Une approche d'ingénierie inverse combinant ontologies et modèles relationnels probabilistes: application aux emballages bio-composites

Mélanie Munch, Patrice Buche, Cristina Manfredotti, Pierre-Henri Wuillemin, Helene Angellier-Coussy

Une approche d'ingénierie inverse combinant ontologies et modèles relationnels probabilistes: application aux emballages bio-composites

M. Munch*[1], P. Buche[2, 3], C. Manfredotti[4], P-H. Wuillemin[5], H. Angellier-Coussy[2]

[1] I2M, U. Bordeaux, INRAE, Talence, France
[2] IATE, U. Montpellier, INRAE, CIRAD, Montpellier SupAgro, Montpellier, France
[3] LIRMM, U. Montpellier, CNRS, INRIA GraphIK, Montpellier, France
[4] UMR MIA-Paris, AgroParisTech, INRAE, Paris-Saclay University, Paris, France
[5] Sorbonne University, UPMC, Univ Paris 06, CNRS UMR 7606, LIP6, 75005 Paris, France

melanie.munch@u-bordeaux.fr

Résumé

En raison des nombreux paramètres et variables à prendre en compte, la conception de nouveau procédés de transformation pour les emballages alimentaires est un défi économique et écologique. Pour le relever, celui-ci nécessite (1) l'intégration de sources hétérogènes de données et (2) de pouvoir raisonner causalement. Dans cet article, nous présentons POND (Process and observation ONtology Discovery), un workflow dédié à l'étude de questions expertes sur des domaines modélisés par la Process and Observation Ontology (PO²), ontologie dédiée à la représentation de procédés de transformation. Nous illustrons son fonctionnement à travers un problème concret d'ingénierie inverse à partir d'une base de données inédite dans le cadre de la conception d'emballages alimentaires bio-composites.

Mots-clés

Graphe de Connaissance, Modèle probabiliste, Causalité, Emballage alimentaire

Abstract

Designing new processes for bio-based and biodegradable food packaging is an environmental and economic challenge. Due to the multiplicity of the parameters, such an issue requires an approach that proposes both (1) to integrate heterogeneous data sources and (2) to allow causal reasoning. In this article, we present POND (Process and observation ONtology Discovery), a workflow dedicated to answering expert queries on domains modeled by the Process and Observation Ontology (PO2), an ontology dedicated to represent transformation processes. The presentation is illustrated with a real-world application on bio-composites for food packaging to solve a reverse engineering problem, using a novel dataset composed of data from different projects.

Keywords

Knowledge graph, Probabilistic model, Causality, Food packaging

Acronymes

BC : Base de Connaissance ; **CD :** Contrainte Dure ; **CL :** Charge Lignocellulosique ; **CS :** Contrainte Souple ; **GE :** Graphe Essentiel ; **MRP :** Modèle Relationnel Probabiliste ; **QCc :** Question de Connaissance causale ; **RB(C) :** Réseau Bayésien Causal ; **SR :** Schéma Relationel

1 Introduction

Chaque année, l'utilisation massive de plastique résulte en une accumulation constante de déchets environnementaux, avec des conséquences désastreuses à la fois pour les écosystèmes et la santé humaine. Face à l'épuisement grandissant des énergies fossiles et de la production croissante de résidus organiques (agricoles, urbains, forestiers ou d'industries agro-alimentaires) non récupérés, des technologies innovantes ont été développées pour la production de matériaux recyclables, biosourcés et biodégradables. Parmi ces solutions, le poly(3-hydroxybutyrate-co-3-hydroxyverate) (PHBV) est un biopolymère bactérien prometteur, dégradable dans le sol et les océans, pouvant être synthétisé à partir de toutes sortes de résidus carbonés. Afin de diminuer son coût et empreinte carbone et d'agir sur le réchauffement climatique, le développement de biocomposites de PHBV avec des produits lignocellulosiques a été travaillé [7]. Néanmoins, cette augmentation dans la composition de fibres lignocellulosiques a également un impact sur la fragilité du composite final et ses aptitudes de mise-en-œuvre. Ainsi, un compromis doit être trouvé entre la quantité maximale intégrable de charge et les caractéristiques finales du bio-composite. Pourtant, trouver des liens causaux entre les différentes variables présentées depuis le jeu de données seul peut être ardu. Si les travaux précédents sur la causalité suggèrent l'usage d'interventions (i.e. changer une variable en gardant les autres constantes pour évaluer les effets) pour construire des modèles causaux [26], celles-ci peuvent de-

venir très vite chronophages et onéreuses. Dans cet article, nous présentons une alternative, POND (Process and observation ONtology Discovery), un workflow basé sur la combinaison entre une ontologie dédiée à la représentation de procédés de transformation, PO² [18], et des modèles probabilistes. L'idée principale est de combiner les connaissances expertes contenues dans le graphe de connaissance [12] avec celles prodiguées par un expert pour guider l'apprentissage d'une extension des réseaux Bayésiens (RB), le modèle relationnel probabiliste (MRP) [15]. Le modèle ainsi appris sous contraintes est alors capable de raisonner causalement sur le problème étudié, afin de répondre à des questions expertes. Les contributions originales de cet article sont (1) l'intégration complète de PO² dans un pipeline pour répondre à des questions expertes ; (2) un outil pour répondre à des questions causales permettant une approche d'ingénierie inverse ; (3) une méta-analyse de différents projets menés sur l'étude des emballages alimentaires biocomposites. La section 2 de cet article présente les prérequis utiles à la compréhension de POND, couvrant à la fois l'ontologie PO², les modèles graphiques probabilistes et la découverte causale. La section 3 introduit POND et souligne ses contributions à l'état de l'art sur la combinaison entre bases de connaissances et modèles probabilistes dans le cadre d'intégration de connaissances expertes. La section 4 illustre POND à travers l'exemple des emballages biocomposites et d'un exemple concret. Pour ce faire, nous basons notre étude sur une base de connaissances novatrice, composée à partir de différents projets.

Cet article est une version traduite d'un travail préalablement publié par les auteurs [20], et récompensé du *Best Paper Award*.

2 Travaux antérieurs

2.1 Process and Observation Ontologie PO²

PO² est une ontologie générique dédiée à la représentation des procédés de transformation. Initialement dédiée à la science des aliments [18], elle a été développée via le scénario 6 de la méthodologie NeON [30], en retravaillant une ontologie préexistante dédiée à l'éco-conception de procédés de transformation [10]. Elle a récemment été utilisée pour des produits biosourcés tels que les emballages alimentaires biocomposites. La Figure 1 présente un aperçu de ses différents concepts, décrits par 67 concepts et 79 relations. Un **procédé de transformation** est représenté ici comme une succession d'**étapes**, mises en relation par des **entités temporelles**, auxquelles sont rattachés des **résultats** expérimentaux pouvant être mesurés à de multiples échelles et unités sur différents **composants** (représentant des facteurs d'intérêts). La version 2.0 de PO² est implémentée en OWL 2 [1], et publiée sur AgroPortal [2] en licence publique Creative Commons Attribution International (CC BY 4.0) [3].

1. https ://www.w3.org/TR/owl2-overview/
2. http ://agroportal.lirmm.fr/ontologies/PO2
3. https ://creativecommons.org/licenses/by/4.0/

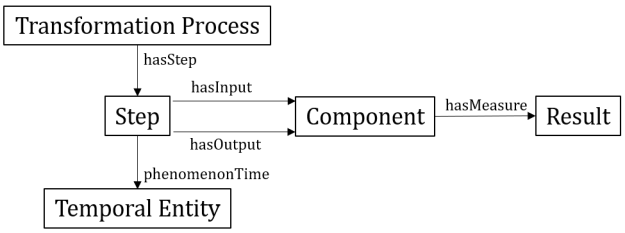

FIGURE 1 – Représentation simplifiée des composants principaux de l'ontologie PO².

2.2 Modèles Probabilistes : RB et MRP

2.2.1 Réseaux Bayésiens

Un RB est la représentation de probabilités jointes sur un ensemble de variables aléatoires dont les dépendances probabilistes sont encodées par un graphe acyclique orienté. Ainsi, chaque nœud représente une variable, et chaque flèche une relation probabiliste. L'apprentissage d'un RB se fait généralement en deux temps : la structure d'abord, puis les dépendances probabilistes. Dans notre cas, cet apprentissage se fait en utilisant l'algorithme classique Greedy Hill Climbing [5] guidé par un score BIC [27]. Un RB dont chaque relation traduit dans le domaine représenté une relation causale est appelé **RB causal** (RBC). Un exemple de RBC est présenté par la Figure 4, tandis que la Table 2 montre un exemple de dépendance probabiliste sous la forme d'un tableau de dépendance conditionnelle (dans ce cas, l'évolution des probabilités de la variable **Contrainte à la rupture** en fonction des valeurs de la variable **Charge**).

2.2.2 Graphes Essentiels

Pour chaque RB, il est possible de déduire le graphe essentiel (GE), graphe acyclique semi-orienté exprimant la classe d'équivalence de Markov du réseau. Bien que RB et GE partagent la même structure (mêmes nœuds et mêmes liens), certaines relations ne sont pas orientées dans le GE, se traduisant par une arête sans flèche. Cette présence (ou non) d'orientation traduit la nécessité de la conserver pour préserver les relations d'indépendances encodées dans le graphe. Plus généralement, si une relation n'est pas orientée dans le GE, alors celle-ci peut être inversée dans le RB sans modification des relations de dépendance sous-jacentes ; en revanche, si elle est orientée, alors son inversion nécessite le réapprentissage de toute la structure du RB. La Figure 2 illustre deux exemples de classes d'équivalence de RB et leur GE associé. Comme nous le verrons par la suite, le GE est donc une source d'information pouvant être utilisée, dans un certain cadre (que nous définirons par la suite) pour la découverte causale.

2.2.3 Modèles Relationels Probabilistes

Le MRP est une extension orientée objet des RB. Là où un RB ne nécessite qu'une seule couche d'information pour être décrit, le MRP en nécessite deux : (1) le schéma relationnel (SR), qui fournit une description qualitative des classes et de leurs variables (appelées dans ce cadre attributs) composant sa structure ; et (2) le modèle relationnel

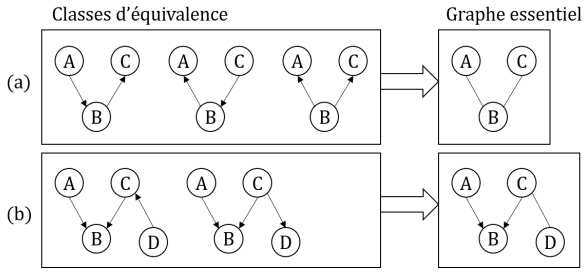

FIGURE 2 – Exemples de deux classes d'équivalence de Markov et leur graphe essentiel attribué. (a) Dans le premier exemple, les arcs des RB peuvent être orientés dans toutes les directions sans incidence sur les relations entre variables : le GE est donc complètement non-orienté. (b) Dans le second en revanche les trois variables A, B et C forment une structure particulière ne pouvant être modifiée : elle est donc marquée dans le GE.

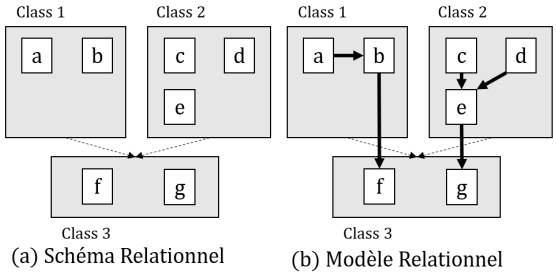

FIGURE 3 – Exemple d'un MRP et de la définition de ses (a) hautes et (b) basses structures.

(MR), qui contient toutes les informations quantitatives sur la distribution probabiliste entre les attributs. Deux classes peuvent être liées dans le SR par des liens relationnels, qui indiquent la direction des liens probabilistes entre les attributs : ainsi, dans la Figure 3, le lien de **Classe 1** vers **Classe 3** dans le SR force l'orientation de la relation entre les attributs b et f dans le MR. En revanche, en l'absence de lien relationnel (comme par exemple entre **Classe 1** et **Classe 2**), aucun lien ne peut être appris. Enfin, l'orientation des liens au sein d'une même classe n'est pas forcée. L'apprentissage entièrement automatique d'un MRP, en deux temps (pour le SR puis le MR) est complexe : dans notre cas, nous construisons à la main le SR, réduisant la complexité d'apprentissage du MR à celle d'un RB [15]. Une fois le SR et le MR définis, le MRP est similaire à un RB.

2.2.4 Apprentissage sous contraintes

L'apprentissage sous contraintes (*i.e.*, guidé par des connaissances restreignant l'espace de recherche) dans le cadre des RB permet d'améliorer grandement la précision des résultats, qu'il s'agisse de celle de la structure [9] ou des paramètres [8]. Cela se vérifie d'autant mieux dans le cadre de petites bases de données [23]. Dans cette optique, des travaux précédents ont proposé des méthodes pour imposer un ordre total [6] ou partiel [25] sur les nœuds, créant ainsi des contraintes directionnelles sur leurs relations. Dans notre cas, la définition manuelle du SR permet de créer un ordre partiel entre les variables. Celui-ci transcrit les connaissances expertes apportées par la base de connaissance (BC) et l'expert, favorable à un contexte de découverte causale comme présenté dans [21]. Puisque dans notre cas nous intégrons des connaissances expertes lors de l'apprentissage, celui-ci est considéré comme étant fait sous *contraintes causales*.

2.3 Découverte de Connaissances

2.3.1 Découverte Causale

Il est bien connu qu'une corrélation n'implique pas forcément de causalité : il est donc important de répondre à certains critères précis lorsque l'on se place dans un contexte de découverte causale. Parmi les principaux, on retrouve l'absence de facteurs extérieurs non mesurés (la **suffisance causale** [29]) ; et l'absence de données erronées, de biais de sélection ou de cas déterministique [16]. D'une façon générale, le jeu d'apprentissage doit être représentatif du domaine que l'on veut modéliser : si un évènement n'est pas représenté, ou si les proportions ne sont pas conformes à la réalité (par la sur-représentation d'un cas par exemple), il devient alors impossible de tirer de bonnes conclusions causales. La découverte de causalité à partir d'un jeu de données peut se faire par des calculs d'indépendance entre variables [29, 31] ; ces méthodes néanmoins ne permettent pas d'intégrer de connaissances extérieures. Certains travaux ont proposé l'utilisation du GE pour apprendre des modèles causaux : [17] propose deux stratégies optimales pour suggérer des interventions afin d'apprendre des modèles causaux ; [28] et [4] utilisent le GE pour construire un RBC tout en maintenant un nombre limité d'interventions possibles. Tout comme les méthodes à base de calculs d'indépendances, ces méthodes n'intègrent pas de sources de connaissance extérieures. Notre but étant au contraire de pouvoir intégrer les connaissances expertes contenues dans des sources externes, nous avons donc choisis de combiner l'apprentissage avec des ontologies.

2.3.2 Combinaison avec les Ontologies

La découverte causale à partir de données seules est une tâche ardue. Pour cette raison, de nombreux travaux ont entrepris d'intégrer les connaissances contenues dans les ontologies pour apprendre des modèles probabilistes et découvrir de nouvelles relations. Par exemple, différentes extension d'ontologies permettent l'intégration directe de raisonnement probabilistes (comme BayesOWL [11, 32], ou HyProb-Ontology [19]). Si elles permettent de raisonner, elles n'autorisent néanmoins pas l'apprentissage de nouvelles relations. Au contraire, d'autres travaux partent de la structure de l'ontologie pour construire un RB, en considérant par exemple les propriétés objets comme des relations probabilistes [14] ou causales [1] ; cet a priori ne peut néanmoins pas s'appliquer à PO², où de nombreuses relations ne peuvent être directement traduites causalement. D'autres méthodes, finalement, ont été développées pour répondre à certains cas précis, ne pouvant être étendus à d'autres cas : ainsi, [2] utilise des modèles prédéfinis pour

réaliser des diagnostics médicaux, qui ne peuvent être étendus à d'autres applications médicales. Il est à noter que bien que POND se base sur une unique ontologie (PO²), la complexité de celle-ci lui permet d'aborder de nombreux problèmes sur des applications plus larges qu'une simple ontologie de domaine. Dans notre cas, il est ainsi possible de raisonner sur n'importe quel procédé de transformation, pour peu qu'il soit descriptible par l'ontologie.

3 POND : PO² Ontology Discovery

Dans cette section nous présentons POND, dont le but est l'intégration de connaissances expertes dans l'apprentissage d'un modèle probabiliste afin de raisonner dessus. Nous nous concentrerons ici sur les différentes sources pouvant être utilisées pour répondre à de complexes questions probabilistes et causales. Notre application finale étant un problème d'ingénierie inverse, nous mettrons un accent particulier sur la découverte causale et les applications possibles offertes par cette dernière.

3.1 Intégration de Connaissances

Expression. Les connaissances expertes peuvent venir : (1) de données expérimentales, récupérées auprès de différentes sources (telles que des publications, livres ou données produites au sein de projets) ; (2) d'entretiens directs sollicités auprès d'experts du domaine. La plupart de ces informations peuvent ensuite être directement structurées dans l'ontologie PO² : cela concerne les données factuelles, descriptives du domaine. L'intérêt de cette sémantisation est qu'elle permet d'établir un vocabulaire cohérent et complet, indispensable pour la suite. À partir de celui-ci, l'expert peut ainsi formuler des **Questions Expertes** de deux types : certaines restent à un niveau descriptif, et peuvent être répondues en requêtant directement l'ontologie (**Questions de Compétences**) ; d'autres requièrent en revanche un raisonnement probabiliste, et nécessitent l'apprentissage d'un modèle (**Questions de Connaissance**). Dans cet article, nous nous focaliserons sur les questions de connaissance causales (QCcs), qui peuvent être formalisées de deux façons, en définissant X_i et X_j deux groupes de variables du domaine :

QCc_1 Est-ce que X_i a une influence sur X_j ?
QCc_2 Quel est l'impact de X_i sur X_j ?

Ces deux questions illustrent la double lecture offerte par les RBC : alors que QCc_1 se concentre sur l'aspect descriptif des relations apprises (pouvant être déduit directement du graphe), QCc_2 interroge plutôt leur nature (pouvant être analysée à partir des tables de probabilités conditionnelles).

Intégration. Une fois la QCc définie, le modèle peut être construit. Comme décrit en Section 2.2, le but ici est de transcrire les connaissances expertes exprimées au préalable dans le SR du MRP. L'originalité de notre approche repose sur la façon dont cette connaissance est intégrée :

1. **Par l'alignement des variables de l'ontologie dans le SR.** Grâce au vocabulaire commun défini dans PO², l'expert peut facilement extraire les variables intéressantes pour sa question. La définition sémantique permet également de récupérer de façon automatique les valeurs associées, même si elles ont été mesurées dans différents contextes, et les unifier (ou non). Cette question de l'unification est posée par la structure même de certains procédés. Ainsi, supposons un procédé de transformation où, dépendant des itérations, une température donnée est mesurée soit à l'étape A, soit à l'étape B. Dans ce cas, l'expert peut alors décider si ces températures sont similaires (i.e. elles peuvent être comparées, et donc unifiées), ou au contraire si elles représentent deux mesures différentes. Ce genre de problème ne peut être résolu par l'ontologie directement : dépendant de la QCc, la distinction entre les températures des étapes A ou B peut être pertinente ou non. La structuration sémantique du vocabulaire est donc ici importante : elle permet à l'expert de construire le modèle, tout en utilisant un vocabulaire qui lui est accessible, pour décrire des connaissances que seul lui peut fournir. À partir de PO², l'expert peut alors spécifier les attributs à représenter dans le SR, en spécifiant à chaque fois l'itinéraire, l'étape et le composant sur lequel a eu lieu la mesure. Cela permet de lier à chaque variable ses valeurs, à savoir ici les datatypes, qui permettent de composer la base d'apprentissage du RBC.
2. **Par la définition des contraintes de précédence.** Par définition, les relations entre classes du SR établissent les contraintes de précédence : si une telle contrainte existe entre la classe comprenant la variable A vers la classe comprenant la variable B, alors un lien appris sera forcément orienté de A vers B. Au contraire, deux variables définies dans une même classe n'auront aucune contrainte de direction quant à leur potentielle relation. Ces contraintes de précédences traduites en liens relationnels entre classes peuvent être déduites des informations temporelles contenues dans le BC (par exemple, une mesure faite à une étape au temps t peut avoir une influence sur une mesure faite à une étape au temps $t + n$, mais pas l'inverse). Elles peuvent être également fournies par l'expert, en tant que causalités potentielles du type "Je sais que la variable A peut expliquer la variable B" ou, au contraire, "A et B ne peuvent pas partager de lien."

Notre contribution dans cette section est la formalisation de l'intégration de connaissances dans un workflow : grâce à PO², tout procédé de transformation peut être aisément intégré dans un SR, grâce au vocabulaire commun défini avec l'expert. Ce dernier permet également la constitution automatique de la base d'apprentissage (utilisée lors de la définition du MR) directement à partir des faits contenus dans la BC.

3.2 Validation Causale

Une fois le SR construit par l'expert par l'intégration de ses contraintes de précédences, le MR peut être appris ; il devient alors possible d'instancier le MRP (défini par le SR et

le MR) en un RB. Dans notre cas, nous considérons que les connaissances expertes intégrées ont permis d'apprendre le modèle sous contraintes causales, permettant ainsi d'utiliser le RB pour déduire des connaissances causales [22]. Cela est dû au fait que nous considérons ce modèle appris comme l'intersection entre (1) toutes les contraintes contenues dans le jeu de données utilisé par l'apprentissage (exprimé par le GE) ; et (2) toutes les connaissances expertes intégrées dans l'apprentissage (exprimées par le SR). Encore une fois, il est important de rappeler que ces déductions se basent sur le fait que nous nous considérons en contexte favorable à la découverte causale tel que décrit en Section 2.3.1. La validation se déroule alors de la façon suivante :

- Si une relation est apprise entre deux variables ayant une contrainte de précédence définit par l'expert, alors la causalité est validée par la connaissance experte.
- Si une relation apprise est représentée par un arc orienté dans le GE, alors la causalité est validée par les données à travers le GE. Cette déduction suit le raisonnement suivant : si la base de faits utilisée pour l'apprentissage est fiable, alors cette relation ne peut être orientée que de cette façon pour respecter les contraintes appliquées lors de l'apprentissage et les indépendances calculées entre les variables inhérentes à la base d'apprentissage. De plus, cette relation est validée même si aucune contrainte de précédence n'a été placée par l'expert entre ces variables.
- Si une relation est apprise mais qu'elle n'est validée ni par l'expert ni par le GE, alors il est impossible d'en déduire de la causalité.

Idéalement, cette découverte causale a pour but de valider causalement toutes les relations du RB, permettant ainsi de définir un RBC. Néanmoins, même si toutes les relations ne peuvent être validées, elle permet également :

- **D'aider l'expert à critiquer.** Puisque nous cherchons à modéliser des domaines réels, l'évaluation directe est parfois très compliquée (voire impossible) à réaliser directement. En revanche, en présentant les relations causales apprises à l'expert, nous lui donnons un outil pour les critiquer et les questionner à partir de ses connaissances propres, en suggérant par exemple de nouvelles hypothèses à vérifier expérimentalement.
- **De répondre aux QCcs.** Une QCc dépend de la découverte causale pour être résolue : QCc_1 regarde la présence (ou l'absence) de relation entre les variables concernées, tandis que QCc_2 utilise ces relations pour raisonner sur les tables de probabilités conditionnelles. Si les relations nécessitant d'être vérifiées ne sont pas validées, alors il est impossible de répondre aux QCcs dans un cas comme dans l'autre.

Il est important de noter qu'en cas de non-validation, plusieurs solutions sont possibles : l'expert peut fournir de nouvelles connaissances (sous la forme de données supplémentaires, ou de nouvelles contraintes de précédences) ; de nouvelles expériences peuvent être suggérées, pour par exemple compléter des trous de connaissance identifiés ; ou bien la QC peut être redéfinie. Si néanmoins aucune de ces solutions ne fonctionne, alors la BC est jugée insuffisante pour répondre aux questions actuelles de l'expert.

Une application de ce protocole de validation est donné en exemple en Section 4.2.2.

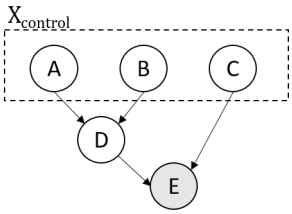

FIGURE 4 – Exemple d'un court RBC. L'ensemble $X_{control}$ représente les variables de contrôle, i.e. les variables sur lesquelles l'expert peut intervenir. E représente la variable cible.

3.3 Inférences Causales

Nous avons vu jusqu'à présent comment intégrer les connaissances expertes pour apprendre un modèle, et comment valider celui-ci. Si cela est généralement suffisant pour répondre directement aux questions de type QCc_1 (en vérifiant la présence ou absence de relations validées causalement), les QCc_2 et leurs dérivées nécessitent une analyse plus approfondie. Pour illustrer ce besoin, nous considérerons le RBC présenté en Figure 4, et supposerons qu'il a été validé causalement. Nous considérons également la QCc QCc_{ex} : *"Quelle intervention dois-je faire pour maximiser la variable E ?"*, combinaison entre la QCc_1 (*"Quelles variables ont un impact sur E ?"*) et la QCc_2 (*"Quelle est l'influence de ces variables ?"*).

Afin de répondre à QCc_{ex}, nous devons tout d'abord identifier les variables sur lesquelles il sera possible d'intervenir. Dénotées comme *variables de contrôles*, elles se distinguent des variables ayant également un impact sur la variable cible, mais sur lesquelles aucune intervention n'est envisageable. Dans notre exemple, nous pouvons ainsi voir que les parents directs de E sont D et C. Néanmoins, D n'appartient pas à $X_{control}$: il est donc nécessaire de remonter jusqu'à ses parents, A et B. Nous définissons donc pour la variable E l'ensemble $X_{inter} = X_{control} \cap Parents(E)$ (dans notre cas, $\{A, B, C\}$), qui contient les variables sur lesquelles on peut intervenir pour répondre à QCc_{ex}. En effet, puisque nous considérons un RBC, alors intervenir sur un parent aura un effet sur les enfants. En pratique, cela signifie que pour chaque combinaison de valeurs de A, B ou C, les valeurs de E et leurs probabilités seront affectées : cela constitue une base de tous les scénarios potentiels, dont il faudra extraire celui permettant de répondre à QCc_{ex}. Pour faire cela, l'expert est à nouveau sollicité pour définir ses propres critères d'acceptabilité, tels que "Quelles valeurs sont préférables pour la variable cible ?", ou "Quelles conditions devraient s'appliquer sur X_{inter} ?".

Ces critères sont classifiés en deux sortes :

— **Critères durs.** Certaines valeurs ou combinaisons de valeurs sont impossibles à atteindre : les scénarii correspondants sont alors automatiquement retirés. Par exemple, les experts peuvent souhaiter que la somme des variables de X_{inter} ne dépassent pas un certain seuil ; ou bien ils souhaitent exclure certaines valeurs de E. Dans notre cas, puisque l'on souhaite maximiser E, alors on ne considère pas ses valeurs les plus basses.

— **Critères souples.** Dans certains cas, l'expert a besoin de classer ses préférences dépendant du contexte. Peut-être qu'avoir une haute valeur de E n'est pas intéressant si A est également élevée ; ou peut-être qu'une valeur plus basse de E, mais avec une plus haute probabilité de réalisation, est un cas plus intéressant qu'un meilleur scénario n'ayant aucune chance d'arriver.

Définir ces critères permet de mieux définir les besoins de l'expert, et ainsi trouver la réponse à QCc_{ex} y correspondant le mieux. Nous montrerons en Section 4.3 comment cette approche peut être appliquée concrètement dans le cadre d'un problème d'ingénierie inverse.

4 Application aux Emballages Biocomposites

Pour la suite, nous définissons la QCc_{bio} : "Quelles caractéristiques de la charge ligno-cellulosique permettent d'optimiser les propriétés mécaniques de l'emballage ?".

4.1 Présentation de la Base de Connaissance

Nous avons collecté des données de cinq projets tournés vers le développement de biocomposites à partir de PHBV et de charges lignocellulosiques (CLs) provenant de déchets organiques tels que des rejets de cultures (*Chercheur d'avenir région Languedoc-Roussilon MALICE* et *H2020 NoAW*), des dérivées de l'industrie agro-alimentaire (*FP7 EcoBioCAP*) ou de déchets urbains (*H2020 Resurbis*). Les CLs de ces projets ont été obtenues par fractionnement à sec de la biomasse pure. Enfin, pour comparer, des fibres de cellulose pure ont également été utilisées comme références dans le cadre du projet *H2020 Usable*. Au final, cela constitue une base d'apprentissage de 88 formulations décrites par 15 attributs différents [24].

4.2 Intégration des Connaissances Expertes

L'intégration des connaissances expertes se déroule en deux temps : (1) la correspondance des attributs intéressants de la BC jusqu'au SR, et (2) la définition des contraintes de précédence potentielles. Dans cette section, nous présentons les résultats principaux utilisés pour apprendre le modèle final, ainsi qu'un exemple d'intégration de critiques expertes.

4.2.1 Définition du SR

Sélection des Attributs.[4] L'expert décrit une CL par trois catégories d'attributs : la composition biochimique (**cellulose**, **hémicellulose**, **cendres** et **lignine**) ; le diamètre médian apparent (**D50**) ; la **charge de remplissage** (i.e. la quantité ajoutée au produit final). Le produit final, quant à lui, est décrit par quatre catégories distinctes d'attributs : les propriétés mécaniques (**contrainte à la rupture**, **stress à la rupture** et **module de Young**), la **perméabilité** (à la vapeur d'eau), les propriétés thermiques (températures de **cristallisation** et de **fusion**), et la dégradation thermale (températures de **début** et de **pic**). Parmi ces dernières catégories, seules les propriétés mécaniques sont nécessaires pour décrire QCc_{bio} ; néanmoins, dans un contexte de découverte causale (pas d'attributs manquants) et dans une optique de facilitation des critiques expertes, nous intégrons dans un premier temps les autres catégories.

4. Pour le reste de l'article, tous les attributs représentés dans le modèle seront indiqués en **caractères gras**.

Lignine	[0 ;19] *(32)*	]19.4 ;26.4] *(30)*	]26.4 ;49] *(23)*	
Charge	[2 ;4] *(10)*	]4 ;11] *(34)*	]11 ;21] *(22)*	]21 ;50] *(19)*
CR	[0.2 ;0.5] *(19)*	]0.5 ;0.8] *(44)*	]0.8 ;1] *(15)*	]1 ;1.07] *(3)*

TABLE 1 – Extrait de la discrétisation utilisée dans notre application pour la **Lignine**, la **charge de remplissage** et la **contrainte à la rupture** (CR) *(quantité de données pour une catégorie).*

Discrétisation des Attributs. Puisqu'il s'agit pour la plupart de mesures réalisées sur les produits, les données contenues dans notre BC sont pour la plupart continues (i.e. elles ne peuvent être automatiquement rangées dans des catégories distinctes discrètes). Cependant, de par leur nature, les RB classiques ne peuvent apprendre à partir de ce type de données ; il est donc nécessaire de passer par une phase de discrétisation des variables considérées. Cette étape est importante, car pouvant influencer l'apprentissage des différentes relations entre les variables et ainsi changer l'interprétation du modèle. Elles sont ainsi très sensibles aux retours prodigués par les experts : il est donc important de convenir d'une discrétisation proposant une description équilibrée des classes (pas de sur-représentation d'une valeur pouvant déséquilibrer le modèle), mais présentant également une cohérence avec le domaine cible. Dans notre cas, nous cherchons par exemple à déterminer si les caractéristiques cibles sont dégradées (ratio valeur finale sur valeur initiale strictement inférieure à 1), conservées (valeur égale à 1) ou améliorées (valeur strictement supérieure à 1). Un exemple de la discrétisation appliquée dans notre application est donnée dans la Table 1 : les variables de contrôle sont réparties de façon équilibrées par rapport au jeu de données initial (exemple de la **lignine**), tandis que les variables cibles ont une discrétisation choisie par l'expert (exemple de la **charge** ou de la **contrainte à la rupture**).

Définition des Contraintes de Précédence. Dans notre cas, l'expert définit initialement deux contraintes de précédence qui seront précisées par la suite :

— Entre les variables du CL et les caractéristiques finales de l'emballage. Ainsi, les premières sont considérées comme des variables de contrôle pouvant avoir un impact sur les secondes, qui décrivent le résultat final. On définit donc deux classes dans le SR, avec un lien relationnel allant de la classe des

variables de contrôle vers la classe des variables de description des caractéristiques.

— Entre les différentes catégories des variables de description des caractéristiques. Cette distinction permet de considérer chaque catégorie comme un sous-groupe indépendant des autres (e.g., les caractéristiques mécaniques n'ont aucune influence sur les caractéristiques thermales). Pour modéliser ceci, la classe définie précédemment est elle-même compartimentée en différentes sous-classes indépendantes les unes des autres.

4.2.2 Retour Expert

Une fois le premier modèle appris, une discussion avec l'expert est requise pour critiquer à la fois (1) les relations apprises et (2) les dépendances probabilistes. En illustration, nous considérons le modèle appris présenté en Figure 5. Dans cette situation, l'expert a relevé plusieurs incohérences :

— La somme des constituants de la CL vaut 100 : il est donc cohérent que des corrélations soient apprises entre eux. Néanmoins, celles-ci n'ont aucun sens d'un point de vue causal. Par conséquent, il est décidé de les placer dans des sous-classes indépendantes : cela traduit le fait que l'expert a un total contrôle sur elles et peut aidément les faire varier de façon indépendante. Ce choix de modélisation conduira en Section 4.3 à l'élaboration de la contrainte CD_1 garantissant qu'une formulation soit techniquement possible (i.e. la somme des constituant vaut toujours 100).

— La température de **fusion** ne peut pas expliquer la température de **cristallisation** : la relation apprise est une corrélation, pas une relation causale. Pour y remédier, les deux paramètres sont séparés dans des sous-classes indépendantes.

— La **contrainte à la rupture** n'est contre toute attente expliquée par aucun paramètre. Une nouvelle discrétisation est testée pour tenter de mieux représenter la variable.

Le retour expert permet d'identifier des trous de connaissances (i.e. des cas non représentés dans la BC pouvant entrainer des apprentissages incomplets). Ainsi, le modèle décrit montre que lorsque la **charge de remplissage** $\in$]21 ;50[, alors la température de **fusion** $\notin$]1 ;1.02[. Cela peut être dû à deux raisons : (1) il s'agit bien d'un trou de connaissance, et la BC doit être complétée ; (2) il s'agit d'un cas normal, ne nécessitant pas d'ajouts de données supplémentaires. Dans cet exemple, l'expert a bien pu confirmer que l'absence d'amélioration de la température de **fusion** était cohérent dans le cas d'une augmentation de la **charge**.

Après quelques allers-retours, un RS a été retenu pour l'étude de la QCc. Celui-ci est présenté en Figure 6.

4.3 Résolution de la Question de Connaissance

Nous considérons maintenant un RBC entièrement validé et permettant de répondre à la QCC, présenté en Figure

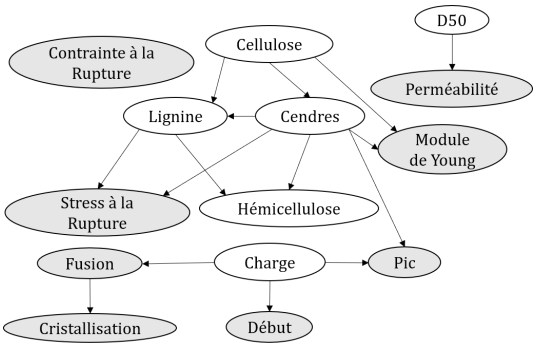

FIGURE 5 – Modèle appris après une itération et critiqué par l'expert en Section 4.2. Les variables de contrôles sont indiquées en blanc, et les variables à expliquer en gris.

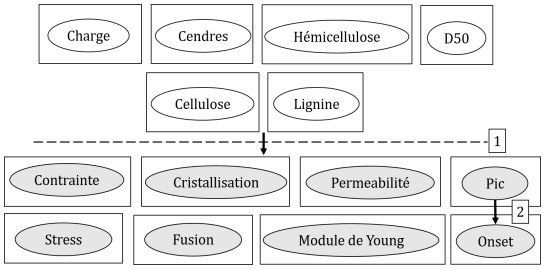

FIGURE 6 – Schéma relationnel finalement retenu après validation experte. Chaque rectangle représente une classe du SR, tandis que chaque ovale représente une variable. Deux types de liens relationnels sont indiqués : (1) indique le lien depuis toutes les classes au dessus de la ligne vers celles en dessous ; (2) indique le lien établi depuis la température de **pic** vers celle du **début**.

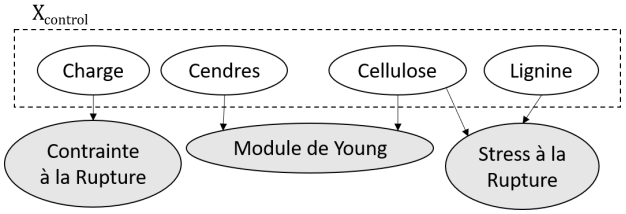

FIGURE 7 – Extrait du RB final sélectionné pour répondre à QCc_{bio}. Puisque toutes les relations représentées sont influencées par des contraintes de précédence expertes, on considère ce modèle *causal*.

	Contrainte à la rupture			
Charge	]0.24 ;0.5]	]0.5 ;0.8]	]0.8 ;1]	]1 ;1.07]
]2 ;4]	0.0076	0.4924	***0.4924****	0.0076
]4 ;11]	0.002	***0.770****	0.1620	***0.0660****
]11 ;21]	0.3624	0.4522	0.1826	0.0028
]21 :50]	***0.5747****	0.2630	0.1071	0.0552

TABLE 2 – Probabilités conditionnelles pour la **contrainte à la rupture**. (***Vraisemblance maximale****).

7, appris à partir du RS présenté en Figure 6. Dans une optique de simplification, nous présentons ici une version simplifiée où toutes les variables non pertinentes ont été retirées. QCc_{bio} nécessite deux interventions pour maximiser l'amélioration des propriétés mécaniques considérées : (1) la **charge de remplissage** et (2) la composition du CL.

4.3.1 Charge Optimale

D'après la Figure 7, la **charge** a un impact sur les propriétés mécaniques uniquement à travers la contrainte à la rupture. À partir de la table des probabilités conditionnelles (Table 2), plusieurs lectures sont possibles dépendant des critères experts élicités :

- Si l'on recherche la plus haute valeur possible de la **contrainte à la rupture** (]1 ;1.07]), les probabilités sont toutes quasi-nulles. Il s'agit donc d'un critère non envisageable réalistiquement.
- En fixant un critère dur visant la deuxième meilleure valeur de la **contrainte à la rupture** (entre 0.8 et 1), une **charge** $\in$]2 ;4] pourrait être considérée : elle garantit en effet une probabilité de 0.49, signifiant qu'en moyenne un produit final sur deux atteindra la valeur souhaitée (et obtiendra une valeur comprise entre 0.5 et 0.8 dans la plupart des autres cas).
- Dans le cas d'un procédé industriel, en revanche, l'expert pourrait souhaiter placer un critère dur sur la probabilité de réussite plutôt que sur la valeur cible, afin de garantir une stabilité des résultats. Dans ce cas, il semble plus raisonnable de considérer une **charge** $\in$]4 ;11], garantissant une probabilité de 0.77 d'obtenir une **contrainte** $\in$]0.5 ;0.8].

4.3.2 Prospections de CLs

D'après le RB présenté en Figure 7, le **module de Young** et le **stress à la rupture** dépendent de la composition du CL considéré (de la teneur en **cendres**, **cellulose** et **lignine**). Dans cette partie, nous utilisons ce résultat pour proposer de nouvelles CLs permettant de maximiser les valeurs de ces paramètres. Pour cela, nous cherchons dans @Web [3], une BC contenant les informations de composition de nombreuses biomasses [13], afin d'en proposer de nouvelles pertinentes. D'une façon similaire au paragraphe précédent, nous introduisons de nouveaux critères d'acceptabilité durs (CD) et souple (CS) :

- CD_1 Pour proposer de nouvelles CLs réalistes, il est important que la somme de ses constituants ne dépasse pas 100 (en d'autres termes, la biomasse simulée doit être possible). En fixant $x \in$ {**Cendres**, **Cellulose**, **Lignine**, **Hémicellulose**} et l'intervalle associé $[x_{min}; x_{max}]$ déterminé par la discrétisation, nous fixons CD_1 tel que $\sum_x x_{min} < 100$.
- CD_2 Nous voulons que les valeurs cibles soient intéressantes : on fixe donc **Module de Young** > 0.8 $\cap$ **Stress à la rupture** > 0.8.
- CD_3 La probabilité de réalisation doit être supérieure à 0.25.
- CS_1 Dans le cas où aucune CL potentielle n'est trouvée, nous voulons étendre la recherche à de potentiels candidats proches de la composition cible. Afin d'évaluer la qualité de tels substituts, nous définissons un score de qualité. Soit une CL m contenue dans @Web, sa composition x_m et un intervalle cible $[x_{min}; x_{max}]$ déterminé par le RB ($x \in$ {**Cendres**, **Cellulose**, **Lignine**}). On définit le score $S_m = \sum_x \sigma(m, x)$ avec $\sigma(m, x) = min(abs(x_m - x_{min}), abs(x_m - x_{max}))$.
  Plus S_m est bas, plus la CL proposée est proche de la recommandation.

La requête effectuée sur @Web retourne quinze résultats, présentés partiellement dans la Table 3. Chacun de ces scénarios évalue la probabilité de succès de CD_2, sachant que l'on respecte CD_1 et CD_3. Parmi les deux scénarios présentés dans cet article, le plus probable ($p = 0.99$) n'est pas une correspondance exacte : la biomasse s'approchant le plus est l'écorce de pin, avec un S-score de 5.26 (dû notamment à son taux de cendres trop bas par rapport à la recommandation). Le second scénario présenté, plus bas en terme de probabilité de réalisation ($p = 0.82$), est quant à lui une correspondance parfaite avec l'enveloppe de riz. Malgré cette différence de probabilité donnant l'écorce de pin comme une réalisation quasi-certaine, l'enveloppe de riz serait néanmoins à privilégier pour les tests. En effet, il est important de garder en tête que, comme introduit plus haut, l'une des limites des RB est son traitement des discrétisations : le comportement autour des valeurs limites peut donc être plus compliqué à prédire. Dans le cadre de l'écorce de pin, nous avons vu que son taux de cendres est trop bas (1.44 en moyenne), ce qui mitige déjà grandement les résultats prédits ; mais sa composition en lignine (27.33 en moyenne) le place tout juste au-dessus de la quantité de lignine recommandée par le modèle, rendant plus incertain son comportement. L'enveloppe de riz, au contraire, présente des compositions plutôt éloignées des limites des catégories de discrétisation. Il semble donc plus sûr de tester

p	0.99	p	0.82
Cendres	[6.7 ;24.7]	Cendres	[6.7 ;24.7]
Cellulose	[10.9 ; 25.6]	Cellulose	[25.6 ;33]
Lignine	[26.4 ; 49]	Lignine	[19.4 ; 26.4]
Exact	$\emptyset$	**Exact**	Enveloppe de Riz
Similaire	Ecorce de Pin	**Similaire**	$\emptyset$
S_{Pin}	5.26	S_{Riz}	0

TABLE 3 – Exemple de résultats correspondants aux critères d'acceptabilité définis, et leur propabilité p de réalisation. Lors de l'absence de correspondance exacte, un S-score a été calculé pour trouver le CL le plus proche de la cible.

	Cendres	Cellulose	Lignine
Ecorce de Pin	1.44	20.6	27.33
Enveloppe de Riz	14.5	31.9	25.7

TABLE 4 – Composition des deux nouvelles charges lignocellosiques candidates.

dans un premier temps ce matériel. Les compositions complètes des deux CLs sont présentées dans la Table 4.

En conclusion, si le choix des discrétisations établies a un sens pour le domaine, il introduit également des biais : l'appartenance d'une valeur à certaines catégories peut parfois être compliqué à distinguer, et certaines catégories semblent ainsi gonflées artificiellement par rapport à d'autres non représentées dans la BC. Cela souligne encore une fois l'importance de la représentativité d'une BC dans l'apprentissage automatique : une plus grande diversité de cas et d'exemples permettrait de limiter ces effets de bords.

5 Conclusion

Dans cet article, nous avons présenté POND, un workflow complet dédié à la réponse de questions expertes sur des bases de connaissance représentant des procédés de transformation modélisés par l'ontologie PO². Nous nous sommes focalisés ici sur la causalité, et les outils offerts par la découverte causale (comme l'ingénierie inverse), en présentant l'introduction de connaissances expertes à différents embranchements de la modélisation. Celle-ci se base sur deux points : l'établissement d'un vocabulaire commun standardisé à travers l'ontologie PO², et la formalisation de connaissances expertes ne pouvant pas être exprimées directement dans une BC car dépendantes du contexte. Nous avons ensuite illustré cette approche à travers une application concrète sur les emballages bio-composites. Grâce à l'ontologie, ce workflow permet à l'expert de facilement manier les connaissances expertes à intégrer d'une part, ainsi que l'ajout et la modification de celles-ci à la volée. Enfin, nous avons défini une formalisation de différentes contraintes expertes permettant de guider la lecture du RB appris afin d'éliciter les réponses les plus intéressantes du point de vue de l'utilisateur. Ainsi, à travers notre illustration, nous avons présenté plusieurs réponses possibles, et identifié de potentiels nouveaux matériaux à tester, encore non testés dans la base d'origine.

Comme dans toute analyse causale, il est important de considérer le contexte dans lequel l'apprentissage a lieu, et que nous avons détaillé en Section 2.3. Dans les travaux présentés ici, nous avons également supposé que l'expert consulté a une connaissance fiable du domaine, et où aucune contradiction n'est considérée (ce qui ne se serait pas forcément vérifié dans le cas d'un panel d'experts où des divergences peuvent se trouver). L'intégration de ces possibles dissensions et leur modélisation afin d'établir l'apprentissage d'un modèle optimal est une piste de recherche que nous souhaitons explorer.

De même, les recommandations établies par le RB et générées de façon automatique permettent d'établir une liste de règles établissant des scénarios plus ou moins crédibles. Par exemple, si l'on regarde la Table 2, il parait hautement improbable qu'une charge élevée (entre 21 et 50) permette d'obtenir une contrainte à la rupture améliorée (avec une probabilité quasi-nulle de 0.06). L'utilisation de ces règles pour évaluer la crédibilité de nouvelles informations ou la pertinence de la BC est une autre piste à étudier dans la poursuite de ces travaux.

Remerciements

Nous tenons à remercier Claire Meyer (Equipe PhyProDiv, INRAE IATE) pour nous avoir fourni les données pour la prospection de biomasses. Le travail présenté dans ce papier a été financé partiellement par l'Agence Nationale de Recherche dans le cadre des projets D2KAB (ANR-18-CE23-0017) et DataSusFood (ANR-19-DATA-0016).

Références

[1] Montassar Ben Messaoud, Philippe Leray, and Nahla Ben Amor. Semcado : A serendipitous strategy for learning causal bayesian networks using ontologies. *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 182–193, 2011.

[2] Giacomo Bucci, Valeriano Sandrucci, and Enrico Vicario. Ontologies and bayesian networks in medical diagnosis. *HICSS*, pages 1–8, 2011.

[3] Patrice Buche, Juliette Dibie-Barthelemy, Liliana L. Ibanescu, and Lydie Soler. Fuzzy Web Data Tables Integration Guided by a Termino-Ontological Resource. *IEEE Transactions on Knowledge and Data Engineering*, 25(4) :805–819, 2013.

[4] Federico Castelletti and Guido Consonni. Discovering causal structures in bayesian gaussian directed acyclic graph models. *Journal of the Royal Statistical Society Series A, Royal Statistical Society*, 183 :1727–1745, 2020.

[5] David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3(null) :507–554, mar 2003.

[6] Gregory F. Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4) :309–347, 1992.

[7] Grégoire David, Giovanna Croxatto Vega, Joshua Sohn, Anna Ekman Nilsson, Arnaud Hélias, Nathalie Gontard, and Helene Angellier-Coussy. Using life cycle assessment to quantify the environmental benefit of upcycling vine shoots as fillers in biocomposite packaging materials. *International Journal of Life Cycle Assessment*, 2020.

[8] C. P. De Campos and Q. Ji. Improving bayesian network parameter learning using constraints. In *ICPR*, pages 1–4, 2008.

[9] C.P. De Campos, Z. Zhi, and Q. Ji. Structure learning of bayesian networks using constraints. In *ICML*, pages 113–120, 2009.

[10] Juliette Dibie, Stéphane Dervaux, Estelle Doriot, Liliana Ibanescu, and Caroline Pénicaud. $[MS]^2O$ - A multi-scale and multi-step ontology for transformation processes : Application to micro-organisms. In *ICSS*, pages 163–176, 2016.

[11] Zhongli Ding, Yun Peng, and Rong Pan. *BayesOWL : Uncertainty Modeling in Semantic Web Ontologies*, pages 3–29. Springer Berlin Heidelberg, 2006.

[12] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. In *SEMANTiCS (Posters, Demos, SuCCESS)*, 2016.

[13] Charlène Fabre, Patrice Buche, Xavier Rouau, and Claire Mayer-Laigle. Milling itineraries dataset for a collection of crop and wood by-products and granulometric properties of the resulting powders. *Data in Brief*, 33, 2020.

[14] Stefan Fenz. Exploiting experts' knowledge for structure learning of bayesian networks. *Data And Knowledge Engineering*, 73 :73 – 88, 2012.

[15] Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. Learning probabilistic relational models. In *IJCAI*, page 1300–1307. Morgan Kaufmann Publishers Inc., 1999.

[16] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10 :524, 2019.

[17] Alain Hauser and Peter Bühlmann. Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, pages 926—-939, 2014.

[18] Liliana Ibanescu, Juliette Dibie, Stéphane Dervaux, Elisabeth Guichard, and Joe Raad. Po2- a process and observation ontology in food science. application to dairy gels. *Metadata and Semantics Research*, pages 155–165, 2016.

[19] Abdul-Wahid Mohammed. Knowledge-oriented semantics modelling towards uncertainty reasoning. *SpringerPlus*, 5, 2016.

[20] Melanie Munch, Patrice Buche, Cristina Manfredotti, Pierre-Henri Wuillemin, and Helene Angellier-Coussy. A process reverse engineering approach using Process and Observation Ontology and Probabilistic Relational Models : application to processing of biocomposites for food packaging. In *15th International Conference on Metadata and Semantics Research*, Madrid, Spain, November 2021.

[21] Melanie Munch, Juliette Dibie, Pierre-Henri Wuillemin, and Cristina E. Manfredotti. Towards interactive causal relation discovery driven by an ontology. In *FLAIRS*, pages 504–508, 2019.

[22] Melanie Munch, Juliette Dibie-Barthélemy, Pierre-Henri Wuillemin, and Cristina E. Manfredotti. Interactive causal discovery in knowledge graphs. In *Semex@ISWC 2019*, volume 2465 of *CEUR Workshop Proceedings*, pages 78–93. CEUR-WS.org, 2019.

[23] Melanie Munch, Pierre-Henri Wuillemin, Cristina Manfredotti, Juliette Dibie, and Stephane Dervaux. Learning probabilistic relational models using an ontology of transformation processes. In *OTM 2017 Conferences*, pages 198–215, 2017.

[24] Mélanie Munch, Patrice Buche, Stéphane Dervaux, and Hélène Angellier-Coussy. Itinerary Description for biocomposites from poly(3-hydroxybutyrate-co-3-hydroxyvalerate) and lignocellulosic fillers, 2021.

[25] Pekka Parviainen and Mikko Koivisto. Finding optimal bayesian networks using precedence constraints. *Journal of Machine Learning Research*, 14 :1387–1415, 2013.

[26] Judea Pearl. *Causality : Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.

[27] Gideon Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2) :461 – 464, 1978.

[28] Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G Dimakis, and Sriram Vishwanath. Learning causal graphs with small interventions. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[29] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

[30] Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, and Mariano Fernández-López. The neon methodology for ontology engineering. In Mari Carmen Suárez-Figueroa, Asunción Gómez-Pérez, Enrico Motta, and Aldo Gangemi, editors, *Ontology Engineering in a Networked World*, pages 9–34. Springer, 2012.

[31] Louis Verny, Nadir Sella, Séverine Affeldt, Param Singh, and Hervé Isambert. Learning causal networks with latent variables from multivariate information in genomic data. *PLOS Computational Biology*, 13, 2017.

[32] Shenyong Zhang, Yi Sun, Yun Peng, and Xiaopu Wang. Bayesowl : A prototype system for uncertainty in semantic web. *ICAI*, 2 :678–684, 2009.