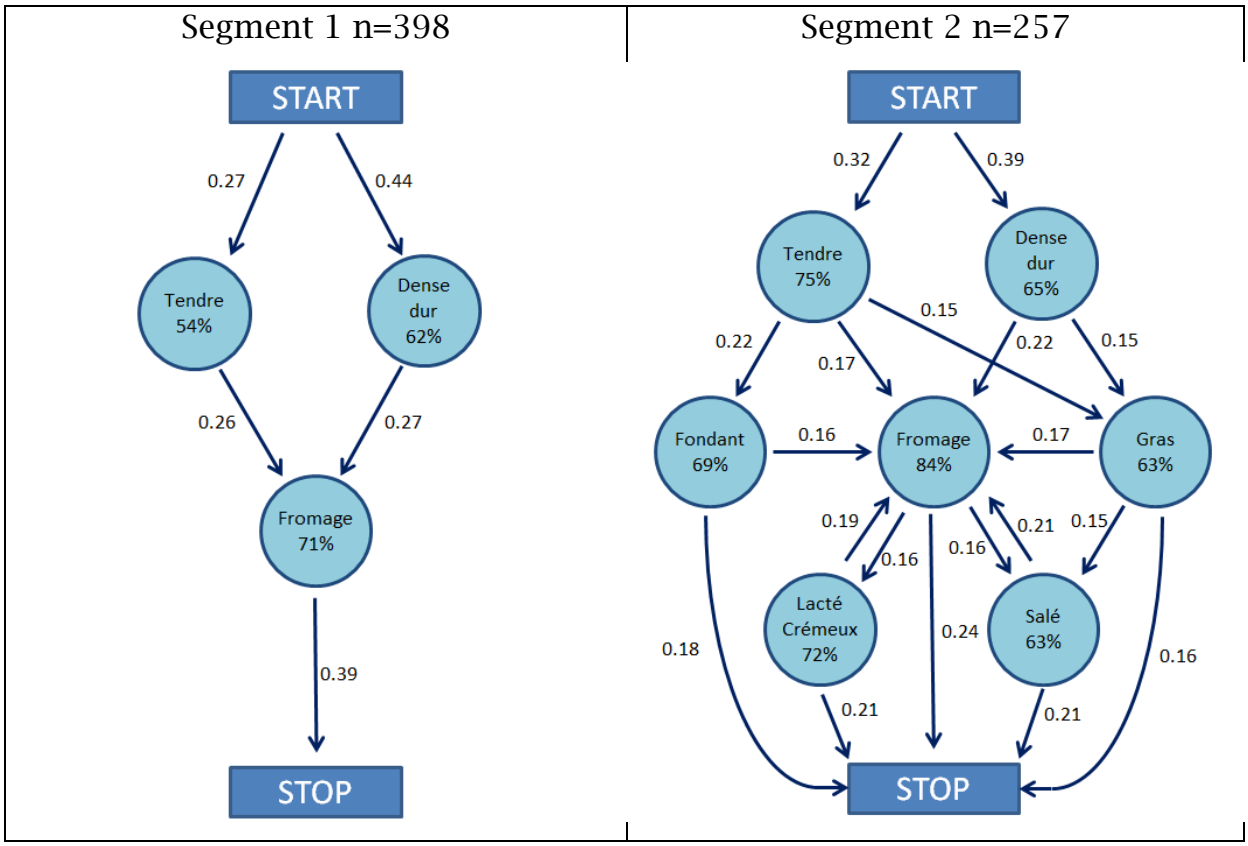


Figure 51 Graphes DTS des 2 segments lorsque le panel est segmenté en 2 pour le Gouda 4wk30.

La segmentation en 2 segments semble se baser sur des différences de comportement. Dans le graphe du premier segment il y a peu de descripteurs alors qu'il y en a beaucoup dans le deuxième segment (Figure 51). Les panélistes du segment 1 ont semble-t-il choisis peu de descripteurs et ce sont contentés des descripteurs les plus évidents. Au contraire les panélistes du deuxième segment ont choisis beaucoup de descripteurs avec des changements fréquents.

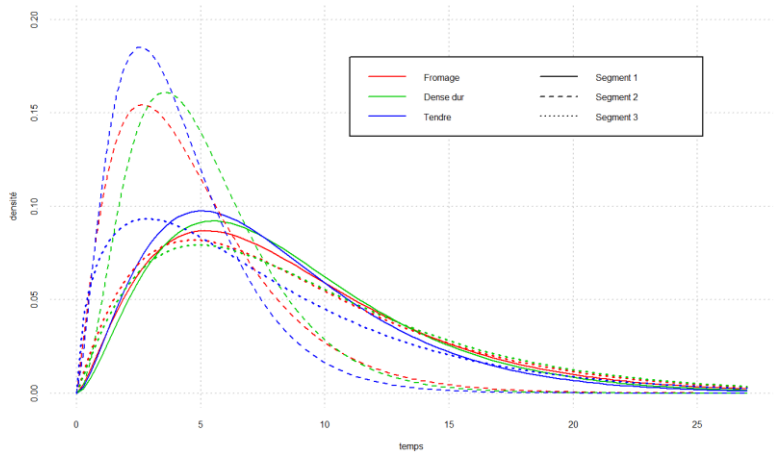


Figure 52 Temps de séjour estimés pour les descripteurs "Fromage", "Dense dur", et "Tendre" pour un mélange avec 3 composantes pour le Gouda 4wk30.

La Figure 52 présente les distributions Gamma pour la segmentation en 3 segments. Le segment 2 semble similaire au segment 2 de la segmentation en 2 segments. Les segments 1 et 3 sont très proches du segment 1 de la segmentation en 2 segments sauf pour le descripteur Tendre du segment 3 qui a des probabilités plus élevées d'avoir des durées de dominance courtes.

Pour la segmentation en 3 segments nous constatons que le graphe du segment 2 (Figure 53) est très similaire au segment 2 de la segmentation en 2 segments. Les segments 1 et 2 correspondent à deux perceptions très différentes. Les panélistes du segment 1 ont majoritairement choisi Tendre comme premier descripteur puis Lacté-Crémeux, Gras ou Fondant alors que ceux du segment 3 ont choisi Dense dur puis Fromage et éventuellement Salé.

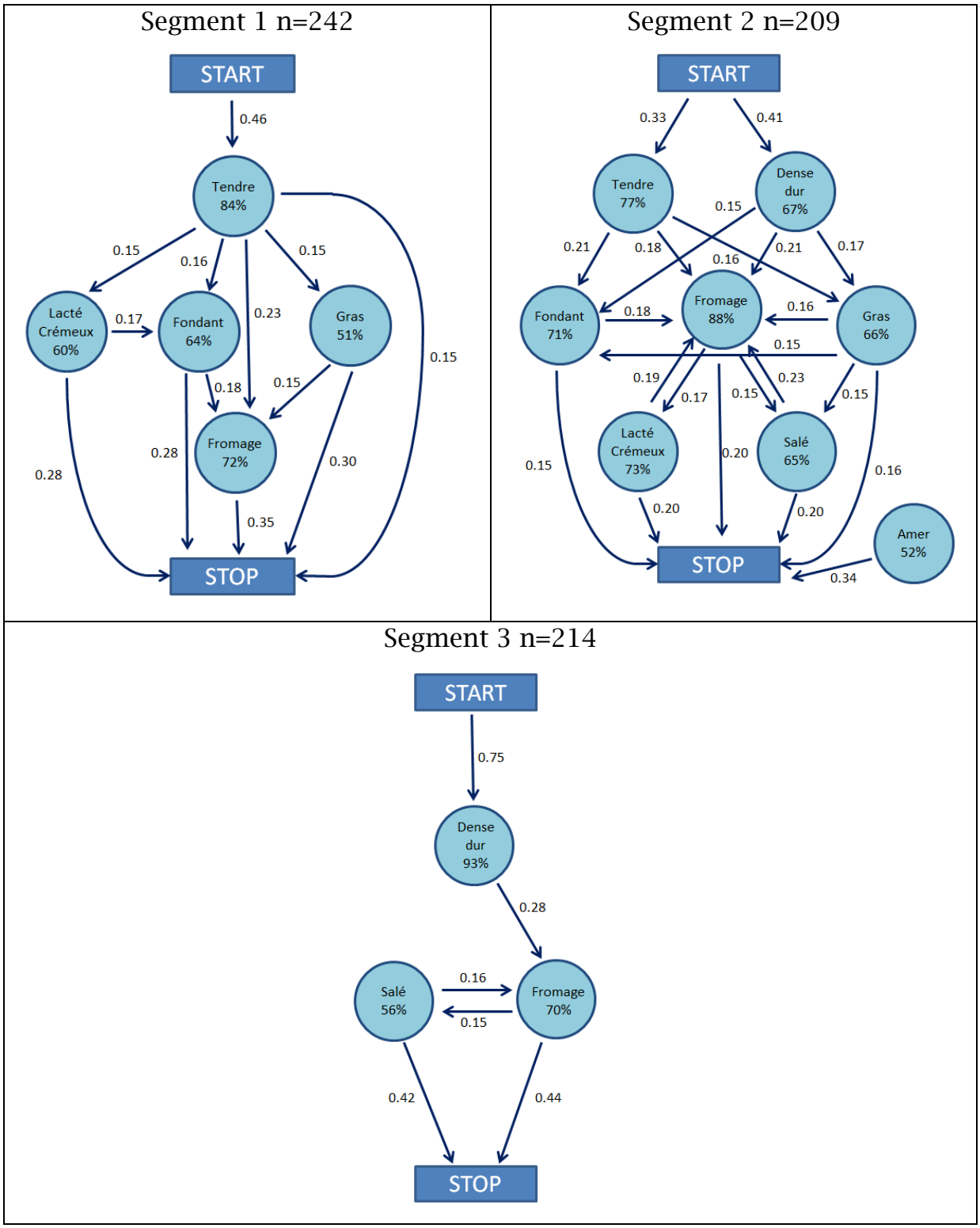


Figure 53 Graphes DTS pour la segmentation du Gouda en 3 segments pour le Gouda 4wk30.

Bien que les critères d'information indiquent de segmenter en 2 ou 3 segments, il peut être intéressant d'observer la segmentation obtenue avec un plus grand nombre de segments. En effet, ces critères offrent un compromis entre amélioration de la modélisation et parcimonie quant au

nombre de paramètres, mais, comme le modèle comprend un grand nombre de paramètres, les critères risquent de choisir peu de segments même s'il y a une grande diversité de perceptions au sein du panel. Nous allons donc observer la segmentation obtenue pour 4 et 5 segments.

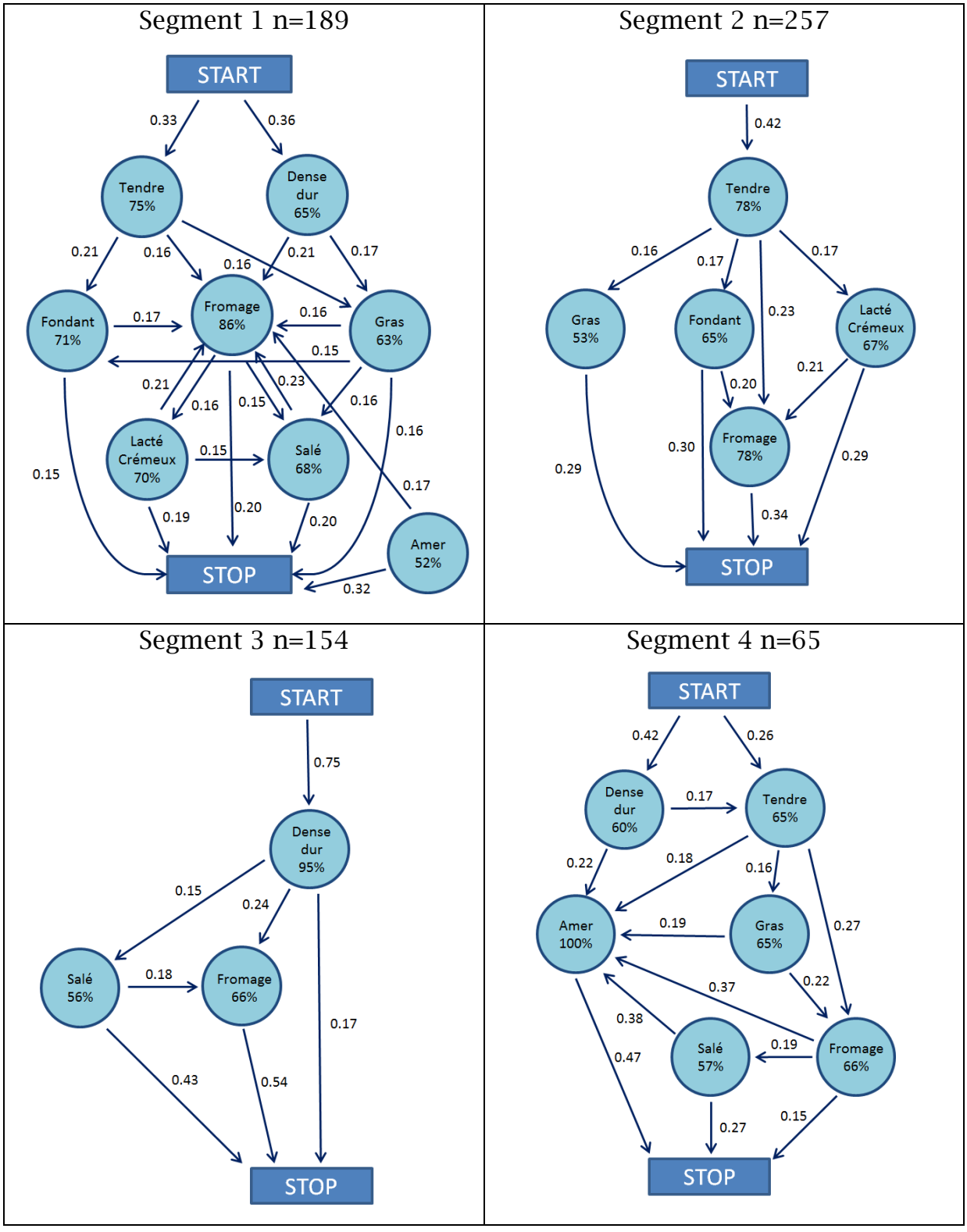
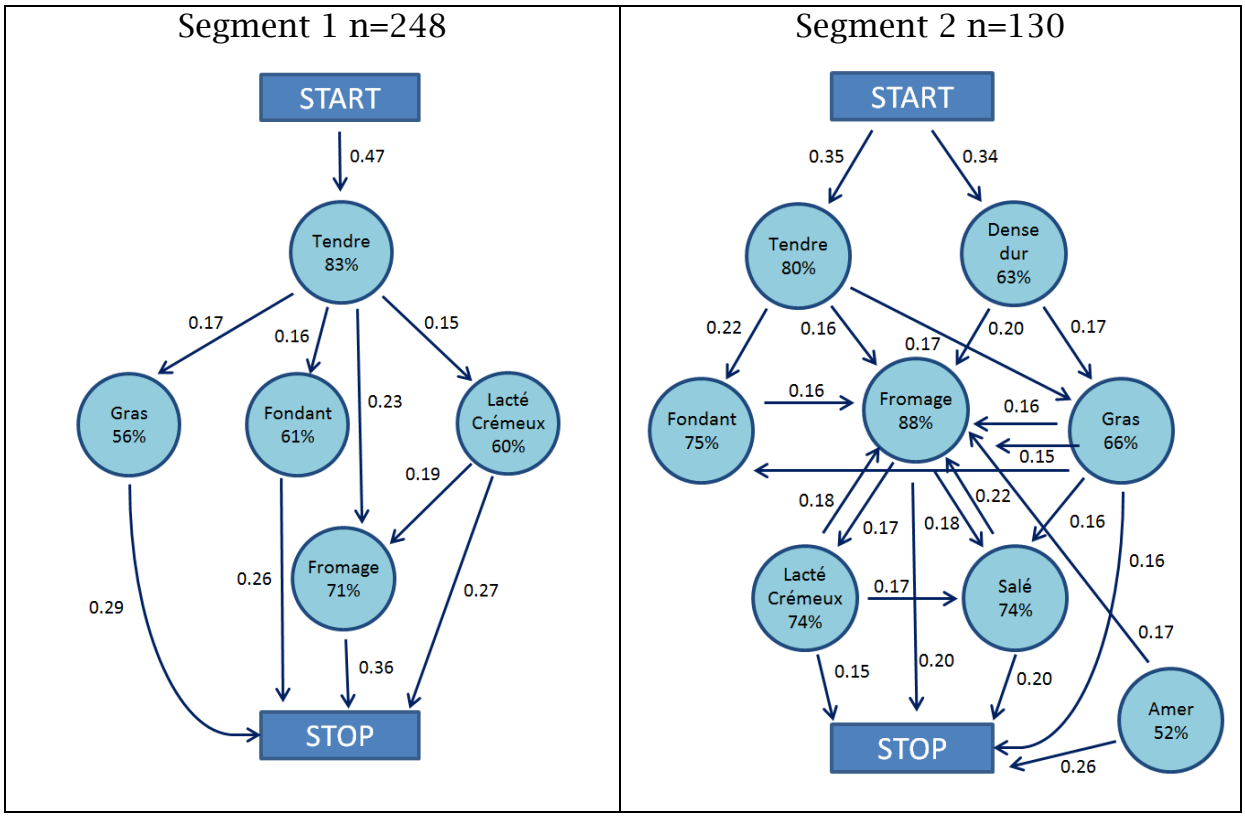


Figure 54 Graphes DTS pour la segmentation du Gouda en 4 segments pour le Gouda 4wk30.

En segmentant le panel en 4, nous constatons que le segment 1 avec 189 panélistes est similaire au segment 2 lorsque le panel est segmenté en 3, que le segment 2 qui comprend 257 panélistes est similaire au segment 1 et que le segment 3 composé de 154 panélistes est similaire au segment 3 vu précédemment (Figure 54). Le segment 4 en revanche se distingue de ce qui a été vu lors des segmentations précédentes par le descripteur Amer qui est utilisé par l'ensemble des panélistes composant ce groupe et vers lequel il y a le plus de transitions. Tous les descripteurs ont une probabilité supérieure à 0,15 d'être suivi par Amer et cette probabilité est particulièrement élevée pour Fromage et Salé avec respectivement 0,37 et 0,38. Ce groupe n'est toutefois composé que de 65 panélistes.



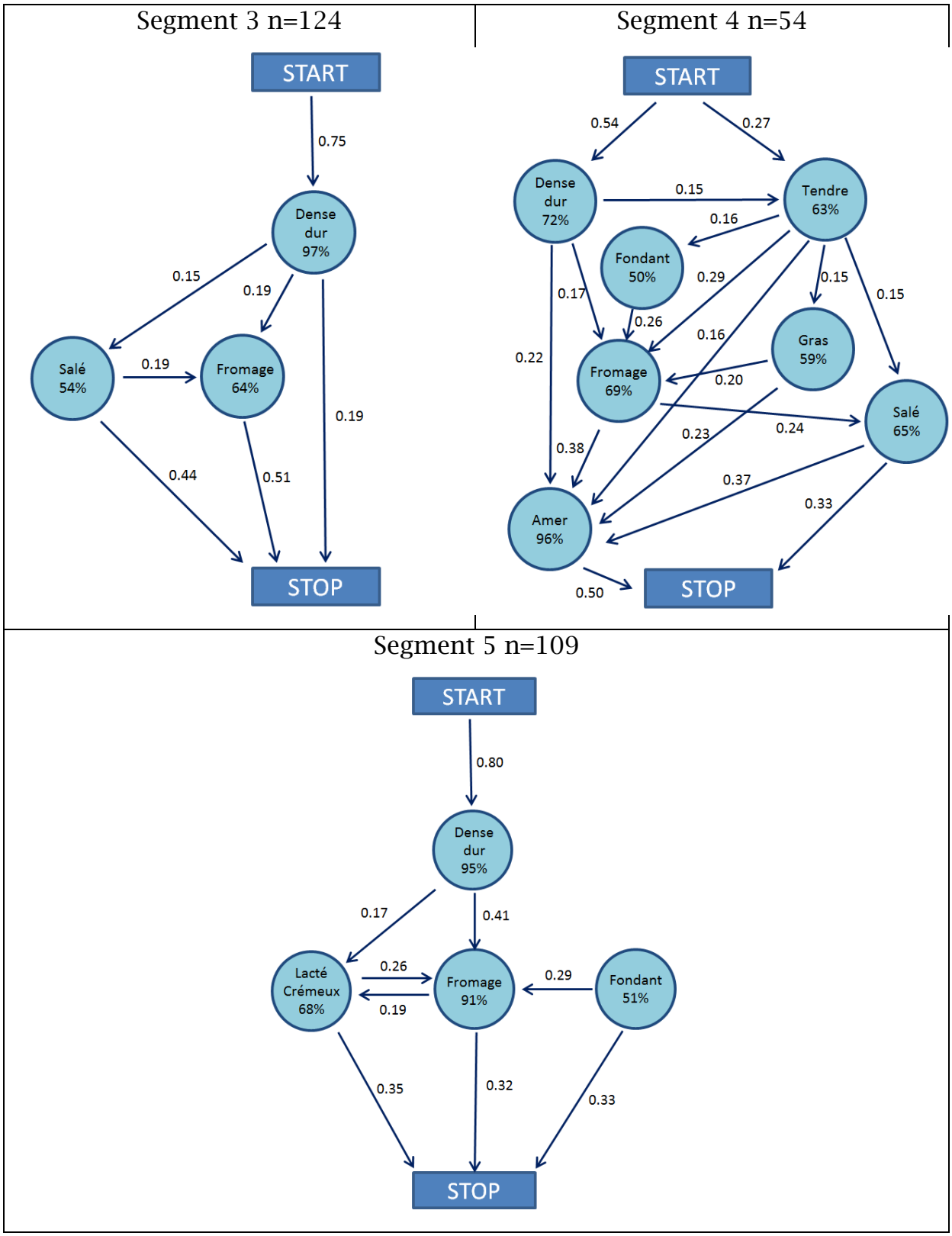


Figure 55 Graphes DTS pour la segmentation du Gouda en 5 segments pour le Gouda 4wk30.

En segmentant en 5, seul le segment 5 composé de 109 panélistes se distingue de ce qui a été vu précédemment avec Dense-dur comme premier descripteur puis Fromage ou Lacté-crémeux et également la présence de

Fondant (Figure 55). Il est à noter que le segment 4 ne regroupe que 54 panélistes ce qui représente moins de 10% du panel. Cette segmentation en 5 est en plus difficile à interpréter d'un point de vue sensoriel. Il faut donc sans doute sélectionner 4 segments pour ce produit.

Corrélation entre segmentation et note hédonique

En plus du protocole DTS, les panélistes ont donné une note hédonique pour ces Goudas et nous allons donc pouvoir observer si les différences de perception observées avec la segmentation ont une influence sur l'appréciation. Pour cela nous avons réalisé une ANOVA expliquant les notes hédoniques par les segments pour chacune des segmentations.

Tableau 18 Notes hédoniques moyennes observées pour le Gouda 4wk30 avec une segmentation du panel allant de 2 à 5 segments et p-values obtenues pour l'ANOVA expliquant les notes de liking par le segment auquel les panélistes appartiennent.

	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	P-value
G=2	5,54	5,66				0,328
G=3	5,77	5,68	5,31			0,004
G=4	5,72	5,83	5,24	5,11		$< 10^{-3}$
G=5	5,56	5,80	5,16	5,24	6,07	$< 10^{-3}$

Le Tableau 18 nous donne les résultats. Nous constatons qu'avec 2 segments les moyennes sont très proches et la p-value de l'ANOVA est égale à 0,328 ce qui signifie que l'appartenance à l'un ou l'autre des segments n'a pas d'effet sur l'appréciation de ce Gouda. Ce résultat semble logique puisque cette segmentation a mis en avant des différences de comportement plus que des différences de perception. Avec 3 segments, si les deux premiers segments ont des notes hédoniques moyennes proches de 5,70, le troisième segment a en revanche une note moyenne plus faible égale à 5,31 et l'ANOVA indique que l'effet du segment sur la note hédonique est significatif. Le troisième segment ce caractérise par une perception Dense dur de ce produit ce qui semble avoir un effet négatif sur son appréciation. La segmentation en 4 segments oppose les deux premiers segments aux deux derniers en termes de note hédonique. Les deux premiers segments sont ceux qui ont le plus

apprécié ce Gouda et ils se distinguent par la présence de Fondant et de Lacté crémeux alors que les segments 3 et 4 sont respectivement les consommateurs ayant trouvé ce fromage Dense-dur et Amer. Finalement la moyenne hédonique la plus élevée observée est celle du segment 5 lors de la segmentation en 5 segments avec 6,07. Ce segment est caractérisé par Dense-dur puis Lacté crémeux, Fromage et Fondant.

3.6.2 Fromages frais

Cette partie présente la segmentation pour le fromage frais P6.

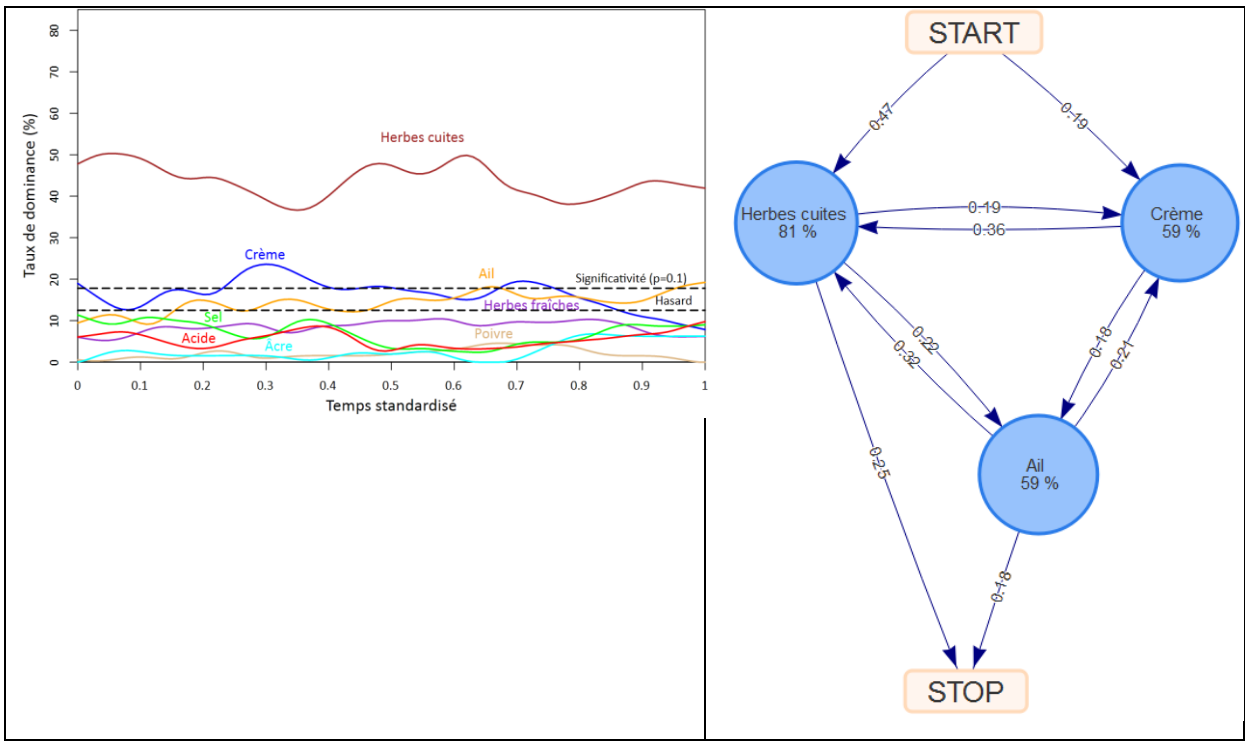


Figure 56 Courbes et graphe DTS du fromage frais P6.

Le fromage frais P6 est particulièrement intéressant à segmenter puisque sa perception semble, au premier regard des courbes, être extrêmement simple en se limitant au descripteur Herbes cuites. Toutefois les descripteurs Crème et Ail ressortent également légèrement dans les courbes et sont présents dans le graphe (Figure 56). Le graphe montre que Herbes cuites ou Crème sont sélectionnés comme premier descripteur et qu'ensuite les descripteurs Herbes cuites, Crème et Ail semblent être sélectionnés alternativement et que l'ordre a peu d'importance.

Tableau 19 Valeurs prises par les critères d'information pour un nombre de segment G allant de 1 à 5 lors de la segmentation du fromage frais P6.

	G=1	G=2	G=3	G=4	G=5
$\boldsymbol{BIC}$	4931,90	5182,04	5435,17	5765,79	6075,57
$\boldsymbol{AIC}$	4744,08	4804,23	4867,39	5008,02	5127,822
$\boldsymbol{AIC_C}$	4106,08	4254,23	4173,07	4150,02	4100,375

Nous avons vu précédemment que du point de vue de l'analyse sensorielle il peut être intéressant de ne pas se limiter à la segmentation avec le nombre de segments sélectionnés par les critères d'information. Le nombre plutôt faible de panélistes limite aussi la confiance que l'on peut avoir dans les critères d'information. Pour le fromage frais P6 les critères AIC et BIC indiquent qu'un seul segment serait suffisant alors que AIC_c prend des valeurs proches pour 1 et 5 segments (Tableau 19). Nous proposons d'utiliser de 2 à 4 segments et d'observer si ces segmentations ont du sens.

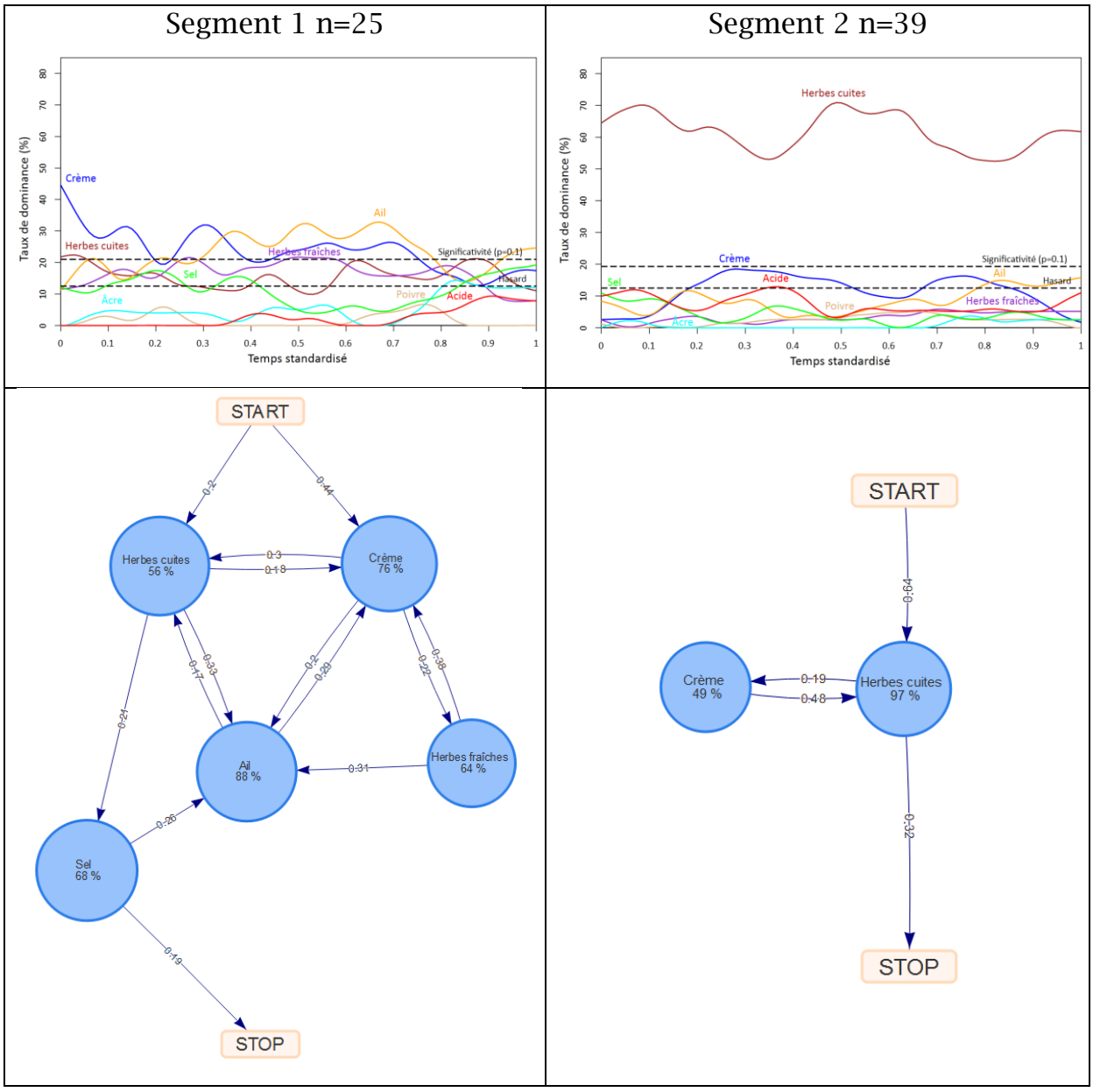


Figure 57 Courbes et graphes DTS des segments obtenus pour le fromage frais P6 lors de la segmentation en 2 segments.

Pour la segmentation en 2 groupes, le segment 2 est conforme à l'image donnée par l'ensemble du panel avec la perception presque uniquement du descripteur Herbes cuites (Figure 57). En revanche le segment 1 montre une perception plus riche de ce produit avec Herbes fraîches et Sel qui apparaissent dans le graphe et Herbes cuite qui est beaucoup moins présent.

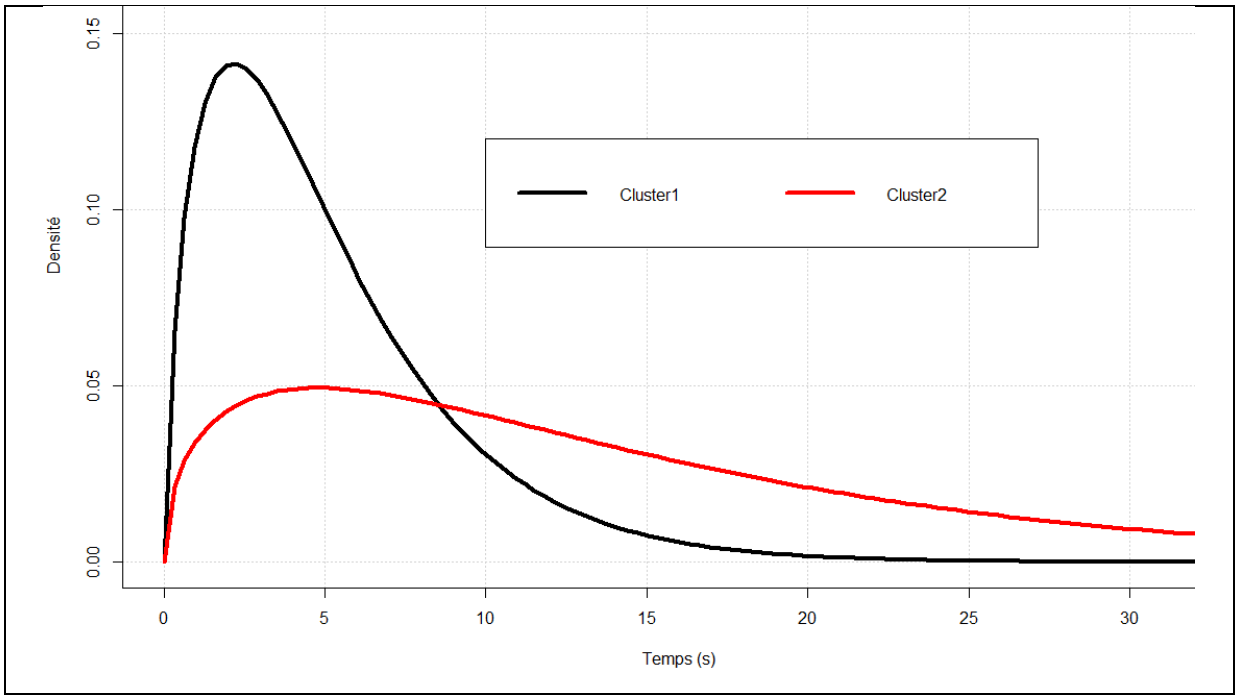
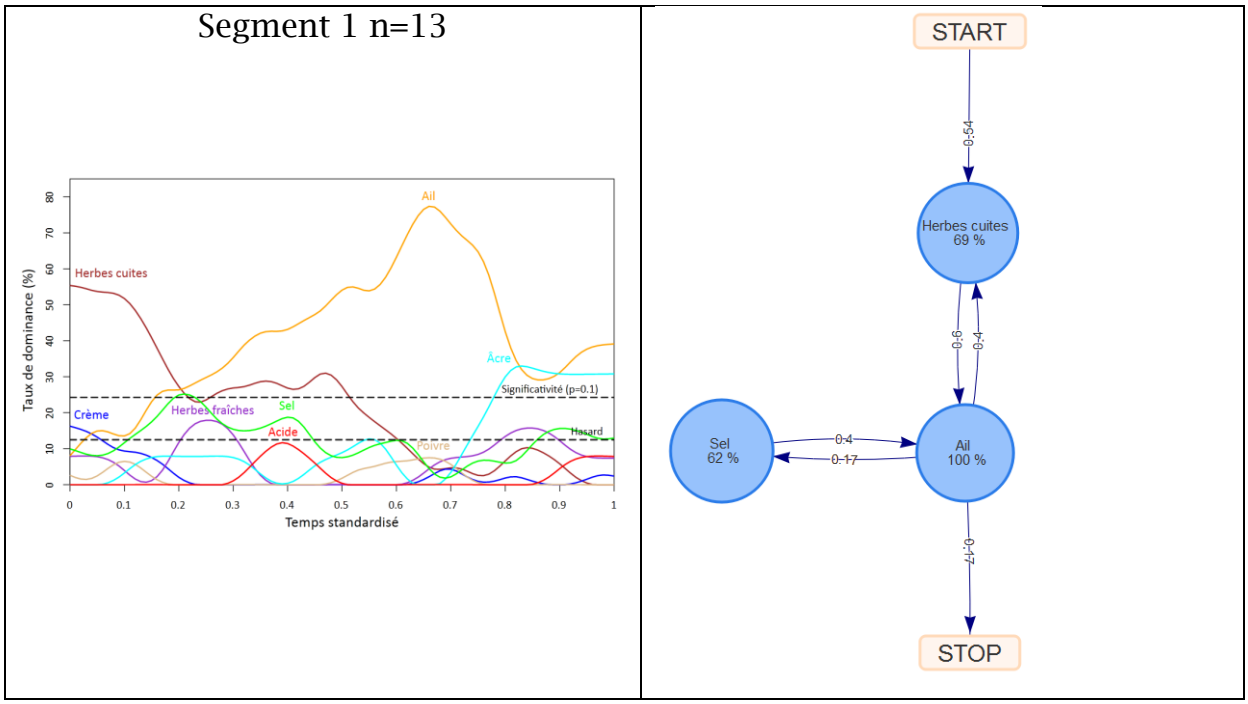


Figure 58 Lois Gamma estimées pour le descripteur Herbes cuites pour les 2 clusters obtenus pour le Gouda 4wk30.

Nous observons que les durées de dominance sont beaucoup plus longues pour les panélistes appartenant au deuxième segment (Figure 58). Les panélistes de ce segment ont choisi majoritairement Herbes cuites et ont conservé ce descripteur dominant plus longtemps.



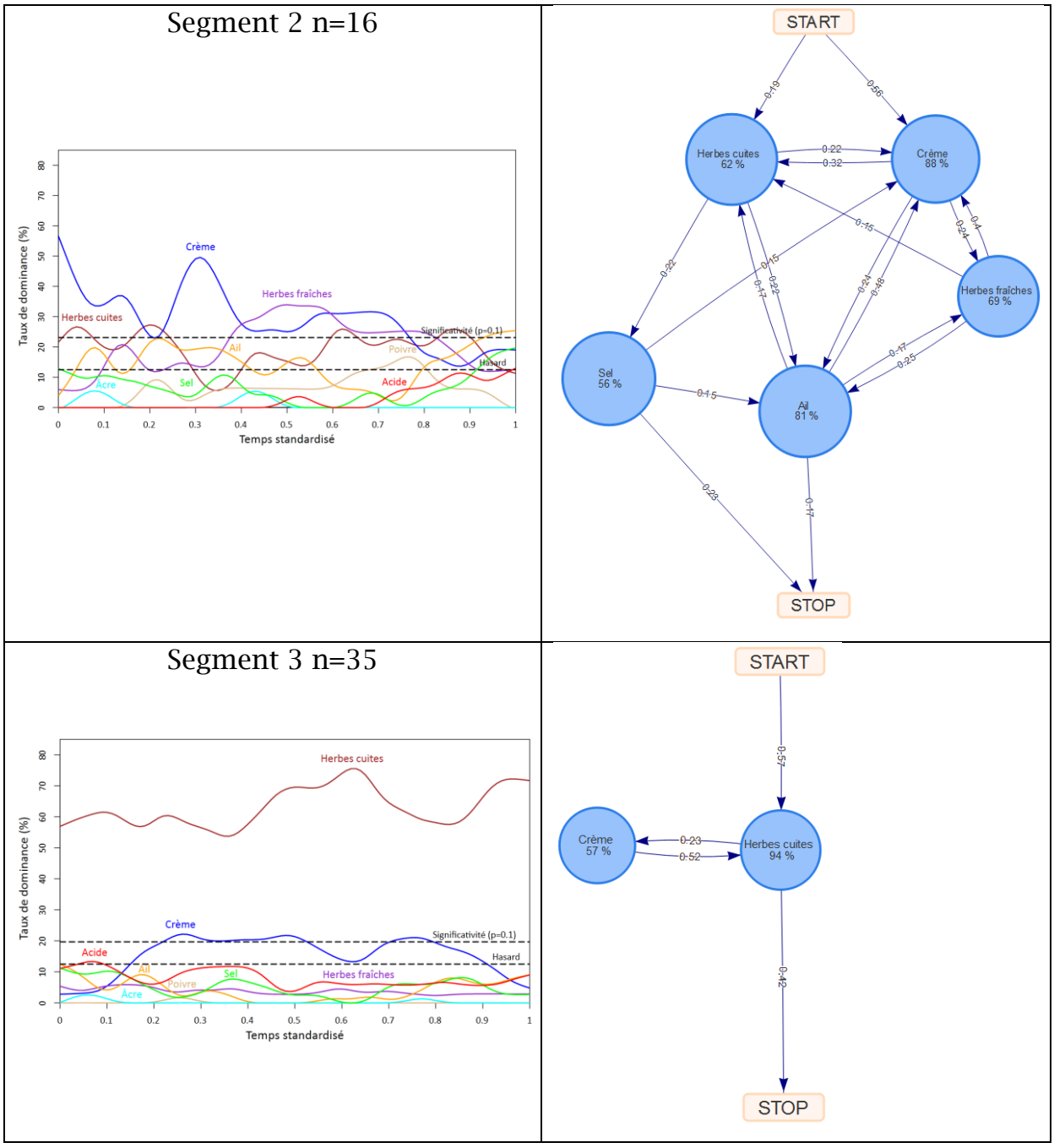
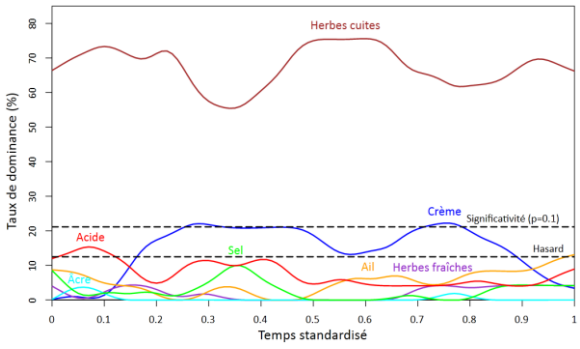


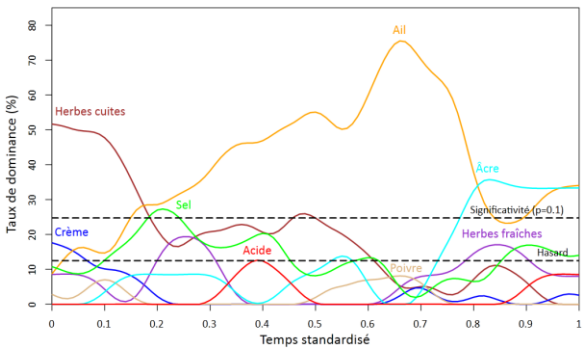
Figure 59 Courbes et graphes DTS des segments obtenus pour le fromage frais P6 lors de la segmentation en 3 segments.

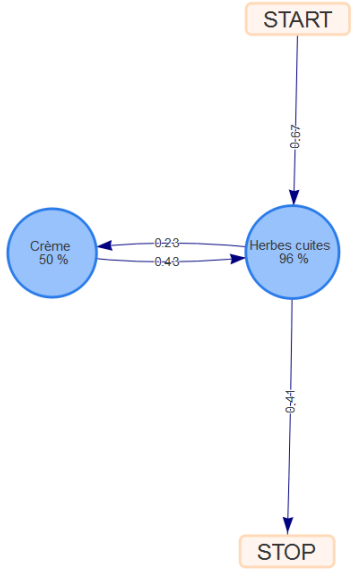
Pour la segmentation en 3 segments (Figure 59), nous constatons que le segment 3 est similaire au segment 2 vu précédemment avec la prédominance du descripteur Herbes cuites. Le premier segment se distingue particulièrement par le descripteur Ail alors que le deuxième segment lui se distingue par Crème et Herbes fraîches.

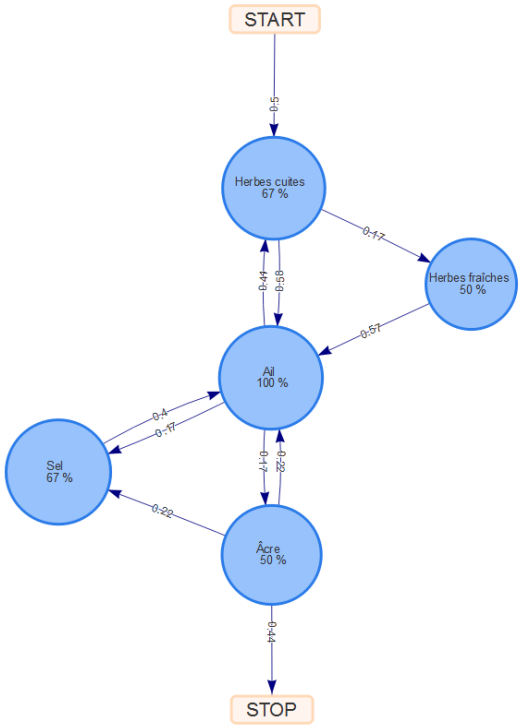
Segment 1 n=24



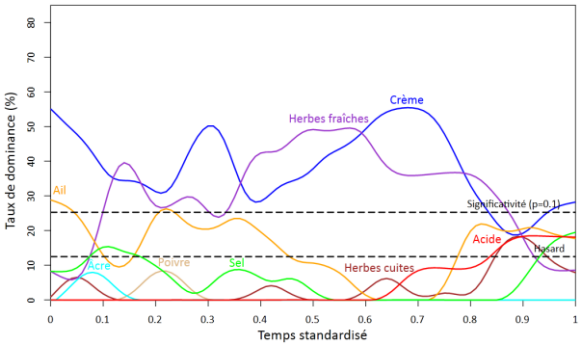
Segment 2 n=12



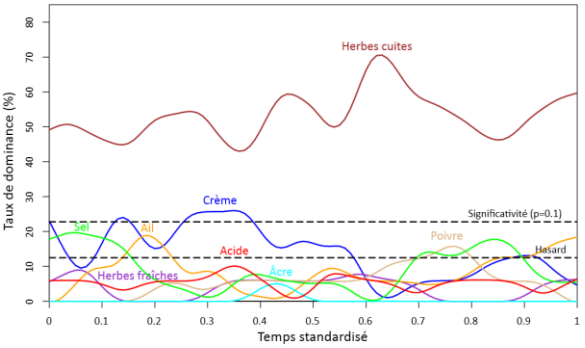




Segment 3 n=11



Segment 4 n=17



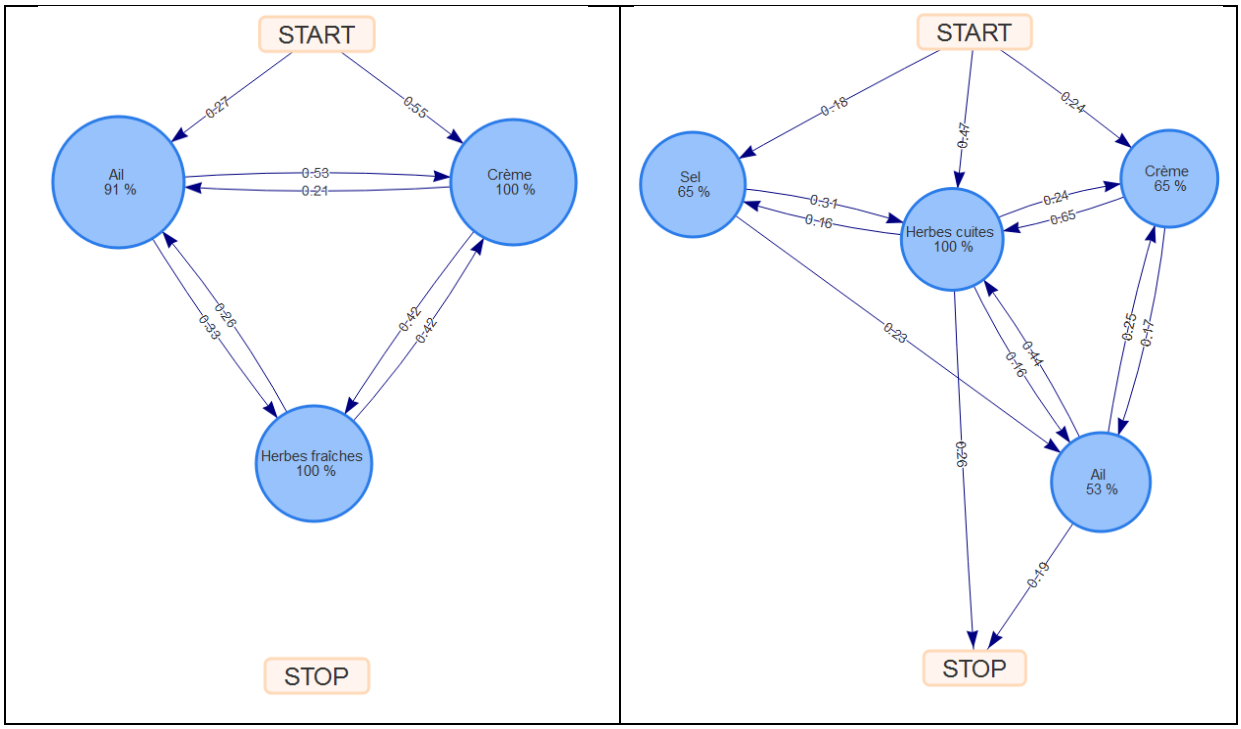


Figure 60 Courbes et graphes DTS des segments obtenus pour le fromage frais P6 lors de la segmentation en 4 segments.

La segmentation en 4 segments consiste, semble-t-il, à segmenter le segment 3 vu précédemment, ce qui donne les segments 1 et 4 (Figure 60) qui sont tous les deux largement caractérisés par Herbes cuites mais se distinguent par la présence de Sel et Ail dans le segment 4 et pas dans le 1. Cette différence est tout de même assez faible et il n'est donc peut-être pas nécessaire d'aller jusqu'à 4 segments.

Tableau 20 Notes hédoniques moyennes observées pour le fromage frais P6 avec une segmentation du panel allant de 2 à 4 segments et p-values obtenues pour l'ANOVA expliquant les notes hédoniques par le segment auquel les panélistes appartiennent.

	Segment 1	Segment 2	Segment 3	Segment 4	P-values
G=2	4,38	3,64			0,205
G=3	4,04	4,41	3,67		0,558
G=4	3,98	4,21	4,68	3,18	0,356

Des différences importantes d'appréciation existent entre les segments mais elles ne sont pas statistiquement significatives (Tableau 20) certainement à cause de la taille des échantillons et de la forte dispersion des valeurs des notes hédoniques au sein du panel. En regardant l'ensemble des

segmentations, il existe une tendance avec des notes hédoniques plus faibles pour les segments ayant principalement perçu Herbes cuites alors que les segments caractérisés par Crème et Herbes fraîches ont les notes les plus élevées mais ces valeurs sont calculées sur des petits échantillons et si elles sont montrées ici ce n'est qu'à titre illustratif.

3.7 Bilan

Ce chapitre présente les applications des méthodes vues dans le chapitre 1 à différents jeux de données : des chocolats Barry-Callebaut et Lindt Excellence, des fromages frais et des Goudas.

Dans un premier temps, nous avons vu que l'hypothèse markovienne est adaptée aux données DTS et que la loi Gamma améliore la modélisation des durées de dominance comparée à la loi exponentielle. Nous avons également observé que les données simulées à partir du modèle semi-markovien ressemblent aux données utilisées pour estimer le modèle ce qui confirme la qualité de la modélisation. Nous avons ensuite vu l'intérêt de cette modélisation à travers une nouvelle représentation graphique, le graphe DTS, qui permet une meilleure compréhension des données DTS en suggérant différentes séquences de perception des sensations procurées par un même produit. Le chapitre s'est poursuivi par le découpage de la durée de dégustation en périodes qui met en avant la temporalité de la perception et améliore encore la compréhension de la perception des produits. Nous avons alors vu les résultats pour le test de différence entre 2 échantillons DTS mais ils sont assez mitigés avec des résultats différents de ceux obtenus par les analyses conventionnelles. Finalement, nous avons vu des résultats de segmentation du panel basée sur les différences de perception d'un même produit. L'existence d'importantes différences de perception temporelle a ainsi pu être mise en évidence au sein d'un même panel.

L'approche proposée utilisant les processus semi-markoviens permet de répondre de manière cohérente aux différentes problématiques posées par les données DTS. Cet ensemble d'outils permet une exploration plus complète des données DTS par les animateurs de panel, notamment en prenant en compte la diversité des perceptions au sein du panel.

Chapitre 4.

Discussion

Dans cette partie, nous allons discuter les différents résultats obtenus au chapitre précédent ainsi que les choix de méthodes. Nous reprendrons la logique des deux chapitres précédents en discutant successivement l'adéquation du modèle aux données DTS, le graphe DTS, le découpage en périodes, le test de différence entre échantillons DTS et enfin la segmentation par perception. Une dernière partie traitera des remarques plus générales sur les travaux présentés dans cette thèse.

4.1 Adéquation du modèle aux données DTS

Conditions d'applicabilité

La modélisation proposée utilise un grand nombre de paramètres, ce qui nécessite d'avoir des panels de grande taille pour que l'estimation puisse être faite correctement. En réalité, au-delà du nombre de panélistes, ce qui est important est le nombre de transitions observées dans le jeu de données puisque les transitions et les durées de dominance associées constituent les individus statistiques utilisés pour l'estimation. La DTS est de plus en plus utilisée avec des consommateurs, ce qui permet d'avoir des tailles de panels relativement élevées. Avec des panélistes entrainés, il est commun de réaliser des répétitions de la dégustation de chaque produit, ce qui va multiplier le nombre d'observations, si du moins, comme à l'accoutumée, elles sont considérées comme des observations indépendantes.

Par exemple, pour un jeu de données avec 10 descripteurs, il y a 10*11=110 paramètres dans la matrice de transition (avec les transitions vers STOP) mais il est impossible de rester dans le même descripteur donc il n'est pas nécessaire d'estimer la diagonale de la matrice de transition. Chaque ligne se sommant à 1, il est possible de déduire un paramètre à partir des autres, il y a donc 2*10=20 paramètres qui ne nécessitent pas d'être estimés. Il y a ainsi 90 paramètres pour les transitions, 9 pour les probabilités initiales et 20 pour estimer les lois Gamma pour chaque descripteur soit un total de 119 paramètres. Pour un jeu de données avec 60 panélistes et une moyenne de 5 transitions observées par séquence, il y a 60 observations pour estimer les 9 paramètres des probabilités initiales, 300 observations pour estimer les 90

paramètres des transitions et 300 observations pour estimer les 20 paramètres des lois des durées de dominance. Si le nombre d'observations semble raisonnable pour estimer les probabilités initiales et les durées de dominance, il serait souhaitable d'avoir des panels avec au moins 100 panélistes pour estimer les transitions, ce qui permettrait d'avoir au moins 500 observations pour estimer 90 paramètres.

Des exemples sont donnés dans le Tableau 21 pour illustrer le nombre de panélistes nécessaires pour avoir au moins 5 observations en moyenne par paramètre à estimer. Nous observons que pour les valeurs couramment observées de 10 descripteurs et 4 ou 5 transitions en moyenne, le nombre de panéliste nécessaire est effectivement une centaine.

Tableau 21 Nombre de panélistes nécessaire pour avoir en moyenne 5 observations par paramètre de la matrice de transition à estimer en fonction du nombre de descripteurs D et du nombre moyen de transitions observé dans l'étude NbTrans.

	D=6	**D=8**	**D=10**	**D=12**
NbTrans =3	n=50	n=93	n=150	n=220
NbTrans =4	n=38	n=70	n=113	n=165
NbTrans =5	n=30	n=56	n=90	n=132
NbTrans =6	n=25	n=47	n=75	n=110

Ces effectifs sont à mettre en perspective avec les simulations réalisées dans notre deuxième article (Annexe 2, p207) qui montrent que l'estimation du modèle de mélange peut être compliquée avec seulement 60 panélistes lorsque les segments sont proches et encouragent donc à avoir des panels de plus grande taille.

Standardisation

Dans les travaux présentés ici, les séquences DTS ont été standardisées à gauche puisque le temps avant le premier clic n'est pas modélisé. En revanche, nous déconseillons de standardiser les durées des séquences, c'est-à-dire pour chaque séquence de diviser les durées de dominance par la durée totale de la séquence, car cette standardisation déforme les durées de

dominance alors que nous cherchons à les modéliser. Cette standardisation est réalisée uniquement pour découper le temps de dégustation en périodes afin de faciliter l'implémentation informatique et de pouvoir représenter les frontières des périodes sur les courbes DTS. Le découpage en périodes étant basé uniquement sur les transitions et pas sur les durées de dominance et les frontières étant déterminées comme un pourcentage de la durée totale de chaque séquence, cette standardisation n'a aucun effet ni sur la détermination du nombre de périodes, ni sur la position des frontières entre périodes.

Transitions vers STOP

Nous avons fait le choix d'inclure dans le modèle les transitions vers STOP. L'ajout de l'état STOP diminue mécaniquement les probabilités de choisir les autres descripteurs mais cette déformation des données correspond certainement à la réalité puisque pour les descripteurs perçus à la fin de la dégustation il y a une certaine probabilité que rien ne soit perçu ensuite (ce qui correspond à la probabilité de choisir STOP). Ce choix permet de prendre en compte un maximum d'information afin d'exploiter le modèle pour tester des différences entre produits ou segmenter le panel par perception. L'état STOP a toutefois été exclu pour découper la durée de dégustation en périodes car il y aurait alors mécaniquement une dernière période constituée simplement du dernier instant de dégustation qui correspond à l'instant d'observation de toutes les transitions vers STOP. L'état STOP est également exclu du modèle lorsque nous réalisons des simulations. En effet, si nous réalisons les simulations en prenant en compte l'état absorbant STOP, alors la logique voulait que nous simulions jusqu'à ce que cet état soit atteint. Dans ce cas, nous avons observé une grande disparité dans la longueur des séquences simulées qui ne correspondait pas à ce qui est observé dans un panel. Une autre approche, que nous n'avons pas encore explorée, consisterait à réaliser un grand nombre de simulations et à ne sélectionner que les séquences simulées répondant à certains critères, par exemple, nombre de transitions ou durée totale.

Différences de durées de dominance entre descripteurs

Nous avons vu que les distributions des durées de dominance peuvent parfois être extrêmement proches d'un descripteur à un autre. On pourrait donc, comme première étape de la modélisation, tester si les distributions des durées dominance sont similaires et, si c'est le cas, modéliser ensemble les durées de dominance des descripteurs correspondants. Ceci permettrait de diminuer le nombre de paramètres du modèle.

Pondération par séquence

Nous avons vu précédemment que l'estimation se faisait sur les transitions et les durées de dominance associées, mais chaque séquence n'a pas le même nombre de transitions et n'a donc pas la même influence sur l'estimation du modèle. Une solution peut être d'estimer le modèle en pondérant chaque séquence en fonction de l'inverse du nombre de transitions observées. Toutefois, nous considérons qu'il est anormal que certaines transitions aient plus de poids que d'autres et cette solution ne ferait donc que déporter le problème.

4.2 Graphe DTS

Le graphe DTS permet une représentation plus précise de la perception d'un produit par un panel que les courbes DTS en suggérant différentes séquences de perception des descripteurs dominants. Les courbes peuvent être mal interprétées en considérant que la séquence de descripteurs successivement les plus dominants correspond à la séquence perçue par la plupart des panélistes, alors que ce sont généralement des panélistes différents qui perçoivent les descripteurs au moment des pics observés dans les courbes, comme nous avons pu le voir grâce à la segmentation. Les graphes permettent également de suggérer des associations de descripteurs, comme par exemple Cacao et Amer pour des chocolats noirs, signifiant que l'un ou l'autre a pu être employé pour désigner la même sensation. Dans ce cas, l'animateur de panel peut s'interroger sur la pertinence de conserver autant de descripteurs pour ses prochaines études sur le même type de produits. La méthode DTS étant contrainte quant au nombre total de descripteurs, une réduction de la taille de la liste de ces derniers est souvent appréciable. Le graphe permet aussi de montrer des situations dans lesquelles les descripteurs sont perçus au même moment, plutôt que selon une séquence temporelle, comme par exemple pour les fromages frais.

Descripteurs significatifs dans les courbes et absents dans le graphe

Nous avons vu que certains descripteurs qui dépassent légèrement la ligne de significativité dans les courbes ne sont pas présents dans les graphes. Ceci doit nous conduire à considérer avec précaution l'information convoyée par les courbes car il est fort probable que celle-ci ne soit pas robuste. A l'inverse, il peut s'agir du premier descripteur, comme Sec pour le chocolat 811NV qui n'a certes été utilisé que par environ 30% des panélistes mais dans un laps de temps très court et qui constitue donc une information pertinente. Une solution pourrait être d'avoir un seuil de sélection des descripteurs incluant ceux ayant une probabilité initiale dépassant T_{seuil}.

Seuils de sélection

Le choix des seuils pour décider ce qui est affiché dans le graphe est complexe et nous n'avons à l'heure actuelle pas trouvé de solution optimale. Les seuils proposés dans Lecuelle et al. (2018) se sont finalement avérés inadaptés pour certains jeux de données car ils conduisaient à des graphes surchargés et donc illisibles ou à des graphes avec très peu ou pas du tout d'informations. Notre approche a alors consisté à choisir les seuils en fonction des données. Celle-ci s'est avérée pertinente d'un point de vue sensoriel, bien qu'elle n'ait aucune valeur en termes de significativité statistique.

Alternativement, nous pourrions par exemple utiliser la moyenne du pourcentage de panélistes ayant choisi au moins une fois chaque descripteur comme valeur pour D_{seuil}. Cette approche permettrait d'éviter d'avoir un seuil subjectif et éviterait aussi d'avoir à étudier le tableau des pourcentages de panélistes ayant utilisé chaque descripteur pour chaque produit.

Le choix de ne pas inclure dans le graphe les mêmes descripteurs pour tous les produits d'une étude nous empêche de comparer les produits uniquement sur la base des différences de transition ; cependant, les plus grosses différences entre produits sont en général basées sur la perception de descripteurs différents. De plus, un nouveau problème apparaîtrait si nous voulions inclure les mêmes descripteurs dans tous les graphes : si un descripteur est présent dans un graphe alors que très peu de panélistes l'ont utilisé, les probabilités de transition à partir de ce descripteur seraient probablement non robustes.

Variabilité de perception au sein d'un panel

Finalement, l'avantage principal du graphe sur les courbes DTS est certainement de permettre l'observation de la variabilité des perceptions au sein d'un panel. Après avoir utilisé le graphe pour analyser un grand nombre de jeu de données, nous avons observé qu'il n'y avait que rarement une façon unique de percevoir un produit. L'idée de segmenter le panel sur la base des différences interindividuelles de perception temporelle a donc tout son sens.

4.3 Découpage en périodes

Modèle

Nous avons fait le choix de découper la durée de dégustation pour que l'hypothèse d'homogénéité des probabilités de transition soit plus acceptable. Ce découpage permet d'avoir une information sur l'existence d'une évolution des liens entre descripteurs au cours du temps. Précisément, le nombre de périodes ainsi que la position de leurs frontières constituent une information intéressante pour l'animateur de panel.

Il existe (Vergne, 2008) une généralisation de notre approche, connue sous le nom de « Drifting Markov model », qui consiste à remplacer les matrices de transition par période par une fonction linéaire ou polynomiale d'elles-mêmes. D'un point de vue sensoriel, l'interprétation de cette fonction modélisant la « dérive » de la chaine de Markov nous a semblée, si ce n'est impossible, du moins trop ardue.

Descripteurs présents dans les graphes

Les descripteurs affichés dans les graphes pourraient être différents d'une période à l'autre. Cette approche permettrait certainement de prendre en compte les descripteurs choisis uniquement à un moment précis comme Sec pour le chocolat 811NV. Néanmoins, il nous a semblé préférable d'avoir les mêmes descripteurs afin de pouvoir observer l'évolution des transitions entre les périodes. Il faudrait peut-être changer les valeurs inscrites dans les bulles des descripteurs, qui correspondent au pourcentage de panélistes ayant choisi au moins une fois le descripteur, en réalisant ce calcul par période, cependant cette valeur serait grandement affectée par la durée de la période.

Limites du découpage

Le découpage en période est utilisé pour obtenir plus d'informations sur les produits mais cette approche multiplie le nombre de paramètres et, bien qu'elle améliore nettement la qualité de la modélisation, il nous semble

déraisonnable de vouloir l'employer conjointement aux outils développés sur la base de la vraisemblance, tels que le test de différence et la segmentation.

4.4 Test de différence

Méthode

La méthode présentée ici pour réaliser le test de différence permet d'obtenir des résultats cohérents mais qui peuvent être améliorés. Par exemple, les chocolats Excellence Puissant et Prodigieux qui semblent très proches en observant les courbes ou les graphes DTS sont considérés comme largement différents par le test. Pour les permutations présentées dans la partie 3.5.1, des produits très différents obtiennent des p-values autour de 5% alors que les produits identiques dépassent difficilement les 10%. Ces observations surprenantes pourraient être dues au fait qu'il y a peu d'individus, mais aussi à une faiblesse dans l'estimation de la statistique de test. En effet, cette statistique est estimée grâce à des simulations basées sur un modèle estimé sur les deux échantillons comparés. Plus les échantillons sont différents, plus il y a de variabilité dans les simulations et plus la distribution de la statistique de test est « étendue ». Frascolla et al. (2019) ont montré récemment que l'estimation de la statistique de test peut être améliorée en la réalisant par permutation plutôt que par simulation.

Nature des différences

Le test de différence permet de savoir si deux modèles sont significativement différents mais, si c'est le cas, ne donne aucune information sur la nature de ces différences. Une solution serait de réaliser un test d'adéquation entre les distributions de chaque descripteur pour les deux échantillons comparés d'une part, et d'étudier si des différences significatives existent entre les deux matrices de transition d'autre part. Toutefois, comparer des probabilités est risqué, notamment lorsqu'il y a peu d'observations. Imaginons qu'il n'y ait qu'une transition partant d'un descripteur pour un échantillon et aucune pour l'autre, alors la différence entre les deux matrices de transition montrerait une différence égale à 100%. Il serait alors sans doute préférable de comparer des matrices contenant les effectifs plutôt que les probabilités.

La compréhension des causes et des effets des différences observées peut bien sûr ne pas uniquement reposer sur les données DTS et nécessite d'autres informations. Par exemple, les différences observées pour la perception des Goudas au Royaume-Uni par rapport aux autres pays peuvent amener à s'interroger sur les habitudes de consommation et leur impact sur la perception. Le test de différence entre échantillons DTS dans son utilisation peut ainsi permettre de faire émerger de nouvelles questions de recherche.

Répétitions et multi-bouchées

Dans le cas où l'étude comprend plusieurs répétitions ou plusieurs bouchées successives pour un même produit, le test de différence peut être utilisé pour les comparer. Dans le cas où il n'y aurait pas de différence, alors ces répétitions ou bouchées peuvent être considérées comme des observations indépendantes et ainsi augmenter utilement la taille de l'échantillon.

4.5 Segmentation

Sélection du nombre de segments

Nous avons vu dans les exemples de segmentation présentés que le nombre de segments sélectionnés varie selon le critère d'information utilisé et ne semble pas correspondre au nombre de segments pour lesquels l'interprétation sensorielle est la plus pertinente. En attendant de trouver une approche statistique adaptée, nous proposons donc d'utiliser une méthode pas à pas jusqu'à obtenir 2 segments qui semblent similaires d'un point de vue sensoriel ou qu'un segment soit constitué de trop peu de panélistes.

Pénalisation de la vraisemblance

La pénalisation de la vraisemblance introduite dans la partie 2.5.2 améliore la qualité de l'estimation comme le confirme les simulations présentées dans l'article de l'annexe 2 (p207). Néanmoins, le choix de la pénalité peut sans doute être amélioré et nécessite de nouvelles investigations. La convergence asymptotique de l'estimateur mérite également d'être étudiée.

Segmentation par produit basée sur la perception temporelle

Si des méthodes existent pour segmenter un panel par perception (Dahl & Naes, 2004; Llobell, Cariou, Vigneau, Labenne, & Qannari, 2019), à notre connaissance, notre approche de segmentation est la première à segmenter un panel en se basant sur des différences de perception temporelle. Plus largement, la segmentation en analyse sensorielle se fait généralement sur l'ensemble des produits là où nous proposons une segmentation par produit. Nous proposons de segmenter sur la perception puis d'observer si les différences de perception, quand elles existent, ont un effet sur l'appréciation d'un produit. Notre démarche est donc l'inverse de celle de la cartographie des préférences qui segmente sur les notes hédoniques avant d'éventuellement observer s'il y a des différences de perception entre les segments obtenus. Nous pensons que notre nouvelle approche est pertinente d'un point de vue sensoriel et qu'elle devrait s'avérer complémentaire de celle de la cartographie des préférences.

Segmentation multi-produits

Nous avons présenté dans le second chapitre une méthode pour réaliser la segmentation sur tous les produits simultanément qui permet de regrouper les panélistes qui sont en accord en termes de perception pour les différents produits composant une étude. Nous n'avons pas présenté d'application de cette méthode qui nécessite encore d'être développée. Pour cette méthode, le choix du nombre de segments est encore plus compliqué que dans le cas simple de la segmentation puisqu'ici l'intérêt sensoriel du nombre de segment peut varier d'un produit à l'autre et est donc difficilement lisible. L'hypothèse supposée par cette méthode, qui est l'existence de groupes de panélistes ayant une perception similaire pour l'ensemble des produits, reste à valider. Cette méthode devra donc être appliquée sur de nombreux jeux de données pour établir son intérêt d'un point de vue sensoriel.

Une autre façon de réaliser la segmentation est d'assembler tous les jeux de données et de considérer qu'ils ne forment qu'un échantillon qui sera modélisé par un seul et unique modèle. Cette approche peut être utilisée pour observer si des groupes d'individus se distinguent par leur comportement en DTS, par exemple par l'utilisation d'une partie seulement de la liste de descripteurs. Cette méthode pourrait également permettre d'étudier l'existence de liens entre descripteurs indépendamment des produits.

4.6 Divers

Application à TCATA

Les travaux de cette thèse ont montré ce que l'approche stochastique peut apporter à l'analyse des données DTS. L'une des suites attendue est certainement d'étendre ces travaux au Temporal Check-All-That-Apply (TCATA), une autre méthode sensorielle temporelle introduite plus récemment (Castura et al., 2016). Les deux méthodes se distinguent par le fait qu'un seul descripteur peut être sélectionné à chaque instant pour la DTS, alors que plusieurs descripteurs peuvent l'être pour TCATA. Les modèles présentés dans cette thèse, se basant sur les changements de descripteurs, ne peuvent donc pas s'appliquer directement à TCATA. Nous avons développé une méthode applicable à TCATA, cependant elle n'est pas suffisamment avancée pour être présentée ici.

Application à d'autres domaines

Bien que cette approche ait été développée spécifiquement pour l'analyse sensorielle, elle peut être réutilisée dans de nombreux domaines. Des applications peuvent par exemple être envisagées pour des problèmes de fiabilité ou en informatique pour, par exemple, étudier les comportements de navigation sur internet avec les passages (transitions) d'une page à l'autre.

Temps de calcul

Les méthodes présentées dans cette thèse nécessitent un nombre important d'opérations qui peuvent conduire à des temps de calcul relativement longs. Pour un jeu de données avec 106 panélistes sans répétition avec en moyenne 6 transitions, l'estimation du processus semi-Markovien, la construction du graphe de Markov et la détermination de la position des frontières entre périodes sont quasiment immédiates pour un ordinateur avec un processeur à 2,4GHz et 8Gb de RAM. En revanche, la détermination du nombre de périodes et le test de différence qui nécessitent un nombre important de simulations nécessitent respectivement un peu plus de 5 et 15 minutes pour 1000 simulations. La segmentation en 3 segments dure environ 1 minute.

Toutefois, le code peut encore être amélioré et les durées être ainsi réduites en faisant par exemple appel à la parallélisation. Ces améliorations seront nécessaires pour que les méthodes puissent être utilisées en routine.

Individus aberrants

La segmentation permet d'observer au sein du panel des différences de perception mais aussi des différences de comportement en DTS. Par exemple, la segmentation du panel en 2 groupes ayant dégusté le gouda 4wk30 distingue les panélistes selon le temps entre deux clics. La DTS étant souvent utilisée avec des consommateurs, il y a un risque d'avoir au sein du panel des comportements aberrants (par exemple, très peu de clics ou au contraire énormément). La vraisemblance des séquences observées pour chaque individu pourrait permettre de détecter les individus ayant les comportements les plus éloignés du panel.

Perception moyenne et variabilité de perception au sein d'un panel

A travers les graphes DTS et la segmentation, nous avons pu observer l'existence de différences de perception pour un même produit au sein d'un panel de consommateurs. Bien que cette observation puisse sembler assez évidente, la majorité des études sensorielles visent à mesurer uniquement une réponse moyenne. Nous pensons que cette approche peut conduire à des conclusions erronées puisque la réponse moyenne peut ne correspondre à la perception d'aucun panéliste. Les informations obtenues risquent en effet d'être de piètre qualité voir contradictoire comme c'est le cas pour le Gouda 4wk30 qui est perçu au même moment Tendre et Dense-dur.

Conclusion

Cette thèse visait à proposer une nouvelle approche pour l'analyse des données de Dominance Temporelle des Sensations (DTS). A notre connaissance, il n'existait pas de travaux utilisant les processus stochastiques pour modéliser des données d'analyse sensorielle. Nous avons pourtant pu voir que cette approche est naturelle pour les données DTS puisqu'il s'agit d'observer l'évolution au cours du temps de la variable aléatoire « descripteur dominant », ce qui constitue par définition un processus stochastique. Cette approche a permis pour la première fois une modélisation des données DTS prenant en compte toute leur complexité : les choix de descripteurs, l'ordre des choix et les durées de dominance. Nous avons pu voir que cette modélisation est bien adaptée aux données DTS avec des tests d'adéquation pour l'hypothèse markovienne et le choix de distribution pour modéliser les durées de dominance.

Le graphe DTS permet une meilleure compréhension de la perception à l'échelle individuelle et permet ainsi d'observer la variabilité de perception pour un produit. Il permet de savoir quels descripteurs sont associés et permet également de visualiser les descripteurs importants pour le produit sans toutefois présenter de temporalité. Le découpage en périodes met en évidence une évolution du modèle de la perception durant la dégustation de certains produits. La prise en compte de cette évolution permet une compréhension plus précise de la perception du produit. Le nombre de périodes donne une information sur l'existence d'une temporalité pour le produit étudié. Le test de différence entre échantillons DTS est un outil important qui peut être utilisé pour déterminer statistiquement si des produits sont différents ou si des sous-échantillons du panel sont différents. La méthode que nous avons proposée est une première approche prometteuse mais doit encore être améliorée. Finalement, la segmentation des panélistes par similarité de leur perception a été développée suite aux observations de variabilité au sein des panels objectivée grâce aux graphes DTS. La problématique de la segmentation d'un panel DTS a permis le développement d'un modèle nouveau, évidemment en analyse sensorielle, mais aussi plus largement en statistique. La segmentation a mis en avant des

différences importantes au sein d'un panel avec des perceptions parfois opposées d'un même produit. L'existence d'un lien entre différences de perception et différences d'appréciation hédonique a également pu être montrée.

La modélisation proposée utilise un grand nombre de paramètres ce qui nécessite des échantillons de grande taille, mais comme nous avons pu le voir la perception d'un produit peut être complexe et il peut y avoir une importante variabilité au sein d'un panel. C'est pourquoi nous considérons qu'étudier uniquement la perception moyenne du panel est en règle générale une erreur et la richesse de la méthode que nous proposons est de prendre en compte la variabilité au sein de la population.

Finalement, bien que cette approche ait été développée spécifiquement pour l'analyse sensorielle, elle présente l'avantage de pouvoir être réutilisée dans d'autres domaines tels que la fiabilité ou l'informatique.

Références bibliographiques

Ahlström, R., Berglund, B., Berglund, U., Engen, T., & Lindvall, T. (1987). A Comparison of Odor Perception in Smokers, Nonsmokers, and Passive Smokers. *American Journal of Otolaryngology, 8*(1), 1-6.

Akaike, H. (1974). New Look at Statistical-Model Identification. *Ieee Transactions on Automatic Control, Ac19*(6), 716-723.

Albert, A., Salvador, A., Schlich, P., & Fiszman, S. (2012). Comparison between temporal dominance of sensations (TDS) and key-attribute sensory profiling for evaluating solid food with contrasting textural layers: Fish sticks. *Food Quality and Preference, 24*(1), 111-118.

Ares, G., Castura, J. C., Antunez, L., Vidal, L., Gimenez, A., Coste, B., et al. (2016). Comparison of two TCATA variants for dynamic sensory characterization of food products. *Food Quality and Preference, 54*, 160-172.

Banfield, J. D., & Raftery, A. E. (1993). Model-Based Gaussian and Non-Gaussian Clustering. *Biometrics, 49*(3), 803-821.

Barbu, V. S., & Limnios, N. (2008). *Semi-Markov chains and hidden semi-Markov models toward applications : their use in reliability and DNA analysis*. New York: Springer Science + Business Media.

Berchtold, A. (1998). *Chaînes de Markov et modèles de transition applications aux sciences sociales*. Paris: Hermès.

Bierlaire, M. (2015). *Optimization: Principles and Algorithms*: EPFL Press.

Blissett, A., Hort, J., & Taylor, A. J. (2006). Influence of chewing and swallowing behavior on volatile release in two confectionery systems. *Journal of Texture Studies, 37*(5), 476-496.

Boubakri, R., Lecuelle, G., Schlich, P., Visalli, M., & Ben Hassine, K. (2017). Rapport de stage. Développement statistique en analyse sensorielle : Application des chaines de Markov à la Dominance Temporelle des Sensations. In.

Cairncross, S. E., & Sjöström, L. B. (1950). Flavour profiles: A new approach to flavour problems. *Food Technology, 4*, 308-311.

Cardot, H., Lecuelle, G., Schlich, P., & Visalli, M. (2019). Estimating finite mixtures of semi-Markov chains: an application to the segmentation of temporal sensory data. *Journal of the Royal Statistical Society Series C-Applied Statistics*.

Castura, J. C., Antunez, L., Gimenez, A., & Ares, G. (2016). Temporal Check-All-That-Apply (TCATA): A novel dynamic method for characterizing products. *Food Quality and Preference, 47*, 79-90.

Caumel, Y. (2015). *Probabilités et processus stochastiques*. Paris: Lavoisier-Hermès.

Celeux, G., & Govaert, G. (1992). A Classification Em Algorithm for Clustering and 2 Stochastic Versions. *Computational Statistics & Data Analysis, 14*(3), 315-332.

Chen, J. H., Li, S. T., & Tan, X. M. (2016). Consistency of the penalized MLE for two-parameter gamma mixture models. *Science China-Mathematics, 59*(12), 2301-2318.

Clark, C. C., & Lawless, H. T. (1994). Limiting Response Alternatives in Time-Intensity Scaling - an Examination of the Halo-Dumping Effect. *Chemical Senses, 19*(6), 583-594.

Coombs, C. H. (1964). ***A theory of data*****. New York ; London: Wiley.**

Dahl, T., & Naes, 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.**

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via Em Algorithm. ***Journal of the Royal Statistical Society Series B-Methodological, 39*****(1), 1-38.**

Dinnella, C., Masi, C., Naes, T., & Monteleone, E. (2013). A new approach in TDS data analysis: A case study on sweetened coffee. ***Food Quality and Preference, 30*****(1), 33-46.**

Dodge, Y. (2007). ***Statistique dictionnaire encyclopédique*****. Paris Berlin Heidelberg [etc.]: Springer.**

Droesbeke, J.-J., Saporta, G., & Thomas-Agnan, C. (2013). ***Modèles à variables latentes et modèles de mélange. Journées d'étude en statistique organisées par la Société française de statistique [éditées par] Jean-Jacques Droesbeke,... Gilbert Saporta,... Christine Thomas-Agnan*****. Paris: Éd. Technip.**

Duizer, L. M., Bloom, K., & Findlay, C. J. (1997). Dual-attribute time-intensity sensory evaluation: A new method for temporal measurement of sensory perceptions. ***Food Quality and Preference, 8*****(4), 261-269.**

Eddelbuettel, D., & Balamuta, J. J. (2018). Extending R with C plus plus : A Brief Introduction to Rcpp. ***American Statistician, 72*****(1), 28-36.**

Fisher, R. A. (1912). On an absolute criterium for fitting frequency curves. ***Messenger of Mathematics, 41*****, 155-160.**

Franczak, B. C., Browne, R. P., McNicholas, P. D., Castura, J. C., & Findlay, C. J. (2015). A Markov Model for Temporal Dominance of Sensations (TDS) data. In, ***11th Pangborn symposium*****. Gothenburg, Sweden.**

Frascolla, C., Lecuelle, G., Cardot, H., Schlich, P., & Visalli, M. (2019). Comparaison de trajectoires qualitatives avec des chaînes semi-markoviennes : une application en analyse sensorielle. In, ***51es Journées de Statistique*****. Vandœuvre-lès-Nancy.**

Frühwirth-Schnatter, S. (2006). ***Finite mixture and Markov switching models*****. New York: Springer.**

Frydman, H. (2005). Estimation in the mixture of Markov chains moving with different speeds. ***Journal of the American Statistical Association, 100*****(471), 1046-1053.**

Galmarini, M. V., Visalli, M., & Schlich, P. (2017). Advances in representation and analysis of mono and multi-intake Temporal Dominance of Sensations data. ***Food Quality and Preference, 56*****, 247-255.**

Guttorp, P. (1995). ***Stochastic modeling of scientific data*****. London Glasgow Weinheim [etc.]: Chapman & Hall.**

Hartigan, J., & Wong, M. (1979). Algorithm as 136: A k-means clustering algorithm. *Applied statistics, 28*, 100-108.

Holway, A. H., & Hurvich, L. M. (1937). Differential gustatory sensitivity to salt. *American Journal of Psychology, 49*, 37-48.

Hort, J., Kemp, S. E., & Hollowood, T. (2017). *Time-dependent measures of perception in sensory evaluation*.

Hubert, L., & Arabie, P. (1985). Comparing Partitions. *Journal of Classification, 2*(2-3), 193-218.

Hutchings, S. C., Foster, K. D., Grigor, J. M. V., Bronlund, J. E., & Morgenstern, M. P. (2014). Temporal dominance of sensations: A comparison between younger and older subjects for the perception of food texture. *Food Quality and Preference, 31*, 106-115.

Jaeger, S. R., Hort, J., Porcherot, C., Ares, G., Pecore, S., & MacFie, H. J. H. (2017). Future directions in sensory and consumer science: Four perspectives and audience voting. *Food Quality and Preference, 56*, 301-309.

Jellinek, G. (1964). Introduction to and critical review of modem methods of sensory analysis (odor, tatse and flavour evaluation) with special emphasis on descriptive analysis (flavour profile method). *Journal of Nutrition and Dietetics, 1*, 219-260.

Jones, L. V., Peryam, D. R., & Thurstone, L. L. (1955). Development of a scale for measuring soldiers' food preferences. *Food Research, 20*, 512-520.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya: The Indian Journal of Statistics, Serie A, 62*, 49-66.

Koster, E. P. (2009). Diversity in the determinants of food choice: A psychological perspective. *Food Quality and Preference, 20*(2), 70-82.

Krut, L. H., Brontestewart, B., & Perrin, M. J. (1961). Taste Perception in Smokers and Non-Smokers. *British Medical Journal, 1*(522), 384-&.

Lawlor, S., & Rabbat, M. G. (2017). Time-Varying Mixtures of Markov Chains: An Application to Road Traffic Modeling. *Ieee Transactions on Signal Processing, 65*(12), 3152-3167.

Lecuelle, G., Visalli, M., Cardot, H., & Schlich, P. (2018). Modeling Temporal Dominance of Sensations with semi-Markov chains. *Food Quality and Preference, 67*, 59-66.

Lee, W. E., & Pangborn, R. M. (1986). Time-Intensity - the Temporal Aspects of Sensory Perception. *Food Technology, 40*(11), 71-&.

Lenfant, F., Loret, C., Pineau, N., Hartmann, C., & Martin, N. (2009). Perception of oral food breakdown. The concept of sensory trajectory. *Appetite, 52*(3), 659-667.

Lepage, M., Neville, T., Rytz, A., Schlich, P., Martin, N., & Pineau, N. (2014). Panel performance for Temporal Dominance of Sensations. *Food Quality and Preference, 38*, 24-29.

Levy, P. (1954). Processus semi-Markoviens. In, *Proceedings of International Congress of Mathematics*. Amsterdam, Netherlands.

Llobell, F., Cariou, V., Vigneau, E., Labenne, A., & Qannari, E. (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.**

McLachlan, G. J., & Krishnan, T. (2008). ***The EM algorithm and extensions*****. Chichester: Wiley.**

McLachlan, G. J., & Peel, D. (2000). ***Finite Mixture Models*****. New York.**

McNicholas, P. D. (2016). Model-Based Clustering. ***Journal of Classification, 33*****(3), 331-373.**

Meilgaard, M., Civille, G. V., & Carr, B. T. (1999). ***Sensory evaluation techniques*****. Boca Raton, Fla.: CRC Press.**

Meiselman, H. L. (2013). The future in sensory/consumer research: evolving to a better science. ***Food Quality and Preference, 27*****(2), 208-214.**

Melnykov, V., & Maitra, R. (2010). Finite mixture models and model-based clustering ***Statistics Surveys, 4*****, 80-116.**

Meyners, M. (2011). Panel and panelist agreement for product comparisons in studies of Temporal Dominance of Sensations. ***Food Quality and Preference, 22*****(4), 365-370.**

Meyners, M., & Pineau, N. (2010). Statistical inference for temporal dominance of sensations data using randomization tests. ***Food Quality and Preference, 21*****(7), 805-814.**

Monterymard, C., Visalli, M., & Schlich, P. (2010). The TDS-band plot: A new graphical tool for Temporal Dominance of Sensations data. In, ***2nd conference of the society of sensory professionals*****. Napa, CA, USA, 27-29th October.**

Muñoz, A. M., & Civille, G. V. (1992). The Spectrum descriptive analysis method. In, ***ASTM Manual Series MNL 13, Manual on Descriptive Analysis Testing*****. Hootman, R.C., Ed. Am. Soc. Testing and Materials, West Conshohocken, PA.**

Muñoz, A. M., & Civille, G. V. (1998). Universal, product and attribute specific scaling and the development of common lexicons in descriptive analysis. ***Journal of Sensory Studies, 13*****(1), 57-75.**

Nachtsheim, R., & Schlich, E. (2013). The influence of 6-n-propylthiouracil bitterness, fungiform papilla count and saliva flow on the perception of pressure and fat. ***Food Quality and Preference, 29*****(2), 137-145.**

Neilson, A. J. (1957). Time-intensity studies. ***Drug & Cosmetic Industry, 80*****, 452-453.**

Norris, J. R. (1997). ***Markov chains*****. Cambridge: Cambridge University Press.**

Pamminger, C., & Fruhwirth-Schnatter, S. (2010). Model-based Clustering of Categorical Time Series. ***Bayesian Analysis, 5*****(2), 345-368.**

Pardoux, É. (2007). ***Processus de Markov et applications algorithmes, réseaux, génome et finance cours et exercices corrigés*****. Paris: Dunod.**

Pawitan, Y. (2013). ***In all likelihood : statistical modelling and inference using Likelihood*****. Oxford: Oxford University Press.**

Peltier, C., Visalli, M., & Schlich, P. (2015). Canonical Variate Analysis of Sensory Profiling Data. ***Journal of Sensory Studies, 30*****(4), 316-328.**

Peryam, D. R. (1954). ***Food acceptance testing methodology : a symposium sponsored by the Quartermaster Food and Container Institute for the Armed Forces, Quartermaster Research and Development Command, U.S. Army Quartermaster Corps [at the] Palmer House, Chicago, 8-9 October 1953*****. Washington, D.C.: National Research Council.**

Pineau, N., Cordelle, S., & Schlich, P. (2003). Temporal dominance of sensations : A new technique to record several sensory attributes simultaneously over time. In, ***5th Pangborn symposium*****. Boston, USA.**

Pineau, N., de Bouille, A. G., Lepage, M., Lenfant, F., Schlich, P., Martin, N., et al. (2012). Temporal Dominance of Sensations: What is a good attribute list? ***Food Quality and Preference, 26*****(2), 159-165.**

Pineau, N., Schlich, P., Cordelle, S., Mathonniere, C., Issanchou, S., Imbert, A., et al. (2009). Temporal Dominance of Sensations: Construction of the TDS curves and comparison with time-intensity. ***Food Quality and Preference, 20*****(6), 450-455.**

Pingel, J., Ostwald, J., Pau, H. W., Hummel, T., & Just, T. (2010). Normative data for a solution-based taste test. ***European Archives of Oto-Rhino-Laryngology, 267*****(12), 1911-1917.**

Poisson, S.-D. (1837). ***Recherches sur la probabilité des jugements en matière criminelle et en matière civile précédées des Règles générales du calcul des probabilités*****. Paris: Bachelier.**

Prutkin, J., Duffy, V. B., Etter, L., Fast, K., Gardner, E., Lucchina, L. A., et al. (2000). Genetic variation and inferences about perceived taste intensity in mice and men. ***Physiology & Behavior, 69*****(1-2), 161-173.**

Pyke, R. (1961). Markov Renewal Processes - Definitions and Preliminary Properties. ***Annals of Mathematical Statistics, 32*****, 1231-&.**

R Development Core Team. (2019). R: A language and environment for statistical computing. In. Vienna, Austria: R Foundation for Statistical Computing.

Rozin, P., Ebert, L., & Schull, J. (1982). Some Like It Hot - a Temporal Analysis of Hedonic Responses to Chili Pepper. ***Appetite, 3*****(1), 13-22.**

Saporta, G. (2011). ***Probabilités, analyse des données et statistique*****. Paris: Éd. Technip.**

Schiffman, S. S., & Graham, B. G. (2000). Taste and smell perception affect appetite and immunity in the elderly. ***European Journal of Clinical Nutrition, 54*****, S54-S63.**

Schlich, P. (2017). Temporal Dominance of Sensations (TDS): a new deal for temporal sensory analysis. ***Current Opinion in Food Science, 15*****, 38-42.**

Schoumacker, R., Martin, C., Thomas-Danguin, T., Guichard, E., Le Quere, J. L., & Laboure, H. (2017). Fat perception in cottage cheese: The contribution of aroma and tasting temperature. ***Food Quality and Preference, 56*****, 241-246.**

Schwarz, G. (1978). Estimating Dimension of a Model. ***Annals of Statistics, 6*****(2), 461-464.**

Sjöström, L. B. (1954). The descriptive analysis of flavour. In D. R. Peryam, F. J. Pilgrim & M. S. Peterson, ***Food Acceptance Testing Methodology*****. Chicago, IL: U.S. QUARTERMASTER FOOD AND CONTAINER INSTITUTE.**

Smith, W. L. (1955). Regenerative stochastic processes. ***Proceedings of the Royal Society, Series A 232*****, 6-31.**

Song, Y., Keromytis, A. D., & Stolfo, S. J. (2009). Spectrogram: A Mixture-of-Markov-Chains Model for Anomaly Detection in Web Traffic. ***Proceedings of the Network and Distributed System Security Symposium, NDSS*****.**

Stone, H., Sidel, J., Oliver, S., Woolsey, A., & Singleton, R. C. (1974). Sensory Evaluation by Quantitative Descriptive Analysis. ***Food Technology, 28*****(11), 24-+.**

Taylor, D. E., & Pangborn, R. M. (1990). Temporal aspects of hedonic reponses. ***Journal of Sensory Studies, 4*****, 241-247.**

Thomas, A., Chambault, M., Dreyfuss, L., Gilbert, C. C., Hegyi, A., Henneberg, S., et al. (2017). Measuring temporal liking simultaneously to Temporal Dominance of Sensations in several intakes. An application to Gouda cheeses in 6 Europeans countries. ***Food Research International, 99*****, 426-434.**

Thomas, A., van der Stelt, A. J., Prokop, J., Lawlor, J. B., & Schlich, P. (2016). Alternating temporal dominance of sensations and liking scales during the intake of a full portion of an oral nutritional supplement. ***Food Quality and Preference, 53*****, 159-167.**

Thomas, A., Visalli, M., Cordelle, S., & Schlich, P. (2015). Temporal Drivers of Liking. ***Food Quality and Preference, 40*****, 365-375.**

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). ***Statistical analysis of finite mixture distributions*****.**

Van der Vaart, A. W. (1998). Asymptotic Statistics. ***Cambridge University Press*****.**

Veldhuizen, M. G., Wuister, M. J. P., & Kroeze, J. H. A. (2006). Temporal aspects of hedonic and intensity responses. ***Food Quality and Preference, 17*****(6), 489-496.**

Vergne, N. (2008). Drifting Markov models with polynomial drift and applications to DNA sequences. ***Statistical Applications in Genetics and Molecular Biology, 7*****(1).**

Ye, Z. S., & Chen, N. (2017). Closed-Form Estimators for the Gamma Distribution Derived From Likelihood Equations. ***American Statistician, 71*****(2), 177-181.**

Zimoch, J., & Gullett, E. A. (1997). Temporal aspects of perception of juiciness and tenderness of beef. ***Food Quality and Preference, 8*****(3), 203-211.**

Annexes

Annexe 1

Article “Modeling Temporal Dominance of Sensations with semi-Markov chains”

Food Quality and Preference 67 (2018) 59–66

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Modeling Temporal Dominance of Sensations with semi-Markov chains

G. Lecuelle[a,*], M. Visalli[a], H. Cardot[b], P. Schlich[a]

[a] Centre des Sciences du Goût et de l'Alimentation, AgroSup Dijon, CNRS, INRA, Univ. Bourgogne Franche-Comté, F-21000 Dijon, France
[b] Institut de Mathématiques de Bourgogne, CNRS, Univ. Bourgogne Franche-Comté, Dijon, France

ARTICLE INFO

Keywords:
Temporal Dominance of Sensations
Stochastic modeling
Semi-Markov processes
Maximum likelihood

ABSTRACT

Temporal Dominance of Sensations (TDS) data are usually represented by TDS curves of dominance rates and analyzed by linear models of dominance durations. Such approaches do not properly take into account the fact that the selection of a new dominant attribute likely depends on the current dominant attribute.

Thus, modeling TDS data with a stochastic process seems natural, as recently proposed by Franczak et al. (2015) who used discrete time Markov chains. This approach gives the probabilities of transition from one dominant attribute to another. However Markov chains present some limitations when applied to TDS data. As an alternative, this paper considers semi-Markov chains (SMC), a generalization of Markov chains, which allow the duration of the dominant attribute to be distributed arbitrarily. Because probabilities of transition from one attribute to another one can also depend on time, SMC are applied on sequences split into time periods with specific durations, with one model per time period.

Graphs built upon this stochastic pattern can be plotted to represent chronological main transitions between attributes. Contrarily to the TDS curves which summarize a mean panel overview, these graphs can be interpreted as individual's most probable paths and contribute to a better understanding of consumer perception.

1. Introduction

Temporal Dominance of Sensations (TDS) is a temporal sensory method which consists in presenting to panelists a list of attributes from which they are asked to select the most dominant at each moment of tasting (Pineau, Cordelle, & Schlich, 2003). A dominant attribute is generally defined as the most striking perception at a given time (Pineau et al., 2009). TDS provides rather complex data composed of individual selections of sensory attributes against time. As TDS data are influenced by the individuality of each panelist (chewing/swallowing times, comprehension of the task, difference in perception or experience, etc.), sequence of dominances and their duration are subject to randomness.

Most panel leaders rely on graphical tools like TDS curves (Pineau et al., 2009) or TDS bandplots (Galmarini, Visalli, & Schlich, 2017) to analyze TDS data. Each of these tools provides a mean panel overview, with no information about individual variability. To gain statistical insights some authors proposed to analyze TDS data by comparing the dominance durations using ANOVA (Galmarini et al., 2017). To account for sequence of dominances, others split TDS sequences into successive time periods (Lepage et al., 2014), but the choice of the number and the size of these time periods remains arbitrary. Recently, Castura and Li (2016) proposed to study sequences of attributes with TDS monads, dyads, triads and tetrads. To date, no approach taking into account simultaneously attribute selections, order of these selections and dominance durations seems to exist.

A stochastic approach which considers all individual selections over time and takes into account the random nature of TDS data was proposed by Franczak, Browne, McNicholas, Castura, and Findlay (2015) to model TDS data with discrete time Markov chains (DTMC). DTMC have been widely used to model time processes in scientific areas such as physics, chemistry and speech recognition. Transposed to TDS, the stochastic approach models transitions from one attribute to another at each discrete time point under the assumption (called "Markov memoryless property") that the attribute experienced at time point $t + 1$ only depends on the attribute experienced at time point t. This property implies that the probability distribution of dominance durations is a geometric distribution. However, we show that this distribution is unrealistic because it does not provide a good fit for TDS data.

The aim of the present work is to model TDS data using a more suitable stochastic approach. First, some DTMC theory will be presented, and limitations when applying such a model to TDS data will be shown. Then, semi-Markov chains (SMC) will be introduced on sequences split into optimal time periods, and the advantages of such an approach will be explained. Finally, the proposed model will be illustrated by TDS graphs, an "upgraded" graphical representations of

* Corresponding author.
E-mail address: guillaume.lecuelle@inra.fr (G. Lecuelle).

http://dx.doi.org/10.1016/j.foodqual.2017.06.003
Received 18 November 2016; Received in revised form 1 June 2017; Accepted 6 June 2017
Available online 09 June 2017
0950-3293/ © 2017 Elsevier Ltd. All rights reserved.

Markov graphs adapted to TDS data.

2. Material and methods

A stochastic process is a probabilistic model representing the evolution over time of a variable whose change is subject to random variation. TDS data can be modeled by a stochastic process, where the variable of interest, the dominant attribute, is modeled over time. To achieve this, a simple and widely used model is Markov chains (see Norris, 1997).

2.1. Markov chains

A Markov chain is a random process that undergoes transitions from one state to another on a state space. For TDS data the finite state space is composed of the TDS attributes. The fundamental hypothesis is that the probability to be in state i at time $t + 1$ only depends on the current state at time t. This hypothesis of "memorylessness" is called the Markov property. Markov chains can be used on discrete or continuous time data: discrete time Markov chains model state of the system at each discrete time while continuous time Markov chains model transitions between states and times of occurrences of these transitions using exponential distributions.

For a single panelist evaluating a single product, TDS data correspond to a sequence of attributes selected over time with errors assumed random. Therefore, TDS data can be modeled with a Markov chain where the transition probabilities between attributes have to be estimated. Time is continuous but measurements are discrete and it is convenient to model discrete time data with a finite number of possible durations. Markov chains will be applied on a fine grid of discrete times although the true phenomenon is a continuous process. We denote by n the total number of attributes. Probabilities of each attribute to change to another one are aggregated in the transition matrix $P(n \times n)$ (Fig. 1A) where $P_{s_1 s_2}$ is the probability to move from state s_1 to state s_2. The Markov graph (Fig. 1B) provides the same information but in a more readable way. A Markov graph is a directed graph having vertices representing attributes and arrows representing transitions, labeled by the probabilities to go from one sensory attribute to another one.

For the model given in Franczak et al. (2015), dominance durations are estimated by the number of instants of time the chain stays in the same state. As a consequence, the time spent in state s is distributed according to a geometric distribution with parameter $p = 1-P_{ss}$, where P_{ss} is the probability to stay in state s, corresponding to the number of "trials" needed to move from state s to another state. Based on the transition matrix shown in Fig. 1A), if a panelist selects the attribute Sweet, then at each time there will be a probability equal to 0.4 for this panelist to change of dominant attribute and a probability equal to 0.6 to keep Sweet as dominant. At each time, staying (or not staying) in state Sweet is modeled by the Bernoulli law with a probability of success (change of dominant attribute) equal to 0.4. The number of times spent in the state Sweet is then the number of independent Bernoulli trials needed to get one success and thus is geometrically distributed with a parameter equal to 0.4.

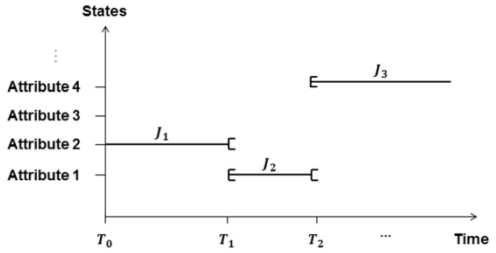


Fig. 2. Representation of an attribute sequence. T_k represents the time of the $k + 1$th change and J_k the k^{th} dominant attribute.

However, this distribution assumption is unrealistic with TDS data. Indeed, a minimum duration is required before choosing another dominant attribute: the time needed to make another selection of dominant attribute and to click on the corresponding button. This observation has been verified based on the investigation of 10 datasets with different product types (chocolates, cheeses, wines) and descriptors. Further examples are illustrated using the dominance duration of Cream and Cooked herbs attributes in a fresh cheese dataset (Thomas, Visalli, Cordelle, & Schlich, 2015) and Crunchy attribute in a chocolate dataset (Fig. 2). It means that the probability of changing from one dominant attribute to another is not the same at every time point; therefore, dominance durations cannot be geometrically distributed.

2.2. Semi-Markov chains

As an alternative to Markov chains, semi-Markov processes were first studied in the 1950's by Lévy (1954) and Smith (1955) independently. More recently, a detailed theoretical analysis of discrete time semi-Markov chains was given in Barbu and Limnios (2008). Semi-Markov chains (SMC) are a generalization of Markov chains which relax the restriction on the distribution of the dominance durations. SMC separately model the transitions and the durations to better fit the data, unlike Markov chains which model both in one single model. Thus, a

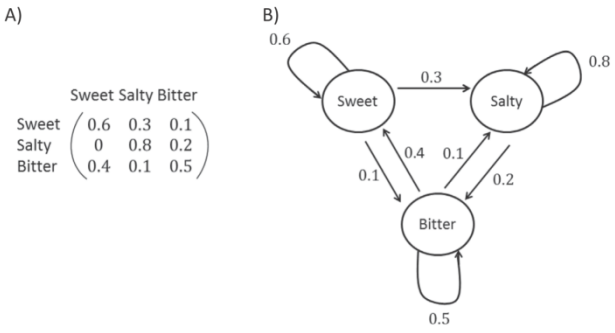


Fig. 1. A) Transition matrix with a finite state space (Sweet, Salty, Bitter). B) Markov graph corresponding to the transition matrix shown in A. In this example, at each time the probability to stay in sweet is equal to 0.6 and the probability to go from sweet to salty is equal to 0.3.

SMC is composed of two models, one for transitions between attributes and a second for dominance durations. The Markovian hypothesis is relaxed: the state of the chain at the next time point now depends on the state at present time, like in Markov chains, but also on the duration since the last change of dominant attribute. The state space is still composed of the attributes.

In this paper, we consider a negative binomial law to estimate dominance durations because of its flexibility in fitting empirical data and particularly overdispersed data (Lloyd-Smith, 2007). Estimations are made using the maximum likelihood principle.

A Markov chain J is used to model the change from one dominant attribute to another (Fig. 2). We denote by $(J_k)_{k=1,2,...}$ the sequence of states taken by the Markov chain, with J_k the state of the chain after k transitions. The transition probabilities are estimated using maximum likelihood. The estimated probability to move from state s_1 to state s_2 is the empirical number of transitions from state s_1 to state s_2 divided by the total number of transitions from state s_1.

The renewal process T models the time of transitions and $(T_k)_{k=0,1,...}$ is the time of the $(k+1)$th transition.

The dominance duration $(X_k)_{k=1,2,...}$ is the duration spent in the kth selected attribute: $X_k = T_k - T_{k-1}$.

2.3. Likelihood computation

Likelihood function (Pawitan, 2013) is a function of the parameters of the model given observed data. Maximum likelihood is computed using estimates of the transition matrix and dominance duration distributions by multiplying probability to observe each event (transitions and dominance durations). Assuming independence between sequences, meaning that selections in each sequence have no influence on selections in other sequences, the likelihood of a dataset is obtained by multiplying the likelihood of each sequence observed in this dataset. A realization of a SMC, corresponding to one sequence during M times, is composed of the successive attribute selections and their dominance duration and is defined as follows:

$$\mathscr{H}(M) = (J_1, X_1, ..., J_{N(M)-1}, X_{N(M)-1}, J_{N(M)}, u_M),$$

where $N(M)$ is the number of transitions within the considered time interval and $u_M := M - S_{N(M)}$ is the censored dominance duration of the last dominant attribute. Indeed, according to the model, duration of the last dominant attribute is censored because the chain is stopped. This makes sense only if the timer is stopped, thus the assessor could not complete the evaluation. But usually the evaluation stops when the assessor clicks STOP meaning that there is nothing more to observe. In this case, u_M is replaced by $X_{N(M)}$. The likelihood of $\mathscr{H}(M)$ is then calculated by multiplying the transition probabilities ($P_{s_1 s_2}$) from state s_1 to state s_2, and the probabilities ($f_s(d)$) of state s to remain dominant during d times as follows:

$$L(\mathscr{H}(M)) = \begin{cases} \alpha_{J_1} * \prod_{k=2}^{N(M)} P_{J_{k-1}J_k} f_{J_{k-1}}(X_{k-1}) * \overline{H}_{J_{N(M)}}(u_M), \\ \quad \textit{if the timer stopped}, \\ \alpha_{J_1} * \prod_{k=2}^{N(M)} P_{J_{k-1}J_k} f_{J_{k-1}}(X_{k-1}) * f_{J_{N(M)}}(X_{N(M)}), \\ \quad \textit{if the assessor clicks STOP}, \end{cases}$$

where α_{J_1} is the initial probability to be in state J_1 and $\overline{H}_{J_{N(M)}}(\cdot)$ is the survival function of the last state of the sequence, $J_{N(M)}$, defined as follows:

$$\overline{H}_{J_{N(M)}}(g) = \mathbb{P}(X_{N(M)} > g | J_{N(M)} = s) = 1 - \sum_{k=1}^{g} f_s(k), \quad g = 0,1,...$$

$\overline{H}_{J_{N(M)}}(g)$ is the probability to stay more than g times in the state s.

2.4. SMC on split sequences

A SMC model is expected to be better than classical Markov chains, but in some cases it will not fit TDS data well. Indeed, the heterogeneity of transition probabilities over time is an important source of error on the model: the probability to move from one attribute to another can be very different during each phase of tasting. Several explanations can be put forward. TDS is a multi-modal protocol, and due to product in-mouth destructuring, aroma descriptors have a greater chance to appear after texture ones. In the same way, trigeminal descriptors will be most likely dominant at the end of the sequence. Furthermore, the probability to move from one attribute to another can depend on time even if all attributes being evaluated belong to the same sensory modality. For some products such as chocolate, texture sensations appear in a logical order, hard or crunchy and then soft or melty: probabilities to move to these attributes clearly depend on time. So, probabilities of transition from one attribute to another one can change over time.

Considering the number of observations for a TDS dataset, estimating one SMC model per discrete time point is unrealistic. In the TDS literature (Lepage et al., 2014; Pineau, Neville, & Lepage, 2011), splitting sequences into time periods has already been done, providing meaningful results. The cutting was made arbitrarily with equal durations and with a number of time periods usually fixed to three, considering there are “beginning”, “middle” and “end” phases during tasting. Likelihood offers a way to determine the number of time periods and their frontiers taking into account product specificities according to the SMC model.

2.4.1. Automatic determination of time periods with more homogeneous transition probabilities

Because of the nature of the data, the variations of the likelihood function are not monotonous when moving the location of frontiers. Thus, searching for the maximum likelihood cannot be done using standard optimization methods. For this reason, all possible cutting points have to be considered, which is computationally expensive. For each cutting, data are modeled using one SMC per time period. Then, the likelihood of data is calculated by multiplying likelihoods of all time periods according to associated SMC. Finally, the cutting points that maximize the global likelihood are selected.

2.4.2. Test for homogeneous transition probabilities

In order to check if transition probabilities are homogeneous over time for one product, a statistical test is used. Sequences are split into two time periods with an optimized choice of the frontier as described previously and a transition matrix is estimated for each time period without transitions from START and to STOP (because transition from START is always realized during the first time period and transition to STOP always during the last). The null hypothesis assumes the equality of these two transition matrices, meaning that transition probabilities are homogeneous over the two time periods.

To estimate the distribution of the test statistic we use a Monte Carlo approach. A large number of datasets for the studied product are simulated using the model without splitting (sequentially, transitions between attributes and dominance durations are randomly selected according to the model). For each simulated dataset, sequences are split into two time periods with the same location for the frontier. By construction the null hypothesis is true for these datasets: transition probabilities are homogeneous over time periods because they have been generated by the same transition matrix. For each simulated dataset, one transition matrix is estimated by time period and the distance between these two matrices is calculated using the Manhattan distance. The distribution of these distances is an estimation of the distribution of the test statistic assuming the null hypothesis of no difference in transition probabilities between time periods is true.

Finally, the distance between the two transition matrices of actual

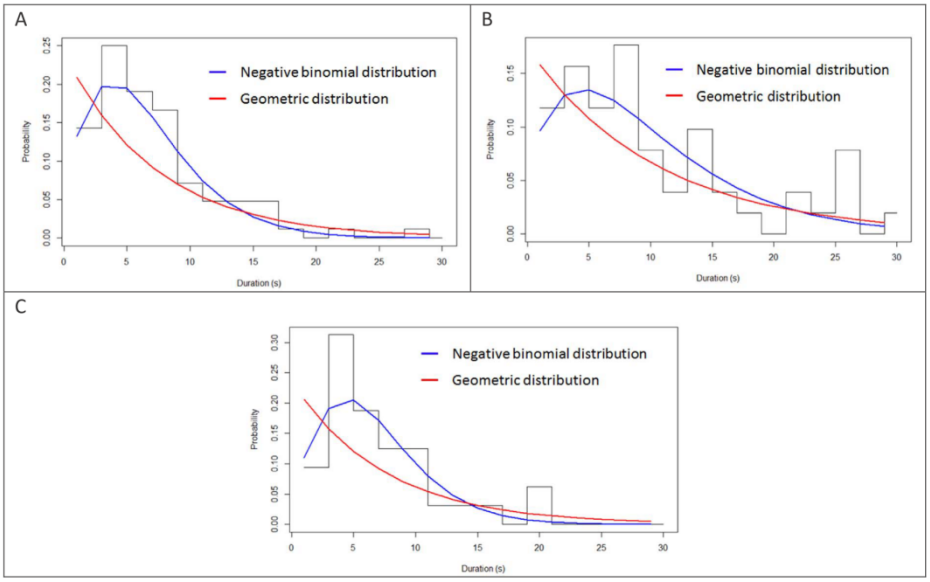


Fig. 3. Histogram of dominance durations for the attribute Cream of fresh cheese P2 (A), the attribute Cooked herbs of fresh cheese P5 (B) and the attribute Crunchy of chocolate with 70% cocoa (C). The red curve is the estimation of these durations by a geometric distribution, corresponding to Markov chains estimation, and the blue curve is the estimation by a negative binomial distribution.

data is calculated. The p-value is the percentage of simulated datasets having a distance larger or equal to this observed distance. If the obtained p-value is smaller than α, the two matrices are considered as being significantly different and then splitting into two time periods is justified. In the other case the null hypothesis is not rejected and no splitting is necessary.

When considering more than two splitting time periods, this test can be performed iteratively. For two time periods, if transition matrices are significantly different, the test is performed with three time periods with optimized choice of frontiers by testing adjacent time periods. The number of time periods grows until at least two adjacent time periods have not been found significantly different. For example, if sequences are split into 4 time periods but transition matrices of time periods 3 and 4 are not significantly different, 3 time periods are retained. It is to be noticed that allowing the number of time periods to vary not only improves the model fitting but also gives interesting information about product perception to the panel leader.

2.5. TDS graphs

To visualize the estimated model, we use the classical Markov graph, which is a graphical representation of the transition matrix providing a representation of the most probable sequences of attributes given by the panelists during the TDS task.

For the sake of clarity, the choice has been made to only display in the graph attributes and transition probabilities considered as "sufficiently representative". To keep the balance between clarity and accuracy, the following criteria are proposed: i) attributes have to be quoted by at least one half of the panelists for the current product; ii) transition probabilities have to be greater than $\frac{1}{n+2}$, n being the number of attributes.

If all transitions are displayed, the sum of the probabilities indicated on the arrows coming away from one attribute is equal to 1. However the same does not work in reverse. The sum of the probabilities indicated on the arrows coming to one attribute can exceed 1 because they are conditional on the current dominant attribute.

When sequences are split into time periods, arrows are colored to represent the time period while transitions occurred. The size of attribute circles is proportional to the mean dominance duration of the attribute and the width of arrows is proportional to the value of the corresponding transition probability.

2.6. Example datasets

The proposed models were applied on several datasets using R (R Development Core Team, 2017). Results for two datasets are presented in this paper. First, a dataset with 3 chocolates evaluated by 18 consumers with two replications and on 10 attributes: one chocolate with 70% of cocoa, another with 70% of cocoa expected to be sweeter and softer, and one with 90% of cocoa. Replications have been considered as additional assessors.

Second, a dataset presented in the Data Analysis Workshop at the 2014 Sensometrics meeting (Thomas et al., 2015). This dataset includes TDS data on 6 flavored fresh cheeses tasted by 64 consumers according to 8 attributes. In this dataset, consumers picked a dominant attribute which faded after a few seconds. This trick aimed at inciting consumers to click, but as with regular TDS the previous dominant attribute remained dominant until the next dominant attribute had been selected.

Time is discretized with a pace of one second in order to avoid long computation time without losing too much information.

First step to model TDS data is to determine the number of time periods. Test for heterogeneity is applied to each product of the datasets. Sequences are split into an iterative number of time periods until two transition matrices of two different time periods are similar. At each iteration, one thousand simulations of datasets are made from a single transition matrix to estimate the distribution of the test statistic. Then frontiers locations are determined according to the maximum likelihood principle. Finally, TDS graph are drawn and compared to

Table 1
P-values for chocolates of the statistical tests for transition probabilities heterogeneity over time. If the p-value is lower than 0.05, then the two transition matrices of the two periods are significantly different. Otherwise these two matrices are considered as similar (p-values in bold), the test is stopped and the chosen number of periods is equal to the actual tested number of periods minus one.

	2 periods	3 periods		4 periods		
	1st vs 2nd	1st vs 2nd	2nd vs 3rd	1st vs 2nd	2nd vs 3rd	3rd vs 4th
70%	0.0106	0.0019	0.0020	**0.2844**	0.0641	0.0231
70% sweet	0.0117	**0.1506**	0.0123			
90%	0.0031	0.0267	0.0008	0.0268	0.0150	**0.0874**

usual TDS curves.

3. Results

3.1. Distribution of dominance durations

For both chocolate and fresh cheese datasets, we observe that dominance durations of attributes are not geometrically distributed because the probabilities to select another dominant attribute are not the same over time. There are few short dominance durations (Fig. 3). The negative binomial law offers a better modeling for these variables. Hence, semi-Markov chain modeling is better than Markov chain for these datasets.

3.2. Number of time periods

Table 1 shows that chocolates 70% and 90% are cut in three time periods and that chocolate 70% sweet is cut in 2 time periods.

Table 2 shows that fresh cheeses do not require cutting. Tests for heterogeneity show that number of time periods depends on the product.

3.3. Frontiers of time periods

Graphical representation of splitting is shown on standardized TDS curves for convenience but computations are done on raw data in order to conserve real dominance durations distributions.

For the chocolate 70% (Fig. 4A), frontiers are located at 42% and 72% of the total duration of sequences corresponding on average to 18 and 32 s with a mean dominance duration of 44 s. For the chocolate 70% sweet (Fig. 4C), frontier is located at 19% of the total duration of sequences corresponding on average to 8 s with a mean dominance duration of 44 s. For the chocolate 90% (Fig. 4E), frontiers are located at 36% and 71% of the total duration of sequences corresponding on average to 17 and 35 s with a mean dominance duration of 49 s.

The frontiers of the time periods give information about the variations in time of transition probabilities and thus about the different phases during tasting and the temporal perception of the product.

Table 2
P-values for fresh cheeses of the statistical tests for transition probabilities heterogeneity over time. If the p-value is lower than 0.05, then the two transition matrices of the two periods are significantly different. Otherwise these two matrices are considered as similar (p-values in bold), the test is stopped and the chosen number of periods is equal to the actual tested number of periods minus one.

	2 periods 1st vs 2nd
P1	**0.0831**
P2	**0.1443**
P3	**0.1581**
P4	**0.1162**
P5	**0.0502**
P6	**0.1277**

3.4. TDS graph

For chocolate 70% (Fig. 4B), the probability for a panelist to select Crunchy as the first attribute is equal to 0.81. It is to be noticed that this value exactly matches the dominance rate of Crunchy on the y-axis at time 0 in Fig. 4A because contrarily to other transitions it is not a conditional probability as all sequences begin by START. If Crunchy is selected as dominant during the first time period there is a probability of 0.41 to select Cocoa and a probability of 0.28 to select Sweet. Transitions from Crunchy to Sweet or to Cocoa and from Sweet to Cocoa are significantly observed during the first time period. The transition from Sweet to Melting is significantly observed during the third time period. The TDS graph shows that there is no significant transition during the second time period, meaning that there is no agreement among panelists during this time period.

For chocolate 70% sweet (Fig. 4D), most of the panelists selected Crunchy and then moved to Sweet. During the second time period, panelists selected among Sweet, Cocoa and Melting in a random order. The Cocoa to Fatty transition occurs with probability 0.17, whereas the probabilities for Cocoa to Sweet and for Cocoa to Melting are higher (0.27 and 0.29). But it is interesting that Fatty to Cocoa occurs with probability 0.32, suggesting a link between fatty and cocoa.

In the first time period, the TDS graph of chocolate 90% (Fig. 4F) shows that panelists selected mostly Crunchy and then Bitter or Dry. Then, those who selected Dry mostly selected Bitter. After that, some panelists selected Cocoa and then Bitter again. In the second time period, only the transition from Bitter to Cocoa is significant. Finally, in the last time period, some of the panelists selected the attribute Astringent. The transition from Bitter to Cocoa is significant during the first and the second time periods with two different values of transition probability which illustrates the need to split sequences into time periods that are modeled separately.

The TDS graph of fresh cheese P6 (Fig. 5B) shows that the first selected attribute was Cooked herbs with probability 0.47 and Cream with probability 0.19. After that first selection, panelist mostly selected Cooked herbs, Cream and Garlic but in a random order. Cooked herbs and Garlic are the most selected attributes before the tasting end.

The TDS graph in Fig. 5B provides more information than the TDS curve in Fig. 5A. Examining the TDS curve, the only attribute that seems important is Cooked herbs for the fresh cheese P6 and the other attributes seem to be picked at random. On the contrary, the TDS graph shows that panelist did not keep Cooked herbs during the whole tasting but significantly selected Cream and Garlic too. For chocolate 70%, examining only the curve (Fig. 4A), we think that panelists perceived Crunchy, then Sweet, Cocoa and at the end Melting but with the graph (Fig. 4B), we observe that this sequence is in reality a combination of several perceptions. TDS curves only give a mean panel overview whereas TDS graphs allow the visualization of individual perceptions.

4. Discussion and conclusions

TDS sequences can be modeled with a stochastic process and more

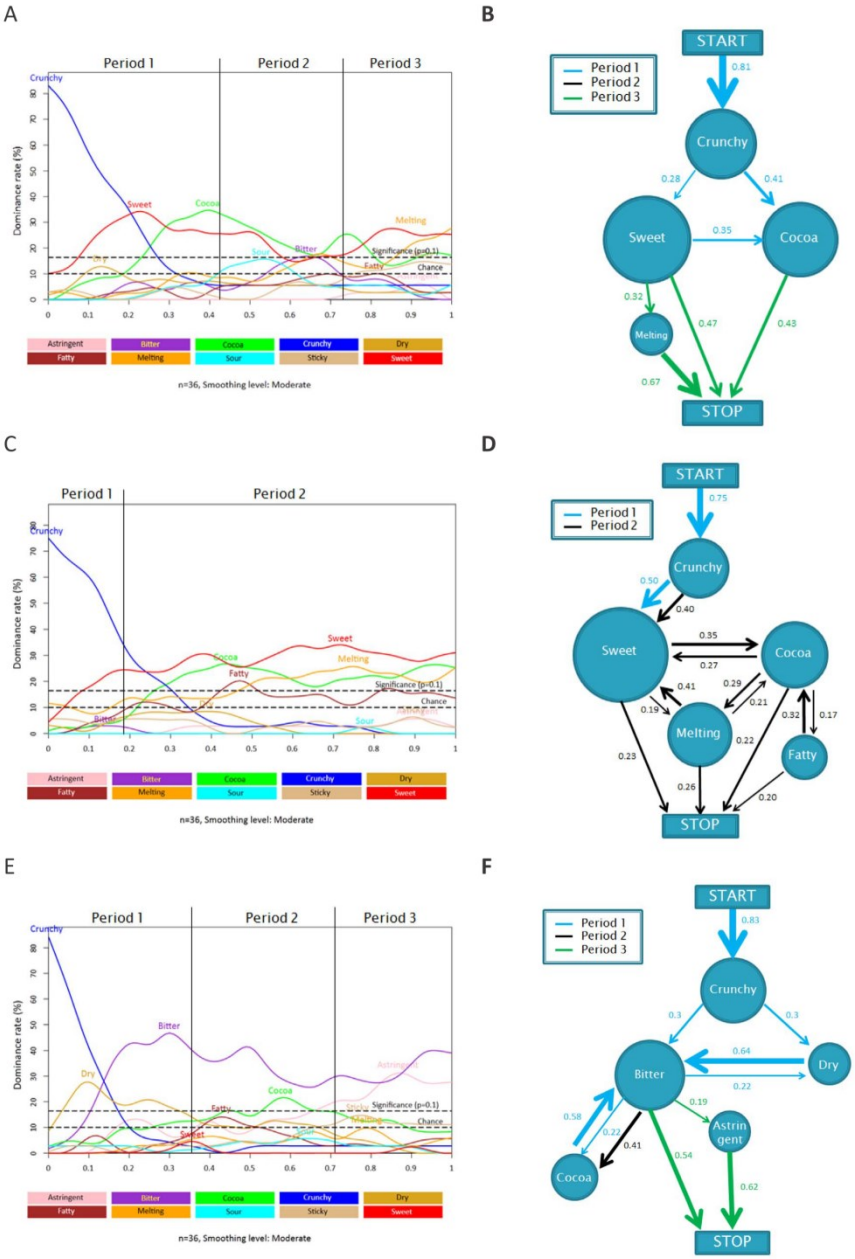


Fig. 4. The TDS curves and the TDS graphs of chocolates with 70% cocoa (A, B), 70% sweet cocoa (C, D) and 90% cocoa (E, F).

precisely with a semi-Markov chain. Castura and Li (2016) showed that it can be of interest to study dominance sequences of several attributes especially in case of concurrent perception. It suggests that probabilities of transition can depend on the present and the recent past. Increasing the order of the chain (the number of previous states used to calculate transition probabilities) can improve the goodness of fit but requires a

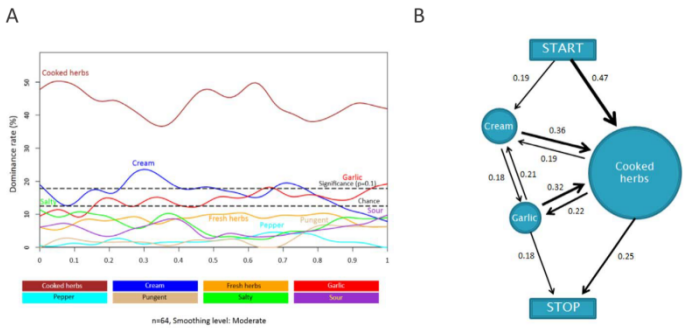


Fig. 5. The TDS curves and the TDS graph of the fresh cheese P6.

huge amount of data. Furthermore, when sequences are split into time periods, the number of transitions during each time period is quite small. Thus, time periods are easy to model and a chain of order one provides sufficient precision.

Panelists with a different number of dominance selections will contribute differently to the model estimation. One solution could be to weight panelist by the number of selected attributes when estimating the model.

As we showed in Section 2, dominance durations are better modeled using a negative binomial distribution than a geometric distribution, and transition probabilities are time-dependent, which justifies separating results into time periods. The number of time periods and the location of the frontiers between time periods provide interesting information about the different relevant temporal phases during the tasting of a product. For the three chocolates, the first time period includes a lot of information and agreement among panelists. The second time period is a phase with panelist disagreement. Observing the TDS curves, few attributes are dominant during the second time period. It can seem counterintuitive that the method for determining the number of time periods and location of frontiers find time periods without agreement but it is justified because that kind of time period requires a different modeling. When there is a third time period (chocolates 70% and 90%), it is a phase with agreement among panelists.

For the purpose of simplicity and because of the small sample size, time has been discretized. The choice of step size influences the goodness of fit of dominance durations estimation and choosing a too large step size can lead to hide some dominant attributes with too small dominance duration. This choice is a compromise between loss of information and computation performances. A better estimation of dominance durations could have been obtained using kernel density estimation but a parametric model is easier to estimate and to handle with a known density function.

In order to compute likelihood, independence between sequences is assumed. This assumption is justified because the experimental design ensures that consecutive products are different and balanced over the experiment. With replications, panelists do not know they are retasting a previously evaluated sample. On the other hand, in case of multi-intake tasting, this assumption cannot be verified because previous bites certainly influence perception so intakes should be modeled separately.

Splitting is done by computing all the time period possibilities but it could be of interest to specify a minimum length for time periods. Really short time periods which do not make sense would be avoided and computation time would be shorter. Nevertheless, this minimum length is to be subjectively defined with a risk of information loss. However, computation is quite simple and can easily be done with a programming language offering high performance such as C++. By combining R and C++, calculation of frontiers for fresh cheese dataset took less than 3 s with 2 and 3 time periods and took 2 min and 30 s with 4 time periods on a computer with a 2.4 GHz CPU and 8 Gb RAM (no parallel tasking).

As SMC is a parametric model, an information criterion like the Akaike Information Criterion could have been used to determine the number of time periods. However the large number of parameters ($p-1$ frontiers locations, pn^2 transition probabilities and $2p$ parameters for the estimation of dominance durations) and the small sample size compared to the number of attributes make the AIC penalty high when the number of time periods increases. As a result the AIC rarely selects more than 1 time period.

Finally, the TDS graph can be used to represent the data in a different but complementary way than TDS curve. Indeed, a TDS curve shows a mean panel overview which can be wrongly understood as the only way to perceive the product. On the contrary, a TDS graph suggests the most probable individual successions of attributes and gives valuable information about individual differences in qualitative perception. TDS graphs do not summarize a product as a one unique sequence of dominances, but emphasize that there can be several different perceptions of a product. In addition, confusion between dominance and intensity can occur when interpreting TDS curves, but not with TDS graphs.

This probabilistic way to explore TDS data offers the opportunity to apply usual likelihood methods. In the future, applications such as product comparison and segmentation of a panel by perception will be developed.

References

Barbu, V. S., & Limnios, N. (2008). *Semi-Markov chains and hidden semi-Markov models toward applications : Their use in reliability and DNA analysis*. New York: Springer Science+Business Media.

Castura, J. C., & Li, M. (2016). Using TDS dyads and other dominance sequences to characterize products and investigate liking changes. *Food Quality and Preference*, *47*, 109–121. http://dx.doi.org/10.1016/j.foodqual.2015.06.019.

Franczak, B. C., Browne, R. P., McNicholas, P. D., Castura, J. C., & Findlay, C. J. (2015). *A Markov Model for Temporal Dominance of Sensations (TDS) data*. Paper presented at the 11th Pangborn symposium, Gothenburg, Sweden.

Galmarini, M. V., Visalli, M., & Schlich, P. (2017). Advances in representation and analysis of mono and multi-intake Temporal Dominance of Sensations data. *Food Quality and Preference*, *56*, 247–255. http://dx.doi.org/10.1016/j.foodqual.2016.01.011.

Lepage, M., Neville, T., Rytz, A., Schlich, P., Martin, N., & Pineau, N. (2014). Panel performance for Temporal Dominance of Sensations. *Food Quality and Preference*, *38*, 24–29. http://dx.doi.org/10.1016/j.foodqual.2014.05.002.

Lévy, P. (1954). *Processus semi-Markoviens*. Paper presented at the Proceedings of International Congress of Mathematics, Amsterdam, Netherlands.

Lloyd-Smith, J. O. (2007). Maximum likelihood estimation of the negative binomial dispersion parameter for highly overdispersed data, with applications to infectious diseases. *PLoS One*, *2*(2), http://dx.doi.org/10.1371/journal.pone.0000180.

Norris, J. R. (1997). *Markov chains*. Cambridge: Cambridge University Press.

Pawitan, Y. (2013). *In all likelihood: Statistical modelling and inference using likelihood*. Oxford: Oxford University Press.

Pineau, N., Cordelle, S., & Schlich, P. (2003). *Temporal dominance of sensations: A new technique to record several sensory attributes simultaneously over time*. Paper presented at the 5th Pangborn symposium, Boston, USA. July 20–24.

Pineau, N., Neville, T., & Lepage, M. (2011). *Panel performance tool for Temporal*

Dominance of Sensations studies. Paper presented at the 9th Pangborn symposium, Toronto, Canada.
Pineau, N., Schlich, P., Cordelle, S., Mathonniere, C., Issanchou, S., Imbert, A., ... Koster, E. (2009). Temporal Dominance of Sensations: Construction of the TDS curves and comparison with time-intensity. *Food Quality and Preference*, *20*(6), 450–455. http://dx.doi.org/10.1016/j.foodqual.2009.04.005.
R Development Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Smith, W. L. (1955). Regenerative stochastic processes. *Proceedings of the Royal Society, Series A*, *232*, 6–31.
Thomas, A., Visalli, M., Cordelle, S., & Schlich, P. (2015). Temporal drivers of liking. *Food Quality and Preference*, *40*, 365–375. http://dx.doi.org/10.1016/j.foodqual.2014.03.003.

Annexe 2

Article “ Estimating finite mixtures of semi-Markov chains: an application to the segmentation of temporal sensory data “

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2019)

Estimating finite mixtures of semi-Markov chains: an application to the segmentation of temporal sensory data

Hervé Cardot, Guillaume Lecuelle, Pascal Schlich and Michel Visalli

Université Bourgogne Franche-Comté, Dijon, France

[Received June 2018. Revised April 2019]

Summary. In food science, it is of great interest to obtain information about the temporal perception of aliments to create new products, to modify existing products or more generally to understand the mechanisms of perception. Temporal dominance of sensations is a technique to measure temporal perception which consists in choosing sequentially attributes describing a food product over tasting. This work introduces new statistical models based on finite mixtures of semi-Markov chains to describe data collected with the temporal dominance of sensations protocol, allowing different temporal perceptions for a same product within a population. The identifiability of the parameters of such mixture models is discussed. Sojourn time distributions are fitted with a gamma probability distribution and a penalty is added to the log-likelihood to ensure convergence of the expectation–maximization algorithm to a non-degenerate solution. Information criteria are employed for determining the number of mixture components. Then, the individual qualitative trajectories are clustered with the help of the maximum *a posteriori* probability approach. A simulation study confirms the good behaviour of the estimation procedure proposed. The methodology is illustrated on an example of consumers' perception of a Gouda cheese and assesses the existence of several behaviours in terms of perception of this product.

Keywords: Bayesian information criterion; Categorical time series; Expectation–maximization algorithm; Gamma distribution; Identifiability; Markov renewal process; Model-based clustering; Penalized likelihood; Temporal dominance of sensations

1. Introduction

The development of food products is usually based on the measurement of product sensory perceptions from panels of consumers. Sensory perception while eating a food product has been acknowledged as a temporal process for 60 years (Neilson, 1957). Measuring temporal sensory perception is a complex task and various approaches have been developed in sensory science (see Hort *et al.* (2017)). Recently a technique called temporal dominance of sensations (TDS) has been introduced by Pineau *et al.* (2009). A review on TDS can be found in Schlich (2017). The panellists must describe the tasted product by choosing which attribute, among a list composed of about 10 items, corresponds to the most striking perception at a given time. This task results in sequences of attributes with choices and time of the choices. When an attribute has been selected as dominant, it is considered as dominant until the panellist selects another dominant attribute. At each time only one attribute can be dominant. An example of such an experiment for chocolate tasting is presented in Fig. 1 with data represented as band plots.

Address for correspondence: Hervé Cardot, Institut de Mathématiques de Bourgogne, Université Bourgogne Franche-Comté, 9 avenue Alain Savary, F-21000 Dijon, France.
E-mail: herve.cardot@u-bourgogne.fr

© 2019 Royal Statistical Society 0035–9254/19/68000

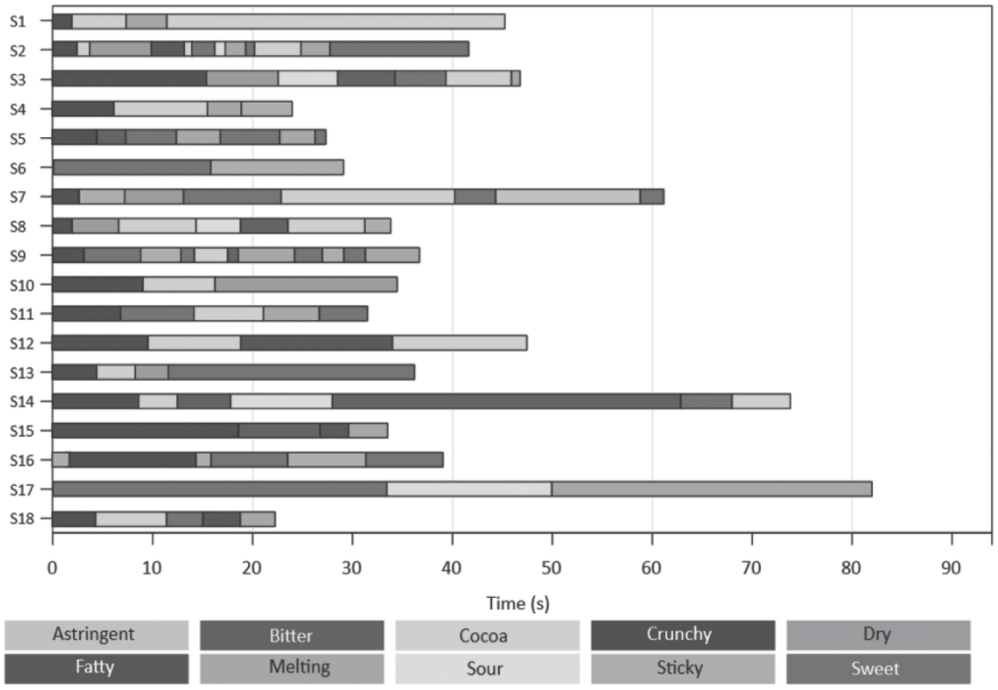


Fig. 1. Tasting of a chocolate with 70% of cocoa by 18 panellists, denoted by S_1–S_{18}, with 10 attributes: the bands represent the succession over time of the dominant attributes selected by each panellist while tasting this chocolate; the figure has been produced by means of the TimeSens© software (www.timesens.com)

Some simple methods are currently used to describe such qualitative temporal data. Most of them rely on the observation of TDS curves, which consist in representing the evolution along time of the proportions of the dominant attributes at a panel level. Even if this statistical approach can be very informative, such a tool provides only a mean panel overview and no information about individual variability. Some quantitative analyses are used as a complement (Galmarini *et al.*, 2017) but these methods consider only dominance durations (the time that is spent as dominant for each attribute). None of these approaches takes into account the whole complexity of TDS data: choices of dominant attribute, order of the choices and dominance durations that are sojourn times in the successive dominant attributes. Recently, Franczak *et al.* (2015) proposed modelling TDS data with Markov chains. The Markov hypothesis, meaning that the probability of the next choice of dominant attribute depends on only the current dominant attribute, seems to be reasonable from a sensory perspective. However, the Markov hypothesis imposes strong restrictions on the sojourn time distribution which should be geometrically distributed when considering a discrete time process, or exponentially distributed when considering a continuous time process (see for example Norris (1998) for a general presentation of Markov chains). Recently Lecuelle *et al.* (2018) noted that the sojourn time distributions were not distributed according to a geometric law. Consequently, it has been proposed to model TDS data with semi-Markov chains and it has been shown that allowing arbitrarily distributed sojourn times achieves a better fit to the data. Note that approaches based on multivariate categorical data are not adapted for TDS data since we observe sequences with a random number of visited states (see Fig. 1). Semi-Markov chain or Markov renewal processes,

which were introduced more than 60 years ago (Lévy, 1956; Smith, 1955), are now widely used in numerous fields of science such as queuing theory, reliability and maintenance, survival analysis, performance evaluation, biology, deoxyribonucleic acid analysis, risk processes, insurance and finance or earthquake modelling (see for example Barbu and Limnios (2008) and references therein).

It has often been suggested by sensory scientists (Jaeger *et al.*, 2017) that consumers form non-homogeneous populations and heterogeneity in consumers' food products perception was established by Prutkin *et al.* (2000). To take into account heterogeneity among individuals and to avoid conclusions on a non-existing 'average consumer', consumer segmentation is a recommended strategy (Köster, 2009; Meiselman, 2013). Introducing mixtures for modelling the different perceptions of a sample of panellists for a same product can be of real interest.

A mixture model (McLachlan and Peel, 2000; Melnykov and Maitra, 2010) is a probabilistic model enabling us to represent the presence of subpopulations within an overall population. Finite mixture models are widely used in numerous fields of science such as biology or economy because they offer probabilistic tools for performing clustering. Mixture models are commonly used with the Gaussian distribution but they can also be used with any parametric model (see the numerous examples in Frühwirth-Schnatter (2006) as well as Banfield and Raftery (1993) or McNicholas (2016)). For temporal data, mixtures of Markov chains have been used in different fields such as finance (Frydman, 2005), computer science (Song *et al.*, 2009), road traffic estimation (Lawlor and Rabbat, 2017) or labour economy (Pamminger and Frühwirth-Schnatter, 2010). In continuous time and continuous response, Delattre *et al.* (2016) introduced mixtures of stochastic differential equations and used a classification rule based on estimated posterior probabilities to cluster growth curves. However, as far as we know, the present work is the first that considers mixtures of semi-Markov processes. The purpose of this paper is to estimate mixtures of semi-Markov chains, in discrete or continuous time, to perform a segmentation of a sample of panellists into groups with similar perceptions. The methodology that is developed in this paper can be useful in many domains for which the aim is to analyse and perform a segmentation of panels of categorical trajectories.

Identifiability is a crucial issue for mixture models (see Titterington *et al.* (1985) and Frühwirth-Schnatter (2006)) and we show under general conditions that, when identifiable parametric models are considered for the distribution of sojourn times, the parameters of the model are identifiable up to label swapping. The estimation of the parameters is performed with the expectation–maximization (EM) algorithm (McLachlan and Krishnan, 2008) in which a penalty may be added to avoid degenerate solutions. In our sensory analysis example, sojourn times are fitted with gamma distributions and, as explained in Chen *et al.* (2016), the likelihood is generally unbounded in the case of mixtures of gamma distributions. We thus consider a penalized likelihood criterion that leads to more stable estimates and enables us to avoid degenerate solutions. As the number of mixture components is generally unknown, an information criterion is employed to select the number of subpopulations that should be considered (see Pamminger and Frühwirth-Schnatter (2010) for a discussion about model selection in the context of mixtures of Markov chains). Then, the observed trajectories can be clustered thanks to the *maximum a posteriori* (MAP) probability classification approach (see Frühwirth-Schnatter (2006)).

The method proposed is illustrated on a data set from the European Sensory Network (Thomas *et al.*, 2017). This data set includes TDS data for four Gouda cheeses tasted by 665 consumers according to 10 attributes. A mixture of semi-Markov chains with gamma sojourn time distributions is adjusted to fit the data.

The paper is organized as follows. Section 2 presents the mixture models and discusses the identifiability issue. Section 3 presents the EM algorithm that was employed for the estimation of the

parameters of the mixture, the proportions and the number of components. Section 4 evaluates the performances of the statistical methods through a simulation study and Section 5 provides an illustration of the proposed method on cheese tasting data. Concluding remarks and discussion are given in Section 6. Proofs, additional details on the simulation study and the estimated parameters for the cheese tasting data are gathered in an on-line supplementary document.

2. Stochastic model and notation

2.1. Markov renewal processes and finite mixtures of Markov renewal processes

Consider a finite state homogeneous Markov chain $(J_p)_{p\geqslant 1}$, taking values in the finite state space $\mathcal{S}=\{1,\ldots,D\}$, with transition matrix $\mathbf{P}$, whose generic elements are $P_{lj}=\Pr(J_{p+1}=j|J_p=l)$, $l,j\in\mathcal{S}$. Consider the random sequence $(X_p)_{p\geqslant 1}$ made by the successive sojourn times in the states visited. For each $p\geqslant 1$, X_p represents the sojourn time in state J_p and takes values in $T=1,2,\ldots$ if time, denoted by t, is discrete and in $T=[0,\infty[$ if time is continuous. For $j\neq l$, we denote by $\Phi_{lj}(t)=\Pr(X_p\leqslant t\mid J_p=l,J_{p+1}=j)$ the cumulative distribution function of the sojourn time given the current and the next states of the random process $(J_p)_{p\geqslant 1}$. We suppose that the random process $(J_p,X_p)_{p\geqslant 1}$ satisfies the Markov property, for all $t\in T$, $l\in\mathcal{S}$ and $j\neq l$,

$$\Pr(J_{p+1}=j,X_p\leqslant t\mid J_p=l,J_{p-1},\ldots,J_1,X_{p-1},\ldots,X_1)=P_{lj}\,\Phi_{lj}(t). \tag{1}$$

The process $(J_p,X_p)_{p\geqslant 1}$ is called a Markov renewal process, whereas the stochastic process giving the state of the system at every time $t\in T$ is called a semi-Markov process (see for example Pyke (1961) or Barbu and Limnios (2008)). For identifiability, it is also supposed that $P_{jj}=0$, for all $j\in\mathcal{S}$, so that, at each jump, the system cannot remain in the same state. To avoid trajectories with an infinite number of visited states, we also suppose that the semi-Markov chain is regular (see Pyke (1961)). This is true for gamma-distributed sojourn times that are considered in the application, and more generally under the very weak condition that the cumulative distribution function is continuous at 0 with $\lim_{t\to 0_+}\Phi_{lj}(t)=0$. Finally, to characterize the law of $(J_p,X_p)_{p\geqslant 1}$ completely we define the vector $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_D)$ of initialization probabilities:

$$\alpha_j=\Pr(J_1=j),\qquad j\in\mathcal{S}. \tag{2}$$

The example that is given in Fig. 2 describes the representation, in terms of semi-Markov trajectory, of the fourth TDS sequence of the data set that was presented in Fig. 1.

The distribution of the semi-Markov process $(J_p,X_p)_{p\geqslant 1}$ is completely characterized by the set of parameters $(\boldsymbol{\alpha},\mathbf{P},\Phi_{lj},l,j\neq l\in\mathcal{S})$ and in what follows its probability law is denoted by $\mathrm{Law}(\boldsymbol{\alpha},\mathbf{P},\Phi_{lj},l,j\neq l\in\mathcal{S})$.

Consider now G independent semi-Markov processes taking values in the same state space $\mathcal{S}$ and, for $g=1,\ldots,G$, the initialization vector of probabilities $\boldsymbol{\alpha}^g$, the transition matrix $\mathbf{P}^g$ and the cumulative distribution functions for the sojourn times $\Phi^g_{lj}(t)$, $t\in T$. Denoting by $\pi_g>0$ the probability of observing a Markov renewal process with parameters $(\boldsymbol{\alpha}^g,\mathbf{P}^g,\Phi^g_{lj},l,j\neq l\in\mathcal{S})$, we consider the finite mixture process $(J^\pi_p,X^\pi_p)_{p\geqslant 1}$ whose law is given by

$$\sum_{g=1}^{G}\pi_g\,\mathrm{Law}(\boldsymbol{\alpha}^g,\mathbf{P}^g,\Phi^g_{lj},l,j\neq l\in\mathcal{S}). \tag{3}$$

The following proposition states that a finite mixture of Markov renewal processes is a Markov renewal process.

Proposition 1. The process $(J^\pi_p,X^\pi_p)_{p\geqslant 1}$ is a Markov renewal process with parameters

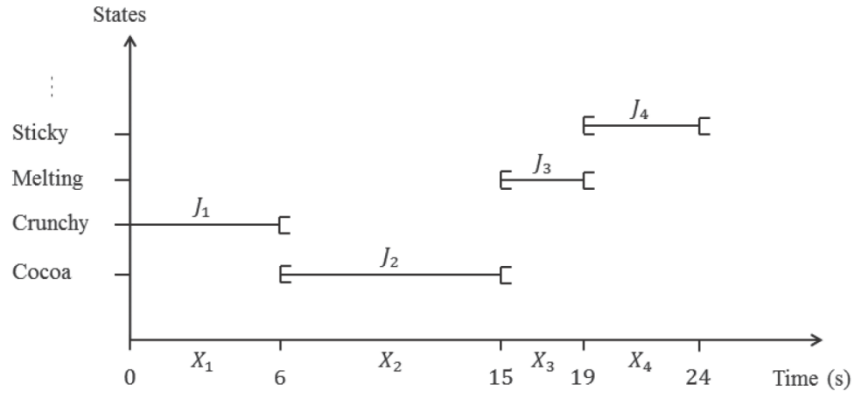


Fig. 2. Modelling of sequence S_4 (see Fig. 1) with a Markov renewal process $(J_p, X_p)_{p\geqslant 1}$: the successive states chosen by the panellist are J_1, crunchy, J_2, cocoa, J_3, melting and J_4, sticky

$$\left(\sum_{g=1}^{G} \pi_g \boldsymbol{\alpha}^g,\ \sum_{g=1}^{G} \pi_g \mathbf{P}^g,\ \sum_{g=1}^{G} \pi_g \Phi_{lj}^g, l, j \neq l \in \mathcal{S} \right).$$

2.2. The identifiability issue

Identifiability of mixture models can be a complicated issue (see for example Teicher (1963), Yakowitz and Spragins (1968), Titterington *et al.* (1985) or Allman *et al.* (2009)). However, identifiability of the parameters of a stochastic model is a very important condition to ensure the convergence of estimation algorithms to a unique value. We consider here a parametric framework and we are interested in models that are defined by a family of distributions $\mathcal{F}(\boldsymbol{\Theta}) = \{\text{Law}(\boldsymbol{\theta}), \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ where $\boldsymbol{\Theta} \subset \mathbb{R}^q$ is the parameter space and $\boldsymbol{\theta}$ is a vector of parameters characterizing the probability distribution. We consider convex combinations of probability laws in $\mathcal{F}(\boldsymbol{\Theta})$, $\Sigma_{g=1}^{G} c_g \text{Law}(\boldsymbol{\theta}_g)$, with $\Sigma_g c_g = 1$, $c_g > 0$ and $\boldsymbol{\theta}_g \in \boldsymbol{\Theta}$, for $g = 1, \ldots, G$.

Adopting the same definition as in Yakowitz and Spragins (1968), we say that the finite mixtures are identifiable in the family $\mathcal{F}(\boldsymbol{\Theta})$ if and only if the convex hull of $\mathcal{F}(\boldsymbol{\Theta})$ has the uniqueness representation property:

$$\sum_{g=1}^{G} c_g \text{Law}(\boldsymbol{\theta}_g) = \sum_{h=1}^{H} c'_h \text{Law}(\boldsymbol{\theta}_h) \tag{4}$$

implies that $G = H$ and for each $g \in 1, \ldots, G$ there is some $h \in \{1, \ldots, G\}$ such that $c_g = c'_h$ and $\boldsymbol{\theta}_g = \boldsymbol{\theta}_h$.

Moreover, it was proved by Yakowitz and Spragins (1968) that finite mixtures of a family $\mathcal{F}(\boldsymbol{\Theta})$ are identifiable if and only if the cumulative distribution functions of the elements are linearly independent.

We suppose from now that the family of distributions of the sojourn times is parametric: $\Phi_{lj}(t) = \Phi(t, \boldsymbol{\Gamma}_{lj})$, with $\boldsymbol{\Gamma}_{lj} \in \mathbb{R}^d$. Classical parametric distributions of sojourn times are the negative binomial distribution if time is discrete ($d = 2$) and exponential ($d = 1$) or gamma distributions ($d = 2$) if time is continuous. In our renewal Markov processes framework, a parameter $\boldsymbol{\theta}_g$ will be of the form $\boldsymbol{\theta}_g = (\boldsymbol{\alpha}^g, \mathbf{P}^g, \boldsymbol{\Gamma}_{lj}^g, l \in \mathcal{S}, j \neq l \in \mathcal{S})$. It was shown by Teicher (1963) that finite mixtures of gamma distributions are identifiable whereas it was proved by Yakowitz and Spragins (1968) that finite mixtures of exponential distributions as well as negative binomial

distributions are also identifiable. More recently, it has been shown by Gupta *et al.* (2016), under technical assumptions, that almost all finite mixtures of Markov chains with D states are identifiable provided that at least two consecutive transitions can be observed and the number of mixture components is not too large compared with the number of states; more precisely $D \geqslant 2G$.

As shown in the next proposition, the identifiability of mixtures of renewal Markov processes can be assessed under weaker conditions than those required for mixtures of Markov chains because the sojourn time distributions are not directly related to transition probabilities and there is no need to introduce any particular condition on the number of mixture components or on the number of states of the Markov chain. See also Gassiat *et al.* (2016) for an intermediate identifiability result stated for hidden Markov chains, which does not impose any condition on the number of states and mixture components but is based on knowledge of the law of at least two consecutive transitions.

For simplicity we assume that all the initialization probabilities α_l^g and all the transition probabilities P_{lj}^g are strictly positive. This ensures that all the sojourn time distributions can be observed by considering the law of $(J_1^\pi, X_1^\pi, J_2^\pi)$.

Hypothesis 1. $\forall\, g \in \{1, \ldots, G\}, \forall\, l \in \mathcal{S},\ \alpha_l^g > 0$ and $\forall\, j \neq l,\ P_{lj}^g > 0$.

We also need to add the following hypothesis which means that two subpopulations g and g' cannot have exactly the same set of parameters for the distributions of duration times.

Hypothesis 2. $\forall\, g \in \{1, \ldots, G\}$ and $\forall\, g' \neq g, \exists\, l \in \mathcal{S}$ and $j \neq l$ such that $\mathbf{\Gamma}_{lj}^g \neq \mathbf{\Gamma}_{lj}^{g'}$.

If this condition is not fulfilled, we may have two mixture components whose sojourn time distributions are exactly the same. In that case, we cannot distinguish the two corresponding subpopulations according to their sojourn times.

Proposition 2. Suppose that the family of sojourn time distributions is identifiable and that hypotheses 1 and 2 hold. Then all the finite mixtures from the family $\mathcal{F}(\mathbf{\Theta})$ can be identified when the law of the sequence $(J_1^\pi, X_1^\pi, J_2^\pi)$ drawn from a mixture of renewal Markov processes is known.

In other words, it is possible to identify the parameters of a finite mixture from $\mathcal{F}(\mathbf{\Theta})$ provided that we can observe at least one transition and the first sojourn times and the first state. The condition $\boldsymbol{\alpha}_l^g > 0$, which was also required in Gupta *et al.* (2016), ensures that all the possible transitions can be observed during the first transition. This hypothesis could be weakened by considering the law of mixture sequences with more than one transition. The condition on the transition probabilities that must be strictly positive is essentially of a technical nature and enables us to simplify the demonstration. Note that $P_{lj}^g = 0$ means that the transition from l to j is never observed so we cannot associate a duration time distribution with the transition from state l to state j in mixture component g. Without hypothesis 1, we should restrict the set of indices that are related to the sojourn times to the set corresponding to strictly positive transition probabilities.

3. Maximum likelihood estimation and model selection

Suppose that we have a sample of n independent consumers, for which we may consider B independent and identically distributed replications of the tasting experiment. For each consumer i, with $i = 1, 2, \ldots, n$, we thus obtain B sequences S_i^b, for $b = 1, \ldots, B$, observed for $t \leqslant T_i^b$ and denoted by

$$S_i^b = (J_1^{i,b}, X_1^{i,b}, \ldots, J_{N(T_i^b)-1}^{i,b}, X_{N(T_i^b)-1}^{i,b}, J_{N(T_i^b)}^{i,b}, X_{N(T_i^b)}^{i,b}), \tag{5}$$

where $N(T_i^b)$ is the random number of states visited by consumer i during replication b. We suppose that $N(T_i^b) \geqslant 2$.

We suppose that the observed trajectories $S_1^1, \ldots, S_1^B, \ldots, S_n^1, \ldots, S_n^B$ are drawn from a mixture of G semi-Markov processes whose law is given in equation (3) and we aim at estimating the parameters which characterize the law of the mixture: the vector of mixture proportions $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$ and $(\boldsymbol{\alpha}^g, \mathbf{P}^g, \Phi_{lj}^g, l \in \mathcal{S}, j \neq l \in \mathcal{S})$, for $g = 1, \ldots, G$, which characterize the law of the semi-Markov processes for each mixture component. We suppose in this section that the number G of components is known.

3.1. The particular case of gamma-distributed sojourn times with replications and no anticipation

In our sensory examples, sojourn times are positive and continuous random variables and we suppose that they are distributed according to gamma distributions. The choice of the gamma distribution is motivated by its simplicity and its ability to fit sojourn time distributions with many different shapes. The density depends on two parameters, the shape parameter $a > 0$ and $\lambda > 0$, and is defined as follows:

$$f(t, a, \lambda) = \frac{t^{a-1}\lambda^a \exp(-\lambda t)}{\Gamma(a)}, \qquad t \geqslant 0,$$

where $\Gamma(a)$ is the gamma function. The corresponding expected value is a/λ and the variance a/λ^2.

We suppose, as in Lecuelle *et al.* (2018), that the sojourn time distribution depends on the current state only:

$$\Pr(X_p^{\pi} \leqslant t | J_p = l, J_{p+1} = j, Z = g) = \Pr(X_1^g \leqslant t | J_1 = l) \tag{6}$$

so there is no anticipation, in some sense, of the next dominant attribute. This assumption, which seems relevant in a food tasting context, also enables us to deal with moderate size samples by reducing significantly the number of parameters to be estimated. In that case, hypothesis 2 means that, for each mixture component g and g', there is at least one state l such that the two cumulative distributions $\Pr(X_1^g \leqslant t | J_1 = l)$ and $\Pr(X_1^{g'} \leqslant t | J_1 = l)$ are not equal. If we denote by d the number of parameters that are required to characterize each sojourn time distribution, we need only, with this simplification, to estimate GDd parameters to characterize the sojourn time distributions instead of $GD(D-1)d$ in the more general setting that was studied in the previous section. Note that from now on $d = 2$, which corresponds to the particular case of gamma-distributed sojourn times.

3.2. The likelihood

By successive conditioning, the likelihood that is related to a statistical unit i with B independent replications drawn from a Markov renewal process with parameters $\boldsymbol{\theta}_g = (\boldsymbol{\alpha}^g, \mathbf{P}^g, (a_{lg}, \lambda_{lg}), l \in \mathcal{S})$ can be written

$$\begin{aligned} L_g(S_i^1, \ldots, S_i^B; \boldsymbol{\theta}_g) &= \prod_{b=1}^{B} L_g(S_i^b; \boldsymbol{\theta}_g) \\ &= \prod_{b=1}^{B} \left\{ \alpha_{J_1^{i,b}}^g \phi_{J_1^{i,b}}^g (X_1^{i,b}) \prod_{k=2}^{N(T_i^b)} \mathbf{P}_{J_{k-1}^{i,b} J_k^{i,b}}^g \phi_{J_k^{i,b}}^g (X_k^{i,b}) \right\}, \end{aligned} \tag{7}$$

where $\phi_l^g(x) = f(x, a_{lg}, \lambda_{lg})$ is the density function evaluated at x of a gamma random variable with parameters $a = a_{lg}$ and $\lambda = \lambda_{lg}$.

If we do not suppose anymore that the mixture component from which unit i arises is known, the log-likelihood under the mixture model of the nB trajectories becomes

$$\ln\{L(S_1^1, \ldots, S_n^B; \boldsymbol{\theta})\} = \sum_{i=1}^{n} \ln\left\{ \sum_{g=1}^{G} \pi_g \prod_{b=1}^{B} L_g(S_i^b; \boldsymbol{\theta}_g) \right\}, \tag{8}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$ is the set of parameters of the mixture model. A direct maximization of the log-likelihood (8), according to $\boldsymbol{\theta}$, is cumbersome and classical optimization algorithms are generally not suitable to deal with that kind of problem (see for example McLachlan and Krishnan (2008)). The EM algorithm, which is presented below, is preferred because it enables the optimization procedure to be decomposed into two simple steps.

3.3. The expectation–maximization algorithm

The EM algorithm is a very useful algorithm that was first designed to perform maximum likelihood estimation for incomplete-data problems (see Dempster *et al.* (1977)). It is an iterative optimization technique of the likelihood that can be very effective for estimating mixture models by considering the unknown mixture components as missing observations (see McLachlan and Peel (2000)).

Introduce the missing mixture component indicators Z_i, for $i = 1, \ldots, n$, which are vectors with G elements, composed of one 1 and $G - 1$ 0s and that indicates from which component of the mixture the trajectory S_i arises. In other words, if S_i has been generated by the gth mixture component then $Z_{ig} = 1$ and $Z_{il} = 0$ for $l \neq g$. The complete-data log-likelihood can be written as

$$\begin{aligned} \ln\{L_c(S_1^1, \ldots, S_1^B, Z_1, \ldots, S_n^1, \ldots, S_n^B, Z_n; \boldsymbol{\theta})\} &= \sum_{i=1}^{n} \sum_{g=1}^{G} Z_{ig} \ln\left\{ \pi_g \prod_{b=1}^{B} L_g(S_i^b; \boldsymbol{\theta}_g) \right\} \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} Z_{ig} \ln(\pi_g) + \sum_{i=1}^{n} \sum_{g=1}^{G} Z_{ig} \sum_{b=1}^{B} \ln\{L_g(S_i^b; \boldsymbol{\theta}_g)\}. \end{aligned} \tag{9}$$

This function is much easier to maximize, according to $\boldsymbol{\theta}$, than the log-likelihood function that is given in equation (8).

An initial value $\boldsymbol{\theta}^{(0)}$ of the parameters must be carefully chosen before starting the algorithm. The choice of the starting point can be of great importance and is discussed in Section 3.4. The EM algorithm proceeds iteratively according to the following scheme. Suppose that an estimate of $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}^{(m-1)}$, has been calculated at step $m - 1$, with $m \geqslant 1$.

3.3.1. Expectation step

The expectation step consists in computing the expected log-likelihood of the complete data given the observed trajectories and the value of the parameters estimated during the previous iteration. We define

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m-1)}) &= E[\ln\{L_c(S_1, Z_1, \ldots, S_n, Z_n; \boldsymbol{\theta}) | S_1, \ldots, S_n, \boldsymbol{\theta}^{(m-1)}\}] \\ &= \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{Z}_{ig}^{(m)} \sum_{b=1}^{B} \ln\{L_g(S_i^b; \boldsymbol{\theta}_g^{(m-1)})\} + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{Z}_{ig}^{(m)} \ln(\pi_g^{(m-1)}), \end{aligned} \tag{10}$$

with $\hat{Z}_{ig}^{(m)} = E[Z_{ig}|S_1, \ldots, S_n, \boldsymbol{\theta}^{(m-1)}]$, the conditional probability for S_i to be generated by the component g of a mixture model with parameters $\boldsymbol{\theta}^{(m-1)}$, where $\boldsymbol{\theta}^{(m-1)}$ is the value of the set of parameters computed in the previous iteration. We obtain, with Bayes theorem,

$$\begin{aligned}\hat{Z}_{ig}^{(m)} &= \Pr(Z_{ig} = 1 | S_i; \boldsymbol{\theta}^{(m-1)}) \\ &= \frac{\pi_g^{(m-1)} \prod_{b=1}^{B} L_g(S_i^b; \boldsymbol{\theta}^{(m-1)})}{\sum_{j=1}^{G} \pi_j^{(m-1)} \prod_{b=1}^{B} L_j(S_i^b; \boldsymbol{\theta}^{(m-1)})}.\end{aligned} \tag{11}$$

3.3.2. Maximization step

The maximization step consists in updating the value of parameter $\boldsymbol{\theta}$ given the expected values of $\hat{Z}_{ig}$, for $g = 1, \ldots, G$ and $i = 1, \ldots, n$, by looking for the maximum, according to $\boldsymbol{\theta}$, of the function $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(m-1)})$ defined in equation (10). The mixture probabilities π_g appear only in the second term on the right-hand side of equation (10). The new estimates at step m are obtained by solving

$$\frac{\partial}{\partial \pi_g} \left\{ \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{Z}_{ig}^{(m)} \ln(\pi_g) + \lambda \left(\sum_{g=1}^{G} \pi_g - 1 \right) \right\} = 0, \tag{12}$$

where λ is the Lagrange multiplier that is associated with the constraint $\Sigma_{g=1}^{G} \pi_g = 1$. We obtain the standard solution $\pi_g^{(m)} = n^{-1} n_g^{(m)}$, with $n_g^{(m)} = \Sigma_{i=1}^{n} \hat{Z}_{ig}^{(m)}$.

The G Markov chain transition matrices and initialization probabilities $(\boldsymbol{\alpha}_g, \mathbf{P}_g, g = 1, \ldots, G)$ as well as the parameters that are related to the sojourn time distributions (a_{lg}, λ_{lg}) are updated by maximizing the first term on the right-hand side of equation (10).

Thanks to the multiplicative structure of likelihood equation (7) given the mixture component, the first term on the right-hand side of equation (10) can be written as the sum of two distinct functions, where the first depends only on the semi-Markov chains parameters $(\boldsymbol{\alpha}_g, \mathbf{P}_g, g = 1, \ldots, G)$ whereas the second depends only on the sojourn time distributions $(a_{lg}, \lambda_{lg}, l \in \mathcal{S}, g = 1, \ldots, G)$. Thus, these two sets of parameters can be estimated separately by maximizing each part of the log-likelihood during the maximization step.

Introducing again Lagrange multipliers, this yields the standard solution for the transition probabilities estimators as well as the initialization probabilities:

$$\begin{aligned}\hat{\alpha}_j^{g(m)} &= \frac{\sum_{i=1}^{n} \hat{Z}_{ig}^{(m)} \sum_{b=1}^{B} \mathbb{1}_{\{J_1^{i,b} = j\}}}{B \sum_{i=1}^{n} \hat{Z}_{ig}^{(m)}}, \\ \hat{\mathbf{P}}_{hj}^{g(m)} &= \frac{\sum_{i=1}^{n} \hat{Z}_{ig}^{(m)} \sum_{b=1}^{B} n_{hj}^{ib}}{\sum_{l=1}^{D} \sum_{i=1}^{n} \hat{Z}_{ig}^{(m)} \sum_{b=1}^{B} n_{hl}^{ib}},\end{aligned} \tag{13}$$

where n_{hj}^{ib} is the number of $h \to j$ transitions for trajectory S_i^b.

It was shown by Chen *et al.* (2016) that for gamma mixture models the log-likelihood is not bounded. Intuitively, the degeneracy comes from the fact that, if the ratio a_{lg}/λ_{lg}, which

corresponds to the expected sojourn time in state l for mixture g, is kept constant, while a_{lg} is tending to ∞, then the corresponding variance a_{lg}/λ_{lg}^2 will tend to 0 and the corresponding gamma density, mimicking the Dirac distribution at a_{lg}/λ_{lg}, will not be bounded. Consequently, to avoid such a degenerate solution, it may be preferable to introduce a penalization in the maximization step that prevents the parameters a_{lg} from becoming too large. Thus, we add to the function Q, defined in equation (10), a penalty that is similar to the penalty given in Chen *et al.* (2016) and defined as follows:

$$\text{Pen}(a_{lg}, l \in \mathcal{S}, g = 1, \ldots, G) = -\frac{1}{\sqrt{\left\{ \sum_{i=1}^{n} \sum_{b=1}^{B} N(T_i^b) \right\}}} \sum_{g=1}^{G} \sum_{l \in \mathcal{S}} \{a_{lg} + \ln(a_{lg})\}. \tag{14}$$

Note that this penalty does not need to take into account the parameters λ_{lg} of the gamma distributions. Its effect decreases as the sample size and the number of observed transitions increase.

Finally, the parameters of the sojourn time distributions can be estimated by maximizing the following expected partial penalized log-likelihood:

$$\sum_{i=1}^{n} \sum_{g=1}^{G} \hat{Z}_{ig}^{(m)} \sum_{b=1}^{B} \sum_{k=1}^{N(T_i^b)} \ln\{\phi_{J_k^{i,b}}^{g}(X_k^{i,b})\} + \text{Pen}(a_{lg}, l \in \mathcal{S}, g = 1, \ldots, G), \tag{15}$$

with classical optimization procedures.

Once the algorithm has converged, model-based clustering of the observed sequences is performed by considering the MAP probability criterion, which is defined as $\text{MAP}(\hat{Z}_{ig}) = 1$ if $g = \arg\max_h(\hat{Z}_{ih})$ and $\text{MAP}(\hat{Z}_{ig}) = 0$ otherwise.

3.4. Choosing the starting point of the expectation–maximization algorithm

A crucial issue for the EM algorithm is the choice of the value of the starting point $\boldsymbol{\theta}^{(0)}$. It was shown in Galmarini *et al.* (2017) that the time that is spent in each state provides an interesting indicator to study TDS data. Thus, we have chosen to select the initial values of the EM algorithm by considering the Hartigan–Wong k-means algorithm (Hartigan and Wong, 1979) applied to the D-dimensional vector of mean sojourn times in each state, with the Euclidean distance and $k = G$ clusters. A heuristic justification can be given by the fact that the identification of the mixture components seems to be easier for sojourn times. Indeed as seen in the proof of proposition 2, the sojourn time distribution of any finite mixture of Markov renewal processes is identifiable when the family of sojourn time distributions is identifiable. Then, the method of moments is employed to obtain the initial values for the transition matrices and for the parameters of the gamma distributions.

3.5. Selection of the number of mixture components

When the number of components G is unknown, an information criterion can be used to select the number of mixture components (see McLachlan and Peel (2000) for a detailed presentation of the various approaches that have been developed in the literature). Such information criteria rely on a compromise between the fit to the data and the complexity of the model considered, more complex models being less desirable. The Bayes information criterion (BIC), which has good asymptotic properties (see Keribin (2000)), is simple to compute and seems effective to select the number of components. We choose to define the BIC as follows:

$$\mathrm{BIC}(G) = q\ln(nB) - 2\ln[L\{S_1^b, \ldots, S_n^b, b = 1, \ldots, B;\ \hat{\boldsymbol{\theta}}(G)\}], \tag{16}$$

where $\hat{\boldsymbol{\theta}}(G)$ is the estimator of $\boldsymbol{\theta}$ when a mixture of G components has been considered and $q = q\{\boldsymbol{\theta}(G)\}$ is the number of free parameters to be estimated. nB corresponds to the number of independent observations in classical mixture models and could be different in our setting of temporal data. Indeed, it is not completely clear which value should be considered for the sample size since, for each trajectory, we have observations that are correlated over time (see Pamminger and Frühwirth-Schnatter (2010) for a discussion in a similar context of mixtures of Markov chains) and we could also take account of the number of observed transitions. In the particular setting that was described in Section 3.1 and taking into account the fact that $P_{ll}^g = 0$ for all $g = 1, \ldots, G$ and all $l \in \mathcal{S}$, we obtain $q = G - 1 + G\{D - 1 + D(D-2) + Dd\} = GD(D + d - 1) - 1$, with $d = 2$ for the two-parameter gamma distribution. If there is one absorbing state in $\mathcal{S}$, as in the example in Section 5, and we suppose that it is not possible that the first observed state is this absorbing state, then $q = G - 1 + G\{D - 2 + (D-1)(D-2) + (D-1)d\}$.

Other popular criteria are the Akaike information criterion (AIC), in which the term $q\ln(nB)$ in equation (16) which penalizes the complexity of the model is replaced by $2q$ and the corrected AIC, denoted by $\mathrm{AIC_c}$, in which the term $q\ln(nB)$ is replaced by $2q + 2q(q+1)/(nB - q - 1)$.

4. A simulation study

A simulation study is conducted to evaluate the performances of the penalized and unpenalized EM algorithms under various mixture scenarios. We also measure the ability of the AIC and the BIC to select the correct number of mixture components. Simulations are performed by using the R language (R Core Team, 2018) and C++ with the `Rcpp` package (Eddelbuettel and François, 2011). Programs are available on request to the authors.

4.1. Simulation protocol and indicators of performance

To obtain realistic simulations, we simulated qualitative trajectories based on semi-Markov chains whose parameters were estimated on the real data set that was presented in Section 1. In that experiment, panellists evaluated three chocolates, with a list of $D = 10$ attributes, the first chocolate with 70% of cocoa, the second with 70% of cocoa also but sweeter than the first and the third with 90% of cocoa (see Visalli *et al.* (2016) for a more detailed presentation of the data). An experiment that was related to the tasting of the first chocolate is presented in Fig. 1. The components of the renewal Markov process corresponding to each chocolate are estimated by maximum likelihood (see Lecuelle *et al.* (2018)), considering gamma-distributed sojourn times with no anticipation effect, as in Section 3.1. Estimated parameters of the semi-Markov chains associated with the three chocolates are reported in section B of the on-line supplementary file.

First, we consider a known number of components equal to 2, with simulated sequences of four or 10 transitions. Note that to be able to control the number of transitions we do not introduce any absorbing state. We study two cases of mixtures: the first with two well-separated subpopulations (the chocolates with 70% and 90% of cocoa) and the second with two populations with similar distributions (the two chocolates with 70% of cocoa).

Second, we assume that the number of components is unknown to evaluate the ability of the various information criteria that were presented in Section 3.5 to recover the true number of components in the population. We consider three configurations: one with only one component (chocolate with 70% of cocoa), one with two well-separated components (the chocolates with 70% and 90% of cocoa) and one with two similar components (the two chocolates with 70% of cocoa). The selection of the number of components is a difficult task and the information

criteria do not always give good results with stochastic processes (see for example Celeux and Durand (2008)).

Thanks to our knowledge of these chocolates, we can assume that some transitions are not possible (occur with a probability 0). Taking this information into account, we can reduce the number of transition parameters to be estimated. We have 49 unknown probability transition parameters for the chocolate with 70% of cocoa, 62 unknown parameters when considering the two chocolates with 70% of cocoa and 69 unknown parameters when considering the chocolate with 70% of cocoa and the chocolate with 90% of cocoa.

For simulating mixtures with $G = 2$ components, the number of individuals belonging to each component is randomly selected thanks to the binomial law $B(n, 0.5)$, meaning that $\pi_1 = \pi_2 = \frac{1}{2}$. Then, for each type of chocolate, individual trajectories are simulated sequentially by selecting randomly the successive states and durations according to the estimated transition probabilities and dominance duration distributions. For each case, we simulated 500 data sets with samples of sizes $n = 60$, $n = 200$ and $n = 600$ and $B = 3$ replications.

To avoid computation issues when estimating the parameters that are related to the gamma distributions, the values of $\hat{Z}_{ig}^{(m)}$ are rounded to 10^{-4} and the maximum likelihood estimation is performed only when there are more than seven observations. Otherwise the gamma parameters are set to the values that were estimated on all the observations belonging to the corresponding mixture, independently of the state.

The number of maximal iterations of the EM algorithm is set to 100. *A posteriori* this was sufficiently large because, for all the designs considered, convergence was achieved before 100 iterations.

To check whether the transition matrices are well estimated, we consider the following relative error between the estimated transition matrices $\hat{\mathbf{P}}^g$ and the transition matrices $\mathbf{P}^g$ that were used to generate the simulated data for component g:

Table 1. Parameter estimation errors when considering unpenalized EM for two clusters with $n = 60$, $n = 200$ and $n = 600$ and with simulated sequences with four and 10 transitions and $B = 3$ repetitions†

n	$Err(\boldsymbol{\alpha}^1)$	$Err(\boldsymbol{\alpha}^2)$	$Err(\mathbf{P}^1)$	$Err(\mathbf{P}^2)$	$Err(a)$	$Err(\lambda)$	$\pi_1 = 0.5$
With 4 transitions							
Well-separated components							
60	0.01 (0.01)	0.01 (0.02)	0.26 (0.12)	0.18 (0.10)	0.23 (0.75)	0.40 (0.91)	0.55 (0.14)
200	<0.01 (<0.01)	<0.01 (<0.01)	0.06 (0.04)	0.04 (0.02)	0.04 (0.06)	0.08 (0.14)	0.50 (0.04)
600	<0.01 (<0.01)	<0.01 (<0.01)	0.02 (0.01)	0.01 (<0.01)	0.01 (0.01)	0.02 (0.02)	0.50 (0.02)
Not well-separated							
60	0.01 (0.01)	0.03 (0.04)	0.33 (0.13)	0.47 (0.15)	0.44 (1.77)	0.70 (3.00)	0.61 (0.25)
200	<0.01 (<0.01)	0.01 (0.03)	0.10 (0.08)	0.16 (0.16)	0.29 (2.18)	0.38 (2.93)	0.49 (0.12)
600	<0.01 (<0.01)	<0.01 (<0.01)	0.02 (0.01)	0.01 (0.03)	0.03 (0.06)	0.04 (0.12)	0.50 (0.03)
With 10 transitions							
Well-separated components							
60	0.01 (0.01)	0.01 (0.01)	0.08 (0.06)	0.05 (0.04)	0.07 (0.12)	0.15 (0.21)	0.50 (0.10)
200	<0.01 (<0.01)	<0.01 (<0.01)	0.02 (0.01)	0.01 (<0.01)	0.01 (0.01)	0.02 (0.02)	0.50 (0.04)
600	<0.01 (<0.01)	<0.01 (<0.01)	0.01 (<0.01)	<0.01 (<0.01)	<0.01 (<0.01)	0.01 (<0.01)	0.50 (0.02)
Not well separated							
60	0.01 (0.01)	0.03 (0.07)	0.09 (0.06)	0.23 (0.19)	0.18 (0.41)	0.21 (0.85)	0.62 (0.19)
200	<0.01 (<0.01)	<0.01 (<0.01)	0.02 (0.01)	0.02 (0.06)	0.03 (0.04)	0.03 (0.04)	0.53 (0.07)
600	<0.01 (<0.01)	<0.01 (<0.01)	0.01 (<0.01)	<0.01 (<0.01)	0.01 (0.01)	0.01 (0.01)	0.50 (0.02)

†For each design, the mean and standard deviation, in parentheses, are computed considering 500 simulated data sets.

Table 2. Parameter estimation errors when considering penalized EM for two clusters with $n = 60$, $n = 200$ and $n = 600$ and with simulated sequences with four and 10 transitions and $B = 3$ repetitions†

n	$Err(\boldsymbol{\alpha}^1)$	$Err(\boldsymbol{\alpha}^2)$	$Err(\mathbf{P}^1)$	$Err(\mathbf{P}^2)$	$Err(a)$	$Err(\lambda)$	$\pi_1 = 0.5$
With 4 transitions							
Well-separated components							
60	0.01 (0.01)	0.01 (0.02)	0.26 (0.13)	0.18 (0.10)	0.10 (0.07)	0.24 (0.15)	0.54 (0.15)
200	<0.01 (<0.01)	<0.01 (<0.01)	0.06 (0.04)	0.04 (0.02)	0.03 (0.03)	0.06 (0.07)	0.50 (0.05)
600	<0.01 (<0.01)	<0.01 (<0.01)	0.02 (0.01)	0.01 (<0.01)	0.01 (<0.01)	0.01 (0.01)	0.50 (0.02)
Not well separated							
60	0.01 (0.01)	0.04 (0.08)	0.32 (0.13)	0.48 (0.16)	0.11 (0.11)	0.22 (0.42)	0.61 (0.26)
200	<0.01 (<0.01)	0.01 (0.02)	0.10 (0.07)	0.15 (0.16)	0.09 (0.19)	0.11 (0.22)	0.50 (0.12)
600	<0.01 (<0.01)	<0.01 (<0.01)	0.02 (0.01)	0.01 (0.05)	0.03 (0.22)	0.04 (0.30)	0.50 (0.03)
With 10 transitions							
Well-separated components							
60	0.01 (0.01)	0.01 (0.01)	0.08 (0.07)	0.06 (0.05)	0.06 (0.04)	0.13 (0.11)	0.49 (0.11)
200	<0.01 (<0.01)	<0.01 (<0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.03)	0.50 (0.04)
600	<0.01 (<0.01)	<0.01 (<0.01)	0.01 (<0.01)	<0.01 (<0.01)	<0.01 (<0.01)	0.01 (<0.01)	0.50 (0.02)
Not well separated							
60	0.01 (0.01)	0.02 (0.04)	0.10 (0.06)	0.20 (0.18)	0.09 (0.14)	0.10 (0.20)	0.62 (0.18)
200	<0.01 (<0.01)	<0.01 (<0.01)	0.02 (0.01)	0.01 (0.04)	0.03 (0.02)	0.03 (0.03)	0.53 (0.07)
600	<0.01 (<0.01)	<0.01 (<0.01)	0.01 (<0.01)	<0.01 (<0.01)	0.01 (<0.01)	0.01 (<0.01)	0.50 (0.02)

†For each design, the mean and the standard deviation, in parentheses, are computed considering 500 simulated data sets.

Table 3. Correct classification rate for two clusters with well-separated components and not-well-separated components with $n = 60$, $n = 200$ and $n = 600$ and with length of simulated sequences equal to four and 10 transitions†

	Results for well-separated components			*Results for not-well-separated components*		
	$n = 60$	$n = 200$	$n = 600$	$n = 60$	$n = 200$	$n = 600$
With 4 transitions						
k-means	0.85 (0.07)	0.86 (0.05)	0.86 (0.03)	0.81 (0.09)	0.78 (0.08)	0.76 (0.06)
Mixture model	0.92 (0.07)	0.99 (0.02)	1 (<0.01)	0.82 (0.09)	0.93 (0.06)	0.98 (0.02)
With 10 transitions						
k-means	0.87 (0.07)	0.89 (0.04)	0.90 (0.02)	0.83 (0.07)	0.84 (0.05)	0.85 (0.03)
Mixture model	0.97 (0.05)	1 (0.01)	1 (<0.01)	0.89 (0.09)	0.97 (0.05)	1 (<0.01)

†For each design, the mean and standard deviation, in parentheses, are computed from 500 simulated data sets.

$$\text{Err}(\mathbf{P}^g) = \frac{\|\mathbf{P}^g - \hat{\mathbf{P}}^g\|_2^2}{\|\mathbf{P}^g\|_2^2}, \tag{17}$$

where $\|\mathbf{P}\|_2 = \text{tr}(\mathbf{P}'\mathbf{P})$ is the squared Frobenius norm of matrix $\mathbf{P}$. A similar error is computed for the initial probabilities:

$$\text{Err}(\boldsymbol{\alpha}^g) = \frac{\|\boldsymbol{\alpha}^g - \hat{\boldsymbol{\alpha}}^g\|_2^2}{\|\boldsymbol{\alpha}^g\|_2^2}. \tag{18}$$

Table 4. Choice of the number of components with one component, two well-separated components and two not-well-separated components and four or 10 observed transitions†

	Results for one component selected			*Results for two components selected*					
				Well separated			*Not well separated*		
	$n=60$	$n=200$	$n=600$	$n=60$	$n=200$	$n=600$	$n=60$	$n=200$	$n=600$
With 4 transitions									
BIC									
1	500	500	500	500	5	0	500	491	0
2	0	0	0	0	493	500	0	9	494
3	0	0	0	0	2	0	0	0	6
AIC									
1	500	500	500	93	0	0	491	4	0
2	0	0	0	394	473	498	9	373	487
3	0	0	0	13	27	2	0	123	13
With 10 transitions									
BIC									
1	500	500	500	43	0	0	411	0	0
2	0	0	0	457	497	500	89	431	495
3	0	0	0	0	3	0	0	69	5
AIC									
1	500	500	500	0	0	0	27	0	0
2	0	0	0	407	491	498	245	318	495
3	0	0	0	93	9	2	228	182	15

†The number of clusters selected by the BIC and the AIC are shown for 500 simulated data sets.

We also check whether the estimated parameters of the sojourn time gamma distribution are well estimated by considering the following relative errors:

$$\begin{aligned} \mathrm{Err}(a) &= \frac{\sum_{l=1}^{D}\sum_{g=1}^{G}(\hat{a}_l^g - a_l^g)^2}{\sum_{l=1}^{D}\sum_{g=1}^{G}(a_l^g)^2}, \\ \mathrm{Err}(\lambda) &= \frac{\sum_{l=1}^{D}\sum_{g=1}^{G}(\hat{\lambda}_l^g - \lambda_l^g)^2}{\sum_{l=1}^{D}\sum_{g=1}^{G}(\lambda_l^g)^2}. \end{aligned} \tag{19}$$

4.2. Results

Parameter estimation errors, evaluated with equations (17)–(19), are given in Table 1 for the unpenalized version of the EM algorithm and in Table 2 when the penalized version of the EM algorithm described in equation (15) was employed to estimate the parameters. We note that the introduction of the penalty enables us to improve the accuracy of the estimates, especially for small samples, few transitions or with clusters with a similar distribution of the semi-Markov processes. Without penalty, we observe larger mean errors for the estimated parameters of the sojourn time distribution and high values for the standard deviations of

Table 5. Values taken by the BIC, the AIC and the AIC_c for a number of clusters G ranging from 1 to 4 for the young and low fat Gouda cheese from the European Sensory Network data set

Criterion	*Results for the following values of G:*			
	1	*2*	*3*	*4*
BIC	87525.93	86872.04	87058.21	87599.31
AIC	86859.73	85534.05	85048.43	84917.74
AIC_c	86874.99	85776.07	85794.99	86858.84

Table 6. Estimated initial probabilities for the young and low fat Gouda cheese from the European Sensory Network data set, considering two or three mixture components

Cluster	*Results for the following attributes:*									
	Bitter	*Cheese*	*Dense hard*	*Fatty*	*Melting*	*Milky cream*	*Salty*	*Sharp*	*Sour*	*Tender*
With 2 components										
1	0.03	0.08	0.44	0.05	0.03	0.04	0.03	0.01	0.02	0.27
2	0.02	0.04	0.39	0.08	0.05	0.05	0.03	0.02	0.01	0.32
With 3 components										
1	0.02	0.10	0.15	0.07	0.06	0.07	0.04	0.01	0.02	0.46
2	0.02	0.04	0.41	0.08	0.04	0.04	0.02	0.02	0.01	0.32
3	0.03	0.05	0.75	0.03	0	0	0.03	0.02	0.02	0.06

the errors. For example, when $n = 60$ with only four observed transitions and clusters that are not well separated, we obtain $\mathrm{Err}(\lambda) = 0.70$ without penalty whereas this error is reduced to $\mathrm{Err}(\lambda) = 0.22$ thanks to the introduction of the penalty. When the sample size becomes larger ($n = 200$ or $n = 600$) and the number of transitions is large both estimation procedures lead to similar results.

From now on, we shall consider only estimates obtained with the penalized EM algorithm.

In our simulation context, we know for each trajectory which component of the mixture it belongs to and we can check whether it has been assigned with the MAP criterion to the right component. The rate of correct classification is given in Table 3. We note that, overall, the rate of well-classified trajectories is high with values ranging from 0.83 to 1. Model-based clustering substantially improves the classification accuracy compared with k-means, except for the more difficult case with a small sample ($n = 60$), four transitions and clusters not well separated, where neither approach performs well.

We present in Table 4 the number of components that were selected by the BIC and the AIC. Whatever the number of individuals, the BIC and the AIC select the correct number of components when there is only one component. With two well-separated clusters, the BIC and the AIC generally give good results, except for the case with four transitions and $n = 60$ where the BIC and, to a lesser degree, the AIC select only one component rather than two.

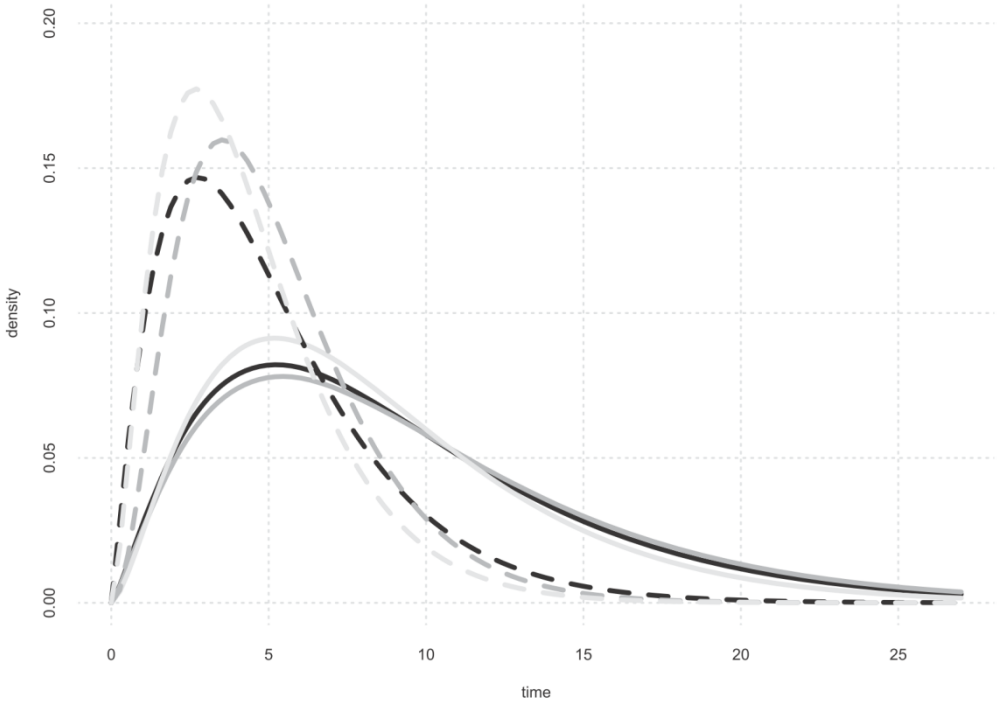


Fig. 3. Estimated sojourn time distributions for the attributes cheese (———), dense hard (———) and tender (———) when considering two mixture components for the young and low fat Gouda cheese: ———, cluster 1; – – –, cluster 2

When the two mixture components are not very different, the BIC and the AIC provide effective criteria for selecting the number of components only when the samples are large. The AIC often selects the same numbers of components as the BIC, but it sometimes selects too many components. For small samples and small numbers of transitions, the BIC is more restrictive and tends to lead to an underestimate of the true number of components. Similar conclusions, in a different context, were found by Celeux and Durand (2008). AIC_c can only be used with large samples because of the too large number of parameters of the model. It does not perform better than the BIC and the AIC in this simulation study and the corresponding results are not shown here.

5. Clustering temporal dominance of sensations for a Gouda cheese

We now study data resulting from an experiment of the European Sensory Network aiming at measuring simultaneously perception and liking of Gouda cheeses (Thomas *et al.*, 2017). A large panel of $n = 665$ consumers from six European countries tasted four Gouda cheeses with different ages and fat content according to the TDS protocol. A list of $D = 10$ attributes was presented to the consumers on a computer screen. Panellists tasted $B = 3$ successive bites so there are three sequences corresponding to the three repetitions for each panellist and for each product. In this sample, the mean number of transitions within an individual sequence is equal to 4.1 (see Table 14 in the on-line supplementary file).

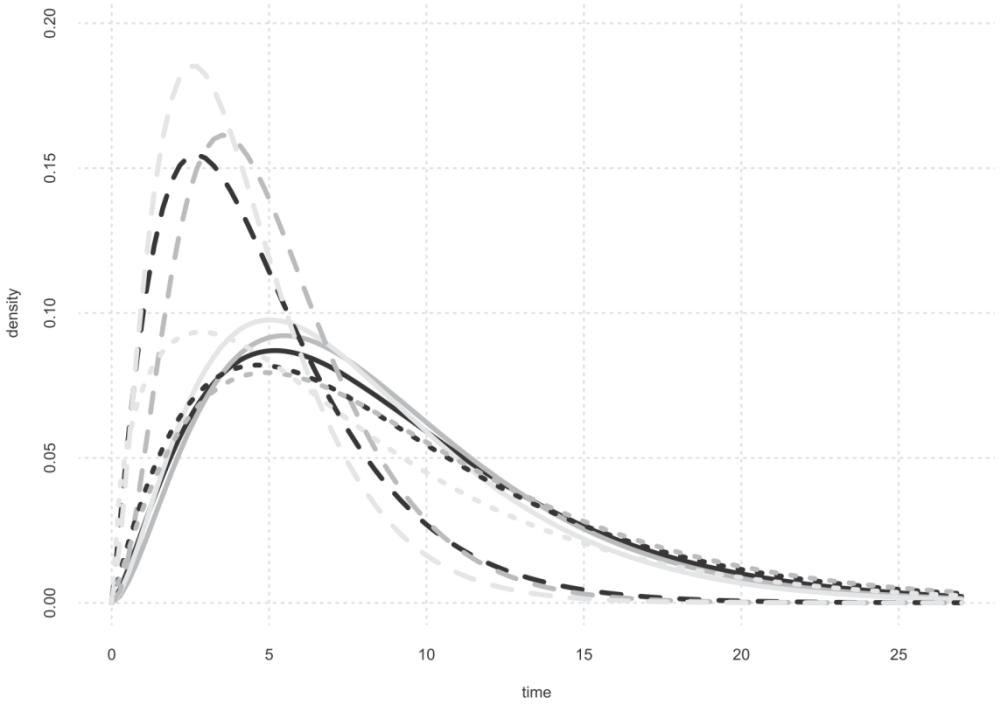


Fig. 4. Estimated sojourn time distributions for the attributes cheese (———), dense hard (———) and tender (———) when considering three components for the young and low fat Gouda cheese: ———, cluster 1; – – –, cluster 2; ------, cluster 3

Our goal in this study is to perform a segmentation of the panel, and to describe, if there are any, the differences of perceptions for a product. We present only the results for a young and low fat Gouda cheese, whose perception by consumers is more complex.

The maximal number of iterations of the EM algorithm was set to 400 because we observed that the algorithm requires more iterations to converge with this data set. This can be explained by a higher complexity of the model than in the simulation study because all transitions are possible with these products.

As shown in Table 5, all the information criteria approaches select at least two mixture components, showing the existence of different behaviours in the panel. The BIC suggests that two clusters should be considered and AIC_c suggests three clusters but both take really close values for two and three components. That is why we shall examine these two cases in what follows. The AIC suggests that at least four clusters should be considered but, as is well known, it is a less parsimonious criterion than BIC and AIC_c. With two components, the clusters obtained are respectively composed of 398 and 267 individuals, whereas, with three components, the clusters obtained are respectively composed of 242, 209 and 214 individuals.

The estimated initial probabilities are shown in Table 6. As expected in sensory studies, most of the panellists chose a texture attribute 'dense hard' or 'tender' as the first dominant attribute. With two components, the initial probabilities are really close for the two components, with only some small differences. With three components, large differences are observed between clusters, especially for the attributes dense hard and tender. In cluster 1, most of the panellists

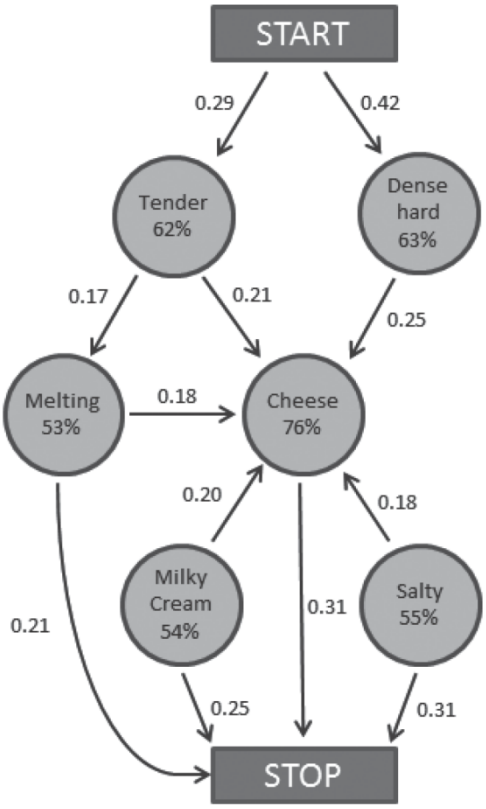


Fig. 5. TDS graph for the young and low fat Gouda cheese with no segmentation

chose tender as the first attribute whereas, in cluster 3, most of the panellists chose dense hard. In cluster 2, both dense hard and tender have a high probability of being chosen as the first attribute.

Fig. 3 presents the estimated gamma distributions of the sojourn times, with two components, for the attributes cheese, dense hard and tender whereas the estimated gamma distributions for the same attributes when considering three components are drawn in Fig. 4. We can note that, for all clusters, there are only small differences between the estimated distributions of the various attributes. We can also observe that, with two components, the estimated distributions are different between the two clusters, with higher probabilities for long durations in cluster 1. With three components, the estimated distributions are really similar for clusters 1 and 3 but are different from cluster 2. Note that the estimated values of all the parameters related to the sojourn times are reported in section C of the on-line supplementary material.

Examining now the dynamic of the tasting process, the most important transitions are represented by using the TDS graphs of Figs 5–7. As in Lecuelle *et al.* (2018), we represent only transition probabilities that are larger than 0.15 and with at least one half of the panellists having actually elicited the corresponding attribute of the product. The estimated values of all the transition probabilities are reported in section C of the on-line supplementary material. The TDS graph drawn from the whole group (Fig. 5, without clustering) suggests

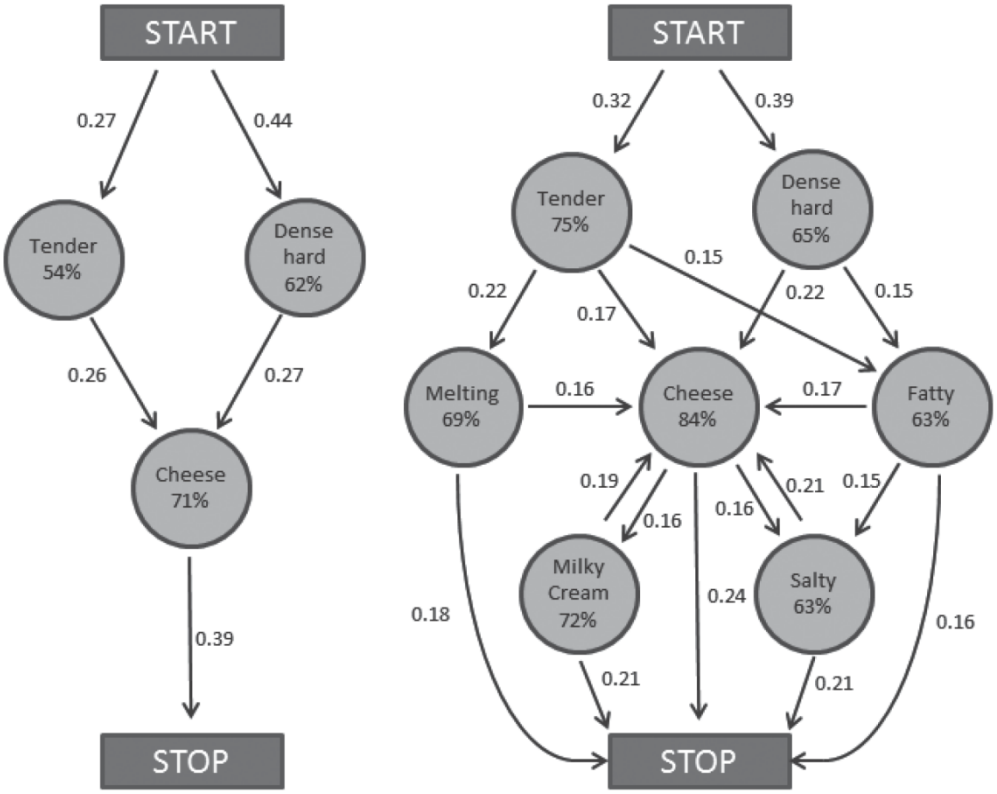


Fig. 6. TDS graphs for the young and low fat Gouda cheese when considering a segmentation into two clusters: (a) first cluster; (b) second cluster

the existence of two different sequences of perception: the first starts with the attribute dense hard, transitions to cheese and ends, whereas the second starts with tender and either goes to cheese and then ends, or goes to melting and then either ends or goes to cheese before ending. It is interesting that the milky cream and salty attributes are represented on the graph, since they were elicited by 54% and 55% of the panellists, but are reached by no arrows, since every transition to them occurred with a probability lower than 0.15. Considering the segmentation into two clusters, Fig. 6 presents the two TDS graphs that are associated with each cluster. Both clusters start with a more or less balanced choice between dense hard and tender. From that point, panellists of cluster 1 move to cheese and then end, whereas panellists of cluster 2 followed a more complex route. Indeed, those on tender can move to melting, cheese or fatty and those on dense hard to the same except to melting. Then their route to the end can be quite complex, using some transitions to milky cream or salty, the two attributes having too small probabilities at the panel level to be reached. The fact that both groups start, like the whole panel, with a choice between opposite attributes tender and dense hard is not satisfying and claims for investigating the decomposition into three clusters. Indeed, clusters 1 and 3 in Fig. 7 (segmentation into three groups) start respectively with tender and dense hard and then follow a different route: cluster 3 goes directly to cheese, whereas

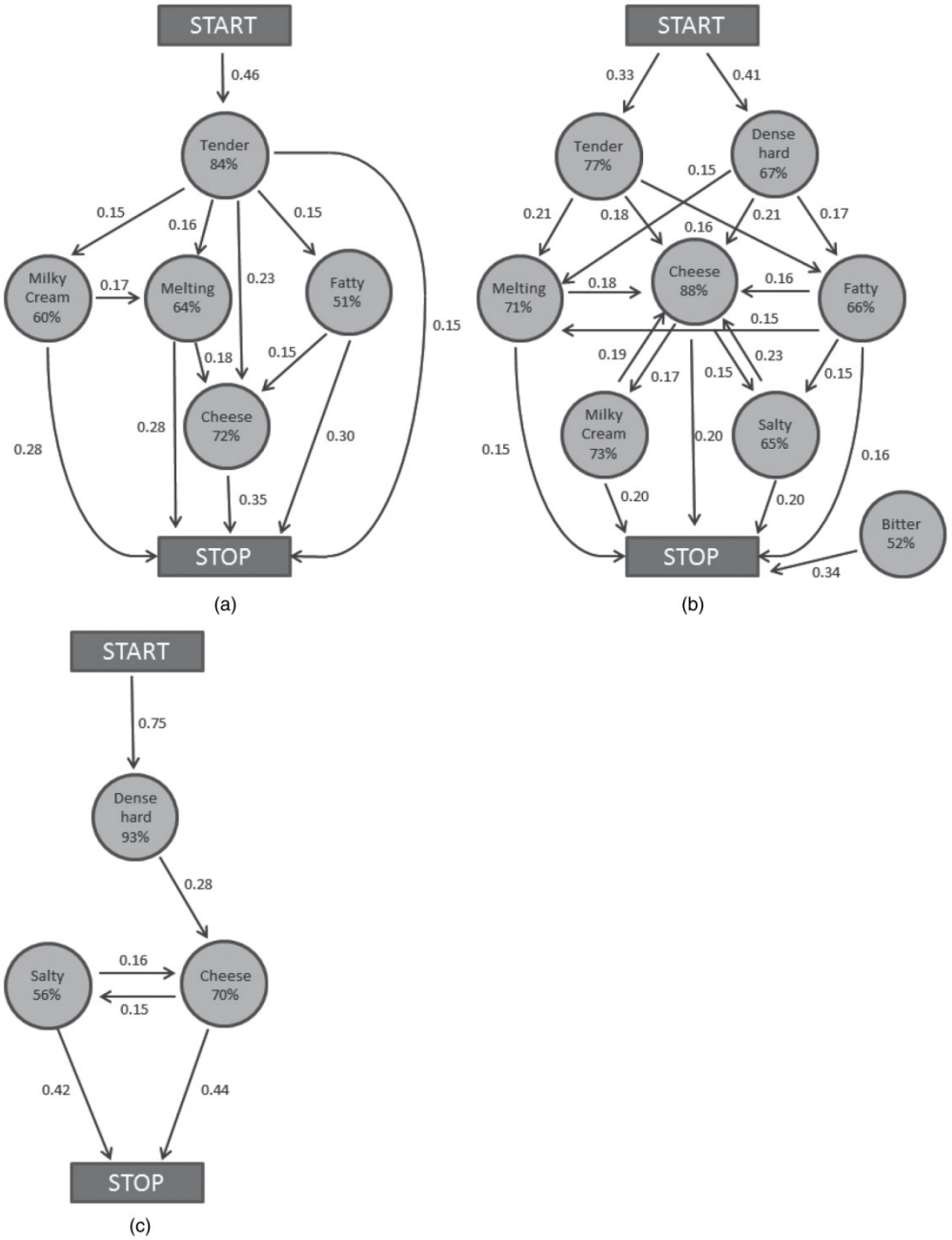


Fig. 7. TDS graphs for the young and low fat Gouda cheese when considering a segmentation into three clusters: transitions for (a) the first cluster, (b) the second cluster and (c) the third cluster

cluster 1 first transitions to melting, milky cream or fatty before reaching the cheese perception. Cluster 1 ends after cheese, whereas cluster 3 can also transition first by salty. Cluster 2 in Fig. 7 is quite similar to cluster 2 in Fig. 6, exhibiting a very complex perception path. Indeed, it can be seen in Table 14 in the on-line supplementary file that panellists in this cluster did an average number of transitions ranging from 5 to 6, whereas panellists in the two other clusters made only an average number of transitions equal to 3. Thus, this cluster is likely to gather panellists with different perceptions, but the algorithm could not split them while still improving the fit. Therefore, it is also possible that this cluster gathers some 'noisy panellists', namely panellists who had not clearly understood the TDS task.

From a sensory perspective, the segmentation enables us to model what is really perceived by the panellists instead of considering a mean panel overview, corresponding to the perception of none of the panellists. The differences observed between clusters can be explained both by real differences of perception and by differences of behaviour with the TDS task. Such mixture models give the opportunity for further investigation on these new questions by for example examining the relationship between perception and other variables such as age, sex or experience.

6. Concluding remarks

This research was motivated by the need for a segmentation method for temporal sensory data. For this we introduced a new mixture of semi-Markov chains which enabled us, thanks to a model-based clustering approach, to gather into homogeneous groups consumers having similar tasting perceptions. A penalized EM algorithm was introduced to estimate the parameters of the semi-Markov chains and the mixture proportions. The evaluation of this estimation method on simulated data shows good performance, improving the segmentation that is obtained by the k-means algorithm, while providing much more information on individual behaviours. The results on real data show interesting progress in TDS data analysis by offering the possibility of exhibiting different perceptions in a panel for a same product. The development of such segmentation approaches opens new perspectives, both for understanding the perception mechanism and for studying how panellists use TDS and understand the TDS protocol.

The models that are presented in this paper may depend on a large number of parameters and so require large samples at hand to be estimated accurately. As usual with unsupervised classification, choosing the number of clusters is a difficult task. The method that is used in this paper relies on information criteria and is not very effective for small samples. The BIC seems to overestimate the model complexity whereas the AIC has a tendency to select models with too large complexity.

From a statistical perspective, this sensory modelling question has given us the opportunity to study a new model for mixtures of qualitative trajectories which may have applications in many fields of science and may be useful for example to study web traffic or qualitative trajectories in sociology or economics. The identifiability issue has been addressed under general conditions, considering parametric families of sojourn time distributions. From a methodological perspective, it also showed that introducing a penalty in the maximization step of the EM algorithm improves the quality of the estimates. However, some further investigations must be done to determine which penalty is the most effective. It would also be of great interest to check rigorously in future work the consistency of such a penalized maximum likelihood approach in the context of mixtures of semi-Markov chains and to study the asymptotic distribution of the estimators. This would enable us to build confidence intervals and to test statistical hypotheses.

Acknowledgements

We thank the two referees and the Associate Editor for their constructive remarks. Guillaume Lecuelle's work is supported by the Bourgogne–Franche Comté Regional Council and the French National Institute for Agricultural Research (Dijon).

References

Allman, E. S., Matias, C. and Rhodes J. A. (2009) Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, **37**, 3099–3132.
Banfield, J. D. and Raftery, A. E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803–821.
Barbu, V. S. and Limnios, N. (2008) *Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications: Their Use in Reliability and DNA Analysis*. New York: Springer Science and Business Media.
Celeux, G. and Durand, J.-B. (2008) Selecting hidden Markov model state number with cross-validated likelihood. *Comput. Statist.*, **23**, 541–564.
Chen, J., Li, S. and Tan, X. (2016) Consistency of the penalized MLE for two-parameter gamma mixture models. *Sci. China Math.*, **59**, 2301–2318.
Delattre, M., Genon-Catalot, V. and Samson, A. (2016) Mixtures of stochastic differential equations with random effects: application to data clustering. *J. Statist. Planng Inf.*, **173**, 109–124.
Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B, **39**, 1–38.
Eddelbuettel, D. and François, R. (2011) Rcpp: seamless R and C++ integration. *J. Statist. Softwr.*, **40**, 1–18.
Franczak, B. C., Browne, R. P., McNicholas, P. D., Castura, J. C. and Findlay, C. J. (2015) A Markov model for temporal dominance of sensations data. *11th Pangborn Symp.*
Frühwirth-Schnatter, S. (2006) *Finite Mixture and Markov Switching Models*. Berlin: Springer.
Frydman, H. (2005) Estimation in the mixture of Markov chains moving with different speeds. *J. Am. Statist. Ass.*, **100**, 1046–1053.
Galmarini, M. V., Visalli, M. and Schlich, P. (2017) Advances in representation and analysis of mono and multi-intake temporal dominance of sensations data. *Food Qual. Pref.*, **56**, 247–255.
Gassiat, E., Cleynen, A. and Robin, S. (2016) Inference in finite space non parametric Hidden Markov Models and applications. *Statist. Comput.*, **26**, 61–71.
Gupta, R., Kumar, R. and Vassilvitskii, S. (2016) On mixtures of Markov chains. In *Advances in Neural Information Processing Systems 29* (eds D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon and R. Garnett), pp. 3441–3449. Red Hook: Curran Associates.
Hartigan, J. A. and Wong, M. A. (1979) Algorithm AS 136: A K-means clustering algorithm. *Appl. Statist.*, **28**, 100–108.
Hort, J., Kemp, S. and Hollowood, T. (2017) *Time-dependent Measures of Perception in Sensory Evaluation*. Chichester: Wiley.
Jaeger, S. R., Hort, J., Porcherot, C., Ares, G., Pecore, S. and MacFie, H. J. H. (2017) Future directions in sensory and consumer science: four perspectives and audience voting. *Food Qual. Pref.*, **56**, 301–309.
Keribin, C. (2000) Consistent estimation of the order of mixture models. *Sankhya* A, **62**, 49–66.
Köster, E. (2009) Diversity in the determinants of food choice: a psychological perspective. *Food Qual. Pref.*, **20**, 70–82.
Lawlor, S. and Rabbat, M. G. (2017) Time-varying mixtures of Markov chains: an application to road traffic modeling. *IEEE Trans. Signl Process.*, **65**, 3152–3167.
Lecuelle, G., Visalli, M., Cardot, H. and Schlich, P. (2018) Modeling temporal dominance of sensations with semi-Markov chains. *Food Qual. Pref.*, **67**, 59–66.
Lévy, P. (1956) Processus semi-Markoviens. In *Proc. Int. Congr. Mathematicians 1954, Amsterdam*, vol. III (eds P. Erven and N. V. Noordhoff), pp. 416–426. Amsterdam: North-Holland.
McLachlan, G. J. and Krishnan, T. (2008) *The EM Algorithm and Extensions*, 2nd edn. New York: Wiley.
McLachlan, G. J. and Peel, D. (2000) *Finite Mixture Models*. New York: Wiley.
McNicholas, P. D. (2016) Model-based clustering. *J. Classificn*, **33**, 331–373.
Meiselman, H. (2013) The future in sensory/consumer research: evolving to a better science. *Food Qual. Pref.*, **27**, 208–214.
Melnykov, V. and Maitra, R. (2010) Finite mixture models and model-based clustering. *Statist. Surv.*, **4**, 80–116.
Neilson, A. (1957) Time-intensity studies. *Drug Cosmtc Indust.*, **80**, 452–453.
Norris, J. R. (1998) *Markov Chains*. Cambridge: Cambridge University Press.
Pamminger, C. and Frühwirth-Schnatter, S. (2010) Model-based clustering of categorical time series. *Baysn Anal.*, **5**, 345–368.

Pineau, N., Schlich, P., Cordelle, S., Mathonniere, C., Issanchou, S., Imbert, A., Rogeaux, M., Etievant, P. and Koster, E. (2009) Temporal dominance of sensations: construction of the TDS curves and comparison with time-intensity. *Food Qual. Pref.*, **20**, 450–455.
Prutkin, J., Duffy, V., Etter, L., Fast, K., Gardner, E., Lucchina, L. A., Snyder, D. J., Tie, K., Weiffenbach, J. and Bartoshuk, L. M. (2000) Genetic variation and inferences about perceived taste intensity in mice and men. *Physiol. Behav.*, **69**, 161–173.
Pyke, R. (1961) Markov renewal processes: definitions and preliminary properties. *Ann. Math. Statist.*, **32**, 1231–1242.
R Core Team (2018) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Schlich, P. (2017) Temporal dominance of sensations (TDS): a new deal for temporal sensory analysis. *Curr. Opin. Food Sci.*, **15**, 38–42.
Smith, W. L. (1955) Regenerative stochastic processes. *Proc. R. Soc. Lond.* A, **232**, 6–31.
Song, Y., Keromytis, A. and Stolfo, S. (2009) Spectrogram: a mixture-of-Markov-chains model for anomaly detection in web traffic. In *Proc. Network and Distributed System Security Symp*.
Teicher, H. (1963) Identifiability of finite mixtures. *Ann. Math. Statist.*, **34**, 1265–1269.
Thomas, A., Chambault, M., Dreyfuss, L., Gilbert, C. C., Hegyi, A., Henneberg, S., Knippertz, A., Kostyra, E., Kreme, S., Silva, A. P. and Schlich, P. (2017) Measuring temporal liking simultaneously to temporal dominance of sensations in several intakes: an application to gouda cheeses in 6 European countries. *Food Res. Int.*, **99**, 426–434.
Titterington, D., Smith, A. and Makov, U. (1985) *Statistical Analysis of Finite Mixture Distributions*. London: Wiley.
Visalli, M., Lange, C., Mallet, L., Cordelle, S. and Schlich, P. (2016) Should I use touchscreen tablets rather than computers and mice in TDS trials? *Food Qual. Pref.*, **52**, 11–16.
Yakowitz, S. and Spragins, J. (1968) On the identifiability of finite mixtures. *Ann. Math. Statist.*, **39**, 209–214.

Supporting information
Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material: Estimating finite mixtures of semi-Markov chains: an application to the segmentation of temporal sensory data'.

Supplementary material : Estimating finite mixtures of semi-Markov chains: an application to the segmentation of temporal sensory data

Hervé Cardot

Institut de Mathématiques de Bourgogne, UMR 5584 CNRS, Université Bourgogne Franche-Comté, F-21000 Dijon, France

E-mail: herve.cardot@u-bourgogne.fr

Guillaume Lecuelle, Pascal Schlich, Michel Visalli

Centre des Sciences du Goût et de l'Alimentation, AgroSup Dijon, CNRS, INRA, Université Bourgogne Franche-Comté, F-21000 Dijon, France

E-mail: guillaume.lecuelle@inra.fr, pascal.schlich@inra.fr, michel.visalli@inra.fr

A. Proofs

Proof of Proposition 2.1.
The proof is immediate. Consider the unobserved latent class variable Z, taking values in $\{1,\ldots,G\}$ and satisfying $\Pr[Z=g]=\pi_g$ for $g=1,\ldots,G$. The law of $(J_p^\pi, X_p^\pi)_{n\geq 1}$ given $Z=g$ can be expressed as Law $\left(\boldsymbol{\alpha}^g, \mathbf{P}^g, \Phi_{\ell j}^g, \ell, j \in \mathcal{S}\right)$. Thus,

$$\alpha_j^\pi = \Pr\left[J_1^\pi = j\right] = \sum_{g=1}^{G} \Pr\left[J_1^\pi = j | Z=g\right] \Pr[Z=g] = \sum_{g=1}^{G} \pi_g \alpha_j^g.$$

We also clearly have that $(J_p^\pi)_{p\geq 1}$ is a Markov chain, with transition probabilities,

$$\mathbf{P}_{\ell j}^\pi = \Pr\left[J_{p+1}^\pi = j | J_p^\pi = \ell\right] = \sum_{g=1}^{G} \Pr\left[J_{p+1}^\pi = j | J_p^\pi = \ell, Z=g\right] \Pr[Z=g] = \sum_{g=1}^{G} \pi_g \mathbf{P}_{\ell j}^g,$$

and, for $t \in T$,

$$\Phi_{\ell j}^\pi(t) = \sum_{g=1}^{G} \Pr\left[X_p \leq t \mid J_{p+1}=j, J_p=\ell, Z=g\right] \Pr[Z=g] = \sum_{g=1}^{G} \pi_g \Phi_{\ell j}^g(t).$$

□

Proof of Proposition 2.2.
We have, for each mixture component g, $\Pr\left[X_1 \leq t, J_1=\ell, J_2=j | Z=g\right] = \alpha_\ell^g \mathbf{P}_{\ell j}^g \Phi(t, \mathbf{\Gamma}_{\ell j}^g)$, for $t \geq 0$, $\ell \in \mathcal{S}$ and $j \neq \ell$. We introduce the $D(D-1)$ dimension (functional) vector,

for $t \geq 0$,

$$\begin{aligned}\mathbf{F}_g(t) &= \left(\alpha_\ell^g \mathbf{P}_{\ell j}^g \Phi(t, \mathbf{\Gamma}_{\ell j}^g), \ell = 1, 2, \ldots, D, j \neq \ell\right) \\ &= \left(\alpha_1^g \mathbf{P}_{12}^g \Phi(t, \mathbf{\Gamma}_{12}^g), \cdots, \alpha_1^g \mathbf{P}_{1D}^g \Phi(t, \mathbf{\Gamma}_{1D}^g), \cdots, \alpha_D^g \mathbf{P}_{D(D-1)}^g \Phi(t, \mathbf{\Gamma}_{D(D-1)}^g)\right)\end{aligned}$$

With assumption ***(H1)****, all the coefficients* $\alpha_\ell^g \mathbf{P}_{\ell j}^g$ *are strictly positive. Since the family of sojourn time distributions is identifiable, we can deduce, with assumption* ***(H2)*** *that* $\pi_1 \mathbf{F}_1(t), \pi_2 \mathbf{F}_2(t), \ldots, \pi_g \mathbf{F}_G(t)$ *are* G *linearly independent vectors of functions. Considering the characterization of identifiability established in Yakowitz and Spragins (1968), this implies that, up to label swapping, there is a unique way of writing the mixture distribution* $\mathbf{F}^\pi(t) = \Pr\left[X_1^\pi \leq t, J_1^\pi = \ell, J_2^\pi = j\right]$,

$$\mathbf{F}^\pi(t) = \sum_{g=1}^{G} \pi_g \mathbf{F}_g(t), \quad t \geq 0.$$

For $g = 1, \ldots, G$, *denote by* $\mathbf{u}^g(t) = \pi_g \mathbf{F}_g(t)$ *the* $D(D-1)$ *dimensional vector of functions that can be identified from the knowledge of* $\mathbf{F}^\pi(t)$. *Since, by assumption* ***(H1)****,* $\pi_g \alpha_\ell^g \mathbf{P}_{\ell j}^g > 0$, *we can determine the value of the set of parameters* $\mathbf{\Gamma}_{\ell j}^g, \ell = 1, 2, \ldots, D, j \neq \ell$ *by comparing* $\pi_g \alpha_\ell^g \mathbf{P}_{\ell j}^g \Phi(t, \mathbf{\Gamma}_{\ell j}^g)$ *with* $\mathbf{u}_{\ell j}^g(t)$, *and we can write each component of* $\mathbf{u}^g(t)$ *as follows* $u_{\ell j}^g(t) = \gamma_{\ell j}^g \Phi(t, \mathbf{\Gamma}_{\ell j}^g)$.

We prove now that the mixture probability π_g, *the initialization probabilities* $\alpha_1^g, \ldots, \alpha_D^g$ *and the transition probabilities* $\mathbf{P}_{\ell j}^g$ *are uniquely determined when the set of coefficients* $\{\gamma_{\ell j}^g, \ell \in \mathcal{S}, j \neq \ell\}$ *is known. Since* $\gamma_{\ell j}^g = \pi_g \alpha_\ell^g \mathbf{P}_{\ell j}^g$, *we first note that*

$$\begin{aligned}\sum_{\ell \in \mathcal{S}} \sum_{j \neq \ell} \gamma_{\ell j}^g &= \pi_g \sum_{\ell \in \mathcal{S}} \sum_{j \neq \ell} \alpha_\ell^g \mathbf{P}_{\ell j}^g \\ &= \pi_g\end{aligned}$$

because $\sum_{j \neq \ell} \mathbf{P}_{\ell j}^g = 1$ *and* $\sum_{\ell \in \mathcal{S}} \alpha_\ell^g = 1$. *Using the same trick again, we get that, for each* $\ell \in \mathcal{S}$,

$$\begin{aligned}\frac{1}{\pi_g} \sum_{j \neq \ell} \gamma_{\ell j}^g &= \sum_{j \neq \ell} \alpha_\ell^g \mathbf{P}_{\ell j}^g \\ &= \alpha_\ell^g.\end{aligned}$$

Finally, we deduce the values of the transition probabilities, for each $\ell \in \mathcal{S}$ *and* $j \neq \ell$,

$$\mathbf{P}_{\ell j}^g = \frac{1}{\pi_g \alpha_\ell^g} \gamma_{\ell j}^g$$

and the proof is complete. □

Table 7. Estimated initial probabilities for the 3 chocolates.

Chocolate	Astringent	Bitter	Cocoa	Crunchy	Dry	Fatty	Melting	Sour	Sweet	Sticky
70	.00	.00	.00	.81	.03	.00	.03	.00	.11	.03
70 Sweet	.00	.00	.00	.75	.03	.00	.11	.00	.06	.06
90	.00	.03	.03	.83	.08	.00	.00	.03	.00	.00

B. Description of the semi-Markov chains used for the simulation study

We report, in this supplementary Section, additional information about the law of the semi-Markov chains used in the Simulation study. These laws are based on estimations made with real tasting experiences for three different chocolates. Table 7 gives the estimated initialisation probabilities whereas the transition probabilities are presented in Table 8. The parameters of the gamma distributed sojourn times are presented in Table 9. Note that, in order to be able to control the number of observed transitions, we do not consider any absorbing state "STOP".

Table 8. Estimated transition probabilities for the 3 chocolates.

	Chocolate with 70% of cocoa									
	Astringent	Bitter	Cocoa	Crunchy	Dry	Fatty	Melting	Sour	Sweet	Sticky
Astringent	.00	.33	.00	.00	.00	.00	.00	.33	.33	.00
Bitter	.06	.00	.25	.00	.00	.06	.13	.06	.31	.11
Cocoa	.00	.15	.00	.00	.15	.09	.21	.06	.27	.06
Crunchy	.00	.07	.40	.00	.17	.00	.03	.00	.27	.07
Dry	.00	.15	.15	.00	.00	.15	.00	.15	.38	.00
Fatty	.00	.13	.38	.00	.00	.00	.38	.00	.13	.00
Melting	.00	.00	.21	.00	.00	.00	.00	.11	.58	.11
Sour	.09	.36	.27	.00	.00	.00	.18	.00	.00	.09
Sweet	.03	.03	.28	.05	.03	.08	.28	.10	.00	.13
Sticky	.00	.00	.00	.10	.20	.00	.20	.10	.40	.00
	Chocolate with 70% of cocoa sweet									
	Astringent	Bitter	Cocoa	Crunchy	Dry	Fatty	Melting	Sour	Sweet	Sticky
Astringent	.00	.00	.00	.00	.00	.00	1.00	.00	.00	.00
Bitter	.00	.00	.00	.00	.00	1.00	.00	.00	.00	.00
Cocoa	.03	.03	.00	.03	.00	.23	.34	.00	.34	.00
Crunchy	.00	.00	.23	.00	.16	.03	.10	.00	.45	.03
Dry	.00	.00	.29	.14	.00	.29	.00	.00	.14	.14
Fatty	.05	.00	.36	.00	.05	.00	.27	.00	.23	.05
Melting	.00	.00	.25	.04	.00	.18	.00	.00	.54	.00
Sour	.00	.00	.00	.00	.00	.00	1.00	.00	.00	.00
Sweet	.00	.00	.46	.02	.00	.17	.22	.02	.00	.10
Sticky	.00	.00	.12	.00	.00	.38	.12	.00	.38	.00
	Chocolate with 90% of cocoa									
	Astringent	Bitter	Cocoa	Crunchy	Dry	Fatty	Melting	Sour	Sweet	Sticky
Astringent	.00	.53	.00	.00	.00	.18	.00	.06	.00	.24
Bitter	.19	.00	.30	.00	.11	.14	.07	.04	.09	.07
Cocoa	.00	.48	.00	.03	.10	.07	.17	.00	.03	.10
Crunchy	.06	.29	.13	.00	.32	.13	.03	.00	.00	.03
Dry	.23	.55	.18	.00	.00	.00	.00	.05	.00	.00
Fatty	.17	.44	.06	.00	.00	.00	.22	.00	.00	.11
Melting	.14	.57	.14	.00	.00	.07	.00	.00	.00	.07
Sour	.20	.60	.00	.00	.00	.00	.00	.00	.00	.20
Sweet	.00	.17	.50	.00	.00	.17	.17	.00	.00	.00
Sticky	.25	.50	.00	.08	.00	.08	.08	.00	.00	.00

Table 9. Estimated parameters of the gamma distributions for the 3 chocolates.

	Chocolate with 70% of cocoa									
	Astringent	Bitter	Cocoa	Crunchy	Dry	Fatty	Melting	Sour	Sweet	Sticky
a	1.90	1.38	1.53	2.83	2.12	1.78	1.72	1.18	1.73	3.45
λ	0.29	0.21	0.21	0.41	0.38	0.21	0.35	0.15	0.26	0.77
	Chocolate with 70% of cocoa sweet									
	Astringent	Bitter	Cocoa	Crunchy	Dry	Fatty	Melting	Sour	Sweet	Sticky
a	1.76	1.69	1.30	2.04	1.87	1.50	1.51	1.69	2.28	3.51
λ	0.14	0.25	0.20	0.32	0.33	0.22	0.22	0.25	0.31	0.62
	Chocolate with 90% of cocoa									
	Astringent	Bitter	Cocoa	Crunchy	Dry	Fatty	Melting	Sour	Sweet	Sticky
a	1.86	1.52	1.67	2.40	2.05	3.29	2.88	1.73	3.86	3.70
λ	0.20	0.20	0.27	0.50	0.27	0.81	0.70	0.21	1.45	0.63

Table 10. Gouda cheese example: estimated transition probabilities with 2 clusters.

	Cluster 1										
	Bitter	Cheese	Dense hard	Fatty	Melting	Milky cream	Salty	Sharp	Sour	Tender	STOP
Bitter	0.00	0.08	0.05	0.05	0.05	0.04	0.08	0.04	0.06	0.03	0.53
Cheese	0.07	0.00	0.05	0.07	0.09	0.08	0.13	0.04	0.04	0.05	0.39
Dense	0.09	0.26	0.00	0.07	0.06	0.09	0.11	0.05	0.08	0.04	0.13
Fatty	0.07	0.16	0.04	0.00	0.11	0.10	0.07	0.05	0.06	0.07	0.27
Melting	0.04	0.20	0.00	0.09	0.00	0.14	0.13	0.02	0.06	0.06	0.26
Milky	0.04	0.20	0.01	0.07	0.13	0.00	0.11	0.02	0.04	0.08	0.30
Salty	0.09	0.14	0.03	0.07	0.05	0.05	0.00	0.05	0.07	0.02	0.42
Sharp	0.09	0.13	0.06	0.09	0.05	0.01	0.13	0.00	0.07	0.03	0.34
Sour	0.12	0.09	0.06	0.04	0.05	0.03	0.11	0.06	0.00	0.03	0.41
Tender	0.04	0.26	0.02	0.15	0.12	0.13	0.09	0.03	0.03	0.00	0.13
	Cluster 2										
	Bitter	Cheese	Dense hard	Fatty	Melting	Milky cream	Salty	Sharp	Sour	Tender	STOP
Bitter	0.00	0.13	0.02	0.08	0.07	0.10	0.12	0.06	0.06	0.03	0.34
Cheese	0.10	0.00	0.02	0.06	0.10	0.16	0.16	0.04	0.06	0.06	0.24
Dense	0.07	0.23	0.00	0.15	0.12	0.08	0.10	0.07	0.04	0.10	0.04
Fatty	0.05	0.17	0.04	0.00	0.14	0.13	0.14	0.03	0.05	0.09	0.16
Melting	0.09	0.17	0.04	0.07	0.00	0.13	0.11	0.05	0.03	0.12	0.18
Milky	0.07	0.19	0.03	0.09	0.11	0.00	0.13	0.03	0.06	0.08	0.21
Salty	0.13	0.21	0.03	0.06	0.08	0.09	0.00	0.05	0.09	0.05	0.22
Sharp	0.14	0.12	0.05	0.09	0.11	0.09	0.14	0.00	0.04	0.05	0.17
Sour	0.11	0.15	0.01	0.04	0.07	0.11	0.13	0.06	0.00	0.04	0.28
Tender	0.03	0.17	0.05	0.15	0.22	0.14	0.06	0.04	0.05	0.00	0.10

C. Additional information on the Gouda cheese example

We report in Table 10 (resp. in Table 11) the estimated values of the transition probabilities (resp. sojourn time distributions), for the Gouda cheese, when considering a segmentation into two clusters. Estimated parameters for the segmentation into three clusters are reported in Table 12 and Table 13. The average number of transitions are reported in Table 14 for the case of no segmentation, as well as within each cluster when clustering has been done.

References

Yakowitz, S. and J. Spragins (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics 39*, 209–214.

Table 11. Gouda cheese example: estimated gamma distributions with 2 clusters.

	Cluster 1									
	Bitter	Cheese	Dense hard	Fatty	Melting	Milky cream	Salty	Sharp	Sour	Tender
a	1.91	2.30	2.28	2.19	2.53	2.57	2.24	2.19	2.17	2.57
λ	0.20	0.25	0.24	0.25	0.29	0.32	0.25	0.23	0.26	0.30
	Cluster 2									
	Bitter	Cheese	Dense hard	Fatty	Melting	Milky cream	Salty	Sharp	Sour	Tender
a	2.44	2.17	3.17	2.77	2.36	2.23	2.46	3.42	3.08	2.59
λ	0.46	0.43	0.61	0.63	0.46	0.45	0.52	0.73	0.65	0.59

Table 12. Gouda cheese example: estimated transition probabilities with 3 clusters.

	Cluster 1										
	Bitter	Cheese	Dense hard	Fatty	Melting	Milky cream	Salty	Sharp	Sour	Tender	STOP
Bitter	0.00	0.07	0.02	0.06	0.10	0.11	0.00	0.05	0.04	0.08	0.47
Cheese	0.05	0.00	0.01	0.10	0.13	0.09	0.13	0.02	0.01	0.10	0.35
Dense	0.05	0.25	0.00	0.12	0.12	0.09	0.05	0.04	0.01	0.12	0.15
Fatty	0.04	0.15	0.00	0.00	0.13	0.12	0.07	0.05	0.05	0.10	0.29
Melting	0.05	0.19	0.01	0.10	0.00	0.14	0.08	0.03	0.04	0.09	0.28
Milky	0.01	0.17	0.01	0.07	0.17	0.00	0.10	0.03	0.04	0.11	0.28
Salty	0.07	0.11	0.03	0.07	0.09	0.10	0.00	0.03	0.05	0.06	0.39
Sharp	0.06	0.17	0.00	0.11	0.10	0.04	0.13	0.00	0.02	0.06	0.32
Sour	0.11	0.04	0.00	0.10	0.09	0.05	0.05	0.01	0.00	0.04	0.50
Tender	0.02	0.23	0.01	0.15	0.16	0.16	0.07	0.03	0.02	0.00	0.15
	Cluster 2										
	Bitter	Cheese	Dense hard	Fatty	Melting	Milky cream	Salty	Sharp	Sour	Tender	STOP
Bitter	0.00	0.13	0.01	0.08	0.07	0.10	0.12	0.06	0.06	0.02	0.34
Cheese	0.10	0.00	0.02	0.06	0.11	0.17	0.16	0.04	0.07	0.06	0.21
Dense	0.05	0.22	0.00	0.16	0.14	0.08	0.10	0.07	0.05	0.11	0.02
Fatty	0.05	0.17	0.04	0.00	0.14	0.13	0.15	0.03	0.05	0.08	0.16
Melting	0.09	0.18	0.04	0.06	0.00	0.13	0.13	0.05	0.04	0.12	0.15
Milky	0.08	0.19	0.02	0.10	0.10	0.00	0.13	0.03	0.07	0.08	0.20
Salty	0.14	0.23	0.02	0.07	0.07	0.09	0.00	0.05	0.09	0.04	0.21
Sharp	0.15	0.11	0.05	0.10	0.11	0.08	0.11	0.00	0.05	0.05	0.18
Sour	0.12	0.15	0.01	0.04	0.07	0.11	0.14	0.06	0.00	0.04	0.25
Tender	0.03	0.18	0.05	0.16	0.21	0.12	0.07	0.04	0.05	0.00	0.09
	Cluster 3										
	Bitter	Cheese	Dense hard	Fatty	Melting	Milky cream	Salty	Sharp	Sour	Tender	STOP
Bitter	0.00	0.09	0.07	0.04	0.02	0.00	0.15	0.03	0.07	0.00	0.52
Cheese	0.09	0.00	0.09	0.04	0.03	0.07	0.14	0.05	0.06	0.00	0.43
Dense	0.11	0.27	0.00	0.06	0.04	0.09	0.12	0.06	0.09	0.02	0.13
Fatty	0.12	0.21	0.14	0.00	0.08	0.06	0.08	0.02	0.09	0.03	0.19
Melting	0.04	0.17	0.00	0.04	0.00	0.13	0.27	0.00	0.11	0.03	0.22
Milky	0.08	0.27	0.03	0.04	0.06	0.00	0.12	0.00	0.04	0.01	0.35
Salty	0.11	0.16	0.05	0.06	0.03	0.02	0.00	0.07	0.08	0.01	0.42
Sharp	0.10	0.10	0.11	0.06	0.02	0.02	0.18	0.00	0.10	0.01	0.30
Sour	0.10	0.13	0.09	0.00	0.02	0.03	0.14	0.08	0.00	0.02	0.39
Tender	0.13	0.27	0.10	0.08	0.00	0.09	0.16	0.00	0.09	0.00	0.07

Table 13. Gouda cheese example: estimated gamma distributions with 3 clusters.

	Cluster 1									
	Bitter	Cheese	Dense hard	Fatty	Melting	Milky cream	Salty	Sharp	Sour	Tender
a	2.80	2.44	2.76	2.41	2.41	2.54	3.51	2.25	2.67	2.67
λ	0.36	0.28	0.32	0.28	0.29	0.32	0.44	0.26	0.33	0.33
	Cluster 2									
	Bitter	Cheese	Dense hard	Fatty	Melting	Milky cream	Salty	Sharp	Sour	Tender
a	2.43	2.22	3.22	2.91	2.40	2.19	2.52	3.35	3.17	2.59
λ	0.45	0.46	0.63	0.69	0.49	0.47	0.55	0.74	0.68	0.62
	Cluster 3									
	Bitter	Cheese	Dense hard	Fatty	Melting	Milky cream	Salty	Sharp	Sour	Tender
a	1.55	2.07	2.11	1.67	2.75	2.87	1.76	2.24	1.96	1.58
λ	0.16	0.23	0.23	0.21	0.32	0.39	0.20	0.23	0.24	0.20

Table 14. Gouda cheese example: mean number of transitions for an individual sequence for the different scenarios (no clustering, one cluster and two clusters).

	Cluster 1	Cluster 2	Cluster 3
No segmentation	4.13		
2 clusters	3.30	5.37	
3 clusters	3.46	5.79	3.26