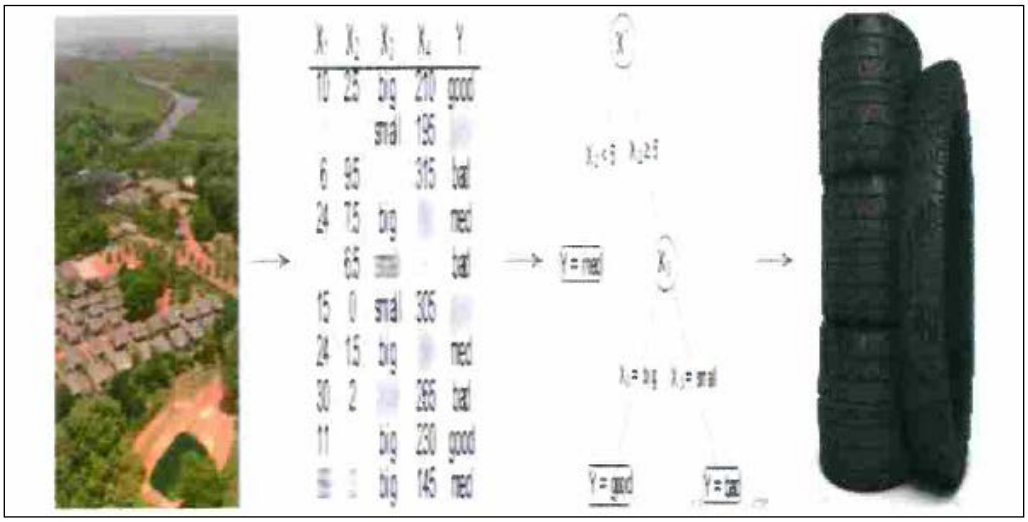
Apprentissage à partir de données et de connaissances incertaines : application à la prédiction de la qualité du caoutchouc

Nicolas Sutton-Charani

Par **Nicolas SUTTON-CHARANI**

Apprentissage à partir de données et de connaissances incertaines : application à la prédiction de la qualité du caoutchouc

Thèse présentée
pour l'obtention du grade
de Docteur de l'UTC



Soutenue le 28 mai 2014
Spécialité : Technologies de l'Information et des Systèmes

D1835

Apprentissage à partir de données et de connaissances incertaines.

Application à la prédiction de la qualité du caoutchouc

Nicolas SUTTON-CHARANI

Thèse soutenue le 28 Mai 2014 devant le jury composé de :

Président :

Yves GRANDVALET
Directeur de Recherche
Univ. de Technologie de Compiègne

Rapporteurs :

Didier DUBOIS
Directeur de Recherche CNRS
Univ. Paul Sabatier

Arnaud MARTIN
Professeur
Univ. de Rennes 1

Examinateurs :

Jérôme SAINTE BEUVE
Chargé de Recherche
INRA SupAgro

Mathieu SERRURIER
Maître de Conférences
Univ. Paul Sabatier

Tristan MARY-HUARD
Chargé de Recherche
INRA AgroParis Tech

Eric GOHET
Chargé de Recherche
CIRAD

Brigitte CHARNOMORDIC
Ingénieur de Recherche
INRA

Directeurs de Thèse :

Thierry DENOEUX
Professeur
Univ. de Technologie de Compiègne

Sébastien DESTERCKE
Chargé de Recherche
Univ. de Technologie de Compiègne

Université de Technologie de Compiègne

Laboratoire Heudiasyc UMR CNRS 7253

28 Mai 2014

A ma mère

Table des matières

Préface

Remerciements

Cette thèse a été financée par le $CIRAD$ (Centre de coopération internationale en recherche agronomique pour le développement).

Je tiens tout d'abord à remercier vivement mon encadrant : Sébastien Destercke. Au delà de ses qualités techniques et pédagogiques, il a su me transmettre un véritable goût pour la recherche. Ses qualités humaines et sa disponibilité m'ont été d'un grand secours. J'ai eu la sensation d'avoir bénéficié grâce à lui d'un vaccin contre certains états de détresse dans lequels se trouvent beaucoup de doctorants.

Je suis aussi reconnaissant envers mon directeur de thèse Thierry Denoeux, qui a bien initié les orientations de ce travail et surtout qui a su être réceptif aux pistes que je proposais et techniquement réactif lorsque ces pistes pouvaient aboutir à quelque chose.

Ce travail s'étant déroulé en partie à l'Université Technologique de Compiègne (UTC) et au $CIRAD$ à Montpellier, je fus accueilli au laboratoire $MISTEA$ (Statistique et Informatique) sur le campus de l'$INRA - SupAgro$ de Montpellier où ma référente a été Brigitte Charnomordic. Je tiens donc à remercier chaleureusement cette dernière sans qui cette thèse n'aurait pu aboutir. De par sa polyvalence scientifique et son énergie inépuisable elle m'a permis de pouvoir travailler de façon pluri-disciplinaire (informatique, statistique, agronomie) et de rester en phase avec le monde de la recherche et ses évolutions actuelles.

Par la même occasion je remercie tout le laboratoire $MISTEA$ qui m'a accueilli dans des conditions plus qu'agréables ce qui m'a permis d'utiliser mes connais-

sances initiales (principalement statistiques) dans un cadre informatique centré sur la notion d'incertitude. Je suis également très reconnaissant envers Nicolas Verzelen avec qui nous avons travaillé avec Brigitte Charnomordic sur les notions de vraisemblances crédibilistes dans un cadre statistique simple (voir Chapitre 2.4.1). J'ai eu la chance de travailler dans une belle atmosphère où règne une bonne humeur énergisante. Merci Lamia puis à Cheikh de m'avoir supporté dans leur bureau, merci à Pascal Neveu et Christophe Abraham pour m'avoir bercé avec du jazz de qualité pendant certaines soirées laborieuses. Merci Nadine, Anne, Véronique, Maria, Meili, Bénédicte, Patrice, Alexandre, Isabelle, Martine, Yuan, Philippe, Damien, Tito, Aunur, Nikolay, Fabien et Coralie pour tous vos précieux conseils et toutes les discussions passionnantes que l'on a pu avoir.

Je suis bien évidemment plus que reconnaissant envers l'UTC et plus particulièrement le laboratoire $HEUDYASIC$ où j'ai effectué ma première année de thèse. Ce laboratoire m'a permis d'acquérir une certaine rigueur scientifique, et une curiosité grandissante pour beaucoup de notions liées à l'incertitude. Merci à Benjamin Quost, Yves Grandvalet et Mylène Masson qui ont toujours été disponibles pour m'aider à avancer. J'ai eu la chance d'y rencontrer des doctorants vraiment formidables, tant sur le plan technique et de l'entraide que sur le plan humain. Merci Luis, Felipe, Sawsan, Nicole, Li Qiang, Jiqiong, Julien, Vincent, Adam, Mouldy, Liu Xiao, Ada, Hoda, Karim, Marek, Jennifer etc.

Lors de mon séjour à Montpellier, même si j'étais localisé dans le laboratoire $MISTEA$, je faisais partie du laboratoire $IATE$ (UMR dépendant du $CIRAD$). Même si je n'ai malheureusement pas eu l'occasion d'y passer beaucoup de temps je suis très reconnaissant envers ceux avec qui j'ai pu collaborer ou échanger, notamment Patrice Buche et Karim. Par la même occasion je remercie bien évidemment le $CIRAD$ pour avoir financé ma thèse mais surtout pour m'avoir matériellement donné la possibilité de la terminer.

Cette thèse comprend une application (*caoutchouc*) dont les données ont été fournies par $MICHELIN$ au $CIRAD$. Je tiens ici à remercier $MICHELIN$ pour avoir partagé avec nous ces données qui sont d'une rareté et d'une richesse non-négligeable, plus particulièrement M^{me} Maria Emilia Barcellos de Menezes (Responsable des laboratoires des usines de production de caoutchouc naturel Michelin

au Brésil). Pour cette application j'ai eu la chance de travailler avec Jérôme Sainte-Beuve ($IATE - CIRAD$) et Eric Gohet ($CIRAD$) que je remercie également pour avoir encadré ce travail applicatif qui était loin d'être simple a priori. Venant d'horizons scientifiques (agronomie, chimie, ...) différents des miens, ils ont su me guider, me transmettre quelques connaissances de base concernant le caoutchouc naturel et surtout m'intéresser à la problématique. Au début de ce travail applicatif j'ai encadré une stagiaire, Emilie Doge, qui m'aida largement à pré-traiter les données et à réaliser les études statistiques préliminaires, je la remercie donc humblement.

Je suis également très reconnaissant envers tous les services administratifs avec lesquels j'ai collaboré. Le rôle joué par les secrétaires dans le monde de la recherche est trop souvent sous-estimé, je suis de ceux qui aiment rappeler que sans elles/eux rien n'est possible. Merci donc à Céline Ledent, Nathalie Alexandre, Julie Jarek, Bérengère Guermonprez, Sabine Vidal, Magalie Colignon, Véronique Moisan (pour le support informatique) pour l'UTC, un grand merci à Clara Lachgar, Coralie Bertrand et une pensée très chaleureuse à Valérie Aboudarham pour $IATE$, et enfin je souhaite remercier très chaleureusement Véronique Sals-Vettorel et Maria Trouche de $MISTEA$.

Même si un doctorat constitue essentiellement un travail de recherche, j'ai tiré beaucoup de plaisir et appris beaucoup lors des enseignements que j'ai pu dispenser. Je remercie donc ceux qui m'ont permis d'enseigner, Thierry Denoeux à l'UTC, Catherine Trottier et Christian Lavergne à l'Université de Montpellier 3, et enfin je suis infiniment reconnaissant envers Nicolas Molinari pour la confiance qu'il m'a donné lors des enseignements que j'ai effectués à l'Université de Montpellier 1.

L'informatique étant une discipline où les échanges sont vitaux, je remercie de manière générale tous les chercheurs avec qui j'ai pu discuter, notamment au sein de la communauté des fonctions de croyances. J'en profite pour remercier le Pr. Dempster pour avoir répondu à certaines questions que j'ai pu lui adresser un peu spontanément par mail.

La vie scientifique n'est qu'un aspect de la vie d'un chercheur, derrière tout chercheur se cache une personne et une vie privée. Pour ma part, j'ai eu la chance

d'être très bien entouré durant ma thèse mais aussi avant. Je remercie donc ma mère pour l'éducation qu'elle m'a donnée, son soutien perpétuel, l'amour dans lequel j'ai grandi et la curiosité qu'elle a su cultiver en moi. Merci à mon père pour la bonne humeur infaillible qu'il a su me transmettre, pour ses conseils et ses mises en garde ainsi que son soutien. Je remercie bien évidemment ma femme Yasmine pour m'avoir accompagné au jour le jour tout au long de mon doctorat et pour m'avoir donné un merveilleux fils, Malik que je remercie pour avoir infiniment augmenté l'amour et la confiance que j'ai en la vie. Par la même occasion je tiens à remercier amplement les familles Sutton et Charani qui ont complété très agréablement mon éducation.

Je veux aussi remercier fraternellement tous mes amis, Julien, Fabrice, Samir, Thomas (H et Z), Johana, Audrey, Eloine, Ludovic, Clara, Adrien, etc.

Résumé français de la thèse

Pour l'apprentissage de modèles prédictifs, la qualité des données disponibles joue un rôle important quant à la fiabilité des prédictions obtenues. Ces données d'apprentissage ont, en pratique, l'inconvénient d'être très souvent imparfaites ou incertaines (imprécises, bruitées, etc). Ce travail de doctorat s'inscrit dans ce cadre où la théorie des fonctions de croyance est utilisée de manière à adapter des outils statistiques classiques aux données incertaines.

Le modèle prédictif choisi est l'arbre de décision qui est un classifieur basique de l'intelligence artificielle mais qui est habituellement construit à partir de données *précises*. Le but de la méthodologie principale développée dans cette thèse est de généraliser les arbres de décision aux données incertaines (floues, probabilistes, manquantes, etc) en entrée **et** en sortie. L'outil central d'extension des arbres de décision aux données incertaines est une vraisemblance adaptée aux fonctions de croyance récemment proposée dans la littérature dont certaines propriétés sont ici étudiées de manière approfondie. De manière à estimer les différents paramètres d'un arbre de décision, cette vraisemblance est maximisée via l'algorithme E^2M qui étend l'algorithme EM aux fonctions de croyance.

La nouvelle méthodologie ainsi présentée, les arbres de décision E^2M, est ensuite appliquée à un cas réel : la prédiction de la qualité du caoutchouc naturel. Les données d'apprentissage, essentiellement culturales et climatiques, présentent de nombreuses incertitudes qui sont modélisées par des fonctions de croyance adaptées à ces imperfections. Après une étude statistique standard de ces données, des arbres de décision E^2M sont construits et évalués en comparaison d'arbres de décision classiques. Cette prise en compte des incertitudes des données permet ainsi d'améliorer très légèrement la qualité de prédiction mais apporte surtout des informations concernant certaines variables peu prises en compte jusqu'ici par les experts du caoutchouc.

Résumé anglais de la thèse

During the learning of predictive models, the quality of available data is essential for the reliability of obtained predictions. These learning data are, in practice very often imperfect or uncertain (imprecise, noised, etc). This PhD thesis is focused on this context where the theory of belief functions is used in order to adapt standard statistical tools to uncertain data.

The chosen predictive model is decision trees which are basic classifiers in Artificial Intelligence initially conceived to be built from *precise* data. The aim of the main methodology developed in this thesis is to generalise decision trees to uncertain data (fuzzy, probabilistic, missing, etc) in input **and** in output. To realise this extension to uncertain data, the main tool is a likelihood adapted to belief functions, recently presented in the literature, whose behaviour is here studied. The maximisation of this likelihood provide estimators of the trees' parameters. This maximisation is obtained via the E^2M algorithm which is an extension of the EM algorithm to belief functions.

The presented methodology, the E^2M decision trees, is applied to a real case : the natural rubber quality prediction. The learning data, mainly cultural and climatic, contains many uncertainties which are modelled by belief functions adapted to those imperfections. After a simple descriptiv statistic study of the data, E^2M decision trees are built, evaluated and compared to standard decision trees. The taken into account of the data uncertainty slightly improves the predictive accuracy but moreover, the importance of some variables, sparsely studied until now, is highlighted.

Publications durant le doctorat

Articles de conférences internationales :

- **Classification Trees Based on Belief Functions**, *N. Sutton-Charani, S. Destercke, T. Denoeux*, in Proc. of the 2^{nd} International Conference on Belief Functions (BELIEF 2012), Compiègne, France, Avril, 2012
- **Learning decision trees from uncertain data with an evidential EM approach**, *N. Sutton-Charani, S. Destercke, T. Denoeux*, 12^{th} International Conference on Machine Learning and Applications (ICMLA 13), Miami, USA, Décembre 4-7, 2013
- **Application of E2M decision trees to rubber quality prediction**, *N. Sutton-Charani, S. Destercke, T. Denoeux*, 15^{th} International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2014), Montpellier, France, Juillet 15-19, 2014
- **Training and evaluating classifiers from evidential data : application to E2M tree pruning**, *N. Sutton-Charani, S. Destercke, T. Denoeux*, in Proc. of the 3^{rd} International Conference on Belief Functions (BELIEF 2014), Oxford, UK, Septembre, 2014

Articles de conférence francophone :

- **Arbres de classification construits à partir de fonctions de croyance**, *N. Sutton-Charani, S. Destercke, T. Denoeux*, $21^{ème}$ Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2012), Compiègne, France, Novembre 15-16, 2012

Introduction Générale

Une grande partie des domaines de l'intelligence artificielle d'une part, et des statistiques d'autre part, se concentrent sur la problématique d'apprentissage de modèles à partir des données disponibles. De manière générale, les données sur lesquelles s'appuient ces méthodes sont supposées précisément connues. Si cette hypothèse est raisonnable dans de nombreux cas, il en existe nombre d'autres où les données disponibles sont souvent imparfaites, incomplètes, incertaines ou pas totalement fiables. C'est en particulier le cas pour les sciences du vivant [1], les problèmes de fiabilité des systèmes [2] ou encore les sciences climatiques [3].

L'incertitude de ces données, lorsqu'elle est prise en compte, est le plus souvent modélisée par des probabilités et/ou des intervalles. Il existe au moins deux raisons de ne pas se restreindre à ces deux modèles :

– d'une part, l'incertitude des données est avant tout épistémique (i.e., elle est réductible via l'obtention de nouvelles informations), et de nombreux auteurs mettent en doute l'universalité des probabilités pour modéliser ce genre d'incertitude (notamment pour modéliser l'ignorance)
– d'autre part, en pratique nous posséderons souvent plus d'information qu'un simple intervalle ou ensemble de valeurs possibles, mais moins d'information qu'il n'en faudrait pour obtenir un modèle probabiliste précis.

Il est donc intéressant d'étudier des modèles d'incertitude assez expressifs pour intégrer ces deux extrêmes (intervalles/ensembles et probabilités) dans un cadre unique, et de proposer des méthodes d'apprentissage pouvant gérer ce type de données incertaines. Le but principal de telles méthodes est d'utiliser au mieux l'information dans l'apprentissage du modèle, tout en ne contraignant pas trop le modèle d'incertitude.

Le présent manuscrit se propose de traiter cette problématique dans le cadre des fonctions de croyance, qui offrent cette flexibilité que nous venons d'évoquer. L'objectif principal de ce travail est donc d'apprendre un modèle à partir d'exemples constitués de couples *entrée/sortie* dans le but de pouvoir prédire la sortie d'un nouvel exemple d'entrée (*apprentissage supervisé*). Les données (entrées et sorties), habituellement constituées de valeurs précises ont, dans ce manuscrit, la particularité d'être éventuellement constituées de fonctions de croyance. L'apprentissage à partir de données incertaines (représentées par des fonctions de croyance ou par d'autres modèles) est un problème complexe car même s'il existe différents modèles d'incertitude, rares sont les méthodologies d'apprentissage prévues pour gérer ces

modèles.

Le choix des arbres de décision comme modèle d'apprentissage s'est assez rapidement imposé, vu qu'à l'objectif de prédiction s'ajoute, dans le problème appliqué qui sera traité ici, le besoin d'obtenir des modèles explicatifs lisibles par les experts agronomes. En effet, la compréhension des mécanismes qui entraînent une plus ou moins grande qualité du caoutchouc naturel est au moins aussi (sinon plus) essentielle à la capacité de bien prédire cette qualité.

Le caoutchouc représente en effet un enjeux économique important, à la fois pour les sociétés (comme Michelin) utilisant cette matière première et pour les producteurs. Les scientifiques et experts du caoutchouc n'ont pas encore une connaissance suffisante des facteurs impactent cette qualité pour bien la contrôler, il est donc nécessaire de raffiner cette connaissance. De plus, les données expérimentales récupérées dans les plantations (souvent de plusieurs centaines d'hectares) sont entachées de fortes incertitudes. Les méthodes développées dans ce travail sont donc bien adaptées pour tirer le meilleur parti des données (incertaines) afin de mieux comprendre ces phénomènes (et éventuellement diriger les efforts expérimentaux futurs).

La première partie de ce manuscrit est dédiée à un rappel des différentes notions générales d'incertitude, des principales théories de l'incertain, et plus précisément de la théorie des fonctions de croyance (cadre de travail choisi dans ce manuscrit). Elle se concentre ensuite sur la notion de vraisemblance dans cette théorie.

La seconde partie revoit la littérature sur les arbres de décision avant de présenter deux nouvelles méthodes de construction adaptées aux données incertaines. La première est une extension d'un travail existant pour prédire des classes binaires (ne pouvant prendre que deux valeurs différentes) au cas multi-classes. La deuxième constitue le cœur du manuscrit et sera donc détaillée de façon approfondie.

La troisième partie comprend les différentes expériences réalisées sur des jeux de données standards pour les deux méthodologies précédemment proposées, ces expériences nous servant à valider les méthodes. Elle présente ensuite en détail l'application agronomique : la structure des données, les modèles d'incertitude développés pour ces données, ainsi qu'une étude des arbres E^2M obtenus à partir de ces données incertaines.

Première partie .

Incertitude et prédiction

CHAPITRE 1

Incertitudes

La notion d'*incertitude* est centrale dans le présent manuscrit. Après avoir défini les principaux types d'incertitude et la manière dont ils interviennent dans ce travail, nous passons ici en revue les différentes théories de l'incertain, l'interprétation qu'elles donnent à cette notion d'incertitude, quelques très bref rappels historiques ainsi que les principaux outils nécessaires à la compréhension de la suite du manuscrit.

Sommaire

1.1. Introduction

Le terme d'*incertitude* est défini littéralement comme l'absence de certitude, c'est-à-dire l'absence d'information précise rendant la réalisation d'un événement parfaitement connue.

1.1.1. Types d'incertitude

Cette connaissance relative à la réalisation de l'événement futur peut être de qualité ou de fiabilité variable. Une information imprécise, par exemple, sera de moindre qualité qu'une information précise.

Il existe donc un lien entre les notions d'*incertitude* et d'*imprécision*. En effet si on sait avec certitude que la valeur d'une variable W définie sur l'espace Ω_W appartient à un ensemble de valeur A, i.e $W \in A \subseteq \Omega_W$, notre connaissance relative à W est alors imprécise ce qui rend la valeur de W incertaine. Les notions d'incertitude et d'imprécision sont donc liées par nature, l'imprécision étant vue ici comme un type d'incertitude. En réalité ces deux notions interviennent à des niveaux différents : c'est l'imprécision de la connaissance qui est à l'origine d'incertitudes relatives à la valeur de W. Par abus de langage, on considérera cependant l'imprécision comme un type d'incertitude même si elle est en réalité à l'origine d'un type d'incertitude donné.

De nombreux auteurs les utilisent ainsi de manière indifférenciée, parlant par exemple de données *précises* lorsque les valeurs des dites données sont connues avec certitude. Dans [4] par exemple, Périnel parle de données *imprécises* lorsque l'incertitude de ces dernières est décrite par des lois de probabilité au lieu de valeurs précises. Dans le présent manuscrit, ces deux notions sont séparées, et nous parlerons plus volontiers d'incertitude dans les cas où elles pourraient être employées toutes les deux, le terme d'imprécision étant réservé à une dimension purement ensembliste.

Même si de nombreux types d'incertitudes ont été caractérisés par différents auteurs [5], deux d'entre eux sont très souvent utilisés : l'incertitude *épistémique* et l'incertitude *aléatoire* [6].

Le terme *épistémique* fait référence à la connaissance, l'*incertitude épistémique* pourra donc être comprise comme une incertitude ayant pour origine une connaissance insuffisante. Une donnée imprécise implique donc de l'incertitude épistémique quant à sa valeur.

De par sa nature, l'incertitude épistémique peut être annulée (et donc transformée en certitude) par une recherche de connaissances supplémentaires. L'âge du président brésilien peut ainsi être épistémiquement incertain par manque de connaissance, le président brésilien ayant par ailleurs un âge précis qu'il est pos-

sible de découvrir.

L'*incertitude aléatoire*, elle, est généralement relative à une population ou un événement dont la réalisation n'est pas connue. Il n'est donc pas possible d'obtenir de l'information permettant de prédire à coups sûr cette réalisation future, et ce, du fait de l'intervention de certains facteurs liés au hasard. La météo ou la valeur du $CAC40$, par exemple, sont aléatoirement incertaines, même si une meilleure connaissance des paramètres influençant ces phénomènes peut améliorer toute prédiction, on ne pourra jamais les prévoir avec une totale certitude.

La *statistique* est la science qui étudie les populations à partir d'échantillons de ces populations [7], [8]. On peut parler d'*incertitude statistique* dès lors que l'on travail dans un cadre statistique et donc à partir d'échantillons qui constitue une répétition d'exemples. Le but de la démarche statistique est de généraliser des résultats obtenus sur des échantillons aux populations d'où sont issues ces échantillons. Une population est en général considérée comme de très grande taille en comparaison avec celle des échantillons. Historiquement la statistique s'est développée depuis l'antiquité, où elle consistait surtout à recenser le bétail et à l'élaboration de son cours. Elle prend une dimension prédictive dès le $XIII^{ème}$ siècle dans un contexte d'étude des durées de vie par les compagnies d'assurance (ce qui correspondrait aujourd'hui à de l'actuariat). Même si la statistique est aujourd'hui un domaine très vaste, la prédiction en constitue toujours une partie importante. Le sujet de cette thèse s'inscrivant dans une démarche statistique de prédiction (on cherche à prédire à partir d'un échantillon de données), on parlera ainsi d'incertitude aléatoire dès qu'on fera référence à l'incertitude directement liée à la problématique prédictive.

On pourra dire que la démarche statistique vise donc à caractériser les incertitudes aléatoires inhérentes aux populations à partir d'échantillons dont la taille constitue un critère déterminant de la qualité des conclusions obtenues sur les échantillons. D'un point de vue technique, la statistique s'appuie essentiellement sur la théorie des probabilités.

Dans la littérature, on peut aussi retrouver une différenciation de l'incertitude en *subjective /objective* [9]. Cette différenciation pourrait se confondre avec celle ici présentée mais est néanmoins assez différente. En effet, certains auteurs considèrent qu'une connaissance imparfaite ou incomplète ne permet pas une réelle objectivité, cette dernière étant réservée au cas ou la répétitivité d'une expérience (un

nombre de fois suffisamment grand) peut permettre d'infirmer ou de confirmer une valeur de probabilité préalablement calculée ou estimée. Les probabilité subjectives, elles, correspondent à des degrés de confiance, et peuvent donc varier selon l'individu ou l'agent qui éprouve cette *confiance*. Il est même envisageable d'éprouver une confiance plus ou moins grande (on peut donc en tirer une probabilité subjective) quant à la modélisation d'une probabilité objective ou aléatoire.

En réalité, les différenciations d'incertitude en *épistémique/aléatoire* ou en *subjective/objective* ne se font pas au même niveau. On parlera plus volontiers d'incertitude épistémique ou aléatoire d'un coté, et de sources d'information subjectives ou objectives. Si l'information dont on dispose provient d'un échantillon (et donc d'une population) la source est alors objective, alors qu'un avis d'expert est propre à l'expert et est donc subjectif par nature. Ces différentes notions peuvent aussi s'entremêler, par exemple dans le cas où on aurait des données relatives à un échantillon (la source est donc objective), mais où ces données seraient incomplètes, ce qui crée de l'incertitude épistémique dans un cadre aléatoire.

1.1.2. Identification des différentes incertitudes dans ce travail

Dans le présent manuscrit, l'objectif est l'obtention d'un modèle prédictif (arbres de décision) à partir de données incertaines, de l'incertitude interviendra donc à différents niveaux. L'incertitude des données, comprise ici comme une *non-fiabilité* partielle de ces dernières est *épistémique*. En effet, cette incertitude résulte généralement d'une imperfection des observations. On peut donc considérer que si ces observations s'étaient bien déroulées, les données ne seraient alors pas incertaines.

Le modèle prédictif utilisé dans ce manuscrit, i.e. les *arbres de décision* (voir Section 4), traite d'incertitudes *aléatoires*. Comme beaucoup de classifieurs statistiques, il s'inscrit dans un cadre incertain où de nombreuses estimations sont réalisées en utilisant les fréquences. L'utilisation des fréquences comme estimateurs de probabilités est typique pour les cas d'incertitudes *aléatoires*. On suppose généralement que ces fréquences sont suffisamment proches des probabilités recherchées pour des échantillons suffisamment grands. De manière générale, les modèles statistiques supposent que les sorties (classes) sont des fonctions *bruitées* des entrées, i.e. qu'on peut expliquer les sorties avec les entrées mais qu'il se peut qu'un *bruit blanc*, par exemple, vienne modifier la sortie prédite ; ce bruit blanc, agissant comme une variable aléatoire classique, est donc naturellement *aléatoirement* incertain.

Dans ce manuscrit, le but est donc d'apprendre un modèle d'incertitude *aléatoire* à partie de données *épistémiquement* incertaines. Il se trouve que la méthodologie proposée dans ce travail (*arbres de décision* E^2M), conserve cette incertitude épistémique pendant tout le processus d'apprentissage (voir Section 7) et la *combine* avec l'incertitude *aléatoire* inhérente au modèle prédictif à l'aide d'un algorithme défini dans un cadre idéal pour les modélisations d'incertitudes *épistémiques* (cadre des *fonctions de croyance* voir Chapitre 2) : l'algorithme E^2M (voir Section 3.2).

1.2. Théories de l'incertain

La science s'intéresse depuis longtemps aux *incertitudes*, à leur compréhension, à leur traitement ainsi qu'à leur propagation. Pour se faire, l'incertitude doit être quantifiée par une *mesure de confiance* qu'on notera μ dans le cas général. On dénombre plusieurs de ces mesures dans différentes théories, néanmoins elles partagent toutes certaines propriétés de base. Cette mesure $\mu : 2^{\Omega_W} \rightarrow [0,1]$ sera telle que $\forall A \subseteq \Omega_W, \quad \mu(A)$ est le degré de confiance dans le fait que $W \in A$. Elle vérifiera les propriétés suivantes :

Propriété 1. $\mu(\emptyset) = 0$, $\mu(\Omega_W) = 1$

Cette proposition signifie que l'événement *nul* est impossible et que l'événement $w \in \Omega_W$ est certain.

Propriété 2. $\forall A \subseteq B \subseteq \Omega_W$, $\mu(A) \leq \mu(B)$

La Proposition 2 équivaut à dire qu'en considérant deux événements $A, B \subseteq \Omega_W$, si A implique B alors B est au moins aussi certain que A.

Aujourd'hui plusieurs théories étudient les incertitudes, nous en donnons ici un très bref aperçu à travers quelques rappels historiques et quelques définitions.

1.2.1. Théorie des probabilités

La notion de *probabilité* fut initialement dédiée à l'expression d'une certaine forme de consensus d'opinion [10]. Le terme de *probabilité* fut donc d'abord lié à de l'incertitude purement épistémique (ou subjective).

La théorie des probabilités s'est développée très progressivement depuis l'emploi du terme de *probabilité* par Aristote pendant l'antiquité, et donc avec une dimension initiale subjective, jusqu'à aujourd'hui où la nature subjective de certaines probabi-

lités est re-devenue un centre d'intérêt d'une partie des scientifiques étudiant l'incertain. Différentes étapes marquèrent l'histoire de ce développement. Parmi elles :

- les travaux de Girolamo Cardano au début du XVI^{ème} siècle [11]
- les travaux et échanges entre Blaise Pascal, Pierre de Fermat et Christian Huygens sur le problème des *partis* [12]
- les développements mathématiques par Jakob Bernoulli [13]
- le cours de logique de Gabriel Cramer qui inspira l'encyclopédie de Diderot XVIII^{ème} siècle [14]
- les travaux de Bayes [15] et Laplace [16] qui furent déterminants pour le développement des théories des probabilités et de la statistique (Bayésienne entre autre)
- le développement de la théorie de la mesure par Borel [17] et Lebesgue [18] au début du XX^{ème} siècle
- l'axiomisation de Kolmogorov s'appuyant sur cette théorie de la mesure [19]

L'axiomatique de Kolmogorov [19] pose les premières bases de l'actuelle théorie. La mesure de confiance est ici la *probabilité* $\mu = P$. D'après ces axiomes, la probabilité P doit, par définition, vérifier les Propositions 1 et 2. Nous nous intéressons ici plus particulièrement au troisième axiome de Kolmogorov :

$$\forall A, B \subseteq \Omega_W \text{ tels que } A \cap B = \emptyset, \quad \text{on a} \quad P(A \cup B) = P(A) + P(B) \qquad \textit{additivité}$$

Ce dernier axiome, appelé additivité des probabilités (ou σ–additivité en référence aux σ-algèbres, i.e les tribus de Borel), représente une propriété très importante des probabilités. Il faut cependant noter que même si cette additivité peut être souhaitable (notamment par le lien existant entre probabilité et fréquence), elle n'est pas forcément nécessaire dans le cas d'incertitude épistémique. Bernoulli travaillait initialement sur des probabilités *subjectives* mesurant la crédibilité de propositions qui n'étaient pas additives [20]. Le passage d'une théorie tolérant la continuité des probabilités (et donc l'infiniment petit) à la réalité physique d'échantillons statistique de tailles finies peut aussi nécessiter des approximations aboutissant à la non-additivité des probabilités [21, 22]. Tout en travaillant sur ces probabilités *subjectives*, Bernoulli développa aussi la théorie des chances (qui traite elle d'incertitudes purement aléatoires). Il proposa dans ce cadre la fameuse loi des grands

nombres [13]. En étendant cette loi aux probabilités *subjectives* (traitant donc d'incertitude épistémiques), il mit quelque peu fin à cette différenciation entre *épistémique* et *aléatoire*, et la théorie des probabilités actuelle traite toutes les différentes incertitudes de la même manière.

Le caractère subjectif des probabilités épistémiques ne sera re-étudié qu'avec la mathématisation de la logique modale dès la fin du $XIX^{ème}$ siècle [23], période pendant laquelle Keynes et Boole rediscutèrent de la notion de subjectivité relative à l'incertitude [24], [25].

En dépit de cette confusion adoptée par la plupart à l'époque, certains auteurs ré-étudièrent par la suite les *probabilités subjectives*. L'un des premiers fut probablement Ramsey qui étudia l'idée de construction volontaire de probabilité subjectives [26], puis De Finetti dont les travaux sur ces mêmes probabilités subjectives [27] furent généralisés par Savage [28]. Ce dernier formalisa ainsi une *théorie des probabilités subjectives et personnelles* [29]. De Finetti et Savage, cependant, considéraient l'additivité pour les probabilités subjectives comme une propriété désirable. Avec Lindley [30], ils finirent par adopter la théorie statistique Bayésienne, qui elle aussi comporte une notion subjective, et en furent des illustres représentants.

La statistique Bayésienne est une branche de la statistique dans laquelle l'incertitude épistémique est quantifiée à l'aide de probabilités dites *Bayésiennes* qui attribuent des degrés de croyance aux différents événements considérés. Ces probabilités sont additives et peuvent donc être combiner avec n'importe quelle probabilité objective (mesurant de l'incertitude aléatoire). Cette combinaison se fait justement à l'aide la fameuse règle de Bayes :

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

Cette règle permet alors une mise a jour d'une information dite *a priori* et représentée par le terme $P(A)$, basée en général sur la connaissance experte, à partir d'observations empiriques $P(B|A)$ correspondant donc à la vraisemblance conditionnelle de B sachant A (souvent d'un échantillon connaissant un paramètre) et d'un facteur multiplicatif $\frac{1}{P(B)}$. L'information *a priori* est épistémiquement incertaine, en effet un avis d'expert expriment en général une opinion personnelle, et donc subjective, relativement à l'événement A. La vraisemblance conditionnelle $P(B \mid A)$ peut, elle, être considérée comme quantifiant une incertitude aléatoire étant directement liée au modèle probabiliste (qui concerne une incertitude généralement aléatoire)

sous-jacent. Le terme normalisateur $P(B)$ est, lui, souvent plus difficile à calculer, différentes techniques sont alors proposées par les statisticiens Bayésiens. L'information ainsi obtenue *a posteriori* contient donc à la fois de l'incertitude épistémique (de part le choix subjectif de la probabilité *a priori*) et de l'incertitude aléatoire (provenant du terme de vraisemblance $P(B \mid A)$).

On peut aussi arguer que tout comme le président brésilien a un age précis et bien réel, le paramètre d'un modèle a en réalité une valeur qui n'est pas aléatoire, même si elle peut être inconnue. En l'absence d'information, la loi uniforme est généralement utilisée comme loi de probabilité *à priori*.

On peut donc comprendre la statistique Bayésienne, comme au moins partiellement subjective, elle peut donc être inclue dans les théories de l'incertain utilisant des mesures de confiance permettant de représenter l'incertitude épistémique. Parmi ces théories, les principales sont la théorie des possibilités, celle des probabilités imprécises et enfin celle des fonctions de croyance dans laquelle s'inscrivent nos travaux. Nous verrons que de forts liens existent entre ces différentes théories. Une des motivations majeures de ces autres théories est que, comme discuté à de nombreuses reprises par les tenants de ces théories [5], [31], les probabilités, qu'elles soient subjectives ou aléatoires, ne peuvent pas capturer tous les états d'incertitude. La complète ignorance, par exemple, est en général modélisée par une distribution uniforme en théorie des probabilités, or cette distribution est en réalité très spécifique, et l'utiliser pour modéliser l'ignorance totale peut aboutir à divers paradoxes.

Par exemple supposons que la valeur w d'une variable $W \in [1,2]$ est totalement inconnue, son inverse $V = \frac{1}{W} \in [\frac{1}{2}, 1]$ est alors aussi totalement inconnue, donc si W suit une loi de probabilité uniforme sur $[1,2]$, V devrait suivre une loi uniforme sur $[\frac{1}{2}, 1]$.

$$\text{Or } W \sim \mathcal{U}_{[1,2]} \iff P(W \leq w) = \begin{cases} 0 & \text{si} \quad w < 1 \\ w - 1 & \text{si} \quad 1 \leq w \leq 2 \\ 1 & \text{si} \quad w > 2 \end{cases} \tag{1.1}$$

La densité f_W de W est donc donnée par $f_W(w) = \frac{\partial}{\partial w} P(W \leq w) = \begin{cases} 0 & \text{si} \quad w \notin [1,2] \\ 1 & \text{si} \quad w \in [1,2] \end{cases}$

FIGURE 1.1.: Densité de $W \sim \mathcal{U}_{[1,2]}$

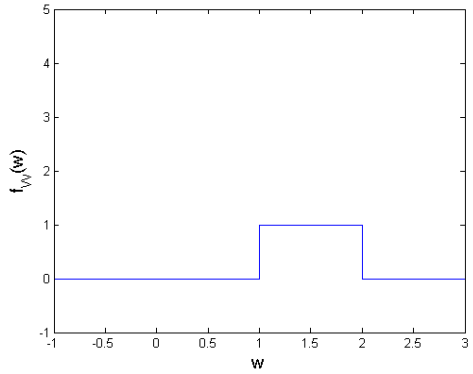
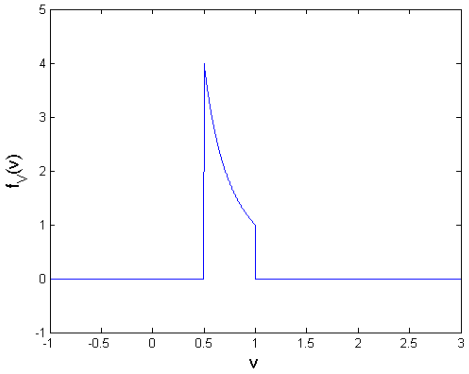


FIGURE 1.2.: Densité de $V = \frac{1}{W}$



De plus, si on avait $V \sim \mathcal{U}_{[\frac{1}{2},1]}$, on aurait $P(V \leq v) = \begin{cases} 0 & \text{si} \quad v < \frac{1}{2} \\ 2v-1 & \text{si} \quad \frac{1}{2} \leq v \leq 1 \\ 1 & \text{si} \quad v > 1 \end{cases}$

$$\begin{aligned} \text{Or on a } P(V \leq v) &= P(\frac{1}{W} \leq v) = P(W \geq \frac{1}{v}) = 1 - P(W < \frac{1}{v}) \\ &= \begin{cases} 1 & \text{si} \quad \frac{1}{v} < 1 \\ 2 - \frac{1}{v} & \text{si} \quad 1 \leq \frac{1}{v} \leq 2 \\ 0 & \text{si} \quad \frac{1}{v} > 2 \end{cases} \\ &= \begin{cases} 1 & \text{si} \quad v > 1 \\ 2 - \frac{1}{v} & \text{si} \quad \frac{1}{2} \leq v \leq 1 \\ 0 & \text{si} \quad v < \frac{1}{2} \end{cases} \neq \begin{cases} 0 & \text{si} \quad v < \frac{1}{2} \\ v-1 & \text{si} \quad \frac{1}{2} \leq v \leq 1 \\ 1 & \text{si} \quad v > 1 \end{cases} \end{aligned}$$

V ne suit donc pas une loi uniforme sur $[\frac{1}{2}, 1]$ et sa densité f_V est donc donnée par $f_V(v) = \frac{\partial}{\partial v} P(V \leq v) = \begin{cases} 0 & \text{si} \quad w \notin [\frac{1}{2}, 1] \\ \frac{1}{v^2} & \text{si} \quad w \in [\frac{1}{2}, 1] \end{cases}$

On conclut donc que si W suit une loi de probabilité uniforme, il en est autrement de son inverse (ou pour tout autre transformation non-affine). Pourtant on est tout aussi ignorant quant à la valeur de W que quant à celle de $\frac{1}{W}$.

Un autre inconvénient de l'utilisation de la loi probabilité uniforme pour modéliser l'ignorance est qu'il devient alors difficile de distinguer les cas où on n'a aucun information des cas d'équiprobabilité fréquentiste (résultant d'un certain nombre d'expériences) pour lesquels cette équiprobabilité constitue une information forte.

Un des principaux aspects que défend ce manuscrit est que la vraie ignorance quant à la valeur d'une variable W ne peut qu'être modélisée de façon imprécise par : $\forall w \in \Omega_W$, $P(W = w) \in [0,1]$. On obtient alors des probabilités inférieures et supérieures (valant respectivement 0 et 1 ici).

Alors que l'origine de la statistique Bayésienne remonte aux travaux de Bayes et Laplace au $18^{ème}$ siècle, ce ne fut qu'à la moitié du $20^{ème}$ siècle qu'elle fut revisitée et développée par De Finetti, Savage, Lindley ainsi que de très nombreux statisticiens.

La statistique Bayésienne mit cependant un certain temps avant d'être acceptée par une partie de la communauté statistique souvent qualifiée de *fréquentiste* par opposition aux Bayésiens [29, chapter 4], [24, chapter 3], [32, chapter 2], [33], [34, pages 18-26].

1.2.2. Théorie des possibilités

Les théories présentées dans la suite de ce chapitre manipulent des mesures non-additives(en général). Cette non-satisfaction du troisième axiome de Kolmogorov rend obsolète de nombreux raisonnements et règles de calcul très utilisés par les *probabilistes classiques*. Il convient alors de se poser la question de savoir si il vaut mieux ignorer l'incapacité des probabilités à modéliser le manque de connaissance et les *contraindre* arbitrairement à être additives, ou au contraire mettre en évidence ce caractère non-additif, et ainsi modéliser l'incertitude de manière plus fidèle.

Comme Bernoulli (ou Boole, Keynes et un certain nombre d'autres scientifiques) pensait les probabilités en terme de degrés de confiance à partir d'informations contenues dans des propositions (ou témoignages), les notions de *possibilité* et de *certitude* relatives à l'occurrence d'événements furent utilisées dans ses travaux [13]. La quantification de ces deux notions implique alors deux mesures d'incertitude inférieures et supérieures $\underline{\mu}$ et $\overline{\mu}$.

En effet, même si la météo du lendemain est par essence incertaine, si d'après l'information dont dispose un agent (épais nuages de pluie intensives recouvrant une grande région autour de Stockholm depuis trois semaines), il sait qu'il ne fera pas

beau le lendemain à Stockholm, on peut alors dire que l'événement *"il pleuvra demain à Stockholm"* est *possible* et l'événement *"il ne fera pas plein soleil à Stockholm demain"* est *certain*.

On peut par ailleurs convenir que le complémentaire d'un événement *certain*, i.e un événement *impossible*, est *certainement* impossible. Cette dualité entre les notions de *certitude* et de *possibilité* s'exprime de la manière suivante :

$$\forall A \subseteq \Omega, \quad \underline{\mu}(A) = 1 - \overline{\mu}(\overline{A}) \tag{1.2}$$

$$\forall A \subseteq \Omega, \quad \underline{\mu}(A) = 1 \iff \overline{\mu}(\overline{A}) = 0 \tag{1.3}$$

1.3 exprime le fait que si A est certain, alors $\overline{A}$ est impossible.

La théorie des possibilités, proposée par Zadeh en 1978 [35] puis largement développée par Dubois et Prade [36], vise à généraliser ces deux notions à un ensemble de degrés, un événement sera ainsi plus ou moins possible (resp. plus ou moins certain ou nécessaire).

Définition 1. Une *mesure de nécessité* (ou de certitude) N (correspondant à $\underline{\mu}$) sur Ω est une application de 2^Ω dans $[0,1]$ définie par les deux conditions :

$$N(\Omega) = 1, \qquad N(\emptyset) = 0 \tag{1.4}$$

$$N(A \cap B) = \min(N(A), N(B)) \qquad \forall A, B \subseteq \Omega \tag{1.5}$$

Définition 2. Une *mesure de possibilité* Π (correspondant à $\overline{\mu}$) sur Ω est une application de 2^Ω dans $[0,1]$ définie par les deux conditions :

$$\Pi(\Omega) = 1, \qquad \Pi(\emptyset) = 0 \tag{1.6}$$

$$\Pi(A \cup B) = \max(\Pi(A), \Pi(B)) \qquad \forall A, B \subseteq \Omega \tag{1.7}$$

Pour une mesure de possibilité donnée Π, en notant $\overline{A}$ le complémentaire de A dans Ω, la fonction N_Π définie de 2^Ω dans $[0,1]$ par

$$N_\Pi(A) = 1 - \Pi(\overline{A}) \tag{1.8}$$

est une mesure de nécessité appelée *mesure de nécessité duale de* Π

Lemme 1. $\Pi(A) \geq N_\Pi(A) \quad \forall A \subseteq \Omega$

On dit alors que toute mesure de possibilité Π *domine* sa mesure de nécessité duale.

Définition 3. Une *distribution de possibilité* $\pi : \Omega \to [0,1]$ est la restriction d'une mesure de possibilité aux singletons de Ω :

$$\pi(x) = \Pi(\{x\}) \quad \forall x \in \Omega$$

De la même manière, on peut définir une mesure de possibilité Π (resp. de nécessité) à partir d'une distribution de possibilité π :

$$\forall A \subseteq \Omega, \quad \Pi(A) = \max_{x \in A} \pi(x) \tag{1.9}$$

Une particularité importante des mesures de possibilité et de nécessité est qu'elles ne sont pas additives. En effet, ce résultat est directement obtenu pour Π par (1.7) et pour N par le lemme 1 et par l'équation (1.7).

Ces deux mesures ne sont donc pas additives et modélisent des incertitudes qui sont, tout comme celles utilisées par Bernoulli dans [13], de nature épistémique (c'est en fonction des connaissances que l'on a relativement à un événement qu'on peut lui attribuer des degrés de possibilité ou de nécessité). De nombreuses opérations de combinaisons, marginalisation, extensions sont alors proposées dans le cadre de la théorie des possibilités. Nous ne les détaillons pas car ce travail utilise principalement une autre théorie plus générale que celle des possibilités (la théorie des fonctions de croyance) mais des compléments peuvent être trouvés dans [36].

Lemme 2. *Si une mesure de possibilité* Π *sur* Ω *domine une mesure de probabilité* P *sur* Ω*, la mesure de nécessité* N_Π *duale de* Π *est alors dominée par* P

Il peut donc être intéressant, pour une mesure de possibilité Π donnée, de s'intéresser à l'ensemble des mesures de probabilité dominées par Π (et donc dominant N_Π), cet ensemble sera noté

$$\mathscr{P}(\Pi) = \{P : \forall A \subseteq \Omega \, P(A) \leq \Pi(A)\} \tag{1.10}$$

Soient P^* et P_* les bornes de $\mathscr{P}(\Pi)$ définies par

$$P^*(A) = \max_{P \in \mathscr{P}(\Pi)} P(A) \quad \forall A \subseteq \Omega \tag{1.11}$$

$$P_*(A) = \min_{P \in \mathscr{P}(\Pi)} P(A) \quad \forall A \subseteq \Omega \tag{1.12}$$

Lemme 3. *Soient Π et N_Π deux mesures duales de possibilité et de nécessité et $\mathscr{P}(\Pi)$ l'ensemble de probabilités associé, alors*

- $P^* = \Pi$
- $P_* = N$

Il vient alors que P^* et P_*, elles non-plus, ne sont pas additives (car ni Π ni N ne le sont).

Lien avec la *théorie des sous-ensembles flous*

En réalité, la théorie des possibilités fut proposée par Zadeh [35] en tant qu'interprétation d'un ensemble flou comme modèle d'incertitude. L'idée principale de la *théorie des sous-ensembles flous*, aussi proposée par Zadeh [37], est de graduer l'appartenance d'éléments aux ensembles à l'aide de fonctions d'appartenance

$$a : \Omega \to [0,1]$$

.

Soit $E \subseteq \Omega$ un ensemble classique, on peut alors modéliser E par sa fonction indicatrice $\mathbb{1}_E$ ($\mathbb{1}_E(w) = 1 \leftrightarrow w \in E$) , en effet connaître toutes les valeurs de $\mathbb{1}_E$ pour tout élément w de Ω est équivalent à connaître E (avec précision).

Si E est considéré comme un ensemble *flou*, il est alors modélisé par sa fonction d'appartenance a_E. $a_E(w) = 0.4$, par exemple, pourra être interprété comme une appartenance partielle de w à E. Dans ce travail, cette représentation permet d'exprimer la notion d'*imprécision* et donc d'une certaine forme d'*incertitude* (épistémique) de manière naturelle.

Dans ce cadre, relativement à l'ensemble flou E, la possibilité d'un événement $A \subseteq \Omega$ correspond alors au degré de compatibilité de A avec E, on a donc : $\Pi(A) = max\{a_E(w) : w \in A\}$.

La théorie des possibilités (avec la théorie des sous-ensembles flous) permet donc

une modélisation de l'incertitude qui peut tenir compte d'une imprécision ensembliste, on peut l'interpréter comme un moyen de définir des mesures d'incertitude sur des espaces imprécis. Elle représente donc une extension de la théorie des ensembles.

Une autre approche est d'étendre la théorie des probabilités aux connaissances ensemblistes, c'est ce que fait la théorie des probabilités imprécises.

1.2.3. Théorie des probabilités imprécises

Les travaux de De Finetti et de Savage traitaient les probabilités comme totalement subjectives. De Finetti utilisait une approche *par paris* consistant à calculer la probabilité $P(A)$ d'un événement A comme le prix que quelqu'un serait prêt à payer pour participer à un pari lui faisant gagner une unité de monnaie en cas d'occurrence de l'événement A ou, de manière équivalente, comme le prix que cette même personne serait prête à demander pour vendre le même pari.

Le caractère subjectif des probabilités vient du fait que c'est cette personne qui définit les prix des paris selon son degré de confiance en la réalisation des événements. Il faut toutefois remarquer que le raisonnement de De Finetti est symétrique qu'il s'agisse d'achat ou de vente du pari. Savage proposa alors une axiomisation de ces probabilités subjectives définies par De Finetti [29].

Dans [5], Walley reprend une approche *par paris* similaire à celle de De Finetti mais en relâchant certains de ses axiomes, ce qui lui permet d'obtenir un écart entre les prix que quelqu'un serait prêt à payer pour acheter des paris et ceux auquel il serait prêt à les vendre. Il choisit ainsi de considérer le prix maximum $\underline{P}$ qu'une personne serait prête à payer pour un pari, et le prix minimum $\overline{P}$ auquel elle serait prête à le vendre. Il est évident que, les prix d'achat sont inférieurs aux prix de ventes des paris, on a donc ici $\overline{P} = \overline{\mu}$ et $\underline{P} = \underline{\mu}$. Une conséquence directe de cette dissymétrie fut que les probabilités subjectives $\underline{P}$ et $\overline{P}$ ainsi calculées ne sont plus forcément additives.

Une autre interprétation importantes des probabilités imprécises est l'approche par *ensembles crédaux* où un *ensemble crédal* $\mathscr{P}$ contient l'ensemble convexe de probabilités à valeurs entre $\underline{P}$ et $\overline{P}$:

$$\mathscr{P} = \{P : \forall A \in \Omega, \underline{P}(A) \leq P(A) \leq \overline{P}(A)\}$$

Différents critères sont définis par Walley relativement à ces ensembles de probabilités :

– $\underline{P}$ et $\overline{P}$ *évitent les pertes sûres* (i.e. absence d'opportunité d'arbitrage) si $\mathscr{P} \neq \emptyset$

– $\underline{P}$ et $\overline{P}$ sont *cohérentes* si $\forall A \subseteq \Omega$, $\underline{P}(A) = P_*(A)$ avec $P_*(A) = \inf_{P \in \mathscr{P}} P(A)$

Au lieu de considérer, une unique mesure de probabilité comme c'est le cas en théorie des probabilités, Walley préfère ainsi injecter de l'imprécision quand à la mesure de probabilité à utiliser. Un des arguments qu'il met en avant est que dans de nombreux cas, les informations disponibles ne nous permettent pas de choisir une unique mesure de probabilité. Dans le cas extrême d'absence d'information, i.e. d'ignorance totale, au lieu de l'utilisation habituelle de la probabilité uniforme (habituellement utilisée en théorie des probabilités pour modéliser l'ignorance totale), Walley propose de modéliser l'ignorance totale tout simplement par une probabilité inférieure $\underline{P}$ nulle et par une probabilité supérieure $\overline{P}$ égale à 1, il obtient donc l'ensemble crédal

$$\mathscr{P} = \{P : \forall A \in \Omega, 0 \leq P(A) \leq 1\}$$

Cette approche offre en outre des représentations géométriques attractives. Il est cependant à noter que, pour pouvoir modéliser tous les ensembles convexes de probabilités, Walley doit utiliser la notion de *prévision*, correspondant à l'espérance en théorie des probabilités classique.

Exemple de modélisation probabiliste imprécise :
le Modèle de Dirichlet Imprécis (*MDI*)

En statistique fréquentiste, on détermine habituellement la loi de probabilité d'un phénomène à partir des fréquences observées.

Soit W une variable aléatoire à valeur dans $\Omega = \{A, B, C\}$. Si lors d'une expérience, on observe deux réalisations de l'événement *"$W = A$"*, trois réalisations de *"$W = B$"* et cinq réalisations de *"$W = C$"*, on en déduit naturellement la loi de probabilité suivante :

- $P(A) = \frac{2}{10}$
- $P(B) = \frac{3}{10}$
- $P(C) = \frac{5}{10}$

Cette loi est cependant très spécifique au vu du faible nombre de réalisations de l'expérience. C'est dans ce contexte que Walley proposa un modèle probabiliste imprécis basé sur un raisonnement Bayésien [5].

En statistique Bayésienne, la loi de Dirichlet est généralement utilisée comme loi *à priori* dans le cas multinomial. Ce choix est justifié entre autre par le fait que cette loi possède la caractéristique de donner aussi une loi de Dirichlet comme loi *à posteriori* (le fait de rester dans une même classe de loi de probabilité lors d'une mise à jour des probabilités constitue un avantage très apprécié des praticiens).

Cependant, tout comme les fréquentistes, les Bayésiens aboutissent toujours à une unique loi de probabilité. C'est pour cette raison que Walley proposa d'utiliser un ensemble de lois de Dirichlet (une loi de Dirichlet possède deux paramètres : $s \in \mathbb{R}$ et $t \in [0,1]^{|\Omega|}$, il s'agit donc ici de choisir l'ensemble des lois de Dirichlet pour tous les paramètres t possibles avec s fixé) au lieu d'une seule.

L'inférence obtenue sur $P(\{w\})$ par ce choix donne, $\forall w \in \Omega$:

$$\underline{P}(\{w\}) = \frac{n_w}{n+s} \tag{1.13}$$

$$\overline{P}(\{w\}) = \frac{n_w + s}{n+s} \tag{1.14}$$

où n_w est le nombre de réalisations de l'événement "$W = w$", n est le nombre total d'observations et $s > 0$ est un hyper-paramètre (Walley suggère de prendre $s = 1$ ou $s = 2$).

Il est évident que la probabilité estimée de façon fréquentiste $P(\{w\}) = \frac{n_w}{n}$ se situe entre $\underline{P}$ et $\overline{P}$ et que plus le nombre d'observations n est grand, plus l'ensemble crédal $\mathscr{P} = \{P : \forall A \in \Omega, \underline{P}(A) \leq P(A) \leq \overline{P}(A)\}$ sera réduit et tendra vers P quand n tendra vers l'infini.

Pour l'exemple précédent, le MDI avec $s = 1$ donne :

- $\underline{P}(A) = \frac{2}{11} \quad \overline{P}(A) = \frac{3}{11}$
- $\underline{P}(B) = \frac{3}{11} \quad \overline{P}(B) = \frac{4}{11}$
- $\underline{P}(C) = \frac{5}{11} \quad \overline{P}(C) = \frac{6}{11}$

Ce modèle, comme la plupart des modèles probabilistes imprécis, suggère que moins on a d'information disponible, plus on est imprécis dans nos estimations (plus $\mathscr{P}$ est large). Cet aspect est justement un avantage selon Walley qui estime qu'en théorie de la décision, il y a des cas où on ne devrait pas aboutir à une décision précise mais plutôt à un ensemble de décisions, surtout dans le cas où on dispose de peu d'information.

Lien avec la théorie des possibilités

En théorie des probabilités imprécises, l'incertitude est donc quantifiée à l'aide de deux mesures duales ($\underline{P}$ et $\overline{P}$), tout comme en théorie des possibilités (avec Π et N_Π). D'après (1.10), (1.11) et (1.12), il est donc envisageable de considérer les mesures de nécessité et de possibilité comme les bornes inférieures et supérieures d'un ensemble de probabilités, la théorie des possibilités est alors un cas particulier de la théorie des probabilités imprécises défini par la structure ensembliste (floue) considérée.

Bilan

La théorie des probabilités imprécises permet donc le traitement de probabilités épistémiques, qui ne sont alors pas nécessairement additives. Elle travaille avec des ensembles convexes de probabilités ou des modèles équivalents. Il ressort néanmoins des applications que son cadre étant relativement large (en pratique il n'est pas toujours évident de calculer les bornes de cet ensemble de probabilités), d'autres théories connexes à celle-ci peuvent éventuellement être privilégiées de manière à pouvoir réaliser différentes opérations plus aisément. La théorie des possibilités en fait partie, mais présente cependant un inconvénient : son spectre d'application est quant à elle assez réduit, ne capturant pas les probabilités classiques. La théorie des fonctions de croyance, se situant à un niveau intermédiaire entre ces

deux dernières théories, permet aussi un traitement de l'incertitude épistémique et offre un bon compromis entre généralité et applicabilité.

CHAPITRE 2

Théorie des fonctions de croyance

Dans cette partie, après un bref rappel historique, les principales notions de la théorie des fonctions de croyance sont ici présentées. Le *Modèle des Croyances Transférables* de Smets est ensuite brièvement passé en revue avec un exemple de modèle de prédiction crédibiliste. Enfin, quelques points sont discutés concernant certains aspects philosophiques de cette théorie.

Sommaire

2.1. Généralités

Le principal reproche que les fréquentistes font généralement aux Bayésiens est l'attribution d'une loi de probabilité *à priori* aux paramètres inconnus de leur modèles. Ils ne remettent cependant pas en cause le raisonnement de mise à jour des probabilités conditionnelles. Un autre type d'inférence statistique propose d'ailleurs une manière de procéder à cette mise à jour sans loi *à priori*, il s'agit de l'inférence fiduciaire proposée par Fisher au début du $XX^{ième}$ siècle [38]. Outre le débat entre fréquentistes et Bayésiens, un autre débat existe à l'intérieur de la communauté statistique, il oppose les Bayésiens aux défenseurs de l'*inférence fiduciaire*. En effet l'inférence fiduciaire a la particularité d'aboutir à des distributions (dites *fiduciaires*) *non-additives*, qui ne peuvent donc être considérées comme des mesure de probabilité *usuelles* des inférences fréquentistes ou bayésiennes [39].

C'est dans ce contexte que Dempster fit en 1967 une première présentation de la théorie des fonctions de croyance [40] avec comme objectif entre autre, de réconcilier inférences Bayésienne et fiduciaire [41]. Cette théorie fit donc d'abord son apparition dans un contexte purement statistique. Il proposa ainsi d'inférer des probabilités inférieures et supérieures à partir de données statistiques. Ces probabilités ayant la caractéristique notoires de ne pas être additives.

En 1976, Shafer formalisa et étendit cette théorie à une multitude de raisonnements et d'opérations sur différentes incertitudes épistémiques [42]. La théorie est ainsi appelée *théorie des fonctions de croyance* ou *théorie de Demster-Shafer* ou encore *théorie de l'évidence* (*evidence theory* en anglais). Dans la vision de Shafer, les fonctions de croyance sont interprétées comme des fonctions mesurant différents niveaux de jugements ou de croyance à partir d'éléments d'information (d'où le terme anglais "*evidence*") pouvant éventuellement être eux-mêmes incertains (imprécis ou peu fiables par exemple).

En réalité, même si les travaux initiaux de Dempster étaient totalement inscrits dans un cadre statistique (avec des modèles sous-jacents probabilistes donc), ce furent les travaux de Shafer, vulgarisant et étendant ceux de Dempster, qui permirent le vrai essor de la théorie qui se développe désormais dans d'autres domaines

que la statistique, tels que la fusion d'information ou l'intelligence artificielle. Progressivement, la communauté des fonctions de croyance s'est plutôt concentrée sur ces aspects non-statistiques. Néanmoins, les travaux de Dempster et Shafer, étant complémentaires, permettent d'appliquer la théorie aux incertitudes aléatoires statistiques et épistémiques.

Au début des années 90 Smets proposa le Modèle des Croyances Transférables (MCT), un modèle *crédibiliste* (i.e. de fonctions de croyance) où la notion de probabilité est totalement absente et où les incertitudes sont plutôt épistémiques.

La théorie des fonctions de croyance peut également être interprétée comme un cas particulier de *probabilités imprécises*, mais cette vision n'est pas consensuelle. Ce qui est sûr, c'est que les théories des possibilités et des probabilités peuvent être vues comme des cas particuliers de cette théorie qui est donc plus générale, capturant entre autre la logique ensembliste d'une manière élégante.

Dans ce manuscrit, nous travaillons dans un contexte de prédiction à partir de données incertaines, nous avons donc affaire à des incertitudes aléatoires (problème statistique de prédiction) mais aussi à des incertitudes épistémiques (incertitude sur les données). De manière à modéliser ces différents types d'incertitude de manière spécifique nous avons choisi la théorie des fonctions de croyance comme cadre de modélisation.

2.2. Définitions et propriétés de base

Soit W une quantité incertaine à valeurs dans Ω_W qu'on appellera le *cadre de discernement* ou *univers*.

Définition 4. Une *fonction de masse* m^w relative aux réalisations de W est une fonction définie sur l'ensemble des sous-parties [1] de Ω_W noté 2^{Ω_W} et à valeurs dans $[0,1]$ vérifiant

$$\sum_{B \in 2^{\Omega_W}} m^w(B) = 1 \tag{2.1}$$

Dans ce manuscrit nous supposons que $m^w(\emptyset) = 0$, même si certains auteurs ne

1. Les expressions $B \subseteq \Omega_W$ et $B \in 2^{\Omega_W}$ sont équivalentes.

font pas cette hypothèse et utilisent le terme $m^w(\emptyset)$ appelé *conflit* pour mesurer les contradictions entre sources d'informations par exemple [43, 44].

Définition 5. A partir de m^w, deux fonctions mesurant l'incertitude sont définies : la fonction de *croyance* Bel^w et la fonction de *plausibilité* pl^w. Leurs expressions sont les suivantes :

$$\forall A \in 2^{\Omega_W} \qquad Bel^w(A) = \sum_{B \subseteq A} m^w(B) \tag{2.2}$$

$$pl^w(A) = \sum_{B \cap A \neq \emptyset} m^w(B) \tag{2.3}$$

Bel somme les masses des connaissances (B) qui impliquent A ($B \subseteq A$), tandis que Pl somme celles qui sont consistantes ($A \cap B \neq \emptyset$) avec A. Les fonctions de croyance et de plausibilité expriment donc de manière différente l'information contenue dans m^w, $Bel^w(A)$ représente le degré de croyance que l'on peut attribuer à A compte tenu des éléments d'information fournis pas m^w et pl^w est le niveau plausibilité de A toujours au vu des éléments d'information contenues dans m^w. En théorie des fonctions de croyances, les fonctions de croyance et de plausibilité joueront le rôle des deux mesures d'incertitudes retrouvées en théories des possibilités et des probabilités imprécises, on aura $Bel^w = \underline{\mu}^W$ et $pl^w = \overline{\mu}^W$. Bel et Pl vérifient ainsi les Propriétés 1 et 2. De part leur définition, on aura aussi la propriété suivante (pendant de 1.2) :

Propriété 3. $\forall A \subseteq \Omega$, $pl^w(A) = 1 - Bel^w(\overline{A})$

Il est possible de retrouver m^w à partir de Bel^w ou de pl^w via l'*inverse de Möbius* :

$$\forall A \in 2^{\Omega_W} \qquad m^w(A) = \sum_{\emptyset \neq B \subseteq A} (-1)^{|A|-|B|} Bel^w(B) \tag{2.4}$$

$$m^w(A) = \sum_{\emptyset \neq B \subseteq A} (-1)^{|A|-|B|+1} pl^w(\overline{B}) \tag{2.5}$$

Bel^w et pl^w étant donc reliées bijectivement à m^w, lorsqu'on parlera de *fonction de croyance* en générale, on fera indifféremment référence à m^w, Bel^w ou pl^w.

Définition 6. La fonction de *contour* $pl^w : \Omega_W \to [0,1]$ relative à m^w est définie par $\forall w \in \Omega_W$, $pl^w(w) = Pl(\{w\})$.

Les praticiens des fonctions de croyance commencent généralement par modéliser la fonction de masse m^w, qui représente l'incertitude de la source d'information, pour ensuite s'intéresser aux fonctions de croyance et de plausibilité, cependant d'après (2.4) et (2.5) il est possible de commencer par modéliser Bel^w pour ensuite

travailler sur m^w ou pl^w par exemple.

Exemple 1. *Un fabricant de smartphone, lors d'une étude marketing, désire évaluer les proportions hommes/femmes au sein de ses clients.*

On posera W = sexe des clients, et donc $\Omega_W = \{F, H\}$.

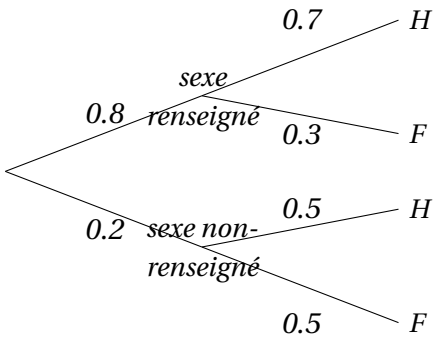
Il se réfère pour cela à la base de données clients dont il dispose, cependant il sait que certains clients ne remplissent pas tous les champs du formulaire d'achat. Malgré tout, il remarque que parmi ceux qui remplissent au moins la case "sexe" (ces derniers représentent 80% *des clients), il y a* 70% *d'hommes et* 30% *de femmes.*

On remarque ici que même si son objectif final est lié à une incertitude aléatoire (l'ensemble de ses clients est une population), il a une incertitude épistémique relative aux clients dont il ignore le sexe.

Dans un cadre probabiliste classique, il pourrait supposer que ses clients ayant renseigné leur sexe étant assez nombreux, il peut alors étendre leur proportions à l'ensemble de ses clients, il obtient donc

$$\begin{cases} P_1(H) &= 0.7 \\ P_1(F) &= 0.3 \end{cases}$$

S'il est toutefois soucieux de prendre en compte son ignorance (épistémique) concernant les autres clients, il pourrait alors supposer qu'étant totalement ignorant concernant leur sexe, il peut partir du principe qu'ils contiennent autant d'hommes que de femmes (ceci revient alors utiliser une distribution uniforme sur ces derniers même si ils peuvent très bien avoir une distribution non-uniforme en réalité), il obtient alors



$$\rightarrow \begin{cases} P_2(H) = 0.8 \times 0.7 + 0.2 \times 0.5 = 0.66 \\ P_2(F) = 0.8 \times 0.3 + 0.2 \times 0.5 = 0.34 \end{cases}$$

Si toutefois nous ne faisons pas cette hypothèse d'uniformité, l'incertitude du fabricant relative à W peut se modéliser par la fonction de masse suivante :

$$\begin{cases} m^w(\{F\}) = 0.3 \\ m^w(\{H\}) = 0.5 \\ m^w(\{F,H\}) = 0.2 \end{cases}$$

On peut alors calculer les fonctions de croyance et de plausibilité comme suit :

$$\begin{cases} Bel^w(\{F\}) = m(\{F\}) = 0.3 \\ Bel^w(\{H\}) = m(\{H\}) = 0.5 \end{cases}$$

et

$$\begin{cases} pl^w(\{F\}) = m(\{F\}) + m(\{F,H\}) = 0.5 \\ pl^w(\{H\}) = m(\{H\}) + m(\{F,H\}) = 0.7 \end{cases}$$

Définition 7. Un ensemble $A \in 2^{\Omega_W}$ est appelé *élément focal* de m^w si $m^w(A) > 0$

Définition 8. La *fonction de masse vide* n'a que le cadre de discernement comme élément focal, elle vérifie donc $m^w(\Omega_W) = 1$ et modélise l'ignorance totale.

On a ici une modélisation intéressante de l'ignorance. En effet, contrairement à sa modélisation probabiliste par une distribution uniforme, la fonction de masse vide est bien moins spécifique, un des objectifs de cette thèse est d'ailleurs de montrer que la modélisation crédibiliste de l'ignorance (éventuellement partielle) peut permettre une plus grande efficacité de prédiction que l'uniformité des probabilités.

Définition 9. Une fonction de masse m^w est dite *consonante* si ses ensemble focaux $A_1, \ldots, A_r$ sont emboîtés (i.e. $A_1 \subset \cdots \subset A_r$)

Propriété 4. *Si une fonction de masse est consonante, les fonctions de croyance et de plausibilité qui en résultent sont alors des mesures de nécessité et de possibilité.*

Cette dernière propriété découle directement des Equations (1.7), (1.5), (2.2) et (2.3). La théorie des possibilités est donc bien un cas particulier de la théorie des fonctions de croyance [42].

Propriété 5. *Si tous les éléments focaux d'une masse m^w sont des singletons, m^w est alors une mesure de probabilité et est donc additive, on a alors $Bel^w = pl^w$*

Par définition, Bel^w et pl^w ne sont pas nécessairement additives au sens où $\forall A, B \in 2^{\Omega_W}$ on a juste

$$Bel^w(A \cup B) \geq Bel^w(A) + Bel^w(B) - Bel^w(A \cap B)$$

Plus généralement on a la propriété de monotonie suivante :

Propriété 6. $Bel^w(\bigcup_{i=1}^{k} A_i) \geq \sum_{I \subseteq \{1,\ldots,k\}} (-1)^{|I|+1} Bel^w(\bigcap_{i \in I} A_i)$

Dans de nombreux problèmes, les croyances des différents agents se font à différents niveaux. Des opérations d'extension ou de projection de fonctions de croyance entre espaces sont alors nécessaires ; Soient $X \in \Omega_X$ et $Y \in \Omega_Y$ deux variables.

Définition 10. On appelle *extension cylindrique* de m^X (relative à l'espace Ω_X) sur l'espace produit $\Omega_X \times \Omega_Y$ la masse de croyance $m^{X\uparrow X,Y}$ définie par :

$$m^{X\uparrow X,Y}(A) = \begin{cases} m^X(B) & \text{si} \quad A = B \times \Omega_Y \\ 0 & \text{sinon} \end{cases} , \forall A \subseteq \Omega_X \times \Omega_Y$$

Définition 11. La projection $m^{X,Y\downarrow X}$ d'une masse de croyance $m^{X,Y}$ (relative à l'espace produit $\Omega_X \times \Omega_Y$) est définie par :

$$m^{X,Y\downarrow X}(B) = \sum_{A\downarrow\Omega_X = B} m^{X,Y}(A) \ , \forall B \subseteq \Omega_X$$

où $A \downarrow \Omega_X$ est la projection de $A \subseteq \Omega_X \times \Omega_Y$ sur Ω_X.

Définition 12. Deux variables $W_1 \in \Omega_1$ and $W_2 \in \Omega_2$ sont dites *cognitivement indépendantes* si $\forall (A \times B) \subseteq \Omega_1 \times \Omega_2,\ Pl^{W_1W_2}(A,B) = Pl^{W_1}(A) Pl^{W_1}(B)$

Cette notion, présentée dans [42], signifie que tout nouvel élément d'information concernant une de ces deux variables n'affecte pas notre connaissance relative à l'autre variable.

Combinaison de masses

Dans les applications, l'information peut provenir de plusieurs sources distinctes. Il est alors nécessaire de combiner ces informations entre elles. Suivant que les informations de deux sources sont conflictuelles ou non, deux attitudes sont alors envisageables pour les combiner, on parlera de combinaison *conjonctive* si on part du principe que les deux sources sont fiables et ne peuvent donc pas se contre-dire, alors que pour une combinaison *disjonctive* on suppose simplement qu'au moins une des deux sources est exacte. En pratique, suivant les cas, on choisira le type de combinaison le plus adapté.

Dans ces travaux, nous supposons les différentes sources d'information indépen-

dantes entre elles.

Cette dualité en 'ET / OU' s'illustre donc par deux grandes règles de combinaison génériques. Une multitude de règles intermédiaires est néanmoins offerte par le grand nombre de travaux de *fusion d'information* réalisés dans le cadre des fonctions de croyance [45], [46], [47], [48]. Voici donc l'expression de ces règles de combinaison conjonctives et disjonctives :

Soient m_1^W et m_2^W les deux masses à combiner et $A \subseteq \Omega_W$ un événement

$$\textit{régle conjonctive} \quad m_1^W \textcircled{\cap} m_2^W(A) = \frac{1}{1-\kappa} \sum_{B \cap C = A} m_1^W(B) m_2^W(C) \tag{2.6}$$

$$\text{avec } \kappa = \sum_{B \cap C = \emptyset} m_1^W(B) m_2^W(C)$$

$$\textit{régle disjonctive} \quad m_1^W \textcircled{\cup} m_2^W(A) = \sum_{B \cup C = A} m_1^W(B) m_2^W(C) \tag{2.7}$$

Propriété 7. *Si P est une probabilité et m^w une fonction de masse quelconque, alors $P \textcircled{\cap} m^w$ est une probabilité.*

Propriété 8. *En notant $pl^w_{1\textcircled{\cap}2}$ la fonction de plausibilité calculée à partir de $m_1 \textcircled{\cap} m_2$, on a*

$\forall w \in \Omega,\ pl^w_{1\textcircled{\cap}2}(\{w\}) = \frac{pl_1^w(\{w\}) pl_1^w(\{w\})}{1-\kappa}$

Exemple 2 (suite de l'Exemple 1)**.** *Si le fabricant de smartphone décide finalement de représenter l'incertitude liée à W en combinant m^w avec les dires d'un de ses employés qui affirme que, contrairement aux apparences, $2/3$ des clients sont des femmes. La masse modélisant l'opinion de l'employé se modélise par :* $\begin{cases} m^w_{employé}(\{F\}) = 2/3 \\ m^w_{employé}(\{H\}) = 1/3 \end{cases}$ *On peut remarquer ici, que $m^w_{employé}$ est une probabilité (n'ayant pour éléments focaux que des singletons).*

La combinaison conjonctive de m^w et de $m^w_{employé}$ donne donc :

$$\begin{aligned}
\kappa &= m^w(\{F\})m^w_{employé}(\{H\} + m^w(\{H\})m^w_{employé}(\{F\} \\
&= 0.3 \times 1/3 + 0.5 \times 2/3 = \frac{1.3}{3} \\
m^w \textcircled{\cap} m^w_{employé}(\{F\}) &= \frac{1}{1-\kappa}[m^w(\{F\})m^w_{employé}(\{F\} + m^w(\{F,H\})m^w_{employé}(\{F\}] \\
&= \frac{3}{1.7}[0.3 \times \frac{2}{3} + 0.2 \times \frac{2}{3}] \\
&= \frac{10}{17} \\
m^w \textcircled{\cap} m^w_{employé}(\{H\}) &= \frac{1}{1-\kappa}[m^w(\{H\})m^w_{employé}(\{H\} + m^w(\{F,H\})m^w_{employé}(\{H\}] \\
&= \frac{3}{1.7}[0.5 \times \frac{1}{3} + 0.2 \times \frac{1}{3}] \\
&= \frac{7}{17}
\end{aligned}$$

On peut remarquer que $m^w \textcircled{\cap} m^w_{employé}$ est bien une probabilité.

2.3. Interprétations

Il existe différentes interprétations des fonctions de croyance. Nous ne revoyons ici que les deux principales.

2.3.1. Fonction multi-valuée

Outre l'interprétation de Bel^w et de pl^w comme niveaux différents d'incertitude obtenus à partir de m^w, Bel^w et pl^w peuvent aussi être vues comme les limites inférieures et supérieures de l'ensemble des probabilités de A résultant de la prise en compte de l'incertitude inhérente à m^w. Plus m^w contiendra un grand niveau d'incertitude, plus cet ensemble sera large.

Une autre interprétation possible des fonctions de croyance est de les définir à partir d'une *fonction multivaluée* $\Gamma : T \mapsto 2^{\Omega_W} \setminus \{\emptyset\}$ et d'une mesure de probabilité P sur T où T est un espace probabilisable quelconque.

Le terme multivaluée signifie qu'à tout élément $\{t\} \in T$, correspond un ensemble d'éléments $\Gamma(t) \subseteq \Omega_W$ (donc un élément de 2^{Ω_W}).

On peut alors définir m^w, Bel^w et pl^w comme ceci :

$$\forall A \in 2^{\Omega_W} \quad \begin{aligned} m^w(A) &= P(\{w \in \Omega_W : \Gamma(w) = A\}) && (2.8)\\ Bel^w(A) &= P(\{w \in \Omega_W : \Gamma(w) \subseteq A\}) && (2.9)\\ pl^w(A) &= P(\{w \in \Omega_W : \Gamma(w) \cap A \neq \emptyset\}) && (2.10) \end{aligned}$$

Cette approche par fonction multivaluée est celle initialement présentée par Dempster dans sa première présentation de la théorie des fonctions de croyance [40]. Malgré le fait qu'elle aide à la compréhension ensembliste des fonctions de croyance, elle est cependant peu utilisée dans la littérature du fait de sa difficulté à être mise en pratique. Il n'est en effet pas aisé, dans les applications, de bien choisir l'espace T et la fonction multivaluée à considérer.

Ces deux approches sont de toute manière équivalente, une masse m^w (ou Bel^w ou pl^w) correspondant exactement à un espace T donné, une mesure de probabilité P sur T et une fonction multivaluée Γ de T dans Ω_W.

2.3.2. Modèle des Croyances Transférables (*MCT*)

En 1994, Smets présenta le Modèle des Croyances Transférables (MCT) [49] qui définit un cadre de travail bien formalisé pour raisonner et décider dans le cadre de la théorie des fonctions de croyance. En effet ce modèle comprend deux niveaux : le niveau *crédal* où les incertitudes sont modélisées et traitées, puis le niveau décisionnel où une décision est prise en fonction des résultats obtenus au niveau précédent.

La particularité majeure de ce modèle est qu'il est purement crédibiliste ou presque, en effet il ne fait intervenir aucune mesure de probabilité, sauf au niveau décisionnel où il propose d'en construire une, la *probabilité pignistique* à partir de fonctions de croyance construites, elles, au niveau crédal, de manière à prendre une décision (en fonction des probabilités pignistiques des événements ainsi calculées).

Même si cette probabilité pignistique construite au niveau décisionnel est aussi une des principales caractéristique de ce modèle, il faut noter que le MCT offre d'autres critères de décision tel que le *maximum de plausibilité* [50].

L'expression d'une probabilité pignistique P^W_{pig} construite à partir d'une masse de croyance m^w est la suivante :

Définition 13.

$$\forall w \in \Omega_W, \quad P_{pig}(\{w\}) = \sum_{A \subseteq \Omega_W : w \in A} \frac{m^w(A)}{|A|} \tag{2.11}$$

Cette probabilité pignistique vérifie dans tous les cas $Bel^w \leq P_{pig} \leq pl^w$, elle a cependant le défaut d'être obtenue en projetant uniformément les masses sur les singletons constituant les éléments focaux, ceci allant totalement à l'encontre de l'avantage de modélisation qu'offre les fonctions de croyance quant à l'ignorance. Elle a été proposée par pragmatisme en regard des cas où même si l'utilisation optimale des informations dont on dispose aboutit à une décision imprécise, des décisions précises doivent être prises (*prise de décision précise à partir d'informations partielles*). Il est à souligner que d'autres probabilités auraient pu être proposées pour un tel passage d'un univers tolérant l'imprécision à un univers de décision précise, la probabilité pignistique présente aussi l'avantage de correspondre à la valeur de Shapley [51] qui est une référence en théorie des jeux pour l'équité des gains notamment.

Approche de Dempster pour l'expérience de Bernoulli dans le *MCT*

L'expérience de Bernoulli consiste à répéter n tirages aléatoires et indépendamment distribués à deux issues possibles A et B, et à estimer la vraie loi de probabilité P^W de la variable $W \in \Omega^W = \{A, B\}$ représentant un tirage à partir des résultats $w = (w_1, \ldots, w_n) \in \{A, B\}^n$. L'approche statistique fréquentiste classique à ce problème est d'estimer donc le paramètre θ de la loi de Bernoulli $Ber(\theta)$ modélisant P^W par $P^W(A) = \theta$ en maximisant la vraisemblance $L(\theta; w)$. L'estimateur obtenu correspond à la fréquence observée $\frac{n_A}{n}$ de A dans w (n_A étant le nombre de A observé dans w).

Il peut paraître évident que plus n est grand, plus le modèle d'incertitude se doit d'être précis (dans ce cas les fréquences empiriques tendent vers la vraie probabilité P^W).

C'est dans ce même cadre que Smets proposa de construire une fonction de croyance prédictive m^w à partir des fréquences empiriques selon le modèle initial

de Dempster [52] adapté au TBM. Voici la fonction de croyance m^w obtenue :

$$\begin{cases} m^w(\{A\}) &= \frac{n_A}{n+1} \\ m^w(\{B\}) &= \frac{n_B}{n+1} \\ m^w(\{A,B\}) &= \frac{1}{n+1} \end{cases} \tag{2.12}$$

Cette masse de croyance exprime clairement une part d'ignorance (représentée par la valeur de $m^w(\{A,B\})$) diminuant avec la taille n de l'échantillon. Pour des explications concernant son obtention le lecteur sera renvoyé à [53, 52, 54].

Les fonctions de croyance et de plausibilité qui en découlent sont :

$$\begin{cases} Bel^w(\{A\}) &= \frac{n_A}{n+1} \\ Bel^w(\{B\}) &= \frac{n_B}{n+1} \\ Bel^w(\{A,B\}) &= 1 \end{cases} \qquad \begin{cases} pl^w(\{A\}) &= \frac{n_A+1}{n+1} \\ pl^w(\{B\}) &= \frac{n_B+1}{n+1} \\ pl^w(\{A,B\}) &= 1 \end{cases}$$

On a donc toujours $Bel^w \leq P^W_{fréq} \leq pl^w$. Il est de plus évident que lorsque n va grandir, les écarts entre Bel^w, $P^W_{fréq}$ et pl^w vont se réduire jusqu'à s'annuler lorsque n va tendre vers l'infini, on aura alors $Bel^w = P^W_{fréq} = pl^w$. Comme d'après le théorème centrale limite, $P^W_{fréq} \underset{n\to\infty}{\to} P^W$, Bel^w et pl^w tendront aussi vers la vraie mesure de probabilité P^W.

On peut remarquer la ressemblance existante entre ce modèle et le modèle de Dirichlet imprécis (en prenant le paramètre s du MDI égale à 1, les deux modèles sont équivalents). Il faut toutefois garder à l'esprit que ces deux modèles s'inscrivent dans des cadres de travail très différents : alors que le MDI est un modèle probabiliste imprécis (il utilise donc des mesures de probabilité), le modèle prédictif de Dempster adapté au MCT est défini sans la moindre mesure de probabilité (même si il utilise une distribution sous-jacente au problème).

2.4. Vraisemblance crédibiliste

Rappel : En théorie des probabilités, la (fonction de) vraisemblance du paramètre d'un modèle paramétrique $\{P_\theta : \theta \in \Theta\}$ relative à un échantillon statistique w, lui même réalisation d'une variable aléatoire discrète W, est la fonction $L(.; w)$ de Θ dans $[0, 1]$ définie par $L(\theta; w) = P_\theta(W = w)$. La recherche du maximum de vraisemblance fournit alors un estimateur naturel du paramètre θ. Les estimateurs du maximum de vraisemblance sont couramment utilisés en Statistique et possèdent de bonnes propriétés [55]. Cette vraisemblance (parfois qualifiée de *classique* dans ce manuscrit) n'est calculable que pour des données d'apprentissage *précises*.

Lorsque notre connaissance relative à l'échantillon w est *imparfaite* ou *incomplète*, au sens où la valeur de certaines données ou de certains paramètres de la loi de probabilité P_θ sous-jacente à w est inconnue, on n'est alors plus dans un cadre classique de données *précises*, on parlera alors de données *imprécises*. Dans ce cadre, Denoeux propose une explicitation de la vraisemblance dite *imprécise* [56] relativement à un tel échantillon connu de façon *imparfaite*. Cette vraisemblance peut être vue comme l'équivalent de la vraisemblance *incomplète* explicitée par Dempster [57]. Voici la définition de la vraisemblance *imprécise* :

Définition 14. La vraisemblance dite *imprécise* du paramètre θ d'un modèle $\{P_\theta : \theta \in \Theta\}$ relativement à un échantillon *imprécis* A (l'information issue de notre observation est de la forme $w \in A$) est définie par $L(\theta; A) = \sum_{w \in A} P_\theta(W = w)$

On peut remarquer une certaine ambiguité de langage car même si cette vraisemblance $L(\theta; A)$ se calcule à partir d'une information imprécise ($w \in A$), elle est structurellement *précise* car correspondant à un unique nombre (donc *précis*). Le terme de "*vraisemblance imprécise*" pourrait renvoyer à une vraisemblance impossible à calculer précisément et étant donc constituée d'un ensemble plutôt que d'un nombre.

Dans le cas où notre connaissance n'est plus seulement imprécise mais aussi incertaine et représentée par une fonction de croyance m^w, différents poids peuvent être affectés aux éléments imprécis A_i. Dans ce contexte d'incertitude des données, Denoeux [56] a récemment proposé de définir la vraisemblance crédibiliste suivante :
Définition 15. La vraisemblance *crédibiliste* du paramètre θ d'un modèle $\{P_\theta : \theta \in \Theta\}$

relativement à un échantillon *crédibiliste* représenté par m^w dont les éléments focaux sont $A_1, \ldots, A_z$ est définie par

$$L(\theta; m^w) = \sum_{i=1}^{z} m^w(A_i) L(\theta; A_i) \tag{2.13}$$

Cette vraisemblance est donc la somme des vraisemblances imprécises des éléments focaux de m^w pondérée par les masses de ces éléments focaux. Dans le cas de données w précises, on a $m^w(\{w\}) = 1$, ce qui donne

$L(\theta; m^w) = m^w(\{w\}) L(\theta; \{w\}) = L(\theta; w)$ et on retrouve donc bien la vraisemblance *classique*.

Il est à noter que d'autres choix de vraisemblances sont possibles dans le cadre des fonctions de croyance. En effet, d'un point de vue probabiliste imprécis par exemple, se ramener à une seule valeur de vraisemblance ne fait pas forcément sens avec des fonctions de croyances. Une alternative aurait pu être de travailler sur des ensembles de vraisemblances à maximiser. Cette alternative pose cependant des problèmes importants d'un point de vue calculatoire.

Dans [56], Denoeux montre le résultat suivant :

$$\begin{aligned} L((\theta; m^w) &= \sum_{i=1}^{z} m^w(A_i) L(\theta; A_i) \\ &= \sum_{i=1}^{z} m^w(A_i) \sum_{w \in A_i} P_\theta(W = w) \\ &= \sum_{w \in \Omega_W} P_\theta(W = w) \sum_{A_i \ni w} m^w(A_i) \\ &= \sum_{w \in \Omega_W} P_\theta(W = w) pl^w(\{w\}) \\ &= \mathbb{E}_\theta[pl^w(W)] \end{aligned} \tag{2.14}$$

L'expression de la vraisemblance crédibiliste ne fait donc intervenir la fonction de croyance m^w qu'à travers sa fonction de contour pl^w.

Proposition 1. *Pour toute fonction de croyance $m^w : 2^{\Omega_W} \to [0,1]$, nous avons*

$$L(\theta; m^w) = \Big(\sum_{u \in \Omega_W} pl^w(u) \Big) . L(\theta; \hat{m}^w)$$

où $\hat{m}^w$ *est la fonction de croyance (bayésienne) définie par* $\forall v \in \Omega_W, \quad \hat{m}^w(\{v\}) = \frac{pl^w(v)}{\sum_{u \in \Omega_W} pl^w(u)}$

Cette proposition est démontrée en Annexe B.

Il est intéressant de remarquer que la fonction de croyance $\hat{m}^w$ n'a pour éléments focaux que des singletons et qu'elle définit donc une distribution de probabilité (voir Propriété 5). Cette proposition montre bien que maximiser une vraisemblance crédibiliste $L(\theta; m^w)$ revient au même que maximiser la vraisemblance crédibiliste correspondante $L(\theta; \hat{m}^w)$. Cette dernière vraisemblance $L(\theta; \hat{m}^w)$ ne fait intervenir m^w qu'à travers sa fonction de contour pl^w. On peut aussi remarquer que maximiser une vraisemblance crédibiliste relativement à une fonction de croyance m^w est donc équivalent à maximiser une autre vraisemblance crédibiliste relativement à une autre fonction de croyance $\hat{m}^w$ qui est, elle, une distribution de probabilité (étant définie sur Ω_W et non sur 2^{Ω_H} elle n'a que des singletons comme éléments focaux).

2.4.1. Etude du comportement de l'estimateur du maximum de vraisemblance crédibiliste (*EMVC*) sur des exemples simples

Nous présentons ici une étude comparative des comportements des estimateurs du maximum de vraisemblance classiques (EMV) et crédibiliste ($EMVC$). Nous nous plaçons tout d'abord dans un cadre très simple d'échantillon binomial, i.e. à 2 classes, où la vraie proportion de chaque classe doit être estimée. L'échantillon est supposé incertain dans le sens où un niveau d'incertitude (épistémique) global ϵ des données est supposée ici connu. Dans ce cadre de travail, après avoir défini un modèle génératif (probabiliste) de données *d'un niveau de fiabilité* $1-\epsilon$ (i.e. d'un niveau d'incertitude de ϵ), nous calculons des EMV classiques et crédibilistes, avec et sans prise en compte de notre connaissance de ce niveau d'incertitude ϵ des données. Nous étudions ensuite certaines propriétés des estimateurs obtenus.

Nous présentons ensuite le comportement de l'$EMVC$ dans le cas d'un échantillon multinomial à 3 classes lorsque l'on fusionne 2 de ses classes. Nous montrons ainsi un comportement potentiellement surprenant qui aura d'importantes conséquences sur la méthodologie des *arbres* E^2M présentée ensuite en Chapitre 7.

Exemple du modèle paramétrique binomial

Le but, ici, est tout d'abord d'illustrer la mise en oeuvre d'estimation par maximum de vraisemblance crédibiliste dans le cas de données incertaines. Le choix crédibiliste de représentation de l'incertitude des données ϵ est comparé à ce qu'aurait donné un choix plus classique de cadre probabiliste. A travers cette comparaison des différentes vraisemblances, une comparaison est effectuée entre les représentations de l'ignorance dans les cadres de représentation de l'incertitude correspondants, i.e. par une *distribution uniforme* dans le cadre probabiliste et une *fonction de masse vide* (voir Définition 8) dans le cadre crédibiliste.

Après avoir défini un modèle génératif de *données incertaines* dont le niveau de fiabilité *globale* ϵ est supposé connu, les différentes modélisations sont présentées, les vraisemblances correspondantes à ces modélisations et les estimateurs les maximisant sont explicités et enfin certaines de leur propriétés asymptotiques sont étudiées.

Modèle génératif :

Soit $v = (v_1, \ldots, v_n)$ un échantillon *précis* mais non observable contenant n réalisations d'une variable aléatoire de Bernouilli $V \sim Ber(\theta_V)$ avec $\Omega_V = \{0,1\}$. On considère par ailleurs un échantillon, observable lui, $w = (w_1, \ldots, w_n)$ comprenant n réalisations de la variable aléatoire W définie par le modèle génératif suivant (de w à partir de v et de ϵ) :

$$W = (1 - B)V + BU \tag{2.15}$$

où

$$\begin{cases} V \sim Ber(\theta_V) \\ B \sim Ber(\epsilon) \\ U \sim Ber(0.5) \quad \text{avec } \Omega_U = \{0,1\} \\ V, B \text{ et } U \text{ indépendantes} \end{cases}$$

Ce modèle génératif est une interprétation possible de la connaissance du niveau ϵ d'incertitude des données. A l'aide de ce modèle, l'objectif est d'estimer θ_V à partir de w et de ϵ. En remarquant que lorsque $B = 0$ on a $W = V$, on peut interpréter le modèle (2.15) comme un modèle probabiliste de fiabilité des données. En effet, ϵ peut ici être interprété comme le niveau de *fiabilité globale* de l'échantillon w. Dès

que $B = 1$, V est aléatoirement remplacé.

Modélisations de l'incertitude des données et interprétations :

Quatre modélisations différentes sont ici proposées, deux dans le cadre probabiliste et deux dans le cadre crédibiliste. Chacune d'entre elle correspond à une vraisemblance différente, on obtiendra donc quatre estimateurs $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ et $\hat{\theta}_4$ en les maximisant.

– **Cadre probabiliste :**

Nous proposons ici d'estimer θ_V par maximisation de sa vraisemblance $L(\theta; w)$ relativement à w. Cette vraisemblance est d'abord calculée sans prise en compte ni de notre connaissance du niveau d'incertitude des données, ni donc du modèle génératif 2.15 des données mais en supposant que w est la réalisation d'une variable de Bernoulli de paramètre θ_V. On notera cette vraisemblance $L_0(\theta; w)$.
Le paramètre θ_V est ensuite estimé par maximisation de sa vraisemblance $L(\theta; w)$ toujours calculée relativement à w mais cette fois avec prise en compte du modèle génératif des données 2.15 (et donc d'ϵ).
Dans ce modèle, on suppose qu'une proportion ϵ des données a préalablement pu être remplacée par une valeur quelconque de $\Omega_W = \{1, 0\}$. Dans ce modèle, l'*ignorance* est interprétée de manière *uniforme* : on estime que lorsqu'on ne sait pas, toutes les possibilités ont le même poids.

– **Cadre crédibiliste :**

A partir de cette connaissance de w et ϵ, nous allons construire deux échantillons crédibilistes m^v et $\tilde{m}^v$ définis sur 2^{Ω_V}. Ces masses de croyance multidimensionnelles m^v et $\tilde{m}^v$ tiendront compte de la fiabilité globale de w de manière à exprimer notre incertitude relative à v.

$$\text{On notera : } m^v = \begin{pmatrix} m_1^v \\ \vdots \\ m_n^v \end{pmatrix} = \begin{pmatrix} m_1^v(\{1\}) & m_1^v(\{0\}) & m_1^v(\{1,0\}) \\ \vdots & \vdots & \vdots \\ m_n^v(\{1\}) & m_n^v(\{0\}) & m_n^v(\{1,0\}) \end{pmatrix}$$

$$\text{et } \tilde{m}^v = \begin{pmatrix} \tilde{m}_1^v \\ \vdots \\ \tilde{m}_n^v \end{pmatrix} = \begin{pmatrix} \tilde{m}_1^v(\{1\}) & \tilde{m}_1^v(\{0\}) & \tilde{m}_1^v(\{1,0\}) \\ \vdots & \vdots & \vdots \\ \tilde{m}_n^v(\{1\}) & \tilde{m}_n^v(\{0\}) & \tilde{m}_n^v(\{1,0\}) \end{pmatrix}$$

Dans le cas de m^v, de manière à exprimer notre connaissance de la fiabilité globale ϵ de w, pour chaque exemple i, on attribuera la masse $1-\epsilon$ à l'ensemble focal $\{w_i\}$ et ϵ à l'ignorance (donc à l'ensemble focal $\{1,0\}$).

$$\text{On aura donc } \forall i = 1,\dots,n, \quad \begin{cases} m_i^v(\{1\}) &= (1-\epsilon)1_{w_i=1} \\ m_i^v(\{0\}) &= (1-\epsilon)1_{w_i=0} \\ m_i^v(\{1,0\}) &= \epsilon \end{cases}$$

L'*ignorance* est ici modélisée par le terme $m^w(\Omega_W)$ et interprétée de manière moins engagée que la manière uniforme. En effet, on estime ici que lorsqu'on ne sait pas, *tout* est possible, nous n'avons pas de raisons de privilégier certaines possibilités, mais nous n'avons pas de raison non plus de les traiter spécifiquement de façon *égalitaire*.

Dans le cas de $\tilde{m}^v$, on attribuera, pour chaque exemple i, $1-\frac{\epsilon}{2}$ à l'ensemble focal $\{w_i\}$ et $\frac{\epsilon}{2}$ à son complémentaire $\{w_i\}^c$.

$$\text{On aura alors } \forall i = 1,\dots,n, \quad \begin{cases} \tilde{m}^v(\{1\}) &= (1-\frac{\epsilon}{2})1_{w_i=1} + \frac{\epsilon}{2}1_{w_i=0} \\ \tilde{m}^v(\{0\}) &= \frac{\epsilon}{2}1_{w_i=1} + (1-\frac{\epsilon}{2})1_{w_i=0} \\ \tilde{m}^v(\{1,0\}) &= 0 \end{cases}$$

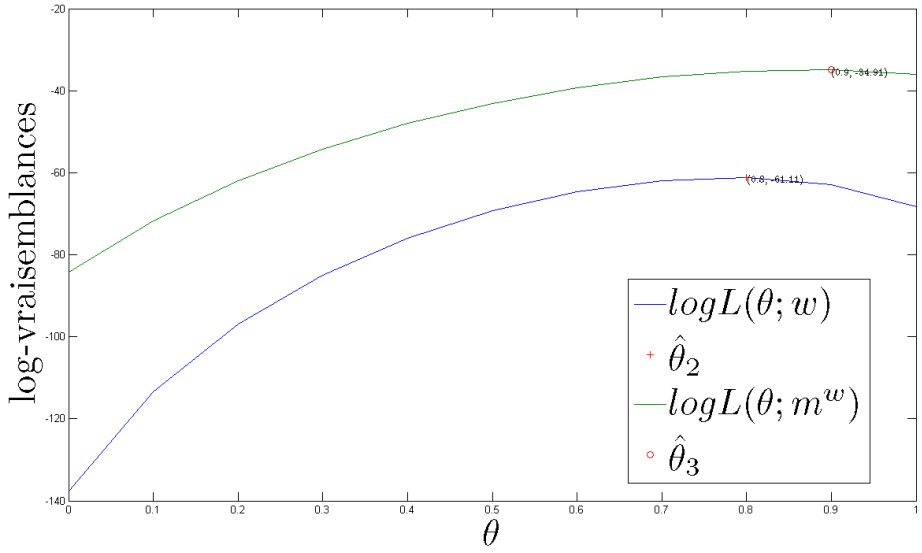
$\tilde{m}^v$ n'ayant pour éléments focaux que des singletons, elle définit une distribution de probabilité. L'*ignorance* est ici modélisée de façon uniforme. On estime ici que lorsqu'on ne sait pas, il toutes les possibilités ont exactement la même probabilité de se produire (la part d'incertitude ϵ des données est également répartie sur 0 et sur 1).

Remarque : Dans tout ce qui suit, on supposera que $\epsilon \in]0,1]$. Pour le cas $\epsilon = 1$, les données w sont totalement non-fiables, il deviendrait alors illusoire de vouloir estimer quoi que ce soit à partir de w, la seule chose à noter est que dans ce cas, $W \sim Ber(\frac{1}{2})$.

TABLE *2.1.:* Estimation du paramètre θ_V d'un échantillon binomial par maximisation de vraisemblance dans un cadre de données incertaines (de niveau d'incertitude connu ϵ)

estimateur	vraisemblance maximisée	cadre	prise en compte de ϵ
$\hat{\theta}_1 = \frac{n_1}{n}$	$L_0(\theta_V; w) = \theta_W^{n_1}(1-\theta_W)^{n-n_1}$	probabiliste	non
$\hat{\theta}_2 = [\frac{n_1}{n(1-\epsilon)} - \frac{\epsilon}{2(1-\epsilon)}]_+ \wedge 1$	$L(\theta_V; w) = [\frac{\epsilon}{2} + (1-\epsilon)\theta_V]^{n_1}.[1 - \frac{\epsilon}{2} - (1-\epsilon)\theta_V]^{n-n_1}$	probabiliste	oui
$\hat{\theta}_3 = [\frac{n_1}{n}\frac{(1+\epsilon)}{(1-\epsilon)} - \frac{\epsilon}{(1-\epsilon)}]_+ \wedge 1$	$L(\theta_V; m^v) = [(1-\epsilon)\theta_V + \epsilon]^{n_1}.[(1-\epsilon)(1-\theta_V) + \epsilon]^{n-n_1}$	crédibiliste	oui
$\hat{\theta}_4 = [\frac{n_1}{n(1-\epsilon)} - \frac{\epsilon}{2(1-\epsilon)}]_+ \wedge 1$	$L(\theta_V; \tilde{m}^v) = [\frac{\epsilon}{2} + (1-\epsilon)\theta_V]^{n_1}.[1 - \frac{\epsilon}{2} - (1-\epsilon)\theta_V]^{n-n_1}$	crédibiliste	oui

FIGURE *2.1.:* Variations des log-vraisemblances *classiques* et crédibilistes en fonction du "vraie" paramètre θ_V pour $n = 100$, $n_1 = 70$ et $\epsilon = 0.3$

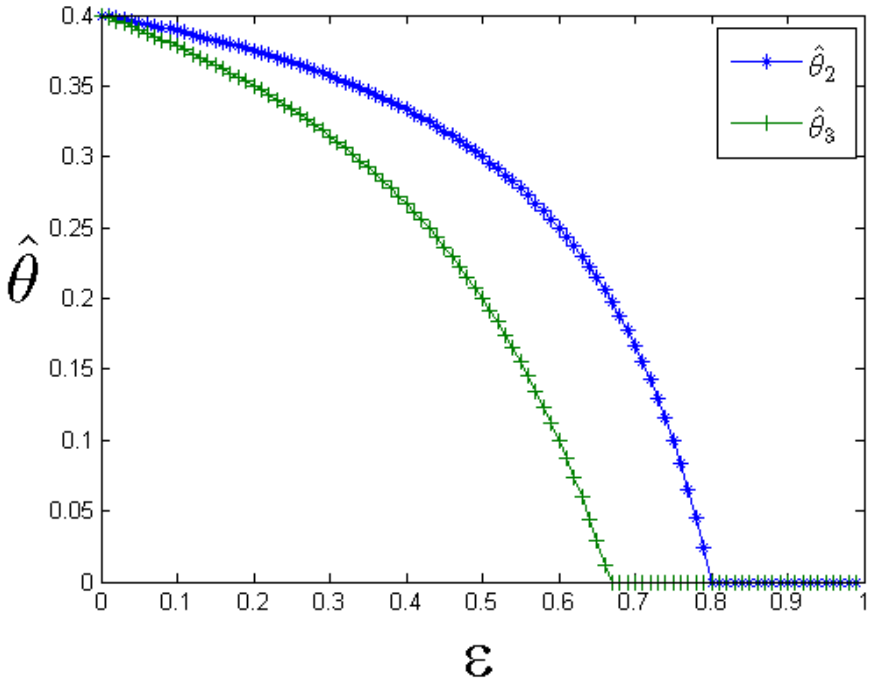


Vraisemblances et estimateurs obtenus :

Le calcul des quatre vraisemblances associées aux quatre modélisations précédemment proposées ainsi que leurs maximisations sont explicités en Annexe B, les résultats sont résumés dans la Table 2.1 où $[.]_+$ est la partie positive et $\wedge$ est l'opérateur $minimum$.

La figure 2.1 représente les variations de la fonction $\theta \rightarrow \log L(\theta; w)$ et de $\theta \rightarrow$

FIGURE 2.2.: Variations de $\hat{\theta}_2$ et $\hat{\theta}_3$ en fonction du niveau d'incertitude ϵ des données pour le cas où $n = 100$ et $n_1 = 40$



$\log L(\theta; m^v)$ dans le cas où $n = 100$, $n_1 = 70$ et où $\epsilon = 0.3$. On peut observer que les deux vraisemblances *classiques* et crédibilistes ont globalement la même forme, mais que les vraisemblances et leurs maximums respectifs $\hat{\theta}_2$ et $\hat{\theta}_3$ ont des valeurs différentes. La vraisemblance crédibiliste $L(\theta; m^v)$ semble supérieure à la vraisemblance *classique* $L(\theta; w)$.

La figure 2.2 représente les variation de $\hat{\theta}_2$ et $\hat{\theta}_3$ en fonction du niveau d'incertitude ϵ des données pour le cas où $n = 100$ et $n_1 = 40$. On observe que plus le niveau d'incertitude est élevé, plus les deux estimateurs auront tendance à favoriser la classe majoritaire ($\{0\}$ ici). On remarque aussi que ce phénomène est encore plus accentué pour l'estimateur crédibiliste $\hat{\theta}_3$. Cet effet en faveur de la classe majoritaire est intuitivement naturel, en effet plus des données sont incertaines, plus on ne retiendra d'elles que le minimum d'information, contenu ici dans la simple classe majoritaire.

Propriétés :

Proposition 2. *En prenant en compte le niveau d'incertitude ϵ des données, on a $L(\theta_V; w) = L(\theta_V; \tilde{m}^v)$, et donc $\hat{\theta}_2 = \hat{\theta}_4$.*

La proposition 2 (démontrée en Annexe B) signifie que dans ce cadre de travail, adopter une approche probabiliste classique (i.e. maximiser $L(\theta_V; w)$ selon le modèle génératif 2.15) est équivalent à adopter une approche crédibiliste (i.e. maximiser $L(\theta_V; \tilde{m}^v)$) avec une fonction de croyance $\tilde{m}^v$ modélisant la part d'incertitude ϵ des données de façon uniforme sur Ω_V.

Proposition 3. *$\hat{\theta}_1$ est un estimateur biaisé de θ_V, et son biais est donné par $E[\hat{\theta}_1] = \theta_W = \frac{\epsilon}{2} + (1-\epsilon)\theta_V$.*

Ce résultat s'obtient très simplement, en remarquant que $\hat{\theta}_1 = \sum_{i=1}^{n} 1_{\{w_i=1\}}$ et à partir de la preuve de la Proposition 1.

Proposition 4. *$\hat{\theta}_2$ est un estimateur consistant de θ_V, i.e. $\hat{\theta}_2$ converge en probabilité vers θ_V.*

Proposition 5. *$\hat{\theta}_3$ est asymptotiquement biaisé θ_V,*

i.e. $\lim_{n\to\infty} E[\hat{\theta}_3] \neq \theta_V$ et son biais asymptotique est donné par

$$\lim_{n\to\infty} E[\hat{\theta}_3] = \begin{cases} 0 & si \quad \theta_V < \frac{\epsilon}{2(1+\epsilon)} \\ (1+\epsilon)\theta_V - \frac{\epsilon}{2} & si \quad \frac{\epsilon}{2(1+\epsilon)} \leq \theta_V \leq \frac{2+\epsilon}{2(1+\epsilon)} \\ 1 & si \quad \theta_V > \frac{2+\epsilon}{2(1+\epsilon)} \end{cases}$$

.

On remarque que pour pour le cas où $\theta_V = \frac{1}{2}$, on a $\lim_{n\to\infty} E[\hat{\theta}_3] = \frac{1}{2} = \theta_V$.

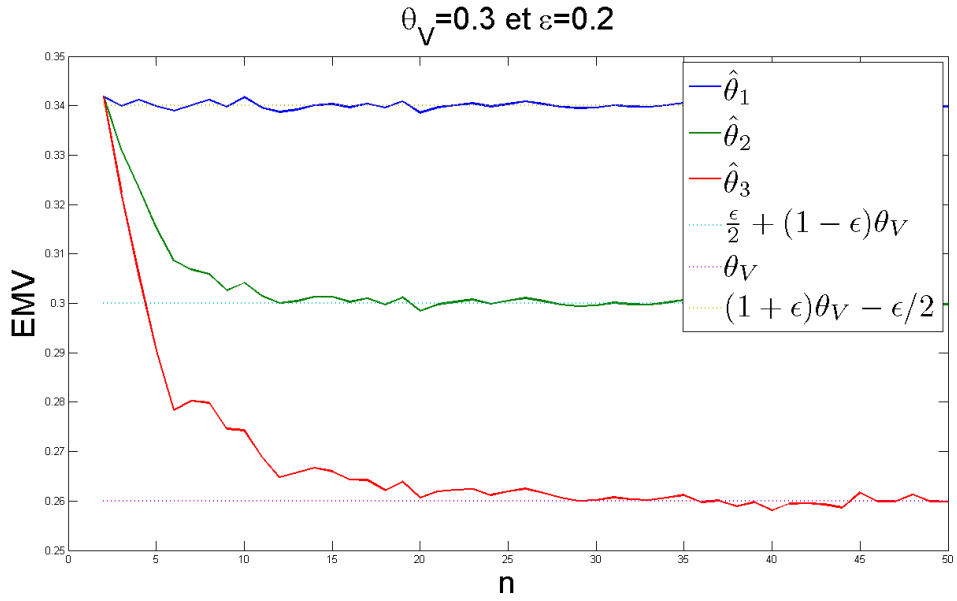
On peut donc ici conclure que la modélisation de l'incertitude des données par $\tilde{m}^v$ est asymptotiquement préférable à celle par m^w, la modélisation de l'ignorance par une loi uniforme apparaît donc ici plus fidèle que par une masse *vide*. Le seul cas où $\hat{\theta}_3$ est consistant est celui d'égalité initiale parfaite entre les deux classes considérées.

Les propositions 4 et 5 sont démontrées en Annexe B.

Remarque :

Il est possible d'utiliser d'autres modèles génératifs pour obtenir des données avec

FIGURE 2.3.: Variations de $\hat{\theta}_2$ et de $\hat{\theta}_3$ en fonction de la taille de l'échantillon pour le cas où $\theta_V = 0.3$ et où $\epsilon = 0.2$



un niveau d'incertitude ϵ. Considèrons le modèle génératif alternatif suivant :

Soit $v = (v_1, \ldots, v_n)$ un échantillon *précis* mais non observable contenant n réalisations d'une variable aléatoire de Bernouilli $V \sim Ber(\theta_V)$ avec $\Omega_V = \{0, 1\}$.

On définit le modèle génératif (de w à partir de v et de ϵ) suivant :

$$W = (1 - B)V + B(1 - V) \tag{2.16}$$

où

$$\begin{cases} V \sim Ber(\theta_V) \\ B \sim Ber(\frac{\epsilon}{1+\epsilon}) \\ V \text{ et } B \text{ indépendantes} \end{cases}$$

Proposition 6. *La vraisemblance $L(\theta_V; w)$ calculée selon le modèle génératif 2.16 a le même maximum que la vraisemblance $L(\theta_V; m^w)$.*

Cette proposition, démontrée en Annexe B, implique qu'il existe un modèle génératif permettant à l'estimateur du maximum de vraisemblance classique d'être égal à l'estimateur du maximum de vraisemblance crédibiliste calculé relativement à m^w. Ce résultat semble très intéressant car il permet une interprétation *probabiliste* du maximum de vraisemblance crédibiliste.

La figure 2.3 illustre les variations de $\hat{\theta}_1$, $\hat{\theta}_2$ et de $\hat{\theta}_3$ en fonction de la taille n de l'échantillon dans le cas où $\epsilon = 0.2$, et où $\theta_V = 0.3$. Les échantillons bruités w sont ici simulés 10000 fois selon le modèle génératif (2.15) et les moyennes des estimateurs obtenus sont ensuite représentées dans les différentes courbes.

On y observe clairement les convergences asymptotiques de $\hat{\theta}_2$ et de $\hat{\theta}_3$ vers respectivement θ_V et $(1+\epsilon)\theta_V - \frac{\epsilon}{2}$ (on a bien $\frac{\epsilon}{2(1+\epsilon)} \leq \theta_V \leq \frac{2+\epsilon}{2(1+\epsilon)}$ ici). Il est intéressant aussi de remarquer que pour des petits échantillons, $\hat{\theta}_3$ semble plus proche de θ_V que $\hat{\theta}_2$.

D'autres simulations ont été réalisées de manière à illustrer les variation de la différence d'erreur quadratique empirique entre $\hat{\theta}_2$ et $\hat{\theta}_3$ en fonction de la taille n de l'échantillon et du niveau d'incertitude ϵ des données pour différentes valeurs de θ_V, les figures correspondantes sont en Annexe B. La conclusion principale de ces simulations est que l'estimateur probabiliste $\hat{\theta}_2$ commet une moins grande erreur dans la majorité des cas. On peut tout de même souligner que pour les cas de très petits échantillons très déséquilibrés (i.e. contenant une classe très largement majoritaire) l'estimateur crédibiliste $\hat{\theta}_3$ commet en moyenne de moins grandes erreurs prédictives.

Bilan de l'exemple du modèle binomial :

De manière générale, pour le modèle génératif 2.15, l'estimateur probabiliste $\hat{\theta}_2$ est préférable à son homologue crédibiliste $\hat{\theta}_3$. Cela n'est pas très étonnant compte tenu du fait que $\hat{\theta}_2$ est obtenu par maximisation de la vraisemblance du paramètre θ_V du modèle ayant justement généré les données servant au calcul de cette vraisemblance. L'avantage crédibiliste est ici apparemment restreint aux cas pathologiques de très petits échantillons très déséquilibrés.

L'intérêt principal de l'estimateur du maximum de vraisemblance crédibiliste réside dans le fait qu'il ne suppose pas d'existence d'un modèle génératif de l'incertitude épistémique des données (il est généralement difficile de générer l'imperfection des données ou le manque de connaissance). Il serait intéressant, par la suite, de comparer les EMV classiques et crédibilistes dans le cas où la vraisemblance classique utilisée ne correspond pas au modèle génératif sous-jacent, i.e. considérer une erreur de modélisation.

Rappelons que, dans cette étude, le niveau d'incertitude est considéré constant pour tout l'échantillon w. Or comme nous le montrerons empiriquement dans le chapitre suivant, et comme le suggère Hüllermeier dans [58], la vraisemblance crédibiliste se comporte mieux lorsque le niveau d'incertitude des données est variable que lorsqu'il est constant.

La vraisemblance crédibiliste de la Définition 15 reste naturelle et relativement simple à utiliser dans un problème où les données sont intrinsèquement faciles à modéliser sous forme de fonctions de croyance, c'est pourquoi nous l'avons choisie dans ce manuscrit comme outil central d'estimation.

Même incertaines, les données contiennent de l'information potentiellement très riche, maximiser la vraisemblance crédibiliste comme nous le ferons dans ce travail est sans aucun doute préférable à ne pas tenir compte du tout de cette incertitude des données.

Fusion de 2 classes dans un échantillon à 3 classes

Nous nous plaçons ici dans un cadre d'estimation du paramètre d'une loi multinomiale à 3 classes à partir d'un échantillon *précis* $w = (w_1, \dots, w_n)$ issu de cette loi et de la connaissance a priori de la fiabilité ϵ des données de cet échantillon. Comme pour l'exemple précédent, en référence au vrai échantillon $v = (v_1, \dots, v_n)$ (v est lui fiable à 100%), on modélisera une masse de croyance m^v telle qu'on attribuera une masse de $1-\epsilon$ aux valeurs initiales de w et ϵ à Ω_W.

Considérons un cas pratique où w contient 2 exemples de classe A, 3 de classe B et 5 de classe C ($n_1 = 2$, $n_2 = 3$ et $n_3 = 5$) et où $\epsilon = 0.1$. La vraisemblance crédibiliste obtenue est $L(\theta, m^v) = [(1-\epsilon)\theta_A + \epsilon]^{n_1}[(1-\epsilon)\theta_B + \epsilon]^{n_2}[(1-\epsilon)\theta_C + \epsilon]^{n_3}$. En utilisant le fait que $\theta_C = 1 - \theta_A - \theta_B$, et en annulant les dérivées partielles (après avoir vérifié la convexité de la vraisemblance crédibiliste), on trouve : $\hat{\theta}_B = [\frac{(n_2 - n_1 - n_3)\epsilon + n_2}{n(1-\epsilon)}]_+ \wedge 1$ et $\hat{\theta}_A = [\frac{n_1(1-\epsilon)(1-\theta_B) + (n_1 - n3)\epsilon}{(1-\epsilon)(n_1 + n3)}]_+ \wedge 1 \approx 0.18$ On remarque que la prise en compte de l'incertitude des données, cette fois encore, favorise les classes majoritaires (sans prise en compte de l'incertitude des données, l'EMV du paramètre de la classe A aurait été directement obtenu par la proportion de classe A dans w, soit $\hat{\hat{\theta}}_A = 0.2$).

Fusionons à présent les classes B et C en une nouvelle classe qu'on appellera BC. Notre échantillon initial contient donc 2 exemples de classe A et 8 exemples de classes BC. En gardant le même niveau d'incertitude global des données ($\epsilon = 0.1$), l'$EMVC$ de la classe A obtenu est ici : $\hat{\theta}_A = 0.11$.

Bilan :

Nous observons ici un comportement très particulier de l'$EMVC$. En effet, après fusion de deux classes, l'estimateur de la classe qui n'a pas été fusionnée est modifié. Il est important d'observer ici le sens de modification de cet estimateur. Il apparaît que cette fusion des deux classes majoritaires a encore plus défavorisé la classe minoritaire (A) déjà défavorisée par la prise en compte de l'incertitude des données. Ce résultat sera déterminant pour l'implémentation de la méthodologie des *arbres* E^2M (voir Section 7).

2.5. Aspects philosophiques et pratiques

La théorie des fonctions de croyance offre donc une souplesse modélisation de l'incertitude intéressante . Elle permet de traiter les incertitudes épistémiques et aléatoires, et même si elle ne les différencie pas concrètement (le traitement est le même dans les deux cas), le simple fait de se demander à quel type d'incertitude on a affaire peut aider à une modélisation qui, elle, peut traiter de manière distincte tous les types d'incertitude en les mettant à différents niveaux. Dans l'exemples 1, les incertitudes épistémiques sont modélisées en amont des incertitudes aléatoires.

Cependant le fait de considérer l'ensemble des sous-parties 2^{Ω_W} pour définir une fonction de masse au lieu du simple espace Ω_W comme c'est le cas en théorie des probabilités implique la non-additivité des fonctions de croyance. Ceci représente un inconvénient majeur pour la conciliation de méthodologies crédibilistes et probabilistes classiques.

En outre, la taille de l'ensemble des sous-parties augmentant de façon exponentielle avec la taille de l'espace considéré, en pratique, le nombre d'opérations nécessaires à l'implémentation d'une inférence crédibiliste et donc le temps de calcul

peuvent vite exploser. Pour éviter cela, il faut bien contrôler les modèles (et les nombres d'ensembles focaux considérés) et proposer des méthodes pratiques d'inférence de complexité limitée. C'est l'objet des chapitres suivants qui se concentreront sur le cas des arbres de classification.

CHAPITRE 3

Algorithmes EM et E^2M

Comme expliqué brièvement dans la partie précédente, il est courant de se retrouver en face d'incertitudes épistémiques à l'intérieur d'un cadre de travail de traitement d'incertitudes aléatoires (apprentissage de modèle probabiliste à partir de données incertaines). En statistique, les modèles d'incertitudes sont probabilistes et traitent donc naturellement des incertitudes aléatoires mais il arrive que les données de départ ou les modèles probabilistes sous-jacents soient mal observés ou mal connus et donc entachés d'incertitudes épistémiques. Les estimations par maximum de vraisemblances peuvent alors être compliquées à réaliser directement. Ces vraisemblances dites *incomplètes* ou *imprécises* sont notées $L(\theta, A)$, A étant l'ensemble des possibilités selon notre connaissance : $w \in A$. Lorsque notre connaissance des données n'est plus seulement imprécise (ensembliste) mais aussi incertaine il devient alors intéressant de chercher à maximiser la vraisemblance crédibiliste $L(\theta, m^w)$ décrite en section 2.4 qui en résulte. Néanmoins, dans le cas général, la vraisemblance d'un modèle paramétrique peut être difficile à maximiser (n'étant pas toujours convexe). C'est encore plus vrai pour les vraisemblances imprécises et crédibilistes. Il faut donc un moyen pratique pour résoudre ce problème. C'est dans ce contexte que Dempster proposa l'algorithme EM permettant la maximisation des vraisemblances *imprécises* qui sera ensuite étendu par Denoeux aux vraisemblances incertaines *crédibilistes*.

Sommaire

3.1. Algorithme *EM*

En 1977, dans un cadre statistique classique cette fois, Dempster présenta un moyen de maximiser certaines vraisemblances *imprécises* $L(\theta, A)$ qui seraient directement maximisables (i.e. par une formule *fermée* analytique) si les données étaient complètes (on parlerait alors de vraisemblance *complètes* ou *complétées*) : l'*algorithme Expectation Maximisation* (*EM*) [57]. Cet algorithme est aujourd'hui très largement utilisé dans différents domaines (statistique, intelligence artificielle, ...).

Dempster prouva dans [57] qu'en maximisant itérativement l'espérance $E[L(\theta, w)|A]$ de la vraisemblance *complète* conditionnellement à notre connaissance A (la vraisemblance *complète* est incertaine dans le cas de données incertaines), la vraisemblance *imprécise* se retrouve alors aussi maximisée.

L'algorithme *EM* est itératif, et chaque itération q comprend deux étapes :

– l'étape *Expectation* (*E*) où est calculé l'espérance de la log-vraisemblance complète conditionnellement à notre connaissance $w \in A$

$$Q(\theta;\theta^{(q)}) = E[\log L(\theta, w)|A;\theta^{(q)}]$$

– l'étape *Maximisation* (*M*) où l'espérance calculée à l'étape *E* est maximisée

$$\theta^{(q+1)} = \arg\max_{\theta\in\Theta} Q(\theta;\theta^{(q)})$$

L'algorithme converge vers un maximum local de la vraisemblance imprécise,i.e. $L(\theta^{(q)}, A) \leq L(\theta^{(q+1)}, A)$.

Cet algorithme peut en réalité s'appliquer à n'importe quelle vraisemblance *incertaine* (le principe est de maximiser l'espérance d'une telle vraisemblance sur l'espace de notre connaissance). Il fait donc désormais partie des méthodes de base pour maximiser des vraisemblance en cas de données manquantes, mais aussi pour estimer les coefficients de pondération des modèles de mélanges [57], [59, Chapter 9.2], [60].

3.2. Extension crédibiliste : l'algorithme E^2M

Dans [56], Denoeux propose une extension de l'algorithme EM aux données crédibiliste : l'algorithme E^2M.

L'idée est la même : maximiser itérativement l'espérance $E[L(\theta, w) \mid m^w]$ de la vraisemblance *complète* conditionnellement à notre connaissance cette fois-ci représentée par m^w. Il montre que la vraisemblance crédibiliste $L(\theta, m^w)$ décrite en section 2.4 converge également vers un maximum local.

La difficulté principale de cette démarche était la manière de calculer ces espérances conditionnelles à des fonctions de croyance. En effet tout calcul d'espérance nécessite le choix d'une mesure de probabilité au sens de Lebesgue [61] et dans un cadre crédibiliste il n'est pas toujours évident de choisir la mesure adaptée au problème. Contrairement aux probabilités qui sont elles définies uniquement sur des singletons, les fonctions de croyance étant définies sur l'espace 2^{Ω_W}, intégrer sur des ensembles n'est pas directement réalisable. De nombreux travaux ont cherché des solutions à ce problème le plus souvent en utilisant l'intégrale de Choquet [62], ce qui aboutit à des bornes inférieures et supérieures pour l'espérance recherchée, or avec de telles bornes l'étape de maximisation peut devenir problématique.

Pour résoudre ce problème, lors de la $q^{ième}$ itération, Denoeux propose de combiner conjonctivement la probabilité $P_{\theta^{(q)}}$ issue du modèle paramétrique $\mathscr{P} = \{P_\theta | \theta \in \Theta\}$ avec la fonction de croyance (multidimensionnelle) m^w représentant les données. Le résultat de cette combinaison est en fait une probabilité notée $P(.|\theta^{(q)}, m^w)$ qui permet alors tout calcul d'intégrale, de plus d'un point de vue intuitif cette mesure tient compte à la fois du modèle P_θ et de notre connaissance m^w il s'agit donc bien en quelque sorte d'une probabilité conditionnelle à notre savoir.

On obtient

$$\begin{aligned} \forall w \in \Omega, \quad P(w \mid \theta^{(q)}, m^w) &= P_{\theta^{(q)}} \textcircled{\cap} m^w(w) \\ &= \frac{P_{\theta^{(q)}}(w) pl^W(w)}{L(\theta^{(q)}, m^w)} \end{aligned}$$

A chaque itération q on a toujours les deux étapes :

– l'étape *Espérance* (E) où est calculé l'espérance de la log-vraisemblance com-

plète qui est calculée avec la mesure de probabilité $P(w|\theta(q), m^w)$

$$Q(\theta;\theta^{(q)}) = E[\log L(\theta, w) \mid \theta^{(q)}; m^w]$$

– l'étape *Maximisation* (M) où l'espérance calculée à l'étape E est maximisée

$$\theta^{(q+1)} = \arg\max_{\theta\in\Theta} Q(\theta;\theta^{(q)})$$

L'algorithme converge ici encore localement vers l'estimateur du maximum de vraisemblance crédibiliste,i.e. $L(\theta^{(q)}, m^w) \leq L(\theta^{(q+1)}, m^w)$.

Notons que si les données sont juste imprécises ($m^w(A) = 1$), alors on retrouve l'algorithme EM classique, il s'agit donc bien d'une extension.

Bilan :

La finalité de ce travail étant l'apprentissage d'un modèle prédictif à partir de données épistémiquement incertaines, le choix des fonctions de croyance comme modèle de représentation de cette incertitude épistémique rend possible et cohérente l'utilisation de l'algorithme E^2M pour maximiser la vraisemblance crédibiliste des données.

Le passage d'un niveau de modélisation crédibiliste (des données) à un niveau de prédiction probabiliste est réalisé par la maximisation de la vraisemblance crédibiliste des données relativement à un paramètre de ce modèle prédictif probabiliste (voir section 4). Cette maximisation n'étant pas directement possible, contrairement à l'exemple du modèle paramétrique de Bernouilli (section 2.4.1), nous utilisons l'algorithme E^2M en modifiant la méthode de clustering d'un exemple présenté par Denoeux dans [56](voir section 7).

D'autres choix auraient pu être faits pour passer d'un niveau modélisation incertaine épistémique crédibiliste à un niveau de modélisation probabiliste. Le MCT (voir section 2.3.2) propose un tel passage avec la probabilité pignistique qui est calculée à partir de fonctions de croyance. Ce modèle cependant, répartissant uniformément les masses des éléments focaux sur les singletons qu'ils contiennent, revient à renoncer à l'avantage de modélisation incertaine offert par les fonctions de croyance en niant l'aspect imprécis de notre connaissance, et ce dès le stade de modélisation des incertitudes des données. Nous partons du principe que, plus

longtemps cette largeur de modélisation est préservée durant le processus d'inférence, plus importants seront les avantages qui en découlent. En maximisant la vraisemblance crédibiliste des données, on conserve ces avantages de modélisation pendant toute la phase de modélisation des incertitudes des données tout en estimant le modèle probabiliste correspondant le mieux à la connaissance qu'on a des données.

De plus une *probabilisation* de l'incertitude des données par le MCT reviendrait à ignorer la nature épistémique de cette incertitude. En effet une probabilité classique est un modèle d'incertitude adéquat surtout pour les incertitudes aléatoires (comme vu en section 1).

On aurait aussi pu faire le choix de garder une dimension incertaine imprécise plus longtemps, en considérant par exemple l'ensemble des modèles probabilistes compatibles avec notre connaissance crédibiliste des données, et donc l'ensemble des prédictions qui en résultent. Un choix (par vote par exemple) entre ces différentes prédictions aurait ensuite été nécessaire pour la phase de décision ou de prédiction. Un ensemble de prédictions aurait même pu être envisagé (prédictions *imprécises*), cependant nous nous plaçons dans un cadre classique de prédiction précises.

Deuxième partie .

Incertitude et arbres de décision

CHAPITRE 4

Arbres de décision

Les arbres de décision sont un modèle de classification/prédiction simple d'utilisation et qui permettent une interprétation très utile aux experts. Aujourd'hui très largement utilisés dans différents domaines ils représentent un classifieur standard, facilement modulable selon les applications. Après une brève description de leur structure et de leur utilisation prédictive, un rappel historique de leur apparition dans la littérature, leur construction est présentée dans un formalisme inhabituel qui sera adéquat pour la méthodologie des *arbres* E^2M présentée en Section 7. Les principes d'élagage et de forêts aléatoires sont ensuite rapidement présentés.

Sommaire

4.1. Cadre général, définitions

L'*apprentissage* consiste à apprendre un modèle à partir de données ou d'observations. Nous considérons ici le problème de *prédiction* où le modèle consiste à prédire une sortie (la prédiction) à partir d'une entrée (les données).

Lorsque cette prédiction porte sur une variable $Y \in \Omega_Y = \{w_1, \ldots, w_K\}$ dite *catégorique* , *nominale* ou *qualitative*, la prédiction rentre dans le cadre général de la

classification. Si Y est une variable *numérique*, il s'agit alors du cadre de la *régression*.

La classification (de même que la régression) a pour but l'attribution d'une classe à un individu à partir de ses J *attributs* $X = (X^1, \ldots, X^J) \in \Omega_X = \Omega_{X^1} \times \cdots \times \Omega_{X^J}$ (ou variables descriptives). Un classifieur est donc une fonction $f : \Omega_X \longrightarrow \Omega_Y$.

Un classifieur se construit à partir d'un échantillon d'apprentissage qui contient N réalisations du couple (X, Y) et que l'on notera :

$$E = \begin{pmatrix} x_1, y_1 \\ \vdots \\ x_N, y_N \end{pmatrix} = \begin{pmatrix} x_1^1, \ldots, x_1^J, y_1 \\ \vdots \\ x_N^1, \ldots, x_N^J, y_N \end{pmatrix}. \tag{4.1}$$

Parmi les classifieurs les plus reconnus en classification supervisée [59] on peut citer le classifieur naïf de Bayes, les arbres de décision, les k-plus-proches-voisins (k-ppv), les réseaux de neurones, les séparateurs à vaste marge (SVM).

En général les classifieurs sont évalués sur leur aptitude à bien *classer* des individus d'un autre échantillon appelé l'*échantillon de test*. Cependant leur stabilité, leur interprétabilité, les temps de calculs qu'ils nécessitent sont autant de critères pouvant rentrer en compte dans leur évaluation suivant le problème. Il n'y a donc pas de meilleur classifieur d'une façon générale, tout dépend de ce qu'on attend d'eux suivant les objectifs de notre problème.

Pour cette thèse nous avons choisi les arbres de décision pour leur sortie graphique très interprétable même par un non-informaticien ou non-statisticien. Grâce à cette interprétabilité les intéractions avec les experts sont facilitées (voir chapitre 9).

Arbres de décision : formalisme

Un arbre de décision est formellement une structure d'arbre comprenant des noeuds, des branches et des feuilles (ou noeuds terminaux). Un arbre à H feuilles sera noté $\mathscr{P}_H = \{t_1, \ldots, t_H\}$ où $t_1, \ldots, t_H$ représentent les feuilles de $\mathscr{P}_H$.

Relativement à $\mathscr{P}_H$, on définit la variable *feuille* $Z_{\mathscr{P}_H} \in \Omega_{Z_{\mathscr{P}_H}} = \{1, \ldots, H\}$.

Un arbre de décision se lit de haut en bas, à chaque noeud est attaché un attribut X^j et chaque branche issue de ce noeud correspond à une sous-partie de l'espace de définition Ω_{X^j} de X^j.

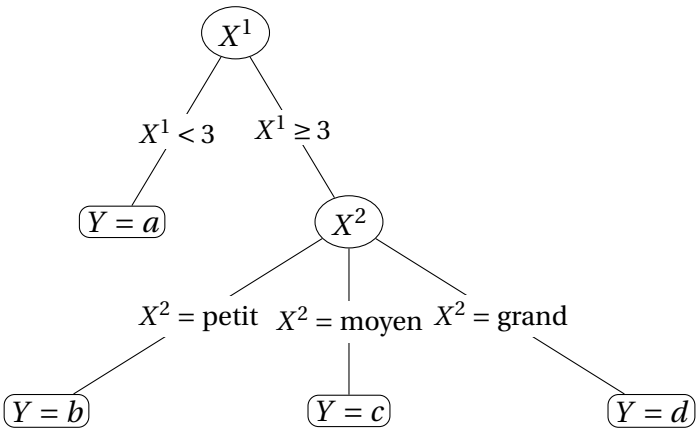


FIGURE *4.1.:* exemple d'arbre de décision

On peut donc voir un arbre de décision comme une partition de l'espace des attributs Ω_X avec une classe attribuée à chaque élément de cette partition, ces éléments de la partition étant les équivalents des feuilles d'un arbre.

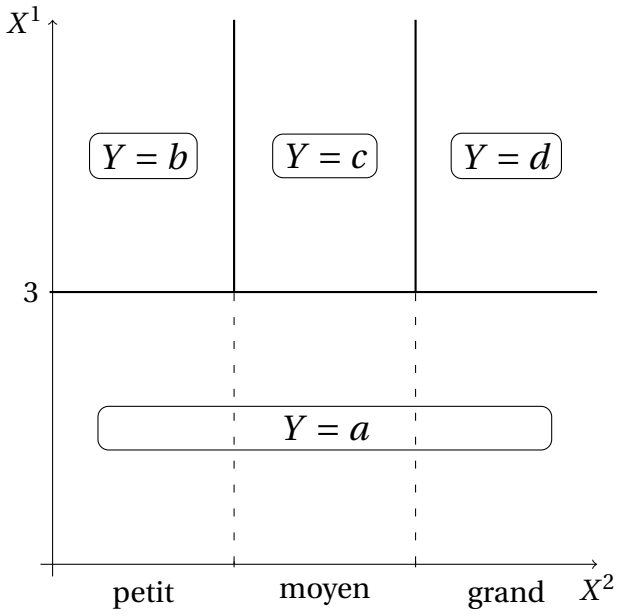


FIGURE *4.2.:* partition correspondante

On notera A_h^j l'élément de partitions de Ω_{X^j} correspondant à la feuille t_h. Un arbre de décision $\mathscr{P}_H$ pourra donc s'écrire

$$\mathscr{P}_H = \{t_1, \ldots, t_H\} = \{A_1^1 \times \cdots \times A_1^J, \ldots, A_H^1 \times \cdots \times A_H^J\}$$

où $\times$ est le produit cartésien.

La figure 4.1 représente un exemple simple d'arbre de décision. Un nouvel individu dont les attributs sont $X^1 = 5$ et $X^2 =$ petit sera donc classifié $Y = b$. Par la suite on confondra les feuilles t_h avec les sous-espaces qu'elles représentent. La figure 4.2 représente la partition de Ω_X correspondant à l'arbre de la figure 4.1.

Remarque :

La structure des arbres de décision et leur lecture de bas en haut leur confère un pouvoir de *sélection de variable*. En effet, les attributs apparaissant en haut d'un arbre ont un *pouvoir explicatif* de la classe supérieur à ceux des attributs apparaissant plus bas dans les mêmes arbres.

Lorsque la classe Y est nominale on parle d'*arbre de classification* par opposition au cas ou Y est numérique auquel cas on parle d'*arbre de régression*. Dans la suite de ce manuscrit nous nous restreindrons aux arbres de classification qu'on appellera indifféremment arbres de décision ou arbres de classification.

4.2. Historique

Les premiers travaux s'apparentant plus ou moins aux arbres de décision furent présentés par Belson en 1956 dans un cadre statistique de *régression* [63]. Un peu plus tard, les premiers algorithmes d'arbres de décision firent leur apparition, le premier fut lui aussi proposé par les statisticiens Morgan et Sonquist en 1963 [64].

Cependant ce n'est qu'en 1984 que les arbres de décision devinrent réellement populaires avec le travail de Breiman et al. [65]. Ce travail, intitulé $CART$ (Classification And Regression Trees), s'inscrit lui aussi dans un cadre purement statistique.

C'est Quinlan qui en 1986 fut le premier informaticien à s'intéresser de près aux arbres de décision avec l'algorithme $ID3$ (dont les versions ultérieures sont $C4.5$, $C5.0$) [66]. Grâce à ce travail, c'est toute une communauté scientifique (celle de l'*apprentissage automatique* qui est essentiellement composé d'informaticiens) qui bénéficia de cet outil très puissant que représentent les arbres de décision.

De très nombreux autres algorithmes d'arbres de décision furent ensuite proposés tel que $CHAID$ [67], $SLIQ$ [68], $QUEST$ [69], $VFDT$ [70].

Malgré tout $CART$ et $C4.5$ restent à ce jour les deux algorithmes de construction d'arbres de décision les plus reconnus et utilisés. Alors que le premier ne produit que des arbres binaires (chaque noeud ne donne naissance qu'à deux branches) et comprend une méthode d'élagage dont l'efficacité n'est plus à démontrer, le second est quant à lui plutôt adapté aux attributs qualitatifs (chaque noeud sur un tel attribut donne naissance à un nombre de branches égal au nombre de modalités de l'attribut) et produit des arbres moins profond (un attribut ne peut apparaître qu'une fois le long d'un chemin partant du noeud initial à une feuille).

On peut remarquer que très souvent les statisticiens ont tendance à privilégier $CART$ alors que leurs confrères informaticiens se tournent plus facilement vers $C4.5$.

4.3. Construction

Dans cette partie, le principe général de la construction d'un arbre de décision est présenté.

Durant la construction d'un arbre de décision, le but est de séparer au mieux les classes de manière à obtenir des feuilles (ou de manière équivalente des éléments de la partition de Ω_X correspondant à l'arbre) le plus "pures" possible en terme de classe. Cette pureté, ou homogénéité, de classe se mesure à l'aide de fonctions d'impureté notées i.

Différentes fonctions d'impureté existent, parmi les plus connues on peut citer l'*entropie de Shanon* (utilisée dans $C4.5$), l'*indice de Gini* (utilisé dans $CART$). Voici leurs expressions pour une feuille t_h contenant potentiellement K classes

$$\text{entropie : } i(t_h) = -\sum_{k=1}^{K} \alpha_h^k \log(\alpha_h^k) \tag{4.2}$$

$$\text{indice de Gini : } i(t_h) = \sum_{k=1}^{K} \alpha_h^k (1 - \alpha_h^k) \tag{4.3}$$

où α_h^k est la probabilité de la classe k au sein de t_h. Il est à noter que le choix du critère de pureté n'a que peu d'influence sur l'efficacité de l'arbre obtenu comme souligné dans [65, Chapter 2.5.2].

Pendant la construction d'un arbre de décision, pour chaque feuille de l'arbre, pour toutes les coupures envisageables, un gain de pureté Δi est calculé, et la coupure donnant le meilleur gain Δi_{max} est choisie. Pour une coupure d'une feuille t_h donnant naissance aux M feuilles $t_{h_1} \dots t_{h_M}$, ce gain de pureté se calcule comme ceci :

$$\Delta i = i(t_h) - \frac{1}{\pi_h} \sum_{m=1}^{M} \pi_m i(t_{h_m}) \tag{4.4}$$

où π_m est la probabilité de la feuille t_{h_m} (et π_h celle de t_h)

Dans les méthodes classiques d'arbres de décision, les termes π_h et α_h^k des équations (4.2), (4.3) et (4.4) sont simplement estimés de manière fréquentiste :

$$\pi_h = P(Z = h) \approx \frac{n_h}{n} \tag{4.5}$$

$$\alpha_h^k = P(Y = \omega_k / Z = h) \approx \frac{n_h^k}{n_h} \tag{4.6}$$

où n_h représente le nombre d'exemples *terminant* dans la feuille t_h et où n_h^k représente le nombre d'exemples de classe k qui *terminent* dans la feuille t_h.

De manière à faciliter la présentation de notre méthodologie, nous résumerons toutes ces probabilités dans un paramètre θ, qui sera donc relatif à un arbre donné. On notera

$$\theta_{\mathscr{P}_H} = \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_H \end{pmatrix} = \begin{pmatrix} \alpha_1 & \pi_1 \\ \vdots & \vdots \\ \alpha_H & \pi_H \end{pmatrix} = \begin{pmatrix} \alpha_1^1 & \cdots & \alpha_1^K & \pi_1 \\ \vdots & \ddots & \vdots & \vdots \\ \alpha_H^1 & \cdots & \alpha_H^K & \pi_H \end{pmatrix} \in \Theta_H$$

où Θ_H est l'espace de tous les paramètres possibles pour un arbre à H feuilles.

La plupart (mais pas toutes) des fonctions d'impureté ne donnent que des gains d'impureté Δi positifs ou nuls. Ceci est une propriété mathématique directement liée à leur définitions.

La construction d'un arbre se fait donc de manière itérative ; des *critères d'arrêt*, appelés aussi *pré-élagage* doivent donc être définis préalablement. Ces critères varient suivant l'algorithme choisi. Ils définissent des situations pour lesquelles une

feuille t_h ne doit plus être coupée. Voici les principales situations pour lesquelles un arrêt peut être envisagé.

- la taille de la feuille (ou de l'un de ses potentiels enfants) est trop petite : $|t_h| < taille_{min}$
- le gain d'impureté maximale est trop petit : $\Delta i_{max} < gain_{min}$
- le nombre de feuilles maximum est atteint : $H = H_{max}$
- la profondeur maximale : $profondeur(\mathscr{P}_H) \geq profondeur_{max}$
- la feuille est suffisamment pure : $i(t_h) < i_{min}$

Les choix des critères d'arrêt est déterminant pour la complexité des arbres obtenus ainsi pour que les temps de calcul. Selon la cadre de travail, ils pourront donc être ajustés. Des critères d'arrêt restrictifs donnant des arbres plutôt petits faciliteront la compréhension et l'interprétabilité par les différents experts impliqués mais auront un pouvoir prédictif limité par rapport à des arbres obtenus avec des critères d'arrêt plus laxistes qui seront donc plus denses mais qui correspondront à un partitionnement plus fin de Ω_X.

4.4. Elagage

D'un point de vue applicatif, la taille d'un arbre de décision peut avoir beaucoup d'importance. Outre le fait qu'un arbre trop *grand* peut être très difficile à interpréter, il risque d'être trop spécifique à l'échantillon d'apprentissage et son pouvoir de généralisation peut diminuer au delà d'une certaine *taille*. Ce phénomène est connu sous le nom de *sur-apprentissage*.

Pour pallier à cet inconvénient, de nombreuses méthodes d'élagage, appelées aussi *post-élagage* ont été proposées pour les différents algorithmes d'arbre de décision. A titre d'exemple, l'élagage par minimisation du critère *Coût-Complexité* proposé par Beiman dans l'algorithme $CART$ [65] permet de réaliser un compromis entre la complexité et l'efficacité des arbres qui restent ainsi suffisamment efficaces (en terme de prédiction) tout en étant de tailles relativement réduites.

Ce critère à minimiser est $R_\gamma(\mathscr{P}_H) = R(\mathscr{P}_H) + \gamma.H$ où $R(\mathscr{P}_H)$ est l'erreur de prédiction commise par $\mathscr{P}_H$ (sur l'échantillon d'apprentissage ou sur un échantillon d'élagage).

Une fois un arbre de décision construit, une suite de sous-arbres emboîtés est obtenue à partir de l'arbre initial (en minimisant $R_\gamma(\mathscr{P}_H)$ et en jouant sur le paramètre

γ), et le sous-arbre minimisant $R(\mathscr{P}_H)$ sur un échantillon test (ou d'élagage) est choisi. Ce critère est notamment utilisé dans la fonction *prune* du package $rpart$ du logiciel R.

Dans $CART$, les critères d'arrêt sont généralement très souples, ils se restreignent à la pureté totale des feuilles ($i(t_h) = 0$) une feuille pouvant ne contenir qu'un exemple de l'échantillon d'apprentissage.

Remarque : Les arbres de décision ont généralement un haut pouvoir prédictif allié à une bonne interprétabilité. Ils présentent cependant un inconvénient non-négligeable : l'instabilité. En effet, en modifiant même très légèrement l'échantillon d'apprentissage, il est courant d'obtenir un arbre très différent même si l'efficacité prédictive reste elle, relativement stable (voir [71]).

4.5. Forêts aléatoires

Si le coté interprétable des arbres de décision n'est pas important et que l'efficacité prédictive est le seul objectif, de nombreuses méthodes faisant intervenir des arbres de décision ont été développées dans le but d'améliorer cette efficacité prédictive. La plus connue est sans doute les *forêts aléatoires* proposées par Breiman en 2001 [72].

Cette méthodologie consiste en l'apprentissage d'un certain nombre d'arbres de décision sur des sous-échantillons de l'échantillon d'apprentissage (obtenus par *bootstrap*). La prédiction se fait ensuite en combinant les prédictions de chaque arbre (par *vote* par exemple). Les conditions associées aux noeuds peuvent concerner plusieurs attributs (tirés aléatoirement) combinés éventuellement de façon non-linéaire.

L'efficacité prédictive des *forêts aléatoires* et leurs propriétés mathématiques [73] en font des classifieurs très puissants et donc très populaires, notamment en *apprentissage automatique*. La méthodologie des forêts aléatoires tire avantage d'incertitudes aléatoires injectés à différents stades de l'apprentissage (ré-échantillonage aléatoire, ensembles aléatoires d'attributs potentiellement associés aux noeuds). D'autres méthodologies d'arbres de décision, que nous allons maintenant détailler, adoptent d'autres modèles d'incertitudes à différentes étapes de la construction et de l'utilisation des arbres.

CHAPITRE 5

Modélisations de l'incertitude dans différentes méthodologies d'arbres de décision

Nous passons ici en revue quelques méthodologies d'arbres de décision ou différentes incertitudes interviennent, soit à l'intérieur du modèle, soit dans notre connaissance des données.

Sommaire

5.1. Approches probabilistes

Dans le cadre probabiliste classique, à part les *forêts aléatoires* (voir Section 4.5) qui tirent profit d'incertitudes aléatoires inhérentes au modèle, différentes méthodologies permettent de gérer certains types de données incertaines. L'incertitude (épistémique) des données est alors exprimée à l'aide de distributions de probabilité (données incertaines probabilistes).

5.1.1. *Décomposition* des exemples d'apprentissage dans l'arbre : *Tsang et al.*

En 2011, Tsang et al. propose dans [74] d'apprendre des arbres de décision à partir de données incertaines. L'incertitude de ces données est exprimée à l'aide de distributions de probabilité et ne concerne que les attributs. Les auteurs proposent pour cela deux solutions : convertir ces distributions en données précises en calculant leurs moyennes (*approche moyennes*) ou alors décomposer les exemples dans toutes les feuilles des arbres en regard de ces distributions sur les attributs (*approche distributions*).

Dans tous les cas, les critères de coupure restent classiques. Pour l'approche distributions, les effectifs utilisés pour les estimations dans les calculs de pureté (voir Equations 4.2 et 4.3) ou de gain de pureté (voir Equation 4.4) sont *remplacés* par des portions du nombre d'exemple total. Ces portions sont obtenues en calculant, pour chaque exemple i et chaque feuille t_h, les probabilités $P(x_i \in t_h)$ d'appartenance à la feuille. Pour une coupure "$X^j < 5$", si un exemple a comme $j^{ième}$ attribut une distribution uniforme sur $[4,7]$, les auteurs considèrent alors que un tiers de cet exemple vérifie la condition de coupure, et que les deux autres tiers ne la vérifie pas. On aura donc, pour cet exemple, une portion de $\frac{1}{3}$ *terminant* dans la feuille correspondant à cette condition et une portion de $\frac{2}{3}$ *terminant* dans la feuille correspondant au complémentaire de cette condition.

Les estimations des probabilités des feuilles π_h et des classes dans les feuilles α_h^k proposées dans les Equations (4.5) et (4.6) ne sont pas modifiées mais les termes n_h et n_h^k sont ici obtenus en sommant les proportions d'exemples, on a $n_h = \sum_{i=1}^{n} P(x_i \in t_h)$

et $n_h^k = \sum_{i:y_i=w_k} P(x_i \in t_h)$.

Dans le cas de données *précises* on connaît avec certitude les feuilles où *tombent* les différents exemples, on a donc $n_h = \sum_{i=1}^{n} \mathbb{1}_{\{x_i \in t_h\}}$ et $n_h^k = \sum_{i:y_i=w_k} \mathbb{1}_{\{x_i \in t_h\}}$. Ces fonctions indicatrices sont ici remplacées par des probabilités reflétant notre connaissance des données.

Même si cette approche a l'avantage d'être simple à mettre en oeuvre et d'être assez naturelle, elle présente cependant deux inconvénients. Premièrement, l'incertitude *épistémique* des données est ici exprimée à l'aide de probabilités classiques, or comme vu en Section 1.2, les probabilités ne sont pas idéales pour représenter toutes les incertitudes épistémiques. Deuxièmement, dans cette approche les exemples se *partagent* dans les différentes feuilles or en réalité un exemple possède des vrais attributs (éventuellement mal observés) et devrait donc *terminer* dans une unique feuille. Fractionner les exemples de cette manière pour obtenir des *pseudo-effectifs* à valeurs réelles ne correspond donc pas à la réalité (même si cette réalité est mal connue). On peut même dire que ce modèle *apprend* de l'incertitude et non la réalité.

5.1.2. Méthodologie Périnel

En 1999, Périnel propose une méthodologie d'arbres de décision gérant les données incertaines probabilistes [4]. Encore une fois, l'incertitude des données est exprimée à l'aide de distributions de probabilité et ne concerne que les attributs. Contrairement à Tsang et al., Périnel ne fractionne pas les exemples d'apprentissage, il propose un autre critère de coupure que la maximisation de la pureté des feuilles (rendue ici incalculable).

Périnel propose d'évaluer un arbre $\mathscr{P}_H$ par la vraisemblance de la classe conditionnellement aux attributs $L(\theta_H; y \mid x)$. Il montre que, dans un cadre de données d'apprentissage précises, chercher à maximiser une telle vraisemblance revient (à un coefficient près) à minimiser l'entropie telle que dans l'Equation 4.4. Il existe donc un lien formel entre vraisemblance (conditionnelle) des données et *qualité* d'un arbre, ou d'une partition, $\mathscr{P}_H$.

Dans le cas de données incertaines probabilistes, le calcul des vraisemblances $L(\theta_H; y \mid x)$ nécessite l'estimation des paramètres θ_H. Ces paramètres sont ici estimés

justement en maximisant cette vraisemblance $L(\theta_H; y \mid x)$ qui est donnée par :

$$L(\theta_H; y \mid x) = \prod_{i=1}^{N} \prod_{h=1}^{H} [P(Z = h \mid X = x_i) P_{\theta_H}(y_i)]^{\mathbb{1}_{z_i=h}} \tag{5.1}$$

Les termes $\mathbb{1}_{z_i=h}$ sont incertains car, les attributs étant incertains, les feuilles dans lesquelles tombent les exemples sont alors aussi incertaines. Cette vraisemblance n'est donc pas directement maximisable, Périnel propose alors d'utiliser l'algorithme EM pour la maximiser [4].

Par rapport à la méthodologie proposée par Tsang et al., cette méthodologie conserve donc toujours ce défaut de représentation d'incertitude *épistémique* (des données) par des distributions de probabilité (adéquates pour modéliser les incertitudes *aléatoires*). Cependant, au lieu d'un fractionnement (assez discutable) des exemples dans les différentes feuilles d'un arbre, cette méthodologie propose d'apprendre le modèle correspondant le mieux aux données (incertaines). Nous reprendrons ensuite cette idée dans la méthodologie principale présentée dans ce manuscrit (voir Chapitre 7).

5.2. Approches probabilités imprécises

Dans le cadre de la théories des probabilités imprécises, certains chercheurs utilisent des *ensembles crédaux* (voir Section 1.2.3)au lieu des simples fréquences dans les estimations des probabilités feuilles et des classes dans les feuilles (voir Equations 4.5, 4.6). Ces *ensembles crédaux* peuvent être obtenus de différentes façons. Un moyen classique est d'utiliser le MDI (voir Section 1.2.3, Equations (1.13), (1.14)). L'utilisation d'un ensemble d'estimateurs de ces probabilités, au lieu d'un seule, implique l'obtention potentielle d'un ensemble de prédictions en sortie du modèle.

Pour ne pas avoir à gérer cet ensemble de prédictions, Abellan et Moral proposent dans [75] d'utiliser des ensembles crédaux (obtenu par le MDI) et de choisir comme critère de coupure le maximum d'entropie obtenu sur l'*ensemble crédal* du MDI.

En considérant une feuille t_h contenant n_h exemples avec $\forall k \in \{1, \ldots, K\}\ n_h^k$ le nombre d'exemples de classe w_k dans t_h et n la taille de l 'échantillon d'apprentis-

sage, on peut définir, à l'aide du MDI, les ensembles crédaux $\mathscr{P}_Z$ et $\mathscr{P}_{Y|Z}$ relativement aux probabilités de feuilles et de classe dans les feuilles :

$\mathscr{P}_Z = \{P : \forall h \in \{1,\ldots,H\}\ \frac{n_h}{n+s} \leq P(Z=h) \leq \frac{n_h+s}{n+s}\}$ et

$\mathscr{P}_{Y|Z} = \{P : \forall h \in \{1,\ldots,H\}\, \forall k \in \{1,\ldots,K\}\ \frac{n_h^k}{n_h+s'} \leq P(Y=w_k|Z=h) \leq \frac{n_h^k+s'}{n_h+s'}\}$

En reprenant ensuite le type de gain de pureté de l'Equation 4.4, mais pour le cas général d'une feuille donnant naissance à M nouvelles feuilles (M est donc le nombre de modalités de l'attribut utilisé dans la coupure), on onbtient :

$$\begin{aligned}
\Delta i &= i(t_h) - \frac{1}{\pi_h}\sum_{m=1}^{M}\pi_m i(t_{h_m}) \\
&= \max_{P_1\in\mathscr{P}_Z,\in P_2\in\mathscr{P}_{Y|Z}} [-\sum_{k=1}^{K} P_2(Y=w_k\,|\,Z=h)\log P_2(Y=w_k\,|\,Z=h) \\
&\quad + \frac{1}{P_1(Z=h)}\sum_{m=1}^{M} P_1(Z=h_m)\sum_{k=1}^{K} P_2(Y=w_k\,|\,Z=h_m)\log P_2(Y=w_k\,|\,Z=h_m)]
\end{aligned}$$

Cette méthodologie permet une nette amélioration de l'efficacité prédictive (voir [75]).

De manière à rendre aussi les arbres moins instables, Crossman et al. proposent dans [76] de choisir pour chaque potentielle nouvelle coupure, un ensemble de coupures (permettant un gain d'information supérieur à un certain seuil), de dupliquer l'arbre appris jusqu'ici et de continuer, pour chacune d'entre elles, l'apprentissage. Les prédictions des arbres obtenus sont ensuite combinées par *vote*. Le critère de coupure est là encore la maximisation du gain de pureté *maximal* des noeuds.

Même si ces méthodologies améliorent empiriquement l'efficacité et la stabilité des arbres, elles ne gèrent pas les entrées incertaines.

5.3. Approches floues

Les arbres de décisions flous apparaissent à la fin des années 70, donc peu de temps après le développement de la théorie des sous-ensembles flous et au même moment que l'apparition des premiers algorithmes d'arbres de décision.

Ils essaient de répondre à deux objectifs :

– adapter l'utilisation des arbres de décision aux données floues

– tirer profit de la logique floue pour augmenter les propriétées d'efficacité et/ou

de stabilité des arbres de décision usuels.

Comme vu en Section 7.4, les arbres de décision sont assez instables. Pour pallier à cet inconvénient, certains chercheurs ont développé des méthodologies d'arbres de décision *flous*. L'idée principale est de remplacer les coupures habituellement *précises* par des coupures dites *douces*. En effet, la coupure $X^j < 4.07$ est différente de la coupure $X^j < 4.09$. Ces deux coupures peuvent donc aboutir à des gains de pureté potentiellement très différents. Pour cette raison, en modifiant même très légèrement l'échantillon d'apprentissage, on peut obtenir des arbres très différents. L'idée de coupure *douce* revient à considérer, pour chaque coupure possible, un degré de satisfaction de la condition de coupure pour chaque exemple (au lieu de simplement considérer la satisfaction, ou non, de cette condition).

De nombreuses méthodologies d'arbres de décision flous ont été proposées [77], [78], nous nous intéressons ici à deux d'entre elles qui sont aujourd'hui considérées comme des références en la matière : le FID de Janikow et les $SoftDecisionTrees$ de Olaru et Wehenkel.

5.3.1. FID de Janikow

Dans un cadre initiale de régression, Janikow propose en 1998 une méthodologie d'arbre de décision flous : le FID [79, 80]. Janikow choisit une structure d'arbre du même type que C4.5 (non binaire). Une étape préliminaire de discrétisation des attributs numériques (dont la classe si régression) ou catégoriques à grand nombre de modalités est requise. Les données d'apprentissage sont ensuite *floutées*, cette méthodologie a donc l'avantage de gérer les données incertaines floues. Janikow utilise l'entropie des probabilités de classe dans les feuilles comme mesure d'impureté. Pour se faire il estime ces probabilités en utilisant la logique floue. Les attributs ayant été floutés, un même exemple tombe ici encore potentiellement dans plusieurs feuilles. Les prédictions des différentes feuilles seront alors combinées disjonctivement.

De nombreuses extensions de FID ont été proposées par Janikow [81]. Elles contiennent notamment des algorithmes d'élagage et de construction de forêts aléatoires floues [82].

5.3.2. Soft Decision Trees de Olaru et Wehenkel

En 2003 Olaru a proposé une méthodologie d'arbres de décision flous appelée "Soft Decision Trees" [83]. Les arbres construits sont binaires (type $CART$), ils ne gèrent que les attributs numériques et les classes à deux modalités. Les données d'apprentissage sont ici floutées dynamiquement pendant la construction des arbres. Comme Janikow, Olaru s'inscrit aussi dans un cadre de régression. Les sortie des arbres sont donc numériques et le critère de pureté est ici remplacé par la minimisation de l'erreur quadratique de prédiction.

Bilan : Le FID de Janikow, tout comme les $Soft\,Decision\,Trees$ d'Olaru améliorent tous les deux à la fois l'efficacité prédictive et la stabilité des arbres construits. Cependant, le FID ne gère les données incertaines qu'à condition qu'elles soient exprimées par des sous-ensemble flous.

5.4. Approches possibilistes

En théorie de l'information, Klir distingue, lui, deux types d'incertitude que sont la *non-spécificité* et le *conflit* [84]. La *non-spécificité* fait référence à une certaine méconnaissance du *vrai* modèle d'incertitude, une incertitude *non-spécifique* est donc compatible avec une multitude de modèles alors qu'une incertitude *spécifique* ne concerne qu'un modèle bien particulier. Le *conflit* d'une incertitude vient du nombre d'événements possibles et d'une certaine forme de *compétition* incertaine. Par exemple, la fonction de masse vide (voir Définition 8) est totalement *non-spécifique* et ne contient donc aucun *conflit* alors qu'une distribution de probabilité uniforme est totalement *spécifique* (voir Section 1.2.1) et contient un *conflit* maximum (entre tous les singletons qui ont donc le même poids).

Ces deux notions ont été largement utilisées par différents chercheurs. Dans [85], Borgelt et al. proposent des nouvelles mesures d'impureté basées sur la mesure de *non-spécificité* (mesure U) proposée par Higashi et Klir dans [86] (en tant que généralisation de l'information d'Hartley [87]) appliquée aux distributions de possibilité correspondant aux fréquences dans les noeuds.

Définition 16. Soit $\pi : \Omega^W \rightarrow [0,1]$ une distribution de possibilité telle que $|\Omega^W| < \infty$ avec $\forall i \in \{1,\ldots,|\Omega^W|-1\}, \exists w \in \Omega^W : \pi(w) = \pi_{(i)}$ et $\pi_{(i)} < \pi_{(i+1)}$. Sa *non-spécificité* est

alors donnée par $U(\pi) = \sum_{i=2}^{|\Omega^W|} \pi_{(i)} \log \frac{i}{i-1}$

Ce choix augmente l'efficacité des arbres obtenus. Dans [88], Ben-Amor et al. présentent une méthode de prédiction de la classe d'exemples pour lesquels les attributs sont incertains et représentés par des distributions de possibilité à partir d'arbres de décision classiques. Dans [89], Jenhani et al. proposent une méthodologie d'apprentissage d'arbre de décision à partir d'exemples dont les classes sont incertaines et exprimées à l'aide de distributions de possibilité. La mesure d'impureté proposée utilise les notion de *non-spécificité* et de proximité entre distributions de possibilité [90].

Toutes ces méthodologies s'inscrivent dans d'autres cadres d'incertitude que celui des probabilités classiques et permettant donc des modélisations d'incertitudes mal capturés par les modèles probabilistes standard (l'incertitude *épistémique* par exemple). En outre, l'utilisation de la notion de *non-spécificité* (voir [84]) permet une amélioration de l'efficacité prédictive des arbres. Il est cependant à noter qu'aucune de ces méthodologie ne peut gérer à la fois des données incertaines en entrée (attributs) et en sortie (classes), et qu'elles ne gèrent pas forcément les données incertaines représentées par des modèles plus larges (fonctions de croyance par exemple).

5.5. Approches crédibilistes

Dans le cadre de la théorie des fonctions de croyance, peu de méthodologies d'arbres de décision ont jusqu'ici été proposées. Nous décrivons ici brièvement deux méthodologies permettant plus ou moins l'apprentissage d'arbres de décision à partir de données incertaines et représentées par des fonctions de croyance où tirant avantage analytiquement des fonctions de croyance.

Dans [91], Elouedi, Mellouli, et Smets proposent une méthodologie d'arbres de décision dans le cadre crédibiliste. Ils s'inscrivent dans le MCT (voir Section 2.3.2) Les classes des exemples peuvent être incertaines et sont représentées par des fonctions de croyance. L'idée principale est d'apprendre des arbres en prenant comme critère de coupure la maximisation de l'entropie des probabilités pignistiques moyennes des classes. Ils proposent aussi une autre approche (*l'approche conjonctive*) utilisant des distances entre fonctions de croyance.

Pour la prédiction de la classe d'un nouvel exemple, les attributs peuvent être incertains et crédibiliste. Les auteurs obtiennent ainsi des fonctions de croyance relatives à la classe pour chaque feuille, et combinent ensuite disjonctivement ces fonctions de croyance pour obtenir une unique fonction de croyance. Après avoir calculé la probabilité pignistique de cette fonction de croyance, la prédiction est ensuite la classe ayant la plus grande probabilité pignistique. Cette approche est intéressante mais le fait de transformer notre incertitude relative aux classes et exprimée par des fonctions de croyance en incertitude probabiliste (pignistique) est critiquable car on perd alors les avantages de modélisation crédibiliste.

Dans [92], Skarstein-Bjanger et Denoeux proposent une méthodologie d'arbres de décision gérant les classes incertaines crédibilistes. Ils se placent dans le cadre de l'algorithme $CART$, les arbres sont donc binaires (deux branches sont issues de chaque noeuds). Ils utilisent comme mesure d'impureté une combinaison linéaire, proposée par Klir dans [86], de la U-mesure de *non-spécificité* adaptée aux fonctions de croyance avec une mesure de *conflit* :

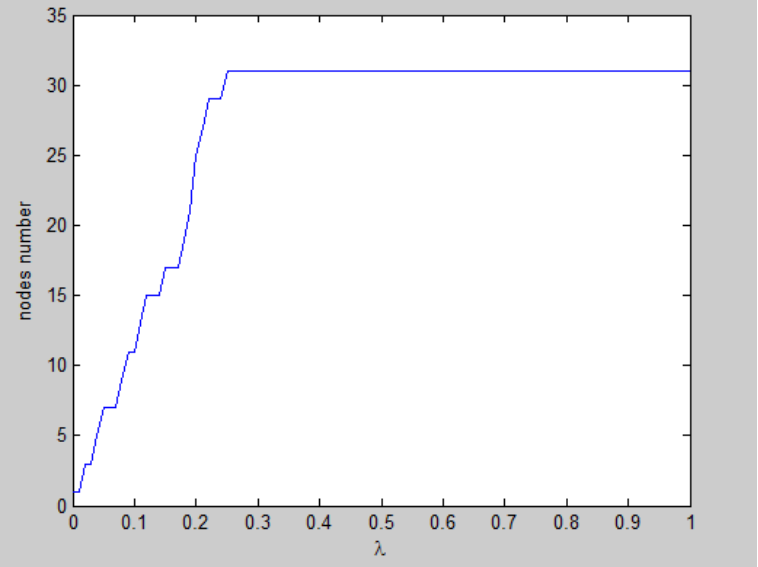
$$U_\lambda(m) = (1-\lambda)N(m) + \lambda D(m) \tag{5.2}$$

où $\lambda \in [0,1]$ est un hyper-paramètre, $N(m) = \sum_{A \subseteq \mathscr{Y}} m(A)\log_2|A|$ mesure la *non-specificité* et $D(m) = -\sum_{A \subseteq \mathscr{Y}} m(A)\log_2 BetP(A)$ le *conflit*.

En utilisant U_λ comme mesure d'impureté $i(t)$, le gain d'information Δi calculé selon (4.4) peut être négatif. Cela constitue un critère d'arrêt naturel pendant la construction de l'arbre, aucune coupure n'est ainsi effectuée quand tous les gains possibles d'information sont négatifs. Même si Klir avait proposé de fixer λ à 0.5, λ est généralement déterminé par validation croisée.

Cette mesure d'impureté U_λ est calculée sur la masse de croyance m^y relative à la classe obtenue en appliquant le modèle prédictif de Dempster pour l'expérience de Bernoulli adapté au MCT (voir Section 2.3.2). Ce modèle est défini uniquement pour deux classes (i.e. $|\Omega_Y| = 2$), il en sera donc de même de cette méthodologie. On notera désormais $\Omega_Y = \{A, B\}$. Pour un noeud de taille n, contenant n_A exemples

FIGURE 5.1.: Nombre de noeuds en fonction de λ pour le jeu de données Pima



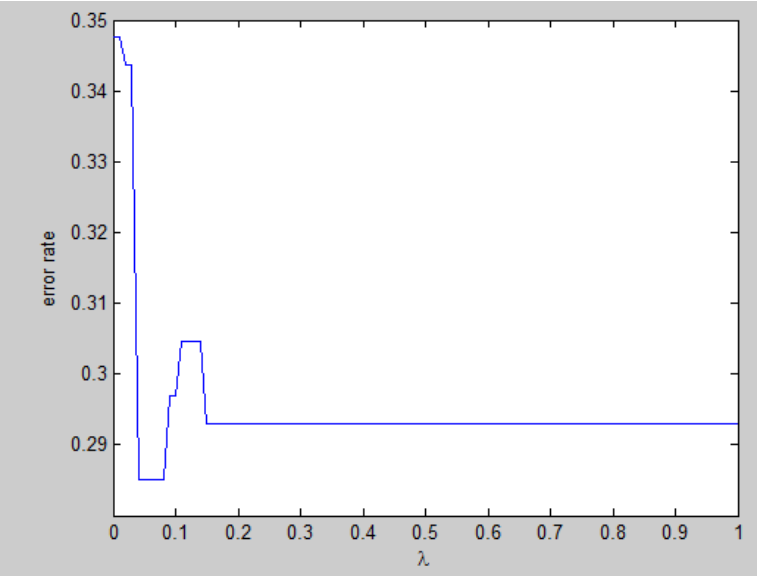
de classe A et n_B exemples de classe B, on aura donc

$$\begin{cases} m^y(\{A\}) & = & \frac{n_A}{n+1} \\ m^y(\{B\}) & = & \frac{n_B}{n+1} \\ m^y(\{A,B\}) & = & \frac{1}{n+1} \end{cases} \tag{5.3}$$

Il est intéressant de remarquer que le modèle de Dempster pour l'expérience de Bernoulli appliqué à la mesure de pureté des feuilles d'un arbre de décision permet de prendre en compte le nombre d'exemples présents dans une feuille (i.e. la *taille* des feuilles). En effet, l'Equation (5.3) implique que plus *grande* sera une feuille, moins grande sera la masse allouée à l'espace entier Ω_W et donc moins *incertaine* sera la masse m^W. Cette prise en compte de la taille des feuilles est réalisée par la mesure de *non-spécificité* N. En effet, cette fonction appliquée à la masse de croyance proposée ci-dessus par Dempster donne $N(m^y) = \frac{1}{n+1}$, le terme $(1-\lambda)$ peut donc être interprété comme un indicateur de l'importance donnée au manque d'exemples dans une feuille.

Les Figures 8.1 et 8.2 montrent l'impact du paramètre λ sur la complexité (en

FIGURE 5.2.: taux d'erreur en fonction de λ pour le jeu de données Pima



nombre de noeuds) de l'arbre obtenu et sur son efficacité sur le jeu de donnée $Pima$. Nous constatons que cette complexité augmente avec λ. Ceci suggère que l'optimisation de ce paramètre devrait intégrer la complexité des arbres comme critère. Le paramètre λ semble n'avoir qu'une faible influence sur l'efficacité des arbres.

Il est à noter que cette méthodologie gère les classes incertaines représentées par des fonctions de croyance.

De manière à évaluer cette méthode dans le cadre de données précises, des expériences sont menées en Section 8.1.1 en comparant l'efficacité prédictive des arbres utilisant $U_\lambda(m^y)$ avec des arbres de décision classique $CART$. Globalement, ces expériences montrent que la méthodologie SBD peut rivaliser avec $CART$ en terme d'efficacité.

De manière générale, dans le cadre de la théorie des fonctions de croyance , les méthodologies d'arbres de décision existantes font intervenir les fonctions de croyance à différents niveaux. Au niveau des données d'apprentissage, les deux méthodologies ici présentées gèrent les classes incertaines, uniquement pour le cas à deux classes pour la méthodologie de Skarstein-Bjanger et Denoeux. Au niveau du

modèle, alors que pour la méthodologie proposée par Elouedi, Mellouli, et Smets une partie de l'information initiale crédibiliste est perdue au moment du passage à la probabilité pignistique, les arbres de Skarstein-Bjanger et Denoeux ont l'avantage de tirer profit du cadre crédibiliste en prenant la taille des feuilles en compte au moment des calculs de gain de pureté des noeuds potentiels. Aucune de ces méthodologies ne gère cependant les données d'apprentissage incertaines en entrée (attributs).

5.6. Bilan

Le tableau 5.1 représente un bilan de cet état de l'art sur les arbres de décision mettant en jeu de l'incertitude.

Auteurs	Cadre de travail	Attributs incertains	Classes incertaines	Modèle d'incertitude imprécis
Tsang et all. [74]	probabiliste	distributions de probabilité → fractionnement des exemples ou moyenne des distributions	non	
Périnel, Diday et Ciampi [4]	probabiliste	distributions de probabilité → EM		critère de coupure : maximiser $L(\theta_H, (y \mid x))$
Abellan et Moral [75]	probabiliste imprécis			MDI → ensembles crédaux → mesure d'impureté : max d'entropie
Janikow [79]	flou	sous-ensembles flous	sous-ensembles flous	logique floue
Olaru [83]	flou			critère de coupure : minimiser l'erreur quadratique de prédiction (classe numérique)
Borgelt et al. [85]	possibiliste			mesure d'impureté : *non-spécificité*
Jenhani et al. [89]	possibiliste		distributions de possibilités	mesure d'impureté : *non-spécificité*
Elouedi, Mellouli, et Smets [91]	crédibiliste		fonctions de croyance	critère de coupure : maximisation de l'entropie de la probabilité pignistique
Skarstein-Bjanger et Denoeux [92]	crédibiliste		fonctions de croyance	mesure d'impureté : combinaison de la *non-spécificité* et du *conflit* (U_λ)

TABLE *5.1.*: Récapitulatif sur les arbres incertains

De nombreuses méthodologies d'arbres de décision ont été proposées dans différentes théories qui tentent de gérer et de tirer profit de différentes incertitudes, à différents niveaux : incertitude des données pendant l'apprentissage et au moment de la prédiction (de la classe d'un exemple à partir de ses attributs ici incertains), à l'intérieur du modèle dans le calcul des gains de pureté.

Dans le premier cas, il s'agit généralement d'une incertitude de type épistémique. En effet, l'imperfection des observations rend souvent incomplète notre connaissance des *vraies* données. Peu de méthodologies permettent de gérer les entrées incertaines. Celles qui le font sont souvent restreintes à des cadres de modélisation limités (probabilistes, flous) et font quelques fois des choix assez questionables pour y parvenir (fractionnement des exemples dans différentes feuilles d'un même arbre).

Au niveau du modèle, l'incertitude est principalement aléatoire. L'utilisation de fréquences dans nombres de modèles prédictifs fait sens surtout pour des échantillons de tailles suffisamment grandes et le fait de faire ainsi tendre cette taille à l'infini suppose des modèles génératifs principalement construits à partir de variables purement aléatoires. C'est à ce niveau que la plupart de ces méthodologies tirent avantage des différentes incertitudes.

Pour les arbres de Skarstein-Bjanger et Denoeux, on pourrait même parler d'incertitude épistémique relative à de l'incertitude aléatoire. Le modèle de Dempster (voir Section 2.3.2) permet ainsi de diminuer la confiance attribuée aux fréquences utilisées dans les calculs de gains de pureté pour des feuilles trop petites. Cette baisse de confiance est épistémique, on estime que pour des petites feuilles on n'a pas suffisamment d'information. L'incertitude aléatoire provient de l'utilisation fréquentiste des proportions (éventuellement modifiées en $\frac{1}{n+1}$) comme estimateurs des probabilités de classes.

Certaines méthodologies proposent des solutions pour prédire la classe d'exemples dont les attributs seraient incertains. Pour les arbres de décision flous, il peut même être souhaitable de partir d'attributs déjà floutés, une étape préliminaire de floutage des données précises étant souvent nécessaire. En général, les différentes types de prédiction à partir de données incertaines sont aisés à mettre en oeuvre mais ne sont pas forcément optimaux. Ce problème reste un problème assez ouvert où de nombreuses solutions pourraient à l'avenir être étudiées.

CHAPITRE 6

Extension de la méthodologie Skarstein-Bjanger et Denoeux (SBD) au cas multi-classes

La méthodologie d'arbres de décision proposée par Skarstein-Bjanger et Denoeux [92] utilise la mesure d'impureté crédibiliste U_λ proposée par Klir dans [86] et appliquée la fonction de croyance issue de l'approche de Dempster à l'expérience de Bernoulli adapté au MCT. Cette approche étant définie uniquement dans le cas de deux classes, nous proposons ici son extension au cas multi-classe de trois manières différentes :

- en transformant les échantillons à plus de deux classes en plusieurs échantillons à deux classes, en construisant un arbre par échantillon d'apprentissage ainsi obtenu, puis en combinant les prédictions obtenues par ces différents arbres
- en remplaçant l'approche de Dempster à l'expérience de Bernoulli par le Modèle de Dirichlet Imprécis (MDI)
- en utilisant le modèle multinomial de Denoeux

Sommaire

6.1. Combinaison de classifieurs binaires selon la méthodologie de *Quost et al.*

Dans [93], Quost et al. présentent une méthode permettant de résoudre un problème de classification multi-classe en combinant des classifieurs entraînés sur des échantillons à deux classes. Ils proposent d'apprendre, pour chaque paire de classes $\{w_i, w_j\}$, une fonction de croyance sur la classe y conditionnellement à cette paire de classes (on obtient $m_{ij} = m^y(. \mid y \in \{w_i, w_j\})$), et de les combiner en une fonction de croyance globale sur Ω_Y en utilisant une procédure d'optimisation.

Nous proposons ici d'utiliser cette méthode à partir des arbres de décision de Skarstein-Bjanger et Denoeux, utilisant ces derniers comme classifieurs de base pour apprendre les fonctions de croyance conditionnelles.

Cette méthode est différente de celle présentée par Vannoorenberghe et Denoeux ([94]) où K arbres à 2 classes sont construits considérant pour chacun d'eux "une classe contre les autres" puis dont les fonctions de croyance de sortie sont combinées en faisant la moyenne des K masses étendues (de $\{w_k; \overline{w_k}\}$ dans Ω_Y).

Les arbres de décision sont bien adaptés à ce genre de combinaisons de par leur simplicité. Cependant, on peut remarquer que l'optimisation sur λ devient alors problématique, étant donné que $K(K-1)/2$ classifieurs doivent être appris à chaque étape de l'optimisation.

6.2. Modèle de Dirichlet Imprécis (*MDI*)

Le MDI fut introduit dans un cadre de "probabilités imprécises" par Walley [5] (voir Section 1.2.3). On peut vérifier que les probabilités inférieures $\underline{P}$ et supérieures $\overline{P}$ auxquelles ce modèle aboutit constituent une fonction de croyance : en posant $Bel^w = \underline{P}$ et $Pl^w = \overline{P}$, on aboutit (via les Equations 2.4 et 2.5) ainsi à la masse

de croyance

$$\begin{cases} m^w(\{w_1\}) &= \frac{n_1}{n+s} \\ \quad\vdots \\ m^w(\{w_K\}) &= \frac{n_K}{n+s} \\ \\ m^w(\Omega_Y) &= \frac{s}{n+s} \end{cases} \tag{6.1}$$

où n_k est le nombre d'occurrences de la classe w_k dans la feuille considérée $k = 1,\ldots,K$.

On peut remarquer que l'on retrouve le modèle de Dempster (Equation 2.12) pour K=2 et s=1. En utilisant m^w ainsi définie, U_λ peut alors calculer l'impureté d'un noeud contenant plus de deux classes et des arbres multi-classes peuvent alors être créés. La forme analytique de U_λ appliquée à m^w est alors :

$$U_\lambda(m^w) = \frac{(1-\lambda)s}{n+s}\log_2(K) - \frac{\lambda}{n+s}\sum_{k=1}^{K} n_k \log_2\left[\frac{Kn_k+S}{K(n+s)}\right] \tag{6.2}$$

Ce modèle est facile à mettre en oeuvre, m^w ne contenant que $K+1$ ensembles focaux. Cependant le fait qu'il ait été défini dans le cadre de la théorie des probabilités imprécises rend son interprétation difficile dans un cadre tel que le MCT. Il est aussi à noter que l'imprécision de m^y (contenue dans le terme $m^w(\Omega_Y)$) ne dépend que de la taille de l'échantillon n, et pas de sa distribution de probabilité sur Ω_Y. Ceci n'est pas le cas pour le modèle multinomial de Denœux qui offre d'intéressantes alternatives.

6.3. Modèle multinomial de Denoeux

Dans [54], Denoeux propose d'utiliser les intervalles de confiance de Goodman [95] pour construire une fonction de croyance prédictive. La première étape est la construction d'intervalles de probabilité [96] (probabilités inférieures et supérieures sur les singletons) puis de les transformer en fonctions de croyance. Pour un échantillon iid $(y_1,\ldots,y_n) \in \Omega_Y^n$, ces intervalles de probabilité $[\underline{P}_k, \overline{P}_k]$ sont don-

nés, pour Y_k (k=1,...,n), par :

$$\underline{P}_k = \frac{q + 2n_k - \sqrt{\Delta_k}}{2(n+q)} \quad et \quad \overline{P}_k = \frac{q + 2n_k + \sqrt{\Delta_k}}{2(n+q)} \quad , \tag{6.3}$$

où q est le quantile d'ordre $1-\alpha$ de la loi du chi-deux à un degré de liberté, et où $\Delta_k = q(q + \frac{4n_k(n-n_k)}{n})$.

Comme démontré dans [54], la mesure de confiance inférieure

(i.e., $\underline{P}(A) = max(\sum_{w_k \in A} \underline{P}_k, 1 - \sum_{w_k \notin A} \underline{P}_k)$) obtenue par ces intervalles dans les cas $K = 2$ où 3 est une fonction de croyance.

On peut remarquer que les fonctions de croyance obtenues suivent le principe de Hacking (voir [54] pour plus de détails), mais que la solution pour $K = 2$ n'est pas équivalente à celle de l'Equation 2.12.

Dans le cas $K > 3$, l'inverse de Möbius (voir Equations (2.4) et (2.5)) de $\underline{P}$ peut être négative et n'est donc pas une fonction de croyance en général. Différentes méthodes incluant de la programmation linéaire sont proposées dans [54] pour l'approximer par une fonction de croyance. En outre, dans le cas particulier où les classes sont ordinales, Denœux propose un algorithme restreint à un certain ensemble d'éléments focaux. Une fonction de croyance valide est ainsi obtenue et peut être utilisée par U_λ pour mesurer la pureté des noeuds et créer alors des arbres multi-classes.

6.4. Bilan

La méthodologie d'arbres de décision *SBD* initialement prévue pour deux classes, est ici étendue aux cas de données à plus de deux classes. Cette extension peut se faire de trois manières différentes. Les combinaisons d'arbres à deux classes (selon Quost et al. [93]) sont une solution mais présente le défaut de générer plusieurs arbres de décision, ce qui rend compliquée l'interprétation de ces derniers. Le *MDI* (de Walley [5]) représente aussi une alternative ressemblant à une extension de l'approche de Dempster à l'expérience de Bernoulli au cas multi-classes, cependant ce modèle est défini dans un autre cadre que la théorie des fonctions de croyance, son interprétation crédibiliste est donc difficile. Le modèle multinomial de Denoeux [54], permet lui de construire des fonctions de croyance multi-classes définies et justifiées en théorie des fonctions de croyance. Pour les cas de deux ou trois classes

ce modèle définit explicitement une fonction de croyance alors que pour plus de trois classes, des approximations sont requises dans le cas général.

Des expériences sont présentées en Section 8.1 de manière à comparer les efficacités de ces différentes extensions possibles avec celles obtenues par des arbres *CART* classiques. Ces expériences sont réalisées à partir de jeux de données *précis* benchmark.

CHAPITRE 7

Arbres de décision E^2M

La méthodologie d'apprentissage principale proposée dans ce manuscrit est ici présentée. Il s'agit d'arbres de décisions construits à partir de *données imparfaites* ou *incertaines*. Plus précisément, l'incertitude des données, généralement de nature épistémique, est modélisée de façon crédibiliste (i.e. par des fonctions de croyance) et les différents paramètres des arbres sont estimés par l'algorithme E^2M. Après avoir résumé la problématique générale des données incertaines, la méthodologie générale proposée est exposée, une méthode d'élagage par évaluation incertaine via l'algorithme E^2M est explicitée, et enfin la prédiction par un arbre E^2M est expliquée.

Sommaire

7.1. Description du problème des données imparfaites

La plus grande partie des problématiques d'apprentissage automatique partent du principe qu'on dispose d'un échantillon d'apprentissage *certain* ou *précis*. Cependant, dans beaucoup d'applications, les données sont entachées d'incertitude, on parle alors de *données imparfaites* ou *incertaines*. Cette incertitude peut être de différents types. Très souvent, ces données sont imprécises, cela pouvant venir d'approximations réalisées par différents capteurs ou appareils de mesure. Le cas extrême de l'imprécision des données est le cas de données manquantes. Ce dernier cas, connu de tous les praticiens, est souvent géré en supprimant les exemples incomplets (i.e. pour lesquels certaines valeurs d'attribut sont manquantes) même si d'autres méthodes ont aussi été proposées [57], [97], [98].

Il reste cependant que la plupart des méthodes gérant ces données incertaines les transforment en données *précises*. L'objectif principal de ce manuscrit est de montrer que cette incertitude des données contient de l'information en soi et qu'utiliser ces données telles quelles (i.e. dans leur forme incertaine) est préférable à leur transformation préalable en données précises du point de vue informatif. L'objectif minimal du présent manuscrit est en tout cas de montrer que ces données incertaines peuvent être utilisées lors d'un apprentissage automatique. Nous montrons ensuite que manier des données incertaines peut se faire assez naturellement à l'aide de fonctions de croyance. Pour se faire, dans une optique d'apprentissage d'arbres de décision, nous avons choisi d'utiliser l'algorithme E^2M pour les estimations requises pendant la phase d'apprentissage, cet algorithme permettant justement la maximi-

sation de critères de vraisemblance à partir de fonctions de croyance.

Il convient ici de préciser que d'autres choix auraient pu être fait pour représenter cette incertitude des données : distributions de probabilité, de possibilité ou ensemble de distributions de probabilité. Le choix crédibiliste que nous avons fait s'explique par le fait que les fonctions de croyance englobent les probabilités et possibilités et par le fait que le cadre de travail qu'elles proposent est à la fois très souple et permet une modélisation de tous les types d'incertitude (en tout cas ceux étudiés dans ce manuscrit) de manière adéquate.

Exemple 3. *En reprenant l'arbre de décision 4.1 et en posant* $\Omega_1 = [1,10]$ *et* $\Omega_2 = \{petit, moyen, grand\}$, *on considére un exemple* i *dont les attributs incertains sont donnés par*

$$\left\{\begin{array}{rcl} m_i^1(\{2\}) & = & 1/3 \\ m_i^1(\{3\}) & = & 1/3 \\ m_i^1([2,5]) & = & 1/3 \end{array}\right. et \left\{\begin{array}{rcl} m_i^2(\{petit\}) & = & 3/4 \\ m_i^2(\{petit, moyen\}) & = & 1/8 \\ m_i^2(\{moyen, grand\}) & = & 1/8 \end{array}\right. .$$

On ne peut a priori pas savoir dans quelle feuille i *termine.*

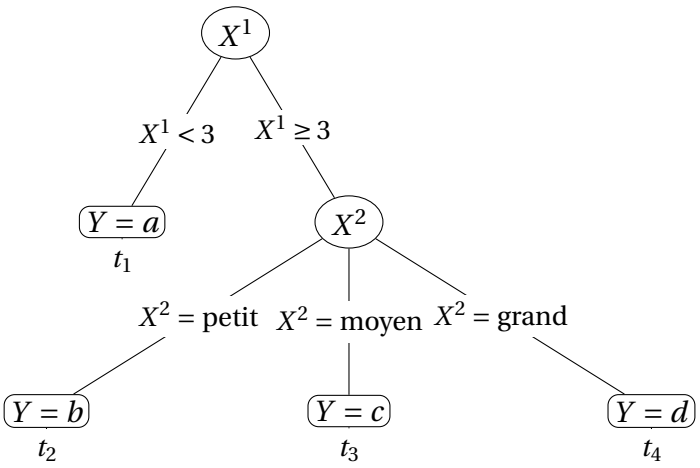


FIGURE *7.1.:* exemple d'arbre de décision

7.2. Méthodologie générale des arbres de décision E^2M

Après une description formelle, les différents calculs et résultats relatifs à la mise en œuvre de l'algorithme E^2M pendant la construction d'un arbre E^2M sont exposés. L'algorithme du choix de la meilleure coupure, pour un arbre donné, est ensuite

présenté et enfin l'algorithme général de construction d'un arbre E^2M est donné, dans un premier temps sans expliciter de technique d'élagage particulière.

7.2.1. Description formelle de la méthodologie

Dans les arbres de décision classiques, le critère à maximiser lors du choix d'une coupure est le gain de pureté Δi. Dans un cadre de données crédibilistes, les différentes puretés ne sont pas directement calculables. En effet, en considérant un arbre $\mathscr{P}_H$ de paramètre θ_H, pour un exemple dont les valeurs des attributs sont incertaines, on ne peut pas connaître de manière certaine la feuille t_h où cet exemple *finit* dans l'arbre. Même s'il aurait été possible d'annuler préalablement cette incertitude des données, nous choisissons de la conserver pendant toute la construction des arbres. Pour calculer les gains de pureté Δi, au lieu de simples estimations fréquentistes, nous estimons les probabilités des feuilles π_h et des classes dans les feuilles α_h^k en maximisant la vraisemblance crédibiliste $L(\theta_H; m^{z,y})$ du couple (Z, Y) (i.e. feuille, classe). L'échantillon d'apprentissage crédibiliste étant constitué de croyance relative au couple (X, Y) (i.e. attribut, classe), une étape préliminaire sera alors nécessaire pour exprimer les croyances relatives aux feuilles à partir de celles relatives aux attributs (de manière à expliciter $L(\theta_H; m^{z,y})$ à partir de m^{XY}). Cette vraisemblance crédibiliste sera ensuite maximisée à l'aide de l'algorithme E^2M.

7.2.2. L'algorithme E^2M appliqué à l'estimation du nouveau paramètre d'un arbre lors d'une coupure

Après avoir décrit la manière d'obtenir m^z à partir de m^x, les éléments de formalisme nécessaires à la présentation de la méthodologie sont exposés, quelques calculs préalables à la mise en œuvre de l'algorithme E^2M sont présentés, et enfin l'algorithme est détaillé.

Transfert des croyances relatives aux attributs relativement à un arbre de décision

Le calcul de la vraisemblance $L(\theta_H; m^{z,y})$ faisant intervenir notre croyance m^z (pour chaque exemple d'apprentissage) relative à l'appartenance aux différente feuilles d'un arbre, nous devons dans un premier temps expliciter m^z à partir de

m^x. En effet notre connaissance relative aux attributs d'un exemple peut directement se traduire en connaissance relative à l'appartenance de l'exemple aux feuilles d'un arbres (suivant les seuils ou ensembles définis par les différentes coupures de l'arbre) au simple prix de l'hypothèse d'indépendance cognitive (voir Définition 12) entre les attributs.

Cette hypothèse est toutefois assez naturelle car tout apport d'élément d'information relatif à un des attributs n'a a priori aucune raison de changer notre connaissance relative aux autres attributs (e.g. remplacer un capteur d'un ensemble de capteurs ne change pas la qualité des mesures effectuées par les autres capteurs).

D'après l'Equation 2.14, une vraisemblance crédibiliste d'une fonction de croyance ne dépend que de sa fonction de contour, pour expliciter $L(\theta_H; m^{z,y})$ nous n'utiliserons donc de $m^{z,y}$ que les fonctions de contour de m^z et de m^y. Pour un exemple i, ces fonctions de contour seront notées $pl_{ih} = Pl_i^z(\{h\})$ et $pl_i^k = Pl_i^Y(\{\omega_k\})$. pl_{ih} sera donc la plausibilité d'appartenance de l'exemple i à la feuille t_h et pl_i^k la plausibilité que l'exemple i soit de classe ω_k.

En reprenant le formalisme du chapitre 4, et donc en définissant toute feuille t_h comme un élément de la partition de Ω_X, on pourra écrire $t_h = \{A_h^1 \times, \ldots, \times A_h^J\}$ où $\forall j = 1, \ldots, J \quad A_h^j \subseteq \Omega_j$, on aura alors $\forall i \in \{1, \ldots, N\}$,

$$\begin{aligned} pl_{ih} &= Pl_i^Z(\{t_h\}) \\ &= Pl_i^X(\times_{j=1}^J A_h^j) \\ &= \prod_{j=1}^J Pl_i^{X^j}(A^j) \text{ par indépendance cognitive des attributs.} \end{aligned} \tag{7.1}$$

Exemple 4. *En reprenant l'exemple 3 et en considérant la classe incertaine de l'exemple i*

donnée par $\begin{cases} m_i^y(\{c\}) & = & \frac{4}{5} \\ m_i^y(\{a,b,c,d\}) & = & \frac{1}{5} \end{cases}$ *, on obtient*

$$\begin{aligned}
pl_{i1} &= Pl_i^z(\{1\}) = Pl_i^1([1,3[) \\
&= m_i^1(\{2\}) + m_i^1([2,5]) = \frac{2}{3} \\
pl_{i2} &= Pl_i^z(\{2\}) = Pl_i^1([3,10]]) \times Pl_i^2(\{petit\}) \\
&= [m_i^1(\{3\}) + m_i^1([2,5])] \times [m_i^2(\{petit\}) + m_i^2(\{petit, moyen\})] = \frac{2}{3} \times \frac{7}{8} = \frac{7}{12} \\
pl_{i3} &= Pl_i^z(\{3\}) = Pl_i^1([3,10]]) \times Pl_i^2(\{moyen\}) \\
&= [m_i^1(\{3\}) + m_i^1([2,5])] \times [m_i^2(\{petit, moyen\}) + m_i^2(\{moyen, grand\})] = \frac{1}{6} \\
pl_{i4} &= Pl_i^z(\{4\}) = Pl_i^1([3,10]]) \times Pl_i^2(\{grand\}) \\
&= [m_i^1(\{3\}) + m_i^1([2,5])] \times m_i^2(\{moyen, grand\}) = \frac{1}{12}
\end{aligned}$$

et

$$\begin{aligned}
pl_i^1 &= Pl_i^y(\{a\}) = m_i^y(\{a,b,c,d\}) = \frac{1}{5} \\
pl_i^2 &= Pl_i^y(\{b\}) = m_i^y(\{a,b,c,d\}) = \frac{1}{5} \\
pl_i^3 &= Pl_i^y(\{c\}) = m_i^y(\{c\}) + m_i^y(\{a,b,c,d\}) = 1 \\
pl_i^4 &= Pl_i^y(\{d\}) = m_i^y(\{a,b,c,d\}) = \frac{1}{5}
\end{aligned}$$

Remarque : L'utilisation de la fonction de contour pl^z suppose l'existence de la fonction de croyance m^z. Il est en réalité possible d'expliciter directement m^z en fonction de m^x par

$$m^z(\{1,\ldots,H'\}) = \sum_{\substack{A \subseteq \bigcup_{h=1}^{H'} A_h \\ A \not\subset \bigcup_{g \in G} A_g \\ |G| = H'-1}} m^x(A) \tag{7.2}$$

La preuve est laissée au lecteur. Nous n'aurons pas besoin d'utiliser directement m^z, nous nous restreindrons donc à utiliser sa fonction de contour pl^z telle qu'explicitée dans l'Equation 7.1.

Formalisme et calculs préliminaires à la mise en œuvre de l'algorithme E^2M

On considère ici un arbre à H feuilles $\mathscr{P}_H$ dont le paramètre est noté

$$\theta_H = \begin{pmatrix} \alpha_1^1 & \cdots & \alpha_1^K & \pi_1 \\ \vdots & \ddots & \vdots & \vdots \\ \alpha_H^1 & \cdots & \alpha_H^K & \pi_H \end{pmatrix} \in \Theta_H.$$

Du fait de l'incertitude des données, on peut parler de *variables latentes ou cachées* exprimant d'une part l'appartenance des exemples aux différentes classes, et d'autre part leur appartenance aux différentes feuilles de l'arbre. On les définira de la manière suivante :

- $y_i^k = \begin{cases} 1 & \text{si le } i^{\text{ème}} \text{ exemple est de classe } \omega_k \\ O & \text{sinon} \end{cases}$
- $z_{ih} = \begin{cases} 1 & \text{si le } i^{\text{ème}} \text{ exemple appartient à la feuille } t_h \\ O & \text{sinon} \end{cases}$

Ensuite, les fonctions de contour des différentes classes $Pl^y(\{\omega_k\})$ s'obtiennent directement à partir des connaissances qu'on a de ces classes, i.e. de m^y, on notera pl_i^k la plausibilité que le $i^{\text{ème}}$ exemple soit de classe ω_k.

On pourra donc écrire $pl_i^k = Pl_i^Y(\{\omega_k\}) = Pl_i^Y(y_i^k = 1)$.

1. En reprenant le raisonnement de l'exemple de *clustering de données catégo-*

riques de [56] (mais en ne considérant qu'un attribut : la *feuille*), on obtient :

$$
\begin{aligned}
L(\theta_H; m^{z,y}) &= E_{\theta_H}[pl^{Z,Y}(Z,Y)] = E_{\theta_H}[\prod_{i=1}^{n} pl_i^{Z,Y}(Z,Y)] \\
&\quad \text{par indépendance cognitive entre les exemples} \\
&= \prod_{i=1}^{n} E_{\theta_H}[pl_i^{Z,Y}(Z,Y)] \\
&\quad \text{par indépendance stochastique entre les exemples} \\
&= \prod_{i=1}^{n} E_{\theta_H}[pl_i^{Y}(Y) pl_i^{Z}(Z)] \\
&\quad \text{par indépendance cognitive entre } Z \text{ et } Y \\
&= \prod_{i=1}^{n} \sum_{h=1}^{H} E_{\theta_H}[pl_i^{Y}(Y) pl_i^{Z}(Z) \mid z_{ih} = 1]\pi_h \\
&= \prod_{i=1}^{n} \sum_{h=1}^{H} E_{\theta_H}[pl_i^{Z}(Z) \mid z_{ih} = 1].E_{\theta_H}[pl_i^{Y}(Y) / z_{ih} = 1]\pi_h \\
&\quad \text{par indépendance cognitive entre } Z \text{ et } Y \\
&= \prod_{i=1}^{n} \sum_{h=1}^{H} E_{\theta_H}[pl_i^{Z}(Z = h) \mid z_{ih} = 1]. \sum_{k=1}^{K} pl_i^{Y}(Y = \omega_k) P(Y = \omega_k \mid z_{ih} = 1)\pi_h \\
&= \prod_{i=1}^{n} \sum_{h=1}^{H} pl_{ih}\pi_h \sum_{k=1}^{K} pl_i^k \alpha_h^k \qquad (7.3)
\end{aligned}
$$

Dans le cas de données précises, ($pl_{ih} = 1$ pour une unique feuille t_h, 0 pour les autres), on retrouve la vraisemblance *classique*.

2. Etant donné un arbre $\mathscr{P}_H$ de paramètre θ_H, la probabilité d'observer (y_i, z_i) s'écrit :

$$
\begin{aligned}
P_\theta(y_i, z_i) &= P_\theta(z_i) P_\theta(y_i \mid z_i) \\
&= [\prod_{h=1}^{H} \pi_h^{z_{ih}}].[\prod_{h=1}^{H} P(y_i \mid z_{ih} = 1; \theta)^{z_{ih}}] \\
&= \prod_{h=1}^{H} (\pi_h \prod_{k=1}^{K} \alpha_h^{k\, y_i^k})^{z_{ih}} \qquad (7.4)
\end{aligned}
$$

3. La plausibilité de l'échantillon précis $E = (y, z)$ (voir Equation 4.1) se factorise

en :

$$\begin{aligned} pl(y,z) &= \prod_{i=1}^{n} pl_i^{Y,Z}(y_i,z_i) \quad \textit{independance cognitive entre les exemples} \\ &= \prod_{i=1}^{n} pl_i^{Y}(y_i)\, pl_i^{Z}(z_i) \quad \textit{independance cognitive entre } Y \textit{ et } Z \\ &= \prod_{i=1}^{n} [\prod_{h=1}^{H} (pl_{ih})^{z_{ih}} \prod_{k=1}^{K} (pl_i^k)^{y_i^k}] \end{aligned} \tag{7.5}$$

4. Finalement, l'espérance de la log-vraisemblance de θ à l'itération q s'écrit (voir [56] pour les détails) :

$$\begin{aligned} Q(\theta;\theta^{(q)}) &= E[\log L(\theta,y,z) \mid m^{z,y};\theta^{(q)}] \\ &= \sum_{i,h} t_{ih}^{(q)} \log \pi_h + \sum_{i,h,k} \beta_{ih}^{k(q)} \log \alpha_h^k \end{aligned} \tag{7.6}$$

où $t_{ih}^{(q)} = E[z_{ih} \mid m^{z,y};\theta^{(q)}]$ and $\beta_{ih}^{k(q)} = E[z_{ih} y_i^k \mid m^{z,y};\theta^{(q)}]$

Mise en œuvre de l'algorithme E^2M

Etape Estimation (E) :

Les termes $t_{ih}^{(q)}$ et $\beta_{ih}^{k(q)}$ seront ici exprimés à partir des données et des paramètres estimés à l'étape précédente, i.e. à partir des pl_{ih}, pl_i^k et des $\pi_h^{(q)}, \alpha_h^{k(q)}$.

Pour commencer,

$$\begin{aligned} t_{ih}^{(q)} &= E[z_{ih} \mid m^{z,y};\theta^{(q)}] \\ &= P(z_{ih}=1 \mid m^{z,y};\theta^{(q)}) \end{aligned} \tag{7.7}$$

Il est ici impossible d'utiliser la définition de $P(. \mid m^{z,y};\theta^{(q)})$ car cette probabilité est uniquement définie pour le couple (Z,Y),i.e. pour des événements qui sont des sous-ensembles de $\Omega_Z \times \Omega_Y$, et z_{ih} ne dépend que de Z et est donc uniquement sous-ensemble de Ω_Z. Ce terme doit donc d'abord être étendu à $\Omega_Z \times \Omega_Y$ pour pouvoir être explicité, ensuite on pourra obtenir le résultat voulu en marginalisant sur Ω_Z.

$$P(y_i, z_{ih}=1 \mid m^{z,y}; \theta^{(q)}) = \frac{P_{\theta^{(q)}}(y_i, z_{ih}=1)\, pl(y_i, z_{ih}=1)}{L(\theta^{(q)}, m_i^{Z,Y})} \tag{7.8}$$

où

$$\begin{cases} P_\theta(y_i, z_{ih}=1) = P_\theta(y_i \mid z_{ih}=1) P_\theta(z_{ih}=1) = \left(\prod_{k=1}^{K} \alpha_h^{k\, y_i^k} \right) . \pi_h \\ pl(y_i, z_{ih}=1) = pl(y_i)\, pl(z_{ih}=1) = \left(\prod_{k=1}^{K} pl_i^{k\, y_i^k} \right) . pl_{ih} \\ L(\theta, m_i^{z,y}) = E_\theta[pl(y_i)\, pl(z_{ih}=1)] = \sum_{h=1}^{H} pl_{ih} \pi_h \sum_{k=1}^{K} pl_i^k \alpha_h^k \end{cases} \tag{7.9}$$

On obtient

$$\begin{aligned} P(y_i, z_{ih}=1 \mid m^{z,y}; \theta^{(q)}) &= \frac{\prod_{k=1}^{K} \alpha_h^{k(q)\, y_i^k} . \pi_h^{(q)} \prod_{k=1}^{K} pl_i^{k\, y_i^k} . pl_{ih}}{\sum_{h=1}^{H} pl_{ih} \pi_h^{(q)} \sum_{k=1}^{K} pl_i^k \alpha_h^{k(q)}} \\ &= \frac{\pi_h^{(q)} pl_{ih} \prod_{k=1}^{K} (\alpha_h^{k(q)} pl_i^k)^{y_i^k}}{\sum_{h=1}^{H} pl_{ih} \pi_h^{(q)} \sum_{k=1}^{K} pl_i^k \alpha_h^{k(q)}} \end{aligned} \tag{7.10}$$

Puis en marginalisant sur Ω_Z, on obtient :

$$t_{ih}^{(q)} = P(z_{ih}=1 \mid m^{z,y}; \theta^{(q)}) = \frac{\pi_h^{(q)} pl_{ih} \sum_{k=1}^{K} \alpha_h^{k(q)} pl_i^k}{\sum_{h=1}^{H} pl_{ih} \pi_h^{(q)} \sum_{k=1}^{K} pl_i^k \alpha_h^{k(q)}} \tag{7.11}$$

Avec le même raisonnement pour les termes $\beta_{ih}^{k(q)}$, on obtient :

$$\begin{aligned}
\beta_{ih}^{k(q)} &= E[z_{ih} y_i^k / m^{z,y}; \theta^{(q)}] \\
&= P(z_{ih}=1, y_i^k = 1 / m^{z,y}; \theta^{(q)}) \\
&= \frac{P_{\theta^{(q)}}(z_{ih}=1, y_i^k=1).pl(z_{ih}=1, y_i^k=1)}{L(\theta, m_i^{Z,Y})} \\
&= \frac{[P_{\theta^{(q)}}(y_i^k=1/z_{ih}=1)P_{\theta^{(q)}}(z_{ih}=1)].[pl(z_{ih}=1).pl(y_i^k=1)]}{L(\theta, m_i^{Z,Y})} \qquad (7.12) \\
&= \frac{\alpha_h^{k(q)} \pi_h^{(q)}.pl_{ih} pl_i^k}{\sum_{h=1}^{H} pl_{ih}\pi_h^{(q)} \sum_{k=1}^{K} pl_i^k \alpha_h^{k(q)}} \qquad (7.13)
\end{aligned}$$

Etape Maximisation (M) :

Cette dernière étape étant assez simple mais longue, elle n'a que peu d'intérêt à être exposée ici, elle sera donc explicitée en Annexe C et résumée dans le bilan qui suit.

Bilan

$$\begin{aligned}
L(\theta_H; m^{z,y}) &= \prod_{i=1}^{N} \sum_{h=1}^{H} pl_{ih}\pi_h \sum_{k=1}^{K} pl_i^k \alpha_h^k \qquad (7.14) \\
Q(\theta_H, \theta_H^{(q)}) &= \sum_{i,h} t_{ih}^{(q)} \log \pi_h + \sum_{i,h,k} \beta_{ih}^{k(q)} \log \alpha_h^k \qquad (7.15)
\end{aligned}$$

et le maximum de $Q(\theta_H, \theta_H^{(q)})$ est atteint pour

$$\theta_H^{(q+1)} := (\alpha_h^{k(q+1)}, \pi_h^{(q+1)})_{\substack{h=1,\ldots,K \\ k=1,\ldots,K-1 \\ j=1,\ldots,H-1}} \qquad (7.16)$$

où $\alpha_h^{k(q+1)} = \frac{\sum_i \beta_{ih}^{k(q)}}{\sum_i t_{ih}^{(q)}}$ and $\pi_h^{(q+1)} = \frac{1}{N} \sum_{i=1}^{N} t_{ih}^{(q)}$

$$\text{avec } t_{ih}^{(q)} = E[z_{ih} \mid m^{x,y}; \theta_H^{(q)}] = \frac{\pi_h^{(q)} pl_{ih} \sum_{k=1}^{K} \alpha_h^{k(q)} pl_i^k}{\sum_{h=1}^{H} pl_{ih} \pi_h^{(q)} \sum_{k=1}^{K} pl_i^k \alpha_h^{k(q)}}$$

$$\text{et } \beta_{ih}^{k(q)} = E[z_{ih} y_i^k \mid m^{x,y}; \theta_H^{(q)}] = \frac{\alpha_h^{k(q)} \pi_h^{(q)} . pl_{ih} pl_i^k}{\sum_{h=1}^{H} pl_{ih} \pi_h^{(q)} \sum_{k=1}^{K} pl_i^k \alpha_h^{k(q)}}$$

7.2.3. Algorithme d'estimation du nouveau paramètre d'un arbre obtenu lors d'une coupure

Nous présentons ici l'algorithme permettant d'estimer le paramètre multidimensionnel θ_H d'un arbre quelconque $\mathscr{P}_H$ à partir de données incertaines m^{XY}, préalablement traduites en pl_{ih} et pl_i^k, et avec une initialisation $\theta_H^{(0)}$.

Algorithm 1: Algorithme de l'estimation du paramètre θ_H d'un arbre $\mathscr{P}_H$

Input: $\theta_H^{(0)}, \epsilon$
Output: paramètre final θ_H

$q = 1$;
repeat
  calcul de $\theta^{(q+1)} = (\alpha_h^{k(q+1)}, \pi_h^{(q+1)}) = \underset{\theta_H \in \Theta_H}{\arg\max} Q(\theta_H, \theta_H^{(q)})$;
  Estimation $L(\theta^{(q+1)}; m^{x,y})$;
  $q = q + 1$;
until $\frac{L(\theta^{(q)}; m^{x,y}) - L(\theta^{(q-1)}; m^{x,y})}{L(\theta^{(q-1)}; m^{x,y})} < \epsilon$;
$\theta_H = \theta^{(q)}$;

Pour une coupure d'une feuille t_h d'un arbre $\mathscr{P}_H$ nous proposons d'initialiser le nouveau paramètre θ_{H+1} à partir du paramètre θ_H (i.e. du paramètre de l'arbre avant coupure) en recopiant tous les paramètres des feuilles de $\mathscr{P}_H$ sauf ceux des nouvelles feuilles t_{h_L} et t_{H_R} qui seront eux issus de celui de la feuille t_h. On aura donc

- $\forall h' \neq h, \forall k \in \{1, \ldots, K\}, \quad (\pi_{h'}^{(0)}, \alpha_{h'}^{k(0)})_{H+1} = (\pi_{h'}, \alpha_{h'}^k)_H$
- $\pi_{h_L}^{(0)} = \pi_{h_R}^{(0)} = \frac{\pi_h}{2}$
- $\forall k \in \{1, \ldots, K\}, \quad \alpha_{h_L}^{k(0)} = \alpha_{h_R}^{k(0)} = \alpha_h^k$

7.2.4. Algorithme général de construction d'un arbre E^2M

La construction d'un arbre E^2M, tout comme un arbre de décision classique, est descendante, i.e. part du nœud initial et construit itérativement des feuilles par coupures des nœuds ou feuilles précédents (on représente généralement les nœuds enfants en dessous de leurs parents).

De même que pour $CART$, toutes les coupures possibles $s \in \mathscr{S}_{\mathscr{P}_H}$ sont considérées, et pour chacune d'entre elles, un gain de pureté Δi classique est calculé à l'aide d'une mesure d'impureté classique (ici l'entropie de Shannon) à partir du paramètre global θ_H de l'arbre initial $\mathscr{P}_H$ et de celui de l'arbre obtenu à partir de s qu'on note $\mathscr{P}_{H+1} = \mathscr{P}_H(s)$: pour une coupure s séparant une feuille t_h en ses enfants t_L et t_R on aura

$$\Delta i(\theta_H(s), \theta_H) = i(t_h) - \frac{\pi_L}{\pi_L + \pi_R} i(t_L) - \frac{\pi_R}{\pi_L + \pi_R} i(t_R)$$

Alors que $i(t_h)$ s'obtient à partir de θ_H, les termes $\pi_L, \pi_R, i(t_L)$ et $i(t_R)$ sont calculés à partir de θ_{H+1}. Le paramètre θ_{H+1} est estimé par maximisation de la vraisemblance crédibiliste $L(.; m^{z,y})$ via l'algorithme E^2M, puis la coupure donnant le meilleur gain de pureté est finalement choisie.

Le déroulement de l'arbre continue jusqu'à ce qu'un critère d'arrêt soit vérifié.

Voici l'algorithme général résumant la construction d'un arbre E^2M :

Cet algorithme ne comprend pas de phase d'élagage, cependant cette dernière peut très bien être ajoutée en fin d'algorithme.

7.2.5. Arbres E^2M approximés

Une différence importante en pratique entre la mise en œuvre de l'algorithme ici proposé et celle des arbres de décision classiques construits à partir de données précises est que la nature précise des données permet de rechercher séparément les meilleures coupures de chaque feuille d'un arbre alors qu'avec des données incertaines, une coupure modifie tous les paramètres, même ceux des autre feuilles (non-coupées), ce qui rend obligatoire la ré-éstimation de tous les paramètres de l'arbre à chaque coupure.

Cet inconvénient rend les temps de calcul beaucoup plus longs en pratique. Nous proposons donc de figer les paramètres des autres feuilles, quand ces temps de

Algorithm 2: Algorithme général d'apprentissage d'un arbre E^2M

Input: $\mathscr{P}_1 = \{t_1\} = \{\Omega_X\}$, données crédibilistes $m^{x,y}$
Output: arbre final $\mathscr{P}_H$

$H = 1$;
$\mathscr{P}_1 = \{t_1\} = \{\Omega_X\}$;
$m^x \rightarrow Pl^z$;
Mise en œvre de l'algorithme 1 appliqué à $\theta_1^{(0)}$;
while *CRITERES d'ARRÊT pas vérifiés* **do**
    foreach *Coupure possible s* **do**
        Estimation de $\theta_H(s)$;
        Calcul de $\Delta i(\theta_H(s), \theta_H)$;
    $s_{optimal} = \underset{s \in \mathscr{S}_{\mathscr{P}_H}}{\arg\max} \Delta i(\theta_H(s), \theta_H)$;
    $\mathscr{P}_{H+1} = \mathscr{P}_H(s_{optimal})$;
    $\theta_{H+1} = \theta_H(s_{optimal})$;
    $H = H + 1$;

calculs deviennent problématiques, nous parlerons alors d'*arbres E^2M approximés*.

Malgré tout, les propriétés de la vraisemblance crédibiliste n'étant pas encore bien établie, même ces arbres E^2M approximés seront appris globalement, i.e. à chaque étape de la construction d'un arbre toutes les coupures potentielles pour toutes les feuilles de l'arbre seront envisagées (tout en figeant les paramètres des feuilles non-coupées).

7.3. Prédiction à l'aide d'un arbre de décision E^2M

Une fois qu'un arbre E^2M $\mathscr{P}_H$ est construit, l'utiliser pour prédire la classe d'un nouvel exemple i^* dont les valeurs de ses attributs x_{i^*} sont précises revient exactement à utiliser un arbre de décision classique. A partir de x_{i^*}, on sait avec certitude la feuille t_{h^*} où i^* aboutit. La classe prédite $y_{i^*}^{pred}$ est alors simplement la classe la plus probable dans t_{h^*} :

$$y_{i^*}^{pred} = \underset{\omega_k \in \Omega_Y}{\arg\max} P_{\theta_H}(\omega_k \mid Z = h) = \underset{\omega_k \in \Omega_Y}{\arg\max} \alpha_{h^*}^k$$

Dans le cas où le nouvel exemple à classifier est incertain, ses attributs étant connus de façon imparfaites, toutes les feuilles sont potentiellement concernées.

Faire une prédiction quant à sa classe ne peut plus se faire directement à partir des prédictions des feuilles, ces prédictions étant conflictuelles. Pour un arbre de décision classique, la sortie est constituée d'une distribution de probabilité prédictive $P_{Y_{i^*}^{pred}}$ dont l'$argmax$ est la prédiction.

Dans le cas d'un arbre E^2M, cette probabilité $P_{Y_{i^*}^{pred}}$ doit être induite à partir de la structure de l'arbre en question, de $m_{i^*}^x$ et de θ_H. Comme expliqué dans 7.2.2, il est facile de transformer $m_{i^*}^x$ en $m_{i^*}^z$, de plus comme on ne sait rien a priori quant à la classe y_{i^*}, on a $m_{i^*}^Y(\Omega_Y) = 1$ (et donc $\forall k \in \{1,\dots,K\}$, $pl_{i^*}^k = 1$), on aura alors $m_{i^*}^{ZY} = m_{i^*}^{Z\uparrow ZY}$ selon la définition 10 (et donc $\forall A, B \subseteq \Omega_Z \times \Omega_Y$, $m_{i^*}^{ZY}(A,B) = m_{i^*}^Z(A)$ si $B = \Omega_Z$, 0 sinon).

La probabilité recherchée $P_{Y_{i^*}^{pred}}$ doit à la fois tenir compte de m_*^z et de la probabilité $P_{\theta_H}^{Z,Y}$ représentant le modèle (i.e. l'arbre), elle correspondra donc à la projection sur Ω_Y de la conjonction (par Dempster) de $m_{i^*}^{ZY}$ et de $P_{\theta_H}^{Z,Y}$, on pourra donc écrire $P_{Y_i^{pred}} = P[.\,|\,\theta_H, m_{i^*}^{ZY}]^{ZY\downarrow Y}$.

On obtient alors $\forall k \in \{1,\dots,K\}$, :

$$
\begin{aligned}
P_{Y_i^{pred}}(\{Y=\omega_k\}) &= P(\{Y=\omega_k\}\,|\,\theta_H, m_{i^*}^{ZY})^{ZY\downarrow Y} \\
&= \sum_{h=1}^{H} P(\{Y=\omega_k, Z=h\}\,|\,\theta_H, m_{i^*}^{ZY}) \\
&= \sum_{h=1}^{H} \beta_{i^*h}^k \\
&= \sum_{h=1}^{H} \frac{\alpha_h^k \pi_h . pl_{i^*h} pl_{i^*}^k}{\sum_{h=1}^{H} pl_{i^*h}\pi_h \sum_{k=1}^{K} pl_{i^*}^k \alpha_h^k} \\
&= \frac{\sum_{h=1}^{H} \alpha_h^k \pi_h . pl_{i^*h}}{\sum_{h=1}^{H} \pi_h pl_{i^*h}}
\end{aligned}
$$

car tous les $pl_{i^*}^k$ sont égaux à 1 et que $\forall h \in \{1,\dots,H\}$, $\sum_{k=1}^{K} \alpha_h^k = 1$

7.4. Elagage : évaluation en classification incertaine

Nous proposons ici une technique d'élagage adaptée aux arbres E^2M. Le principe d'élagage pose le problème plus général d'évaluation de classifieurs incertains. Après avoir explicité ce problème, nous présentons la technique proposée qui est encore une fois basée sur l'algorithme E^2M.

7.4.1. Problématique générale de la classification incertaine

Toute technique d'élagage nécessite l'explicitation d'un critère mettant en œuvre une certaine forme d'évaluation de l'arbre. Dans le cadre de données incertaines, une telle évaluation n'est pas simple. En effet, dans le cas de classes incertaines par exemple, il est difficile de savoir si un exemple est bien classifié ou pas, sa vraie classe n'étant connue que de façon incertaine.

7.4.2. Résolution par l'algorithme E^2M

La vraie classe y_i d'un exemple incertain i n'étant pas *précisément* connue, le taux d'erreur ϵ relatif à un échantillon de test crédibiliste $\hat{m}^{X,Y}$ ne pourra plus être directement calculé, il sera donc estimé. Pour cela, nous considérerons, pour chaque exemple i de cet échantilon de test, une variable aléatoire E_i suivant une loi de Bernouilli de paramètre ϵ valant 1 en cas d'erreur de classification, 0, sinon. On a donc $\forall i \in \{1,\dots,n\}, \quad E_i \sim Ber(\epsilon)$ et $P(E_i = 1) = P(y_i^{pred} \neq y_i) = \epsilon$ et $P(E_i = 0) = P(y_i^{pred} = y_i) = 1-\epsilon$. Le taux d'erreur $R(\mathscr{P}_H) = \epsilon$ d'un arbre $\mathscr{P}_H$ sera donc estimé sur un échantillon de test crédibiliste $\hat{m}^{x,y}$ par ϵ. Comme les réalisations $e = (e_1,...,e_N)$ de $E = (E_1,\dots,E_n)$ (supposés iid) ne sont pas directement observées, on peut estimer, toujours avec l'algorithme E^2M, la valeur de ϵ. Pour se faire, nous maximisons la vraisemblance crédibiliste $L(\epsilon; m^e)$ de m^e relatitivement au paramètre ϵ. On commence pour cela par expliciter cette vraisemblance complétée (précise) $L(\epsilon; e)$:

$$\begin{aligned} L(\epsilon; e) &= P_{E;\epsilon}(e) \\ &= \prod_{i=1}^{N} \epsilon^{e_i}(1-\epsilon)^{1-e_i} \end{aligned}$$

$$\text{On a donc} \quad logL(\epsilon;e) = \sum_{i=1}^{N}[e_i log(\epsilon) + (1-e_i) log(1-\epsilon)]$$
$$= Nlog(1-\epsilon) + log(\frac{\epsilon}{1-\epsilon}) \sum_{i=1}^{N} e_i$$

La vraisemblance crédibiliste (observée) de ϵ relativement à m^e est donnée par :

$$L(\epsilon; m^e) = \prod_{i=1}^{N}[(1-\epsilon) pl_i(0) + \epsilon pl_i(1)]$$

$$\text{où} \quad pl_i(0) = Pl_i^e(\{0\})$$
$$= Pl_i^y(\{y_i^{pred}\}) \quad (y_i^{pred} \text{ est connu})$$

$$\text{et où} \quad pl_i(1) = Pl_i^e(\{1\})$$
$$= Pl_i^y(\{\Omega_Y \setminus y_i^{pred}\})$$
$$= 1 - m_i^y(y_i^{pred})$$

$$\text{On obtient alors} \quad Q(\epsilon;\epsilon^{(q)}) = Nlog(1-\epsilon) + log(\frac{\epsilon}{1-\epsilon}) \sum_{i=1}^{N} \xi_i^{(q)}$$

$$\text{où} \quad \xi_i^{(q)} = E[E_i \mid m_i^e; \epsilon^{(q)}]$$
$$= \frac{\epsilon^{(q)} pl_i(1)}{(1-\epsilon^{(q)}) pl_i(0) + \epsilon^{(q)} pl_i(1)}$$

$$\text{Ce qui donne} \quad \epsilon^{(r+1)} = \underset{\epsilon \in [0,1]}{\text{argmax}}\, Q(\epsilon;\epsilon^{(q)})$$
$$= \frac{1}{N} \sum_{i=1}^{N} \xi_i^{(q)}$$

Remarque :

Une autre solution aurait pu être d'évaluer notre classifieur sur un échantillon incertain $\hat{m}^{X,Y}$ de manière imprécise à l'aide d'un intervalle $[R^-;R^+]$ où

$$R^- = \frac{1}{N}\sum_{i=1}^{N} Bel_i^y(\Omega_Y \backslash \{Y^{pred}\}) \text{ et } R^+ = \frac{1}{N}\sum_{i=1}^{N} Pl_i^y(\Omega_Y \backslash \{Y^{pred}\})$$

Il pourrait être par la suite intéressant de voir si l'estimation obtenue par la première méthode se situe forcément dans cet intervalle ou pas.

7.4.3. Algorithme d'élagage proposé

Par souci de cohérence, nous proposons de choisir une méthode d'élagage dérivée de celle proposée dans $CART$ en utilisant l'évaluation de classifieurs sur données incertaines proposée précédemment dès qu'une évaluation est requise.

Une fois un arbre de décision $\mathscr{P}_H$ construit, la séquence d'arbres emboîtés $\{\mathscr{P}_{H_1},\ldots,\mathscr{P}_{H_p}\}$ est naturellement obtenue (on a $H = H_1 > \cdots > H_p = 1$) directement en considérant (et en gardant en mémoire) les arbres à chaque coupure pendant l'apprentissage de $\mathscr{P}_H$. Ensuite, de même que dans $CART$, le sous-arbre minimisant le taux d'erreur (crédibiliste) sur l'"échantillon test (ou d'élagage) crédibiliste $\hat{m}^{X,Y}$ est choisi. C'est pour calculer les taux d'erreur crédibilistes des sous-arbres qu'on applique notre solution crédibiliste pour la classification incertaine (voir Section 7.4.2).

Algorithm 3: Algorithme d'élagage d'un arbre E^2M

Input: $\mathscr{P}_H$, échantillon d'élagage ou d'apprentissage
Output: arbre final $\mathscr{P}_a$

1 détermination de la séquance d'arbre emboités $\{\mathscr{P}_{H_1},\ldots,\mathscr{P}_{H_p}\}$;
2 arbre final = $\underset{H\in\{H_1,\ldots,H_p\}}{\arg\max}\ R(\mathscr{P}_H)$;

7.5. Conclusion

La méthodologie des arbres de décision E^2M permet d'apprendre des arbres de décision à partir de données incertaines représentées par des fonctions de croyance. Inspirés du cadre de $CART$, ces arbres ont comme critère de coupure la maximisation d'un gain de pureté en terme de classe dans les feuilles réalisé en minimisant l'entropie des classes dans les feuilles. Ces gains de pureté nécessitent l'estimation des probabilités des feuilles et des classes dans les feuilles (π_h et α_h^k). Contrairement au cas de données d'apprentissage précises, l'incertitude des données rend obsolète l'utilisation de fréquences (basées donc sur des effectifs) pour l'estimation de ces probabilités. Nous proposons de les estimer en maximisant la vraisemblance crédibiliste des données (du couple *feuille, classe*). Cette vraisemblance n'étant pas directement maximisable, nous utilisons pour cela l'algorithme E^2M.

La nature de la vraisemblance crédibiliste fait que tous les paramètres d'un arbre doivent être ré-estimés à chaque nouvelle coupure. De manière à réduire les temps de calcul qui en découlent, nous proposons un compromis (temps de calcul / efficacité prédictive) en figeant, pour chaque nouvelle coupure ,les paramètres des feuilles non-coupées.

Un algorithme d'élagage est aussi proposé en tirant parti du coté *global* de la construction des arbres E^2M.

Des expériences sont présentées en Section 8.2 qui montrent l'intérêt de prendre en compte l'incertitude des données d'apprentissage dans un environnement bruité.

Troisième partie .

Expériences et application au caoutchouc

Cette partie est consacrée aux différentes expériences et applications réalisées pendant la thèse. Le premier chapitre présente des expériences réalisées sur des jeux de données benchmark de manière à appliquer les deux méthodologies présentées en Partie II : l'extension de la méthodologie Skarstein-Bjanger et Denoeux au cas multi-classes et enfin les arbres E^2M. Le deuxième chapitre présente l'application principale de ce manuscrit : la prédiction de la qualité du caoutchouc. C'est donc à partir de données réelles, cette fois, que les arbres E^2M seront appliqués.

CHAPITRE 8

Expériences

Nous présentons ici quelques expériences réalisées pour évaluer les deux méthodologies proposées dans ce manuscrit : l'extension de la méthodologie Skarstein-Bjanger et Denoeux au cas multi-classes et la méthodologie principale de ce manuscrit : les arbres de décision E^2M.

Sommaire

8.1. Expériences mettant en oeuvre l'extension multi-classes de la méthodologie Skarstein-Bjanger et Denoeux

Nous présentons tout d'abord ici quelques expériences sur des jeux de données benchmark (UCI) à deux classes. Des expériences sont ensuite réalisées toujours sur des jeux de données benchmark mais à plus de deux classes. Dans les deux cas nous nous intéressons surtout à la différence de performance prédictive entre les différentes méthodes, aucun arbre n'est donc ici élagué.

TABLE 8.1.: Caractéristiques des jeux de données à deux classes UCI utilisés

Jeu de données	# attributs	# exemples	type des attributs
Blood Transfusion (Service Center)	5	748	numérique
Statlog (Heart)	13	270	nominal et numérique
Tic-Tac-Toe Endgame	9	958	nominal
Breast Cancer	9	286	nominal
Pima Indians Diabetes	8	768	numérique
Haberman's Survival	3	306	numérique

TABLE 8.2.: Comparaison des taux d'erreur moyens entre les arbres de Skarstein-Bjanger et Denoeux et $CART$ pour le cas à deux classes

Jeu de données	tx er. moyen $CART$	tx er. moyen SBD
Blood Transfusion (Service Center)	23.5%	24.2%
Statlog (Heart)	28%	25.7%
Tic-Tac-Toe Endgame	21.5%	11.5%
Breast Cancer	5.9%	4.7%
Pima Indians Diabetes	27.3%	25.1%
Haberman's Survival	26.6%	26%

8.1.1. Expériences initiales pour le cas à deux classes

Nous comparons ici l'efficacité prédicitive des arbres de décision obtenus par la méthodologie Skarstein-Bjanger et Denoeux avec celle des arbres de décision suivant le modèle $CART$ sur des jeux de données benchmark à deux classes. Les jeux de données utilisés sont issus du site UCI (http ://archive.ics.uci.edu/ml/). Leurs principales caractéristiques sont résumées dans le Tableau 8.1.

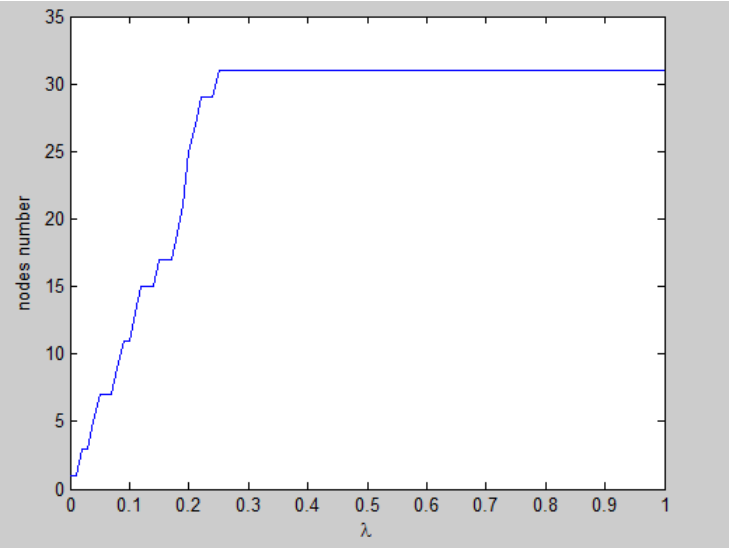
Les critères d'arrêts sont les suivants :

– feuilles de taille inférieure à dix exemples
– gain de pureté minimal : continuer de couper tant que $\Delta i > 0$ pour les arbres de Skarstein-Bjanger et Denoeux (SBD), $\Delta i > \beta$ pour les arbres $CART$

Les paramètres λ de U_λ et β (pour le gain de pureté minimum des arbres $CART$) sont obtenus par validation croisée à dix couches.

Le Tableau 8.2 présente les taux d'erreur moyens obtenus par les deux méthodologies comparées en ré-échantillonnant aléatoirement dix fois de manière à obtenir $2/3$ des exemples dans l'échantillon d'apprentissage et $1/3$ dans l'échantillon test.

FIGURE *8.1.*: Nombre de noeuds en fonction de λ pour le jeu de données $Pima$



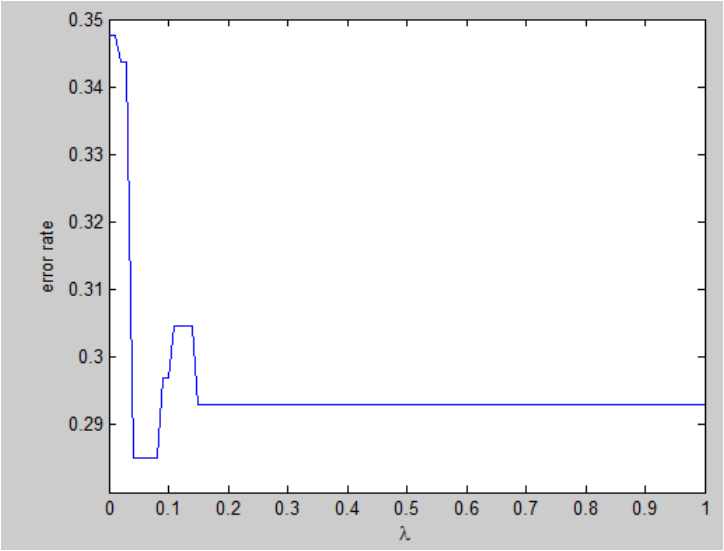
Discussion sur l'hyper-paramètre λ

Les figures 8.1 et 8.2 montrent l'impact de λ sur la complexité de l'arbre obtenu et sur son efficacité sur le jeu de données "Pima". Nous constatons que cette complexité augmente avec λ, ce qui confirme l'intuition que λ peut être interprété comme un indicateur de l'importance donnée au manque d'exemples dans un noeud (i.e. la non-spécificité $N(m)$) et à la tendance de ΔU_λ à être négative. Ceci suggère que l'optimisation (ici une validation croisée à 10 couches) devrait intégrer la complexité des arbres comme critère. Le paramètre λ semble n'avoir que peu d'influence sur l'efficacité des arbres (Klir propose de le fixer à 0.5).

Conclusion : les taux d'erreur obtenus pour les deux méthodologies sont proches. Ceux obtenus par la méthodologie SBD sont inférieurs à ceux obtenus par $CART$. Ceci provient vraisemblablement du fait que l'utilisation de l'approche de Dempster pour l'expérience de Bernoulli permet une prise en compte de la taille des feuilles (i.e. du nombre d'exemples qu'elles contiennent) dans les calculs de gain de pureté.

Remarque : Il est ici nécessaire de citer un auteur ayant récemment proposé une mesure d'impureté prenant en compte la taille des feuilles dans le cadre probabiliste classique, il s'agit de Zighed [99].

FIGURE 8.2.: taux d'erreur en fonction de λ pour le jeu de données $Pima$



Jeux de données	# d'attributs	# classes	# exemples
Iris	4	3	150
Balance Scale	4	3	625
Wine	13	3	178
Car Evaluation	6	4	1728
Page Blocks Classification	10	5	5473
Forest-fires	12	6	517

TABLE 8.3.: Jeux de donnée multi-classes utilisés lors des expériences

8.1.2. Expériences avec l'extension multi-classes

Nous comparons ici les efficacités des arbres de décision obtenus en mettant en oeuvre les trois extensions possibles (voir Section 6) de la méthodologie SBD (voir Section 5.5) entre elles et avec des arbres de décision classiques construit selon l'algorithme $CART$. A titre de rappel, ces trois extensions sont les combinaison de classifieurs binaires selon la mé- thodologie de Quost et al. ("Combinaisons"), le modèle de Dirichlet imprécis (MDI) et le modèle multinomial de Denoeux ("Modèle multinomial").

Les jeux de données utilisés sont toujours issus du site UCI (http ://archive.ics.uci.edu/ml/). Leurs principales caractéristiques sont résumées dans le Tableau 8.3.

Les critères d'arrêts sont les suivants :

– feuilles de taille inférieure à dix exemples

Jeux de données	$CART$	SBD		
		Combinaisons	MDI	Modèle multinomial
Iris	2.0%	2.0%	2.0%	2.0%
Balance Scale	20.2%	25.0%	17.8%	15.9%
Wine	11.9%	8.5%	13.6%	13.6%
Car Evaluation	17.7%	17.7%	15.6%	32.3%
Page Blocks Classification	4.8%	4.7%	5.0%	5.2%
Forest Fires	43.6%	43.0%	43.0%	43.0%

TABLE *8.4.:* Efficacité (taux d'erreur moyens) des arbres en fonction du modèle de construction des masses de croyance

– gain de pureté minimal : continuer de couper tant que $\Delta i > 0$ pour tous les arbres
– profondeur maximale de cinq noeuds (i.e. un chemin menant à une feuille comprendra donc au maximum cinq noeuds *intermédiaires*)

Pour les arbres SBD, l'hyper-paramètre λ est fixé à 0.5. Etant donné que nous privilégions ici l'efficacité à la simplicité des arbres, aucun arbre n'est ici élagué.

Pour les jeux de données à trois classes, le modèle multinomial utilise simplement la fonction de croyance induite par $\underline{P}$ alors que pour *Page blocks* (5 classes), les fonctions de croyance sont approximées par programmation linéaire et l'algorithme des classes ordinales est utilisé pour *Forest fires* (6 classes).

Les Tableaux 8.4, 8.5 et 8.6 présentent les taux d'erreur moyens, le nombre moyen de noeuds par arbre et les temps de calcul [1] (en secondes) obtenus par $CART$ et par les trois extensions de la méthodologie SBD comparées en ré-échantillonant aléatoirement dix fois de manière à obtenir $2/3$ des exemples dans l'échantillon d'apprentissage et $1/3$ dans l'échantillon test.

Conclusions : La méthodologie SBD étendue au cas multi-classes semble avoir une efficacité prédictive comparable à celle de $CART$. Pour certains jeux de données, certaines de ces extensions paraissent clairement plus efficaces que les arbres classiques $CART$, ce n'est cependant pas toujours le cas, et l'extension la plus efficace n'est de plus pas toujours la même. On remarque aussi que les temps de calcul sont allongés pour le modèle multinomial, du fait de l'utilisation de nombreux éléments focaux et de la programmation linéaire.

1. le processeur utilisé pour ces expériences est Intel(R) Core(TM) i7 CPU M 640 @ 2.80GHz 2.80GHz.

Jeux de données	$CART$	SBD	
		MDI	Modèle multinomial
Iris	5	5	5
Balance Scale	27	27	27
Wine	7	9	22
Car Evaluation	17	17	1
Page Blocks Classification	23	27	25
Forest Fires	13	21	1

TABLE *8.5.:* Nombre de noeuds moyen des arbres en fonction du modèle de construction des masses de croyance

Jeux de données	$CART$	SBD		
		Combinaisons	MDI	Modèle multinomial
Iris	0	0	1	6
Balance Scale	0	0	2	29
Wine	0	0	1	19
Car Evaluation	1	1	9	8
Page Blocks Classification	53	38	140	1801
Forest Fires	1	1	15	81

TABLE *8.6.:* Temps de construction (en secondes) des arbres en fonction du modèle de construction des masses de croyance

8.2. Expériences mettant en oeuvre les arbres de décision E^2M

Du fait du manque de benchmark pour les jeux de données incertaines, de manière à évaluer la méthodologie des arbres de décision $E2M$ (voir Section 7), nous proposons ici d'utiliser des jeux de données benchmark classiques et d'injecter artificiellement de l'incertitude dans les données en nous inspirant du modèle génératif (2.15) utilisé en Section 2.4.1 mais sans spécification de la loi utilisée pour simuler la variable V qui correspondra à une donnée *initiale* (d'un de ces jeux de données benchmark).

Ce modèle est le suivant :

$$W = (1-B)V + BU \tag{8.1}$$

où

$$\begin{cases} V \text{ donnée initiale} \\ B \sim Ber(\epsilon) \\ U \sim \mathcal{U}_{\Omega_V} \\ V,B \text{ et } U \text{ indépendantes} \end{cases}$$

La variable V représentera donc une donnée initiale du jeu de données benchmark considéré, et W sera alors la donnée finale (après remplacement éventuel). L'idée sous-jacente est de considérer que pour certaines données, on peut connaître le niveau de fiabilité (en procédant à des tests de validation pour les capteurs par exemple) et en tirer parti. On modélisera l'incertitude des données par des fonctions de croyance de la même manière qu'en Section 2.4.1 (comme m^V). Pour toute donnée w, on posera donc $\begin{cases} m^V(\{w\}) = 1-\epsilon \\ m^V(\Omega_V) = \epsilon \end{cases}$. La différence majeure avec l'étude menée en Section 2.4.1 est qu'ici, le niveau d'incertitude des données ϵ est variable, il sera uniformément simulé sur [0,1].

Le Tableau 8.7 présente les différents jeux de données benchmark utilisés (il proviennent tous du site http ://archive.ics.uci.edu/ml/). Les Tableaux 8.8, 8.9 et 8.10 présentent les taux d'erreur moyens obtenus sur cinq validations croisées à 10 couches, d'une part par des arbres $CART$ et d'autre part par des arbres E^2M. Ils correspondent respectivement aux cas d'attributs incertains, de classe incertaines et de manière générale de données incertaines (attributs et classe confondus). Les taux d'erreur *naïfs* sont ceux obtenus par prédiction systématique de la classe majoritaire. Les critères d'arrêt utilisés sont les mêmes pour les deux méthodologies, ils sont constitués par :

- un maximum de 5 feuilles,
- un gain minimum de pureté relatif de 5%

. Notons que les arbres $CART$ ne prennent donc pas en compte l'incertitude des données lors de leur apprentissage.

jeu de données	# attribut	# classes	# exemples
Iris	4	3	150
Balance scale	4	3	625
Wine	13	3	178
Glass	9	7	214
E.Coli	8	7	336

TABLE *8.7.:* Caractéristiques des jeux de données

Un test de Wilcoxon d'un niveau de 95% est réalisé de manière à conclure quant

à la significativité des différences de taux d'erreur obtenus entre les arbres $CART$ et les arbres E^2M. Lorsque ces tests confirment la significativité statistique des différences de taux d'erreur, les meilleurs taux d'erreur sont mis en gras dans les tables.

algorithm	$naif$	$CART$	E^2M
iris	0.67	0.22	**0.14**
balance	0.54	0.54	**0.37**
wine	0.60	0.28	**0.19**
glass	0.65	**0.54**	0.58
E.Coli	0.57	0.34	**0.28**

TABLE *8.8.:* Attributs incertains

algorithm	$naif$	$CART$	E^2M
iris	0.67	**0.15**	0.16
balance	0.54	0.54	**0.39**
wine	0.63	0.23	0.20
glass	0.66	0.57	0.58
E.Coli	0.57	0.27	0.29

TABLE *8.9.:* Classes incertaines

algorithm	$naif$	$CART$	E^2M
iris	0.67	0.44	**0.17**
balance	0.54	0.54	**0.39**
wine	0.63	0.46	**0.22**
glass	0.65	0.65	**0.58**
E.Coli	0.39	0.38	**0.22**

TABLE *8.10.:* Attributs et classes incertains

Conclusions : Même si le nombre de jeux données utilisés n'est pas très grand, ces résultats sont assez satisfaisants car les arbres E^2M apparaissent presque toujours plus efficaces (en terme de prédiction) que les arbres $CART$. Les arbres E^2M présentent donc l'avantage de tenir compte de l'incertitude des données (en accordant le plus d'importance aux données les plus fiables). Cette différence semble surtout marquée dans le cas où toutes les données sont incertaines (attributs et classe). Ceci n'est pas surprenant si on considère le nombre de variables incertaines. En effet, quand les données sont incertaines, de par leur construction, les arbres E^2M semblent plus efficaces que les arbres $CART$, on peut donc s'attendre à ce que cet avantage croisse avec le niveau d'incertitude des données.

Remarque : Il aurait aussi été intéressant de faire les mêmes expériences mais en construisant différemment les fonctions de croyance d'apprentissage. En effet, on aurait pu modéliser l'incertitude des données à l'aide de probabilités (comme pour $\tilde{m}^v$ dans la Section 2.4.1). On aurait probablement obtenu des arbres encore plus efficaces, car l'incertitude des données aurait alors été modélisée à l'aide de leur modèle génératif. Il faut cependant noter que, dans les applications un tel modèle génératif est rarement disponible.

8.2.1. Illustration de la mise en oeuvre des arbres E^2M approximés

De manière à avoir une idée des avantages (en terme de temps de calcul) et des inconvénients (en terme d'efficacité prédictive) de l'utilisation d'arbres E^2M approximés plutôt que de vrais arbres E^2M, des expériences sont ici présentées. Le modèle génératif de l'incertitude des données, tout comme les critères d'arrêt utilisés sont exactement les mêmes que précédemment. La table 8.11 présente donc une comparaison entre les taux d'erreur moyens (et leur intervalles de confiance) obtenus par prédiction naïve, par des arbres $CART$ (qui ne tiendront donc pas compte de l'incertitude des données), par les arbres E^2M et par les arbres E^2M approximés (tels que présentés en Section 7.2.5) dans le cas de données (attributs et classe) incertaines. Ils sont obtenus pour quatre validations croisées à 10 couches. La table 8.12 présente les temps de calcul moyens correspondants (avec leur écarts-type).

	$naif$	$CART$		E^2M		E^2M approximés	
données	tx er.	tx er.	IC(95%)	tx er.	IC(95%)	tx er.	IC(95%)
iris	0.67	0.45	[0.35 ; 0.54]	0.18	[0.10 ; 0.25]	0.31	[0.22 ; 0.41]
balance	0.54	0.54	[0.44 ; 0.64]	0.37	[0.27 ; 0.46]	0.37	[0.27 ; 0.46]
wine	0.64	0.48	[0.38 ; 0.58]	0.24	[0.15 ; 0.32]	0.28	[0.19 ; 0.36]
glass	0.65	0.65	[0.56 ; 0.75]	0.54	[0.45 ; 0.64]	0.54	[0.44 ; 0.64]

TABLE *8.11.:* Taux d'erreur moyens obtenus par les arbres E^2M approximés avec attributs et classe incertains

Conclusions : On observe clairement que les arbres E^2M approximés représentent

	$CART$		E^2M		E^2M approximés	
données	temps moyen	écart-type	temps moyen	écart-type	temps moyen	écart-type
iris	72	42	2663	903	403	36
balance	6	1	4548	935	196	47
wine	841	547	13564	4978	2454	200
glass	47	1	23161	7562	1273	736

TABLE *8.12.:* Temps de calcul moyens obtenus par les arbres E^2M approximés avec attributs et classe incertains

un vrai compromis entre efficacité prédictive et temps de calcul. En effet, ils présentent des taux d'erreur généralement supérieurs à ceux des arbres E^2M mais tout de même inférieurs à ceux obtenus pour les arbres $CART$. De plus, même si ils sont plus longs à construire que les arbres $CART$, ils sont quand même bien plus rapides à apprendre que les arbres E^2M. On pourra donc faire le choix de les utiliser à la place des arbres E^2M lorsque les temps de calcul sont trop grands pour une application donnée ou quand l'efficacité prédictive requise n'est pas trop élevée.

Remarque : Ces temps de calcul sont élevés, et ce même pour les arbres $CART$. Ceci est en partie dû au fait que les programmes utilisés pour $CART$ ou pour les arbres E^2M sont ici les mêmes, seule la modélisation des données diffère (précise pour $CART$, crédibiliste pour les arbres E^2M). Même s'il aurait été possible de construire rapidement les arbres de décision $CART$ (avec le paquet *rpart* du logiciel R par exemple), le fait que pour chaque nouvelle coupure, les arbres E^2M testent toutes les coupures possibles, associé à un nombre maximale de feuille aurait rendu une telle comparaison quelque peu biaisée. En effet, les arbres $CART$ s'inscrivant dans un cadre de données précises, l'apprentissage peut très bien se faire *localement*, ce qui réduit considérablement les temps de calcul. Nous avons donc fait le choix d'un apprentissage global, même pour les arbres $CART$ car le but était ici d'illustrer l'applicabilité des arbres E^2M et de montrer que si les données présentent un certain niveau d'incertitude (0.5 ici en moyenne), il est alors important de tenir compte de cette incertitude lors de l'apprentissage.

CHAPITRE 9

Application : Prédiction de la qualité du caoutchouc naturel

Nous présentons ici l'application principale de ce travail de thèse.

Sommaire

9.1. Introduction

Il existe aujourd'hui deux types de caoutchouc : le caoutchouc naturel, dérivé du *latex* obtenu par saignée du tronc de l'arbre *Hevea brasiliensis*, et le caoutchouc

synthétique d'origine pétro-chimique fabriqué industriellement. Les deux se partagent le marché du caoutchouc de façon relativement équitable (42% de naturel contre 10% de synthétique). Le caoutchouc naturel, longtemps négligé par rapport au synthétique, est redevenu un enjeu capital social, environnemental, économique et politique. Le caoutchouc naturel se distingue du caoutchouc synthétique par sa supériorité en terme d'élasticité et de résistance à la chaleur. Ce caoutchouc naturel présente cependant une grande faiblesse : l'irrégularité de sa qualité, encore mal contrôlée par les experts. Le but de cette application est donc de tenter de mettre en évidence certains facteurs pouvant expliquer ou impacter la qualité du caoutchouc.

Les données sont issues d'une plantation expérimentale MICHELIN au Brésil, la Plantation Edouard Michelin (PEM) sur laquelle le $CIRAD$ est intervenu au début des années 2000 de manière à améliorer la qualité du caoutchouc produit. Ces données regroupent des indices de qualité, des valeurs chronologiques de variables météorologiques, ainsi que différentes données culturales.

La problématique de la qualité du caoutchouc naturel et les données issues de la plantation PEM sont tout d'abord présentées en 9.2. Une première étude statistique prédictive est ensuite réalisée en 9.3 dans un cadre classique de données *précises*. Enfin, une seconde étude statistique mettant en œuvre les *arbres* E^2M vient compléter la première en 9.4, tout en proposant deux modèles de représentation d'incertitudes des données à l'aide de fonctions de croyance.

9.2. Description de la problématique

La culture de l'hévéa est originaire d'Amérique du sud (essentiellement du Brésil). L'établissement en Asie de plantations encadrées par des européens ainsi que la découverte de la vulcanisation (procédé chimique permettant d'augmenter l'élasticité) permirent un développement considérable de l'*hévéa-culture* au milieu du $XX^{ième}$ siècle. L'hévéa est aujourd'hui principalement cultivé dans les pays d'Asie. Ceci s'explique par la disponibilité de grande superficies culturales et aussi par le développement d'une maladie fongique incurable, le $SALB$, qui contamina et qui contamine encore les plantations d'hévéa sud-américaines. Le développement récent de certaines variétés d'hévéa permettent désormais de reconsidérer le retour de l'hévéa-culture en Amérique du Sud, notamment grâce à la découverte de génotypes résistants au $SALB$.

Les plantations d'hévéa asiatiques sont pour la plupart de petites plantations appartenant aux paysans locaux. Qu'elles soient industrielles ou non, toutes les plantations sont confrontées au problème de variabilité de la qualité du caoutchouc obtenu. Pour pallier cela, le latex provenant de différentes bennes, et donc correspondant à différentes qualités est souvent mélangé, ce qui permet de diminuer cette irrégularité qualitative.

Même si les experts de l'hévéa ont une connaissance approfondie du caoutchouc naturel, ils ne contrôlent pour l'instant qu'assez difficilement cette qualité très variable. Ceci peut sans doute s'expliquer par le regain industriel du caoutchouc synthétique pendant une assez longue période. Ce n'est en effet qu'avec l'essor de l'automobile, et de l'aéronautique, que le caoutchouc naturel est redevenu un centre d'intérêt pour les industriels qui y voyaient alors le meilleur moyen d'obtenir des pneus de bonnes qualité, ne risquant pas l'éclatement en cas de chaleur excessive (pendant le freinage notamment et à l'atterrissage des avions).

Il n'est donc pas étonnant qu'une société de l'envergure de MICHELIN s'investisse dans la recherche agronomique sur l'hévéa.

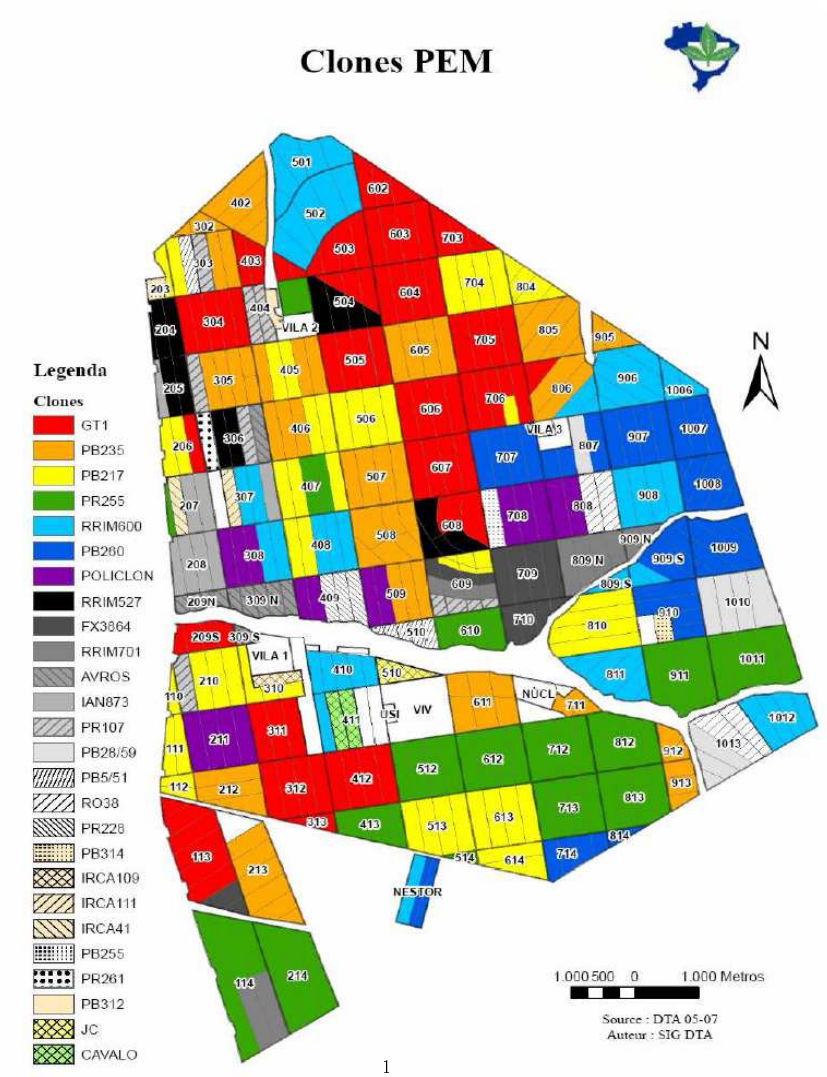
9.2.1. Plantation PEM

La plantation *PEM* se situe dans l'état brésilien du Matto Grosso (latitude 17°C Sud). Le climat y est tropical mais avec des températures basses en hiver. Cette plantation, achetée par Michelin en 1984, recouvre une surface d'environ 9000 hectares. Elle est divisée en parcelles, chacune d'entre elle contenant des hévéas issus d'un même clone (voir figures 9.1 et 9.3).

Lors de la saignée, le matin, le *saigneur* effectue une incision au niveau du tronc des arbres ce qui provoque l'écoulement d'une matière liquide visqueuse blanche : *le latex*. Ce latex est véhiculé dans des vaisseaux laticifères qui entourent les troncs des arbres et qui représente un système d'autodéfense par coagulation lors de l'agression de l'hévéa. Les experts estiment que la plus grosse partie de latex s'écoulant lors d'une saignée est fabriquée approximativement pendant la semaine précédent cette saignée, appelée période de *fabrication du latex*.

FIGURE 9.1.: Photo de la plantation *PEM*

FIGURE 9.2.: Carte du découpage parcellaire de la plantation *PEM* avec indication des clones



Une *tasse* est généralement fixée sur le tronc de manière à recevoir tout l'écoulement de latex. Le saigneur passe ainsi d'arbre en arbre pour la saignée puis repasse quelques heures après ramasser les tasses dont il versera les contenus sous forme de *coagulum* dans une grande benne stationnée à l'intersection de plusieurs parcelles.

Entre temps, en plus de s'être écoulé dans les tasses, le latex aura en partie fermenté avec le développement de bactéries qui joueront un rôle important dans cette période appelée *maturation en tasse*.

Les contenus de plusieurs parcelles sont ainsi mélangés dans une même benne. Ceci induit donc le mélange de toutes les caractéristiques culturales parcellaires. En effet, chaque parcelle ne contient qu'un type clone et les saignées y sont pratiquées suivant des procédés propres à chaque parcelle (*systèmes de saignée, panneau*, etc).

En début de soirée, les bennes remplies de coagulum partent à l'usine où le coagulum sera transformé en caoutchouc par différents procédés qui ne seront pas étudiés ici. Chaque contenu de benne est pesé, analysé, répertorié à l'aide de sa date d'arrivée à l'usine, de son numéro de réception ainsi que de l'ensemble de parcelles correspondant.

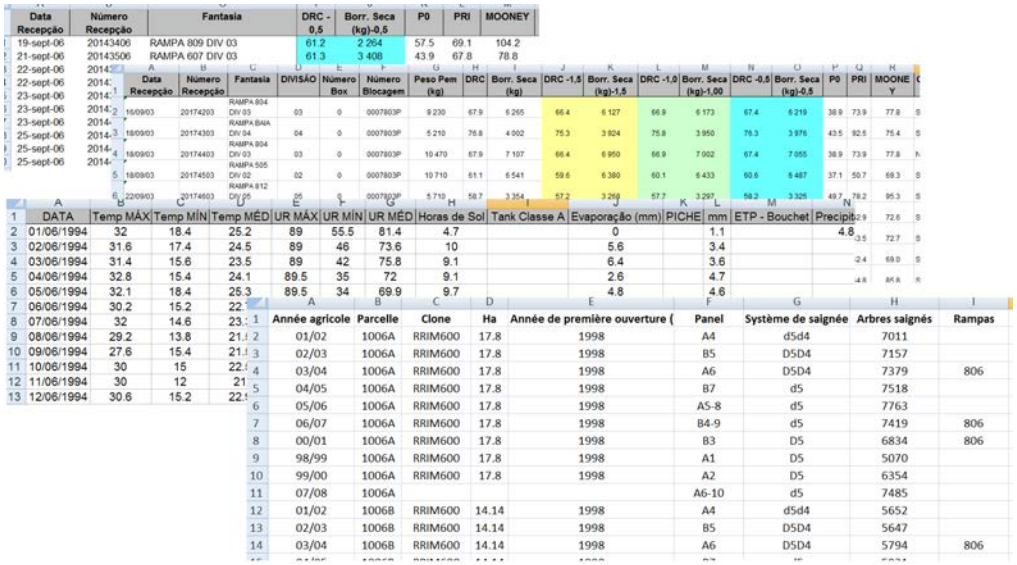
9.2.2. Données

Les *données initiales* fournies par MICHELIN au CIRAD sont constituées de quatre fichiers au format tableur excel, composés de :

- la base de *données parcellaires*
- les deux bases de *données qualité* correspondant à deux périodes d'étude allant respectivement du 16/09/2003 au 18/08/2004 et du 04/09/2006 au 23/07/2007
- la base de *données météorologiques* composées de données climatologiques quotidiennes recouvrant une période allant du 01/06/1994 au 28/02/2009

Une première étape fut d'étudier ces données et de constituer une unique base de données regroupant l'ensemble de ces données. Cette étape fut réalisée dans le cadre d'un stage effectuée par Emilie Doge, une étudiante de l'Institut Universitaire Professionnalisé, département Statistique et Informatique Décisionnelle de l'Université Paul Sabatier à Toulouse.

FIGURE 9.3.: Données initiales



Durant ce stage effectué dans les locaux du laboratoire *MISTEA* à l'*INRA* de Montpellier (campus de Supagro), M^{elle} Doge fut encadrée par Brigitte Charnomordic, Anne Tireau du laboratoire *MISTEA* et moi-même en collaboration avec les experts en caoutchouc Jérôme Sainte-Beuve et Eric Gohet du CIRAD pour réaliser les tâches suivantes :

- étudier dans un premier temps les différentes variables, établir leurs définitions et leurs corrélations
- constituer une base de données unique regroupant les 4 fichiers excel à l'aide des outils MySQL et phpMyAdmin

9.2.3. Etudes statistiques préliminaires

Création d'une base de données unique

Dans un premier temps nous avons répertorié les définitions des variables auprès des experts. Ce dictionnaire est présenté dans la Table 9.1.

Les bases de données *qualité* contiennent trois indices de qualité : le P_0, le PRI et $MOONEY$. Le PRI est en réalité obtenu en faisant le rapport du P_0 avec un autre indice de qualité : le P_{30}. On a $PRI = 100 * \frac{P_{30}}{P_0}$. Nous avons donc recalculé le P_{30} pour chaque exemple des bases de données *qualité*.

Comme la classe à prédire (ou à expliquer) est calculée à partir des indices de

Type de variable	Nom de variable	Définition
Variables Qualité	$DataRecepçao$	Date à laquelle le latex est réceptionné à l'usine
	$NùmeroRecepçao$	Numéro de réception des bennes à l'usine
	$Fantasia$	Parcelle + division
	$Parcelle$	Numéro de la parcelle
	$NùmeroBlocagem$	Numéro du mélange qui a servi pour la benne
	$PesoPem(kg)$	Poids des bennes en kg
	DRC	Teneur en caoutchouc
	$Borr.Seca(kg)$	Poids du contenu des bennes après séchage du latex
	P_0	Plasticité initiale
	PRI	Indice de rétention de la plasticité (Plasticity Retention Index)
	$MOONEY$	Indice de viscosité de $MOONEY$
	$Controle$	Contrôle (oui ou non)
	$InspecionadoPor$	Nom de l'inspecteur
	$Contaminantes$	Contaminants
	$Annéeagricole$	Année agricole (exemple 01/02)
Variables Parcellaires	$P.Monitora$	Regroupement de parcelles
	$Clone$	Type de clone utilisé dans une parcelle
	Ha	Nombre d'hectares de la parcelle
	$Mois/Annéedeplanting$	Mois/année de mise en terre des hévéa d'une parcelle
	$Annéedepremièreouverture$	Année des premières saignées des hévéa d'une parcelle
	$Panel$	Panneau (= localisation de la saignée sur le tronc des arbres)
	$Systèmedesaignée$	Système de saignée utilisé sur une parcelle
	$Arbressaignés$	Nombre d'arbres saignés par parcelles
	$kgsec$	Poids à sec de latex produit par parcelle et par année
	$RAMPAS$	Benne de ramassage
Variables météorologiques	$TempMAX$	température maximale
	$TempMIN$	Température minimale
	$TempMED$	Température médiane
	$URMAX$	Humidité relative maximale
	$URMIN$	Humidité relative minimale
	$URMED$	Humidité relative médiane
	$HorasdeSol$	Heures d'ensoleillement
	$Evaporaçao(mm)$	Evaporation en mm
	$PICHE$	Evaporation Piche
	$ETP-Bouchet$	Evapo-transpiration
	$Precipitaçao$	Précipitation

TABLE *9.1.:* Dictionnaire des variables hévéa

qualité du caoutchouc, les bases de données *qualité* ont servi de point de départ à l'élaboration d'une base de données unique. Chaque ligne de la base de données correspond à un prélèvement d'une benne arrivant à l'usine. La date qui lui est associée est renseignée de même que les données météorologiques correspondantes.

Du fait des mélanges effectués dans les bennes (qui contiendront en général plusieurs coagulums de plusieurs parcelles différentes), les données parcellaires ont dû recevoir un traitement particulier. Ces mélanges sont source d'incertitude sur la composition du coagulum, incertitude qui pourra être prise en compte par les arbres E^2M (voir Section 9.4.1). Des études sur les bennes "pures", ne contenant que les coagulums d'une seule parcelle, auraient été préférables, mais celles-ci sont en trop petit nombre.

Certaines variables quantitatives parcellaires (Ha, $Arbres\,saignés$ et kg_sec) ont été sommées pour les parcelles correspondant à des mêmes bennes. Les autres variables ont été pondérées. Cette pondération tient compte des quantités de coagulum produites par parcelle et par année agricole. Ensuite, pour les variables parcellaires quantitatives ne pouvant simplement être sommées

($Mois/Annéedeplanting$ et $Annéedepremièreouverture$) leur moyennes furent calculées pour chaque benne à l'aide de cette pondération. Pour les variables parcellaires nominatives ($Clone$, $Panel$ et $Systèmedesaignée$), chacune de leurs modalités fut transformée en variable quantitative correspondant à la proportion de la parcelle dans la benne. Par la même occasion, les variables nominatives parcellaires possédant un grand nombre de modalités furent transformées en variables numériques, ce qui constitue un avantage non-négligeable pour l'implémentation d'arbres de décision $CART$, qui doivent considérer, pour les variable nominales, tous les sous-ensembles de modalités possibles.

Exemple : Supposons que la benne B récolte le coagulum des parcelle p_1, p_2, p_3 et p_4, qu'en 2003/2004, les quantités de coagulum produites par ces parcelles furent respectivement de 1, 2, 3 et 4 tonnes, que la variable $clone$ vaut pour ces parcelles respectivement $PR255$, $GT1$, $PB217$ et $PB217$.

Les pondérations de ces quatre parcelles seront donc, pour l'année agricole 2003/2004 respectivement $\frac{1}{10}$, $\frac{2}{10}$, $\frac{3}{10}$ et $\frac{4}{10}$. Dans la nouvelle base de données, chaque exemple correspondant à la benne B et à cette année agricole aura alors, comme proportion de chaque clone, 0 pour tous les clones sauf ces trois clones utilisés dans ces

quatre parcelles. Pour ces derniers, leur proportions vaudront respectivement $\frac{1}{10}$, $\frac{2}{10}$ et $\frac{3}{10} + \frac{4}{10} = \frac{7}{10}$.

Les dimensions de la base de données ainsi construite sont environ de 3000 prélèvements contenant chacun 380 attributs. Une première étude statistique fut réalisée pour observer les premières tendances, et connaître les éventuelles corrélations entre les variables.

Première étude statistique descriptive

Voici un premier résumé des caractéristiques de la base de données obtenue notée désormais BDD :

	Début	**Fin**	**Nombre de prélèvements**
première période	16/09/2003	18/08/2004	1818
deuxième période	04/09/2006	17/07/2007	1235
Total			3053

TABLE *9.2.:* Etalement des deux périodes d'étude

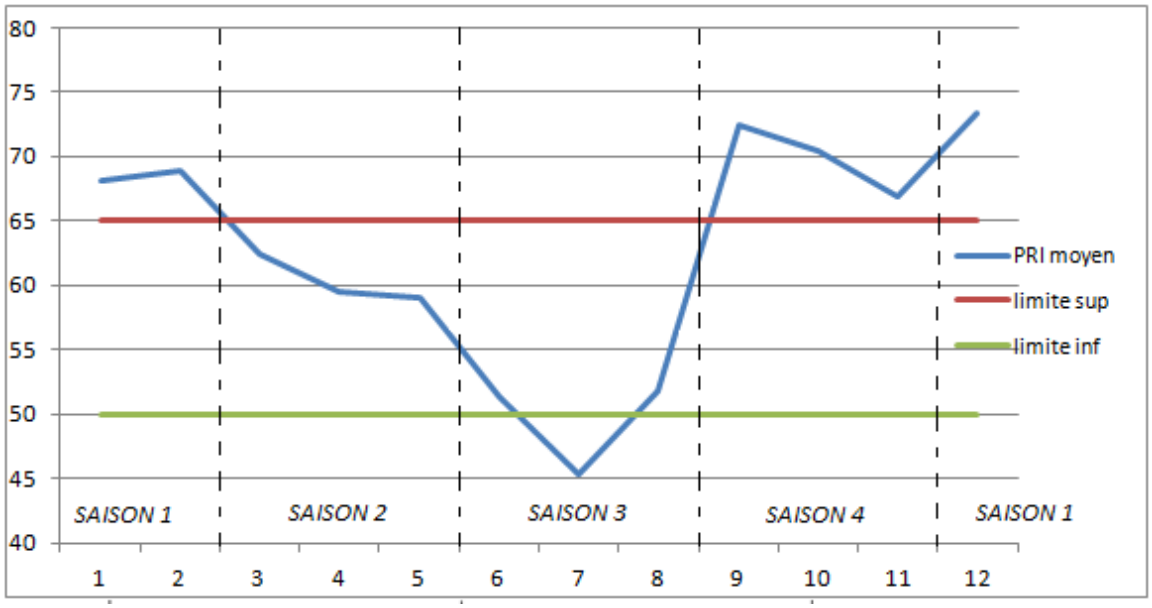
	P_O	PRI	P_{30}	$MOONEY$
P_O	1	0.55	0.82	0.88
PRI	0.55	1	0.91	0.38
P_{30}	0.82	0.91	1	0.64
$MOONEY$	0.88	0.38	0.64	1

TABLE *9.3.:* Matrice de corrélation des indices de qualité

De manière à résumer toutes les variables météorologiques, une variable "Saison" a été créée en fonction du niveau de PRI moyen et de son sens d'évolution sur les conseils de Mr. Sainte-Beuve. Pour cela, des niveaux frontaliers de PRI à 50 et 65 ont été retenus. La figure 9.4 représente l'évolution du PRI moyen en fonction des mois de l'année.

Il n'y a que le mois de Juillet où le PRI moyen est en dessous de 50, cependant pour les mois de Juin et d'Août, les niveaux de PRI moyen étant proches de 50, on pourra réunir ces trois mois en une saison. Ensuite, pour les mois de Septembre,

FIGURE 9.4.: Evolution du PRI moyen en fonction des mois de l'année



Octobre et Novembre, le PRI moyen est au dessus de 65 et est décroissant, on réunira donc ces mois dans une autre saison. Les mois de Décembre, Janvier et Février ayant aussi un PRI moyen supérieur à 65 (et décroissant), ils constitueront une autre saison. Les mois de Mars, Avril et Mai ayant un PRI moyen compris entre 50 et 65 et étant décroissant ils constitueront la dernière saison. Voici un résumé de la nouvelle variable "saison" créée artificiellement :

Saison	mois	PRI
1	Septembre, Octobre, Novembre	> 65 décroissant
2	Décembre, Janvier , Février	> 65 plutôt décroissant
3	Mars, Avril, Mai	$\in [50, 65]$
4	Juin, Juillet, Août	≤ 50

TABLE 9.4.: Création de la variable "saison" en fonction du PRI moyen

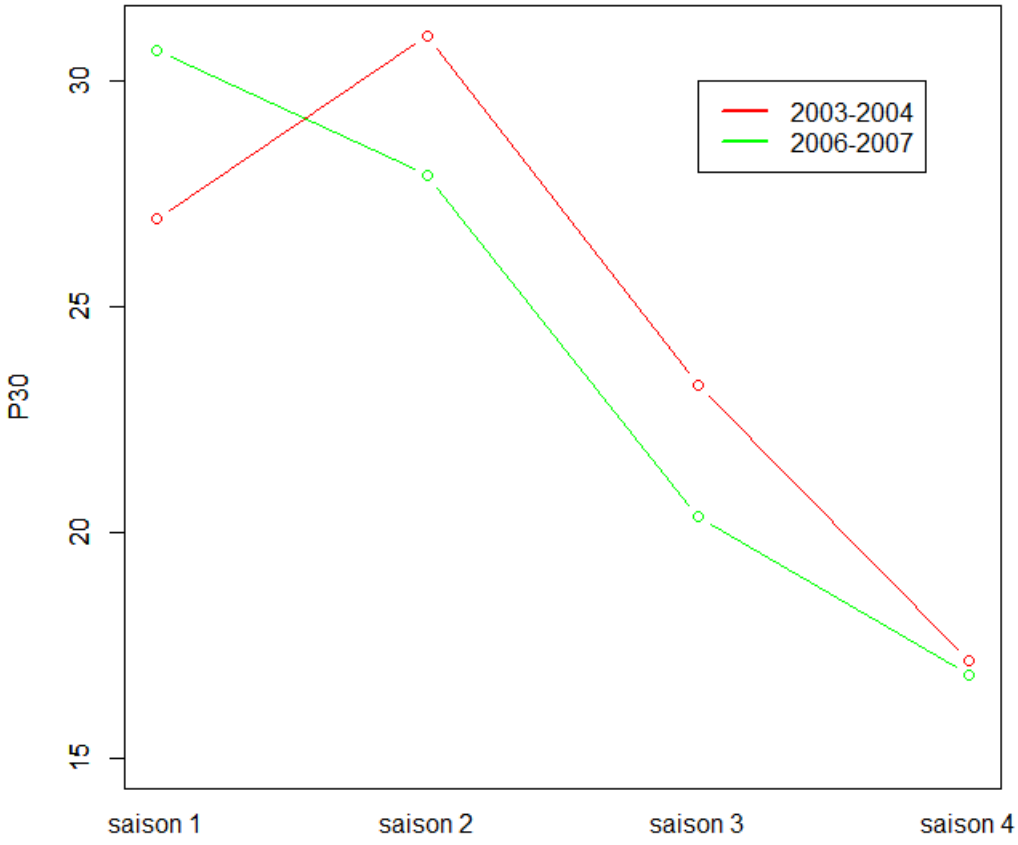
Voici un résumé des principales variables climatiques en fonction de la variable "saison" (les graphiques correspondants se trouvent en Annexe D) :

- Température médiane : à peu près constante de la saison 1 à la saison 3 (≈ 26°C) puis chute en saison 4 (≈ 22°C)
- Humidité relative médiane et ensoleillement : la saison 2 est la plus humide et la moins ensoleillée, la saison 4 est la moins humide et la plus ensoleillée

La Figure 9.5 représente l'évolution du P_{30} en fonction de la saison sur les 2 périodes

d'étude.

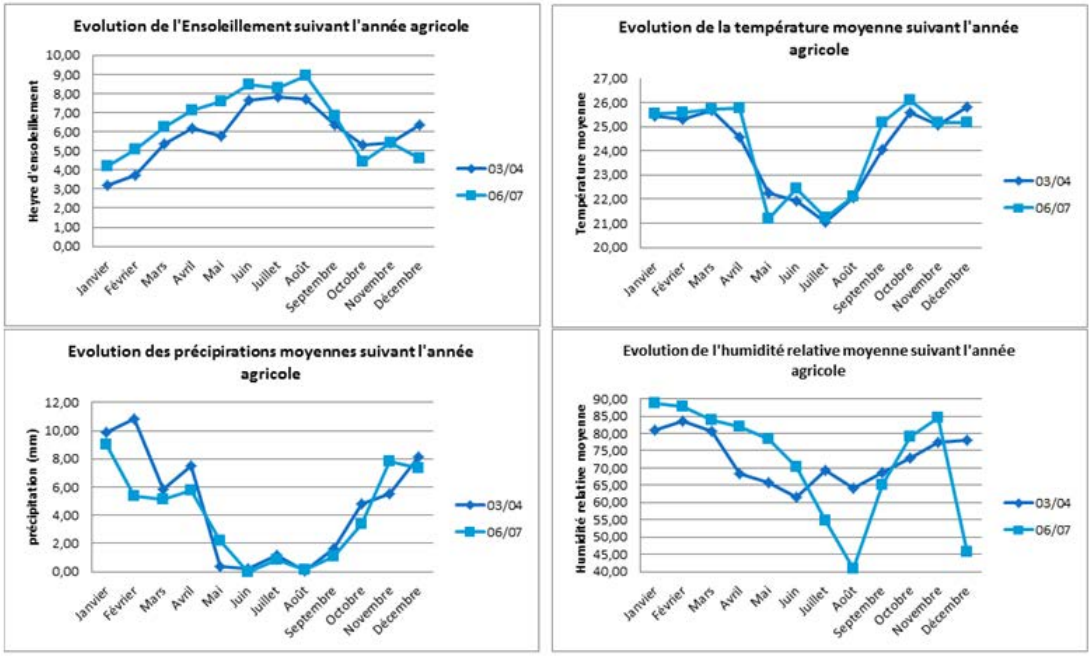
FIGURE 9.5.: Evolution du P_{30} en fonction de la saison sur les 2 périodes d'étude



Le P_{30} a donc été plus élevé en 2003/2004 qu'en 2006/2007 pour chaque saison sauf la saison 1, pour laquelle le P_{30} moyen a été supérieur en 2006/2007. De plus le P_{30} a évolué de la même façon (décroissante) sur les deux périodes d'étude. De manière à identifier le phénomène pouvant avoir expliqué cette inversion en saison 1, la Figure 9.6 représente l'évolution de quatre variables climatiques en fonction du mois et par période d'étude :

La seule évolution climatique apparaissant comme pouvant expliquer cette inversion des valeurs de P_{30} moyen par saison entre les deux périodes d'études en saison 1 semble être l'ensoleillement. En effet, d'après la Figure 9.5 le P_{30} moyen par sai-

FIGURE 9.6.: Evolution du climat en fonction du mois sur les 2 périodes d'étude



son a été supérieur en 2003/2004 par rapport à 2006/2007 sauf en saison 1 (i.e. de Septembre à Novembre), or d'après la Figure 9.6 l'année agricole 2003/2004 à été plus ensoleillée que l'année 2006/2007 sauf justement pendant la saison 1.

Premières conclusions

- la saison 4 donne les moins bon caoutchoucs ce qui s'explique par la définition de la variable "Saison"
- Contrairement à la température, l'humidité et les précipitations, l'ensoleillement semble avoir un impact négatif sur la qualité du caoutchouc.

L'ensemble des statistiques descriptives est disponible en annexe D.

9.3. Etude statistique prédictive sans incertitude sur les données

Une étude statistique prédictive est ici présentée. Les différents indices de qualité, initialement continus, sont discrétisés de différentes manières et des arbres de décision sont construits à partir de la BDD. Ces arbres sont ensuite éventuellement

évalués en terme d'erreur prédictive et interprétés. Du fait de l'instabilité des arbres de décision, de nombreuses expériences ont été réalisées. Celles qui nous permirent d'aboutir à nos principales conclusions sont ici résumées, d'autres études sont présentées en Annexe D.

Deux types d'expériences sont réalisées :

- les arbres sont appris sur une proportion de $\frac{2}{3}$ de la BDD, puis évalués sur le $\frac{1}{3}$ restant.
- les arbres sont appris sur la totalité de la BDD (apprentissage "total").

Pour chaque configuration (i.e. chaque découpage en classes), l'échantillon initial est aléatoirement divisé en échantillon d'apprentissage (2/3 de l'échantillon initial) et échantillon de test (1/3 de l'échantillon initial), et un arbre est construit avec la fonction rpart du logiciel R (paramétrée selon le modèle CART, i.e. l'arbre est entièrement déroulé, pas de pré-élagage donc, puis est élagué selon le critère de coût-complexité dont le paramètre est fixé par validation croisée à 10 couches) à partir de l'échantillon d'apprentissage et est évalué sur l'échantillon de test. Les arbres représentés sur les figures sont parfois ré-élagués de manière à être plus facilement lisibles. Les taux d'erreur sont ensuite calculés comme la proportion de *mauvaise* classification des exemples de l'échantillon test.

Dans le premier cas, pour chaque configuration de découpage des classes, 100 tests sont effectués. De manière à définir les variables attributs les plus *pertinentes* pour l'explication de la qualité du caoutchouc, pour chaque attribut la moyenne des profondeurs d'apparition de l'attribut dans les arbres est calculée. Les attributs présentant les profondeurs moyennes les plus faibles sont ceux expliquant le mieux cette qualité du caoutchouc. Les attributs ainsi sélectionnés sont : la Saison, le clone, l'ensoleillement, le système de saignée, la température et le panneau. Les interactions entre attributs se lisent le long des différents chemins (partant du noeud initial et allant jusqu'aux feuilles).

Les figures 9.7, 9.8, 9.9 et 9.10 représentent des arbres de décision ainsi obtenus à l'aide du logiciel R et du paquetage *rpart*.

La Saison

Le caoutchouc produit pendant les saisons 1 et 2 (Septembre à Février), donc pendant les périodes chaudes et humides est de meilleure qualité que celui produit pendant les saisons 3 et 4 (Mars à Juin), donc pendant des périodes plutôt sèches,

fraîches et ensoleillées (voir par exemple la Figure 9.7).

Le clone

Les différents clones apparaissent comme ayant des impacts différents sur la qualité du caoutchouc. Le clone le plus bénéfique est le PR255. Les clones PB235, B7 et PB217 apparaissent aussi comme de bons clones. Le clone GT1 par contre, semble clairement mauvais (voir par exemple les Figures 9.7 et 9.10).

L'ensoleillement (variable "Horas de Sol")

Même si en période fraiche (saison 4) l'ensoleillement semble bénéfique à la qualité du caoutchouc (voir Figure D.21 en Annexe D), il apparaît que pendant les période de fabrication et de maturation du coagulum (i.e. environ de une semaine avant la saignée à une semaine après) l'ensoleillement ne doit idéalement pas être excessif (voir Figures D.16 et D.24 en Annexe D).

L'humidité relative

De manière générale, l'humidité a un impact positif sur la qualité du caoutchouc (voir Figure 9.10). Comme l'ensoleillement a tendance a faire diminuer l'humidité de l'air, il n'est pas surprenant qu'ensoleillement et humidité présentent des impacts contraires sur la qualité du caoutchouc. Cependant ces deux phénomènes n'étant pas totalement corrélés (-50% en moyenne sur les deux périodes d'étude), ainsi alors que l'impact de l'ensoleillement devenait positif en saison 4, celui de l'humidité semble positif pour toutes les saisons (voir Figure D.19).

Le système de saignée

Le choix du système de saignée présente un impact sur les indices P_0 et $MOONEY$. Lors de son intervention (entre les deux périodes recouvrant la BDD), le CIRAD a recommandé l'arrêt de l'utilisation du système de saignée $D4D5$ qui consiste à stimuler les arbres comme si ils étaient saignés tous les 5 jours mais à les saigner tous les 4 jours. Même si ce système vise à augmenter la production ce système

est néfaste à la qualité du caoutchouc (voir Figure D.19), contrairement au $D4$ qui semble, lui, avoir un impact positif sur la qualité du caoutchouc (voir Figure 9.10).

La température

La température présente une omniprésence dans les arbres obtenus. Pendant les périodes de fabrication et de maturation du coagulum, il semble que la température ne doive pas descendre en dessous d'un certain seuil (voir par exemple Figures 9.7 et 9.10).

Le panneau

Le panneau utilisé pour les saignées indique la position sur le tronc des arbres. Le tronc d'un arbre est divisé latéralement et horizontalement, on peut donc compter 4 faces : A, B, C et D. Les panneaux A et B (recouvrant tout le tronc dans sa partie inférieure) sont descendants, les saignées s'y pratiquent donc successivement de haut en bas. Les panneaux C et D sont inversés et sont donc saignés de bas en haut. Il apparaît que d'une part les panneaux B donnent du caoutchouc de meilleure qualité que les panneaux A, et que d'autre part les panneaux hauts sont préférables aux panneaux A (en considérant les panneaux A,B et C). D'un point de vue physiologique ceci n'est pas très surprenant car lorsque le panneau A est saigné (de bas en haut), l'hévéa répond à cette *agression* en fabricant des sucres qui sont amenés au niveau des incision par le bas et montent donc jusqu'au niveau des premières incision (donc hautes). Lorsque le panneau B est utilisé, ces stocks de sucres ayant débordé au niveau des panneaux A,B et C, ils favorisent certains métabolismes de fabrication du latex. Ceci est un résultat très intéressant car même si un phénomène équivalent était bien connu des agronomes pour ce qui est de la production de latex (qui est donc plus importante pour les positions hautes du panneau B), il se trouve que les experts de la qualité ne différenciaient jusqu'ici que les panneaux descendants (A et B) des panneaux inversés (C et D).

La densité des arbres sur la plantation (variable as/ha)

Il semble que les bennes provenant de parcelles contenant une grande densité d'arbres (la variable "as/ha", signifiant le nombre d'arbres saignés par hectares,

aura alors des grandes valeurs) contiennent du coagulum de moins bonne qualité que celles provenant de grandes parcelles où les arbres sont vraisemblablement plus éloignés les uns des autres (voir Figures 9.7 et 9.8).

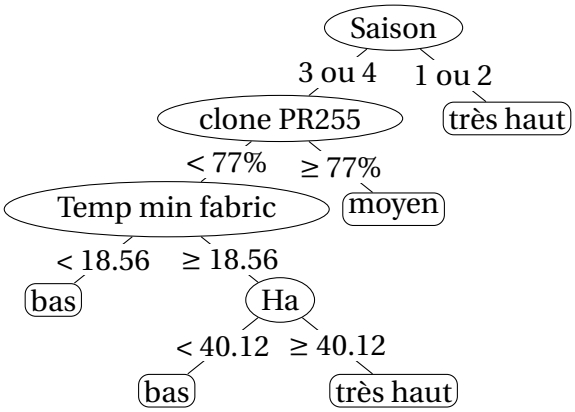


FIGURE *9.7.:* Arbre *CART* prédisant le P_0 découpé en 5 classes équiprobables

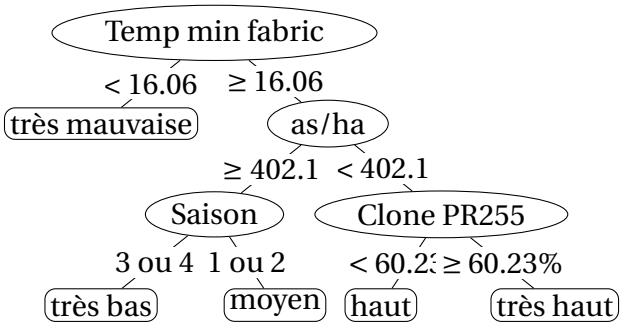


FIGURE *9.8.:* Arbre *CART* prédisant le *PRI* découpé en 5 classes équiprobables

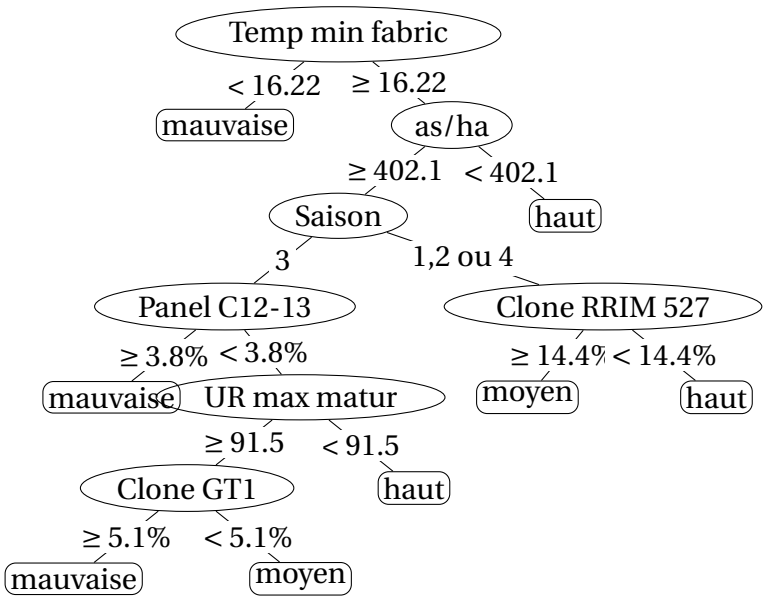


FIGURE *9.9.:* Arbre $CART$ prédisant le PRI découpé en 3 classes : mauvaise ([9.2 ;50[), moyenne ([50 ;65[) et bonne([65 ;113.3])

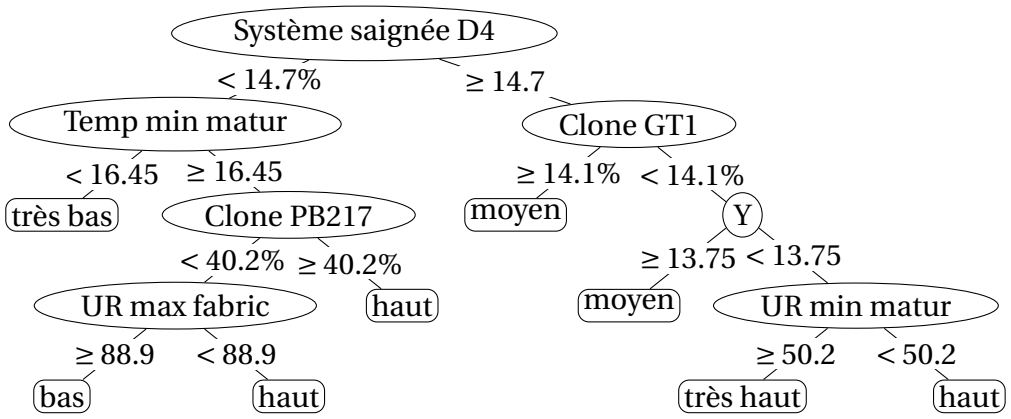


FIGURE *9.10.:* Arbre $CART$ prédisant le $MOONEY$ découpé en 5 classes équiprobables : période 2003/2004

9.4. Etude statistique prédictive avec incertitude sur les données

Après avoir défini un modèle d'incertitude des données faisant intervenir des fonctions de croyance, nous appliquons ici la méthodologie des arbres E^2M au problème de la prédiction de la qualité du caoutchouc naturel.

9.4.1. Modèles d'incertitude des données

De manière à améliorer le potentiel prédictif, et pour tenir compte des incertitudes relatives aux données *hévéa*, nous proposons ici un modèle d'incertitude pour les données qui sera utilisé pour la *crédibilisation* des données précises de la BDD en vue de l'apprentissage d'arbres de décision E^2M.

Les principales incertitudes relatives aux données "*hévéa*" issues de la plantation PEM sont les suivantes :

- incertitude des données de pluie en fonction de la distance de la benne à la station météo
- incertitude des données parcellaires du fait des mélanges dans les bennes

Incertitude des données de pluie

La pluie étant un phénomène très variable géographiquement, et la plantation PEM étant relativement étendue (environ 9000 Hectares), il est très probable qu'il puisse pleuvoir à différents endroits de la plantation sans qu'il pleuve pour autant (de la même manière) sur l'ensemble de la plantation. De plus, la station météorologique ayant mesuré les précipitations se situe à l'un des coin de la plantation, cela rend ces données de pluie encore plus incertaines pour l'ensemble des bennes de la plantation.

Notre modèle d'incertitude sera ici une distribution de possibilité. Pour chaque exemple i, une partie de la masse sera allouée au singleton correspondant à la donnée de pluie initiale précise notée ici w, le reste sera distribué à quatre intervalles emboîtés contenant cette donnée initiale. Les cinq intervalles obtenus seront définis par $[w(1-\delta), w(1+\delta)]$ pour $\delta \in \Delta = \{0, 0.25, 0.5, 0.75, 1\}$.

On définit une fonction $g : [0, d_{max}] \times \Delta \rightarrow [0,1]$ telle que, pour une donnée de pluie initiale w, correspondant à une benne provenant d'un ensemble de parcelles dont le centre de gravité est situé à une distance d de la station météorologique, on ait : $g(d,\delta) = m^W([w(1-\delta), w(1+\delta)])$.

Nous considérons deux types d'éléments focaux : les plus précis ($\delta < 0.5$) et les moins précis ($\delta \geq 0.5$). Nous partons du principe que plus loin sont les parcelles de la station météorologique, plus petite sera la masse allouée aux éléments focaux

considérés comme relativement précis et plus grande sera l amasse allouée aux éléments focaux très imprécis.

Les contraintes devant être vérifiées par $g(d,\delta)$ sont les suivantes :

$$\forall (d,\delta) \in [0, d_{max}] \times \Delta, \quad g(d,\delta) \geq 0 \tag{9.1}$$

$$\forall d \in [0, d_{max}], \quad \sum_{\delta \in \Delta} g(d,\delta) = 1 \tag{9.2}$$

$$\begin{cases} \delta < 0.5 & \rightarrow & \frac{\partial g}{\partial d} < 0 \\ \delta \geq 0.5 & \rightarrow & \frac{\partial g}{\partial d} > 0 \end{cases} \tag{9.3}$$

Les contraintes (9.1) et (9.2) viennent de la définition d'une masse de croyance. Une solution simple à ce problème est d'utiliser une combinaison convexe de deux fonctions linéaires, une croissante avec d pour les éléments focaux les plus précis ($\delta < 0.5$) et une décroissante avec d pour les plus imprécis ($\delta \geq 0.5$), d'où la contrainte (9.3). De cette manière nous proposons la solution suivante :

$$g(d,\delta) = \delta(\frac{2d}{5d_{max}}) + (1-\delta)(\frac{2}{5} - \frac{2d}{5d_{max}}) \tag{9.4}$$

Exemple 5. *Considérons trois données de pluie initiales $w_1 = 0$, $w_2 = 10$ et $w_3 = 30$ issues d'ensembles de parcelles dont les centres de gravité sont respectivement situés à $20km$, $50km$ et $2km$ de la station météorologique. Avec $d_{max} = 80$, on obtient*

$$\left\{ m_1(\{0\}) = 1 \right. \quad \begin{cases} m_2(\{10\}) = 0.15 \\ m_2([7.5,12.5]) = 0.175 \\ m_2([5,15]) = 0.200 \\ m_2([2.5,17.5]) = 0.225 \\ m_2([0,20]) = 0.250 \end{cases} \quad \begin{cases} m_3(\{30\}) = 0.39 \\ m_3([22.5,37.5]) = 0.295 \\ m_3([15,45]) = 0.2 \\ m_3([7.5,52.5]) = 0.105 \\ m_3([0,60]) = 0.01 \end{cases}$$

L'absence de pluie (m_1) est donc ici considérée certaine alors que les parcelles les plus distantes de la station météorologique donneront des masses très incertaines (m_2) contrairement aux parcelles les plus proches de la station qui, elles, donneront des masses surtout concentrées sur des éléments focaux relativement précis (m_3).

Incertitude des données parcellaires du fait des mélanges dans les bennes

Dans la base de données précise, toutes les variables parcellaires catégoriques (et les deux variables quantitatives *Age de saignée en fermeture de la parcelle* et *Age de l'arbre lors de l'ouverture*) ont été transformées en ensemble de variables quantitatives dont chaque élément correspond à une modalité de la variable initiale, en pondérant les valeurs de ces variables prises par les différentes parcelles d'une même benne par les quantités de coagulum produites annuellement par ces parcelles.

Même si cette pondération semble assez naturelle, elle reste discutable, notamment du fait que pour une de ces variables, le clone par exemple, il est tout à fait possible (et même assez probable) que certains clones produisent plus de coagulum que d'autres. Une benne contenant deux parcelles de clones respectifs A et B et de quantités respectives de coagulum produites de 1000kg et de 100kg, pourrait en réalité contenir autant d'arbres de clone A que d'arbre de clone B si A produit 10 fois moins de latex que B. On a donc une incertitude relative à ces proportions de nouvelles variables.

Nous partons du principe que plus équilibrées sont les modalités d'une variable parcellaire, plus incertaines doivent être les masses représentant ces modalités. Pour ce faire, nous utilisons de simples intervalles autour de ces proportions (initialement calculées précisement par le rapport des rendements annuels) dont la largeur croit avec le degré de mélange de ces données parcellaires dans les bennes. Pour mesurer ce degré de mélange nous utilisons l'entropie de Shannon, notée ici ent, calculée sur l'ensemble des proportions de chaque parcelle au sein des bennes.

Nous choisissons le modèle suivant :

Pour chaque variable parcellaire X^j ayant r modalités de proportions $\{p_1,\dots,p_r\}$,

$$\begin{cases} m^{j_1}([\max(p_1 - \frac{ent(p_1,\dots,p_r)}{I.r}, 0), \min(p_1 + \frac{ent(p_1,\dots,p_r)}{I.r}, 1)]) &= 1 \\ \vdots & \\ m^{j_r}([\max(p_r - \frac{ent(p_1,\dots,p_r)}{I.r}, 0), \min(p_r + \frac{ent(p_1,\dots,p_r)}{I.r}, 1)]) &= 1 \end{cases} \tag{9.5}$$

où I est une constante permettant d'éviter les éléments focaux constitués des valeurs extrêmes de 0 et de 1 (notamment dans le cas binaire). Dans la suite, nous choisissons $I = 1$.

Nombre feuilles	$CART$	E^2M
5	0.6685	0.6690
10	0.6521	0.6348

TABLE 9.5.: Taux d'erreur obtenus par les arbres $CART$ et E^2M avec prise en compte de l'incertitude des données

Exemple 6. *On considère une benne contenant 75% de clone A et 25% de clone B. L'entropie calculée sur ces proportions est de 0.8113. On obtient alors les masses :*

$$\begin{cases} m^{clone\,A}([34.43\%, 100\%]) &= 1 \\ m^{clone\,B}([0\%, 65.57\%]) &= 1 \end{cases}$$

9.4.2. Expériences

De manière à observer les conséquences de l'utilisation du modèle d'incertitude des données proposé en Section 9.4.1, deux types d'expériences sont ici présentées de manière à comparer les arbres de décision construits selon l'algorithme $CART$ appliqué à la BDD avec ceux construits à partir de l'échantillon crédibiliste résultant de l'application de ces deux modèles d'incertitude des données à la BDD selon la méthodologie des arbres de décision E^2M. :

- validation croisée (apprentissage sur $2/3$ de la BDD puis évaluation sur le $1/3$ restant)
- apprentissage *total* (sur toute la BDD)

Pour les deux méthodologies ici comparées, dans tous les cas les même critères d'arrêt ont été utilisés. Ils sont composés d'un nombre maximum de feuilles (5 puis 10) et d'un gain de pureté relatif de 0.05. Les taux d'erreur moyen sont ici obtenus par une validation croisée à 3 couches. Ces taux d'erreur sont classiquement calculés (proportion de mauvaise classification), la classe n'étant ici pas incertaine il aurait été inutile d'utiliser les taux d'erreur crédibilistes proposés en Section 7.4. Du fait de la taille de la BDD (environ 3000 exemples contenant 300 attributs) et du fait qu'on ne recherche pas ici d'optimisation particulière, aucun arbre n'est élagué.

La table 9.5 représente les taux d'erreur obtenus par les deux méthodologies pour 5 puis 10 feuilles maximum. Les Figures 9.11 et 9.12 représentent les arbres obtenus cette fois en apprentissage sur toute la base BDD et son équivalent crédibiliste par les méthodologie $CART$ et E^2M pour 10 feuilles maximum (avec les mêmes critères d'arrêt que pour les tests avec échantillons d'apprentissage et de test).

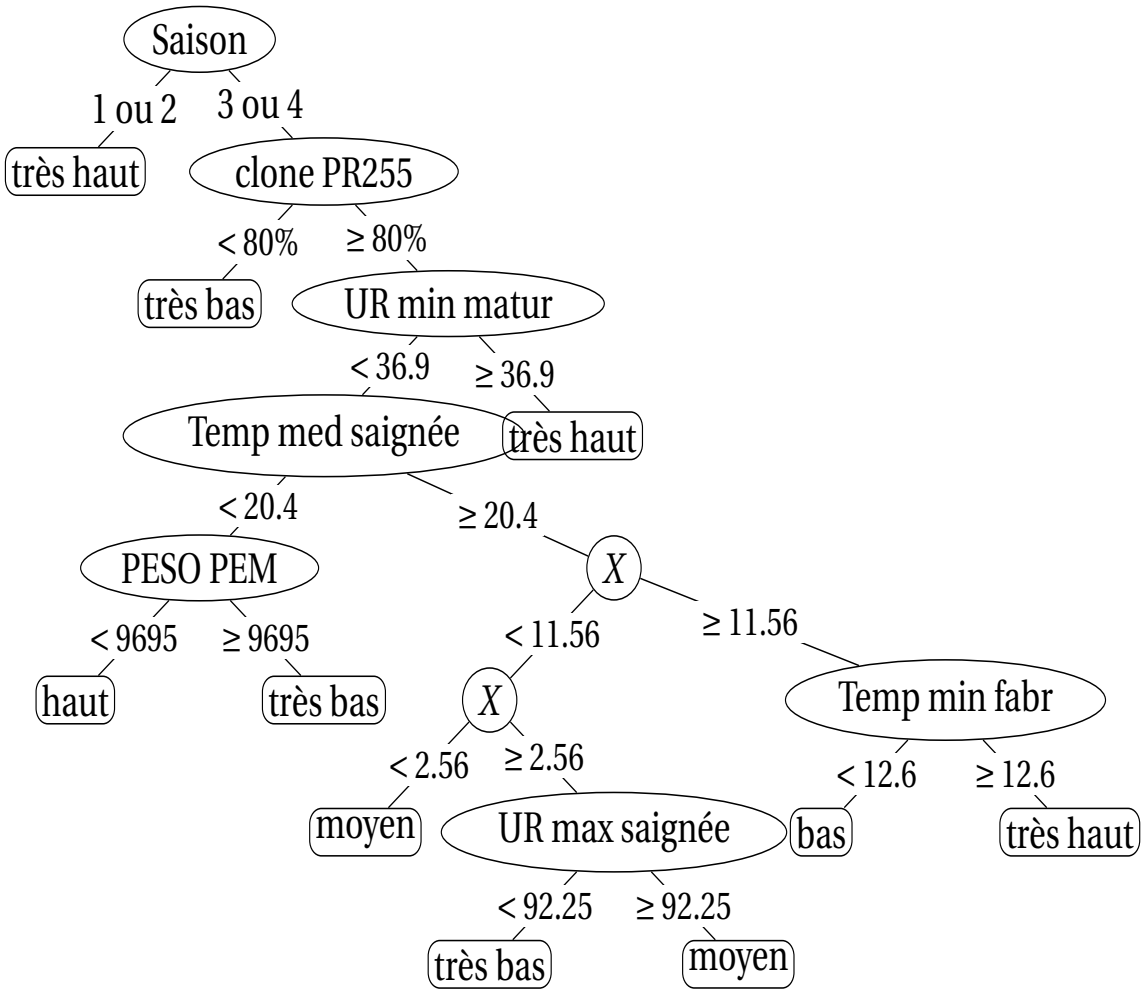


FIGURE *9.11.:* Arbre de décision *CART* à 10 feuilles maximum appris sur la totalité de la *BDD*

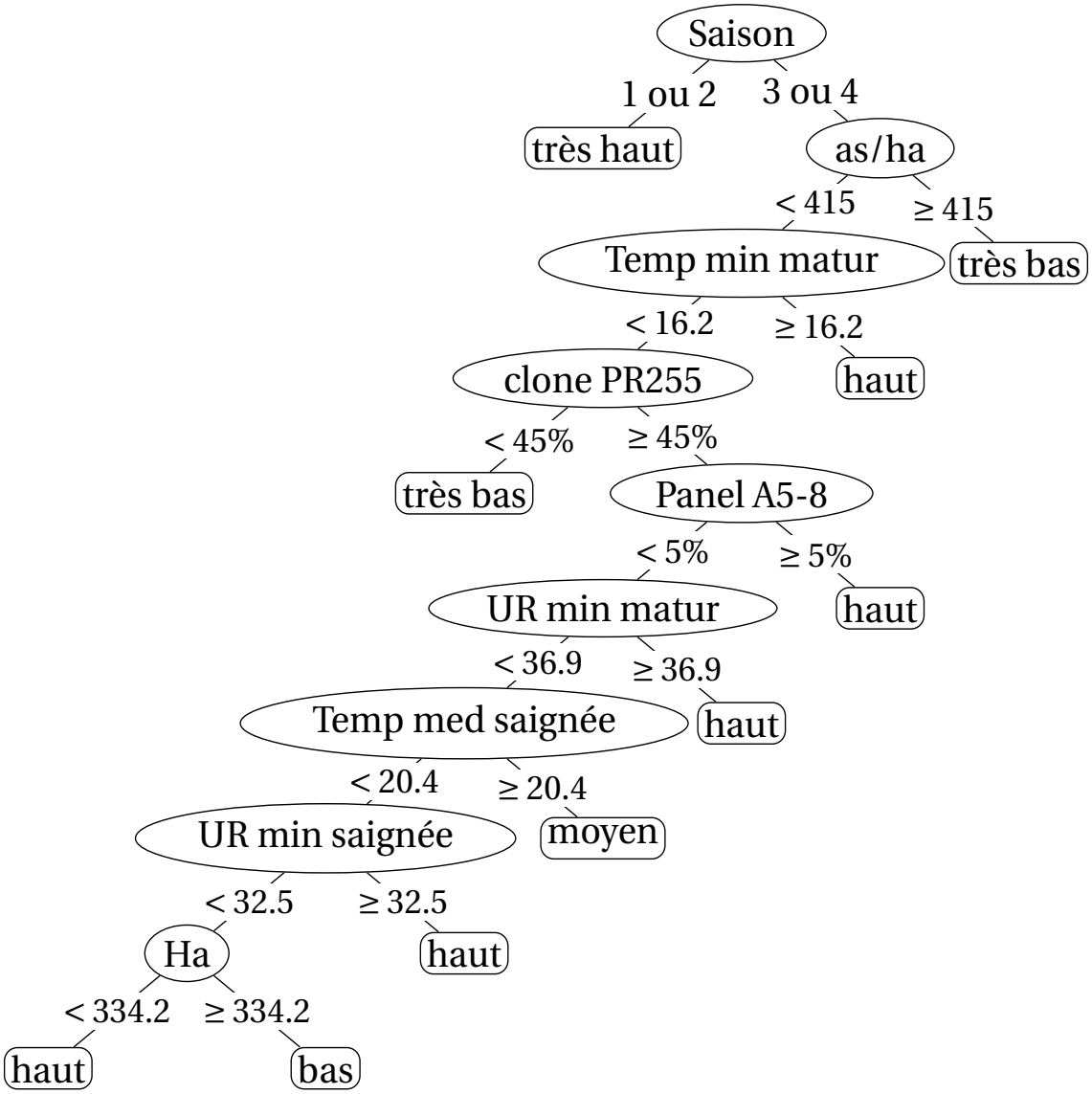


FIGURE *9.12.:* Arbre de décision E^2M à 10 feuilles maximum appris sur la totalité de la BDD

9.4.3. Conclusions

On observe tout d'abord sur le tableau 9.5 que les taux d'erreur des deux méthodologies semblent assez proches, les arbres E^2M à 10 feuilles sont toutefois légèrement plus efficaces que les arbres CART à 10 feuilles.

Si nous comparons les arbres ne prenant pas en compte de l'incertitude des données (Figure 9.11) avec ceux la prenant en compte (Figure 9.12), nous observons que la prise en compte de l'incertitude des données a ici pour conséquence de faire apparaître l'attribut "as/ha" qui représente la densité d'arbres sur la plantation. Les arbres de décision E^2M mettent donc en évidence une relation entre qualité et quantité de caoutchouc produites et permettent l'obtention de seuils de densités *acceptables* sur la plantation. Le clone PR255, même si il est toujours présent dans les arbres E^2M, apparaît plus *bas* dans l'arbre et a donc une moindre importance que pour les arbres $CART$. On peut aussi remarquer que contrairement aux arbres E^2M, les arbres $CART$ font apparaître les attributs "Peso PEM" (poids des bennes) et "X" (abscisse sur la carte de PEM). On peut donc conclure que la prise en compte de l'incertitude des données relativise le poids de ces variables. De la même manière, il est à noter que la variable panneau apparaît uniquement dans les arbres E^2M.

Conclusion

Nous rappelons ici les principales conclusions des différents travaux réalisés durant cette thèse. Nous présentons ensuite certaines perspectives de recherche qui pourraient conforter ou compléter ces travaux.

9.5. Bilan

Incertitude et fonctions de croyance

L'incertitude peut avoir différentes origines et donc être de différentes natures. La différenciation de l'incertitude en *épistémique/aléatoire* n'étant pas prise en compte en théorie des probabilités, il peut être intéressant d'envisager d'autres cadres de travail qui, eux, la prennent en compte.

La théorie des fonctions de croyance propose un tel cadre où les incertitudes épistémiques peuvent être modélisées de façon propre (souvent en amont de la modélisation des incertitudes aléatoires). L'approche de Dempster (voir Section 2.3.2), par exemple, permet de modéliser l'incertitude aléatoire par les fréquences *modifiées* (i.e. en $\frac{1}{n+1}$) tout en relativisant le modèle d'incertitude aléatoire par une incertitude épistémique à l'aide du terme $\frac{1}{n+1}$. Cette incertitude vient du manque de données (incertitude liée à la taille de l'échantillon), elle est donc épistémique car si on avait *théoriquement* une infinité de donnée, elle serait annulée. Les modèle d'incertitude crédibilistes permettent ainsi un traitement propre des incertitudes épistémiques. L'incertitude des données étant de nature épistémique, travailler dans le cadre des fonctions de croyance est donc assez naturel.

Cette théorie étant assez récente, beaucoup de travail reste encore à faire. La notion de vraisemblance, très utilisée par les probabilistes/statisticiens, a pendant longtemps été peu utilisée pour l'apprentissage à partir de données incertaines dans les autres théories de l'incertain (autres que probabilités classiques). Pour estimer un modèle, il est cependant possible d'étendre les raisonnements *par maximum de vraisemblance* aux données crédibilistes, donc épistémiquement incertaines.

L'étude de la vraisemblance crédibiliste présentée en Section 2.4.1 permet d'une part l'illustration d'estimations réalisées dans un cadre de données incertaines crédibilistes et d'autre part de mieux connaître certains des comportements de la vraisemblance crédibiliste. Au delà de ça, elle permet de faire un pont entre probabilités

et fonction de croyance notamment par la proposition 6 qui fait ressortir une certaine correspondance entre estimateur du maximum de vraisemblance crédibiliste et modèle génératif *probabiliste*. Ce type de *pont* est un moyen d'aider à une compréhension/interprétation des fonctions de croyance par la communauté statistique.

Arbres de décision crédibilistes

La première méthodologie d'arbre de décision présentée au Chapitre 6 s'inscrit dans le traitement de l'incertitude au sein d'un modèle de prédiction crédibiliste, les données d'apprentissage ne sont pas incertaines. Ce travail met en valeur l'apport analytique des fonctions de croyance en tant qu'outil de modélisation des incertitudes au sein de modèles prédictifs que sont les arbres de décision.

Dans les algorithmes classiques d'arbres de décision ($CART$, $C4.5$, etc), les calculs de gain de pureté mettent en oeuvre des estimations basées sur les fréquences qui sont strictement équivalentes à des maximisations de vraisemblance. L'idée des arbres de décision E^2M est de généraliser les arbres de décision classiques aux données incertaines crédibilistes. Cette généralisation englobe bien évidemment les données imprécises au sens ensembliste ($x \in A \quad \leftrightarrow m(A) = 1$), probabilistes (une distribution de probabilité définit une unique masse de croyance dont les éléments focaux sont des singletons) ou même les données possibilistes (une distribution de possibilité correspond à une unique masse de croyance dont les éléments focaux sont emboîtés). De par la largeur de son spectre d'application, cette méthodologie peut s'avérer coûteuse en terme de temps de calcul, nous avons donc aussi proposé une version des arbres de décision E^2M non optimale en terme d'erreur de prédiction mais où seuls certains paramètres sont ré-éstimés à chaque nouvelle coupure, ce qui réduit considérablement les temps de calculs (voir Section 7.2.5).

Pour élaguer un arbre de décision à partir de données incertaines, il faut être en mesure d'évaluer les capacité prédictives d'un arbres sur des données incertaines (de classe notamment) lors du choix du sous-arbre optimal. On est donc ramené à un problème dépassant largement le simple cadre d'élagage d'arbre de décision, il s'agit ici du problème général d'évaluation de classifieurs à partir de données incertaines. Nous proposons une solution à ce problème de manière cohérente avec l'esprit du reste du manuscrit : en maximisant la vraisemblance crédibiliste de l'échantillon binomiale représentant les erreurs de classification.

Cette vraisemblance crédibiliste, outil central dans ce manuscrit, permet de généraliser de nombreux raisonnements et méthodologies statistiques aux données incertaines représentées par des fonctions de croyance. La modélisation par des fonctions de croyance de l'incertitude des données est une étape très importante. En effet, pour qu'un classifieur tire pleinement parti de l'information contenue dans des données incertaines, le modèle de construction de ces fonctions de croyance doit être bien établi en considérant le type de fonction de croyance le plus adapté au problème (intervalles, distribution de probabilité, etc). Bien choisir la forme des éléments focaux est alors une étape décisive. Dans ce manuscrit nous proposons deux modèle crédibilistes d'incertitude des données (voir Section 9.4.1).

Caoutchouc naturel

L'application présentée au Chapitre 9 a différents objectifs. Le premier est d'identifier les variables ayant le plus d'impact sur la qualité du caoutchouc naturel et la *manière* ou le *sens* avec lequel elles agissent. Pour se faire, une étude statistique mettant en oeuvre les algorithmes classiques d'arbres de décision est tout d'abord réalisée. Il en ressort l'impact de différentes variables. Il apparaît que la variable "saison" est celle qui est le plus déterminante pour la qualité du caoutchouc. Cette variable, crée artificiellement, peut s'interpréter comme un résumé des variables climatologiques, on peut donc en conclure que le climat a un effet majeur sur la qualité du caoutchouc naturel.

Le second objectif de l'application *caoutchouc* était d'appliquer la méthodologie des arbres de décision E^2M. Après avoir *crédibiliser* les données de la BDD (elle-même construite à partir de plusieurs bases de données provenant de la plantation PEM), des arbres de décision E^2M ont été construit et comparer à ceux obtenus par l'algorithme $CART$. Il en est ressorti qu'en terme d'erreur de prédiction, les arbres E^2M sont aussi efficaces que les arbres $CART$ pour des *petits* arbres (5 feuilles) et même un peu plus efficaces pour des arbres un peu plus *grands* (10 feuilles). Au delà de ce constat, le résultat majeur apporté par les arbres E^2M est l'interprétation alternative (à celle des arbres $CART$) qu'ils permettent. En apprenant les arbres sur toute la BDD, les arbres de décision E^2M mettent ainsi en valeur une relation rendement/qualité (par l'apparition de la variable "as/ha" qui représente le nombre d'arbres saignés par hectare) et l'influence de la variable "Panneau".

L'apparition de cette variable "Panneau" est très intéressante car en considérant

aussi les autres arbres $CART$ obtenus lors de toutes les études statistiques réalisées dans ce manuscrit, il apparaît que la position haute du "Panneau" donne des caoutchoucs de meilleur qualité que la position basse. Ce constat est d'autant plus intéressant qu'il a permis une meilleure concertation des experts de l'hévéa coté plantation (qui eux connaissait bien un tel phénomène et ses origines concernant la quantité de latex produite) et coté usine (qui eux différenciaient surtout les panneaux descendants des panneaux inversés).

9.6. Perspectives

Extension au cas de classes numériques (i.e. d'arbres de régression)

La première perspective méthodologique des arbres de décision E^2M pourrait être d'étendre cette méthodologie au cas de classes numériques. Les arbres obtenus seraient alors des arbres de régression (et non plus de classification). Le premier obstacle à cette extension est qu'en général le critère de coupure des arbres de régression classiques est de minimiser la variance des classes, or les calculs de variance nécessitent une mesure de probabilité (de manière à pouvoir intégrer sur la classe) et il est bien connu qu'en théorie des foncions de croyance une telle mesure (unique) est difficile à trouver. Si par exemple les données de classe sont incertaines (et représentées par des fonctions de croyance), calculer des variances à l'aide d'une unique mesure de probabilité n'est pas évident.

Il est cependant tout à fait envisageable d'utiliser pour celà la même mesure de probabilité utilisée à l'étape (E) de l'algorithme E^2M : la combinaison de Dempster de la masse de croyance représentant les données de classe m^y avec la mesure de probabilité (paramétrique) P_θ du modèle (i.e. de l'arbre).

Une telle extension pourra bien évidemment être appliquée aux données "hévéa" qui contiennent justement des indices de classe numériques. Lors des études présentées dans ce manuscrit, ces classes étaient discrétisées de différentes manières mais il est à noter que les comportements des manufacturiers spécialisés dans le caoutchouc naturel peuvent être assez différents les uns des autres pour des mêmes valeurs prises par ces indices de qualité. Il serait donc assez pratique d'obtenir des prédictions de classe numériques qui pourront alors être interprétées différemment

selon les critères du manufacturier considéré.

Vraisemblance crédibiliste et algorithme E^2M

Il serait aussi très intéressant d'approfondir l'étude de la vraisemblance crédibiliste présentée en Section 2.4.1. Cette étude propose deux modélisations différentes de masses de croyance lorsqu'on connaît le niveau ϵ d'incertitude des données. Il serait donc intéressant, dans un premier temps, de trouver d'autres modélisations dans le même cadre. En considérant les deux modélisations proposées (l'une répartissant l'incertitude sur tous les singletons et l'autre l'allouant entièrement au cadre de discernement), il serait possible d'envisager toutes les modélisations intermédiaires (i.e. en répartissant une partie de l'incertitude sur les singletons et une autre sur le cadre de discernement). On pourrait de cette manière considérer un ensemble de modélisation crédibilistes (ou une fonction de croyance *imprécise*) plutôt qu'une seule. Cet ensemble de modèles donnant lieu à différentes vraisemblances crédibilistes et donc à différents estimateurs, adopter une approche ensembliste, et éventuellement choisir l'estimateur correspondant au maximum (calculé sur l'ensemble de ces vraisemblances) des maximums de vraisemblance crédibiliste pourrait être une solution.

Dans l'étude présentée en Section 2.4.1, l'estimateur du maximum de vraisemblance classique est obtenu à l'aide du modèle génératif des données. Il n'est donc pas surprenant que cet estimateur ait de meilleurs propriétés globales que son homologue crédibiliste. Dans ce cadre, les fonctions de croyance présentent surtout un intérêt lorsque le modèle génératif des données n'est pas ou mal connu. Il serait donc intéressant de faire varier le vrai modèle génératif des données tout en le gardant inchangé pour l'estimation par maximum de vraisemblance classique (on considérerait alors la notion d'erreur de modélisation et ses conséquences) et de comparer ensuite les estimation probabilistes et crédibilistes.

La généralisation de cette étude au cas multinomial reste aussi à faire (seule la fusion de classe est étudiée dans le cas de 3 classes). De cette manière on pourrait avoir une idée de la sensibilité des résultats à la complexité du modèle considéré (i.e. au nombre de classes). Une étude de la sensibilité des résultats au niveau

d'incertitude devrait aussi être réalisée pour les cas binomiaux et multinomiaux.

De manière générale, d'autres vraisemblances (ou ensembles de vraisemblances) pourraient aussi être considérées dans le cadre des fonctions de croyance. Les propriétés d'optimalité des estimateurs obtenus pourraient alors être comparées. En effet la vraisemblance crédibiliste proposée par Denoeux (dans [56]), de par son caractère *unique*, permet l'inférence de loi de probabilité à partir de fonctions de croyance (le paramètre θ_V estimé dans cette étude définit la loi de probabilité des données initiales) et semble *favoriser* les classes majoritaires (au sein de l'échantillon disponible). Il existe cependant d'autres moyens de passer d'un stade crédibiliste à un stade probabiliste, la probabilité pignistique proposée par Smet (voir Section 2.3.2) est sans doute le plus naturel. Cette probabilité, en répartissant uniformément l'incertitude sur toutes les classes, favorise les classes minoritaires. Il serait donc intéressant de comparer ces deux approches, ainsi que leur interprétations.

Il pourrait aussi être envisagé d'étendre toutes les *optimisations* et variantes connues de l'algorithme EM à l'algorithme E^2M (voir [100, 101, 102, 103, 104]).

Caoutchouc naturel

D'un point de vue purement agronomique, la plantation PEM étant de grande taille, il est fort probable que les sols présentent de grandes variabilités. Obtenir les données de sols, pourrait donc bien améliorer les prédictions tant en terme de qualité prédictive qu'en terme d'extraction d'information. De même, la variable "système de saignée" ayant un impact sur la qualité du caoutchouc, obtenir les données de stimulation permettrait sans doute une amélioration prédictive. La variable "Panneau" ayant aussi un impact non-négligeable sur la qualité du caoutchouc, fusionner certaines modalités (codées différemment pour les deux périodes d'études) pourrait permettre une meilleure compréhension de cet impact.

Il serait aussi assez important de trouver un modèle d'incertitude relatif à la date de saignée, et surtout à toutes les variables qui en découlent. En effet, pour les variables météorologiques considérées dans les différentes études, des moyennes sont calculées pour trois périodes définies en fonction de cette date de saignée qui

est supposée être le cinquième jour précédent l'arrivée à l'usine d'une benne. Cependant, comme souligné par les experts, cette date est en réalité assez variable. Souvent des bennes récoltent le latex provenant de plusieurs saignées successives lorsqu'elles ne sont pas suffisamment remplies lors de la première saignée. En observant les laps de temps entre deux passage à l'usine pour des bennes provenant de même groupement de parcelles il apparaît en effet une certaine variabilité de ces durées. Un modèle d'incertitude crédibiliste sur la date de saignée pourrait éventuellement utiliser ces laps de temps entre deux passage à l'usine. Il est à noter qu'au vu des données, une détection des *intrus* ou données aberrantes devra alors être mise en oeuvre de manière à ne pas tenir compte de tels laps de temps lorsqu'ils dépassent une certaine durée (des remaniements sur la plantation font apparaître des bennes n'allant pas à l'usine pendant plusieurs mois). Les jours de congés des saigneurs devraient aussi être pris en compte par le modèle. Une autre difficulté sera ensuite de transférer l'incertitude relative à la date de saignée au calculs de moyenne effectuées sur les périodes de *fabrication* et de *maturation* du latex qui sont définies à partir de cette date de saignée.

D'un point de vue méthodologique il serait évidemment intéressant d'utiliser éventuellement d'autres classifieurs que les arbres de décision pour expliquer la qualité du caoutchouc et de comparer les résultats obtenus avec ceux obtenus par les différents arbres de décision. Au delà de ça, la prise en compte de l'incertitude des données (grâce aux modèles d'incertitude proposés en Section 9.4.1) apportant des informations supplémentaires et améliorant la qualité prédictive, étudier la sensibilité de ces informations (et des qualités prédictives) aux différents paramètres de ces modèles sera aussi nécessaire.

Quatrième partie .

ANNEXES

Annexe A

A.1.

Preuve du lemme 1. Soit $A \subseteq \Omega$, on a alors

$$\Pi(A) - N_\Pi(A) \quad = \quad \Pi(A) - 1 + \Pi(\overline{A}) \quad \text{d'après (1.8)}$$

or d'après (1.7)

$$\begin{aligned} 1 &= \Pi(\Omega) = \Pi(A \cup \overline{A}) \\ &= max(\Pi(A), \Pi(\overline{A})) \quad \text{d'après (1.7)} \\ &\leq \Pi(A) + \Pi(\overline{A}) \end{aligned} \tag{A.1}$$

On obtient donc

$$\Pi(A) - N_\Pi(A) \quad \geq \quad 1 - 1 = 0$$

□

Preuve du lemme 2. Comme Π domine P, alors pour tout $A \subseteq \Omega$ on a

$$\Pi(A) \quad \geq \quad P(A)$$

et donc

$$\begin{aligned} N_\Pi(A) &= 1 - \Pi(\overline{A}) \\ &\geq 1 - P(\overline{A}) \\ &= P(A) \end{aligned}$$

□

Preuve du lemme 3. Il suffit de montrer que $\forall A \subseteq \Omega, \quad P^*(A) = \Pi(A)$

Soit $A \subseteq \Omega$, $x_A \in A$ tel que $x_A = \arg\max_{x \in A} \pi(x)$, $x^* \in \Omega$ tel que $\Pi(x^*) = 1$ (on suppose que $x^* \notin A$) et soit la mesure de probabilité P définie par

- $P(x_A) = \Pi(A)$
- $P(x^*) = 1 - \Pi(A)$

Il est alors évident que P est dominée par Π, donc que $P \in \mathscr{P}_\Pi$ et on a

$$P(A) = P(x_A) = \Pi(A)$$

Le raisonnement est similaire pour montrer que $\forall A \subseteq \Omega, \quad P_*(A) = N(A)$ en définissant P à l'aide de l'argmin de N □

Annexe B

B.1.

Preuve de la Proposition 1 : Par définition, $m'(\{w\}) = \frac{pl^w(w)}{\sum_{u\in\Omega_W} pl^w(u)}$

$$
\begin{aligned}
L(\theta; m^w) &= \mathbb{E}_\theta[pl^w(W)] \text{ par l'Equation 2.14} && \text{(B.1)}\\
&= \sum_{w\in\Omega_W} pl^w(w) P_\theta(W=w) && \text{(B.2)}\\
&= \Big(\sum_{u\in\Omega_W} pl^w(u)\Big). \sum_{w\in\Omega_W} \frac{pl^w(w)}{\sum_{u\in\Omega_W} pl^w(u)} P_\theta(W=w) && \text{(B.3)}\\
&= \Big(\sum_{u\in\Omega_W} pl^w(u)\Big). \sum_{w\in\Omega_W} m'(w) P_\theta(W=w) && \text{(B.4)}\\
&= \Big(\sum_{u\in\Omega_W} pl^w(u)\Big). \sum_{w\in\Omega_W} pl'(w) P_\theta(W=w) \text{ car } m' \text{ est une probabilité} && \text{(B.5)}\\
&= \Big(\sum_{u\in\Omega_W} pl^w(u)\Big). \mathbb{E}_\theta[pl'(W)] && \text{(B.6)}\\
&= \Big(\sum_{u\in\Omega_W} pl^w(u)\Big). L(\theta; m') && \text{(B.7)}
\end{aligned}
$$

□

Calcul de $\hat{\theta}_1$ ***:*** $\hat{\theta}_1$ s'obtient sans tenir compte d'ϵ (i.e. de notre connaissance relative à l'incertitude des données). Pour le calculer, on fait donc comme si w correspondait au

vrai échantillon v, on cherche donc ici à estimer θ_W.

$$\begin{aligned}
L(\theta_V; w) &= \prod_{i=1}^{n} P_{\theta_W}(w_i) \\
&= \theta_W^{n_1}(1-\theta_W)^{n-n_1} \\
\leftrightarrow \log L(\theta_W; w) &= n_1 \log(\theta_W) + (n-n_1)(1-\theta_W) \\
\rightarrow \frac{\partial}{\partial \theta_W} \log L(\theta_W; w) &= \frac{n_1}{\theta_W} - \frac{n-n_1}{1-\theta_W} \\
\text{on a donc } \frac{\partial}{\partial \theta_W} \log L(\hat{\theta}_1; w) &= 0 \\
\leftrightarrow n_1(1-\hat{\theta}_1) &= \hat{\theta}_1(n-n_1) \\
\leftrightarrow \hat{\theta}_1 &= \frac{n_1}{n}
\end{aligned}$$

□

Calcul de $\hat{\theta}_2$ ***:*** Il s'agit ici d'estimer θ_V à partir de w, ϵ et du modèle génératif 2.15. Dans un premier temps on exprimera θ_W en fonction de θ_V (selon 2.15) :

$$\begin{aligned}
\theta_W &= P_{\theta_V}(W=1) \\
&= P(B=1, U=1) + P_{\theta_V}(B=0, V=1) \\
&= P(B=1)P(U=1) + P(B=0)P_{\theta_V}(V=1) \\
&= \frac{\epsilon}{2} + (1-\epsilon)\theta_V
\end{aligned} \tag{B.8}$$

$$\begin{aligned}
L(\theta_V; w) &= \prod_{i=1}^{n} P_{\theta_V}(w_i) \\
&= P_{\theta_V}(W=1)^{n_1} P_{\theta_V}(W=0)^{n-n_1} \text{ (où } n_1 = |\{i : w_i = 1\}|) \\
&= (\theta_W)^{n_1}(1-\theta_W)^{n-n_1} \\
&= [\frac{\epsilon}{2} + (1-\epsilon)\theta_V]^{n_1}[1 - \frac{\epsilon}{2} - (1-\epsilon)\theta_V]^{n-n_1} \\
\log L(\theta_V; w) &= n_1 \log[\frac{\epsilon}{2} + (1-\epsilon)\theta_V] + (n-n_1)\log[1 - \frac{\epsilon}{2} - (1-\epsilon)\theta_V]
\end{aligned} \tag{B.9}$$

$$\frac{\partial}{\partial \theta_V} \log L(\theta_V; w) = \frac{n_1(1-\epsilon)}{\frac{\epsilon}{2} + (1-\epsilon)\theta_V} - \frac{(n-n_1)(1-\epsilon)}{1 - \frac{\epsilon}{2} - (1-\epsilon)\theta_V}$$

$$
\begin{aligned}
\text{On a donc } \frac{\partial}{\partial\hat{\theta}_2}\log L(\hat{\theta}_2;w) &= 0 \\
\Leftrightarrow \frac{n_1}{\frac{\epsilon}{2}+(1-\epsilon)\hat{\theta}_2} &= \frac{n-n_1}{1-\frac{\epsilon}{2}-(1-\epsilon)\hat{\theta}_2} \\
\Leftrightarrow n_1(1-\frac{\epsilon}{2})-n_1(1-\epsilon)\hat{\theta}_2 &= (n-n_1)\frac{\epsilon}{2}+(n-n_1)(1-\epsilon)\hat{\theta}_2 \\
\Leftrightarrow \hat{\theta}_2[(n-n_1)(1-\epsilon)+n_1(1-\epsilon)] &= n_1(1-\frac{\epsilon}{2})-(n-n_1)\frac{\epsilon}{2} = n_1-\frac{n_1\epsilon}{2}-\frac{n\epsilon}{2}+\frac{n_1\epsilon}{2} \\
\Leftrightarrow \hat{\theta}_2 n(1-\epsilon) &= n_1-\frac{n\epsilon}{2} \\
\Leftrightarrow \hat{\theta}_2 &= \frac{n_1}{n(1-\epsilon)}-\frac{\epsilon}{2(1-\epsilon)} \qquad \text{(B.10)}
\end{aligned}
$$

Et comme $\theta_V \in [0,1]$, on obtient

$$\hat{\theta}_2 = [\frac{n_1}{n(1-\epsilon)}-\frac{\epsilon}{2(1-\epsilon)}]_+ \wedge 1 \qquad \text{(B.11)}$$

où $[.]_+$ est la partie positive et $\wedge$ est l'opérateur $minimum$. □

Calcul de $\hat{\theta}_3$ ***:*** Dans un premier temps, on peut remarquer que selon la Définition 15, on a

$$L(\theta_V;m_i^V) = (1-\epsilon)L(\theta_V;\{w_i\})+\epsilon L(\theta_V;\{1,0\})$$

or $L(\theta_V;\{w_i\}) = P_{\theta_V}(V=\{w_i\}) = \begin{cases} \theta_V & \text{si } w_i=1 \\ (1-\theta_V) & \text{si } w_i=0 \end{cases}$,

de plus, $L(\theta_V;\{1,0\}) = P_{\theta_V}(V=\{1\})+P_{\theta_V}(V=\{0\}) = 1$ on obtient alors

$$L(\theta_V;m_i^V) = \begin{cases} (1-\epsilon)\theta_V+\epsilon \text{ si } w_i=1 \\ (1-\epsilon)(1-\theta_V)+\epsilon \text{ si } w_i=0 \end{cases}$$

$$
\begin{aligned}
\text{On a donc } L(\theta_V;m^V) &= \prod_{i=1}^{n} L(\theta_V;m_i^V) \\
&= [(1-\epsilon)\theta_V+\epsilon]^{n_1}[(1-\epsilon)(1-\theta_V)+\epsilon]^{n-n_1} \\
\log L(\theta_V;m^V) &= n_1\log[(1-\epsilon)\theta_V+\epsilon]+(n-n_1)\log[(1-\epsilon)(1-\theta_V)+\epsilon]
\end{aligned}
$$

$$\frac{\partial}{\partial\theta_V}\log L(\theta_V;m^V) = \frac{n_1(1-\epsilon)}{(1-\epsilon)\theta_V+\epsilon} - \frac{(n-n_1)(1-\epsilon)}{(1-\epsilon)(1-\theta_V)+\epsilon}$$

$$\begin{aligned}
\text{On a donc } \frac{\partial}{\partial\hat{\theta}_3}\log L(\hat{\theta}_3;w) &= 0 \\
\Leftrightarrow \frac{n_1}{(1-\epsilon)\hat{\theta}_3+\epsilon} &= \frac{(n-n_1)}{(1-\epsilon)(1-\hat{\theta}_3)+\epsilon} \\
\Leftrightarrow n_1(1-\epsilon)(1-\hat{\theta}_3)+n_1\epsilon &= (n-n_1)(1-\epsilon)\hat{\theta}_3+(n-n_1)\epsilon \\
\Leftrightarrow [-n_1(1-\epsilon)-(n-n_1)(1-\epsilon)]\hat{\theta}_3 &= (n-n_1)\epsilon-n_1\epsilon-n_1(1-\epsilon) \\
\Leftrightarrow [-n(1-\epsilon)]\hat{\theta}_3 &= n\epsilon-n_1\epsilon-n_1\epsilon-n_1+n_1\epsilon \\
\Leftrightarrow \hat{\theta}_3 &= \frac{n\epsilon-n_1\epsilon-n_1}{-n(1-\epsilon)} \\
\Leftrightarrow \hat{\theta}_3 &= \frac{n_1}{n}\frac{(1+\epsilon)}{(1-\epsilon)} - \frac{\epsilon}{(1-\epsilon)}
\end{aligned}$$

Et comme $\theta_V \in [0,1]$, on obtient

$$\hat{\theta}_3 = [\frac{n_1}{n}\frac{(1+\epsilon)}{(1-\epsilon)} - \frac{\epsilon}{(1-\epsilon)}]_+ \wedge 1 \tag{B.12}$$

où $[.]_+$ est la partie positive et $\wedge$ est l'opérateur $minimum$.

□

Preuve de la proposition 2. Dans un premier temps, on peut remarquer que selon la Définition 15, on a

$$L(\theta_V;m_i'^V) = (1-\frac{\epsilon}{2})L(\theta_V;\{w_i\}) + \frac{\epsilon}{2}L(\theta_V;\{w_i^c\})$$

où $\{w_i^c\}$ représente le singleton complémentaire de $\{w_i\}$ dans Ω_V (si $\{w_i\}=\{1\}$, alors $\{w_i^c\}=\{0\}$) or $L(\theta_V;\{w_i\}) = P_{\theta_V}(V=\{w_i\}) = \begin{cases} \theta_V & \text{si } w_i=1 \\ (1-\theta_V) & \text{si } w_i=0 \end{cases}$,

on obtient alors

$$L(\theta_V;m_i'^V) = \begin{cases} (1-\frac{\epsilon}{2})\theta_V + \frac{\epsilon}{2}(1-\theta_V) & \text{si } w_i=1 \\ (1-\frac{\epsilon}{2})(1-\theta_V) + \frac{\epsilon}{2}\theta_V & \text{si } w_i=0 \end{cases}$$

$$\begin{aligned}
\text{On a donc } L(\theta_V; m'^V) &= \prod_{i=1}^{n} L(\theta_V; m_i'^V) \\
&= [(1-\frac{\epsilon}{2})\theta_V + \frac{\epsilon}{2}(1-\theta_V)]^{n_1}[(1-\frac{\epsilon}{2})(1-\theta_V) + \frac{\epsilon}{2}\theta_V]^{n-n_1} \\
&= [(1-\frac{\epsilon}{2}-\frac{\epsilon}{2})\theta_V + \frac{\epsilon}{2}]^{n_1}[(-1+\frac{\epsilon}{2}+\frac{\epsilon}{2})\theta_V + (1-\frac{\epsilon}{2})]^{n-n_1} \\
&= [(1-\epsilon)\theta_V + \frac{\epsilon}{2}]^{n_1}[(1-\frac{\epsilon}{2}) - (1-\epsilon)\theta_V]^{n-n_1} \\
&= L(\theta_V; w) \text{ (voir Equation B.9)}
\end{aligned}$$

L'argmaximum de $L(\theta_V; m'^V)$ est donc l'argmax de $L(\theta_V; w)$. □

Preuve de la proposition 4. Soit f la fonction définie de $[0,1]$ dans $[0,1]$ par $\forall x \in [0,1], \quad f(x) = [\frac{x}{1-\epsilon} - \frac{\epsilon}{2(1-\epsilon)}]_+ \wedge 1$ où $[.]_+$ est la partie positive, $\wedge$ est l'opérateur $minimum$, et $\epsilon \in [0,1]$ est une constante. f est donc continue, en tant que combinaison de fonctions continues.

De plus, par la loi forte des grands nombre, $\frac{n_1}{n} = \frac{1}{n}\sum 1_{\{w_i=1\}}$ converge en probabilité vers $\theta_W = P(W=1)$.

On peut donc conclure que $\hat{\theta}_2 = f(\frac{n_1}{n})$ converge en probabilité vers $f(\theta_W) = \theta_V$. □

Preuve de la proposition 5. En reprenant la preuve de la proposition 4 avec la fonction $g : [0,1] \to [0,1]$ tel que $\forall x \in [0,1], g(x) = [\frac{(1+\epsilon)}{(1-\epsilon)}x - \frac{\epsilon}{(1-\epsilon)}]_+ \wedge 1$ qui est du même type que f, on trouve que $\forall x, y \in [0,1]$,

$$\begin{aligned}
|g(x) - g(y)| &\leq |\frac{(1+\epsilon)}{(1-\epsilon)}x - \frac{\epsilon}{(1-\epsilon)} - \frac{(1+\epsilon)}{(1-\epsilon)}y + \frac{\epsilon}{(1-\epsilon)}| \\
&= \frac{(1+\epsilon)}{(1-\epsilon)}|x-y|
\end{aligned}$$

g est alors lipschitzienne et donc uniformément continue, ce qui permet d'écrire en

remarquant que $\hat{\theta}_3 = g(\frac{n_1}{n})$:

$$\begin{aligned}
\lim_{n\to\infty} E[\hat{\theta}_3] &= \lim_{n\to\infty} E[g(\frac{n_1}{n})] \\
&= E[g(\theta_W)] \text{ par continuité uniforme de } g \\
&= g(\theta_W) \\
&= [\frac{(1+\epsilon)}{(1-\epsilon)}\theta_W - \frac{\epsilon}{(1-\epsilon)}]_+ \wedge 1 \\
&= \{\frac{(1+\epsilon)}{(1-\epsilon)}[\frac{\epsilon}{2} + (1-\epsilon)\theta_V] - \frac{\epsilon}{(1-\epsilon)}\}_+ \wedge 1 \\
&= \{\frac{1}{(1-\epsilon)}[\frac{\epsilon}{2} + (1-\epsilon)\theta_V + \frac{\epsilon^2}{2} + \epsilon(1-\epsilon)\theta_V - \epsilon]\}_+ \wedge 1 \\
&= \{\frac{1}{(1-\epsilon)}[(1-\epsilon)(1+\epsilon)\theta_V + \frac{\epsilon^2}{2} - \frac{\epsilon}{2}]\}_+ \wedge 1 \\
&= \{\frac{1}{(1-\epsilon)}[(1-\epsilon)(1+\epsilon)\theta_V - \frac{\epsilon}{2}(1-\epsilon]\}_+ \wedge 1 \\
&= [(1+\epsilon)\theta_V - \frac{\epsilon}{2}]_+ \wedge 1 \\
&\neq \theta_V
\end{aligned} \tag{B.13}$$

Il est ensuite aisé de vérifier que

$$\begin{aligned}
(1+\epsilon)\theta_V - \frac{\epsilon}{2} < 0 \qquad & \text{si } \theta_V < \tfrac{\epsilon}{2(1-\epsilon)} \\
(1+\epsilon)\theta_V - \frac{\epsilon}{2} \in [0,1] \quad & \text{si } \tfrac{\epsilon}{2(1+\epsilon)} \leq \theta_V \leq \tfrac{2+\epsilon}{2(1+\epsilon)} \\
(1+\epsilon)\theta_V - \frac{\epsilon}{2} > 1 \qquad & \text{si } \theta_V > \tfrac{2+\epsilon}{2(1+\epsilon)}
\end{aligned}$$

□

Preuve de la proposition 6.

$$\begin{aligned}
\theta_W &= P_{\theta_V}(W=1) \\
&= P(B=1, V=0) + P_{\theta_V}(B=0, V=1) \\
&= P(B=1)P_{\theta_V}(V=0) + P(B=0)P_{\theta_V}(V=1) \text{ par indépendance de } V \text{ et } B \\
&= \frac{\epsilon}{1+\epsilon}(1-\theta_V) + \frac{1}{1+\epsilon}\theta_V
\end{aligned} \tag{B.14}$$

$$\begin{aligned}
L(\theta_V; w) &= \prod_{i=1}^{n} P_{\theta_V}(w_i) \\
&= P_{\theta_V}(W=1)^{n_1} P_{\theta_V}(W=0)^{n-n_1} \text{ (où } n_1 = |\{i : w_i = 1\}|) \\
&= (\theta_W)^{n_1}(1-\theta_W)^{n-n_1} \\
&= [\frac{\epsilon}{1+\epsilon}(1-\theta_V) + \frac{1}{1+\epsilon}\theta_V]^{n_1}[1 - \frac{\epsilon}{1+\epsilon}(1-\theta_V) + \frac{1}{1+\epsilon}\theta_V]^{n-n_1} \qquad \text{(B.15)} \\
\log L(\theta_V; w) &= n_1 \log[\frac{\epsilon}{1+\epsilon}(1-\theta_V) + \frac{1}{1+\epsilon}\theta_V] + (n-n_1)\log[1 - \frac{\epsilon}{1+\epsilon}(1-\theta_V) + \frac{1}{1+\epsilon}\theta_V]
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial\theta_V}\log L(\theta_V; w) &= n_1 \frac{\frac{1-\epsilon}{1+\epsilon}}{\frac{\epsilon}{1+\epsilon}(1-\theta_V) + \frac{1}{1+\epsilon}\theta_V} - (n-n_1)\frac{\frac{1-\epsilon}{1+\epsilon}}{1 - \frac{\epsilon}{1+\epsilon}(1-\theta_V) + \frac{1}{1+\epsilon}\theta_V} \\
&= \frac{n_1(1-\epsilon)}{\epsilon + (1-\epsilon)\theta_V} - \frac{(n-n_1)(1-\epsilon)}{1-(1-\epsilon)\theta_V} \qquad \text{(B.16)}
\end{aligned}$$

Puis en notant $\hat{\theta}_5$ l' argmax de $\log L(\theta_V; w)$, on a

$$\begin{aligned}
\text{On a donc } \frac{\partial}{\partial\theta_V}\log L(\hat{\theta}_5; w) &= 0 \\
\Leftrightarrow \frac{n_1}{\epsilon + (1-\epsilon)\hat{\theta}_5} &= \frac{n-n_1}{1-(1-\epsilon)\hat{\theta}_5} \\
\Leftrightarrow n_1 - n_1\hat{\theta}_5(1-\epsilon) &= (n-n_1)\epsilon + (n-n_1)\hat{\theta}_5(1-\epsilon) \\
\Leftrightarrow \hat{\theta}_5 &= \frac{n_1 - (n-n_1)\epsilon}{n(1-\epsilon)} = \frac{n_1(1+\epsilon)}{n(1-\epsilon)} - \frac{\epsilon}{1-\epsilon}
\end{aligned} \qquad \text{(B.17)}$$

Et comme $\theta_V \in [0,1]$, on obtient

$$\hat{\theta}_5 = [\frac{n_1(1+\epsilon)}{n(1-\epsilon)} - \frac{\epsilon}{1-\epsilon}]_+ \wedge 1 = \hat{\theta}_3 \qquad \text{(B.18)}$$

où $[.]_+$ est la partie positive et $\wedge$ est l'opérateur $minimum$. □

Différence d'erreur quadrative de prédiction entre les estimateurs du maximum de vraisemblances classique $\hat{\theta}_2$ et crédibiliste $\hat{\theta}_3$.

FIGURE *B.1.*: Variations de la différence d'erreurs quadratiques Δ_{c-p} en fonction de n et de ϵ pour $\theta_V = 0.05$

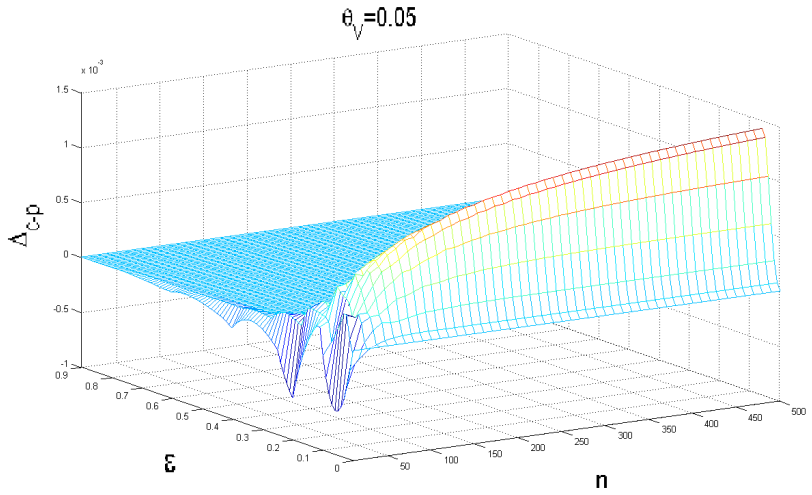
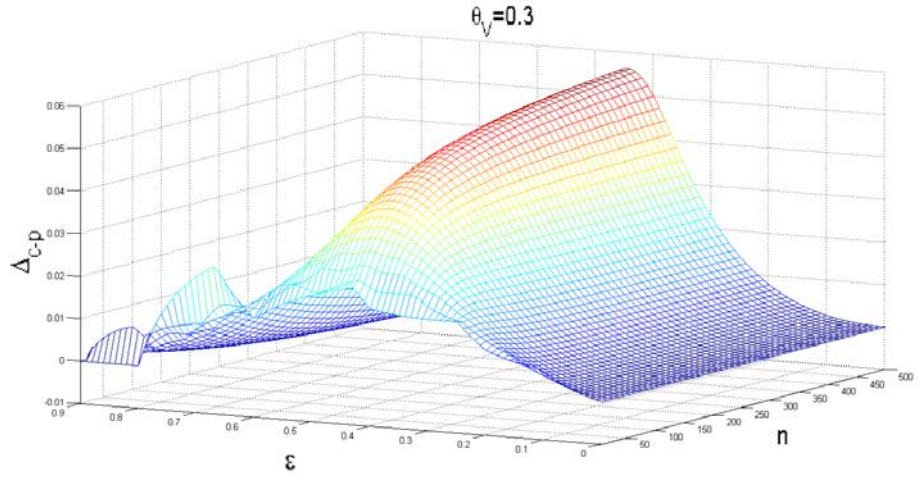


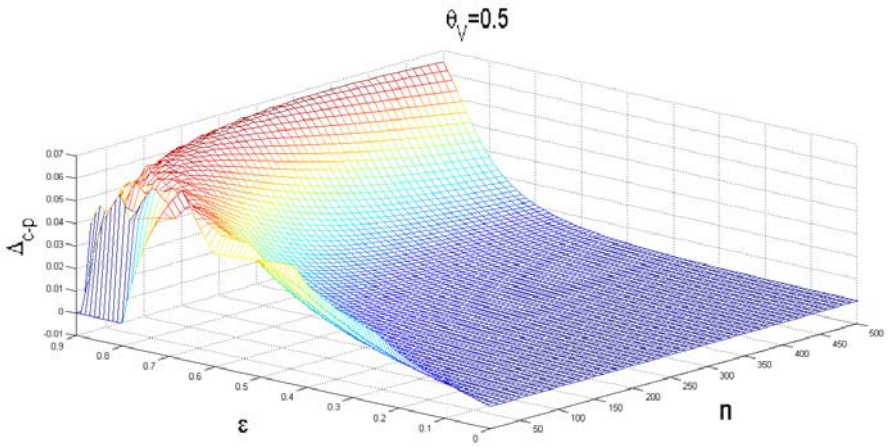
FIGURE *B.2.*: Variations de la différence d'erreurs quadratiques Δ_{c-p} en fonction de n et de ϵ pour $\theta_V = 0.3$



Les figures B.1, B.2 et B.3 représentent les variations de la différence d'erreurs quadratiques Δ_{3-2} entre $\hat{\theta}_3$ et $\hat{\theta}_2$ en fonction de la taille n de l'échantillon et du niveau d'incertitude ϵ des données, pour les cas $\theta_V = 0.05$, $\theta_V = 0.3$ et $\theta_V = 0.5$.

On a $\Delta_{3-2} = E[(\hat{\theta}_3 - \theta_V)^2] - E[(\hat{\theta}_2 - \theta_V)^2]$. On remarque que dans la plupart des configurations $\Delta_{c-p} > 0$, et donc que $\hat{\theta}_2$ est préférable à $\hat{\theta}_3$ (car commettant une moindre

FIGURE *B.3.:* Variations de la différence d'erreurs quadratiques Δ_{c-p} en fonction de n et de ϵ pour $\theta_V = 0.5$



erreur). Cependant on peut aussi remarquer que dans le cas $\theta_V = 0.05$, on a $\Delta_{c-p} < 0$ pour les petites valeurs de n ($n < 100$), et donc que $\hat{\theta}_3$ devient préférable à $\hat{\theta}_2$ pour ces configurations.

Dans toutes les configurations, l'effet du niveau d'incertitude ϵ sur cette différence d'erreurs quadratiques se manifeste par une vague haute (rendant $\hat{\theta}_2$ donc préférable à $\hat{\theta}_3$), pratiquement parallèle à l'axe des n (donc indépendant de n), et surgissant plus ou moins rapidement suivant la valeur du vrai paramètre θ_V considéré (plus θ_V est petit plus cet effet se fait ressentir pour des petites valeurs de ϵ).

Les cas où $\theta_V > 0.5$ ne sont pas ici exposés car étant parfaitement symétriques aux cas où $\theta_V < 0.5$. En effet il est équivalent de vouloir estimer le nombre de 1 dans un échantillon, puis de prendre la différence entre la taille de l'échantillon et l'estimateur obtenu pour estimer le nombre de 0, que de commencer par estimer ce nombre de 0, pour ensuite calculer la différence entre la taille de l'échantillon et l'estimateur obtenu pour estimer le nombre de 1. □

Annexe C

C.1.

Etape Maximisation de l'algorithme E^2M ***appliqué à l'estimation du paramètre d'un arbre*** E^2M. (voir 7.16)

$$Q(\theta;\theta^{(q)}) = \sum_{i,h} t_{ih}^{(q)} \log \pi_h + \sum_{i,h,k} \beta_{ih}^{k(q)} \log \alpha_h^k \quad \text{(C.1)}$$

$$= : f(\pi_1,...,\pi_H) = f(\pi) \text{ par définition de } \pi_h, \text{ on a : } \pi_H = 1 - \sum_{h=1}^{H-1} \pi_h \quad \text{(C.2)}$$

$$= : g(\alpha_1^1,...,\alpha_1^K,...,\alpha_H^1,...,\alpha_H^K) = g(\alpha) \quad \text{(C.3)}$$

calcul des $\pi_1,...,\pi_H$ **:**

Soit f une fonction définie par :

$$f(\pi_1,...,\pi_{H-1}) = \sum_{h=1}^{H-1} \log \pi_h \sum_i t_{ih}^{(q)} + \log(1 - \sum_{h=1}^{H-1} \pi_h) \sum_i t_{iH}^{(q)} + \sum_{i,h,k} \beta_{ih}^{k(q)} \log \alpha_h^k \quad \text{(C.4)}$$

Recherche du maximum local : $\frac{\partial f}{\partial \pi_1} = ... = \frac{\partial f}{\partial \pi_H - 1} = 0$

$$\longleftrightarrow \forall j \in \{1,...,H-1\}, \quad \frac{1}{\pi_j}\sum_i t_{ij}^{(q)} + \frac{-1}{1-\sum_{h=1}^{H-1}\pi_h}\sum_i t_{iH}^{(q)} = 0$$

$$\longleftrightarrow \forall j \in \{1,...,H-1\}, \quad \frac{\sum_i t_{ij}^{(q)}}{\pi_j} = \frac{\sum_i t_{iH}^{(q)}}{1-\sum_{h=1}^{H-1}\pi_h} = \frac{\sum_i t_{iH}^{(q)}}{\pi_H} \tag{C.5}$$

$$\longleftrightarrow \forall j \in \{1,...,H-1\}, \quad \pi_j = \pi_H \frac{\sum_i t_{ij}^{(q)}}{\sum_i t_{iH}^{(q)}}$$

$$\longrightarrow \forall (j,h) \in \{1,...,H-1\}, \quad \pi_h = \pi_j \frac{\sum_i t_{ih}^{(q)}}{\sum_i t_{ij}^{(q)}} \tag{C.6}$$

Puis par (C.5) on a : $\forall j \in \{1,...,H-1\}, \frac{\sum_i t_{ij}^{(q)}}{\pi_j} = \frac{\sum_i t_{iH}^{(q)}}{1-\sum_{\substack{h=1\\h\neq j}}^{H-1}\pi_h - \pi_j}$

et par (C.6) on otient :

$$\frac{\sum_i t_{ij}^{(q)}}{\pi_j} = \frac{\sum_i t_{iH}^{(q)}}{1-\sum_{\substack{h=1\\h\neq j}}^{H-1}\pi_j \frac{\sum_i t_{ih}^{(q)}}{\sum_i t_{ij}^{(q)}} - \pi_j}$$

$$\longleftrightarrow \quad \pi_j \sum_i t_{iH}^{(q)} = \sum_i t_{ij}^{(q)}\left(1 - \frac{\pi_j}{\sum_i t_{ij}^{(q)}} \sum_{\substack{h=1\\h\neq j}}^{H-1}\sum_i t_{ih}^{(q)} - \pi_j\right)$$

$$\longleftrightarrow \quad \pi_j\left(\sum_i t_{iH}^{(q)} + \sum_i t_{ij}^{(q)}\right) = \sum_i t_{ij}^{(q)}\left(1 - \frac{\pi_j \sum_i \sum_{\substack{h=1\\h\neq j}}^{H-1} t_{ih}^{(q)}}{\sum_i t_{ij}^{(q)}}\right)$$

$$\longleftrightarrow \quad \pi_j\left(\sum_i t_{iH}^{(q)} + \sum_i t_{ij}^{(q)} + \sum_i \sum_{\substack{h=1\\h\neq j}}^{H-1} t_{ih}^{(q)}\right) = \sum_i t_{ij}^{(q)}$$

$$\longleftrightarrow \quad \pi_j = \frac{\sum_i t_{ij}^{(q)}}{\sum_i \sum_{h=1}^{H} t_{ih}^{(q)}} \tag{C.7}$$

Or $t_{ik} = E[z_{ih}/\theta; pl] \longrightarrow \forall i \in \{1,...,n\}, \sum_{h=1}^{H} t_{ih} = 1$

$\longrightarrow \pi_j = \frac{1}{n}\sum_{i=1}^{n} t_{ij}^{(q)}$

Estimation des paramètres $\alpha_1^1,...,\alpha_1^K,...,\alpha_H^1,...,\alpha_H^K$:

$\alpha_h^k = P(\omega_k / t_h) \longrightarrow \sum_{k=1}^{K} \alpha_h^k = 1$

$\longrightarrow \forall h \in \{1,...,H\}, \qquad \alpha_h^K = 1 - \sum_{k=1}^{K-1} \alpha_h^k$

On obtient donc :

$$g(\alpha) = constant + \sum_i \sum_h (\sum_{k=1}^{K-1} \beta_{ih}^{k(q)} \log \alpha_h^k + \beta_{ih}^{K(q)} \log[1 - \sum_{j=1}^{K-1} \alpha_h^j]) \tag{C.8}$$

Et par annulation du Langrangien, on obtient :

$$\forall (h,k) \in \{1,...,H\} \times \{1,...,K-1\}, \qquad \frac{\partial g}{\partial \alpha_h^k} = \frac{\sum_i \beta_{ih}^{k(q)}}{\alpha_h^k} - \frac{\sum_i \beta_{ih}^{K(q)}}{1 - \sum_{j=1}^{K-1} \alpha_h^j} = 0 \tag{C.9}$$

$$\longleftrightarrow \forall (h,k) \in \{1,...,H\} \times \{1,...,K-1\}, \qquad \frac{\alpha_h^k}{\sum_i \beta_{ih}^{k(q)}} = \frac{\alpha_h^K}{\sum_i \beta_{ih}^{K(q)}}$$

$$\longrightarrow \forall (h,u,v) \in \{1,...,H\} \times \{1,...,K-1\}^2, \qquad \alpha_h^u = \alpha_h^v \frac{\sum_i \beta_{ih}^{u(q)}}{\sum_i \beta_{ih}^{v(q)}}$$

$$\longrightarrow \forall (h,k) \in \{1,...,H\} \times \{1,...,K-1\}, \qquad \alpha_h^k = \frac{\sum_i \beta_{ih}^{k(q)}}{\sum_i \beta_{ih}^{K(q)}} (1 - \sum_{\substack{j=1 \\ j \neq k}}^{K-1} \alpha_h^j - \alpha_h^k)$$

$$\longrightarrow \forall (h,k) \in \{1,...,H\} \times \{1,...,K-1\}, \qquad \alpha_h^k (1 + \frac{\sum_i \beta_{ih}^{k(q)}}{\sum_i \beta_{ih}^{K(q)}}) = \frac{\sum_i \beta_{ih}^{k(q)}}{\sum_i \beta_{ih}^{K(q)}} (1 - \sum_{\substack{j=1 \\ j \neq k}}^{K-1} \alpha_h^j)$$

$$\longrightarrow \forall (h,k) \in \{1,...,H\} \times \{1,...,K-1\}, \qquad \alpha_h^k (\frac{\sum_i \beta_{ih}^{K(q)} + \sum_i \beta_{ih}^{k(q)}}{\sum_i \beta_{ih}^{K(q)}}) = \frac{\sum_i \beta_{ih}^{k(q)}}{\sum_i \beta_{ih}^{K(q)}} (1 - \sum_{\substack{j=1 \\ j \neq k}}^{K-1} \alpha_h^k \frac{\sum_i \beta_{ih}^{j(q)}}{\sum_i \beta_{ih}^{k(q)}})$$

$$\longrightarrow \forall (h,k) \in \{1,...,H\} \times \{1,...,K-1\}, \qquad \alpha_h^k (\sum_i \beta_{ih}^{K(q)} + \sum_i \beta_{ih}^{k(q)}) = \sum_i \beta_{ih}^{k(q)} (1 - \alpha_h^k \frac{\sum_{\substack{j=1 \\ j \neq k}}^{K-1} \sum_i \beta_{ih}^{j}}{\sum_i \beta_{ih}^{k}})$$

pourtant $\beta_{ih}^{k(q)} = E[z_{ih} y_i^k / m^{Y,Z}; \theta^{(q)}]$ et comme $\sum_{k=1}^{K} y_i^k = 1$, on a :

$$\sum_{k=1}^{K} \beta_{ih}^{k(q)} = E[z_{ih} / m^{Y,Z}; \theta^{(q)}] = t_{ih}^{(q)} \tag{C.10}$$

On obtient donc :

$$\forall (h,k) \in \{1,...,H\} \times \{1,...,K-1\}, \qquad \alpha_h^k (\sum_i \beta_{ih}^{K(q)} + \sum_i \beta_{ih}^{k(q)} + \sum_{\substack{j=1 \\ j \neq h}}^{K-1} \sum_i \beta_{ih}^{j}) = \sum_i \beta_{ih}^{k}$$

$$\longrightarrow \forall (h,k) \in \{1,...,H\} \times \{1,...,K-1\}, \qquad \alpha_h^k = \frac{\sum_i \beta_{ih}^k}{\sum_i \sum_{j=1}^{K} \beta_{ih}^j}$$

$$\longrightarrow \forall (h,k) \in \{1,...,H\} \times \{1,...,K-1\}, \qquad \alpha_h^k = \frac{\sum_i \beta_{ih}^k}{\sum_i t_{ih}^{(q)}} \tag{C.11}$$

□

Annexe D

D.1. Statistiques descriptives :

		Moyenne	Ecart-type
P_0	BDD	37.07	7.28
	2003/2004	36.03	7.14
	2006/2007	38.61	7.23
P_{30}	BDD	23.94	9.54
	2003/2004	23.70	9.36
	2006/2007	24.30	9.78
PRI	BDD	62.62	18.25
	2003/2004	63.79	18.09
	2006/2007	60.90	18.34
$MOONEY$	BDD	75.75	12.21
	2003/2004	72.89	11.18
	2006/2007	79.97	12.43

TABLE *D.1.:* Indices de qualité

	BDD		2003/2004		2006/2007	
	moyenne	**écart-type**	**moyenne**	**écart-type**	**moyenne**	**écart-type**
Temp MAX	30.5	3.6	30.0	3.7	30.9	3.4
Temp MIN	18.4	3.9	18.5	3.7	18.4	3.9
Temp MED	24.4	2.7	24.3	2.8	24.7	2.7
UR MAX	92.7	4.1	90.3	3.6	95.2	2.9
UR. MIN	52.7	17.7	52.5	17.1	53.8	17.8
UR.MED	75.6	12.5	70.3	12.2	81.4	10.0
Heures soleil	5.7	3.5	5.4	3.5	5.9	3.4
Tank Class A	73.7	10.7	75.3	11.0	72.3	10.2
Evaporação (mm)	5.4	6.5	5.5	6.1	5.3	7.0
PICHE	15.4	8.0	15.4	8.2	15.4	7.9
mm	2.9	1.7	2.9	1.9	2.9	1.5
ETP - Bouchet	4.7	2.8	4.7	3.1	4.6	2.5
Precipitação	14.8	17.8	15.2	15.3	14.4	20.0

TABLE *D.2.:* Variables météorologiques

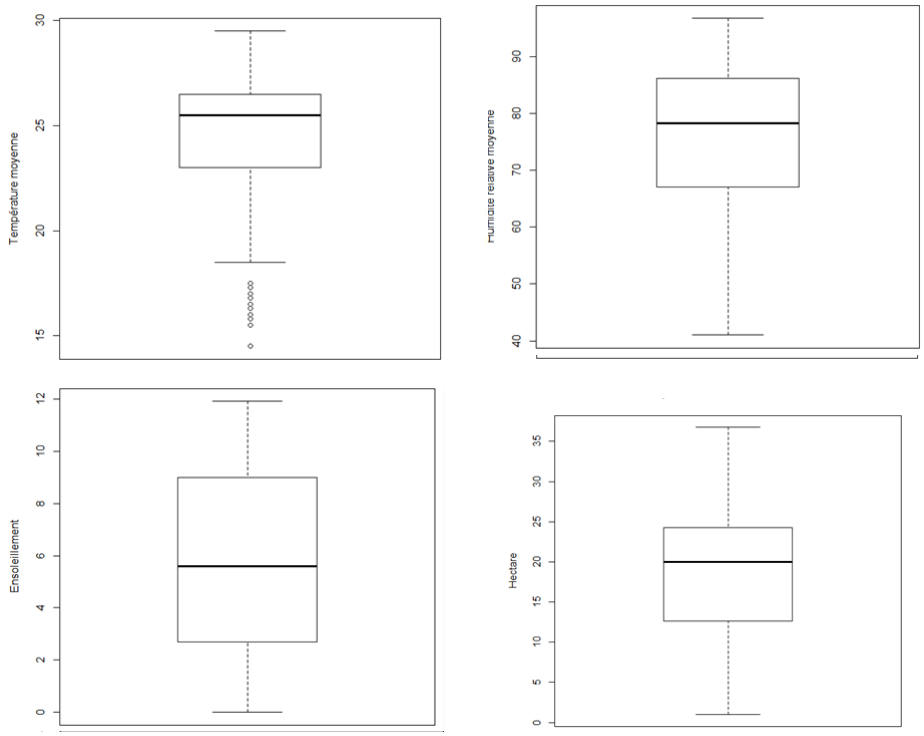


FIGURE *D.1.:* Boxplots des variables Tempéraure Médiane, UR Médiane, Heures d'ensoleillement et de la variable parcellaire *hectares*

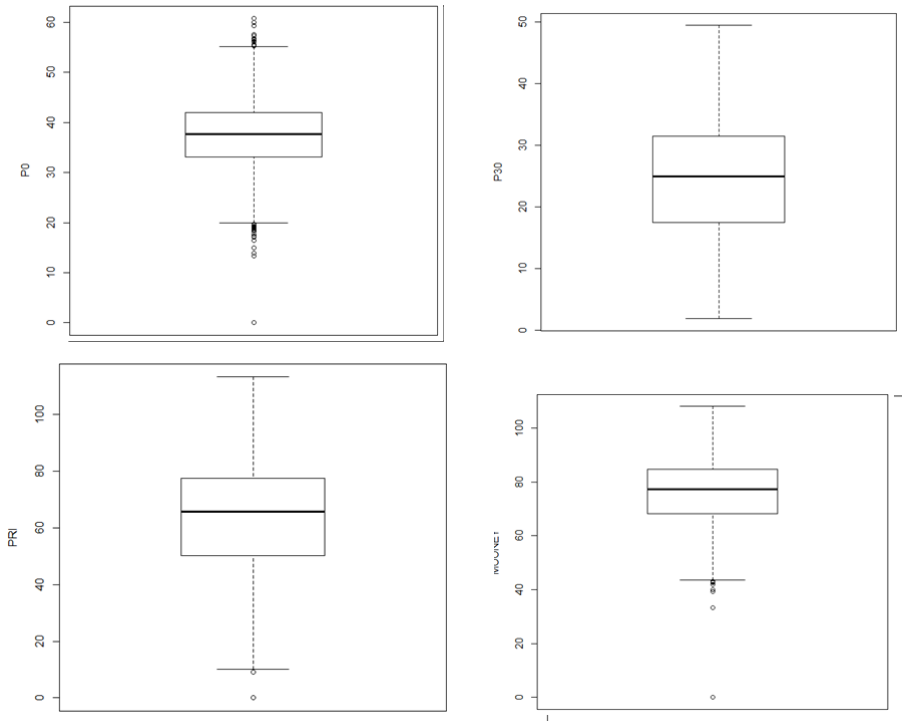


FIGURE *D.2.:* Boxplots des variables P_0, P_{30}, PRI et $MOONEY$

FIGURE *D.3.:* Répartitions de la variable clone dans la base de données parcellaires

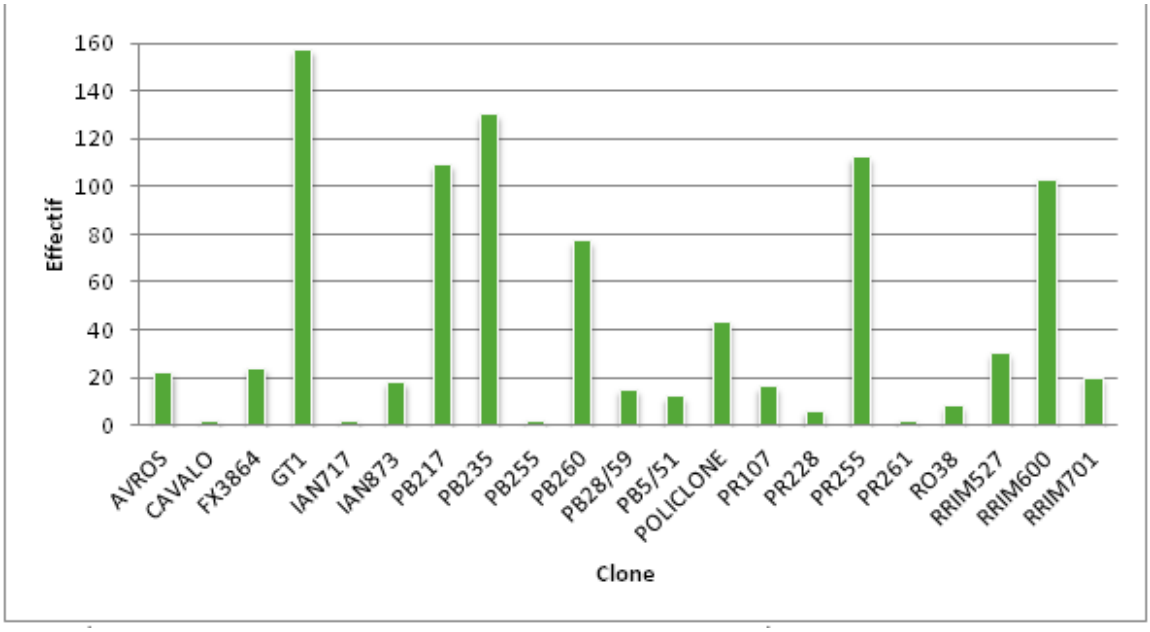


FIGURE *D.4.:* Distribution de la variable parcellaire *Panel*

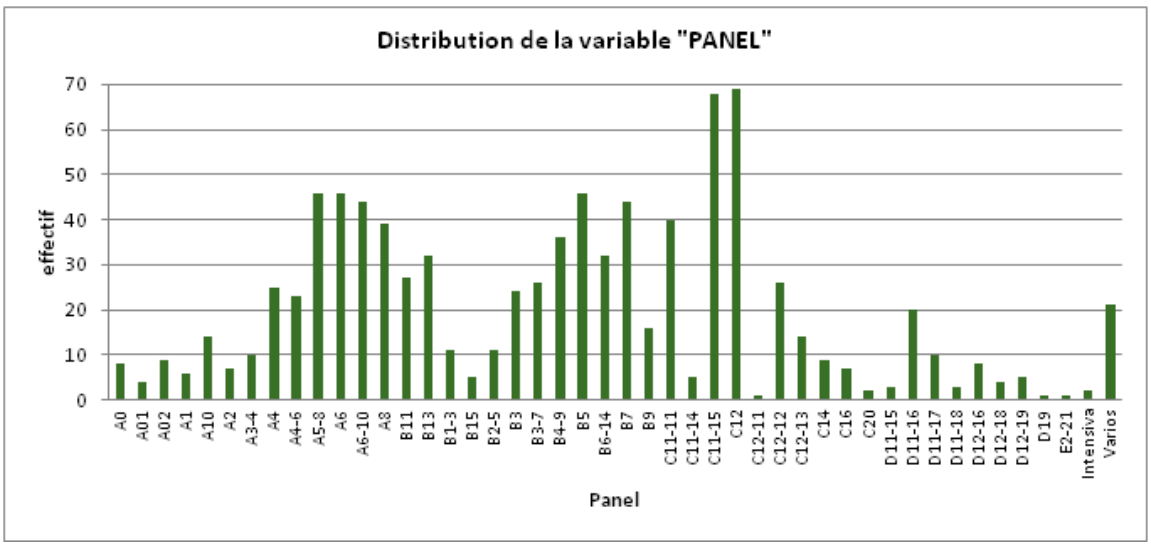


FIGURE *D.5.:* Distribution de la variable système de saignée dans la base de données parcellaires

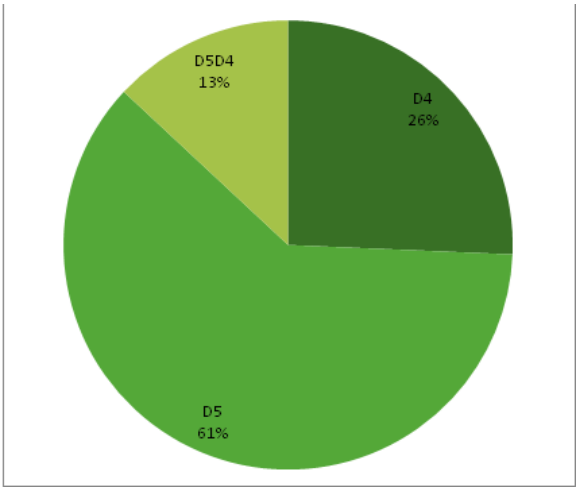
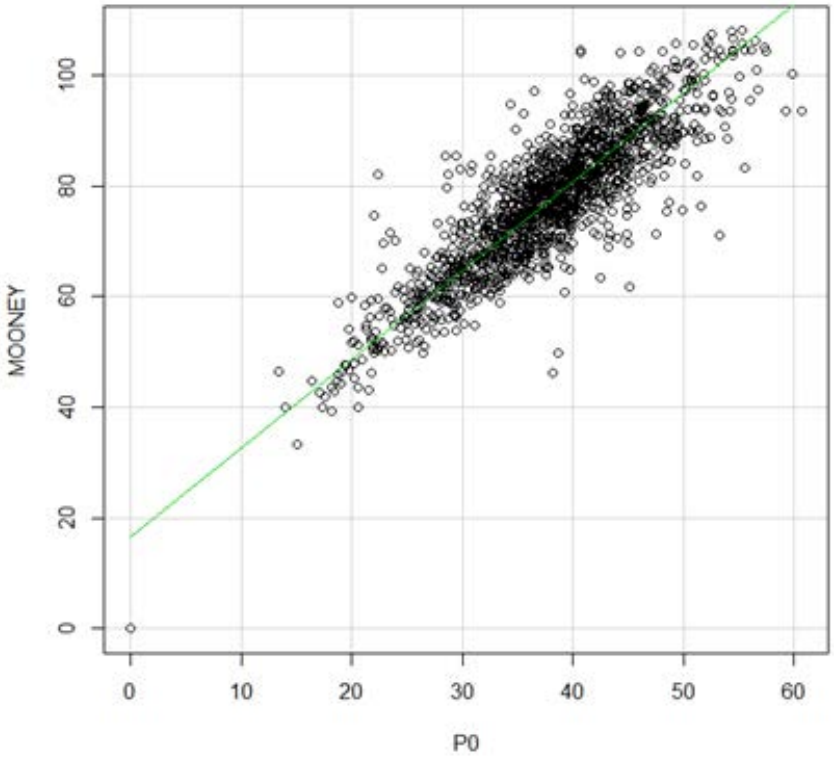


FIGURE *D.6.:* Régression $MOONEY / P_0$



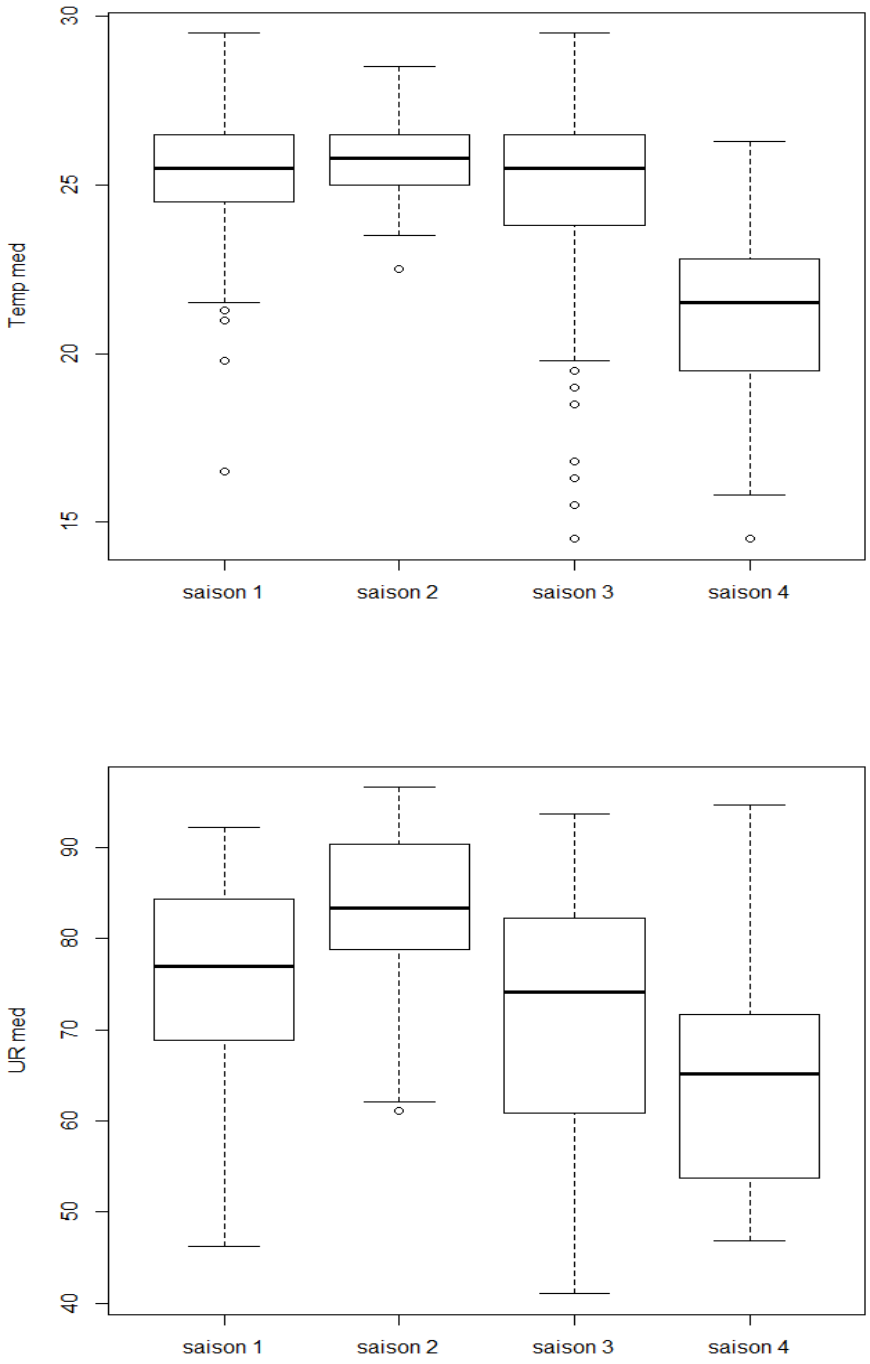
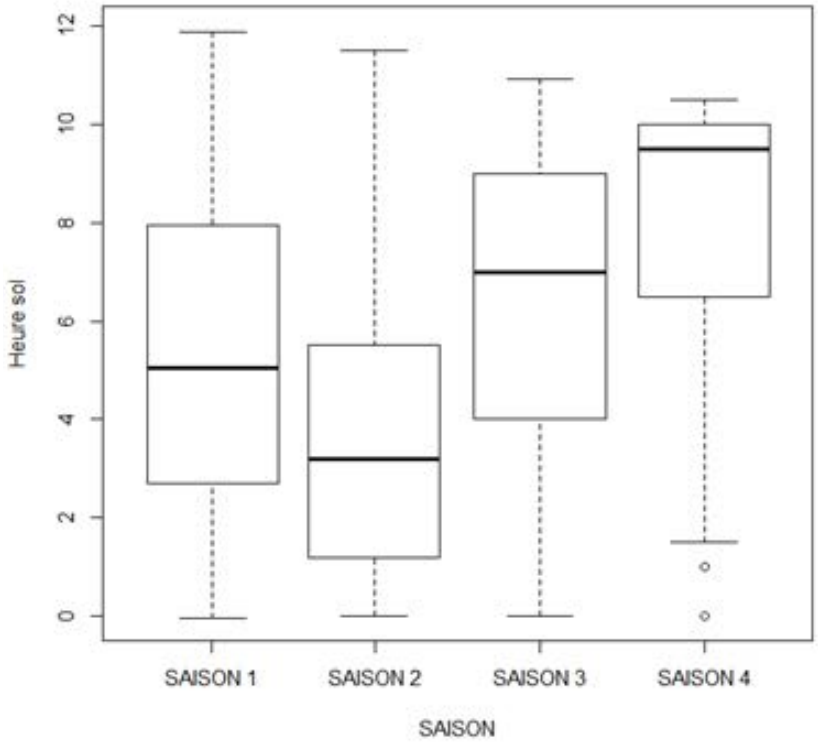


FIGURE *D.7.:* Evolution de la température moyenne et de l'humidité relative moyenne en fonction de la saison

FIGURE *D.8.:* Evolution de l'ensoleillement moyen fonction de la saison



D.2. Etude statistique prédictive sans incertitude sur les données :

Différents découpages de classes (qualité du caoutchouc) sont ici considérés. Pour chaque configuration de classes, 100 tests sont effectués (la structure des arbres de décision étant assez instable, un nombre de tests suffisamment grand était nécessaire). Pour chaque test, l'échantillon initial est aléatoirement divisé en échantillon d'apprentissage (2/3 de l'échantillon initial) et échantillon de test (1/3 de l'échantillon initial), un arbre est construit avec la fonction rpart du logiciel R (paramétrée selon le modèle CART, i.e. l'arbre est entièrement déroulé, pas de pré-élagage donc, puis est élagué selon le critère de coût-complexité dont le paramètre est fixé par validation croisée à 10 couches) à partir de l'échantillon d'apprentissage et est évalué sur l'échantillon de test. Les arbres représentés sur les figures sont quelque fois ré-élagués de manière à être plus facilement lisibles.

Les taux d'erreur seront calculés comme la proportion de mauvaises classifications des exemples de l'échantillon test. On retiendra les taux naïfs d'erreur de prédiction moyens (prédisant systématiquement la classe majoritaire de l'échantillon d'apprentissage), les taux d'erreur moyens de prédiction obtenus par les arbres, ainsi que les noms des variables utilisées pour les coupures des arbres ainsi que leurs scores de profondeur (quand ces derniers sont supérieurs ou égaux à 1). Ces scores correspondent à l'inverse des profondeurs normalisées.

Exemple 7. *Avec la configuration « 5 classes équiprobables, classe = P_{30}» on obtient l'arbre représenté en Figure D.9.*

On obtient alors les scores présentés en Table D.3.

variable	profondeur	inverse profondeur	SCORE
saison	1	1.00	0.44
age.ouv	2	0.50	0.22
Temp.MIN.fabric	3	0.33	0.15
panel.C12.13	4	0.25	0.11
Precipit.matur	5	0.20	0.09

TABLE *D.3.:* Calcul des scores des variables apparaissant sur un arbre de décision

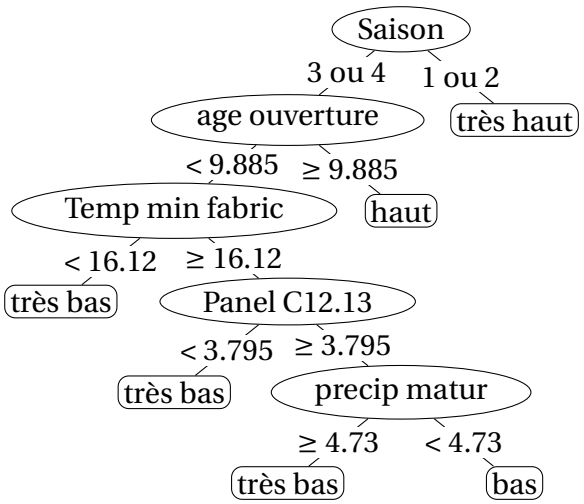


FIGURE *D.9.:* P_{30} découpé en 5 classes équiprobables

Pour les variables apparaissant plusieurs fois sur un même arbre le score maximal est retenu. Les coordonnées géographiques de chaque benne ont été rajoutées à la base de données à partir de la carte de la plantation. Une variable binaire « cirad » a été rajoutée, elle vaut systématiquement 0 en 2003/2004, 1 en 2006/2007. Les tests sont réalisés sur une base de données notée $BDD_{simplifiée}$ contenant les attributs « d'intérêt » suivants :

- La variables « cirad »
- Saison
- PESO PEM (poids de la benne)
- Les coordonnées géographiques (X et Y)
- Proportions de chaque type de clone
- Proportion de chaque panel
- Proportion de chaque système de saignée
- Les variables climatologiques sont calculées pour 3 périodes : la période de fabrication du latex dans l'arbre, la période de saignée, puis la période de maturation dans les tasses.

Ces 3 périodes seront estimées de la façon suivante (J correspond au jour de l'arrivée d'une benne à l'usine) :

- fabrication du latex : $[J-13; J-6]$
- saignée : $\{J-5\}$
- maturation dans les tasses : $[J-4; J]$

nom classe	P_0	PRI	P_{30}	$MOONEY$
très mauvaise	[13.4 ; 31.1[	[9.2 ; 45.6[	[1.87 ; 14.7[	[33.4 ; 65.2[
mauvaise	[31.1 ; 35.7[	[45.6 ; 60[	[14.7 ; 21.6[	[65.2 ; 73.08[
moyenne	[35.7 ; 39.2[	[60 ; 70.4[	[21.6 ; 27.4[	[73.08 ; 79.6[
bonne	[39.2 ; 43.2[	[70.4 ; 79.1[	[27.4 ; 32.9[	[79.6 ; 86.2[
très bonne	[43.2 ; 60.7[	[79.1 ; 113.3[	[32.9 ; 49.5[	[86.2 ; 108[

TABLE *D.4.:* Découpages des classes équiprobables

nom classe	PRI	P_{30}
mauvaise	[9.2 ; 50[	[1.87 ; 18.5[
moyenne	[50 ; 65[	[18.5 ; 24.9[
bonne	[65 ; 113.3[	[24.9 ; 49.5[

TABLE *D.5.:* Découpages des classes ISO

Ces variables climatologiques sont :
- Température minimale, médiane et maximale
- Humidité minimale, médiane et maximale
- Heures de soleil
- Précipitations

Les découpages de classes sont faits de manière à obtenir 5 classes de qualité (« très mauvaise », « mauvaise », « moyenne », « bonne », « très bonne ») avec des effectifs à peu près équivalents pour chaque classe (→ « classe équiprobables »). Un test est aussi réalisé en choisissant comme classe le PRI découpé selon la norme ISO en 3 classes ($PRI < 50$, $PRI \in [50;65]$ et $PRI > 65$) , puis un autre test reprend ce découpage du PRI pour le P_{30} ($P_{30} = \frac{P_0 * PRI}{100}$) en utilisant le P_0 moyen (37.07) calculé sur toute la base de données ($P_{30} < P_0$ moyen$* \frac{50}{100}$; $P_{30} \in [P_0$ moyen$* \frac{50}{100}$; P_0 moyen$* \frac{65}{100}]$; $P_{30} > P_0$ moyen$* \frac{65}{100}$). L'ensemble des découpages de classe obtenus est présenté en Tables D.4 et D.5.

Les tests sont tout d'abord réalisés pour ces différents découpages avec échantillon d'apprentissage et échantillon de test, puis certaine variables sont figées (saison, cirad et clone) pour certaines configurations toujours avec échantillon d'apprentissage et échantillon de test, puis enfin les tests avec découpages en 5 classes équiprobables sont réalisés en utilisant toute la base de données $BDD_{simplifiée}$ comme échantillon d'apprentissage et en calculant les taux d'erreur sur ces mêmes

échantillons d'apprentissage (de façon informative juste).

D.2.1. Résultats pour 5 Classes avec échantillon d'apprentissage et échantillon de test :

	5 classes équiprobables				3 classes ISO	
	P_0	PRI	P_{30}	$MOONEY$	PRI	P_{30}
taux d'erreur naif	0.80	0.81	0.81	0.81	0.50	0.49
taux erreur	0.65	0.61	0.60	0.66	0.38	0.36

TABLE *D.6.:* Taux d'erreur obtenus sur 100 simulations de découpages en 5 classes équiprobables

Seules les variables ayant un score supérieur à 3% sont ici présentées. De manière à voir comment ces variables impactent la qualité du caoutchouc voici six exemples d'arbres de décision obtenus lors de ces simulations (nous partons du principe qu'en général, une variable agit toujours *dans le même sens* sur la classe).

P_0		PRI		P_{30}		$MOONEY$	
variables	scores	variables	scores	variables	scores	variables	scores
Saison	0.41	Temp.MÍN.fabric	0.1	Saison	0.13	Arbres.saignes	0.2
Temp.MÍN.matur	0.1	Saison	0.08	clone.PR255	0.06	saison	0.2
syst.saignee.D4	0.09	as.ha	0.06	clone.GT1	0.05	syst.saignee.D4	0.12
Temp.MÍN.fabric	0.06	X	0.04	Temp.MÍN.fabric	0.04	CIRAD	0.09
age.ouv	0.05	clone.PR255	0.04	clone.RRIM527	0.04	Temp.MÍN.matur	0.09
X	0.04	age.ouv	0.04	Temp.MÍN.matur	0.03	panel.A6	0.07
clone.PR255	0.04	panel.C12.13	0.03	X	0.03	Temp.MÉD.fabric	0.04
Y	0.03	Y	0.03			Temp.MÍN.fabric	0.03
		Temp.MÉD.fabric	0.03				

TABLE *D.7.:* Variables explicatives les plus pertinentes avec scores pour des découpages en 5 classes équiprobables

PRI		P_{30}	
variables	scores	variables	scores
Temp.MÍN.fabric	0.14	saison	0.2
saison	0.14	clone.RRIM527	0.08
as.ha	0.1	Temp.MÍN.fabric	0.07
clone.RRIM527	0.07	clone.GT1	0.07
clone.GT1	0.05	age.ouv	0.05
UR.MÁXmatur	0.04	clone.PR255	0.04
Y	0.03	Temp.MÉD.fabric	0.03
clone.PR255	0.03	Temp.MÍN.matur	0.03
X	0.03	UR.MÍN.matur	0.02
syst.saignee.D4	0.03		

TABLE *D.8.:* Variables explicatives les plus pertinentes avec scores pour des découpages en 3 classes ISO

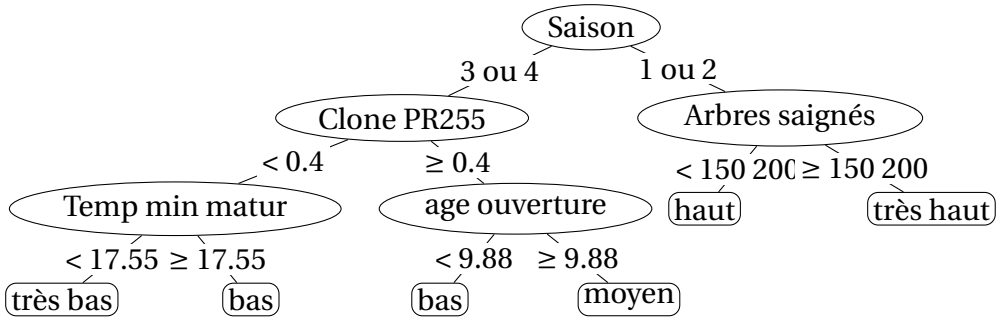


FIGURE *D.10.:* P_0 découpé en 5 classes équiprobables

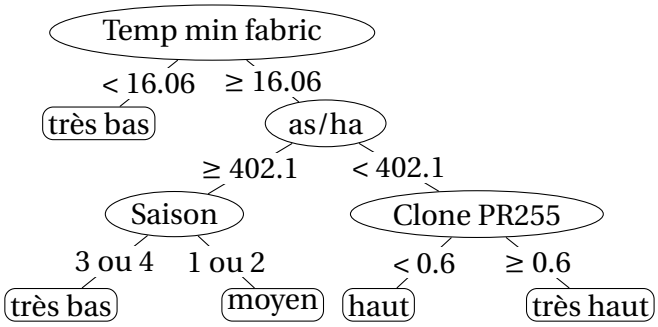


FIGURE *D.11.:* PRI découpé en 5 classes équiprobables

Premières observations :

– Le P_0 semble très dépendant de la variable « saison », il sera donc étudié séparément sur chaque saison. Il apparait néanmoins que les moins bons caoutchoucs sont obtenus en saisons 3 et 4, i.e. de Mars à Aout (Automne-Hiver).

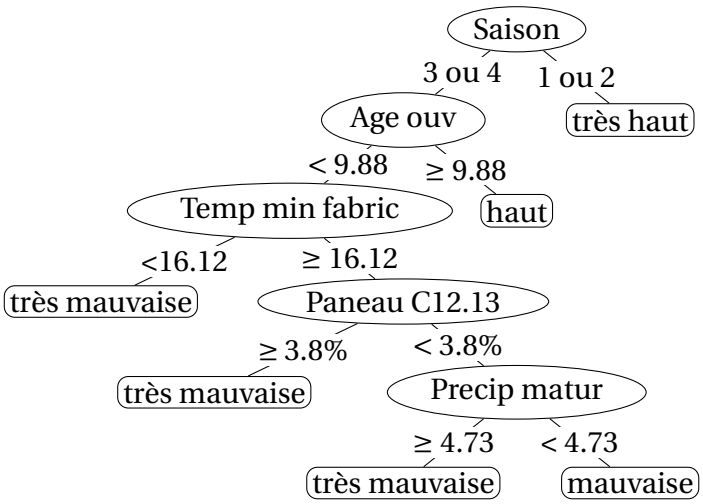


FIGURE *D.12.:* P_{30} découpé en 5 classes équiprobables

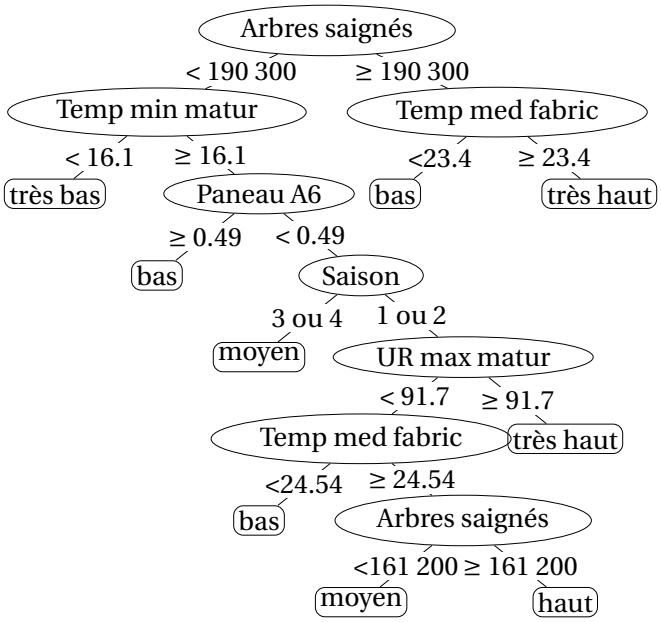


FIGURE *D.13.:* *MOONEY* découpé en 5 classes équiprobables

– Le *MOONEY*, quant à lui, tout en étant assez expliqué par la variable « saison », semble tout autant expliqué par la variable « Arbres saignés », un grand nombre d'arbres saignés (plus de 190000 par benne) étant bénéfique pour la qualité du caoutchouc. Le système de saignée D4 semble aussi avoir des effets positifs sur la qualité du caoutchouc obtenu qui semble meilleure en 2006/2007 qu'en 2003/2004. Il sera donc étudié séparément sur chacune de ces périodes (la forte corrélation entre le P_0 et le *MOONEY* rende superflue l'étude de ce dernier sur chacune des saisons).

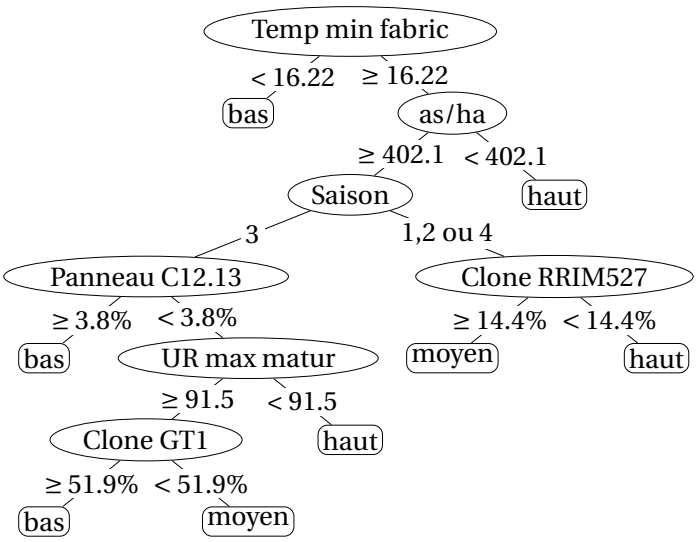


FIGURE *D.14.:* PRI découpé en 3 classes ISO

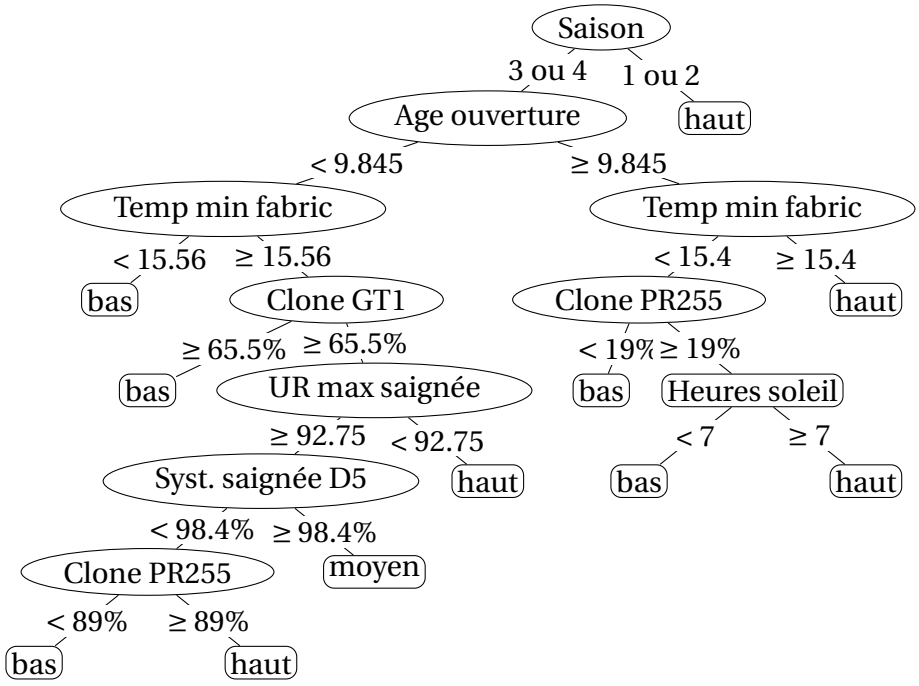


FIGURE *D.15.:* P_{30} découpé en 3 classes ISO

D.2.2. Résultats par saison pour 5 Classes équiprobables de P_0 avec échantillon d'apprentissage et échantillon de test :

Toujours de manière à voir dans quels sens jouent ces variables voici des exemples d'arbres de décisions obtenus lors de ces simulations :

	saison 1	saison 2	saison 3	saison 4
taux d'erreur naif	0.83	0.83	0.82	0.82
taux erreur	0.73	0.69	0.69	0.58

TABLE *D.9.:* Taux d'erreur par saison obtenus sur 100 simulations de découpages du P_0 en 5 classes équiprobables

saison 1		saison 2		saison 3		saison 4	
variables	scores	variables	scores	variables	scores	variables	scores
clone.PR255	0.11	Temp.MÍN.matur	0.2	Ha	0.16	clone.PR255	0.27
age.ouv	0.1	clone.GT1	0.09	syst.saignee.D4	0.1	syst.saignee.D4D5	0.14
PESO.PEM	0.05	CIRAD	0.08	age.ouv	0.08	panel.B5	0.07
clone.GT1	0.05	UR.MÉD.fabric	0.06	X	0.07	Y	0.06
UR.MÁX.fabric	0.04	clone.PB235	0.04	Temp.MÍN.matur	0.05	PESO.PEM	0.06
Arbres.saignes	0.04	UR.MÉD.matur	0.04	Temp.MÍN.fabric	0.05	clone.IAN873	0.05
clone.PB217	0.04	UR.MÁX.fabric	0.04	clone.RRIM600	0.03	X	0.03
syst.saignee.D5	0.04	Precipitação.fabric	0.03	panel.C11.11	0.03	UR.MÉD.fabric	0.03
panel.C12	0.04	Temp.MÁX.fabric	0.03	Precipitação.fabric	0.03		
Temp.MÉD.fabric	0.03	Arbres.saignes	0.03	Temp.MÉD.matur	0.03		
as.ha	0.03	Temp.MÉD.fabric	0.03				
		Temp.MÍN.fabric	0.03				
		clone.PR255	0.03				

TABLE *D.10.:* Variables explicatives les plus pertinentes avec scores par saison pour des découpages du P_0 en 5 classes équiprobables

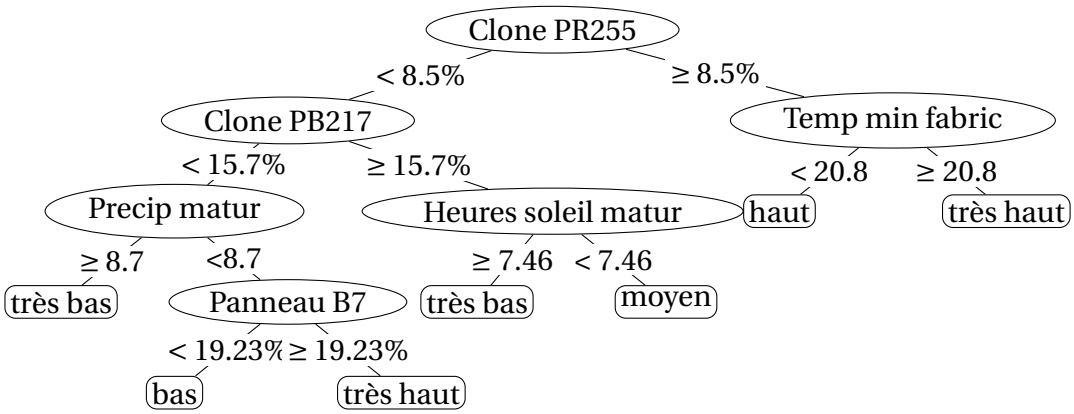


FIGURE *D.16.:* P_0 découpé en 5 classes équiprobables : saison 1

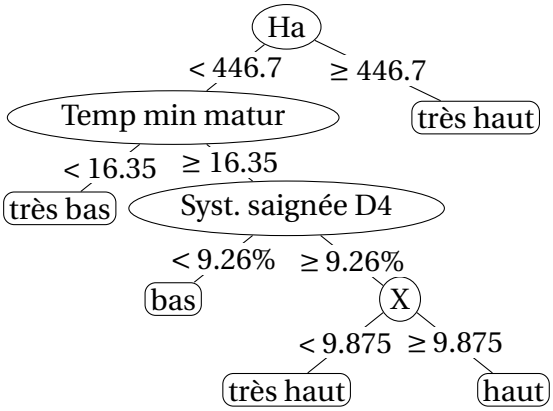


FIGURE *D.17.:* P_0 découpé en 5 classes équiprobables : saison 3

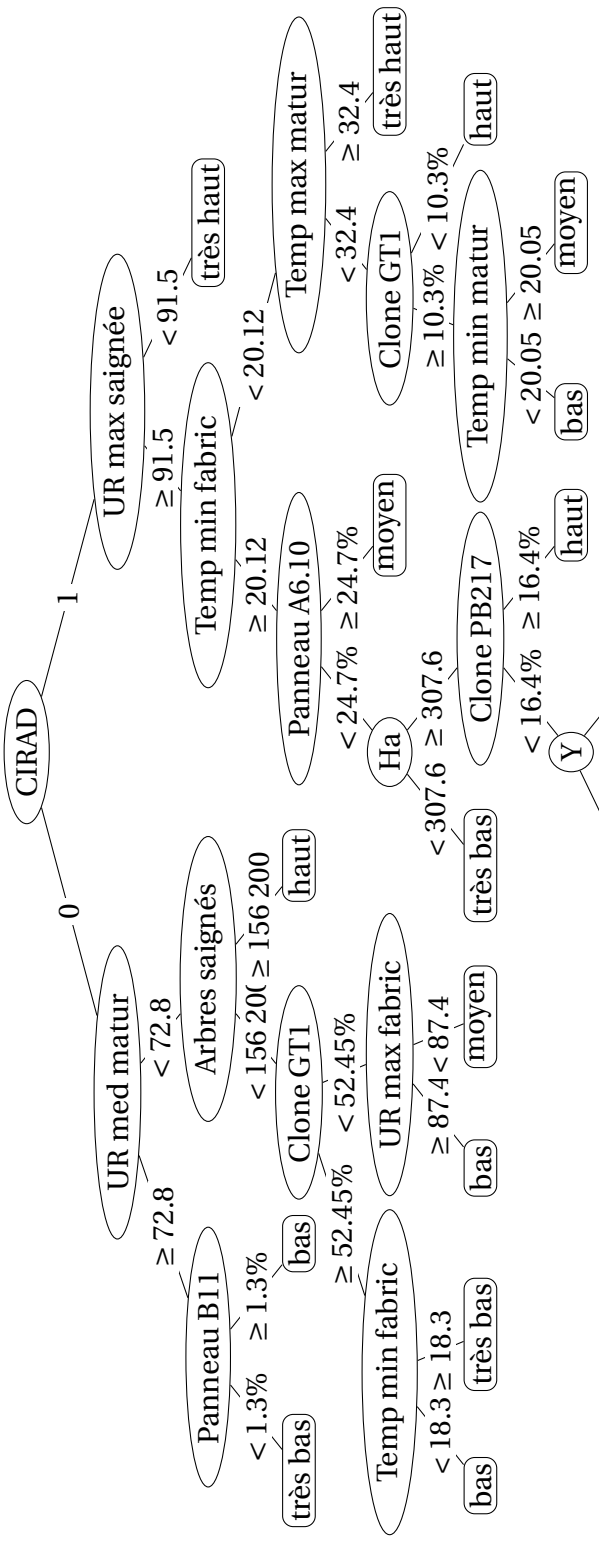


FIGURE *D.18.:* P_0 découpé en 5 classes équiprobables : saison 2

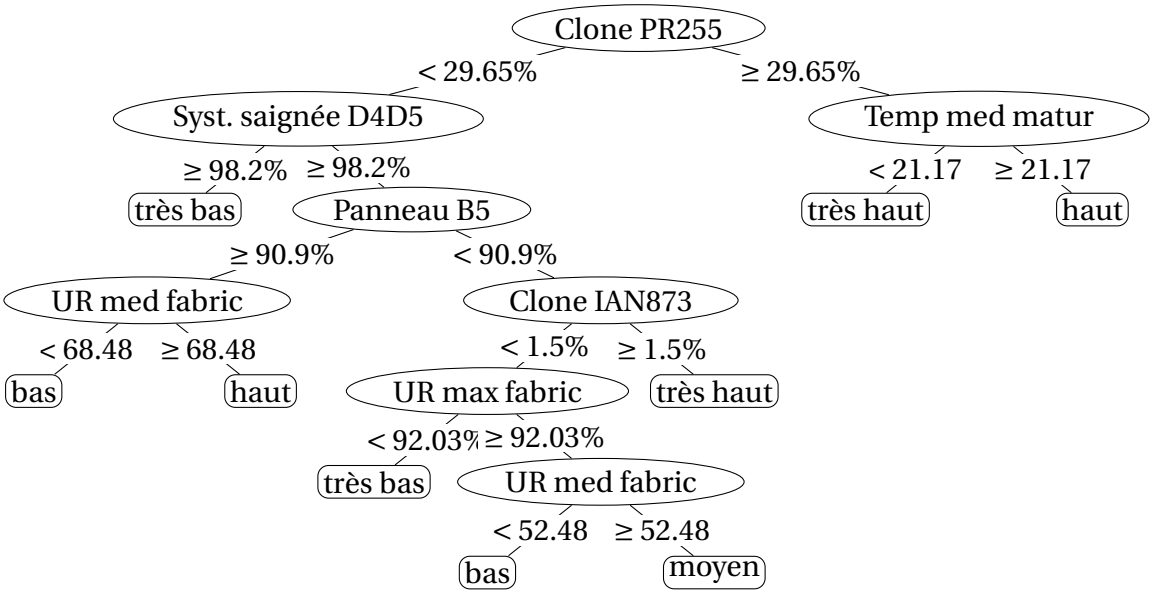


FIGURE *D.19.:* P_0 découpé en 5 classes équiprobables : saison 4

Secondes observations :

- La saison 4 étant la mieux expliquée, elle sera donc ré-étudiée sur la benne pure de manière à figer la variable clone.
- Le clone PR255 (de même que le paneau B7) semble être un bon clone, il a un effet notable en saison 1 et 4. Le clone GT1 lui, semble mauvais.
- La température hautes pendant les périodes de fabrication et de maturation du latex ont un effet positif sur la qualité du latex.
- Le système de saignée D4 semble être profitable à la qualité du caoutchouc obtenu (contrairement au D4D5).
- En période des pluies (saison 2), trop d'humidité peut avoir un effet néfaste sur la qualité du caoutchouc.
- Les grandes parcelles (situées sur la partie ouest de la plantation : petite valeur de X, grandes valeurs de Ha), ont tendance à donner du bon caoutchouc. Ceci est visible surtout en saison 3.

D.2.3. Résultats pour la saison 4 et sur la benne pure (l'unique benne ne contenant qu'un seul clone) pour 5 Classes équiprobables de P_0 avec échantillon d'apprentissage et échantillon de test :

taux d'erreur naif	0.84
taux erreur	0.71

FIGURE *D.20.:* Taux d'erreur obtenus pour la saison 4 et sur la benne pûre sur 100 simulations de découpages du P_0 en 5 classes équiprobables

variables	scores
Temp.MÉD.matur	0.37
Horas.de.Sol.fabric	0.33
Temp.MÁX.matur	0.09
Temp.MÍN.fabric	0.09
UR.MÉD.matur	0.03

TABLE *D.11.:* Variables explicatives les plus pertinentes avec scores pour la saison 4 et sur la benne pûre pour des découpages du P_0 en 5 classes équiprobables

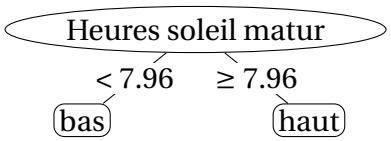


FIGURE *D.21.:* P_0 découpé en 5 classes équiprobables : saison 4 et benne pûre

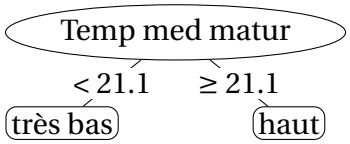


FIGURE *D.22.:* P_0 découpé en 5 classes équiprobables : saison 4 et benne pûre

Observation :

En saison 4, le soleil et les températures élevées sont bénéfiques à la qualité du caoutchouc (hiver).

Résultats par période d'étude pour 5 Classes équiprobables de *MOONEY* avec échantillon d'apprentissage et échantillon de test :

	2003/2004	2006/2007
taux d'erreur naif	0.82	0.82
taux erreur	0.67	0.61

TABLE *D.12.:* Taux d'erreur par période d'étude obtenus sur 100 simulations de découpages du *MOONEY* en 5 classes équiprobables

2003/2004		2006/2007	
variables	scores	variables	scores
syst.saignee.D4	0.28	saison	0.29
clone.PR255	0.12	Temp.MÍN.fabric	0.14
Temp.MÍN.matur	0.11	Arbres.saignes	0.11
saison	0.1	panel.B3.7	0.1
Y	0.08	Temp.MÍN.matur	0.07
clone.GT1	0.06	Temp.MÉD.matur	0.04
clone.PB217	0.05	panel.A6.10	0.03
		Horas.de.Sol.fabric	0.03
		Y	0.03

TABLE *D.13.:* Variables explicatives les plus pertinentes avec scores par période d'étude pour des découpages du *MOONEY* en 5 classes équiprobables

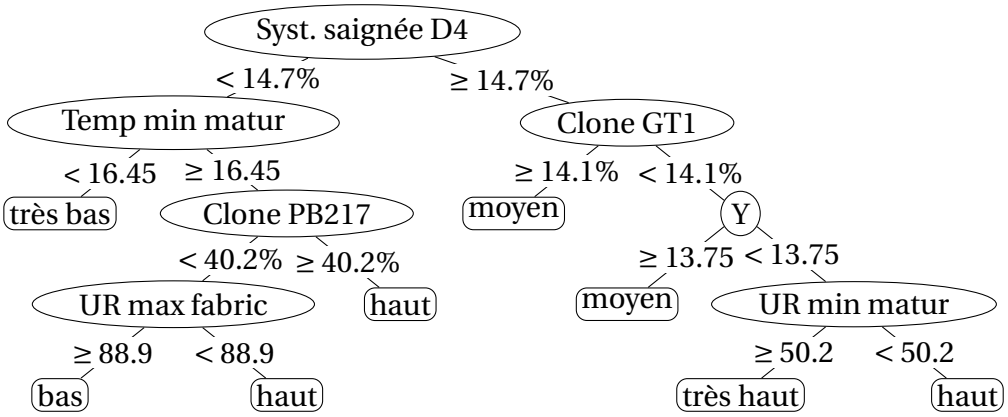


FIGURE *D.23.:* *MOONEY* découpé en 5 classes équiprobables : période 2003/2004

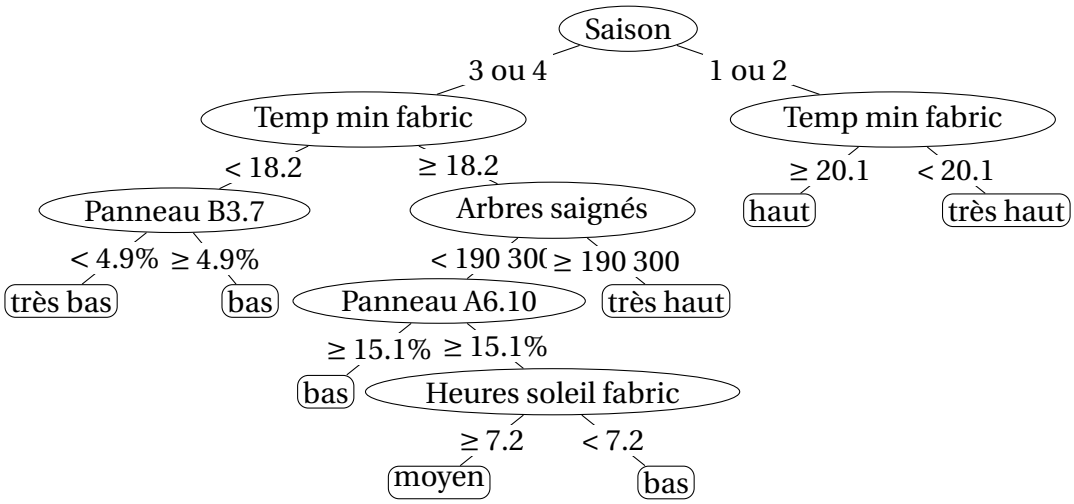


FIGURE *D.24.:* *MOONEY* découpé en 5 classes équiprobables : période 2006/2007

Observations :

- Le système de saignée D4 a eu une bonne influence sur le MOONEY en période 2003/2004 (peut-être son utilisation était-emlle généralisée en 2006/2007 par recommandation du CIRAD)
- Les températures hautes seront encore une fois bénéfiques pour la qualité du caoutchouc
- Les parcelles contenant un grand nombre d'arbres saignés par année agricole (plus de 190000) donneront du bon caoutchouc (particulièrement pendant les saison 3 et 4).

D.2.4. Résultats pour l'ensemble des indices de qualité découpés en 5 classes équiprobables avec apprentissage sur toute la base de données $BDD_{simplifiée}$:

	P_0	PRI	P_{30}	$MOONEY$
taux d'erreur naif	0.79	0.80	0.80	0.82
taux erreur	0.64	0.64	0.64	0.70

TABLE *D.14.:* Taux d'erreur obtenus par des arbres de décision construit sur toute la base de données $BDD_{simplifiée}$ avec découpages en 5 classes équiprobables

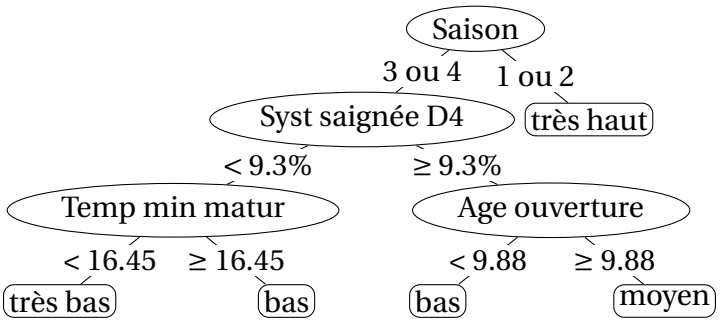


FIGURE *D.25.:* P_0 découpé en 5 classes équiprobables : apprentissage sur toute la base de données $BDD_{simplifiée}$

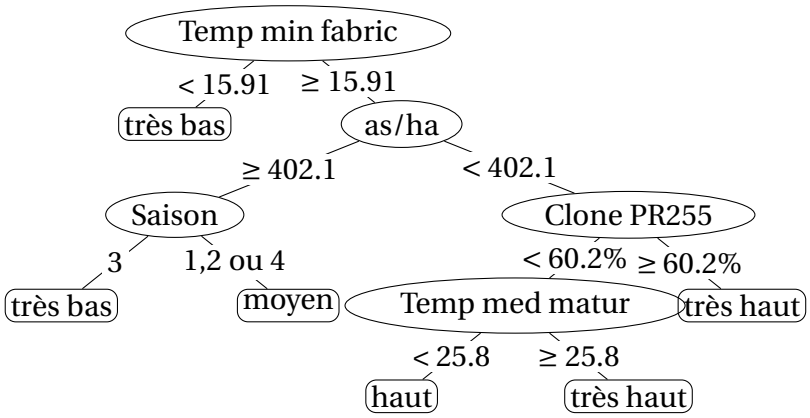


FIGURE *D.26.:* PRI découpé en 5 classes équiprobables : apprentissage sur toute la base de données $BDD_{simplifiée}$

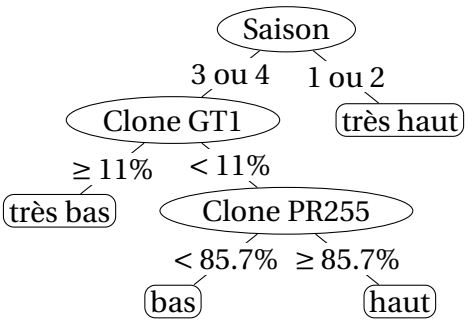


FIGURE *D.27.:* P_{30} découpé en 5 classes équiprobables : apprentissage sur toute la base de données $BDD_{simplifiée}$

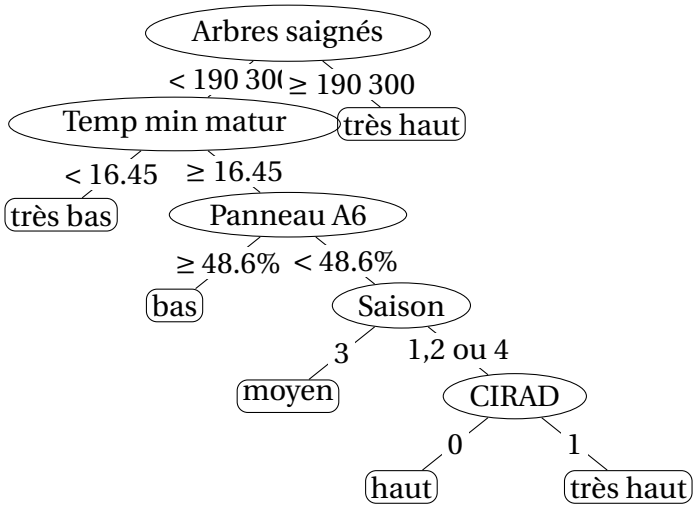


FIGURE *D.28.:* $MOONEY$ découpé en 5 classes équiprobables : apprentissage sur toute la base de données $BDD_{simplifiée}$

BILAN :

La qualité du caoutchouc est ici étudiée via 3 indices de qualité : le P_0, le PRI (ces deux indices sont aussi résumés dans le P_{30}) et le $MOONEY$. Il est intéressant dans un premier temps de constater que ces différents indices n'ont pas évolué dans le même sens entre la première période d'étude (2003/2004) et la deuxième (2006/2007). En effet entre ces deux périodes, le P_0 et le P30 se sont améliorés alors que le PRI s'est dégradé.

Les variables les plus récurrentes dans l'explication de la qualité du caoutchouc sont la « saison », les températures minimales pendant les périodes de fabrication et

de maturation du latex, le rendement (nombre d'arbres saignés par année agricole et nombre d'arbres saignés par hectares) et le clone.

De façon globale, le caoutchouc produit pendant les saisons 1 et 2 (Septembre à Février), donc pendant les périodes chaudes, et humides est de meilleure qualité que celui produit pendant les saisons 3 et 4 (Mars à Juin), donc pendant des périodes plutôt sèches, fraiches et ensoleillées. De plus, les différentes études montrent que pendant les périodes de fabrication et de maturation du latex, la température et l'humidité doivent être suffisamment élevées alors que l'ensoleillement ne doit pas être excessif. La météo au moment de la saignée, estimé à 5 jours avant l'arrivée à l'usine des bennes, ne semble pas avoir un gros impact sur la qualité du caoutchouc obtenu.

Les différents clones apparaissent comme ayant des impacts différents sur la qualité du caoutchouc. Le clone le plus bénéfique est le PR255. Les clones et PB235, B7 et PB217 apparaissent aussi comme de bons clones. Le clone GT1 par contre, semble clairement mauvais.

Le système de saignée semble surtout impacter le P_0 et le $MOONEY$, le D4 apparait comme bénéfique alors que le D4D5 semble néfaste.

Du point de vue de la plantation, il est intéressant de se rendre compte que les grande parcelles, donc avec un grand nombre d'arbres mais suffisamment éloignés (densités pas trop grandes) tendent à donner du bon caoutchouc. Les arbres ouverts trop jeunes (moins de 10 ans) donneront du mauvais caoutchouc.

Table des figures

Liste des tableaux

Bibliographie

[1] Nicolas Sutton-Charani, Sébastien Destercke, and Thierry Denoeux. Application of e^2m decision trees to rubber quality prediction. *IPMU*, 2014.

[2] F. Aguirre, M. Sallak, and W. Schon. Construction of belief functions from statistical data about reliability under epistemic uncertainty. *IEEE Transactions on Reliability*, 62(3) :555–568, 2013.

[3] N. Ben Abdallah, N. Mouhous-Voyneau, and T. Denoeux. Combining statistical and expert evidence using belief functions : Application to centennial sea level estimation taking into account climate change. *International Journal of Approximate Reasoning*, 55(1) :341–354, 2014.

[4] E Périnel. Construire un arbre de discrimination binaire à partir de données imprécises. *Revue de statistique appliquée*, 47 :5–30, 1999.

[5] P. Walley. *Statistical reasoning with imprecise probabilities*. Chapman and Hall, 1991.

[6] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic ? Does it matter ? *Structural Safety*, 31 :105–112, 2009.

[7] M. Dumas. Discussion sur la définition du mot "statistique". *Journal de la société statistique de Paris*, tome 97 :528–535, 1956.

[8] T.H. Wonnacott, R.J. Wonnacott, and P. Cohendet. *Statistique : économie, gestion, sciences, médecine : (avec exercices d'application)*. Economica, 1991.

[9] I. Bloch. Incertitude, imprécision et additivité en fusion de données : point de vue historique. *Traitement du Signal — Volume 13 - n°4*, 1996.

[10] E. Thionville. *De la théorie des lieux communs dans les "Topiques" d'Aristote et des principales modifications qu'elle a subies jusqu'à nos jours*. A. Durand, 1855.

[11] O. Ore. Cardano, the gambling scholar. *Princeton University Press*, 1953.

[12] B. Pascal. *Les lettres de Blaise Pascal : accompagnées de lettres de ses correspondants*. publiées par M. Beaufreton, 1922.

[13] J. Bernoulli. *Ars conjectandi :*. Landmarks of science. Impensis Thurnisiorum, fratrum, 1713.

[14] D. Diderot. *Encyclopédie ou dictionnaire raisonné des sciences des arts et des métiers*. Number vol. 20 in Encyclopédie ou dictionnaire raisonné des sciences des arts et des métiers. Sociétés Typographiques, 1780.

[15] T. Bayes, R. Price, and J. Canton. *An Essay Towards Solving a Problem in the Doctrine of Chances*. C. Davis, Printer to the Royal Society of London, 1763.

[16] P.S. de Laplace. *Théorie analytique des probabilités*. Ve. Courcier, 1814.

[17] É. Borel. *Leçons sur la théorie des fonctions*. Collection de Monographies sur la Théorie des Fonctions. Gauthier-Villars et Fils, 1898.

[18] H.L. Lebesgue. *Leçons sur les séries trigonométriques professées au Collège de France*. Collection de monographies sur la théorie des fonctions. Gauthier-Villars, 1906.

[19] A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933.

[20] G. Shafer. *Non-Additive Probabilities in the Work of Bernoulli and Lambert*, volume 219 of *Studies in Fuzziness and Soft Computing*. Springer Berlin Heidelberg, 2008.

[21] T. Martin. *Probabilités et critique philosophique selon Cournot*. Mathesis (Paris, France). J. Vrin, 1996.

[22] T. Martin. Certitude et probabilité selon Buffon. *revue de l'enseignement philosophique*, 49(3) :5–18, 1999.

[23] G. Frege, C. Besson, and J. Barnes. *Idéographie*. Bibliothèque des textes philosophiques. Librairie philosophique J. Vrin, 1999.

[24] J.M. Keynes. *A Treatise on Probability*. Dover Books on Mathematics Series, (2004 edition). DOVER PUBN Incorporated, 1921.

[25] G. Boole. *An Investigation of the Laws of Thought : On which are Founded the Mathematical Theories of Logic and Probabilities*. George Boole's collected logical works. Walton and Maberly, 1854.

[26] Frank P. Ramsey. Truth and probability. In R. B. Braithwaite, editor, *The Foundations of Mathematics and other Logical Essays*, chapter 7, pages 156–198. McMaster University Archive for the History of Economic Thought, 1926.

[27] Bruno de Finetti. La prévision : ses lois logiques, ses sources subjectives. *Annales de l'institut Henri Poincaré*, 7(1) :1–68, 1937.

[28] J-P. Gayant. L'apport des modèles non-additifs en théorie de la décision dans le risque et l'incertain. *Revue Française d'Économie*, 13(1) :199–227, 1998.

[29] L.J. Savage. *The Foundations of Statistics*. Dover Books on Mathematics Series. DOVER PUBN Incorporated, 1972.

[30] D.V. Lindley. *introduction to probability and statistics from bayesian viewpoint. part 2 inference*. At the University Press, 1965.

[31] Didier Dubois and Henri Prade. Représentations formelles de l'incertain et de l'imprécis. In Denis Bouyssou, Didier Dubois, Marc Pirlot, and Henri Prade, editors, *Concepts et méthodes pour l'aide à la décision - outils de modélisation*, volume 1 of *Traité IC2*, chapter 3, pages 111–171. Lavoisier, http ://www.editions-hermes.fr/, 2005.

[32] R. A. Fisher. *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh, second edition, 1956.

[33] B. De Finetti and L.J. Savage. *sul modo di scegliere la probabilità iniziali*. 1962.

[34] D.V. Lindley. *Making decisions*. Wiley, 1971.

[35] L.A. Zadeh. Fuzzy sets as a basis for theory of possibility. *Fuzzy Sets and Systems*, 1 :3–28, 1978.

[36] D. Dubois and H. Prade. *Fuzzy Sets and Systems : Theory and Applications* , volume V.144, 393 p. Academic Press (APNet), http ://www.apnet.com/, 1980. ks Mathematics in Science and Engineering Series.

[37] L.A. Zadeh. Fuzzy sets. *Information Control*, 8 :338–353, 1965.

[38] R. A. Fisher. Inverse Probability. *Proceedings of the Cambridge Philosophical Society*, 26 :528–535, 1930.

[39] D. V. Lindley. Fiducial Distributions and Bayes' Theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(1) :102–107, 1958.

[40] A. P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38 :325–339, 1967.

[41] Thierry Denœux. Statistical inference from ill-known data using belief functions. In Van-Nam Huynh, Vladik Kreinovich, Songsak Sriboonchitta, and Komsan Suriya, editors, *Uncertainty Analysis in Econometrics with Applications*, volume 200 of *Advances in Intelligent Systems and Computing*, pages 33–48. Springer Berlin Heidelberg, 2013.

[42] Glenn Shafer. *A mathematical theory of evidence*. Princeton-London : Princeton University Press. XIII, 297 p. hbk : $ 22.00 ; pbk : $ 11.25 , 1976.

[43] Sébastien Destercke and Thomas Burger. Toward an axiomatic definition of conflict between belief functions. *IEEE T. Cybernetics*, 43(2) :585–596, 2013.

[44] Weiru Liu. Measuring conflict between possibilistic uncertain information through belief function theory. In Jérôme Lang, Fangzhen Lin, and Ju Wang, editors, *KSEM*, volume 4092 of *Lecture Notes in Computer Science*, pages 265–277. Springer, 2006.

[45] Philippe Smets. Analyzing the combination of conflicting belief functions. *Information Fusion*, 8(4) :387–412, 2007.

[46] Thierry Denœux. Conjunctive and disjunctive combination of belief functions induced by non distinct bodies of evidence. *ARTIFICIAL INTELLIGENCE*, 2007.

[47] Eric Lefevre, Olivier Colot, and Patrick Vannoorenberghe. Belief function combination and conflict management. *Information Fusion*, 3(2) :149–162, 2002.

[48] Kari Sentz and Scott Ferson. Combination of evidence in dempster-shafer theory. Technical report, 2002.

[49] Philippe Smets and Robert Kennes. The transferable belief model. *Artif. Intell.*, 66(2) :191–234, 1994.

[50] A. Nifle and R. Raynaud. Un argument pour le choix entre décision pignistique et maximum de plausibilité en théorie de l'évidence. *GRETSI'97, Grenoble*, 1997.

[51] L. S. Shapley. A value for n-person games. *Contributions to the theory of games*, 2 :307–317, 1953.

[52] Philippe Smets. Belief induced by the partial knowledge of the probabilities. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, UAI'94, pages 523–530, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[53] A. P. Dempster. New methods for reasoning towards posterior distributions based on sample data. *The Annals of Mathematical Statistics*, 37(2) :355–374, 04 1966.

[54] Thierry Denœux. Constructing belief functions from sample data using multinomial confidence regions. *International Journal of Approximate Reasoning*, 42(3) :228 – 252, 2006.

[55] R. A. Fisher. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22 :700–725, 7 1925.

[56] T. Denœux. Maximum likelihood estimation from uncertain data in the belief function framework. *IEEE Trans. on Know. and Data Eng. (to appear)*, 2011.

[57] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 39(1) :1–38, 1977.

[58] Eyke Hüllermeier. Learning from imprecise and fuzzy observations : Data disambiguation through generalized loss minimization. *CoRR*, abs/1305.0698, 2013.

[59] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[60] Iftekhar Naim and Daniel Gildea. Convergence of the em algorithm for gaussian mixtures with unbalanced mixing coefficients. In *ICML*. icml.cc / Omnipress, 2012.

[61] H. Lebesgue. *Sur la mesure des grandeurs*. Monographies de L'Enseignement mathématique. L'enseignement Math., 1935.

[62] G. Choquet. Theory of capacities. *Annales de l'Institut Fourier*, 5 :131–295, 1953.

[63] W. A. Belson. A technique for studying the effects of a television broadcast. 5(3) :195–202, November 1956.

[64] J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data, and a proposal. 58 :415–434, 1963.

[65] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification And Regression Trees*. 1984.

[66] J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1 :81–106, October 1986.

[67] G. V. Kass. An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 29(2) :119–127, 1980.

[68] Manish Mehta, Rakesh Agrawal, and Jorma Rissanen. SLIQ : A Fast Scalable Classifier for Data Mining. In *Extending Database Technology*, pages 18–32, 1996.

[69] Wei-Yin Loh and Yu-Shan Shih. Split selection methods for classification trees, 1997.

[70] Wenhua Xu, Zheng Qin, and Yang Chang. Clustering feature decision trees for semi-supervised classification from high-speed data streams. *Journal of Zhejiang University - Science C*, 12(8) :615–628, 2011.

[71] Christophe Marsala. Fuzzy decision trees to help flexible querying, 2000.

[72] L. Breiman. Random forests. *Statistics*, pages 1–33, 2001.

[73] Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *J. Mach. Learn. Res.*, 9 :2015–2033, jun 2008.

[74] Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-Shing Ho, and Sau Dan Lee. Decision trees for uncertain data. *IEEE Transactions on Knowledge and Data Engineering*, 23(1) :64–78, 2011.

[75] Joaquín Abellán and Serafín Moral. Upper entropy of credal sets. applications to credal classification. *Int. J. Approx. Reasoning*, 39(2-3) :235–255, June 2005.

[76] Richard J. Crossman, Joaquín Abellán, Thomas Augustin, and Frank P. A. Coolen. Building imprecise classification trees with entropy ranges. In F. Coolen, G. de Cooman, Th. Fetz, and M. Oberguggenberger, editors, *ISIPTA'11 : Proceedings of the Seventh International Symposium on Imprecise Probability : Theories and Applications*, pages 129–138, Innsbruck, 2011. SIPTA.

[77] J.-S.R. Jang. Structure determination in fuzzy modeling : A fuzzy cart approach. 1994.

[78] C. Marsala. Fuzzy decision trees for dynamic data. In *Evolving and Adaptive Intelligent Systems (EAIS), 2013 IEEE Conference on*, pages 17–24, April 2013.

[79] C. Z. Janikow. Fuzzy decision trees : issues and methods. *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, 28(1) :1–14, January 1998.

[80] C. Z. Janikow and M. Fajfer. Fuzzy Partitioning with FID3 . 1. *Compute*.

[81] C. Z. Janikow. FID4.1 : an Overview. *Knowledge Creation Diffusion Utilization*, 2000.

[82] C. Z. Janikow. Fuzzy Decision Forest. *Information and Control*.

[83] C. Olaru. A complete fuzzy decision tree technique. *Fuzzy Sets and Systems*, 138(2) :221–254, September 2003.

[84] G.J. Klir and M.J. Wierman. *Uncertainty-Based Information : Elements of Generalized Information Theory*. Studies in Fuzziness and Soft Computing. Physica-Verlag HD, 1999.

[85] Christian Borgelt, Jörg Gebhardt, and Rudolf Kruse. Concepts for probabilistic and possibilistic induction of decision trees on real world data. In *Proc. 4th European Congress on Intelligent Techniques and Soft Computing (EUFIT'96, Aachen, Germany)*, volume 3, pages 1556–1560, Aachen, Germany, 1996. Verlag Mainz.

[86] M. Higashi and G. J. Klir. Measures of uncertainty and information based on possibility distributions. In D. Dubois, H. Prade, and R. R. Yager, editors, *Readings in Fuzzy Sets for Intelligent Systems*, pages 217–232. Kaufmann, San Mateo, CA, 1993.

[87] Ralph Vinton Lyon Hartley. Transmission of information. *Bell Syst. Tech. Journal*, 7 :535–563, 1928.

[88] N Ben Amor, Salem Benferhat, and Z Elouedi. Qualitative classification and evaluation in possibilistic decision trees. volume 2, pages 653–657. IEEE, jul 2004.

[89] Ilyes Jenhani, Nahla Ben Amor, Salem Benferhat, and Zied Elouedi. Sim-pdt : A similarity based possibilistic decision tree approach. In Sven Hartmann and Gabriele Kern-Isberner, editors, *FoIKS*, volume 4932 of *Lecture Notes in Computer Science*, pages 348–364. Springer, 2008.

[90] Masahiko Higashi and George J. Klir. On the notion of distance representing information closeness : possibility and probability distributions. *International Journal of General Systems*, 9(2) :103–115, 1983.

[91] Zied Elouedi, Khaled Mellouli, and Philippe Smets. Belief decision trees : theoretical foundations. *International Journal of Approximate Reasoning*, 28(2–3) :91 – 124, 2001.

[92] Marte Skarstein Bjanger, Fakultet For Fysikk, Informatikk Og Matematikk, and Fag Datateknikk Kunnskapssystemer. Induction of decision trees from partially classified data using belief functions, 2000.

[93] Benjamin Quost, Thierry Denœux, and Marie-Hélène Masson. Pairwise classifier combination using belief functions. *Pattern Recognition Letters*, 28(5) :644 – 653, 2007.

[94] P. Vannoorenberghe and T. Denoeux. Handling uncertain labels in multiclass problems using belief decision trees, 2002.

[95] Leo A. Goodman. On Simultaneous Confidence Intervals for Multinomial Proportions. *Technometrics,*, 7(2) :247–254, May 1965.

[96] LUIS M. DE CAMPOS, JUAN F. HUETE, and SERAFIN MORAL. Probability intervals : A tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 02(02) :167–196, 1994.

[97] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 539. Wiley New York, 1987.

[98] Julie Josse, Marie Chavent, Benot Liquet, and François Husson. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of classification*, 29(1) :91–116, 2012.

[99] Djamel A. Zighed, Simon Marcellin, and Gilbert Ritschard. Mesure d'entropie asymétrique et consistante. In Monique Noirhomme-Fraiture and Gilles Venturini, editors, *EGC*, volume RNTI-E-9 of *Revue des Nouvelles Technologies de l'Information*, pages 81–86. Cépaduès-Éditions, 2007.

[100] Mortaza Jamshidian and Robert I. Jennrich. Acceleration of the EM Algorithm by Using Quasi-Newton Methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(3) :569–587, 1997.

[101] Xiao L. Meng and Donald B. Rubin. Maximum Likelihood Estimation via the ECM Algorithm : A General Framework. *Biometrika*, 80(2) :267–278, 1993.

[102] David R. Hunter, Kenneth Lange, Departments Of Biomathematics, and Human Genetics. A tutorial on mm algorithms. *Amer. Statist*, pages 30–37, 2004.

[103] Yasuo Matsuyama. The alpha-em algorithm : surrogate likelihood maximization using alpha-logarithmic information measures. *IEEE Transactions on Information Theory*, 49(3) :692–706, 2003.

[104] Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, 14(3) :315–332, October 1992.

Résumé de la thèse

Pour l'apprentissage de modèles prédictifs, la qualité des données disponibles joue un rôle important quant à la fiabilité des prédictions obtenues. Ces données d'apprentissage ont, en pratique, l'inconvénient d'être très souvent imparfaites ou incertaines (imprécises, bruitées, etc). Ce travail de doctorat s'inscrit dans ce cadre où la théorie des fonctions de croyance est utilisée de manière à adapter des outils statistiques classiques aux données incertaines.

Le modèle prédictif choisi est l'arbre de décision qui est un classifieur basique de l'intelligence artificielle mais qui est habituellement construit à partir de données *précises*. Le but de la méthodologie principale développée dans cette thèse est de généraliser les arbres de décision aux données incertaines (floues, probabilistes, manquantes, etc) en entrée **et** en sortie. L'outil central d'extension des arbres de décision aux données incertaines est une vraisemblance adaptée aux fonctions de croyance récemment proposée dont certaines propriétés sont ici étudiées de manière approfondie. De manière à estimer les différents paramètres d'un arbre de décision, cette vraisemblance est maximisée via l'algorithme E^2M qui étend l'algorithme EM aux fonctions de croyance.

La nouvelle méthodologie ainsi présentée, les arbres de décision E^2M, est ensuite appliquée à un cas réel : la prédiction de la qualité du caoutchouc naturel. Les données d'apprentissage, essentiellement culturales et climatiques, présentent de nombreuses incertitudes qui sont modélisées par des fonctions de croyance adaptées à ces imperfections. Après une étude statistique standard de ces données, des arbres de décision E^2M sont construits et évalués en comparaison d'arbres de décision classiques. Cette prise en compte des incertitudes des données permet ainsi d'améliorer très légèrement la qualité de prédiction mais apporte surtout des informations concernant certaines variables peu prises en compte jusqu'ici par les experts du caoutchouc.