Validation et enrichissement interactifs d'un apprentissage automatique des paramètres d'un réseau bayésien dynamique appliqué aux procédés alimentaires

Bruno Pinaud, Cédric Baudrit, Mariette Sicard, Pierre-Henri Wuillemin, Nathalie Perrot

Validation et enrichissement interactifs d'un apprentissage automatique des paramètres d'un réseau bayésien dynamique appliqué aux procédés alimentaires

Bruno Pinaud* — **Cédric Baudrit*** — **Mariette Sicard***
Pierre-Henri Wuillemin** — **Nathalie Perrot***

* *UMR782 LGMPA, INRA/AgroParisTech, 78850 Thiverval-Grignon*
{bruno.pinaud, cedric.baudrit, mariette.sicard, nathalie.perrot}@grignon.inra.fr

** *Laboratoire d'Informatique de Paris VI (LIP6 – CNRS UMR7606), 75016 Paris*
pierre-henri.wuillemin@lip6.fr

RÉSUMÉ. *La modélisation des dynamiques des procédés de transformation dans l'industrie agro-alimentaire est un enjeu important pour mieux maîtriser la qualité des produits. L'utilisation de réseaux bayésiens dynamiques est alors une approche candidate pertinente. Malgré la réduction notable du nombre de paramètres, l'information disponible pour réaliser leur apprentissage est lacunaire et incomplète. Les approches classiques d'apprentissage des paramètres atteignent alors leurs limites. Nous proposons, consécutivement à un apprentissage automatique classique des paramètres, d'effectuer de façon interactive et visuelle avec un expert une optimisation a posteriori du contenu des tables de probabilités afin d'améliorer la cohérence du modèle. Nous appliquons nos travaux sur l'affinage de fromages de type camembert.*

ABSTRACT. *Modelling the dynamics of a food transformation process is a major challenge for a good management of the quality of the products. A good candidate approach is then Dynamic Bayesian Network. Despite an important reduction of the number of parameters required, the available information to learn them is tainted with uncertainty and scarce. So, the classical parameters learning methods reach their limits. To overcome these difficulties and after a classical parameters learning, we propose to optimize these parameters interactively with an expert via a specific visualization. Thanks to an application on the ripening of Camembert-type cheese, this new step should help to improve the quality of the model.*

MOTS-CLÉS : *Réseau bayésien dynamique, apprentissage de paramètres, procédés alimentaires, systèmes complexes, intégration de connaissances*

KEYWORDS: *Dynamic bayesian network, parameters learning, food processing, complex system, knowledge integration*

1re soumission à 4e *Journées francophones sur les réseaux bayésiens*, le 29 février 2008

1. Introduction

Les enjeux de compétitivité auxquels sont confrontés les industries agro-alimentaires portent sur la qualité des produits et leur constance et par conséquent la maîtrise des procédés de transformation. Les développements actuels pour atteindre ces objectifs sont la modélisation d'outils d'aide à la décision, afin de reconstruire la dynamique des procédés de transformations et ainsi expliquer le comportement global du produit. Pour ce faire, le formalisme mathématique des réseaux bayésiens dynamiques (RBD) (Murphy, 2002) nous semble pertinent et adapté. Ils permettent de décrire un système qui change et évolue dans le temps en utilisant le formalisme des Réseaux Bayésiens (RB) (Pearl, 1988; Naïm *et al.*, 2007). Les RBD peuvent être vus comme une séquence temporelle qui répète le même RB en fonction de la longueur de la séquence d'observations (Fig. 1). Les arcs entre deux pas de temps peuvent être vus comme la persistance du phénomène dans le temps alors que les autres sont des relations entre variables comme dans les RB classiques.

Compte tenu de la complexité des phénomènes micro-biologiques et physico-chimiques présents au cours des processus de transformation, les connaissances disponibles pour décrire les réactions sont exprimées sous différents formats et à différentes échelles : macroscopique (par ex. couleurs, odeurs, texture), microscopique (dénombrement, concentration de micro-organismes) et aussi en utilisant des lois de transfert physique bien connues. Ces connaissances sont obtenues par des recueils auprès d'experts (technologues, scientifiques, opérateurs) ou bien du recueil de données expérimentales (analyses sensorielles, analyses microbiologiques, ...). Un enjeu important pour effectuer la reconstruction est alors la compréhension et la modélisation de l'ensemble du réseau d'interactions formé par les différents morceaux de connaissances capitalisées à tous les niveaux du processus (micro et macro) et par différents moyens. Malgré les importantes recherches menées depuis les 20 dernières années, les connaissances sur les réactions qui surviennent tout au long du processus de transformation sont encore incomplètes, fragmentées et parfois pauvres (Babin *et al.*, 2005; Perrot *et al.*, 2006; Picque *et al.*, 2006; Trystram *et al.*, 1997). De plus, les coûts humains, techniques et financiers importants pour acquérir les connaissances font que celles qui sont disponibles restent extrêmement lacunaires et en quantité limitée par rapport à la complexité importante des processus à modéliser.

Bien que les RBD permettent de combiner et d'intégrer de la connaissance hétérogène (multi-échelles), leur construction soulève généralement des difficultés car elle requiert des bases de données relativement riches. Le principal verrou scientifique et technologique dans la reconstruction des dynamiques est alors notre capacité à prendre en compte l'incertitude et l'incomplétude de la connaissance du procédé. La figure 1 est un RBD sur 4 pas de temps qui représente l'évolution du micro-organisme *K.marxianus* (Km) et sa consommation de lactose qui sont influencées par la température ambiante et la conséquence au niveau macroscopique sur l'évolution des odeurs durant l'affinage d'un fromage de type camembert. Cette modélisation nécessite deux types d'experts : les experts du domaine (opérateurs, technologues) qui savent décrire les phénomènes d'un point de vue sensoriel (vision macroscopique) et les experts

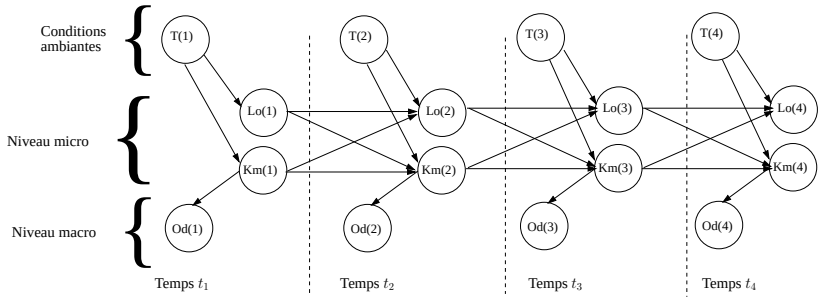


Figure 1. *Représentation simplifiée par un RBD de la dynamique de la croissance de K.marxianus (Km) avec sa consommation de lactose (Lo) qui sont influencées par la température ambiante (T) et la conséquence sur l'évolution des odeurs (Od) sur 4 pas de temps.*

scientifiques qui décrivent les différents aspects microbiologiques. Des relations statiques peuvent alors être formalisées (par ex. Km a un impact sur l'odeur) pour définir la structure du réseau. En revanche la dynamique de ces relations, leurs évolutions, et les interactions entre ces dynamiques qui sont contenues dans les tables de probabilités conditionnelles (CPD) du réseau, sont nettement plus difficiles voire impossibles à formaliser par les experts. Il est alors nécessaire de s'appuyer sur des données d'observations pour réaliser l'apprentissage des paramètres.

En plus de ces problèmes de formalisation, la modélisation d'un système complexe tel qu'un procédé alimentaire, nécessite régulièrement un nombre de paramètres très important qui dépasse bien souvent les capacités humaines de traitement (plusieurs milliers). De plus, la connaissance des experts, comme les observations disponibles sur le système sont incomplètes, imprécises et floues. Les approches classiques d'apprentissage de paramètres avec des données incomplètes qui consistent à utiliser les connaissances *a priori* des experts pour aider au calcul (par ex. sous forme d'*a priori* de Dirichlet ou même d'élicitation de probabilités) ne peuvent donc pas être mises en œuvre. Pour mettre au point un modèle fiable et cohérent avec le processus physique modélisé, nous proposons d'effectuer une validation *a posteriori* avec les experts d'un apprentissage automatique des paramètres réalisé sur les bases de données disponibles. Les différents experts vont ainsi pouvoir valider, contredire, corriger ou bien enrichir le contenu des CPD par l'intermédiaire d'une représentation visuelle interactive adaptée. L'intérêt de l'approche visuelle est qu'elle permet de masquer la complexité sous-jacente des CPD et évite la manipulation directe des probabilités par les experts. Plus généralement, l'intérêt des approches visuelles qui sont largement utilisés dans de nombreux domaines n'est plus à démontrer (Card *et al.*, 1999).

A cet effet, nous présentons dans cet article le développement en cours d'un outil appliqué à la modélisation des cinétiques d'affinage de fromages de type camembert. Le fromage est un écosystème difficile à appréhender dans sa globalité et l'affinage

est un processus complexe à maîtriser. Les différentes réactions qui se produisent sont pour la plupart, inter-dépendantes, et pour certaines mal connues ou parfois même inconnues.

Le reste de cet article est organisé de la façon suivante : le paragraphe 2 présente succinctement la dynamique des réactions principales au cours de l'affinage, le paragraphe 3 présente le RBD mis au point pour reconstruire les cinétiques de l'affinage, le paragraphe 4 présente les développements actuels autour de la représentation visuelle des CPD et la validation de leur contenu. Les perspectives à court terme sont présentées dans le paragraphe 5 et nous concluons dans le paragraphe 6.

2. La dynamique simplifiée de l'affinage du camembert

Après que certaines connaissances sur le processus d'affinage aient été acquises grâce à une phase de recueil bibliographique, et d'autres après consultation d'experts du domaine, sur les 41 jours d'affinage, 4 phases sont perçues par les experts (vision macroscopique du phénomène). Ces 4 phases correspondent alors à certains changements majeurs des caractéristiques microscopiques. Ces travaux sont effectués sur un écosystème simplifié (seulement 4 souches de bactéries) et seules les réactions dominantes sont étudiées pour le moment.

2.1. *Vision macroscopique*

Au début de l'affinage, le fromage passe d'un état très humide à un état plus sec qui correspond à la première phase du procédé. La deuxième phase est caractérisée par une couverture en mycélium (moisissure) qui se développe à la surface du fromage. Ensuite, en phase 3 le mycélium est implanté sur toute la surface, la sous-couche crémeuse (SC) entre le cœur et la surface du fromage commence à se développer et des odeurs de type camembert (ou soufrée) se dégage. Enfin, la phase 4 correspond à l'évolution de SC et des odeurs de type ammoniac. Ces phases, définies d'un point de vue sensoriel, correspondent aux développements et aux activités enzymatiques de micro-organismes plus ou moins dominants dans chaque phase.

2.2. *Vision microscopique*

Les micro-organismes présents dans nos fromages modèles sont : *K.marxianus* (*Km*, levure) qui métabolise principalement le lactose (*Lo*, "sucre du lait") afin d'empêcher le *pH* de descendre à des niveaux trop bas pour éviter des défauts ; *Geotrichum Candidum* (*Gc*, levure) qui consomme le lactate (*La*, forme d'acide lactique) ; *Brevibacterium aurantiacum* (*Ba*, bactéries) qui a un rôle important dans la couleur et le développement des odeurs. Nous avons aussi pris en compte la mesure du pH en surface qui est un bon indicateur de l'activité microbiologique. Nous n'avons pas retenu dans notre modèle *Penicillium camemberti* (*Pc*, moisissures) qui produit la couverture

blanche en surface des camemberts car son implantation est difficile à mesurer avec rigueur. Pour des explications plus précises et complètes, nous invitons le lecteur à se référer à Leclercq-Perlat *et al.* (2004) et à Leclercq-Perlat *et al.* (2006).

La première phase du processus d'affinage est caractérisée par la croissance importante de *Km*, donc la diminution rapide de *Lo* ainsi que la croissance faible de *La* liée à l'activité des bactéries lactiques (LAB). Ensuite, lors de la deuxième phase, *Gc* croit fortement ce qui implique une concentration de *La* qui tend rapidement vers zéro et l'augmentation du *pH*, la croissance de *Km* ralentit puisqu'il ne reste plus beaucoup de *Lo*. Dans la troisième phase, la concentration de *Gc* se stabilise (plus beaucoup de *La* restant), le pH est stable et a atteint un niveau suffisant pour que *Ba* se développe, *SC* commence à apparaître. Enfin, dans la quatrième phase, on assiste une diminution de la concentration de *Km*, *Ba* et *Gc* sont stables, *SC* continue d'évoluer.

3. Modélisation de la dynamique par un RBD

La modélisation des cinétiques d'affinage par un RBD a été effectuée selon l'hypothèse simplificatrice mais réaliste que le processus d'affinage est markovien de 1^{er} ordre. Cela signifie que l'état des variables à $j+1$ jour ne dépend que de l'état au jour j. Cette hypothèse implique que seulement 2 pas de temps sont nécessaires dans le réseau (j et $j+1$) et que sa structure est fixe.

Cette structure a tout d'abord été définie avec les experts. Ensuite, un premier apprentissage automatique des paramètres avec des données issues de différents essais de fabrications de camemberts au sein du laboratoire a été effectué. Nous avons aussi réalisé une première validation du modèle pour mesurer l'impact du manque de données.

3.1. *Définition de la structure du réseau*

À partir d'échanges avec les experts et de recueils bibliographiques, nous avons proposé une structure simplifiée du modèle (Fig. 2) qui met en relation les variables microscopiques et macroscopiques.

Cette structure doit être considérée comme une première approximation qui représente les comportements dynamiques de chacune des variables, non pas d'un point de vue microscopique mais macroscopique : chaque dynamique est déterminée uniquement par la phase dans laquelle se trouve le réseau au jour j. Ce réseau est une façon d'agréger l'information pour la réduire car à chaque pas de temps toutes les variables ne sont pas prises en compte. Nous pouvons ainsi valider les dynamiques sur des jeux de données temporelles où la phase est observée à chaque pas de temps. A terme, le nœud $phase(j)$ sera supprimé pour définir un RBD décrivant la dynamique uniquement à partir de données microscopiques et des variables de conduite de procédé. Mais, la taille et la densité importante de l'imposant réseau ainsi formé font que son apprentissage n'est pas aisé.

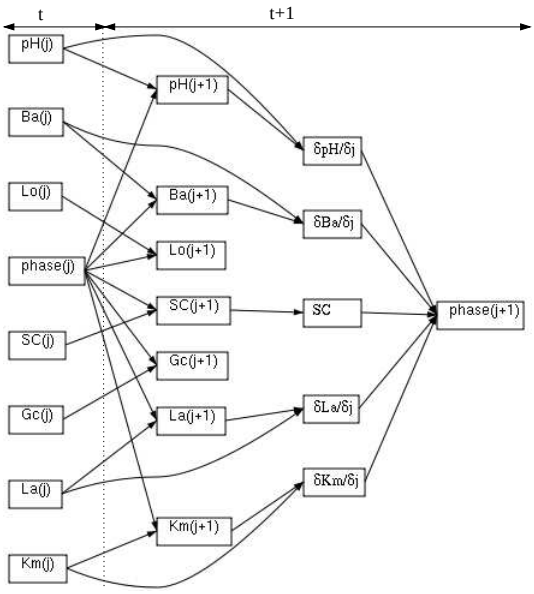


Figure 2. *Structure du RBD définie avec les experts. Le nœud $phase(j)$ est observé.*

3.2. *Les limites pour l'apprentissage des paramètres*

Les différentes méthodes pour réaliser l'apprentissage des paramètres d'un RBD sont principalement des extensions des méthodes des RB (Buntine, 1994; Buntine, 1996). Puisque nos données sont incomplètes et parcellaires, nous avons utilisé l'algorithme EM.

Pour être cohérent avec le phénomène physique à modéliser, les pas de discrétisation de certaines variables du réseau sont faibles. Cela oblige donc à manipuler un nombre importants de valeurs. Par exemple, la variable qui représente le niveau de Ba possède 16 modalités. Sachant qu'il y a 4 phases observées durant l'affinage, pour entièrement définir la CPD du nœud $Ba\,(j+1)$(*i.e.* l'ensemble des $\Pr\left(Ba\,(j+1)\,|Ba\,(j)\,,phase\,(j)\,\right)$), il faudrait donc éliciter $16\times16\times4=1024$ probabilités. En l'état, l'ensemble du réseau nécessite l'élicitation de 7319 probabilités. La base de données de référence contient les résultats de 15 fabrications de camemberts modèles avec une donnée pour chaque nœud du réseau par jour sur les 41 jours que dure l'affinage soit 615 données. Il paraît donc utopique de vouloir renseigner complètement les CPD avec aussi peu de données. Les tables comportent donc des zones parfois importantes de méconnaissances (probabilités calculées à partir de peu d'observations) ou de "non-connaissance" (pas d'observation possible ou observations manquantes car trop difficiles à effectuer).

Le choix retenu actuellement pour représenter la non-connaissance est de se référer au principe de Laplace dit de "raison insuffisante" qui précise que tout ce qui est équiplausible est équiprobable. Cela se traduit par l'utilisation d'une loi de probabilité uniforme pour traiter les différents cas non observés. En effet, pour un nœud X dont les parents sont $Pa\,(X)$, on doit estimer l'ensemble des $\Pr\,(X\,|Pa\,(X)\,)$ avec $\sum_i \Pr\,(X = x_i\,|Pa\,(x) = pa\,) = 1$ et ceci pour chaque combinaison pa possible des valeurs des parents de X qu'il y ait ou pas des observations présentes dans la base de données. Nous verrons que cette propriété de base en probabilité a des conséquences importantes lors de la modélisation d'un processus biologique.

Malgré le manque de données et l'utilisation de loi uniforme, une validation croisée montre que le réseau effectue des prédictions correctes dans en moyenne 70% des cas en tenant compte des erreurs de mesure admises en agro-alimentaire. En revanche, les moments exacts des transitions entre les phases sont plus difficiles à prédire précisément. Afin de mieux prédire la phase à $j + 1$, nous proposons de remettre en cause *a posteriori* le contenu des CPD par des interactions avec les experts à l'aide d'une représentation visuelle adaptée. Cette étape supplémentaire dans l'apprentissage devrait permettre d'injecter de nouvelles connaissances dans le modèle pour améliorer sa cohérence avec le processus physique modélisé.

4. Validation et enrichissement des CPD

Dans la suite, pour un nœud A donné, nous appelons un scénario une combinaison particulière des valeurs des variables parents de A. Par exemple, soit $B = \{b_1, b_2, \ldots, b_m\}$ et $C = \{c_1, c_2, \ldots, c_n\}$ les parents de A ($B \rightarrow A \leftarrow C$). Les n-uplets (b_i, c_j) résultats de $B \times C$ qui forment les lignes de la table de probabilités associées à A, sont les différents scénarios associés au nœud A.

Notre méthode se décompose en 4 parties :

1) apprentissage automatique classique sur la base de données d'observations (algorithme EM),

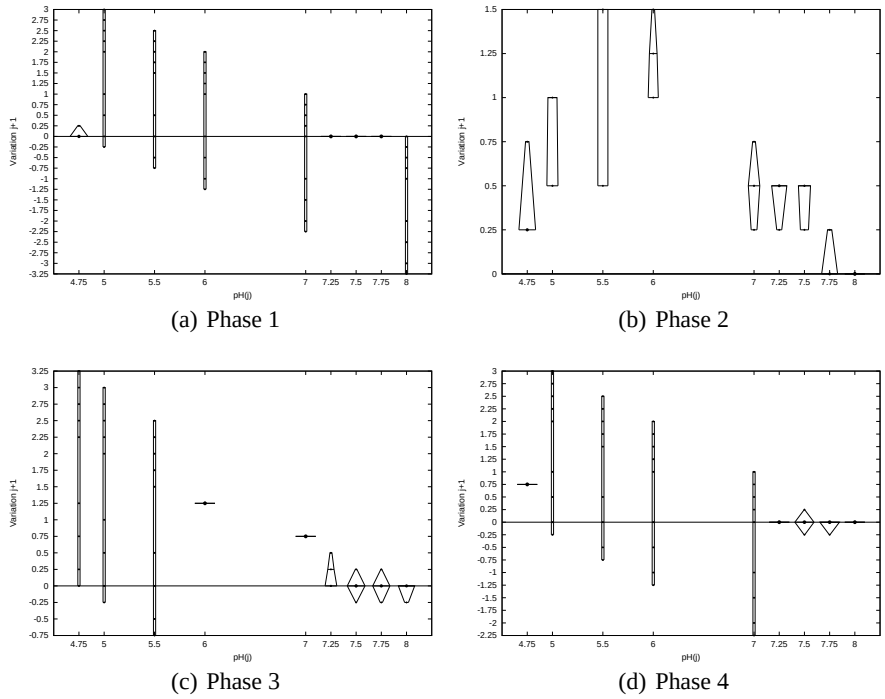
2) représentation graphique des tables de probabilités conditionnelles,

3) caractérisation des différents scénarios identifiés (par ex. scénario non observé, erreur d'apprentissage), et

4) selon les scénarios et les décisions de l'expert, corriger les tables ou la structure pour réduire l'imprécision.

4.1. *Visualisation des tables*

Afin que les experts puissent prendre facilement en main les représentations visuelles des CPD, nous nous restreignons volontairement à des types de représentations qu'ils connaissent comme une courbe d'évolution d'un micro-organisme sur un repère



(a) Phase 1 (b) Phase 2

(c) Phase 3 (d) Phase 4

Figure 3. *Visualisation de la table de probabilités du nœud $pH(j+1)$ découpée selon les 4 phases. Les points représentent les probabilités conditionnelles non nulles. Plus l'enveloppe autour d'un point est large et plus la probabilité associée est importante.*

2D classique en fonction du temps. La visualisation est décomposée selon les 4 phases qui sont clés et structurantes pour les experts.

A titre d'illustration, intéressons nous à $Phase(j) \rightarrow pH(j+1) \leftarrow pH(j)$. La figure 3 est la visualisation de la table de probabilités du nœud $pH(j+1)$. L'axe des abscisses représente les différentes combinaisons des modalités des nœuds parents (ici seulement $pH(j)$) et l'axe des ordonnées représente la variation de la valeur de la variable à $j+1$. Un point indique une probabilité conditionnelle non nulle. A chaque point est associé un segment horizontal dont la longueur est proportionnelle à la probabilité associée. Les extrémités des différents segments sont reliées entre elles afin de former une enveloppe qui entoure les points. Les scénarios représentés par des probabilités uniformes sont facilement identifiables grâce à la forme rectangulaire et allongée de l'enveloppe (par ex. pour $phase(j) = 4$ et $pH(j) = 5.5$).

Quand la structure du réseau sera complétée (ajout d'interactions et de nouvelles variables), il sera nécessaire de filtrer les variables supplémentaires afin que les différents points de l'axe des abscisses soient composés avec le moins de variables

possibles (décomposition fine des CPD). Pour un nombre restreint de variables, une simple sélection sur les modalités semble suffisante. Avec un nombre plus important de variables, nous envisageons l'adaptation de méthodes d'analyses de données telle que l'Analyse en Composante Principale (ACP) afin de ne conserver que les variables les plus significatives.

4.2. *Formalisation des scénarios identifiés*

L'étape suivante est la formalisation de la non-connaissance ou de la méconnaissance qui a été mis en exergue par la visualisation des tables de probabilités. Avec l'aide de deux experts de l'INRA spécialisés sur l'affinage du camembert, et après un court moment nécessaire à la prise en main des visualisations des tables, quelques minutes ont suffit pour mettre en avant différents problèmes issus de l'apprentissage automatique réalisé sur la base de données de référence.

Les différents scénarios sont décomposés en deux groupes : les scénarios non observés qui peuvent être physiquement impossibles ou rares et les autres où certaines probabilités peuvent être précisées.

4.2.1. *Scénarios impossibles et non observés*

De façon générale en dynamique des processus biologiques, certaines combinaisons de valeurs des variables sont physiquement impossibles et ne seront donc jamais observables. Néanmoins, la propriété énoncé en 3.2 implique que l'algorithme d'apprentissage des paramètres doit indiquer au moins une probabilité non nulle. Par exemple, si $phase(j)$ est observée à 3 ou 4 alors on sait que $pH\,(j+1) > 6$. Donc pour les valeurs de $pH\,(j) \leq 6$, ces scénarios sont impossibles et ne seront jamais observés pour de l'affinage de camemberts (le cas de $pH(j) = 4.75$ sera traité par la suite). Par hypothèse, une loi uniforme est alors utilisée pour traiter ces différents scénarios.

4.2.2. *Scénarios possibles mais non observés*

L'incertitude inhérente à tous processus agro-alimentaires ainsi que la complexité des fabrications fromagères (durée, coûts humain et financier importants) fait que certains scénarios de fréquence d'apparitions faible n'ont pas pu être observés par exemple si $phase(j) = 4$ et $pH(j) = 7$. Dans ce cas, une loi uniforme est utilisée dans les CPD.

4.2.3. *Scénarios impossibles observés (biais d'apprentissage)*

La base de données de références comporte peu d'enregistrements par rapport au nombre théorique d'observations requises pour effectuer un apprentissage automatique correct des paramètres. Quelques observations anormales notamment issues d'erreurs de mesures non détectées et d'accident de fabrication peuvent donc produire des probabilités conditionnelles erronées et aberrantes. Par exemple, si $phase(j) = 4$,

on sait qu'il est impossible que $pH\,(j+1) = 4.75$. Ce genre de problème peut aussi se produire si le nombre de valeurs manquantes dans la base de données est trop important. On met ainsi en avant des comportements marginaux.

4.3. *Prise en compte dans le RBD*

Selon le type de scénario identifié et la réponse de l'expert, plusieurs solutions sont possibles pour mettre à jour et/ou enrichir le RBD :

1) l'utilisation des probabilités uniformes due aux contraintes théoriques rend possibles les scénarios impossibles. Mais, nous ne connaissons pas pour le moment leur impact sur les capacités de prédiction du modèle. Néanmoins, il nous semble important que le réseau soit capable de gérer ces scénarios pour ne pas donner de résultats aberrants. Leur prise en compte dans le modèle peut être effectué par une mise à jour de la structure du réseau en ajoutant des nœuds contraintes. Soit une structure de RB $A \rightarrow C \leftarrow B$. L'observation de la valeur de C va imposer des contraintes sur A et B qui ne sont dans ce cas pas indépendants. C peut donc être qualifié de nœud contrainte. Ce type de nœud permet d'exprimer des contraintes existantes entre deux variables pour interdire certaines configurations. Par exemple, pour le nœud Lo, la valeur à $j+1$ ne peut pas être supérieure à celle à j ou encore le pH ne peut pas être inférieur à 7 si $phase(j) = 3$ ou 4.

2) pour les scénarios possibles et non observés, l'expert peut le plus souvent réduire le nombre de modalités en ajoutant quelques probabilités nulles. Une redistribution uniforme est ensuite effectuée sur le nouvel intervalle. Par exemple, si $phase(j) = 4$ et $pH(j) = 7$, la variation de pH est comprise en 0 et 0.5. Nous n'avons plus que 3 probabilités non nulles au lieu des 9 initiales ou plus généralement, on sait que le *pH* ne diminue plus à partir de la phase 2. Ces scénarios peuvent être vus comme des guides pour les expérimentations futures afin d'essayer de réaliser les observations manquantes.

3) Pour les biais d'apprentissage, soit on est sur un scénario impossible (cas n°1) et les probabilités sont redistribués uniformément sur l'ensemble des modalités de la variable concernée, soit on est sur un scénario rare et l'expert devrait pouvoir donner un intervalle de valeur comme dans le cas n°2.

5. Futurs travaux

Nous nous sommes pour le moment surtout concentrés sur les erreurs d'apprentissage et les probabilités uniformes. Mais, certaines probabilités conditionnelles sont certainement calculées à partir de peu d'observations. Nous souhaitons pouvoir associer un indice de confiance à ces probabilités. À partir de la base de données, il est aisé de retrouver le nombre d'observations qui a servi à calculer chaque probabilité conditionnelle. Ce nombre n qui est utilisé dans les algorithmes d'apprentissage de paramètres classiques, tel que le maximum de vraisemblance, est un bon indicateur de

la fiabilité d'une probabilité : si $n = 0$, les parents du nœud concerné n'ont jamais été observés dans la configuration recherchée, donc on retrouve la présence de probabilités uniformes et plus n est proche de 1 moins la probabilité calculée est fiable puisque le nombre d'observations utilisé pour son calcul est réduit. Des tests statistiques qui donnent un intervalle de confiance sur une probabilité en fonction du nombre d'observations ayant servi à la calculer et du degré de liberté de la variable aléatoire associée pourront être ensuite utilisés (Saporta, 2006).

Notre objectif principal est le développement d'une application simple à utiliser par les experts pour valider et enrichir les CPD. Pour cela, les enveloppes autour des probabilités sont interactives. L'expert peut modifier la forme de l'enveloppe et rajouter des points. Les probabilités sont ensuite déduites de la forme géométriques (rapport des longueurs des segments horizontaux et nombre de points). Une extension nécessaire est la gestion des nœuds temporels avec plus de deux parents. Dans ce cas, l'axe des abscisses sera composé de l'ensemble des parents du nœud concerné. Un système de type ACP pourrait permettre de filtrer les variables afin de "découper au plus fin" la table de probabilités pour réduire la complexité de la figure produite.

6. Conclusion

Nous avons présenté dans cet article nos travaux en cours sur la modélisation de la dynamique de l'affinage de fromages de type camembert par un RBD dont la structure a été définie avec les experts du domaine. La principale difficulté lors du calcul des paramètres est la prise en compte correcte dans le modèle de l'incomplétude des données que l'on possède sur le processus et du manque de précision qui en découle. La méthode que nous proposons qui consiste à remettre en cause les probabilités issues d'un apprentissage automatique afin d'injecter aux endroits nécessaires de nouvelles connaissances semble prometteuse. Les premiers retours des experts sur les représentations visuelles et leurs traitements associés sont encourageants. Une comparaison stricte avec un apprentissage automatique classique du RBD reste à effectuer pour mesurer les apports de notre méthode en terme de précision et de cohérence du réseau avec le processus d'affinage.

7. Bibliographie

Babin P., Valle G. D., Dendievel R., Lassoued N., Salvo L., « Mechanical properties of bread crumbs from tomography based finite element simulation », *J. of Material Science*, vol. 40, n° 22, p. 5867-5873, 2005.

Buntine W. L., « Operations for learning with graphical models », *J. of AI Research*, vol. 2, p. 159-225, 1994.

Buntine W. L., « A guide to the literature on learning probabilistic networks from data », *IEEE Trans. on Knowledge and Data Engineering*, vol. 8, n° 2, p. 195-210, 1996.

Card S., Mackinlay J., Shneiderman B., *Readings in Information Visualization ; Using vision to think*, Morgan Keufmann, 1999.

Dubois D., Prade H., « What are fuzzy rules and how to use them », *Fuzzy Sets and Systems*, vol. 84, n° 2, p. 169-185, 1996.

Leclercq-Perlat M.-N., Buono F., Lambert F., Latrille E., Spinnler E., Corrieu G., « Controlled production of Camembert-type cheeses. Part I: Microbiological and physicochemical evolutions », *J. of Dairy Sciences*, 2004.

Leclercq-Perlat M.-N., Picque D., Riahi H., Corrieu G., « Microbiological and Biochemical Aspects of Camembert-type cheeses Depend on Atmospheric Composition in the Ripening Chamber », *J. of Dairy Sciences*, vol. 89, p. 3260-3273, 2006.

Murphy K., Dynamic Bayesian Networks: Representation, Inference and Learning, PhD thesis, University of California, Berkeley, 2002.

Naïm P., Wuillemin P.-H., Leray P., Pourret O., Becker A., *Réseaux Bayésiens*, Eyrolles, 2007.

Pearl J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.

Perrot N., Ioannou I., Allais I., Curt C., Hossenlopp J., Trystram G., « Fuzzy concepts applied to food product quality control: a review », *Fuzzy sets and system*, vol. 157, n° 9, p. 1145-1154, 2006.

Picque D., Leclercq-Perlat M.-N., Corrieu G., « Effects of Atmospheric Composition on Respiratory Behavior, Weight Loss, and Appearance of Camembert-Type Cheeses During Chamber Ripening », *J. of Dairy Science*, vol. 89, p. 3250-3259, 2006.

Saporta G., *Probabilités, analyse des données et Statistiques*, Technip, 2006.

Trystram G., Courtois F., *Computerized Control Systems in the Food Industry*, Marcel Dekker Publishers, chapter Food Process Modelling, p. 55-85, 1997.