



HAL
open science

Approche centrée données massives et intelligence artificielle pour l'évaluation des systèmes industriels durables et circulaires : application à la valorisation des déchets lignocellulosiques

Nancy Prioux

► To cite this version:

Nancy Prioux. Approche centrée données massives et intelligence artificielle pour l'évaluation des systèmes industriels durables et circulaires : application à la valorisation des déchets lignocellulosiques. Génie chimique. Institut National Polytechnique de Toulouse - INPT, 2023. Français. NNT : 2023INPT0066 . tel-04865205

HAL Id: tel-04865205

<https://theses.hal.science/tel-04865205v1>

Submitted on 6 Jan 2025

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

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



Université
de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Institut National Polytechnique de Toulouse (Toulouse INP)

Discipline ou spécialité :

Génie Industriel

Présentée et soutenue par :

Mme NANCY PRIOUX

le jeudi 5 janvier 2023

Titre :

Approche centrée données massives et intelligence artificielle pour
l'évaluation des systèmes industriels durables et circulaires : application à
la valorisation des déchets lignocellulosiques

Ecole doctorale :

Systèmes (EDSYS)

Unité de recherche :

Laboratoire de Génie Chimique (LGC)

Directeurs de Thèse :

M. GILLES HETREUX

M. JEAN PIERRE BELAUD

Rapporteurs :

M. DOMINIQUE MILLET, UNIVERSITE DE TOULON

M. JEAN-MARC COMMENGE, UNIVERSITE LORRAINE

Membres du jury :

M. PASCAL FLOQUET, TOULOUSE INP, Président

M. GILLES HETREUX, TOULOUSE INP, Membre

M. JEAN PIERRE BELAUD, TOULOUSE INP, Membre

MME CAROLINE SABLAYROLLES, TOULOUSE INP, Invitée

MME VALÉRIE LAFOREST, ECOLE NLE SUP DES MINES SAINT ETIENNE, Membre



Remerciements

La Science aime les analogies et les comparaisons alors je vais me permettre de comparer les trois ans de cette thèse à quelque chose qui me calme, un long voyage en mer. Comme un voilier en pleine tempête qui continue à avancer, il m'a fallu traverser le COVID, tel un marin qui fait avancer son voilier sans bouger, et une cyberattaque avec toutes les émotions qui en ont découlé. Mais il y a eu d'un autre côté un voyage calme et paisible plein de joie, de soulagement des fois et de fierté. Et surtout un voyage en mer ne serait rien sans les personnes qui nous accompagnent pour une partie plus ou moins longue du voyage ou que l'on croise lors de brèves escales. Ainsi, je ne saurais finir ce voyage sans les remercier et leur témoigner une sincère reconnaissance. Je vais commencer ces remerciements par la fin de ce voyage. J'adresse tous mes remerciements à l'ensemble des membres de mon jury de thèse. Mes remerciements aux Pr Dominique Millet et Jean-Marc Commenge pour leur rapport complet et étayé ainsi que pour nos échanges constructifs pendant ma soutenance. Merci à Valérie Laforest pour l'intérêt porté à ces travaux et sa participation constructive pendant ma soutenance. Et ce fut un honneur de vous avoir en tant que président du jury Pr Pascal Floquet. Merci d'avoir accepté de finir ce voyage avec moi comme vous l'avez commencé il y a 10 ans en tant que professeur et directeur de filière lors de mes années d'école d'ingénieur en restant simple, brillant et avec une pointe d'humour. Mes remerciements à Gilles Hétreux pour avoir accepté d'être mon directeur de thèse et donc permis à ce que ces travaux voient le jour.

Je tiens à remercier les trois personnes qui m'ont fait confiance, accompagné et guidé au quotidien depuis l'avant-thèse et mon recrutement en tant qu'ingénieur d'études INRA au Laboratoire de Chimie AgroIndustrielle : mon directeur de thèse Jean-Pierre Belaud, Claire Vialle et Caroline Sablayrolles. Claire et Caroline, merci pour vos conseils, votre présence, nos discussions, votre présence, votre attention et votre bienveillance. Jean-Pierre, je ne peux que trop mal écrire ce que je te dois alors merci, merci pour être toujours là quelque soit le problème et quelle que soit l'heure, merci de m'avoir apporté tant de choses.

J'aimerais également remercier les membres du Laboratoire de Génie Chimique, leur gentillesse et leur proximité. Un grand merci tout particulier aux professeurs, aux enseignants-chercheurs et aux chercheurs du département Procédés et Systèmes Industriels « PSI ». Certains ont été mes professeurs pendant mon école d'ingénieur et j'ai eu grand plaisir de les côtoyer en tant que collègues pendant ma thèse. Merci en particulier à Rachid pour toute son aide sur la partie IA, sa gentillesse et sa présence, merci à Stéphane, Rachid, Alessandro, Ludovic, Pascal, Marianne, Vincent et j'en oublie. Je remercie également Patricia Uliana, Danièle « Dany » Bouscary et Angélique Pasoti pour tous les aspects administratifs de la thèse. Merci beaucoup à toi Alain Philip de toujours veiller à ce que chaque doctorant se sente chez lui au laboratoire. Pendant ce voyage, il y a eu de nombreux navires qui ont navigué avec moi avec des fois plus de tempêtes que j'en ai eu et je ne pourrais faire en quelques lignes leur éloge ce sont les futurs docteurs (ou docteurs) qui travaillaient sur divers sujets et qui sont devenus des amis. D'abord merci à Yohann qui était là pendant ces trois ans et avec

qui j'ai partagé les cours et examens en visio pendant la pandémie, avec qui ont discuté de tout et de rien (souvent de tennis, de coloc et de nos étudiants quand même). Merci à tous ceux qui ont partagé mon bureau et qui m'ont fait rire plus d'une fois : Amaya, Asnakech, Erry, Fanny, Guilia, Jana, Matteo, Tina, Yoshra et Vincent. Merci aux « anciens » qui m'ont intégré, fait rire et dont on a partagé tant de bons moments : Adriana, Carlos, Claire, Eduardo, Emilien, Florian, Kalyani, Lise, Manu, Pierre (avec qui j'ai aussi passé de très bons moments à l'A7), Margot, Michelle (nos discussions cuisines me manquent!), Milad, Pauline et Sergio. Et merci aux nouveaux de perdurer la bonne ambiance qu'il y a au labo (et bon courage pour votre fin de voyage) : Abby, Axelle, Boris, Claire, Elsi, Erwann, Éva, Daniel, Gaëlle, Igor, Michele, Mukhlis, Niki, Noha-Lys, Paul, Pauline, Thomas, Vivien et tous ceux que j'ai oublié de citer.

Je me dois aussi de remercier mes amis hors thèse qui sont toujours là, qui m'ont sorti quand j'en avais besoin, qui me supportent (parfois depuis bien longtemps) et qui ont fait de moi ce que je suis aujourd'hui. Alors merci à mes amis de lycée qui sont toujours là 15 ans après, Angéline, Aurélien, Claire, Juliette, Quentin, Rémi et Sarah. Merci à mes amis de l'A7 : Annabelle, ton rire, nos discussions, ta présence, ta constance, toi, Maxime avec qui j'ai partagé une colocation pendant de nombreuses années et une amitié encore plus longue, Pauline pour être toi et pour ton amitié encore maintenant malgré ta nouvelle famille et les kilomètres, Cédric, Luna, Nathan, Valentin et Xavier pour nos soirées JDS/R, nos vacances et tous nos fous rires et cette amitié, Mathilde et Sébastien, Thibaut, Pim et Aliénor, pour tous ces moments de partage, de rire et de joie. Et enfin merci aux amis que j'ai connu à Toulouse dans des circonstances des fois loufoques : Ib' pour être toi et tout ce que tu m'apportes, Flo' pour nos discussions à la fois sérieuses et non, Laura pour tout ce qu'on partage en commun et en différence et nos partages d'anecdotes à n'en plus finir, Nina pour être toujours là pour que je travaille mes réflexes, Marine pour être toi (et pas pour la bibliothèque), Agathe, Aurel', Chroux, Élise, Ludo', Nico', Nicochou, Sofia, Stu', Wanou, Zoé, et tous les autres que j'ai oubliés. Et merci aux framers d'exister.

Je finis ces remerciements par toute ma famille dont aucun n'est scientifique, dont aucun ne comprend vraiment ce que je fais, mais qui ont toujours été là : mes grands-parents que je n'appelle pas assez, mon oncle et ma tante qui m'ont hébergé et subit pendant mes deux années de prépa (et ce n'était pas toujours la joie), mon cousin qui malheureusement pour lui soutient le Racing 92 (rien ne vaut le ST), ma cousine qui a gentiment « accepté » de partager sa chambre pendant deux ans (et dont j'ai eu le plaisir de lui faire découvrir que les sciences ce n'est pas Voldemort). Merci à ma sœur qui a toujours été là même si nous sommes le yin et le yang. Je ne saurais comment remercier mes parents qui m'ont toujours poussé à faire ce que j'avais envie et soutenue dans tout ce que je fais, qui ont toujours fait passer ma sœur et moi avant tout et qui ont toujours pousser à me poser des questions (quitte à m'acheter livre sur livre quand ils n'avaient pas les réponses ce qui est arrivé assez tôt...).

Je ne pourrais finir ces remerciements par dire pour toutes les personnes que j'ai oublié et qui m'ont supporté pendant quelques heures, jours, mois ou années : "*I dedicate this thesis to the Muggles who got me here.*"

Résumé

Ces dernières années ont vu l'émergence de plans d'action pour le climat, le développement durable ou l'environnement comme le plan européen « *Green Deal* » de janvier 2020. L'un des principaux plans d'action, intitulé "Plan d'action pour l'économie circulaire", propose d'utiliser le modèle d'économie circulaire pour simuler l'utilisation de modèles durables. Dans ce contexte, les procédés et systèmes industriels se veulent plus durables et circulaires. Pour évaluer cette durabilité, de nombreuses approches existent et nécessitent un besoin de données, d'informations et de connaissances croissant. Par la transition digitale et l'émergence des données massives, de nombreuses données sont disponibles pour les industries. Qu'elles soient structurées ou non, sous différents formats (PDF, XML, HTML), et de différents types (Web, bases de données, articles scientifiques), les sources de données sont potentiellement des mines de connaissances qui peuvent compléter les analyses de la durabilité. Allant souvent de pair avec les données massives, l'intelligence artificielle apporte des méthodes pour extraire, traiter, analyser et visualiser ces données. De nombreux outils et méthodes sont désormais disponibles à cette fin grâce aux techniques du « *Big Data* » ou de l'intelligence artificielle comme l'apprentissage automatique. L'évaluation des systèmes industriels durables et circulaires peut bénéficier de ces technologies.

Évaluer la durabilité d'un système ou d'un procédé regroupe maintenant un ensemble de méthodes en interaction les unes avec les autres. Cela s'explique par la quantité des données disponibles, la complexité grandissante des procédés industriels et des analyses de la durabilité plus globales.

Les travaux de cette thèse s'intéressent à définir une approche centrée données massives et intelligence artificielle pour l'évaluation des systèmes industriels durables et circulaires. Un cadre méthodologique est développé afin de profiter des articles scientifiques et du web public, d'analyser ces données et de fournir un support pour aider à l'interprétation des résultats. Cette approche induit le couplage de quatre domaines d'intérêt : données massives, intelligence artificielle, analyses de la durabilité et procédés et systèmes industriels. La spécification de ce cadre permet de reconsidérer la façon d'exécuter les analyses de la durabilité. Il donne lieu à son application dans l'agro-industrie. Différentes trajectoires de valorisation des déchets lignocellulosiques sont étudiées selon deux principales dimensions, les intrants de type biomasse et les technologies de procédés. Ce cas d'étude soutient l'ingénierie circulaire et met à l'épreuve le cadre méthodologique proposé.

Mots-clés : Procédés Industriels, Données massives, Développement Durable, Economie Circulaire, Pensée Cycle de Vie, Apprentissage automatique.

Abstract

The last years have seen the emergence of action plans for climate, sustainable development or the environment such as the European "Green Deal" plan of January 2020. One of the main action plans, entitled "Circular Economy Action Plan", proposes to use the circular economy model to simulate the use of sustainable models. In this context, industrial processes and systems are intended to be more sustainable and circular. To assess this sustainability, many approaches exist and require an increasing need for data, information and knowledge. With the digital transition and the emergence of big data, a lot of data is available for industries. Whether structured or not, in different formats (PDF, XML, HTML), and of different types (Web, databases, scientific articles), data sources are potentially wealth of knowledge that can complement sustainability analyses. Often going hand-in-hand with massive data, artificial intelligence brings methods to extract, process, analyze and visualize these data. Many tools and methods are now available for this purpose thanks to "Big Data" techniques or artificial intelligence such as machine learning. Sustainability evaluation and circular industrial systems can benefit from these technologies. Assessing the sustainability of a system or process now brings together a set of methods that interact. This is due to the amount of data available, the increasing complexity of industrial processes, and more comprehensive sustainability analyses.

The work of this thesis is interested in defining a big data and artificial approach driven approach for the evaluation of sustainable and circular industrial systems. A framework is developed in order to take advantage of scientific articles and the public web, to analyze these data and to provide support for the interpretation of the results. This approach induces the coupling of four domains of interest : massive data, artificial intelligence, sustainability analysis and industrial processes and systems. The specification of this framework allows to reconsider the way to perform sustainability analyses. It gives rise to its application in agribusiness. Different trajectories of lignocellulosic waste valorization are studied along two main dimensions, biomass inputs and process technologies. This case study supports circular engineering and tests the proposed methodological framework.

Key-words : Industrial Engineering, Big Data, Sustainability, Circular Economy, Life Cycle Thinking, Machine Learning

Table des matières

Table des matières	ix
Table des figures	xiii
Liste des tableaux	xvii
Préambule	xix
Introduction générale	1
1 Contexte général des travaux	5
1.1 Le développement durable	7
1.2 La transition digitale au service du développement durable	16
1.3 Objectifs et questions scientifiques	22
1.4 Organisation du manuscrit	25
2 État de l'art	27
2.1 Données massives et intelligence artificielle	29
2.2 La pensée cycle de vie pour le développement durable	41
2.3 Développement durable, intelligence artificielle et données massives	47
2.4 Procédés et systèmes industriels et développement durable	50
2.5 Conclusion du chapitre 2	53
3 Définition du cadre méthodologique	55
3.1 Introduction du chapitre 3	57
3.2 But et frontières de l'étude	60
3.3 Architecture des données	61
3.4 Analyse de la durabilité	63
3.5 Analyse et visualisation des résultats	72
3.6 Décision	73
3.7 Conclusion du chapitre 3	73
4 Architecture des données	75
4.1 Introduction du chapitre 4	77
4.2 Collection et extraction des données	78
4.3 Enrichissement et stockage des données extraites	89

4.4	Traitement des données extraites	90
4.5	Analyse des données	90
4.6	Visualisation des données	98
4.7	Conclusion du chapitre 4	98
5	Analyse et visualisation des résultats de l'évaluation de la durabilité	101
5.1	Introduction du chapitre 5	103
5.2	Catégories d'impacts et méthodes	103
5.3	Apprentissage non-supervisé pour l'analyse des résultats de l'évaluation de la durabilité	106
5.4	Positionnement multidimensionnel et k-moyennes	112
5.5	Conclusion sur le chapitre 5	114
6	Cas d'étude : valorisation des déchets lignocellulosiques	115
6.1	Introduction du chapitre 6	117
6.2	Industrie agricole 4.0	118
6.3	Présentation du cas d'étude	121
6.4	Cadre méthodologique appliqué au cas d'étude	124
6.5	Création d'un outil-prototype	142
6.6	Limites du cas d'étude	147
6.7	Conclusion du chapitre 6	147
	Conclusion générale et perspectives des travaux	149
A	Descriptions des principales évaluations environnementales.	III
A.1	Analyse environnementale du cycle de vie	III
A.2	Analyse des flux de matières	IV
A.3	Analyse des impacts environnementaux	IV
A.4	Analyse des risques environnementaux	V
A.5	Analyse des exergies	V
A.6	Apport de matière par unité de service	VI
B	Descriptif détaillé de la RTO utilisée pour extraire les données de la littérature.	VII
C	Liste des articles sélectionnés à titre d'experts avant traitement.	XI
D	Données expérimentales pour le modèle à titre d'expert pour l'énergie d'un broyage.	XIX
E	Visualisations des rendements en glucose prédits par différents modèles.	XXI
F	Résultats des évaluations des impacts du cycle de vie méthode <i>ReCiPe 2016 Endpoint Hierarchical</i>.	XXVII
G	Cartes MDS des quatre dimensions des procédés et des impacts.	XXXV

H	Guide d'utilisation simple de l'outil.	XLI
H.1	Menu principal	XLI
H.2	Extraction des données	XLII
H.3	Données d'arrière-plan (<i>Background data</i>)	XLIV
H.4	Évaluation des impacts environnementaux et économiques	XLIV
H.5	Visualisation des résultats	XLVII
I	Liste des acronymes	XLIX
	Production scientifique	LIII
	Références bibliographiques	LVII

Table des figures

1	Feuille de route du manuscrit.	3
1.1	Positionnement du chapitre 1 dans le manuscrit.	5
1.2	Diagramme de Venn traditionnel du développement durable (adapté de [PURVIS et al., 2019]).	7
1.3	Représentation des mots clés retrouvés dans les articles scientifiques traitant du développement durable entre 2016 et 2021 (dégradé de couleur du bleu pour 2016 au jaune pour 2021) grâce à VOSviewer.	10
1.4	L'économie circulaire selon trois domaines d'action et sept piliers (source : [ADEME, 2014]).	11
1.5	Évaluation, principe, critère et indicateur (adaptée de [MENDOZA et al., 2000]).	13
1.6	Diagramme de présentation du concept de l'Internet de Tout.	18
1.7	Cycle en V adapté de [AFIS, 2004].	20
1.8	Schéma des étapes du cycle de vie du glucose avec le détail des différents systèmes en interaction.	22
1.9	Contexte général de la thèse et questions scientifiques associées.	23
1.10	Questions scientifiques, objectifs et chapitres associés.	24
1.11	Contexte et domaines de notre approche.	25
2.1	Positionnement du chapitre 2 dans le manuscrit.	27
2.2	Science des données vs intelligence artificielle vs données massives vs apprentissage automatique (adapté de [KULIN et al., 2021]).	29
2.3	Pyramide présentant la hiérarchie DIKW (données-information-connaissance-sagesse) (adaptée de [ROWLEY, 2007]).	30
2.4	Schéma décrivant les trois différentes types d'intelligence artificielle (adapté de [KAPLAN et HAENLEIN, 2019]).	34
2.5	Taxonomie des algorithmes d'apprentissage automatique (source : https://cleanpng.com/Rusbert).	35
2.6	Cycle de vie d'un produit dans un contexte d'économie circulaire (source : chaire ELSA-PACT).	42
2.7	La méthode ACV : les quatre étapes et les itérations.	45
2.8	Contexte, domaines, méthodes et outils de notre approche.	54
3.1	Positionnement du chapitre 3 dans le manuscrit.	55
3.2	Les trois domaines d'intérêt du cadre méthodologique.	58

3.3	Schéma général du cadre méthodologique.	59
3.4	Enchaînement des sous-étapes de l'étape d'analyse de la durabilité avec les bases de données associées.	64
3.5	Les trois niveaux d'agrégation des impacts environnementaux.	67
3.6	Étapes du cadre méthodologique et chapitre du manuscrit associé.	74
4.1	Positionnement du chapitre 4 dans le manuscrit.	75
4.2	Cycle de vie de la donnée.	77
4.3	Représentation de relations entre concepts et de l'instanciation de ces relations entre instances.	80
4.4	Représentation de relations entre concepts et de l'instanciation de ces relations entre instances.	81
4.5	Schéma de l'extraction des données d'intérêt des articles scientifiques (adapté de [LOUSTEAU-CAZALET et al., 2016]).	83
4.6	Les principales étapes de l'annotation sémantique d'un tableau guidée par la RTO (adapté de [BUCHE et al., 2013a]).	86
4.7	Ensemble flou correspondant au terme "très fiable" défini sur Θ avec 5 éléments issu de [LOUSTEAU-CAZALET et al., 2016].	88
4.8	Les différentes bases de données lors de l'étape d'Architecture des données.	90
4.9	Schéma général d'un neurone artificiel formel (issu de [CHEKROUN, 2009]).	92
4.10	Graphique présentant la fonction sigmoïde dans l'intervalle [-5; 5].	93
4.11	Illustration du processus de la dépendance partielle jusqu'au PDP (adapté de [BOEHMKE et GREENWELL, 2019]).	97
4.12	Courbes ICE centrées (a) et non-centrées (b) (en noir) avec la courbe PDP (en rouge).	97
5.1	Positionnement du chapitre 5 dans le manuscrit.	101
5.2	Aperçu des catégories d'impact couvertes par la méthode ReCiPe 2016 Midpoint et de leur relation avec les zones de protection issu de [HUIJBREGTS et al., 2017]. La ligne en pointillé signifie qu'il n'y a pas de facteur constant entre le milieu et la fin pour les ressources fossiles.	105
5.3	Étape d'Analyse des résultats avec l'utilisation des méthodes de réduction de dimension et de <i>clustering</i>	106
5.4	Schéma explicatif du fonctionnement du <i>clustering</i> , de sa définition et de la définition du centroïde.	110
5.5	Arbre des algorithmes usuels de <i>clustering</i>	111
6.1	Positionnement du chapitre 6 dans le manuscrit.	115
6.2	Présentation de la composition globale de la biomasse lignocellulosique issu de [ALPER et al., 2020]	122
6.3	Schéma général d'une bioraffinerie de production de glucose ou d'éthanol à partir de la biomasse lignocellulosique	124
6.4	Cadre méthodologique général.	124

6.5	Schéma des frontières du système pour le cycle de vie d'une bioraffinerie ayant pour entrée la biomasse lignocellulosique [PRIOUX et al., 2021]	125
6.6	Étapes du cycle de vie du glucose avec le détail des différents procédés de prétraitement étudiés.	126
6.7	Descriptif des six types d'enchaînement d'opérations unitaires.	127
6.8	Procédé de prétraitement issu de l'article scientifique de [LIU et al., 2013]	128
6.9	Capture d'écran du tableau « description du procédé » d'@Web de l'article de [LIU et al., 2013]	129
6.10	Relation <i>n-aire</i> "Quantité de constituant solide issue d'un broyage"	129
6.11	Schéma général de la régression sous le logiciel Orange.	132
6.12	Visualisation des rendements en glucose prédits par le modèle Réseaux de Neurones Artificiels (RN) en fonction des rendements en glucose réel (valeurs normalisées)	133
6.13	Graphiques centrés de dépendance partielle (en rouge) et d'espérance conditionnelle individuelle (en noir).	134
6.14	Représentation de la matrice [<i>type de procédés</i> × <i>biomasse</i> × <i>impacts</i>] pour les impacts <i>endpoint</i> de la méthode ReCiPe 2016 Endpoint Hierarchical	136
6.15	Étapes du cadre méthodologique avec les boucles d'itération et la volumétrie des données	137
6.16	Indicateurs de dommages des procédés de prétraitement calculés par la méthode Re-CiPe 2016 Endpoint	138
6.17	Diagramme de dispersion de la projection MDS (deux dimensions) et du <i>clustering k-means</i> basé sur la matrice de distance des impacts avec la représentativité entre chaque dimension	139
6.18	Diagramme de dispersion de la projection MDS (deux dimensions) et du <i>clustering k-means</i> basé sur la matrice de distance des procédés avec la représentativité entre chaque dimension	141
6.19	Schéma de fonctionnement de l'outil interne créé pour valider le cadre méthodologique	143
6.20	Architecture des modules VBA-Excel.	144
6.21	Menu principal de l'outil Excel.	145
6.22	Fenêtre du nettoyage des données procédés (ou données de premier-plan).	145
6.23	Vue d'un onglet de résultats de l'évaluation environnementale.	146
6.24	Vue d'une visualisation interactive des résultats de l'évaluation environnementale. . . .	146
6.25	Questions scientifiques, contributions et chapitres associés.	150
6.26	Feuille de route du manuscrit, augmentée des contributions scientifiques (S) et techniques (T) apportées par les travaux de thèse.	151
6.27	Différents coûts nécessaire à une analyse des coûts du cycle de vie	153
B.1	Descriptif de la ressource ontologique et terminologique spécialisée dans la bioraffinerie	IX
E.1	Visualisation des rendements en glucose prédits par le modèle AdaBoost en fonction des rendements en glucose réel (valeurs normalisées).	XXII
E.2	Visualisation des rendements en glucose prédits par le modèle <i>Gradient Boosting</i> en fonction des rendements en glucose réel (valeurs normalisées).	XXIII

E.3	Visualisation des rendements en glucose prédits par le modèle <i>Random Forest</i> en fonction des rendements en glucose réel (valeurs normalisées).	XXIV
E.4	Visualisation des rendements en glucose prédits par le modèle de régression linéaire en fonction des rendements en glucose réel (valeurs normalisées).	XXV
G.1	Cartes MDS des quatre dimensions projetées sur les procédés	XXXVII
G.2	Cartes MDS des quatre dimensions projetées sur les impacts	XXXVIII
G.3	Représentativité cumulée des huit premières dimensions pour les projections sur les impacts et les procédés de la méthode MDS	XXXIX
H.1	Menu principal de l'outil Excel.	XLI
H.2	Fenêtre de la partie "Extraction des données procédés" du menu principal.	XLII
H.3	Fenêtre de choix des fichiers pour une nouvelle extraction des données procédés.	XLII
H.4	Fenêtre du nettoyage des données procédés (ou données de premier-plan).	XLIII
H.5	Vue d'un onglet "Données de premier plan" contenant les données de premier-plan nettoyées.	XLIV
H.6	Description de la partie de données d'arrière-plan du menu principal.	XLIV
H.7	Description de la partie "évaluation des impacts économiques et environnementaux" du menu principal.	XLV
H.8	Vue de l'onglet "Inventaire" contenant les résultats de l'inventaire.	XLV
H.9	Vue d'un onglet de résultats de l'évaluation environnementale.	XLVI
H.10	Vue d'une visualisation interactive des résultats de l'évaluation environnementale.	XLVII

Liste des tableaux

2.1	Liste de méthodes utilisant la pensée "cycle de vie" adaptée de [JOLLIET et al., 2010] . . .	43
3.1	Indices de dispersion selon les différents scores de la matrice de Pedigree (selon la base EcoInvent [®])	70
3.2	Matrice de Pedigree d'après la base EcoInvent [®] [FRISCHKNECHT et al., 2005]	71
5.1	Liste (non exhaustive) de méthodes de calculs des impacts suivant leur orientation . . .	104
5.2	Liste des catégories d'impacts <i>midpoint</i> de la méthode ReCiPe 2016 selon [HUIJBREGTS et al., 2017]	105
6.1	Classement de la biomasse et sa génération adapté de [PTASINSKI, 2016].)	121
6.2	Classement de la biomasse se la génération adapté de [PTASINSKI, 2016))	127
6.3	Nombre d'expériences-types suivant le niveau de granularité de l'étude, le type de procédé et la biomasse	130
6.4	Liste des impacts <i>midpoint</i> de la méthode ReCiPe 2016 Endpoint Hierarchical , abréviations et unité	135
C.1	Liste des articles sélectionnés à titre d'experts avant traitement des procédés de type PM. XII	
C.2	Liste des articles sélectionnés à titre d'experts avant traitement des procédés de type PM-UFM.	XIII
C.3	Liste des articles sélectionnés à titre d'experts avant traitement des procédés de type PM-PC-PS (1/3).	XIV
C.4	Liste des articles sélectionnés à titre d'experts avant traitement des procédés de type PM-PC-PS (2/3).	XV
C.5	Liste des articles sélectionnés à titre d'experts avant traitement des procédés de type PM-PC-PS (3/3).	XVI
C.6	Liste des articles sélectionnés à titre d'experts avant traitement des procédés des type PM-PC-EX-PS, PM-PC-UFM et PM-PC-UFM-PS (non sélectionnés).	XVII
D.1	Énergies de broyage moyenne, minimale et maximale en fonction de la taille de la biomasse en sortie.	XX
E.1	Catégories d'impacts <i>midpoint</i> avec la méthode ReCiPe Endpoint , leur abréviations et leur unité.	XXVIII

E2	Évaluation des impacts <i>midpoint</i> du cycle de vie avec la méthode <i>ReCiPe 2016 Endpoint Hierarchical</i> (P1 à P14).	XXIX
E3	Évaluation des impacts <i>midpoint</i> du cycle de vie avec la méthode <i>ReCiPe 2016 Endpoint Hierarchical</i> (P15 à P30).	XXX
E4	Évaluation des impacts <i>midpoint</i> du cycle de vie avec la méthode <i>ReCiPe 2016 Endpoint Hierarchical</i> (P31 à P46).	XXXI
E5	Évaluation des impacts <i>midpoint</i> du cycle de vie avec la méthode <i>ReCiPe 2016 Endpoint Hierarchical</i> (P47 à P62).	XXXII
E6	Évaluation des impacts <i>midpoint</i> du cycle de vie avec la méthode <i>ReCiPe 2016 Endpoint Hierarchical</i> (P63 à P78).	XXXIII
G.1	Numéro des groupes issus de la méthode hybride MDS - <i>k-means</i> par procédé et référence bibliographique.	XXXVI

Préambule

Tout au long de ce manuscrit, les encadrés suivants sont utilisés pour mettre en lumière certaines informations telles que :



Les définitions importantes



Les problématiques des travaux



Les remarques venant éclairer le discours



Les exemples d'application pratiques d'une notion théorique

Introduction générale

Un message n'est pas délivré tant qu'il n'est pas compris.

Robin Hobb - *L'assassin royal* (1999)

L'industrie est au coeur de deux grands changements :

- la transition digitale avec l'avènement parmi d'autres des données massives, de l'intelligence artificielle et de l'internet des objets.
- l'intégration du développement durable dans les processus.

Tendre vers une industrie plus durable signifie réinventer l'ensemble des chaînes d'approvisionnement et de valeur ainsi que tous les processus de l'industrie. Les procédés de transformation de la matière deviennent plus durables et des solutions innovantes voient le jour. La prise en compte non plus du procédé de transformation, mais du système et son cycle de vie évite aux industries de déporter des impacts à d'autres lieux ou moments. Les procédés cherchent donc à se renouveler, à devenir plus durables et plus flexibles. Cette flexibilité est, aujourd'hui, en pleine expansion grâce aux outils venant de la transition digitale. Tous ces outils et technologies reposent sur le concept d'Industrie 4.0.

L'Industrie 4.0 voit de nombreux nouveaux métiers apparaître comme des experts de la durabilité ou des *data scientists*. Il est difficile pour des métiers venant de domaines aussi différents de communiquer. Or, il serait bénéfique qu'ils le fassent pour aider les décideurs à concevoir des procédés durables et circulaires. L'expert de la durabilité a tout intérêt à utiliser des outils des *data scientists* pour extraire et analyser ses données. L'Industrie 4.0 crée de nombreuses données exploitables plus ou moins facilement. Il est alors intéressant d'utiliser ces données pour servir à la durabilité des industries. Un autre domaine crée de la donnée : la recherche à travers les nombreux articles scientifiques publiés. Ces données souvent textuelles ne sont pas facilement exploitables, mais sont une grande source d'informations. Les industries pour concevoir des procédés plus durables peuvent alors analyser ce qui se fait dans la littérature scientifique. Ces travaux proposent une approche centrée données massives et intelligence artificielle pour l'évaluation des systèmes industriels durables et circulaires.

La problématique principale abordée dans ce manuscrit concerne la gestion des nombreuses données décrivant les procédés et l'apport de ces données dans les analyses et les évaluations de la durabilité grâce aux outils et méthodes issus des données massives et de l'intelligence artificielle. La finalité est de fournir un cadre méthodologique d'aide à la décision pour la conception d'un procédé durable et circulaire. Ce cadre sera testé pour le cas de la transformation de la biomasse

lignocellulosique. Ces travaux sont labélisés par le conseil scientifique de l'EUR BioEco (*Biotechnology Building Bio-based Economy*, <https://bioeco.univ-toulouse.fr/fr/accueil/>) qui s'inscrit dans le programme d'Investissements d'Avenir géré par l'Agence Nationale de la Recherche (ANR-18-EURE-0021). Ces travaux ont donné lieu à diverses actions de dissémination scientifique (Annexe Production Scientifique).

Le manuscrit est composé de six chapitres, chacun illustré sur la figure 1. Cette figure servira de feuille de route tout au long du manuscrit, afin de spécifier les contributions apportées par chaque chapitre.

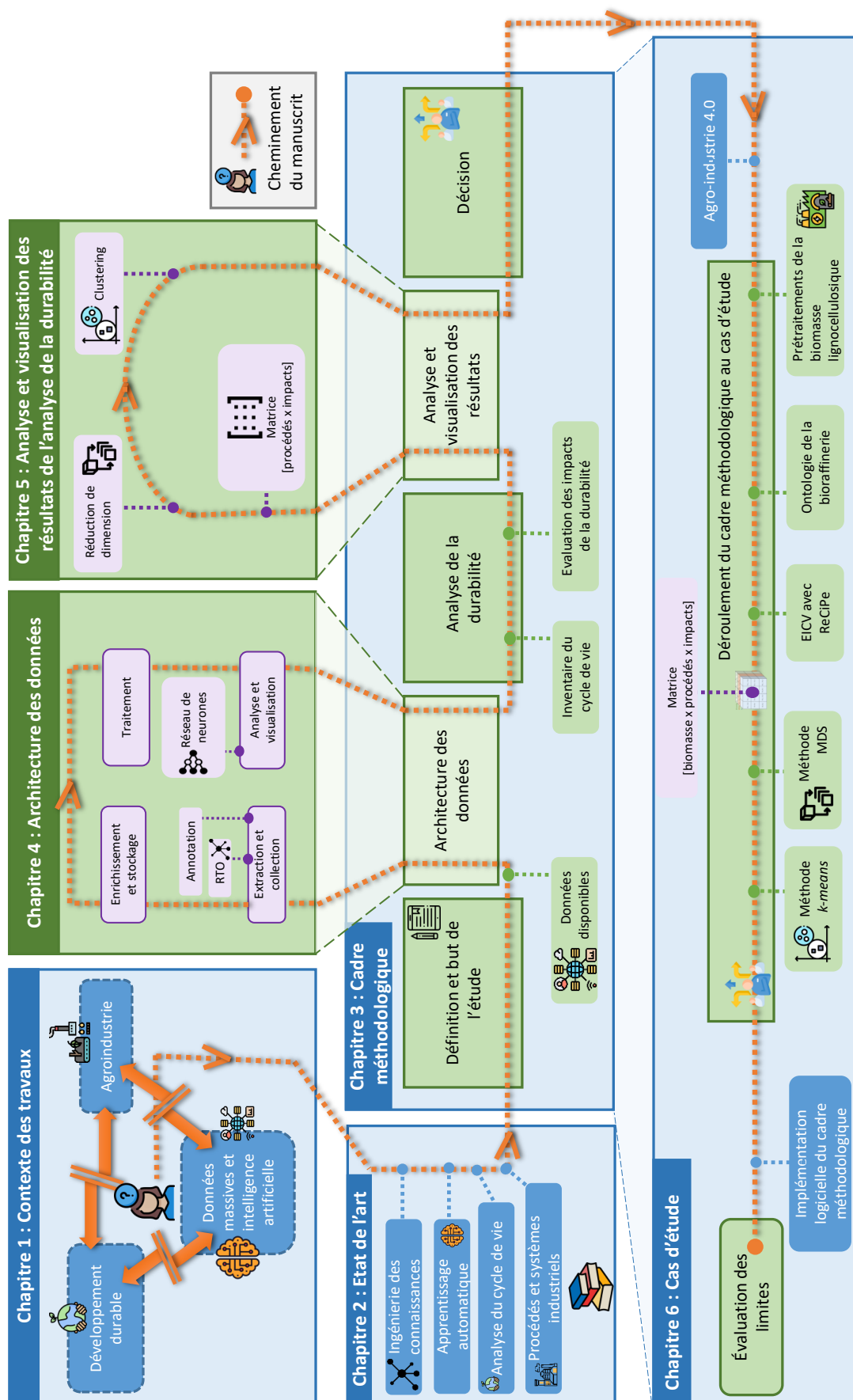


FIGURE 1 – Feuille de route du manuscrit.

Chapitre 1

Contexte général des travaux

Sustainable development is the pathway to the future we want for all. It offers a framework to generate economic growth, achieve social justice, exercise environmental stewardship and strengthen governance.

Ban Ki-moon - *Sécrétaire général des Nations Unis*,
2013

L'objectif de ce chapitre (figure 1.1) est d'une part, de présenter le contexte général dans lequel s'inscrivent ces travaux de thèse et d'autre part, d'identifier les questions scientifiques auxquelles la suite du manuscrit cherchera à apporter des réponses.

Ce chapitre introduit le terme de développement durable à travers un bref historique et la définition de deux concepts associés le cycle de vie et l'économie circulaire. La nécessité pour les décideurs de pouvoir évaluer ce développement durable sera abordée à cette occasion avec notamment l'observation que les données sont un des enjeux principaux de ces évaluations. Les données sont aussi au cœur de la transition digitale qui sera abordée dans la deuxième partie de ce chapitre. Ces deux parties mettent en lumière les différents domaines de recherche et les questions scientifiques.

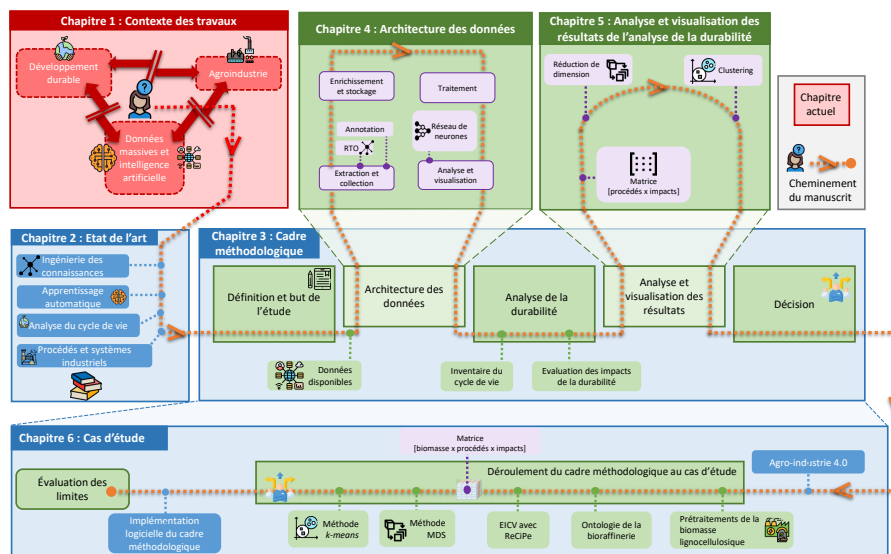


FIGURE 1.1 – Positionnement du chapitre 1 dans le manuscrit.

Sommaire

1.1 Le développement durable	7
1.1.1 Développement durable, cycle de vie et économie circulaire	8
1.1.2 Évaluer le développement durable	12
1.1.3 Les données, enjeux des analyses traditionnelles du développement durable	14
1.1.4 Conclusion sur le développement durable	16
1.2 La transition digitale au service du développement durable	16
1.2.1 L'Industrie 4.0	16
1.2.2 Une vue système orientée développement durable	19
1.3 Objectifs et questions scientifiques	22
1.4 Organisation du manuscrit	25

1.1 Le développement durable

Depuis le début des années 1970, à travers le développement économique, scientifique et technologique de notre société industrielle, l'Homme n'a cessé d'accroître son influence sur la Terre et ses ressources, entraînant un impact grandissant sur son environnement. La prise de conscience de ces problèmes écologiques a suscité de nouvelles idées pour un développement plus respectueux de l'environnement. Le rapport Brundtland de 1987, écrit par la Première ministre de Norvège, Mme Gro Harlem Brundtland, a marqué le début du développement durable, avec la première définition de ce terme comme « *un développement qui répond aux besoins du présent sans compromettre la capacité des générations futures à répondre aux leurs* » [WCED, 1987]. Le terme développement durable (DD), appelé en anglais *sustainable development* ou plus couramment *sustainability*, est officialisé en 1992 par le Sommet de la Terre à Rio qui définit trois piliers ou dimensions : un développement économiquement efficace, socialement équitable et écologiquement soutenable (figure 1.2). La durabilité recherchée est alors la conciliation (ou réconciliation) de ces trois dimensions qui, dès la première révolution industrielle, ont souvent été opposées voire ignorées.

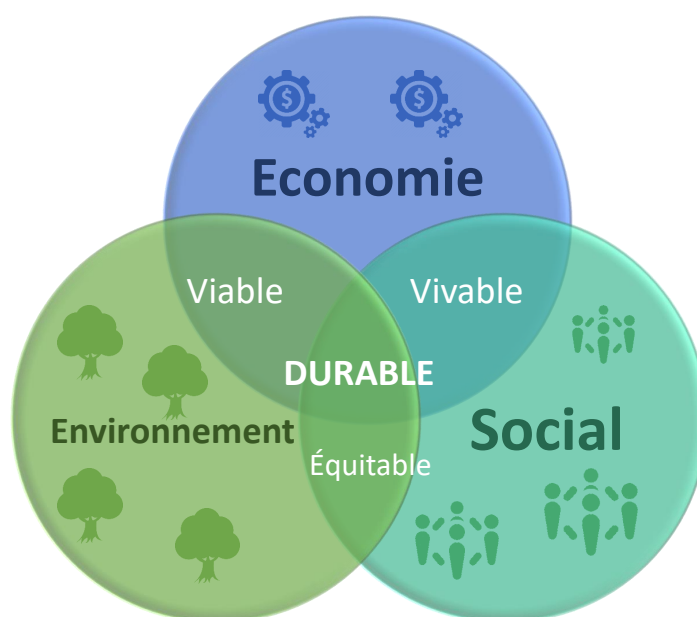


FIGURE 1.2 – Diagramme de Venn traditionnel du développement durable (adapté de [PURVIS et al., 2019]).

Depuis le cinquième rapport d'évaluation du Groupe d'experts Intergouvernemental sur l'Évolution du Climat (GIEC) publié en 2014 alertant les décideurs politiques, mais aussi la communauté scientifique sur l'urgence de la situation climatique, le monde politique met en place de nombreux plans d'action pour le climat, le développement durable ou l'environnement comme le protocole de Kyoto ou les Accords de Paris sur le climat [GIEC, 2014]. Les Accords de Paris sur le climat, signés en 2015 par 193 pays, visent à mettre en place des mesures pour limiter l'augmentation de la température mondiale à 2°C par rapport aux niveaux préindustriels afin d'éviter de mettre en péril l'ensemble de la biodiversité de notre planète. Le rapport du GIEC de 2018 vise plutôt une « action immédiate de la part des gouvernements, entreprises et investisseurs pour infléchir la tendance et limiter le ré-

chauffement climatique à 1,5°C. » [GIEC, 2018]. L'actualité avec le tout récent rapport du GIEC (début 2022) rappelle à quel point les conséquences possibles dues aux changements climatiques de nature humaine sont d'ampleur pour nos sociétés et l'environnement. Selon ce rapport, la température de la planète a augmenté de 1,09°C et la probabilité que nous atteignons une augmentation de 1,5°C à l'horizon 2030 est supérieure à 50 %. Malgré le constat sombre de la situation, le rapport traite également de nombreuses possibilités d'adaptation et de leur faisabilité dont certaines pistes de réflexion sont détaillées dans les paragraphes suivants.



Développement durable (DD) : développement qui répond aux besoins du présent sans compromettre la capacité des générations futures à répondre aux leurs. [ISO 14050 :2020, 2020]

1.1.1 Développement durable, cycle de vie et économie circulaire

La définition classique du développement durable repose sur les trois piliers présentés précédemment. Une autre version caractérise le concept de DD en utilisant la notion de durabilité faible ou de durabilité forte [MANCEBO, 2010]. Le DD est alors caractérisé par cette notion de durabilité :

- La durabilité serait dans sa conception forte « la durabilité d'un mode de développement qui apprécie conjointement une durabilité économique, sociale et environnementale ». L'idée clé est que le capital naturel et le construit ne peuvent être substitués. L'enjeu du DD avec cette durabilité forte est alors de préserver les stocks de capital naturel considéré comme irremplaçable.
- À l'inverse avec sa conception faible, le développement est durable s'il est prospère, si cette prospérité est répartie équitablement dans la population et si elle est « respectueuse » de l'environnement. L'idée, ici, est que les capitaux naturels et construits sont substituables avec pour enjeux de trouver des solutions techniques dites « vertes » pour remplacer les produits et procédés préexistants ou même restaurer l'environnement.

Sur notre planète, aux limites finies, le DD est qualifié par certains comme un « non-sens » [COULLET-DEMAIZIÈRE, 2010]. En effet, un développement avec une durabilité faible implique une croissance infinie de la population ou de l'économie respectueuse de l'environnement, ce qui n'est pas possible physiquement. Il apparaît alors les termes de « société durable » ou simplement de « durabilité », retirant le concept de développement et ne gardant que l'aspect durable. Bien que le concept de DD soit discutable et discuté, il n'en est pas moins aujourd'hui omniprésent dans nos sociétés. Ainsi, de nombreuses parties prenantes (l'Organisation des Nations Unies (ONU), les Organisation non gouvernementale (ONG), l'Union Européenne (UE), les nations, la société civile, les territoires ou l'entreprise) mettent en œuvre ce concept dans leurs différentes exigences.



Partie prenante : Toute personne ou organisation ayant un droit, une part, une revendication ou un intérêt dans un système et donc susceptible d'exprimer des exigences (besoins et contraintes). Définition reprise du domaine de l'ingénierie système [ISO/IEC/IEEE 15288 :2015, 2015].

Le DD exige des changements profonds dans notre façon de penser, de concevoir et d'utiliser les ressources, ainsi que dans les structures économiques et sociales des modes de consommation et de production. Il est aujourd'hui inimaginable de conserver le modèle linéaire de consommation « prendre, fabriquer, utiliser et disposer ». La « pensée cycle de vie » peut contribuer à améliorer les performances environnementales, et des avantages sociaux et économiques sont tirés d'approches prenant en compte le cycle de vie complet de la chaîne d'approvisionnement. En effet, cette approche est utilisée pour minimiser l'impact dans certains domaines, tout en prévenant d'autres impacts dans d'autres domaines. Le DD exige la prise en compte simultanée et stable des aspects économiques, environnementaux et sociaux.



Cycle de vie : Étapes consécutives et liées de l'acquisition des matières premières ou de la génération des ressources naturelles à l'élimination finale [ISO 14050 :2020, 2020]

Les Objectifs du Développement Durable (ODD), au nombre de dix-sept [NATIONS, 2022] et établis par les États membres des Nations Unies en 2015, reprennent ces trois piliers du DD. Un nouveau modèle économique pour un développement plus durable a récemment émergé : l'économie circulaire (EC), qui concilie les aspects économiques et environnementaux. L'EC entre pleinement dans l'ODD12 : « Établir des modes de consommation et de production durables » [GHISELLINI et al., 2016]. La figure 1.3 met en lumière, au travers des mots clés des articles scientifiques ¹, un aperçu de l'évolution des recherches menées sur le DD au cours des cinq dernières années. Au cours des années 2010, le DD est couplé aux politiques de gouvernance, à l'énergie, au changement climatique, aux impacts, à l'Analyse du Cycle de Vie (ACV) ou aux performances (partie basse droite ou haute gauche). Le début des années 2020 marquent le début des recherches sur l'EC et l'Industrie 4.0 (partie gauche de la figure).

En décembre 2019, lors du sommet sur le climat de la COP25 à Madrid [COP25, 2019], le *Green Deal* européen est présenté par la Commission Européenne, et accepté en janvier 2020 [EUROPEAN COMMISSION, 2019]. Ce plan européen comprend plusieurs actions visant à stopper le changement climatique et à réduire le niveau de pollution émis dans l'atmosphère. L'un des principaux plans d'action, adopté en mars 2020, est le *Circular Economy Action Plan*. Il propose d'utiliser le modèle d'EC pour simuler l'utilisation de modèles durables et lance GACERE (*Global Alliance on Circular Economy and Resource Efficiency*) en février 2021. Ce nouveau plan complète le premier plan d'action pour l'économie circulaire de 2015, dont les 54 actions ont été réalisées ou sont en cours de réalisation, et adopte au fur et à mesure de nouvelles actions (dernières en date au 30 mars 2022). L'EC est dé-

1. Une recherche bibliographique a été effectuée sur le terme *Sustainability* et tous les mots-clés associés à ce terme. Seuls les 5000 articles les plus cités sont pris en compte.

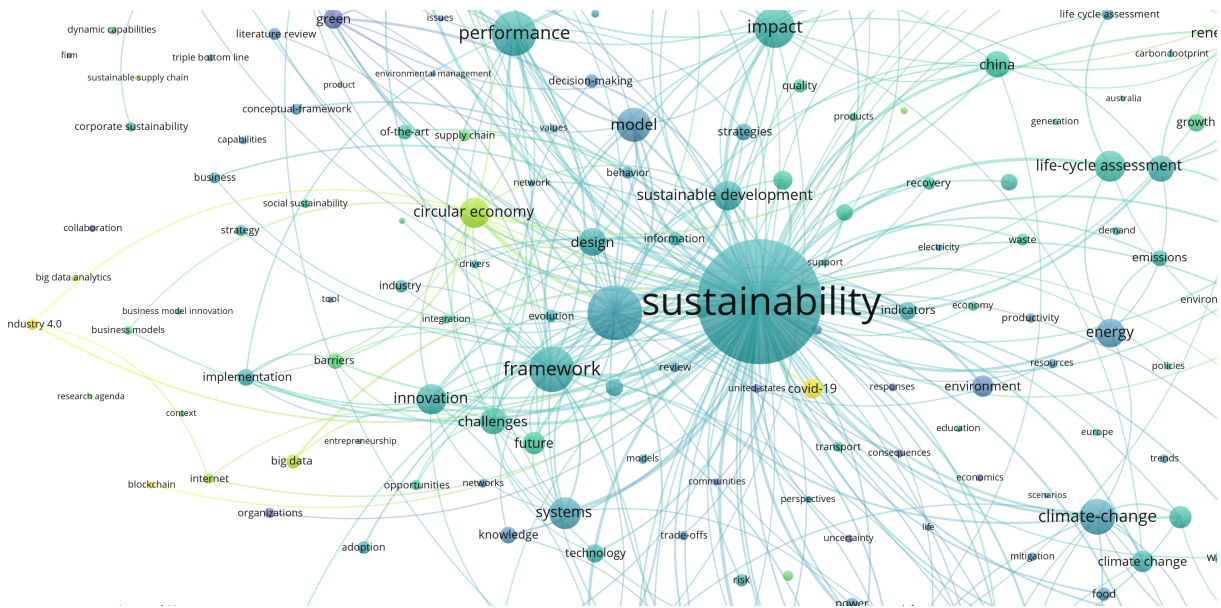


FIGURE 1.3 – Représentation des mots clés retrouvés dans les articles scientifiques traitant du développement durable entre 2016 et 2021 (dégradé de couleur du bleu pour 2016 au jaune pour 2021) grâce à VOSviewer.

finie par la fondation Ellen MacArthur comme « l'économie qui est conçue de façon restauratrice et régénératrice et qui vise à maintenir les produits, les composants et les matériaux à leur plus grande utilité et valeur à tout moment, en distinguant les cycles techniques et biologiques » [FOUNDATION, 2015]. Il est possible de voir une autre définition théorique : l'EC est « un système régénératif dans lequel l'apport des ressources entrantes, des déchets, des émissions et des fuites d'énergies est minimisé en ralentissant, fermant et réduisant les boucles de matières et d'énergie. Cet objectif est atteint grâce à une conception pensée pour le long terme, la maintenance, la réparation, la réutilisation, le réusinage, la remise à neuf et le recyclage. » [GEISSDOERFER et al., 2017]. Une troisième définition de l'EC est un système industriel qui est réparateur ou régénératif par intention et par conception [KIRCHHERR, 2022]. Pour résumer, les objectifs de l'économie circulaire sont :

- Réduire les émissions au niveau mondial,
- Réduire (ou ralentir ou fermer) les boucles de matières et d'énergie,
- Réduire la pression sur les ressources naturelles telles que l'eau, les matériaux « rares » listés par la Commission Européenne et l'utilisation des sols,
- Créer de nouvelles opportunités commerciales et des emplois locaux de qualité,
- Permettre des chaînes de valeur plus résilientes.



La résilience est définie par la capacité d'un système à conserver son fonctionnement malgré une perturbation majeure et/ou à récupérer son fonctionnement après une perturbation majeure [MACMILLAN et al., 2022].

Faisant suite au premier plan en 2015, le gouvernement français a proposé la SNTEDD (Straté-

gie Nationale de Transition Écologique pour un Développement Durable) qui se compose de neuf domaines, dont l'EC. Selon l'Agence Française de l'Environnement et de la Maîtrise de l'Énergie (ADEME), l'EC prend en compte trois domaines d'action : (1) la production et l'offre de biens et de services (2) la consommation au travers de la demande et du comportement du consommateur (économique ou citoyen) et (3) la gestion des déchets avec le recours prioritaire au recyclage qui permet de « boucler la boucle » [BELAUD et al., 2019a]. Ces trois domaines représentent l'ensemble du cycle de vie d'un produit ou d'un processus au travers de sept piliers (Figure 1.4). Pour parvenir à des modèles durables, la pensée cycle de vie peut aider à améliorer les performances environnementales tout en maximisant les avantages économiques et sociaux. À ce titre, des travaux antérieurs au sein de notre équipe de recherche ont traité majoritairement deux piliers : l'éco-conception [MORALES-MENDOZA et al., 2018; NEGNY et al., 2012] et l'écologie industrielle et territoriale [BELAUD et al., 2019a]. Il est aussi possible d'innover tout en gardant un contexte d'EC [LE et al., 2022].

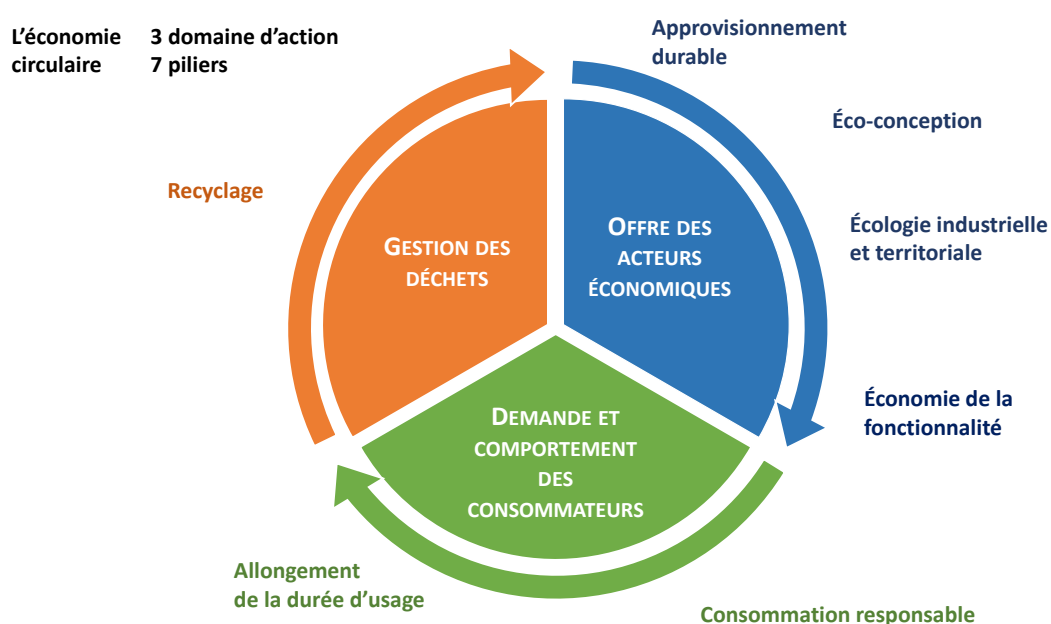


FIGURE 1.4 – L'économie circulaire selon trois domaines d'action et sept piliers (source : [ADEME, 2014]).



Les notions présentées ici sont issues de la norme ISO 14050 [ISO 14050 :2020, 2020] :

- **Produit** : se présenter sous la forme de biens matériels par exemple des produits chimiques, ou de biens immatériels comme des services de logistique.
- **Processus** : ensemble d'activités corrélées ou en interaction qui utilise ou transforme des éléments d'entrée pour fournir un résultat.

1.1.2 Évaluer le développement durable

Suite au Sommet de la Terre de Rio de 1992, la Commission du développement durable des Nations Unies lance en 1995 un programme de travail visant à créer des indicateurs caractérisant le DD. Cependant, les 134 indicateurs testés apparaissent souvent trop généraux et reflètent davantage la juxtaposition d'indicateurs que l'intégration des trois dimensions du DD. En 2006, la stratégie de l'Union Européenne met en exergue une sélection de 11 indicateurs « clés », aujourd'hui toujours utilisés pour indiquer le développement durable d'un pays. La littérature nous donne de nombreux moyens pour évaluer la durabilité d'un pays, d'une région, d'une communauté, d'une organisation, d'un projet ou d'un produit [SINGH et al., 2009]. Il n'est pas possible de décrire tous ces moyens dans l'exercice d'une thèse. Il existe toutefois des revues dans la littérature scientifique essayant de mettre en lumière les grands groupes d'analyses et/ou d'indicateurs. Dès 2007, Ness et al. mettent en exergue trois grands groupes d'évaluation de la durabilité :

- (i) Les indicateurs et les indices, intégrés ou non, sont à l'échelle d'un pays ou d'une région. Ce groupe d'évaluations est le plus rétrospectif.
- (ii) Moins rétrospectif, on trouve les analyses tournées vers un produit ou un projet.
- (iii) Et enfin, les plus prospectifs sont les analyses intégrées à la conception [NESS et al., 2007].

Le type de méthode est aussi un critère de regroupement par lequel on distingue les méthodes analytiques, les méthodes de programmation mathématique ou les méthodes de simulation [CALZOLARI et al., 2022]. La littérature présente aussi de nombreuses revues qui se focalisent sur les évaluations du DD dans un secteur spécifique [TURKSON et al., 2020].

Évaluer le DD a suscité la création d'un tel nombre de moyens que sa compréhension est en réalité devenue difficile pour les décideurs publics ou privés [NY et al., 2006]. Les propriétés, les caractéristiques, les limites et les objectifs de chaque moyen sont devenus flous pour les décideurs. Ainsi, les usages de ces évaluations se sont brouillés, laissant apparaître une idée selon laquelle toutes les évaluations se valent. Ce n'est évidemment pas le cas : tous les moyens créés répondent à des objectifs précis et différents ; ils peuvent être complémentaires, mais ne sont pas interchangeables. Pour chaque situation entraînant une évaluation (ou une analyse) du DD, deux points doivent être pris en compte : le besoin des parties prenantes et les caractéristiques mêmes de la situation. Ces deux points doivent régir le choix des moyens d'évaluation à utiliser. Les moyens d'évaluation ne sont pleinement efficaces que s'ils sont utilisés dans les conditions pour lesquelles ils ont été conçus. La nature et la qualité des données font, par exemple, partie de ces conditions (voir section 1.1.3).



Malgré l'utilisation interchangeable qui est faite des termes « analyse » et « évaluation » dans la littérature, il convient de donner des définitions distinctes.

- **Évaluation** : Action d'apprécier la valeur d'une chose.
- **Analyse** : Examen permettant d'isoler ou de discerner les différentes parties d'un tout. Ces définitions viennent du Centre National de Ressources Textuelles et Lexicales [CNRTL, 2022]. L'évaluation nécessite l'analyse. L'évaluation requiert alors des compétences et un investissement supérieurs à une démarche d'analyse.

Cette évaluation du DD est fréquemment utilisée pour la prise de décision en fournissant aux décideurs des informations utiles et pertinentes. Cependant, il ne faut pas confondre l'apport d'informations pour aider à la décision et la prise de décision elle-même. L'apport d'information est le domaine de l'expert en DD, ce n'est pas le cas de la décision. Si la raison d'évaluer le DD d'un système est d'informer une décision, cette évaluation doit être conçue dans les meilleures conditions possibles (transparence, validité ...), mais ne peut pas se substituer à la décision. Le risque est la création de moyens inadaptés ou le recours à des usages erronés de ces moyens : aucune incertitude sur les résultats, absence de prise en compte des hypothèses, non-respect des conditions d'utilisation ou non-respect du périmètre du système, par exemple. Évaluer le DD c'est, avant toute chose, le défi de bien poser le problème, définir le système, connaître le besoin, les hypothèses et les contraintes. L'évaluation du DD est liée à la notion de critères. Mendoza et al. ajoutent le concept de "principe" qui est défini comme une vérité sur laquelle s'appuie un raisonnement ou une action [MENDOZA et al., 2000]. Les critères sont alors considérés comme un principe de second ordre. Les auteurs mettent à un niveau inférieur à celui des critères les indicateurs qui sont « des variables ou composantes du système étudié utilisées pour caractériser l'état d'un critère particulier ». Enfin, plus bas dans l'échelle de la généralité, on trouve les vérificateurs définis comme « une donnée ou une information qui renforce la spécificité ou la facilité d'évaluation d'un indicateur ». Au bas de l'échelle, nous ajoutons la base de données informationnelles qui regroupe toutes les informations nécessaires à l'évaluation. Rajoutons aussi les trois sphères du DD dans cette organisation hiérarchique qui se trouveraient alors entre les critères et les indicateurs. La figure 1.5 est représentative de cette généralité dans un cadre de durabilité faible, qui est celui de cette thèse, car les indicateurs sont reliés à une seule sphère de la durabilité. Un critère peut s'appuyer sur une ou plusieurs sphères du DD, mais un indicateur ne prend en compte qu'une seule sphère. Par exemple, les critères socio-économiques sont décrits par un ou plusieurs indicateurs sociaux et un ou plusieurs indicateurs économiques.

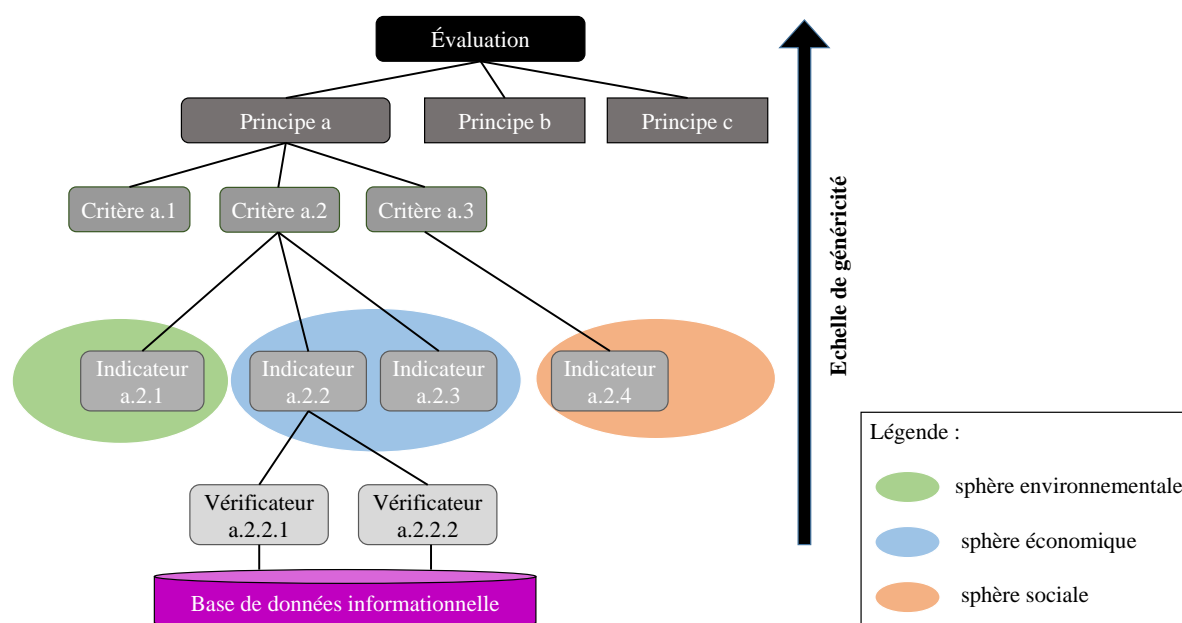


FIGURE 1.5 – Évaluation, principe, critère et indicateur (adaptée de [MENDOZA et al., 2000]).

Dans ces travaux, nous nous intéressons au principe de la pensée cycle de vie dont les évaluations de la durabilité sont les plus couramment utilisées. Ces évaluations systémiques évaluent les critères d'un système sur l'intégralité de son cycle de vie. De plus, évaluer la durabilité d'un système implique nécessairement la prise en compte d'un grand nombre de critères caractérisant des enjeux et objectifs différents, voire contradictoires, propres à chaque partie prenante. Le recours aux analyses multicritères s'impose alors naturellement. Dans le domaine du développement durable, la littérature scientifique n'est ni cohérente, ni harmonisée : les termes méthodes [THABREW et al., 2009], outils [NESS et al., 2007], approches [HACKING et GUTHRIE, 2008], indices [MAYER, 2008] et méthodologies [SINGH et al., 2009] correspondent souvent au même aspect. Dans ces travaux, notre approche se présente sous la forme d'un cadre méthodologique qui intègre des méthodes et outils provenant du domaine de l'Intelligence Artificielle (IA) et des analyses de la durabilité issues du principe de la pensée cycle de vie. Ce cadre sera donc la structure de notre approche.



Nous adoptons dans ce manuscrit la terminologie définie dans [SALA et al., 2015] qui définit chacun des termes parmi une certaine hiérarchie :

- **Framework ou cadre méthodologique/de travail** : base structurelle de l'intégration des concepts méthodologies, méthodes et outils.
- **Méthodologie** : ensemble de méthodes de caractérisation individuelles qui, ensemble, traitent les différentes questions environnementales, économiques et sociales et l'effet/impact associé.
- **Méthodes** : ensemble de modèles, d'outils et d'indicateurs qui permettent de calculer les valeurs des indicateurs pour une certaine catégorie d'impacts.
- **Modèle** : description mathématique, formelle ou informelle du système utilisé pour calculer un indicateur particulier de l'impact des interventions environnementales/-sociales/économiques.
- **Outil** : le logiciel, l'application, la base de données soutenant l'analyse ou l'évaluation effectuée en adoptant une méthode spécifique et les modèles associés.
- **Indicateur** : paramètre mesurable, ou une valeur dérivée de tels paramètres, qui met en évidence, fournit des informations ou cartographie l'état d'un phénomène, avec une signification allant au-delà de celle directement associée à sa valeur [OECD, 2003].
- **Indice ou critère** : mesure résultante de l'agrégation d'indicateurs [NESS et al., 2007].

1.1.3 Les données, enjeux des analyses traditionnelles du développement durable

Au travers des programmes « *Data for Climate* » et « *Big Data and the SDGs* » créés à la fin des années 2010, l'ONU met les données au centre des objectifs du DD [ONU, 2022]. Pour cause, sans données, il est impossible d'évaluer le DD d'un territoire ou d'un système comme schématisé par la figure 1.5. Par exemple, les dirigeants des pays pauvres sans accès aux technologies modernes possèdent très peu d'informations sur leur population contrairement aux pays développés comme les États-Unis où l'ONU a signé un partenariat avec Twitter[®] pour analyser les habitudes des utilisateurs

[NATIONS, 2022]. La Commission européenne suit aussi les avancées de l'Europe sur la disponibilité des données pour le calcul des indicateurs du DD afin d'atteindre les ODD [SABADASH, 2022].

Pour être considéré, le DD a besoin de données environnementales, sociales et économiques, quel que soit le système évalué. Traditionnellement, les données sont contenues dans des bases de données structurées régionales et thématiques et conçues spécialement pour les évaluations de la durabilité. On cite à titre d'exemple les données météorologiques, les données des populations ou les bases de données d'inventaires de cycle de vie de différents produits standards. Ces bases de données sont spécialisées et souvent incomplètes. Dans les analyses conformes à la pensée cycle de vie que nous privilégions dans cette thèse (voir paragraphe 1.1.2), des données tout au long du cycle de vie d'un produit ou d'un procédé sont nécessaires. Or, elles sont compliquées à obtenir : il convient de les compléter avec des données expérimentales, simulées ou extrapolées.

Les arrivées de l'*Open Data* et des données massives (appelées communément *Big Data*) sont devenues à la fois un défi et une opportunité pour le DD [REICHMAN et al., 2011]. Dans un rapport publié en juillet 2020, le Conseil National du Numérique français a rendu un avis pour que les « données environnementales » deviennent des données d'intérêt général. La notion de « données environnementales » repose ici sur l'article L. 124-2 du Code de l'environnement qui en donne une définition très générale : « toute information disponible, quel qu'en soit le support a pour objet l'état des éléments de l'environnement et leurs interactions ; les décisions et activités susceptibles d'avoir des incidences sur ces éléments (énergie, bruit, déchets, émissions, etc.) ; l'état de la santé humaine, dès lors qu'elle peut être altérée par des éléments de l'environnement ; l'analyse des coûts et avantages utilisée pour prendre des décisions ; et les rapports établis par les autorités publiques en lien avec l'environnement ». À travers cette énumération de données se pose la question de la source à utiliser pour l'extraction. C'est ici que l'*Open Data* devient une opportunité pour évaluer le DD. Au-delà de l'*Open Data*, tout capteur de données devient une source de données. Par exemple, souvent nommé « Roi des capteurs », les smartphones sont en capacité de collecter de nombreuses données différentes comme les déplacements des personnes, leur santé, ou encore la température [KRICHEN, 2021]. Si on laisse de côté les questions éthiques sur la récupération de données privées, un autre problème vient se poser sur les données ainsi récoltées : leur traitement. Comme pour les données massives et l'*Open Data*, ces bases de données sont non structurées et hétérogènes. Il est difficile sans des méthodes spécifiques d'extraire des connaissances de toutes ces données disponibles. Des domaines voient le jour comme l'ingénierie des connaissances pour palier à ces difficultés. La première partie du chapitre 2 (paragraphe 2.1.1 et 2.1.2) explore les méthodes disponibles.

1.1.4 Conclusion sur le développement durable

Le DD est difficile à définir et surtout à évaluer. Les évaluations du DD les plus courantes et choisies pour nos travaux proviennent du principe de la pensée cycle de vie. Une base de ces évaluations sont les données. L'arrivée des données massives et de l'*Open Data* permettent un accès facilité et instantané à de très nombreuses données. Cependant ces données sont difficiles à exploiter et à traiter. C'est la première problématique de nos travaux :



Q1 : Comment exploiter et tirer bénéfice des données massives pour aider aux analyses et aux évaluations du développement durable ?

1.2 La transition digitale au service du développement durable

Historiquement, l'industrie est un des plus grands pourvoyeurs de données de par toutes les informations récoltées lors d'une grande partie du cycle de vie d'un produit. Cela s'intensifie aujourd'hui avec l'explosion des capteurs et objets connectés accompagnant la chaîne de production globale. Selon le rapport de l'International Data Corporation de 2021, le volume de données créées et répliquées est passé de 2 zettabytes (ZB) en 2003 à 79 ZB en 2021 et atteindrait 181 ZB en 2025. L'industrie serait pourvoyeur d'environ un tiers de ces données [REINSEL et al., 2018]. En raison du développement de capteurs « intelligents », de l'informatique haute capacité, des technologies d'information, des données massives, de l'Internet des objets (*Internet of Things* (IoT)), l'industrie est confrontée à un grand bond en avant qui est considéré comme une nouvelle révolution industrielle [SHROUF et al., 2014]. Ces données créées par les industries sont hétérogènes. Elles sont en général non structurées et difficiles à intégrer et sous-exploitées. Les pays développés à travers des concepts d'Industrie 4.0 (I4.0) ou de *Cyber Physical Systems* (CPS) commencent à intégrer de plus en plus les données issues de tout le cycle de vie dans les prises de décision [BABICEANU et SEKR, 2016]. Le concept d'Industrie 4.0 n'est pas comme les précédentes révolutions industrielles qui sont à l'origine de nouvelles découvertes technologiques ou énergétiques, mais une combinaison de technologies et de fonctionnalités mises en œuvre dans toute la chaîne de valeur d'une entreprise [EL ZANT et al., 2021]. Les entreprises se voient offrir une proposition complète qui a pour cœur une double transition digitale et écologique.

1.2.1 L'Industrie 4.0

Dans les temps anciens de l'histoire, tous les aspects liés à la production, comme l'agriculture, le transport ou le secteur textile, entre autres, étaient réalisés manuellement, et le développement d'un produit prenait un temps considérable. La première révolution industrielle a été alimentée par la création de la pompe à eau et de la machine à vapeur, qui ont permis de mécaniser la production et de transformer les processus de production manuels en processus plus efficaces [SANTOS et al., 2017]. La deuxième révolution industrielle a été déclenchée par l'électrification et la division du travail dans l'industrie. Les chaînes de montage font leur apparition dans les industries de production et de nouvelles industries comme les industries chimiques ou agroalimentaires émergent. Le charbon était

l'énergie de la première révolution, l'électricité, le pétrole et le gaz sont les énergies de la deuxième [LI et al., 2017]. L'introduction de la microélectronique et de l'automatisation marque le début de la troisième révolution industrielle [ROJKO, 2017]. Les technologies de l'information voient le jour, ainsi que des systèmes de gestion intégrés (dits *Enterprise Resource Planning* (ERP) et Progiciel de Gestion Intégré (PGI)) et les structures des entreprises ont été réorganisées en fonction de leur cœur de métier.

Aujourd'hui, il est considéré que nous vivons la quatrième révolution industrielle à travers un concept introduit par la *Forschungsunion Wirtschaft und Wissenschaft* (association allemande regroupant des industriels et des scientifiques) : l'« Industrie 4.0 » (I4.0) [ALEXOPOULOS et al., 2016] [SANTOS et al., 2017]. Les origines de cette quatrième révolution sont nombreuses, mais on cite les principales : une diminution de la disponibilité des ressources naturelles, l'augmentation des prix de l'énergie, l'augmentation de l'âge des employés, la mondialisation des marchés, l'accès simplifié à de nombreuses technologies digitales et l'avènement des données massives. De plus, les demandes des clients deviennent de plus en plus complexes, l'innovation plus rapide, des produits avec des normes de qualité de plus en plus élevées. L'immédiateté et la satisfaction des commandes sont aussi devenues les normes clients. Dans ce contexte, la numérisation et l'intégration de l'I4.0 sont une solution réfléchie contribuant à améliorer la performance industrielle et répondant aux enjeux économiques et industriels. La transformation des industries en I4.0 est pour l'instant progressive, mais l'impact est considérable sur la performance de l'entreprise, la productivité, les marchés et l'industrie en général. Selon [ROJKO, 2017], l'I4.0 entraînerait une diminution de 10 à 30% des coûts de production et de logistique, et de 10 à 20% des coûts de gestion de la qualité. Cela s'ajoute à d'autres avantages sociaux (environnement de travail plus flexible et convivial), environnementaux (utilisation plus efficace et plus durable des ressources naturelles et énergétiques, meilleure réutilisation des déchets...) et économiques (délais de mise sur le marché plus courts, production plus personnalisée avec des coûts significativement plus bas...).

Plus qu'une révolution technologique, cette révolution se veut novatrice en termes de concepts et d'outils. Plusieurs auteurs débattent sur les caractérisations de cette révolution. On retiendra que les principales sont :

- L'intégration d'outils de l'intelligence artificielle [SANTOS et al., 2017],
- L'intégration de nouvelles technologies tels que les CPS, la 5G, les cobots, les robots ou les IoT,
- Des niveaux élevés d'efficacité opérationnelle, de productivité et d'agilité [LI et al., 2017],
- L'intégration de la durabilité à toutes les échelles,
- Le management des données : de la collecte à l'analyse, et l'intégration des technologies de l'information, des systèmes de fabrication et d'exploitation comme moyen d'acquérir des données de manière plus pertinente, rapide et souple [BRETTEL et al., 2014].

Le concept de durabilité se veut transversal dans l'I4.0, allant des ressources énergétiques et matérielles aux procédés de fabrication et à la logistique amont et aval. Certains voient même pour demain des usines 4.0 (ou usines intelligentes ou digitales) complètement intégrées dans des villes intelligentes et alimentées par des énergies solaire, éolienne et géothermique [STROZZI et al., 2017]. Les usines deviendront à terme connectées, optimisées, transparentes, proactives, agiles, « propres

», sobres et frugales. Les technologies telles que l’IoT, les CPS, l’*Additive Manufacturing*, le *Cloud Manufacturing* ou le *Big Data Analytics* permettent la communication entre les différents acteurs et objets des industries, afin d’obtenir plus de modularité et de personnalisation des produits de masse.

Les CPS sont définis par [ZANERO, 2017] comme un ensemble de ressources informatiques et physiques interconnectées qui utilise une boucle de contrôle intelligente, afin d’adapter et d’améliorer l’autonomie et l’efficacité de l’ensemble du système. Les CPS sont avant tout caractérisés par de fortes interactions entre les composants cybernétiques (logiciels) et les composants physiques. Ces deux éléments combinent étroitement l’IA, l’automatisation et les technologies de communication, afin d’atteindre des niveaux élevés de performance, de fiabilité, d’efficacité et de robustesse dans un environnement physique [GOOSENS et RICHARD, 2013]. Les CPS connectés au Cloud ou à des infrastructures réseau dédiées 0G ou 5G (LoRa, UnaBiz, Sigfox,…) sont souvent appelés IoT. Le *Cloud computing* fournit aux entreprises des ressources informatiques, allant du stockage/traitement des données à leur exploitation par des logiciels, comme des services sur des réseaux internes ou externes [SHILA et al., 2017]. On parle aussi de l’« Internet des Services » par analogie avec les IoT [KAGERMANN et al., 2013]. L’IoT est un concept qui varie suivant les auteurs, mais la vision principale est que l’IoT est un réseau d’objets physiques possédant des caractéristiques propres et pouvant émettre et recevoir des données. Ces objets sont semi-autonomes ou autonomes (intelligents) en interagissant dans et avec leur environnement. Cette possibilité que donne l’IoT de lier les objets entre eux et d’interagir est un grand pourvoyeur de données (on pourra penser à tous les capteurs d’un procédé par exemple). Dans la suite naturelle de l’IoT et de l’Internet des services, les auteurs regroupent maintenant cela sous le concept de l’Internet de Tout (en anglais *Internet of Everything* (IoE)). Cet IoE ne regroupe plus seulement les objets, mais aussi les personnes, les processus (basés sur l’IA) et les données (figure 1.6) pour une connectivité de « n’importe quoi, n’importe où, n’importe quand » pour une santé plus intelligente ou des villes et transports plus intelligents et durables [YANG et al., 2017].

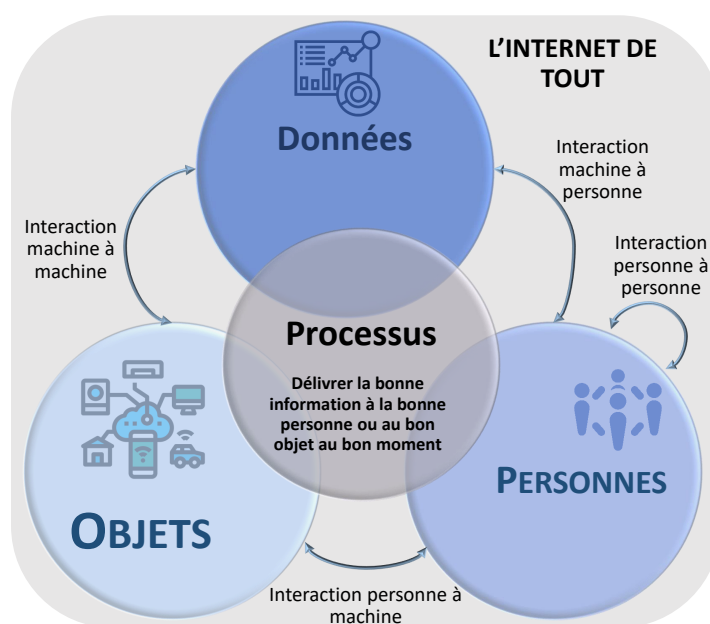


FIGURE 1.6 – Diagramme de présentation du concept de l’Internet de Tout.

Toutes les technologies citées précédemment (IoT, IoE, CPS, robots, cobots, ...) ont comme point commun les données. Ce sont ces données à travers les données massives et les méthodes et outils connexes qui constituent l'un des piliers de la transition digitale des industries [CHEN et ZHANG, 2014]. Elles sont aussi le plus grand défi car elles proviennent de différentes sources souvent hétérogènes et non structurées : il faut alors des processus, des méthodes pour traiter toutes ces données. L'I4.0 puise beaucoup dans le domaine des données massives et de l'intelligence artificielle pour collecter, exploiter, traiter et analyser ces données. Les "processus dans l'IoE" sont toutes les méthodes de l'IA et des données massives servant à délivrer la bonne information au bon moment à la bonne personne ou au bon objet.

L'I4.0 utilise donc les outils et méthodes issus des données massives et de l'IA pour faire face à cette profusion de données dans l'industrie, que ce soit dans les phases de conception ou de maintenance ou avec les fournisseurs ou les clients. Il vient alors la question de savoir si ces méthodes et outils servent aussi à la transition écologique des industries, ce qui nous conduit à établir notre deuxième problématique :



Q2 : Comment bénéficier des méthodes et outils issus de l'intelligence artificielle pour aider les analyses et les évaluations de la durabilité ?

1.2.2 Une vue système orientée développement durable

Comme explicité, le DD devrait exploiter le principe de la pensée cycle de vie. La notion de cycle de vie demeure assez essentielle dans le domaine de l'ingénierie et dans l'industrie. Qu'il s'agisse de système, de service, de produit, de procédé ou de projet, tous les objets génériques, même très différents, ont un cycle de vie [AFIS, 2004]. Pour le cycle de vie complet d'un produit, service, système ou procédé, l'expression admise est du « berceau à la tombe » (« *from cradle to grave* »). Dans l'EC, il est même question de boucle de recirculation entre la fin de vie et l'extraction des matériaux donnant un cycle de vie du « berceau au berceau » (« *from cradle to cradle* »). Pour un système plus complexe, un cycle de vie plus générique incluant toutes les phases préliminaires et indispensables que sont la formulation du besoin, la recherche, la conception et les essais fait son apparition [MEINADIER, 2003]. La notion de cycle de vie, quel que soit le système considéré, décrit l'ensemble des phases de l'existence de ce système depuis l'idée jusqu'à la fin de vie. Ces étapes sont « des périodes du cycle de vie qui correspondent à un état du système ou de sa description ». C'est donc une représentation temporelle du système. Issu du monde de l'industrie, le cycle de vie d'un projet est souvent représenté par le cycle en V (figure 1.7) bien que le paradigme agile a, depuis les années 2000, bousculé ce fait. C'est devenu un standard conceptuel depuis l'apparition de l'Ingénierie Système (IS). Cette représentation apporte la dimension d'un système. En effet, le produit d'un projet n'est plus obligatoirement un produit, mais un système composé de plusieurs modules ou composants plus ou moins complexes.

Avec la 3ème puis la 4ème révolution industrielle, les systèmes sont devenus de plus en plus complexes et les parties prenantes ainsi que les intervenants interagissant avec le système sont de plus en plus nombreux. L'intégration des considérations du DD doit se faire le plus tôt possible, à savoir dès les étapes d'« Exigences du système » et de « Conception générale » ou dès l'analyse du besoin si

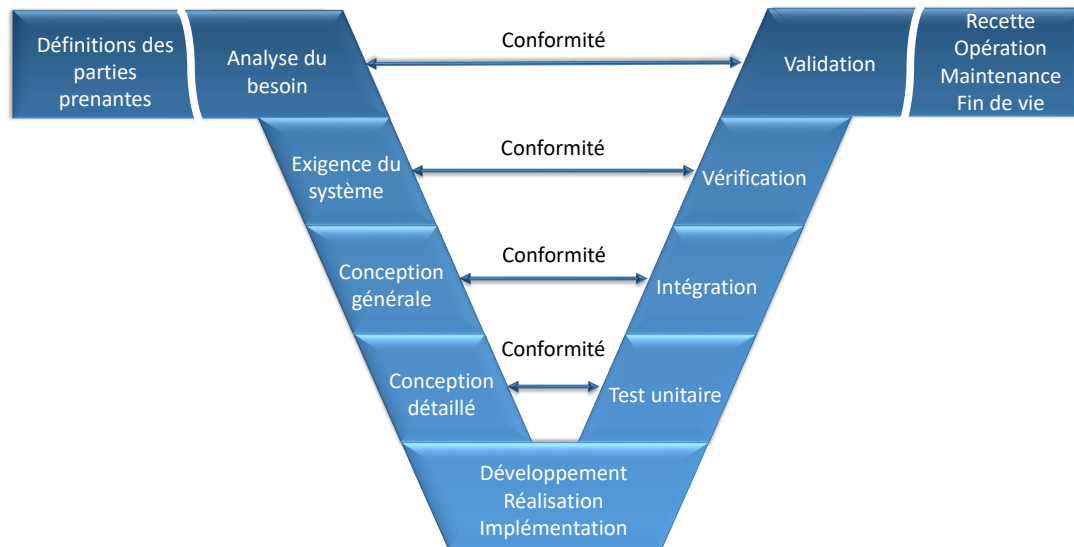


FIGURE 1.7 – Cycle en V adapté de [AFIS, 2004].

le système est intrinsèquement lié à la transition écologique (ferme éolienne, piles à combustible,...) Ces deux étapes font partie du processus « Analyse du système » suivant les processus techniques décrits dans l'ISO/IEC/IEEE 15288 : « le but du processus Analyse du système est de fournir une base de données et d'informations rigoureuse pour la compréhension technique afin d'aider la prise de décision tout au long du cycle de vie » [ISO/IEC/IEEE 15288 :2015, 2015]. L'appel aux concepts de l'I4.0 et aux analyses de la durabilité que nous avons décrits précédemment facilitent cette intégration au plus tôt. Cependant cette intégration au plus tôt entraîne des difficultés. Par exemple, les personnes responsables de ces étapes ne sont pas forcément formées pour prendre en compte les questions de durabilité. Caiardi et al. mettent en œuvre une analyse environnementale dès l'étude d'un projet pour savoir si un procédé de couplage méthanisation-pyrolyse serait plus durable qu'un procédé de méthanisation seul, pour le traitement de biomasses sur le territoire marocain [CAIARDI et al., 2022]. Au lieu de passer directement à la conception générale du procédé la partie "exigences du système" implique que le procédé de couplage soit plus durable que celui de méthanisation pré-existant. Ils utilisent alors une analyse environnementale sur les deux systèmes. Cette intégration de l'analyse environnementale demande, cependant, du temps et de l'investissement humain et financier.



Systeme : combinaison d'éléments en interaction organisée pour atteindre un ou plusieurs objectifs définis.

Définition reprise du domaine de l'ingénierie système [ISO/IEC/IEEE 15288 :2015, 2015].

La complexité due à la multiplicité des individus qui interviennent sur un même système est un des fondements de l'IS. L'IS est née avec les programmes de missile balistique des années 1950 et s'est

développée aux États-Unis lors des grands projets spatiaux des années 1960. Les systèmes intègrent de plus en plus de technologies et disciplines diverses qui ont chacune un langage qui leur est propre. Il a fallu alors développer une approche – qui est maintenant un domaine de recherche – l’IS. C’est une approche qui prend en compte l’ensemble de la vie du système : le cycle en V est devenu une norme au travers de l’IS, et inclut toutes les parties prenantes et toutes les disciplines (technique et management). Par nature, l’IS est une approche systémique et utilise des modèles issus de disciplines différentes : nos travaux s’insèrent dans cette démarche par l’intégration de modèles issus du DD, des données massives et de l’IA. Il est d’ailleurs possible de citer le domaine de recherche de l’Ingénierie Système Basée sur les Modèles (ISBM) (ou *Model-Based Systems Engineering* (MBSE)) qui correspond à une intégration de modèles issus de différents métiers.



« L’IS est une démarche méthodologique (processus, méthodes, outils) multidisciplinaire qui englobe l’ensemble des activités adéquates pour concevoir, développer, vérifier, valider et faire évoluer un système apportant une solution économique et performante aux besoins d’un client tout en satisfaisant l’ensemble des parties prenantes » [ISO/IEC/IEEE 15288 :2015, 2015].

Nos travaux se trouvent à la frontière de nombreux domaines : ceux de l’environnement et de la chimie (analyses du DD et EC) et celui de la science des données (IA et données massives). Le cas d’étude de nos travaux qui s’intéresse à la production de glucose à partir de déchets lignocellulosiques de l’agriculture (détaillé dans le chapitre 6) ajoute aussi les métiers du génie des procédés et des systèmes à cette liste. L’EC prônant la réduction des déchets, les procédés et systèmes industriels (PSI) mettent en place des boucles de recyclage et de valorisation de leurs déchets. Ces boucles sont ajoutées soit a posteriori soit lors de la phase de conception des PSI. Parmi les industries, l’agro-industrie est un des grands pourvoyeurs de déchets. De nombreuses valorisations des déchets agro-industriels voient le jour pour que cette industrie tende vers une économie circulaire et durable. Par exemple, les déchets lignocellulosiques agricoles (la plante hors produit final) servent à fabriquer du glucose. Les déchets agricoles deviennent non plus des déchets, mais des coproduits servant à fabriquer du glucose qui sont ensuite valorisés en biocarburants, biomatériaux ou biomolécules. Cela entre dans une démarche d’EC. La figure 1.8 présente le système de production de glucose à partir des déchets agricoles et les différents systèmes en interaction dans le cycle de vie du système. Le cycle de vie du système est en bas de la figure avec les sous-systèmes que sont l’agriculture (système qui va produire un produit final agricole et un déchet lignocellulosique servant à la production de glucose), les procédés de prétraitement, le transport, la valorisation du glucose et la fin de vie de ces valorisations (énergétique par exemple). Le haut de la figure montre deux autres systèmes en interaction avec le cycle de vie du système.

L’utilisation de déchets agricoles dans une démarche d’EC rentre dans un concept lié à l’EC : la bioéconomie. La bioéconomie désigne toute chaîne de valeur qui utilise comme point de départ des biomatériaux et des produits d’origine agricole, aquatique ou forestière. La Commission Européenne définit la bioéconomie comme : « la production de ressources biologiques renouvelables et leur conversion en denrées alimentaires, aliments pour animaux, produits biosourcés et bioénergie.

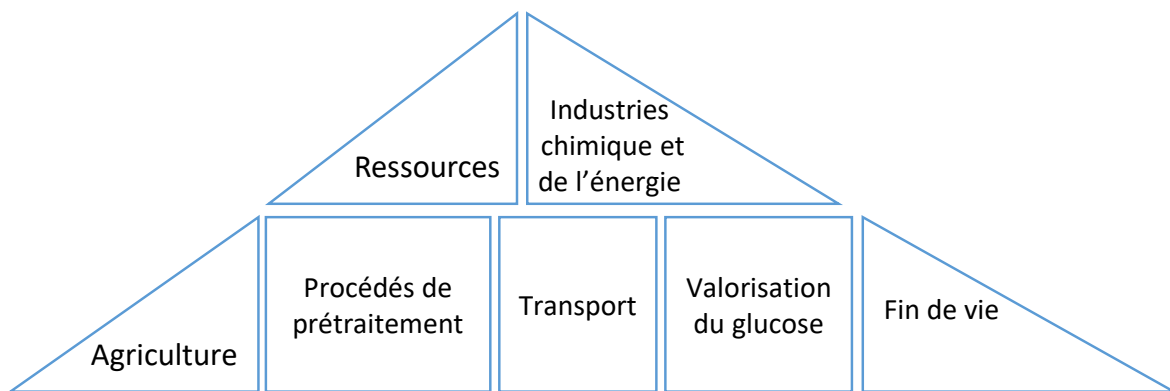


FIGURE 1.8 – Schéma des étapes du cycle de vie du glucose avec le détail des différents systèmes en interaction.

Elle comprend l'agriculture, la sylviculture, la pêche, la production de denrées alimentaires et de pâte à papier, ainsi que certaines parties des industries chimiques, biotechnologiques et énergétiques » [EEA, 2018]. Ce concept englobe également la production de ressources biologiques renouvelables et la conversion de ces ressources et des flux de déchets en produits à valeur ajoutée tels que les denrées alimentaires, les aliments pour animaux, les produits biosourcés et les bioénergies [CARUS, 2016]. Les concepts d'EC et de bioéconomie sont très liés de par le passage des ressources non renouvelables aux biomatériaux. Carus affirme que les deux approches sont différentes, mais complémentaires, car :

« l'EC renforce l'efficacité des procédés et l'utilisation de carbone recyclé pour réduire l'utilisation de carbone fossile supplémentaire. La bioéconomie remplace le carbone fossile par du carbone biologique issu de la biomasse de l'agriculture, de la sylviculture et de l'environnement marin [...] Le concept de bioéconomie va bien au-delà de l'EC, incluant beaucoup plus d'aspects tels que de nouvelles briques chimiques, de nouveaux procédés de transformation, de nouvelles fonctionnalités et propriétés des produits » - [CARUS, 2016].

L'ajout de trajectoires de valorisation des déchets agricoles rentre pleinement dans ces aspects de bioéconomie et d'EC. Cet ajout change le système global de production de produits agricoles qui devient plus durable et circulaire. Cependant pour que le système global reste circulaire et durable, les trajectoires de valorisation se doivent d'être vertueuses. Ce constat nous amène à la troisième et dernière problématique.



Q3 : Comment vérifier que des procédés innovants soient considérés comme des systèmes industriels durables et circulaires ?

1.3 Objectifs et questions scientifiques

Nous remettons ici en perspective les questions scientifiques auxquelles le présent manuscrit va s'employer à répondre en proposant un cadre méthodologique original. La transition digitale est un des moyens qui peut aider à la transition écologique notamment au travers de méthodes développées pour traiter et analyser les données. Il est possible de trouver des technologies usant des don-

nées massives déjà incorporées dans l'économie circulaire ou dans l'écologie industrielle [MING et al., 2015]. Les analyses et les évaluations de la durabilité sont très consommatrices de données notamment de données de premier plan sur le procédé étudié et de données d'arrière-plan pour la transformation des substances en émissions. Or, comme vu dans le paragraphe 1.1.2, pour qu'un procédé soit considéré comme un système industriel circulaire et durable il doit être analysé et évalué. Cette durabilité doit être pensée lors des phases d'études et de conception. Cependant alors que des méthodes d'aide à la décision sont créées pour des éco-parcs industriels [BELAUD et al., 2019a], il n'y a pas ou peu de méthodes développées pour la recherche et l'implémentation des données de premier plan dans une analyse ou une évaluation de la durabilité. La collecte de données peut être longue et fastidieuse. Automatiser cette collecte par l'apport de données issues du web notamment des articles scientifiques et du web public (coût de matière première par exemple) est bénéfique. Ces sources étant très hétérogènes, il faut faire appel à des outils et méthodes spécifiques. Les outils et méthodes développés pour les données massives permettent de les compiler, les traiter et les analyser.

La problématique principale abordée dans ce manuscrit concerne une meilleure gestion des nombreuses données décrivant les procédés et les systèmes ainsi que l'apport de ces données dans les analyses et évaluations de la durabilité grâce aux outils et méthodes issus des données massives et de l'IA. Nous résumons en trois questions de recherche principales les trois grands domaines de recherche. Ces questions sont décrites dans la figure 1.9 et constituent nos objets d'études principaux.

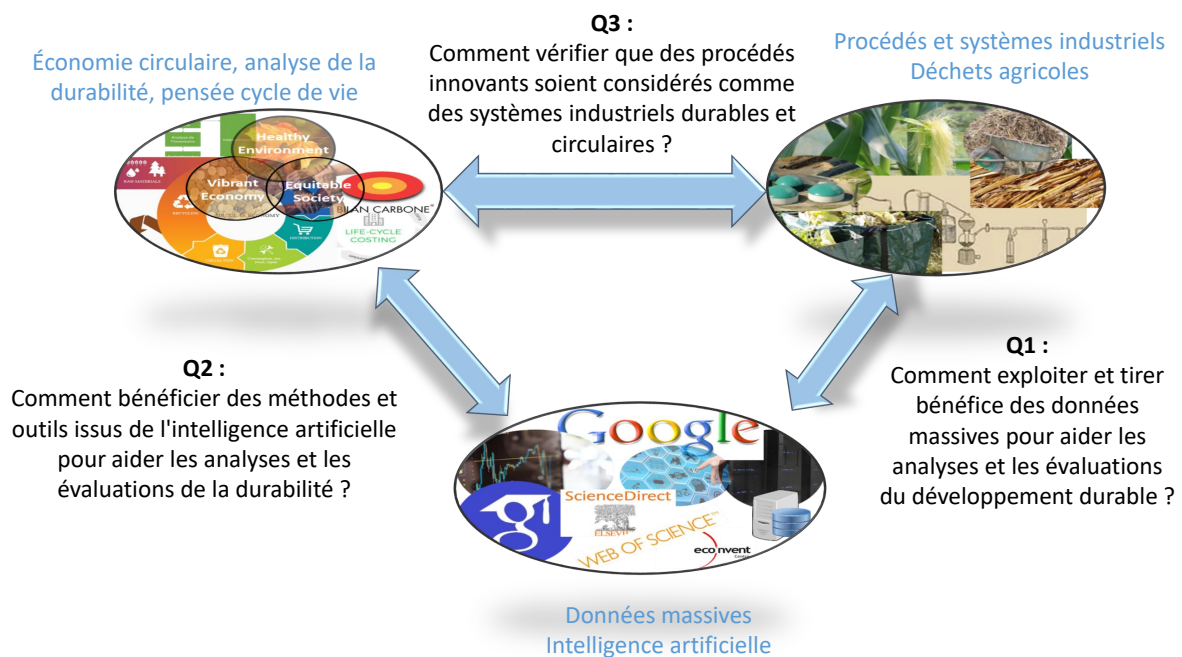


FIGURE 1.9 – Contexte général de la thèse et questions scientifiques associées.



Dans la suite du manuscrit pour simplifier le discours, nous utiliserons « analyse » pour parler d' « analyse et évaluation ».

Les présents travaux abordent la mise en œuvre de méthodes d'IA et de données massives pour aider les analyses de la durabilité dans un contexte industriel ou de recherche permettant aux systèmes à devenir plus durables et circulaires. Ce sujet est d'importance pour les domaines de recherche en écologie industrielle et en ingénierie, ainsi que pour les praticiens industriels. En outre [MURRAY et al., 2017] soulignent la nature multidisciplinaire de l'économie circulaire qui apparaît dans un nombre croissant de domaines de recherche, tel un paradigme décloisonnant les disciplines. Grâce à la conception d'un cadre pour compiler les données des systèmes industriels complexes et analyser les résultats des analyses de la durabilité pour aider l'interprétation, ce sujet de recherche vise à identifier les meilleurs leviers d'action pour aider les chercheurs et les ingénieurs Recherche et Développement (R&D) dans les phases de conception générale et de formulation des exigences à intégrer la durabilité. À chaque question scientifique nous avons défini un objectif. La figure 1.10 reprend ces questions et objectifs et les associe à un chapitre du manuscrit.

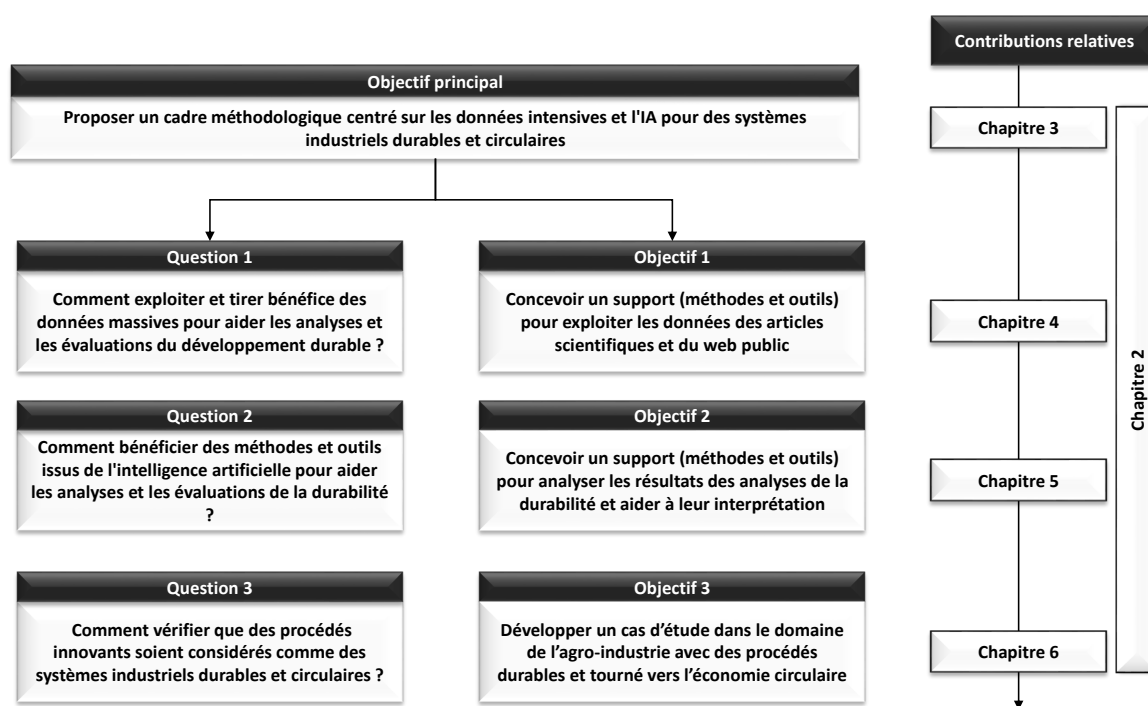


FIGURE 1.10 – Questions scientifiques, objectifs et chapitres associés.

Le chapitre 2 va effectuer un état de l'art des différents domaines de recherches abordant les différents outils et méthodes qui seront utilisés dans notre cadre méthodologique. La figure 1.11 donne une représentation schématique des domaines de recherche. Elle sera complétée avec les outils et méthodes a posteriori de l'état de l'art.

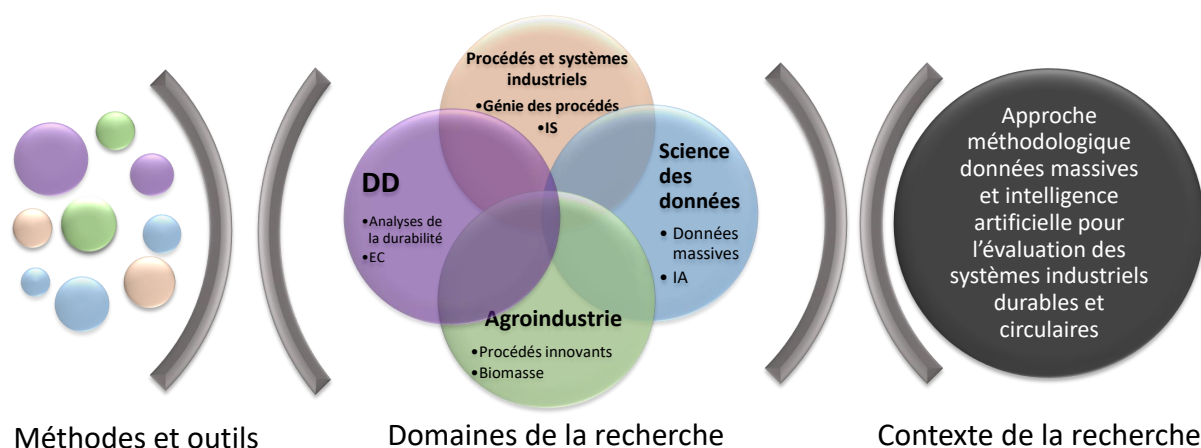


FIGURE 1.11 – Contexte et domaines de notre approche.

1.4 Organisation du manuscrit

La suite du manuscrit est organisée autour de cinq chapitres.

- **Chapitre 1 – Contexte général des travaux :** Ce chapitre a introduit le terme de développement durable à travers un bref historique et la définition de deux concepts associés, le cycle de vie et l'économie circulaire. La nécessité pour les décideurs de pouvoir évaluer ce développement durable sera abordée à cette occasion avec notamment l'observation que les données sont un des enjeux principaux de ces évaluations. Les données sont aussi au cœur de la transition digitale qui sera abordée dans la deuxième partie de ce chapitre. Ces deux parties ont mis en lumière les différents domaines de recherche, les questions scientifiques et les objectifs.
- **Chapitre 2 – État de l'art :** À la lumière de ce chapitre, il dresse le paysage bibliographique dans lequel s'inscrit le cadre méthodologique et plus largement nos travaux. Il parlera des données massives, de l'intelligence artificielle, du développement durable et des procédés et systèmes industriels ainsi que le couplage entre ces quatre domaines.
- **Chapitre 3 – Définition du cadre méthodologique :** Ce chapitre est dédié à la présentation d'un cadre méthodologique mis en place pour coupler les données massives et l'intelligence artificielle aux analyses de la durabilité.
- **Chapitre 4 – Architecture des données :** Ce chapitre précise l'étape 2 de notre cadre méthodologique intitulé "Architecture des données" en décrivant ses sous-étapes. L'apport de l'ingénierie des connaissances pour l'extraction des données nécessaires aux analyses de la durabilité ainsi que les réseaux de neurones et leur interprétabilité pour l'analyse de ces données sont évoqués.
- **Chapitre 5 – Analyse et visualisation des résultats de l'évaluation de la durabilité :** Ce chapitre développe l'étape 4 : "Analyse et visualisation des résultats". Il explore les méthodes issues de l'apprentissage automatique pour analyser et visualiser les résultats des analyses de la durabilité pour aider les décideurs.
- **Chapitre 6 – Cas d'étude : valorisation des déchets lignocellulosiques :** Ce chapitre illustre le cas d'étude et présente la conception et le développement d'un outil.

Enfin, la **conclusion générale** synthétise et reprend les contributions de ces travaux en les confrontant aux questions de recherche et aux objectifs établis dans le Chapitre 1. Elle s'achève sur des perspectives alimentant les réflexions et les évolutions potentielles.

Chapitre 2

État de l'art

"Data! Data! Data! he cried impatiently. I can't make bricks without clay!"

Sherlock Holmes - *"The Adventure in the Copper Beeches"* (1892)

Ce chapitre (figure 2.1) dresse le paysage scientifique et bibliographique dans lequel s'inscrivent le cadre méthodologique et plus largement nos travaux. Il réalise une description précise et appuyée par les définitions de la littérature, d'une part des différents domaines de recherche qui interviennent dans ces travaux et d'autres part les méthodes et outils associés. Il est découpé en quatre grandes parties. La première partie définit les données massives et l'IA, les méthodes qui leur sont attachées ainsi que certaines applications que l'on trouve dans les industries. La seconde partie fait un rapide tour d'horizon des différentes méthodes d'analyses du DD, spécialement les méthodes en adéquation avec la pensée cycle de vie. La troisième partie met en avant la littérature récente sur l'utilisation des données massives, de l'IA pour le DD des industries. Finalement, la quatrième partie s'intéresse aux PSI à travers une rapide littérature de ceux-ci aux services du DD. Ces différentes parties vont mettre en lumière les différentes méthodes sélectionnées pour la définition du cadre méthodologique ainsi que le choix du cas d'étude.

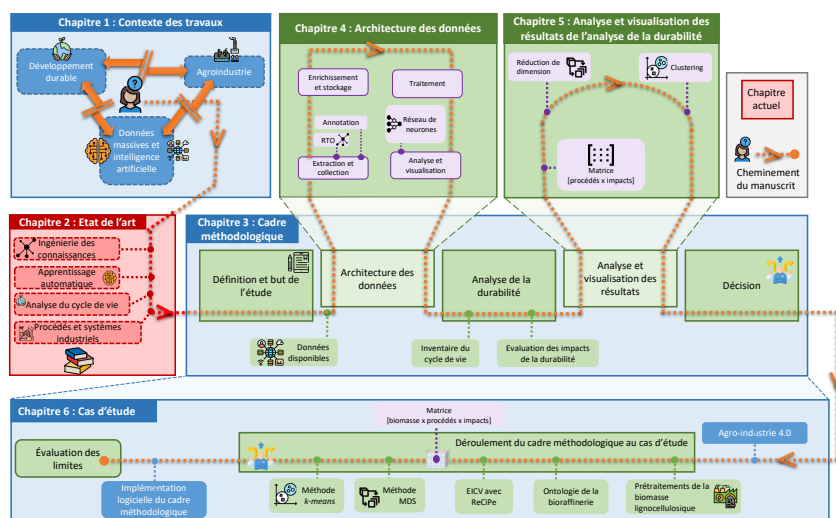


FIGURE 2.1 – Positionnement du chapitre 2 dans le manuscrit.

Sommaire

2.1 Données massives et intelligence artificielle	29
2.1.1 Les données massives	29
2.1.1.1 Donnée, Information, Connaissance et Sagesse	30
2.1.1.2 Notion de "données massives"	31
2.1.1.3 Les données massives dans l'industrie	31
2.1.1.4 Conclusion sur les données massives	33
2.1.2 Intelligence artificielle	33
2.1.2.1 Définition générale de l'intelligence artificielle	33
2.1.2.2 Les domaines de l'intelligence artificielle	34
2.1.2.2.1 Apprentissage automatique ou <i>Machine Learning</i>	34
2.1.2.2.2 <i>Natural Language Processing</i>	37
2.1.2.2.3 Système expert	37
2.1.2.2.4 L'ingénierie des connaissances	37
2.1.2.2.5 Planification	39
2.1.2.2.6 Combinaison avec d'autres technologies	39
2.1.2.3 L'intelligence artificielle dans le secteur industriel	39
2.1.2.4 Conclusion sur l'intelligence artificielle	40
2.2 La pensée cycle de vie pour le développement durable	41
2.2.1 La pensée cycle de vie	41
2.2.2 Méthodes conformes à la pensée cycle de vie	42
2.2.3 Analyse du cycle de vie	43
2.2.3.1 L'analyse environnementale du cycle de vie (AeCV)	44
2.2.3.2 L'analyse économique du cycle de vie (AcCV)	45
2.2.3.3 L'analyse sociale du cycle de vie (AsCV)	46
2.2.3.4 L'analyse de la durabilité du cycle de vie (AdCV)	46
2.2.4 Conclusion sur le choix de l'ACV	47
2.3 Développement durable, intelligence artificielle et données massives	47
2.4 Procédés et systèmes industriels et développement durable	50
2.5 Conclusion du chapitre 2	53

2.1 Données massives et intelligence artificielle

Le développement rapide des technologies de l'information a conduit à la génération d'une énorme quantité de données. Pour traiter et exploiter ces données, des outils et la science autour de ces outils se sont développés. Par l'intermédiaire des *data scientists*, le terme "science des données" est apparu pour décrire un domaine interdisciplinaire ayant pour but d'exploiter des grands ensembles de données. En d'autres termes, la science des données gère le cycle de vie des données. Elle utilise des techniques et méthodes issues des mathématiques, de l'informatique, des visualisations de données, de l'intelligence artificielle et des technologies de l'information, à la fois, pour extraire de la connaissance de ces données et pour les analyser afin d'aider la prise de décisions [WALLER et FAWCETT, 2013]. Cependant, cette science a connu une forte expansion avec l'émergence des données massives (souvent appelées par leur terme anglais *Big Data*). La figure 2.2 présente les recouvrements entre la science des données, les données massives et l'intelligence artificielle. La suite du chapitre s'attarde sur l'intelligence artificielle et les données massives qui sont exploitées au sein de ces travaux.

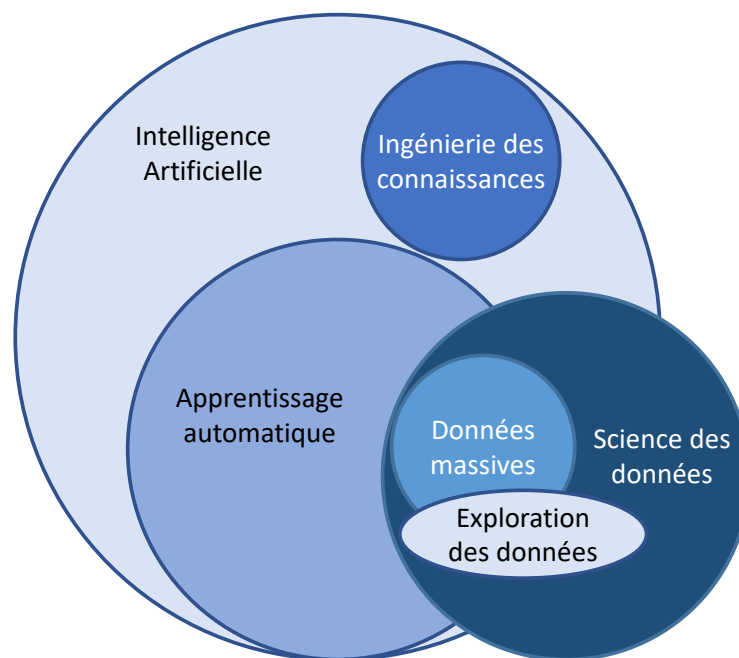


FIGURE 2.2 – Science des données vs intelligence artificielle vs données massives vs apprentissage automatique (adapté de [KULIN et al., 2021]).

2.1.1 Les données massives

Avant de définir les données massives et étudier la littérature associée, les notions de donnée, information, connaissance et sagesse sont explicitées.

2.1.1.1 Donnée, Information, Connaissance et Sagesse

La pyramide de la sagesse définie par [ROWLEY, 2007] et représentée par la figure 2.3 nous propose une hiérarchie entre ces quatre notions (*Data Information Knowledge Wisdom* appelée *DIKW* hiérarchie). Chaque terme est défini individuellement [ROWLEY, 2007] :

« Les données sont définies comme des symboles qui représentent les propriétés des objets, des événements et de leur environnement. Elles sont les produits de l'observation. Mais, elles ne sont utiles que lorsqu'elles sont sous une forme utilisable (c'est-à-dire pertinente). La différence entre les données et les informations est fonctionnelle, et non structurelle. »

« Les informations sont contenues dans des descriptions, des réponses à des questions qui commencent par des mots tels que qui, quoi, quand et combien. Les systèmes d'information génèrent, stockent, récupèrent et traitent les données. L'information est déduite des données. »

« La connaissance est un savoir-faire, et c'est ce qui rend possible la transformation de l'information en instructions. La connaissance peut être obtenue, soit par transmission d'une autre personne qui la possède, soit par instruction, soit par extraction d'une expérience. »

« La sagesse est la capacité d'accroître l'efficacité. La sagesse ajoute de la valeur, ce qui requiert la fonction mentale que nous appelons jugement ou vision. Les valeurs éthiques et esthétiques que cela implique sont inhérentes à l'acteur et sont uniques et personnelles. »

Cette pyramide offre une vision selon laquelle l'information se présente sous la forme de données qui sont structurées de manière à être utiles et pertinentes pour un objectif spécifique [ROWLEY, 2007]. L'identification d'un tel objectif est un élément commun à toutes les applications traitant des données massives examinées par [DE MAURO et al., 2016]. Dans cette perspective, l'information devient un actif de connaissance qui crée de la valeur pour les entreprises [CRICELLI et GRIMALDI, 2008]. Ils en concluent donc que l'information, et non les données, est le carburant fondamental du phénomène des données massives. Précisons que la sagesse est analogue à la compétence. Certains auteurs précisent que la sagesse est la compétence de l'expert [SWARTWOOD, 2013].

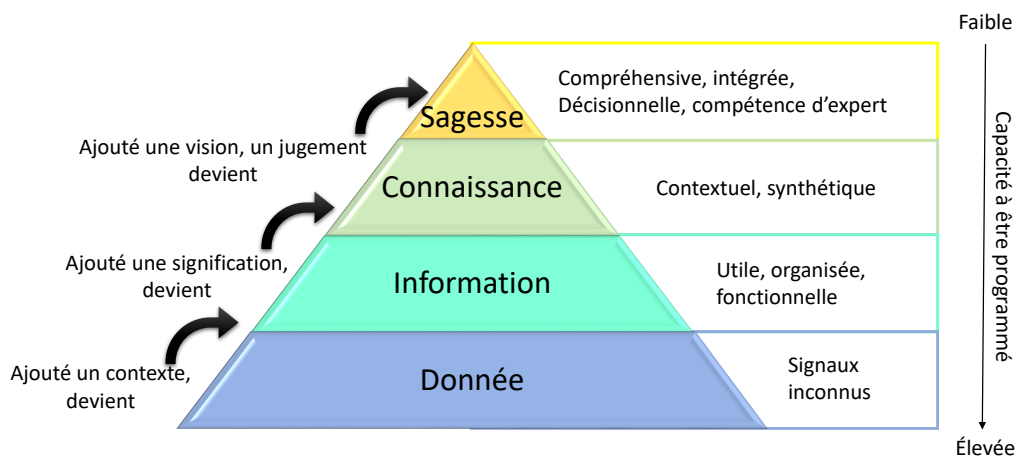


FIGURE 2.3 – Pyramide présentant la hiérarchie DIKW (données-information-connaissance-sagesse) (adaptée de [ROWLEY, 2007]).

2.1.1.2 Notion de "données massives"

Données massives

Les données massives font référence aux ensembles d'informations dont la taille ou le type dépasse la capacité des bases de données relationnelles traditionnelles pour extraire de la valeur à ces ensembles avec une faible latence [DE MAURO et al., 2016]. Pour les caractériser, il est souvent cité les 3V originaux : Volume, Vitesse et Variété auxquels s'ajoutent maintenant la Véracité et la Variabilité [REINSEL et al., 2018]. C'est pourquoi on parle aujourd'hui le plus souvent des 5V. Mais il faut aussi mettre en avant la notion de « Valeur » contenue dans la définition donnée précédente. Cette notion est un aspect important des données massives et décrit la transformation des données en informations susceptibles de créer une valeur économique, environnementale voire sociale pour les entreprises ou la société. Les données massives sont aussi des ressources c'est-à-dire les outils et les méthodes qui vont permettre d'utiliser et traiter les données. L'économie, les États, la société et les industriels s'appuient aujourd'hui fortement sur les données. Les données massives sont donc en train de changer la façon dont nous vivons, travaillons et jouons chaque jour.

Exploration des données ou *data mining*

La légende urbaine des « bières et des couches » est l'exemple le plus connu (car le plus ancien) de mise en œuvre d'exploration des données (appelé plus couramment *data mining*) dans l'industrie. Cette légende concernerait l'entreprise Walmart® qui aurait décidé de combiner les données de son système de carte de fidélité avec celles des systèmes des différents points de vente. Le premier système fournissait alors les données démographiques des clients et le second indiquait quand, où et ce qu'achetaient ces clients. Une fois combinées, ces données ont été exploitées et de nombreuses corrélations ont été découvertes. Par exemple, si un client achète du gin il y a de fortes chances qu'il achète du tonique. Cependant, une corrélation s'est démarquée : le vendredi soir, les hommes américains qui achètent des couches ont une prédisposition à acheter de la bière. L'entreprise aurait alors mis les couches et les bières côte à côte dans leurs magasins entraînant une augmentation des ventes de ces deux produits. Cette légende, dont on trouve les origines en 1992, est souvent reprise pour faire la différence entre interrogation et exploration des données [POWER, 2002]. L'interrogation des données est aisée avec les bases de données relationnelles, contrairement à l'exploration qui vit le jour avec les données massives. En effet, l'idée d'explorer les données vient avec le fait que les entreprises veulent que celles-ci apportent de la valeur à l'entreprise. La définition plus exacte est l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données par des méthodes automatiques ou semi-automatiques [BERRY et LINOFF, 1997]. Ce domaine est un sous-domaine de la science des données. Elle se porte à la frontière entre les données massives, l'intelligence artificielle et la science des données.

2.1.1.3 Les données massives dans l'industrie

Les données massives sont aujourd'hui indispensables dans de nombreux secteurs. Les entreprises du monde entier utilisent les données pour se transformer et devenir plus agiles, améliorer l'expérience client, introduire de nouveaux modèles commerciaux et développer de nouvelles sources d'avantages concurrentiels [REINSEL et al., 2018]. Aujourd'hui, les chercheurs s'efforcent

toujours d'améliorer leurs algorithmes pour traiter les données plus efficacement et introduisent la mise en œuvre des données massives dans de nombreux domaines autres que la finance et le commerce de détail, précurseurs. Les données massives ont changé la façon dont nous faisons des affaires, de la gestion et de la recherche. La science des données massives fait son apparition dans le monde et vise à fournir les outils dont nous avons besoin pour traiter les problèmes liés aux données massives [CHEN et ZHANG, 2014]. [SNIJDERS et al., 2012] pensent que les données massives sont un outil pratique pour creuser des microsavoirs rarement vus par les chercheurs en sciences sociales et les aider à trouver des modèles pour prédire les comportements sociaux ou même les crises sociales et économiques. [BABAR et al., 2019] proposent un système de gestion des données urbaines intelligentes basé sur Hadoop® dans l'optique de villes intelligentes et durables. Les données environnementales comme la quantité de dioxyde de carbone dans l'air sont suivies en temps réel.

Parfois, afin d'améliorer l'efficacité du traitement des données, il est nécessaire de modifier les formats des données pour les rendre plus accessibles aux ordinateurs. [HEINTZ et al., 2014] émet l'idée d'améliorer la gestion des connaissances du cadre décisionnel d'une entreprise chimique basée sur de simples notes en remplaçant ces simples notes par des annotations sémantiques et ontologiques qui sont non seulement lisibles par l'homme, mais aussi adaptées au traitement informatique. [AL-FUQAHA et al., 2015] suggèrent qu'une plateforme analytique données massives généralisée doit être construite. Elle fournit un service à l'application IoT, avec le *cloud computing* pour traiter ainsi qu'extraire des informations des données. Le *Fog Computing* lui sert de pont entre les appareils intelligents et les services de stockage et de *cloud computing* à grande échelle. Il est aussi possible de trouver une plateforme de données massives sémantique qui intègre des données hétérogènes de type *wearable* dans le domaine de la santé [MEZGHANI et al., 2015]. Les *wearables* (de l'anglais *wear* et *able* signifiant littéralement « ce qui peut se mettre ou se porter ») désignent les accessoires et vêtements équipés d'éléments informatiques et électroniques capables, entre autres, de mesurer et transmettre des données sur l'activité de l'individu qui les porte. Le cadre de la plateforme comprend trois couches : une couche de stockage de données dans laquelle la plupart des données sont stockées dans le *cloud*, une couche données massives qui tient compte du traitement des données et une couche de connaissances sémantiques pour comprendre les données par leur ontologie, ainsi que pour préparer les données afin qu'elles puissent être comprises et traitées par l'ordinateur. Les techniques d'analyse sont également revues et améliorées pour rendre les techniques des données massives plus efficaces et plus précises. [GANDOMI et HAIDER, 2015] passent en revue les techniques analytiques pour les données textuelles, audio, vidéo, les données des médias sociaux et l'analytique prédictive. Ils soulignent que l'analyse prédictive éclipse les autres formes d'analyses appliquées aux données de traitement et ils encouragent la promotion de nouvelles techniques statistiques et davantage d'innovations telles que l'analyse en temps réel.

Les données massives sont également utilisées pour aider l'écologie industrielle. [MING et al., 2015] discutent de ce que les données massives pourraient apporter à l'écologie industrielle. Selon eux, les données massives promeuvent la collaboration virtuelle entre les chercheurs, les aider à mieux mesurer les empreintes environnementales de la consommation humaine et les défis liés aux

données dans l'ACV, développer des modèles de systèmes plus réalistes et complexes, mieux capturer les caractéristiques essentielles des comportements dynamiques humains et affiner les modèles de métabolisme pour créer une symbiose réussie. [COOPER et al., 2013] utilisent les données massives pour compléter les données d'arrière-plan d'une analyse environnementale. Le système d'arrière-plan est constitué de tous les autres processus qui interagissent directement avec le système de premier plan.

2.1.1.4 Conclusion sur les données massives

Les données massives font partie intégrante des industries et de la société. À travers les IoT, les CPS, les *Cloud* et les données issues du Web Public, les industries sont entrées dans l'ère des données massives. Cette mise à disposition simple et instantanée des données permet aux industries de se transformer et devenir plus agiles. Cependant, elle entraîne un problème, le traitement de ces données. Les paragraphes précédents nous ont donné des éléments de réponses notamment avec le *data mining* ou certains exemples dans l'industrie d'utilisation de ces données. Si nous nous intéressons plus particulièrement aux données massives dans un souci de développement durable, le paragraphe 2.1.1.3 a mis en lumière l'utilisation des données massives pour aider la collaboration entre chercheurs ou pour compléter les données nécessaires aux analyses de la durabilité. En effet, les données massives, et plus particulièrement le Web Public, regorgent de données facilement disponibles. La littérature a cependant un manque d'utilisation des données massives dans un contexte de DD [BELAUD et al., 2019b]. Pour chaque analyse de la durabilité, les chercheurs n'utilisent que les données de leurs propres expériences ou de celles de leur laboratoire. Pourtant, les données issues des articles scientifiques comportent de nombreuses données de procédés qui nous intéressent particulièrement dans l'optique de comparer des procédés par analyse de la durabilité [LOUSTEAU-CAZALET et al., 2016]. Elles répondent en partie à notre première question : **Q1 - comment exploiter et tirer bénéfice des données massives pour aider les analyses et les évaluations du développement durable?** . Une fois les sources de données identifiées, il convient d'utiliser des méthodes pour extraire ces données et les traiter. Les outils et méthodes des données massives et de l'IA vont permettre cela.

2.1.2 Intelligence artificielle

2.1.2.1 Définition générale de l'intelligence artificielle

Le terme "Intelligence Artificielle" (IA) a été utilisé d'une manière formelle par John McCarthy lors de la conférence de Dartmouth en 1956 qui fait office d'origine de l'IA comme un domaine de recherche spécifique. L'IA est définie alors comme «la science et l'ingénierie de la fabrication d'une machine intelligente qui possède la capacité de calcul lui permettant d'atteindre des objectifs dans le monde humain» [MCCARTHY, 2007]. Il est difficile de définir exactement ce qu'est l'intelligence et cela reste un débat philosophique et scientifique encore aujourd'hui. Dans l'article "*Computing Machinery and Intelligence*" de 1950, Alan Turing affirme que « si la machine peut prétendre avec succès être humaine pour un observateur averti, alors vous devriez certainement la considérer comme intelligente » [TURING, 1950]. Appelée le "*imitation game*" par son créateur, il est maintenant appelé le Test de Turing. Et c'est la notion par laquelle la plupart des scientifiques définissent aujourd'hui l'intelli-

gence d'un programme informatique. Il n'y a aujourd'hui qu'un seul type d'IA : l'IA faible aussi dite étroite, mais la littérature nous présente deux autres types qu'on peut appeler évolution : l'IA forte et la superIA [KAPLAN et HAENLEIN, 2019]. La figure 2.4 décrit les différences entre ces trois types et les dates auxquelles un panel d'experts estime que ces IA vont émerger. La plupart des systèmes d'IA dont nous parlons aujourd'hui, sont des systèmes d'IA faible. Pour l'instant, diverses recherches sont menées en vue de la dissémination des systèmes d'IA, qui sont aujourd'hui fréquemment appliqués dans divers domaines tels que la reconnaissance des formes, l'automatisation, la vision par ordinateur, la réalité virtuelle, le diagnostic, le traitement d'images, le contrôle non linéaire, la robotique, le raisonnement automatisé, l'exploration de données, la planification des processus, les agents et contrôles intelligents, la fabrication, etc. [KUMAR, 2017].

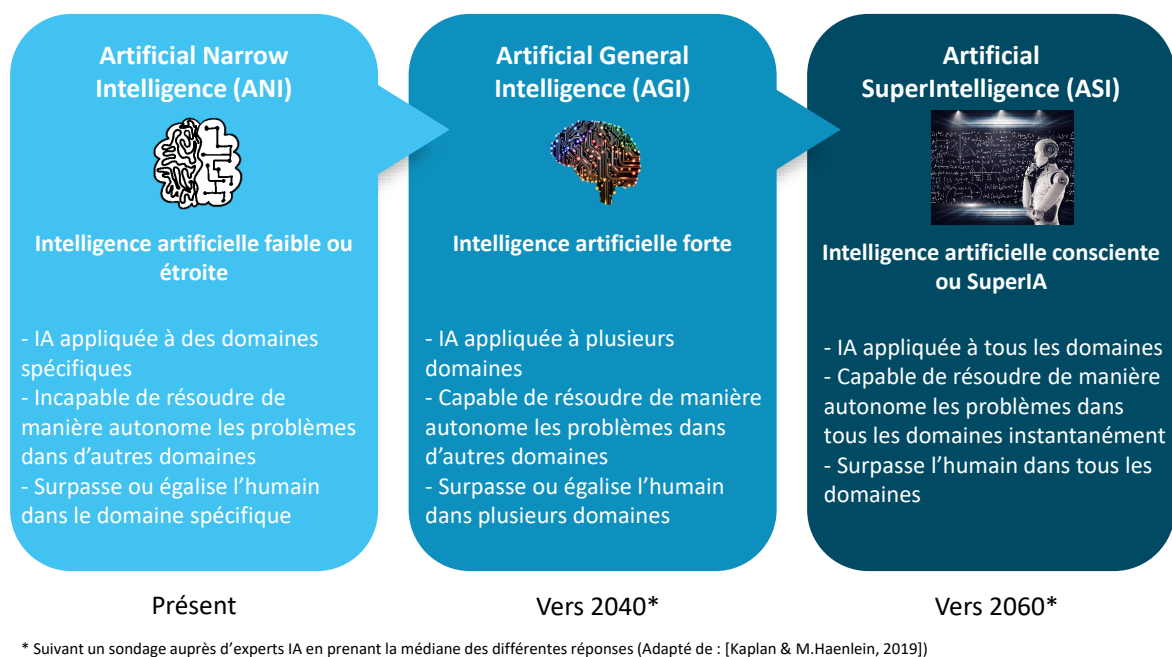


FIGURE 2.4 – Schéma décrivant les trois différents types d'intelligence artificielle (adapté de [KAPLAN et HAENLEIN, 2019]).

2.1.2.2 Les domaines de l'intelligence artificielle

Une revue rapide des principaux domaines de l'IA en détaillant plus précisément ceux qui sont utilisés dans nos travaux, est présentée ci-après.

2.1.2.2.1 Apprentissage automatique ou *Machine Learning*

L'apprentissage automatique, couramment appelé par son terme anglais *Machine Learning* (ML) repose sur l'usage d'algorithmes et de modèles statistiques par des systèmes informatiques (ou des ordinateurs) pour effectuer une ou plusieurs tâches spécifiques sans utiliser d'instructions explicites, mais en se basant plutôt sur des modèles et des inférences [ALPAYDIN, 2010]. Le ML donne aux machines la capacité d'apprendre à partir des données et d'améliorer leurs performances à résoudre des

tâches sans être explicitement programmées pour cela. Les problèmes d'apprentissage automatique sont définis de manière générale comme suit :



Machine Learning : On dit d'un joueur qu'il apprend de l'environnement en ce qui concerne une certaine classe de tâches A (par exemple, des actions) et une mesure de performance r (par exemple, l'état et la récompense), si sa performance aux tâches A, mesurée par r , s'améliore avec l'expérience.

La taxonomie des algorithmes ML est organisée comme dans la figure 2.5, en fonction du résultat souhaité, de l'entrée disponible et du style d'apprentissage de l'algorithme. Il y a trois grands types de ML : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement. Quant au *deep learning* ou apprentissage profond (sous-catégorie du ML), il s'agit d'une méthode qui s'inspire du système nerveux des êtres vivants (réseaux de neurones). Il intervient dans les trois types de ML : on parle par exemple d'apprentissage par renforcement profond. Avec l'énorme quantité de données générées chaque jour, l'apprentissage automatique est une technique clé qui a démontré sa capacité à traduire ces grands ensembles de données en connaissances exploitables [WARING et al., 2020].

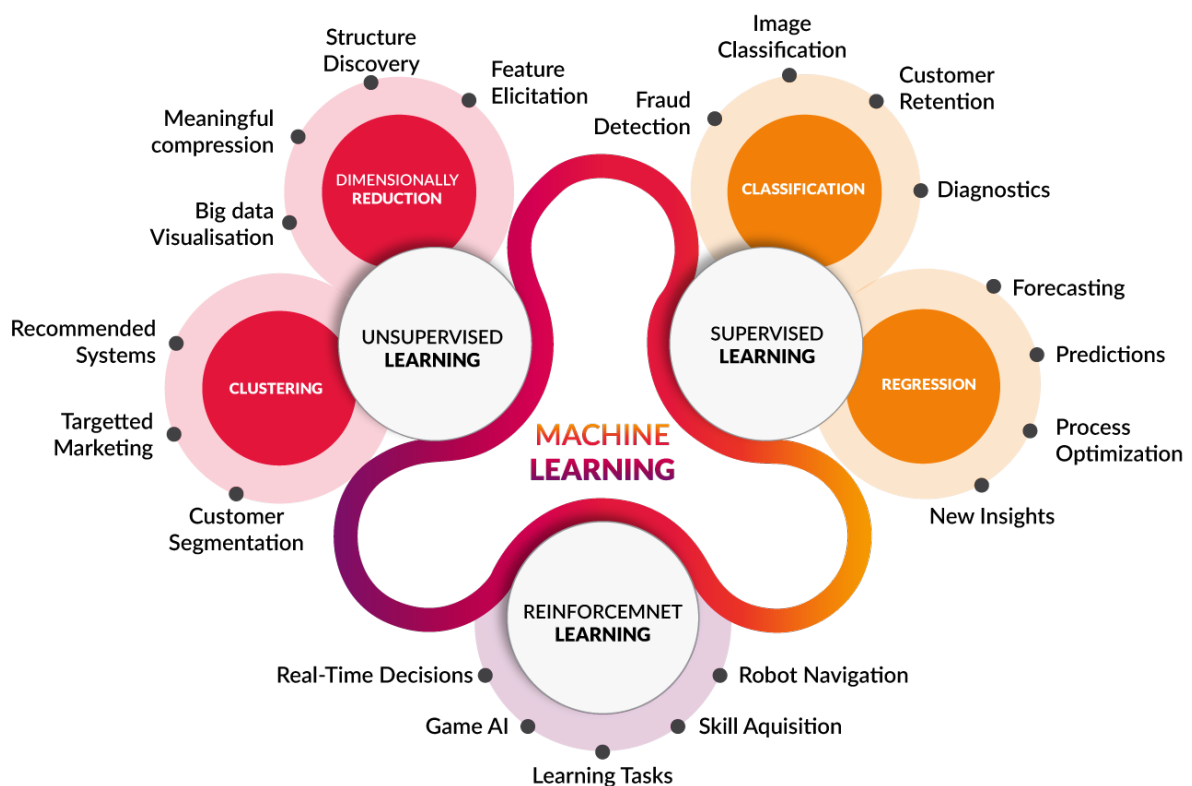


FIGURE 2.5 – Taxonomie des algorithmes d'apprentissage automatique (source : <https://cleanpng.com/Rusbert>).

Apprentissage supervisé Cet algorithme est composé d'une variable cible qui est définie pour être projetée à partir d'un ensemble particulier de variables indépendantes. En utilisant ces variables, des fonctions sont générées pour correspondre à la sortie souhaitée. Il fonctionne comme un processus d'apprentissage qui s'effectue en continu jusqu'à ce que la fonction parvienne à prévoir la sortie correspondante avec le niveau de précision requis [HINTON et SEJNOWSKI, 1999; MITCHELL, 1997]. Il existe de nombreuses méthodes d'apprentissages supervisés [CORNUÉJOLS et al., 2002] dont les machines à vecteurs de support [CORTES et VAPNIK, 1995], les arbres de décisions [MCCULLOCH et PITTS, 1943], les modèles bayésiens naïfs, ainsi que les réseaux de neurones artificiels [QUINLAN, 1986]. Nous aborderons plus en détail les réseaux de neurones dans le chapitre 4.

Apprentissage non supervisé Dans l'apprentissage non supervisé, il n'y a pas d'objectifs souhaitables à atteindre, ni de processus de formation pour corriger l'algorithme. Il est utilisé pour regrouper la population en différents groupes, ce qui est largement appliqué pour segmenter les clients en différents groupes pour une intervention spécifique et trouver des modèles [GIBBS et SU, 2002]. Il faut bien distinguer l'apprentissage supervisé et non supervisé. Dans le premier apprentissage, il s'agit d'apprendre à classer un nouvel individu parmi un ensemble de classes prédéfinies : on connaît les classes à priori. Tandis que dans l'apprentissage non supervisé, le nombre et la définition des classes ne sont pas donnés à priori.

Apprentissage par renforcement L'apprentissage par renforcement est considéré comme un algorithme renforcé de l'apprentissage supervisé. La principale différence entre l'apprentissage par renforcement et l'apprentissage supervisé est qu'il n'y a pas de résultat spécifique à atteindre pour former l'algorithme. Au contraire, l'algorithme est évalué et formé avec une autre fonction de récompense qui évalue le résultat et donne différents niveaux de récompense à l'algorithme en fonction du degré de satisfaction du résultat. Par exemple, aux échecs, le fait de "gagner" ou de "perdre" la partie peut être considéré comme une récompense. En utilisant de tels algorithmes, les machines sont entraînées à prendre des décisions spécifiques.

En complément, nous pouvons citer l'apprentissage semi-supervisé qui se situe entre l'apprentissage supervisé et non supervisé. Il utilise un ensemble de données étiquetées et non étiquetées améliorant la qualité de l'apprentissage. Un des modèles les plus utilisés en ML sont les réseaux bayésiens, combinaison de la théorie des probabilités et de la théorie des graphes. Ils sont des modèles graphiques probabilistes qui représentent de manière compacte la distribution conjointe d'un ensemble de variables. Ils sont surtout employés pour décrire des situations où l'information est incertaine, méconnue ou ignorée. Ils ont deux objectifs principaux :

- Ils permettent grâce à leur structure la découverte de connaissances. C'est ce que nous appelons l'interprétabilité et qui sera développée dans le chapitre 4.
- Ils permettent d'estimer la densité grâce à leurs paramètres.

Les chapitres 4 et 5 détailleront plus précisément des méthodes et algorithmes utilisés dans nos travaux.

2.1.2.2.2 *Natural Language Processing*

Le traitement du langage naturel (NLP) est un sous-domaine de la linguistique, de l'informatique, de l'ingénierie de l'information et de l'intelligence artificielle axé sur la communication entre les ordinateurs et les langues naturelles humaines. Elles visent à résoudre des problèmes tels que la programmation des ordinateurs pour traiter et analyser de grandes quantités de données en langage naturel. Deux missions principales sont associées au traitement du langage naturel : la compréhension du langage naturel et la génération du langage naturel.

2.1.2.2.3 **Systeme expert**

Un système expert est un système informatique qui émule la capacité de prise de décision d'un expert humain. Un système expert se compose, de façon simple, de 3 éléments :

- **Interface utilisateur** : Une interface utilisateur assure l'interaction entre le système expert et l'utilisateur.
- **Base de connaissances** : C'est l'endroit où le système expert stocke les connaissances. Il est principalement utilisé pour collecter et stocker l'expérience, les connaissances et le savoir livresque des experts dans un certain domaine, le bon sens, etc., y compris les pratiques, les faits et les règles réalisables.
- **Moteur d'inférence** : La machine de raisonnement est la partie centrale du système expert, qui est en fait un ensemble de programmes informatiques, utilisé pour simuler le processus de réflexion des experts, contrôler et coordonner le travail de l'ensemble du système expert. Selon les données initiales et les exigences de résolution du problème fournies par l'utilisateur, il utilise les faits et les règles de la base de connaissances, déduit le problème selon certaines méthodes de raisonnement et stratégies de contrôle, et fournit le résultat à l'utilisateur.

2.1.2.2.4 **L'ingénierie des connaissances**

Au cours des dernières années, tous les domaines ont été confrontés à une augmentation exponentielle de leurs données, principalement en raison de l'adoption généralisée de nouvelles technologies et des progrès des sciences de l'information. Cette croissance sans précédent a engendré des défis considérables quant à l'organisation et à l'analyse de ces données, plutôt que leur production. Dans les domaines scientifiques, la difficulté à appréhender et exploiter conjointement les données complique fortement leur traitement informatique. Ces difficultés découlent de divers aspects :

- (1) la grande dispersion des données scientifiques : les données sont éparpillées, elles sont contenues dans des bases de laboratoire, des revues scientifiques, des sites Web, des rapports de projet, des thèses ou des supports de cours ;
- (2) leur hétérogénéité marquée tant au niveau de leurs formats (textes, tableaux, graphiques, images, etc.) que de leurs vocabulaires ;
- (3) leurs échelles temporelles et spatiales variées ;
- (4) leur imprécision due à la prise en compte de la variabilité expérimentale ;

(5) leur fiabilité, en particulier la fiabilité des données du Web.

Depuis longtemps, ces problématiques ont suscité l'intérêt de nombreux domaines de recherche en informatique, notamment les bases de données, l'intégration de données, la représentation des connaissances et l'acquisition de données. Parmi eux, l'Ingénierie des Connaissances (IC) joue un rôle central et est ici considérée comme un sous-domaine de l'IA.

D'après [AUSSENAC-GILLES et al., 2012], l'IC englobe des concepts, méthodes et techniques visant à modéliser, formaliser et acquérir des connaissances au sein des organisations, dans le but de les opérationnaliser, les structurer ou les gérer de manière générale. L'IC trouve ainsi ses applications dans des domaines où l'objectif est de modéliser les connaissances et de les mettre à disposition pour soutenir des activités ou des raisonnements. Ces applications incluent la gestion des connaissances, la recherche d'information sémantique, l'aide à la navigation et l'aide à la décision. L'IC entretient également des relations étroites avec le Web Sémantique, partageant de nombreux outils et méthodes tels que les ontologies, les langages de représentation des connaissances et les raisonnements. Elle s'intéresse particulièrement à l'adaptation et/ou au développement de méthodes et d'outils innovants de capitalisation et de modélisation des connaissances et des données pour permettre leur exploitation, leur traitement, leur visualisation et leur partage par des systèmes informatiques, tant à l'échelle pluridisciplinaire, nationale qu'internationale. Dans un contexte où la gestion des données et des connaissances est en forte croissance, l'IC est devenue un enjeu stratégique majeur. Notamment dans les domaines des sciences du vivant et de l'environnement, où les systèmes informatiques développés devront soutenir les décideurs dans des problématiques mondiales telles que la sécurité alimentaire, la santé humaine, l'impact du changement climatique et la préservation de la biodiversité.

Pour construire un système de capitalisation et de modélisation des connaissances et des données provenant de sources hétérogènes, l'utilisation d'ontologies est l'une des multiples solutions possibles [DOAN et al., 2012; NOY, 2004]. Les ontologies permettent de spécifier des vocabulaires conceptuels standardisés, et l'indexation des sources de données avec le vocabulaire de l'ontologie assure leur interopérabilité. De nombreux travaux portent sur l'étude des ontologies à tous les stades de leur cycle de vie [STAAB S. et STUDER R, 2009] allant de leur construction [CIMIANO P. et al., 2010] à leur évolution en passant par leur alignement [BERNSTEIN P. A. et al., 2011; SHVAIKO P. et EUZENAT J, 2013]. De plus, les ontologies sont de plus en plus utilisées pour répondre à différents besoins. Elles sont de plus en plus utilisées en intégration de données pour garantir l'interopérabilité des sources de données et pour guider l'interrogation de données hétérogènes [CORBY O. et al., 2006]. Elles caractérisent en particulier la fiabilité de sources de données [DESTERCKE et al., 2013] ou tracent la provenance des données. Dans le domaine de la recherche, des ontologies spécifiques à des domaines particuliers sont créées, telles que celle proposée par [BUCHE et al., 2013b] pour l'agroindustrie. Enfin, l'explosion du Web de données dont la promesse est le partage à grande échelle des données liées [HEATH T. et BIZER C, 2011], ne fait que renforcer cette montée en puissance des ontologies dans le domaine de l'IA.

2.1.2.2.5 Planification

La planification et l'ordonnancement automatisés, parfois appelés simplement planification, sont une branche de l'intelligence artificielle qui concerne la réalisation de stratégies ou de séquences d'actions, généralement destinées à être exécutées par des agents intelligents, des robots autonomes et des véhicules sans pilote.

2.1.2.2.6 Combinaison avec d'autres technologies

L'IA est mis en œuvre avec de nombreuses autres technologies pour renforcer sa capacité à extraire diverses informations. Elle est souvent utilisée avec les technologies de "vision par ordinateur" et de "récupération d'informations musicales" pour réaliser la fonction d'observation et d'écoute. Ces dernières années, face à l'explosion des données, elle est aussi fréquemment associée aux techniques des données massives.

2.1.2.3 L'intelligence artificielle dans le secteur industriel

Dans les industries, en particulier les industries de production, où l'automatisation des calculs est le plus souvent nécessaire, l'IA est un atout. Cette partie a pour but de rassembler succinctement les recherches en lien avec nos travaux et qui concernent les utilisations actuelles de l'IA dans différentes industries. L'industrie du génie des procédés est aussi une grande utilisatrice d'IA pour le contrôle des procédés ou l'optimisation des procédés, par exemple. [COMMENGE, 2020] fait un rapide état de l'art de ce qui existe dans la littérature scientifique sur les couplages données massives, IA et génie des procédés.

[MAO et al., 2019] ont discuté de la possibilité d'appliquer l'IA pour promouvoir la fabrication verte et intelligente. Ils ont analysé les principaux problèmes rencontrés aujourd'hui dans le secteur de la fabrication verte et les ont découpés en trois catégories : l'intégration des informations, l'évaluation dynamique des risques et l'aide à la prise de décision ainsi qu'une alerte précoce. Ils ont recommandé l'utilisation de graphes de connaissances pour intégrer les informations et relier toutes les connaissances, l'utilisation de réseaux bayésiens pour l'évaluation des risques et la prise de décision et l'application de l'apprentissage profond pour reconnaître les modèles d'accidents potentiels et émettre des alertes rapides.

De nombreuses recherches ont été menées afin d'améliorer la productivité de différentes industries et de faciliter l'interaction des personnes avec les machines grâce à l'IA. [HUANG et FAN, 1993] ont construit trois systèmes d'IA distincts utilisés à des fins différentes pour le traitement des déchets dans l'industrie de transformation. [PANG et al., 2020] ont développé un mélange de systèmes experts et de systèmes d'autoapprentissage pour détecter et corriger efficacement les défauts qui existent dans les produits en contreplaqué. Grâce à ce système, les ordinateurs trouvent les défauts des produits en comparant les images prises par la caméra à la base de données du système expert, qui proposera ensuite une solution pour réparer automatiquement ces zones défectueuses par des robots et les résumer dans la base de données afin d'améliorer constamment le plan de production. De nouvelles approches de planification IA comme HiTraP-AT (*Hierarchical Transition Planner*

for Automation Technology) pour le domaine de l'automatisation des machines et des usines sont conçues [LEGAT et VOGEL-HEUSER, 2017]. Il s'agit d'un système de planification configurative. Il est basé sur une combinaison d'une formulation de planification ajustée. [HAGEMANN et al., 2019] a présenté une nouvelle méthodologie d'IA qui génère automatiquement la configuration initiale du système de production sur la base de scénarios industriels réels dans le domaine de la production de carrosseries automobiles. L'IA améliore aussi le système de gestion de la construction afin que les utilisateurs puissent s'interfacer directement avec le système en langage naturel [KLASHANOV, 2016]. L'IA n'augmente pas seulement la productivité des industries, elle sert également d'outil essentiel dans la gestion des risques industriels. [HOU et al., 2020] ont tenté d'utiliser une méthode d'apprentissage automatique pour évaluer les facteurs de caractérisation de l'écotoxicité des produits chimiques manquants dans USEtox, un modèle Évaluation des impacts du cycle de vie (EICV) approuvé par les Nations Unies. [AHN et al., 2019] ont appliqué l'IA pour améliorer un système de détection précoce des fuites de pipeline. Dans les cas normaux, il est très difficile de détecter les fissures potentielles qui existent loin des capteurs, car il y a trop de bruits qui pourraient influencer le diagnostic. Afin d'éliminer ces bruits, l'équipe a mis en œuvre un système d'apprentissage automatique basé sur l'algorithme génétique pour sélectionner les caractéristiques et prétraiter les signaux collectés par les capteurs. Elle a comparé ce système avec le système PCA (analyse en composantes principales), plus souvent utilisé, et il s'est avéré qu'il était plus précis.

En outre, l'IA fait également l'objet de nombreuses recherches pour aider à prévoir les conséquences potentielles de différentes décisions et aider les décideurs à prendre des décisions plus prometteuses et plus stables. [CHOU et al., 2010] ont utilisé un système d'IA hybride avec 9 modèles différents pour prévoir le coût de développement des équipements de fabrication des écrans TFT-LCD et ont comparé leurs résultats. Dans certains cas, le système d'IA a amélioré la précision des performances de 46,62 %. [GLIGOR et al., 2018] ont utilisé un réseau de neurones artificiels à action directe pour prévoir la production d'électricité de l'unité de production photovoltaïque afin de soutenir la prise de décision d'une usine de production d'énergie en Roumanie. De plus, puisque l'IA, dans de nombreux cas, doit souvent prendre des décisions nécessaires en temps réel, [TERZIYAN et al., 2018] ont suggéré qu'une technologie Pi-Mind pourrait être introduite et devenir un bon compromis entre la prise de décision pilotée par l'IA et la prise de décision pilotée par l'homme. Pi-Mind est un ensemble de techniques, de modèles et d'outils visant à imiter le comportement décisionnel humain dans différentes situations, ce qui permet de faire des choix en fonction des préférences personnelles uniques.

2.1.2.4 Conclusion sur l'intelligence artificielle

L'intelligence artificielle est donc de plus en plus utilisée dans les industries pour aider la décision, traiter les données, intégrer des informations ou évaluer des risques. L'IA est transverse à plusieurs domaines dont nous avons décrit les objectifs principaux. Deux domaines particuliers ont attiré notre attention : l'ingénierie des connaissances et l'apprentissage automatique. Le paragraphe 2.1.1 a identifié une source de données majeure, les articles scientifiques, mais ne précise pas de méthodes pour extraire les données et les informations. Cela ne répond pas entièrement à notre première question.

La description de l'IC, et notamment des ontologies, nous semble être un bon moyen de répondre à la partie manquante de cette question. Le ML analysant de nombreuses données de différentes dimensions est une autre réponse à cette première question. Le ML complète notre deuxième question en mettant en lumière les algorithmes de réduction de dimension et de *clustering* qui vont aider à l'analyse des résultats des évaluations de la durabilité.

2.2 La pensée cycle de vie pour le développement durable

Le chapitre 1 a abordé succinctement l'évaluation du DD et les très nombreuses évaluations qui existent. Il a été précisé que dans un contexte d'EC les évaluations les plus pertinentes sont les évaluations conformes à la pensée cycle de vie. Dans cette partie, nous décrivons cette « pensée » ainsi que les méthodes associées avant d'expliquer le choix de la méthode utilisée dans ce manuscrit.

2.2.1 La pensée cycle de vie

La pensée cycle de vie (PCV) (en anglais *Life Cycle Thinking* (LCT)) est un état d'esprit dans lequel nous cherchons à identifier les améliorations possibles des biens et services sous la forme d'une empreinte environnementale plus faible et d'une utilisation réduite des ressources à toutes les étapes du cycle de vie. Cela commence par l'extraction et la transformation des matières premières, puis la fabrication et la distribution, jusqu'à l'utilisation et la consommation. Il se termine par la réutilisation, le recyclage des matériaux, la récupération d'énergie et l'élimination finale. Tous les impacts sont pris en compte. La mise en place d'un DD exige de profonds changements dans la façon de penser, dans les structures économiques et sociales des modèles de consommation et de production. Cela signifie l'implication nécessaire du domaine scientifique, technologique et industriel, et la création de méthodologies pour évaluer la durabilité dans chaque cas spécifique. La PCV contribue à améliorer les performances environnementales, les avantages sociaux et économiques des biens et des services en tenant compte de leur cycle de vie complet, en évitant le transfert de charges. Cela signifie minimiser les impacts environnementaux à un stade du cycle de vie d'un produit tout en évitant d'autres impacts par ailleurs. La figure 2.6 montre une approche tournée EC du cycle de vie d'un produit qui part des ressources, va vers la fabrication, la livraison, l'utilisation et l'élimination finale (recyclage ou fin de vie). Chaque étape a des échanges avec les ressources naturelles.

La PCV identifie également les opportunités qui conduisent à des décisions visant à améliorer les performances environnementales, l'image et les avantages économiques. Les entreprises ne prennent pas toujours en compte leurs chaînes d'approvisionnement ou les processus d'utilisation et de fin de vie associés à leurs produits. Les actions gouvernementales se concentrent souvent sur un domaine spécifique, et non sur les impacts ou les avantages qui se produisent dans d'autres domaines et qui sont attribuables à leurs propres niveaux de consommation. Dans ce cas, si l'on ne tient pas compte du cycle de vie complet des biens et des services (c'est-à-dire l'approvisionnement, l'usage et la fin de vie), le seul élément qui souffre est l'environnement, ce qui se traduit par un potentiel de dommages plus élevé à différents niveaux.

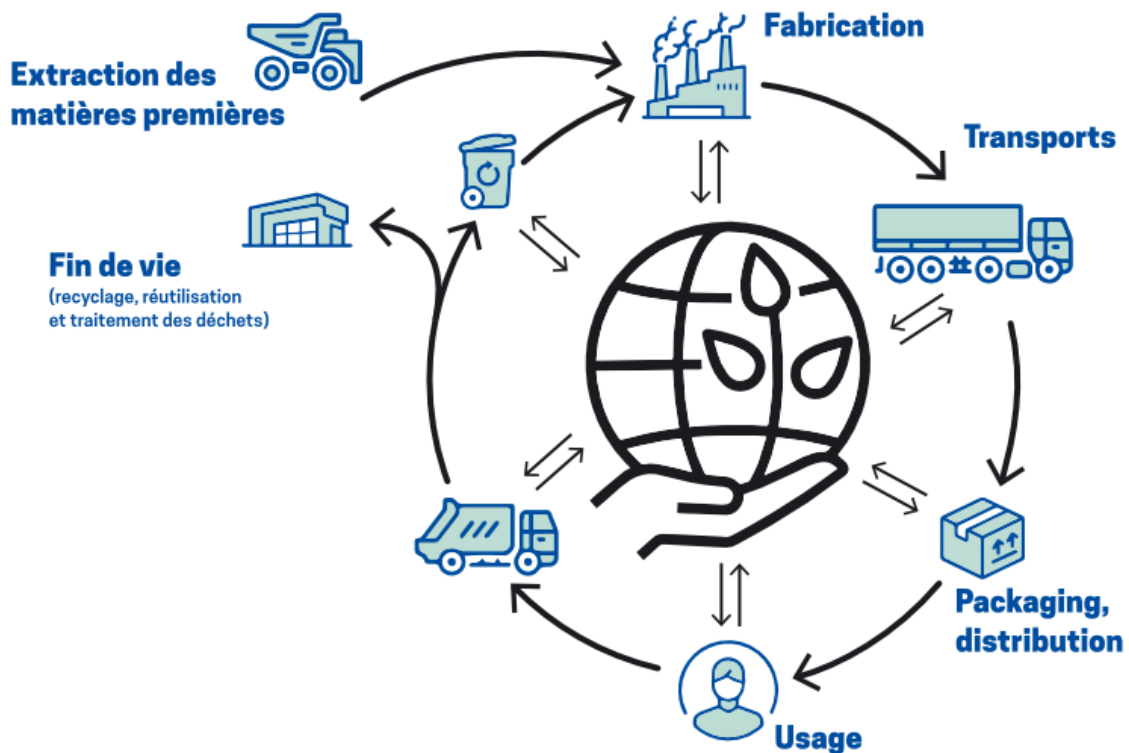


FIGURE 2.6 – Cycle de vie d'un produit dans un contexte d'économie circulaire (source : chaire ELSA-PACT).

2.2.2 Méthodes conformes à la pensée cycle de vie

Selon l'initiative de l'UNEP sur le cycle de vie, la PCV est un mode de pensée qui inclut les conséquences économiques, environnementales et sociales d'un produit ou d'un processus sur l'ensemble de son cycle de vie [GUINÉE et al., 2011]. Il existe plusieurs méthodologies pour évaluer et analyser ces impacts. Certaines d'entre elles sont listées dans le tableau 2.1. Parmi celles-ci, l'Analyse du Cycle de Vie (ACV) est une méthode d'aide à la décision systémique et multicritère. Cette vue systémique de l'ACV est tout l'intérêt de cette méthode, elle prend en compte tous les impacts sur l'environnement, l'économie et la société liés au cycle de vie d'un système (produit, procédé ou service). L'ACV est historiquement orientée vers les impacts environnementaux liés au cycle de vie d'un produit. Les normes qui l'encadrent font d'ailleurs partie des normes de management environnemental. Ces dernières années, elle a été étendue aux impacts sociaux et économiques afin de rencontrer les objectifs de durabilité. Aujourd'hui, le terme « ACV » peut avoir plusieurs significations : il est utilisé à la fois pour caractériser un domaine d'étude, une méthode ou un cas spécifique d'étude. Dans le tableau 2.1, les méthodes ACV spécifiques aux trois sphères sont citées (l'analyse environnementale du cycle de vie, l'analyse des coûts du cycle de vie, l'analyse sociale du cycle de vie) et une méthode ACV les regroupant (l'analyse de la durabilité du cycle de vie).

De manière synthétique, l'annexe A présente les analyses environnementales citées dans le tableau

TABLEAU 2.1 – Liste de méthodes utilisant la pensée "cycle de vie" adaptée de [JOLLIET et al., 2010]

Type des méthodes	Nom des méthodes
Analyse environnementale	Analyse environnementale du cycle de vie Analyse des flux de matières Analyse des impacts environnementaux Analyse des risques environnementaux Analyse des exergies Apport de matière par unité de service
Analyse économique	Analyse coûts-bénéfices Analyses des coûts du cycle de vie Analyse économique des entrées-sortie
Analyse sociale	Analyse sociale du cycle de vie
Analyse de la durabilité	Analyse de la durabilité du cycle de vie

2.1, afin de relever les avantages et les inconvénients de chaque méthode.

2.2.3 Analyse du cycle de vie

L'ACV est une méthode d'aide à la décision basée sur une approche rigoureuse et transparente, en constante évolution et amélioration [JOLLIET et al., 2010]. Son intérêt réside dans la prise en compte des impacts environnementaux, économiques et sociaux tout au long du cycle de vie d'un produit, d'un procédé ou d'un service. Cela inclut toutes les étapes, de l'extraction des matières premières à la fin de vie, en passant par le transport, la fabrication et l'utilisation. L'ACV est encadrée par les normes internationales de management environnemental ISO 14040 et ISO 14044. La première norme définit le cadre général de la méthodologie, tandis que la seconde décrit en détail chacune des étapes [14040 :2006, 2006; 14044 :2006, 2006]. Bien que ces normes se concentrent principalement sur les aspects environnementaux, leur structure peut être étendue pour inclure les aspects économiques et sociaux. Suivant ces normes, l'ACV se divise en 4 étapes interagissant de manière itérative (figure 2.7) :

- 1) **Définition des objectifs et du champ de l'étude :** Cette étape est cruciale car elle oriente toutes les étapes suivantes. Elle consiste à définir le cadre dans lequel l'analyse sera réalisée. On doit tout d'abord identifier les raisons qui motivent la réalisation de l'ACV et décrire l'étendue des résultats attendus. Ensuite, il est nécessaire de délimiter le périmètre de l'étude selon deux aspects. D'une part, il faut sélectionner les limites fonctionnelles du système, généralement définies « de la porte à la porte », « du berceau à la porte » ou du « du berceau à la tombe » [JACQUEMIN et al., 2012]. D'autre part, il faut choisir les limites « stratégiques » représentées par ce qu'on appelle l'ACV attributionnelle ou l'ACV conséquentielle. La première correspond à l'ACV classique, qui évalue l'ensemble des impacts d'un produit (ou d'un procédé ou d'un service) tout au long de son cycle de vie [EARLES et HALOG, 2011].]. L'ACV conséquentielle, quant à elle, s'appuie sur l'ACV attributionnelle et élargit le champ de l'étude aux conséquences d'un choix stratégique qui entraînerait une variation de la demande du produit, entraînant une augmentation ou une diminution des impacts [ZAMAGNI et al., 2012]. Par ailleurs, l'ACV repose sur la quantification du service rendu par le produit (ou le procédé ou le service) grâce à la définition de l'unité fonctionnelle, qui correspond à une unité de service rendu. Cela permet de comparer les impacts de différents produits, services ou procédés en tenant compte du service rendu

équivalent. Enfin, le choix d'une règle d'allocation des flux doit être effectué en fonction des critères d'importance mis en avant entre les différents flux, tels que des critères économiques, environnementaux ou massiques. Toutes les autres hypothèses de l'étude sont formulées à cette étape.

- 2) **Inventaire du Cycle de Vie (ICV) en anglais *Life Cycle Inventory (LCI)*** : Cette étape, parfois fastidieuse, consiste à rassembler et quantifier tous les flux de matière, d'énergie, de devises et de personnes associés à chaque processus élémentaire du cycle de vie. Il est essentiel d'exprimer tous ces flux en fonction de l'unité fonctionnelle. Cette étape est souvent divisée en deux parties : l'inventaire des données d'arrière-plan et l'inventaire des données de premier plan. Le système de premier plan correspond aux processus sous le contrôle du décideur pour lequel l'ACV est réalisée. Le système d'arrière-plan regroupe tous les autres processus qui interagissent directement avec le système de premier plan [CLIFT et al., 1998; ELGHALI et al., 2007]. L'ICV sert de référence pour calculer les impacts associés à chaque flux inventorié. Pour simplifier la complexité des données, il est nécessaire d'effectuer une analyse critique de leur qualité, qui servira de base pour l'analyse de l'incertitude et le calcul de la propagation des erreurs [14040 :2006, 2006].
- 3) **Évaluation des impacts du cycle de vie (EICV) (en anglais *Life cycle impact assessment (LCIA)*)** : Basée sur l'inventaire du cycle de vie et en accord avec les hypothèses formulées lors de la définition initiale de l'ACV, cette étape classe et caractérise les flux préalablement inventoriés dans des catégories d'impacts [14040 :2006, 2006]. Ces catégories d'impact varient en fonction des aspects environnementaux, économiques et sociaux.
- 4) **Interprétation des résultats** : C'est la dernière étape de l'ACV, dont l'objectif est d'établir un lien entre les trois premières étapes. Elle extrait, à partir des résultats d'ICV et d'EICV, des conclusions générales et des recommandations en fonction des objectifs initiaux. À ce stade, il est possible de revoir certains choix effectués lors des trois étapes précédentes. Ainsi, l'ACV apparaît comme une méthode itérative [14040 :2006, 2006]. Dans le cadre général de l'analyse de cycle de vie, chacun des trois piliers du DD conduit à trois types d'ACV.

En outre, il est important de noter qu'une itération de la méthode ACV réduit les incertitudes des résultats jusqu'à ce qu'ils répondent aux exigences fixées lors de la première étape de définition des objectifs de l'étude. La figure 2.7b illustre ces itérations sous la forme d'une spirale de rétrécissement.

2.2.3.1 L'analyse environnementale du cycle de vie (AeCV)

L'Analyse environnementale du Cycle de Vie (AeCV) se concentre sur les impacts environnementaux, qui sont évalués à l'aide d'indicateurs. On distingue deux types d'impacts : les impacts intermédiaires (ou *midpoint*) et les dommages (ou impacts finals ou *endpoint*). Les impacts intermédiaires correspondent à des effets physico-chimiques et biologiques mesurables causés par les flux de matière et d'énergie (par exemple, l'acidification ou le réchauffement climatique). Les dommages, quant à eux, représentent les conséquences de l'augmentation des effets physico-chimiques et biologiques sur des éléments à protéger (les trois plus couramment utilisés étant la santé humaine, les ressources non renouvelables et les écosystèmes) [JOLLIET et al., 2010]. Le choix des indicateurs environnementaux pour chaque catégorie est guidé par des « méthodes » d'évaluation des impacts. En fonction de

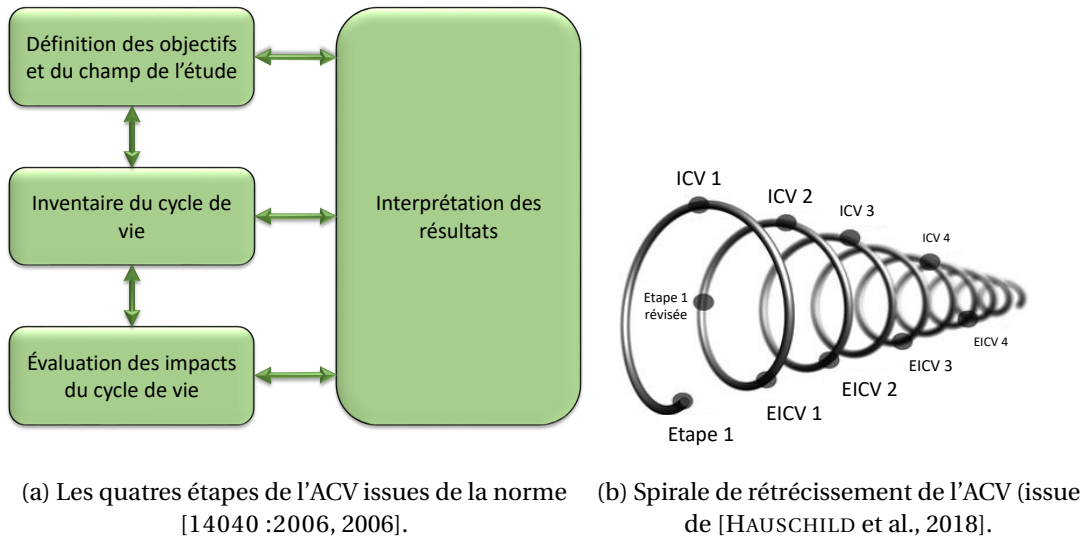


FIGURE 2.7 – La méthode ACV : les quatre étapes et les itérations.

la méthode utilisée, il est possible de calculer les impacts intermédiaires et/ou finaux, puis éventuellement de normaliser, pondérer et agréger les résultats. Le choix d'une méthode spécifique dépend des objectifs initiaux de l'analyse. Cependant, des experts en ACV ont rédigé un article dans lequel ils analysent et comparent les méthodes de caractérisation des impacts. Ils en tirent des recommandations concernant le choix des méthodes de caractérisation à utiliser [HAUSCHILD et al., 2013].

Les impacts *midpoint* concernent le premier niveau d'agrégation d'une méthode d'évaluation des impacts, tandis que les impacts *endpoint* correspondent au second niveau. Pour passer du premier niveau au second, on utilise des facteurs de pondération appelés « facteurs de dommage ». Dans le but d'obtenir un critère environnemental unique, certaines méthodes d'évaluation proposent de normaliser les indicateurs de dommages, puis d'en faire la somme pondérée. La notion d'indicateur et de critère est relative au niveau d'agrégation étudié. En effet, si les critères sont les catégories intermédiaires, leurs indicateurs sont les flux. En revanche, si les critères sont les catégories de dommages, les catégories intermédiaires deviennent des indicateurs.

Il est important de souligner que les catégories les plus robustes à évaluer sont celles liées aux effets globaux, tels que le réchauffement climatique ou l'appauvrissement de la couche d'ozone [HAUSCHILD et al., 2013]. En revanche, les catégories qui suscitent actuellement de nombreuses recherches concernent davantage les problématiques régionales ou locales, telles que l'eutrophisation ou l'épuisement des ressources en eau. De nombreuses études sont également menées sur la monétarisation des différents impacts *midpoint*, en attribuant un poids à chaque impact en fonction de l'étude, permettant ainsi de comparer plusieurs produits ou procédés qui n'ont pas la même unité fonctionnelle [CAIARDI et al., 2022].

2.2.3.2 L'analyse économique du cycle de vie (AcCV)

L'analyse économique du cycle de vie appelée Analyse des Coûts du Cycle de Vie (AcCV) (ou en anglais *Life Cycle Costing* (LCC)) est considérée comme le volet le plus ancien de l'ACV puisqu'elle a été développée et standardisée aux États-Unis à partir de la Seconde Guerre mondiale. D'autres

théories de l'AcCV ont été développées dans le domaine de l'ingénierie système [SWARR et al., 2011]. L'AcCV vise à évaluer les coûts engendrés par un produit (ou un service ou procédé) sur l'ensemble de son cycle de vie. Elle ne se limite pas au prix de vente, car ce dernier ne reflète pas les coûts d'utilisation et de fin de vie. Les coûts évalués par l'AcCV peuvent être internes ou externes. Les coûts internes sont ceux qui reviennent directement au producteur ainsi qu'à l'utilisateur-consommateur. Les coûts externes sont ceux qui incombent aux autres parties prenantes, comme la société ou les concurrents. Ainsi, bien que le cadre méthodologique de l'AcCV soit similaire à celui de l'ACV, le périmètre de l'étude s'élargit dans l'AcCV pour inclure les coûts liés aux phases de R&D ainsi que les coûts associés à des flux non physiques, donc non pris en compte dans AeCV [JOLLIET et al., 2010]. Les coûts du producteur incluent la valeur ajoutée, qui couvre les salaires et les bénéfices. En ce qui concerne le consommateur, les coûts d'utilisation et de fin de vie sont pris en compte.

2.2.3.3 L'analyse sociale du cycle de vie (AsCV)

Par définition, une Analyse sociale du Cycle de Vie (AsCV) (aussi appelée analyse sociétale ou socio-économique) est « une technique d'évaluation des impacts sociaux et socio-économiques (réels et potentiels) positifs et négatifs tout au long du cycle de vie des produits » [UNEP, 2011]. Contrairement à l'AeCV et l'AcCV, qui se concentrent sur les produits, procédés ou services, l'AsCV se focalise sur les comportements [FESCHET, 2014]. Si l'AsCV nécessite plus de temps et de recherche pour être standardisée, mais la plupart des études s'appuient sur le guide de l'UNEP-SETAC [UNEP, 2011], qui décrit les étapes de réalisation calquées sur celles de la norme ISO 14040 [14040 :2006, 2006]. Cependant, peu d'articles ont été publiés sur l'AsCV en raison principalement de sa complexité de mise en œuvre [LEHMANN et al., 2013]. Bien que de nombreux frameworks pour créer une AsCV aient été développés ces dernières années, sa réalisation reste un défi [VAN HAASTER et al., 2017]. Au niveau de l'inventaire, les opinions divergent : certains experts estiment qu'une collecte de données aussi détaillée et précise que celle réalisée dans une AeCV n'est pas applicable à l'AsCV [JØRGENSEN et al., 2010]. D'autres soutiennent que l'utilisation de données génériques provenant de bases de données statistiques conduira à des estimations trop approximatives et trop grossières pour certains impacts sociaux [WEIDEMA, 2006].

2.2.3.4 L'analyse de la durabilité du cycle de vie (AdCV)

L'Analyse de la durabilité du Cycle de Vie (AdCV) (en anglais *Life cycle sustainability assessment* (LCSA)) intègre les résultats de l'AeCV, de l'AcCV et de l'AsCV pour fournir une vision globale d'un système dans le contexte de la durabilité. Selon certains auteurs, l'AdCV représente l'avenir de l'ACV, en soulignant la nécessité de développer simultanément les trois volets et de les intégrer pour promouvoir le développement de produits véritablement durables plutôt que simplement écologiques. Le couplage de ces trois pôles avec d'autres outils d'ingénierie est également fortement recommandé [GUINÉE et al., 2011]. Dans l'ensemble, le cadre établi par la norme ISO 14040, constitué des quatre étapes précédemment décrites, semble convenir à l'application de l'AdCV. Cependant, des défis scientifiques et techniques subsistent quant à son intégration, notamment l'intégration de l'AeCV, AcCV et l'AsCV. Deux grandes propositions s'opposent (ou se complètent) concernant l'AdCV. La première proposition préconise de développer une seule ACV intégrant dès le début de l'inventaire l'environnement, les coûts et les enjeux sociaux [KLOEPFFER, 2008]. La seconde proposition suggère

de réaliser parallèlement les trois ACV pour les trois piliers de la durabilité, en utilisant les mêmes limites et la même unité fonctionnelle pour le système. Dans ce cas, l'AdCV revient à « sommer » les trois ACV : environnementale, des coûts et sociale. Cette somme est pondérée ou non selon les objectifs de l'étude et l'importance accordée à chaque pilier [FINKBEINER et al., 2010; KLOEPFFER, 2008]. Les étapes d'inventaires et d'évaluations d'impacts se retrouvent multipliées par trois et on se retrouve avec de nombreux impacts à analyser pour essayer d'interpréter les résultats. Cette deuxième proposition s'inscrit dans la notion de durabilité faible, contrairement à la première proposition qui s'inscrit davantage dans une perspective de durabilité forte.

2.2.4 Conclusion sur le choix de l'ACV

Par définition, l'ACV est une méthode multicritère et systémique qui évalue les impacts environnementaux. Le fait qu'on puisse utiliser aussi une planification à long terme et que les ACV dynamiques se développent est un avantage à ce type d'approche [ER-RAKI, 2021]. L'ACV permet d'avoir une vue sur de nombreux impacts différents et donc d'identifier les domaines d'amélioration qui ont la plus grande influence sur les impacts totaux du cycle de vie. Le manque de données limite les conclusions que l'on tire d'une étude spécifique et c'est la plus grande limite de cette méthode. L'ajout de méthodes de ML et issues de l'IC va réduire cette limite. Un des avantages de l'ACV est son développement dans les trois sphères du DD à travers des méthodes dédiées à chacune ou à une méthode combinée. Cette dimension n'est pas prise en compte par les autres méthodes qui sont souvent spécifiques à une ou deux sphères maximum.

Avec l'ajout de méthodes d'IA et de ML, la multiplication des bases de données d'inventaires et celle des impacts ne constituent pas forcément un frein et garde le plus d'information possible tout au long de notre cadre méthodologique. L'ACV est un domaine en constante évolution. Des ACV particulières voient le jour comme les ACV territoriales ou les ACV transitoires [VENTURA, 2022]. L'ACV territoriale est une méthode réalisant une évaluation environnementale au niveau d'un territoire et identifiant des orientations d'aménagement plus respectueuses de l'environnement dans un contexte local d'aide à la décision. Ce type d'ACV donne une vision globale du système territorial et évalue les effets cumulatifs dus à la diversité des activités présentes sur ce même territoire [LOISEAU, 2014]. Pour toutes ces raisons, notre choix pour une analyse de la durabilité repose sur l'ACV.

2.3 Développement durable, intelligence artificielle et données massives

En septembre 2015, à l'occasion de son 70ème anniversaire, l'ONU a adopté « Transformer notre monde : Développement durable 2030 » à son agenda. L'objectif est de parvenir à un développement durable mondial qui vise à rassembler tous les pays et les parties prenantes pour interrompre le déclin de la planète. C'est-à-dire qu'ils mettent en avant la production et la consommation durable, et la gestion de l'environnement et des ressources naturelles de la Terre pour permettre à la planète de répondre aux besoins des générations actuelles et futures. L'Agenda 2030 des Nations Unies pour le développement durable est un plan d'action stratégique ambitieux, qui englobe les dimensions économiques, sociales et environnementales. Il se compose de 17 objectifs de développement durable et de 169 cibles. Pour atteindre ces objectifs, il est nécessaire de donner toute sa place à la

science et à la technologie dans la réalisation des objectifs de développement durable. Parmi les nombreuses technologies, celle de la science des données joue incontestablement un rôle prépondérant. Elle est appliquée à chaque phase du cycle de vie pour la durabilité des industries.

Pendant la phase de fabrication, l'IA a de nombreuses utilisations liées à la durabilité. [SCHORR et al., 2020] utilisent la méthode d'apprentissage automatique de la forêt aléatoire pour prédire les caractéristiques de qualité géométriques et dimensionnelles des alésages sur la base des données de processus, ce qui conduit à une fabrication durable. [ÁLVAREZ RODRÍGUEZ et al., 2020] proposent une méthode d'évaluation qui combine l'analyse d'enveloppement des données de réseau dynamique et l'évaluation du cycle de vie pour évaluer la durabilité, l'efficacité économique et sociale de la chaîne d'approvisionnement de détail. [ZANGHELINI et al., 2018] effectuent un examen sur la façon dont l'analyse décisionnelle multicritères est utilisée pour aider l'analyse du cycle de vie dans l'interprétation des résultats. Ils soulignent que l'Analyse Décisionnelle MultiCritères (ADMC) est le plus souvent utilisée à deux endroits pour faciliter l'interprétation de l'ACV : au niveau de l'ICV et au niveau des résultats de l'EICV. D'un point de vue conceptuel, l'ADMC est introduit dans le cadre et les normes de l'ACV en tant qu'étape de "pondération" pour parvenir à un jugement global de la performance environnementale à partir des nombreux indicateurs évalués dans l'ACV. La principale raison d'intégrer l'ADMC à l'ACV est de traiter les résultats de l'ACV en parallèle avec les aspects économiques, sociaux et techniques. À l'heure actuelle, la méthode du surclassement est la plus couramment utilisée dans le cadre du ADMC, tandis que la méthode WSA (*Weighting Sum Approach*) et la méthode AHP (*Analytic Hierarchy Process*) jouent également un rôle important. Il est possible aussi d'utiliser l'analyse de métarégression pour analyser les résultats de diverses études d'ACV sur les émissions de GES afin d'identifier les principaux facteurs de variabilité E-S (*Effect-Size*) et d'estimer à la fois la direction et l'ampleur de leurs effets respectifs dans les études primaires considérées [MENTEN et al., 2013]. Trois facteurs qui pourraient avoir un impact sur cette variation du résultat ont été identifiés et quantifiés : les données techniques, les choix méthodologiques de l'auteur et la typologie de l'étude. [XU, 2020] introduit un cadre de priorisation de la durabilité qui est utilisé pour prendre des décisions rapides dans les systèmes énergétiques et qui intègre des méthodes des groupes flous pour une considération de données hybrides. Un cadre méthodologique couplant l'analyse du cycle de vie et l'analyse d'enveloppement de données de réseau dynamique pour aider à l'évaluation et à l'amélioration de l'efficacité, de la durabilité et de la rentabilité de la chaîne d'approvisionnement de détail a été créé par [ÁLVAREZ RODRÍGUEZ et al., 2020]. [MOHEBALI et al., 2020] appliquent l'analyse d'enveloppement des données pour améliorer l'EICV afin de prendre davantage en compte l'impact économique et social du projet d'usine de lavage du charbon. [LINKE et al., 2019] fournissent un modèle d'exemple sur la façon d'inclure l'IA dans la fabrication. Ils suggèrent qu'une durabilité réussie axée sur les données dans la fabrication devrait se concentrer sur l'utilisation de l'IA dans l'amélioration de trois aspects clés : l'assurance qualité, la conception des machines-outils et l'expertise des travailleurs. La nécessité d'appliquer l'IA pour aider à prédire l'impact du cycle de vie d'un produit agricole a été aussi mise en avant [KAAB et al., 2019]. La comparaison de deux types de techniques d'intelligence artificielle, le réseau neuronal artificiel et le système d'inférence floue neuronal adapté est effectué pour évaluer leur précision dans la prévision des indicateurs environnementaux et de la production d'énergie de sortie pour la production de

canne à sucre dans les fermes plantées et les fermes de ratons.

Pendant la phase de vente au détail et de consommation, l'IA est souvent utilisée pour trouver des modèles cachés entre la durabilité et les comportements de consommation des gens et les stratégies de vente au détail. [YE et al., 2020] appliquent des techniques des données massives, telles que l'analyse sans modèle, pour analyser l'énorme quantité de données fournies par les deux plus grands détaillants en ligne de Chine. Ils étudient l'effet d'entraînement positif que provoque la consommation durable sur les comportements de remboursement des microprêts et la continuité des comportements de consommation durable ultérieurs. [GONG et al., 2019] appliquent l'analyse d'enveloppement des données avec des modèles empiriques pour évaluer la relation entre les activités durables et l'efficacité des performances au niveau des opérations. Ils utilisent deux flux pour caractériser les activités commerciales, les flux financiers et les flux physiques, et découvrent qu'une entreprise avec un niveau plus élevé d'intégration financière et un niveau plus faible d'intégration physique dans la durabilité sont plus susceptibles d'atteindre un niveau plus élevé d'efficacité de vente au détail tout en maintenant leur objectif durable. [ABDELLA et al., 2020] introduisent une évaluation systémique de la durabilité basée sur des techniques d'apprentissage automatique supervisé qui prend en compte les trois lignes de fond de la durabilité, à savoir la société, l'économie et l'environnement, pour analyser la durabilité de la consommation alimentaire.

L'IA est régulièrement utilisée pendant la phase de traitement des déchets et de gestion des affaires afin de promouvoir la durabilité. [EUGENE et al., 2019] exposent trois modèles pour appliquer l'IA en aidant à réaliser l'ingénierie de la molécule au système pour le traitement durable de l'eau qui comprend l'aide à la conception inverse pour les matériaux, l'aide à la conception d'expériences et permet une analyse hiérarchique rapide. [ACCORSI et al., 2018] appliquent un cadre développé à partir d'une architecture multiprofondur pour effectuer une analyse pilotée par les données qui prend en compte les aspects micro et macro de la chaîne d'approvisionnement alimentaire pour soutenir la prise de décision. [NEGRA et al., 2020] proposent de nouvelles stratégies et exigences pour les communautés scientifiques afin de co-développer des indicateurs scientifiques et d'autres outils de décision pour aider les investisseurs du secteur agroalimentaire à gérer plus durablement. [RAUT et al., 2019] utilisent l'analyse des données massives pour analyser les différents prédicteurs de la performance durable des entreprises dans les pays en développement afin de fournir un aperçu pratique aux entreprises manufacturières pour améliorer leur gestion de la durabilité. [BAG et al., 2020] proposent d'utiliser la théorie de la capacité dynamique et le modèle d'équation structurelle des moindres carrés partiels pour analyser les données d'enquête des cadres miniers afin d'évaluer le rôle de l'analyse des données massives dans l'amélioration de la chaîne logistique durable. Leurs résultats montrent que la gestion et l'analyse des données massives jouent un rôle plus important dans la promotion du développement durable que les innovations basées sur les données massives. [GHOLIZADEH et al., 2020] démontre un modèle d'écodurabilité multiobjectif pour la chaîne d'approvisionnement. C'est un programme non linéaire à nombres entiers mixtes qui se base sur un modèle stochastique flou robuste pour résoudre les problèmes des données massives, contrôler les incertitudes et les risques et une méthode ϵ -contrainte augmentée pour résoudre le problème multiobjectif. [WU et al., 2020] discutent de l'utilisation de services basés sur le cloud, intégrés

avec des ressources de crowdfunding dans les données massives pour fournir des solutions moins chères pour la surveillance et le suivi des objectifs de développement durable, et les défis futurs connexes. [MANGINA et al., 2020] étudient les microdonnées anonymisées de l'enquête européenne sur le transport routier de marchandises afin d'acquérir des modèles d'opérations logistiques basés sur plus de 11 millions de trajets, avec différents algorithmes (coopération horizontale, mise en commun et Internet physique), pour analyser l'efficacité des opérations logistiques de la chaîne d'approvisionnement et améliorer la durabilité.

Les applications potentielles du ML en ACV existent dans la littérature, mais les algorithmes utiles dépendent de l'objectif, du type de données et de la taille de l'ensemble de données. Il est ainsi possible de trouver le développement d'une ACV de substitution à l'aide d'algorithmes de réseaux de neurones [PARK et SEO, 2003; SOUSA et WALLACE, 2006], des estimations des catégories d'impacts avec une régression linéaire multivariée [SLAPNIK et al., 2015] ou des analyses de sensibilité avec des régressions linéaires [HEALTH et al., 2009]. En outre, le ML est appliqué en dehors de l'ACV dans le nettoyage des données, la prédiction des flux de sortie ou des performances des systèmes, l'informatique des écosystèmes et l'optimisation des systèmes. À partir de ces méthodes, les algorithmes de ML pourraient avoir de nouvelles applications : filtrage ou nettoyage des ensembles de données pour l'ICV, estimation des données de flux pour les processus unitaires, amélioration de la qualité et de la quantité des données utilisées pour déterminer les flux de matière et génération de scénarios optimisés.

Si nous nous concentrons sur les méthodes d'analyse de la durabilité et les données massives, la littérature apporte des enseignements. [COOPER et al., 2013] proposent l'utilisation de méthodes issues des données massives pour compléter les données d'arrière-plan des ACV. La méthode DILCA (*Data-intensive Life Cycle Assessment*) exploite l'IC et intègre les développements technologiques qui modifient les résultats de l'ACV pour un produit donné au fil du temps [BHINGE et al., 2015]. Les données massives et l'IC sont utilisés pour représenter le cycle de vie d'un produit ou d'un service [ZHANG et al., 2015], avec tous les flux intermédiaires, les émissions et les extractions dans une ACV basée sur l'ontologie.

Le couplage entre l'ACV, l'IA et les données massives commencent à apparaître dans la littérature mais reste très faible et surtout appliqué à une étude particulière. Il n'existe pas d'étude qui propose une approche globale de ce couplage. Or, comme explicité dans les paragraphes précédents l'ACV bénéficie de l'IA et des données massives à n'importe laquelle de ses étapes et il est possible d'envisager un cadre méthodologique global de couplage.

2.4 Procédés et systèmes industriels et développement durable

L'ingénierie des procédés et systèmes industriels (PSI) est un domaine académique et technologique lié aux méthodes d'aide à la décision pour l'ingénierie bio-physico-chimique. Elles sont chargées d'indiquer comment planifier, concevoir, exploiter, contrôler tout type de système de l'opération unitaire à l'industrie elle-même en passant par le procédé chimique [KLATT et MAR-

QUARDT, 2009]. Le domaine des PSI s'est rapidement développé depuis les années 1950, reflétant la croissance des industries pétrolières, gazières et pétrochimiques. La définition élargie des PSI fournie par [GROSSMANN et WESTERBERG, 2000] est la suivante : « l'amélioration des processus de prise de décision pour la création et le fonctionnement de la chaîne d'approvisionnement chimique. Il s'agit de la découverte, de la conception, de la fabrication et de la distribution de produits chimiques dans le contexte de nombreux objectifs contradictoires ». Cette définition englobe l'ensemble de la chaîne d'approvisionnement chimique (du niveau moléculaire au niveau de l'entreprise). Cela a conduit l'intégration de nouveaux facteurs : ceux de la sécurité, de l'environnement et les facteurs économiques. En raison de leur vaste champ d'application et de leur approche systémique sous-jacente, les PSI sont devenus un instrument puissant pour aborder les problèmes de durabilité et guider efficacement l'élaboration des politiques durables [BAKSHI et FIKSEL, 2003; GROSSMANN et HARJUNKOSKI, 2019]. Cela nécessite d'étendre les limites traditionnelles des PSI au-delà du procédé et de l'entreprise pour inclure le cycle de vie et les systèmes économiques et écologiques associés [SIKDAR, 2003]. De plus en plus de méthodes et d'outils efficaces et performants sont créés pour concevoir des systèmes industriels durables [MILLET et al., 2016].

Au niveau moléculaire, certains travaux ont porté sur l'estimation de l'impact environnemental des produits chimiques à partir de descripteurs facilement disponibles dans la pratique, comme première étape vers la conception de molécules ayant une empreinte environnementale minimale [CALVO-SERRANO et al., 2018; WERNET et al., 2008]. L'analyse simultanée de la phase de synthèse et d'utilisation des produits chimiques réduit les impacts environnementaux de la conception intégrée des molécules et des procédés [BURGER et al., 2015]. Les aspects environnementaux de la conception des procédés sont le problème de durabilité le plus étudié en PSI [DIWEKAR, 2005; EL-HALWAGI, 1997] même si l'industrie des procédés chimiques est aussi confrontée à des problèmes économiques et sociaux très importants [COUNCIL, 2003]. La topologie et les conditions de fonctionnement optimales d'une usine de procédés sont déterminées de manière à maximiser simultanément les performances économiques et environnementales dans l'espace de recherche défini par un ensemble de contraintes. Il existe de nombreuses formations qui placent l'intensification des procédés dans leur programme et la recherche est de plus en plus nombreuse sur ce sujet [FERNANDEZ RIVAS et al., 2020]. En effet, l'intensification des procédés est une approche récente de l'ingénierie qui a démontré qu'elle pouvait améliorer considérablement l'efficacité et la sécurité des procédés tout en réduisant les coûts [COMMENGE et al., 2017]. Il existe une grande quantité de littérature basée sur les méthodes MINLP (*Mixed Integer Non Linear Programming*) [GROSSMANN et al., 1999] dans l'intégration de la chaleur [DURAN et GROSSMANN, 1986] et de la matière [KARUPPIAH et GROSSMANN, 2006] pour une consommation minimale d'énergie et d'eau, toutes deux ayant de fortes implications sur la durabilité. En outre, des travaux récents ont porté sur l'intégration des énergies renouvelables dans les usines chimiques afin de réduire davantage la consommation de combustibles fossiles [SÁNCHEZ et MARTÍN, 2018]. Une multitude de mesures ont été employées pour quantifier les performances de durabilité dans la conception des procédés [NEGNY et al., 2012; ČUČEK et al., 2012], mais la tendance récente est de s'appuyer sur des indicateurs d'ACV pour l'écoconception des procédés [MORALES-MENDOZA et al., 2018].

Au niveau de l'entreprise, suivant les approches d'optimisation à l'échelle de l'entreprise

[GROSSMANN, 2005], l'accent a été mis sur l'intégration des activités d'acquisition, de production, de stockage et de distribution des matières premières en tenant compte simultanément des préoccupations environnementales et économiques. Ce sont les méthodes MILP (*Mixed Integer Linear Programming*) qui sont utilisées principalement [MALLIER et al., 2021] même si d'autres méthodes voient le jour [HEINTZ et al., 2014]. À ce niveau, dans un concept d'EC, les chaînes d'approvisionnement en biomasse reçoivent une attention considérable [YOU et WANG, 2011].

La tendance actuelle conduit à appliquer les outils des PSI au-delà du génie chimique et vers de nouveaux domaines d'application. Par exemple dans le secteur agricole, nous pouvons citer la recherche opérationnelle prometteuse sur les problèmes de planification [LÓPEZ-DÍAZ et al., 2018], l'adoption du concept d'EC [BELAUD et al., 2017; COBO et al., 2018], la conception et l'exploitation d'installations de capture et d'utilisation du carbone [CUÉLLAR-FRANCA et AZAPAGIC, 2015], entre autres. L'approche principale de ces études consiste à élargir la portée de l'analyse au-delà des installations de production pour comprendre les vastes implications des décisions d'ingénierie sur l'environnement. La prise en compte des interactions complexes entre tous les sous-systèmes fournit des informations pour la prise de décision. Cela génère aussi des connaissances qui sont intégrées dès les phases de développement des procédés en vue de réduire les impacts à toutes les échelles. Dans le secteur d'automobile apparaissent des outils qui évaluent les impacts d'une innovation sur un procédé pré-existant [GARCIA et al., 2015]. On ne parle plus d'évaluer les impacts seulement pour la conception ou le pilotage mais aussi pour toute innovation ou changement sur un procédé comme dans le cas d'une bioraffinerie [CAIARDI et al., 2022].

Comme mentionné dans la section 2.2.3, l'ACV est une méthode analytique bien établie pour quantifier les impacts de la durabilité des produits ou des procédés. Bien que l'ACV n'ait pas été développée par la communauté PSI, elle partage sa pensée systémique, car elle applique des bilans massique et énergétique à un système de contrôle donné avec des limites élargies, c'est-à-dire qu'elle couvre des étapes au-delà de la phase de production. Cela est un point d'entrée pour un couplage des PSI avec l'ACV [GILLANI et al., 2010; JACQUEMIN et al., 2012]. L'ACV a de nombreuses applications dans de multiples domaines. L'approche standard consiste à comparer un nombre réduit de scénarios, en matière d'impact sur le cycle de vie, afin d'identifier les points chauds et de fournir des recommandations d'amélioration [HELLWEG et MILÀ I CANALS, 2014]. Les premières tentatives d'intégration de critères de durabilité dans les outils PSI se sont concentrées sur la réduction de la consommation d'énergie [LINNHOFF, 1993] et d'eau [WANG et SMITH, 1994]. Au niveau de l'usine, ces consommations ont des impacts environnementaux majeurs. Bien que cette approche ait souvent conduit à des solutions ayant une empreinte environnementale plus faible, elle néglige un large éventail d'impacts et surtout elle ne se concentre que sur l'étape de production. L'émergence de l'ACV dans la communauté des sciences environnementales a donc fourni un cadre puissant pour surmonter ces limitations. Dans un travail pionnier [AZAPAGIC et CLIFT, 1999] ont couplé l'ACV à l'optimisation des procédés, ce qui a conduit à une optimisation multi-objectifs minimisant les impacts du cycle de vie en complément du coût. De nombreux travaux ont depuis continué à fleurir pour lier l'ACV aux outils des PSI. On pourra citer des méthodes alliant la simulation de procédés et l'ACV [JULIO et al., 2017], ou encore l'utilisation d'ACV dans des cadres méthodologiques pour des écoparks industriels [LOISEAU, 2014].

Nous avons pu voir que l'ACV fait partie intégrante des outils PSI. Il reste encore des défis méthodologiques à relever, comme les différences d'échelle entre les ACV et les PSI ou la création de méthodes plus efficaces pour résoudre les algorithmes multiobjectifs par exemple, et de nombreux domaines d'application à explorer. L'ajout d'algorithmes venant de l'IA et notamment du ML est en train d'apparaître pour perfectionner et pour mieux comprendre les problèmes de durabilité dans les PSI. Le secteur de l'agro-industrie et plus particulièrement la conception des bioraffineries est l'un des domaines où les PSI et l'ACV sont le plus couplés [BUSSET et al., 2015; CALVO-SERRANO et al., 2018; GONG et al., 2019; JULIO et al., 2017]. Si nous ajoutons que les bioraffineries rentrent dans une vision d'EC, le choix de notre cas d'étude s'est ainsi tourné vers l'agro-industrie et les bioraffineries avec la valorisation des déchets lignocellulosiques issus de l'agriculture.

2.5 Conclusion du chapitre 2

Dans ce chapitre bibliographique découpé en quatre grandes parties, nous nous sommes intéressés à l'IA, les données massives, le DD et les PSI ainsi que le couplage entre ces quatre domaines. La première partie définit les données massives et l'IA, les méthodes qui leur sont attachées ainsi que certaines applications que l'on trouve dans les industries. Cela nous a permis de proposer comme sources de données décrivant des procédés le Web Public à travers notamment des articles scientifiques, l'IC pour extraire les données et connaissances de ces articles et le ML pour traiter et analyser ces données. C'est à travers ces trois aspects que nous répondrons à nos deux premières questions. La seconde partie a fait un rapide tour d'horizon des différentes méthodes d'analyses du DD, méthodes que l'on avait restreintes aux méthodes PCV. Cela nous a permis de faire le choix de l'ACV comme analyse de la durabilité dans nos travaux et de la définir. La troisième partie a mis en avant la littérature récente sur l'utilisation des données massives, de l'IA pour le DD dans les industries. Le manque de méthode globale utilisant l'IA et les données massives pour « enrichir » l'ACV et servir l'écoconception et la promotion du DD a été ainsi mis en avant. Ces constats ont été les prémisses de ces travaux. Enfin, le dernier paragraphe de cette section s'est intéressé aux PSI à travers une rapide littérature des PSI aux services du DD. Le couplage des PSI et de l'ACV reste récent dans la littérature et le domaine des bioraffineries est un enjeu révélateur. Le choix d'un cas d'étude des procédés de prétraitement de la biomasse lignocellulosique tombe alors sous le sens dans un contexte EC.

De cet état de l'art, la figure du chapitre 1 (section 1.3) illustrant notre approche se complète avec les méthodes et outils retenus ainsi que le cas d'étude (figure 2.8).

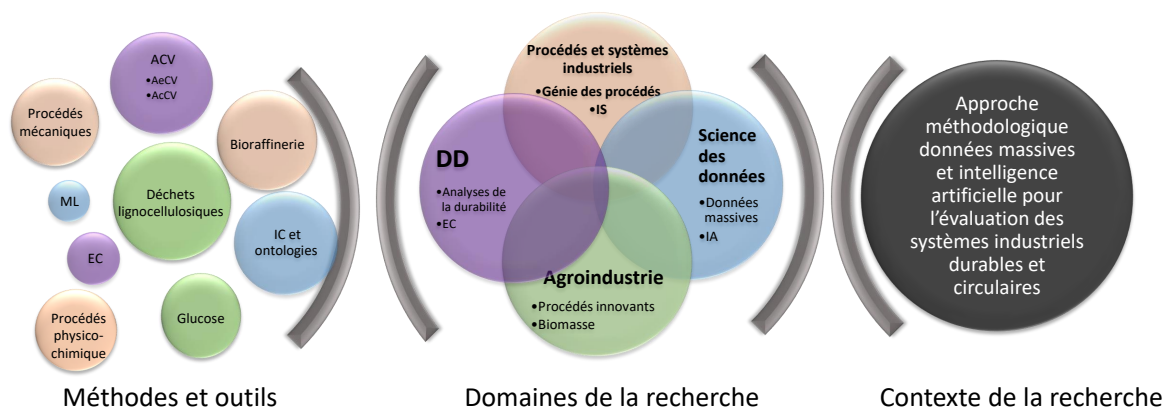


FIGURE 2.8 – Contexte, domaines, méthodes et outils de notre approche.

Chapitre 3

Définition du cadre méthodologique

Space and time are the framework within which the mind is constrained to construct its experience of reality.

Emmanuel Kant - Critique de la Raison pure, 1781

Ce chapitre (figure 3.1) est dédié à la présentation du cadre méthodologique centré sur les données massives et l'intelligence artificielle pour aider les analyses de la durabilité dans un contexte d'économie circulaire. La première partie décrit de façon générale le cadre méthodologique composé de cinq étapes. La deuxième partie détaille ces étapes à la lumière du travail bibliographique présenté dans le chapitre 2. La description est plus succincte pour les étapes 2 (Architecture des données) et 4 (Analyse et visualisation des résultats) comportant l'ajout des méthodes issues des données massives et de l'intelligence artificielle. Elle seront détaillées respectivement dans le chapitre 4 et 5. Ce chapitre est également l'occasion de mettre en avant les bénéfices de ce cadre méthodologique.

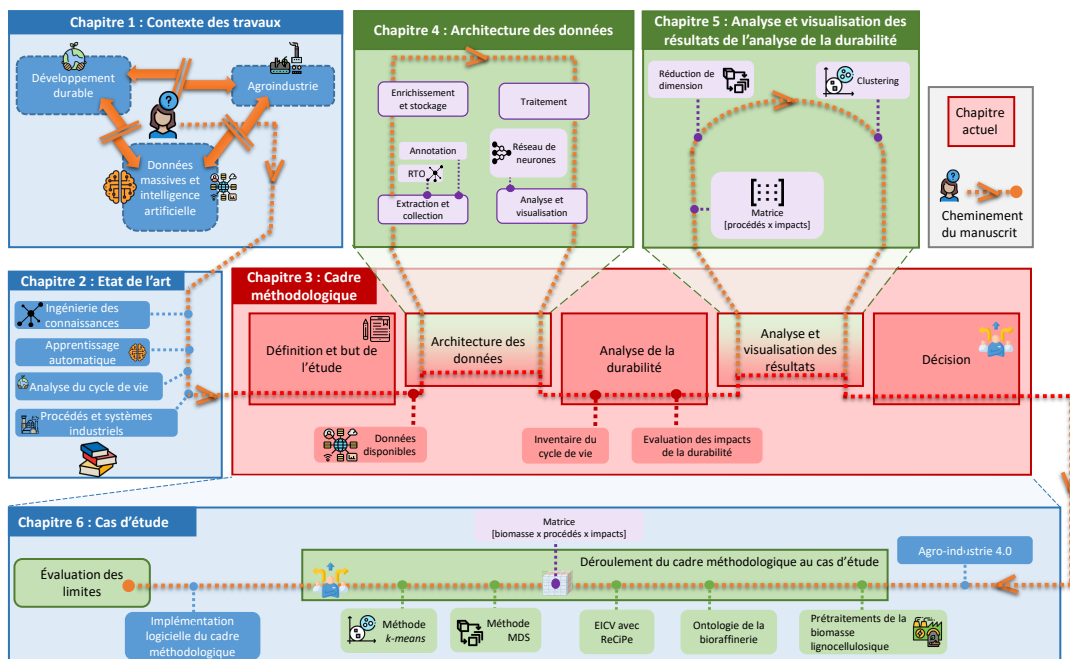


FIGURE 3.1 – Positionnement du chapitre 3 dans le manuscrit.

Sommaire

3.1 Introduction du chapitre 3	57
3.2 But et frontières de l'étude	60
3.3 Architecture des données	61
3.3.1 Données disponibles	61
3.3.2 Présentation générale des sous-étapes de l'étape d'Architecture des données	62
3.4 Analyse de la durabilité	63
3.4.1 Inventaire du cycle de vie	64
3.4.2 Évaluation des impacts du cycle de vie	66
3.4.2.1 Évaluation des impacts environnementaux du cycle de vie	66
3.4.2.2 Les logiciels en ACV	68
3.4.2.3 Qualité des données et analyse de la sensibilité	69
3.4.3 Conclusion sur l'étape d'analyse de la durabilité	70
3.5 Analyse et visualisation des résultats	72
3.6 Décision	73
3.7 Conclusion du chapitre 3	73

3.1 Introduction du chapitre 3

Les chapitres précédents ont évoqué la transition digitale de nos industries ainsi que leur prise en compte du DD. L'IA et les données massives sont au centre de la transition digitale et les analyses de la durabilité une partie du DD. La littérature nous a permis d'observer des couplages entre ces trois domaines, mais qui restent très spécifiques à un domaine d'application spécifique. Peu ou pas de méthodes génériques de couplage existent alors qu'un cadre méthodologique générique serait bénéfique à la fois à l'industrie, mais aussi à la recherche. Un cadre méthodologique étant pour rappel une base structurelle de l'intégration des concepts méthodologiques, méthodes et outils.

Les trois domaines d'intérêt de nos travaux : l'I4.0 (avec les données massives et l'IA), l'industrie et le développement durable sont mis en lumière dans la figure 3.2. Sur cette figure, seule la phase opérationnelle est représentée dans l'industrie. La chaîne d'approvisionnement gouvernée par l'ingénieur industriel peut entamer une transition écologique avec l'ajout d'analyses de la durabilité et une transition digitale avec les données massives et l'IA. Chacune des catégories de la chaîne d'approvisionnement peut être décrite avec des données hétérogènes et peut influencer une autre catégorie. Par exemple, le type d'entrant peut influencer le type de procédés de transformation. Les données influencent les indicateurs sociaux, économiques et environnementaux. Ces données constituent une partie des données de l'entreprise. En ce qui concerne la durabilité de la chaîne d'approvisionnement, l'objectif est d'intégrer toutes les données dans la conception du procédé. Par exemple, pour les évaluations environnementales conformes à la « pensée cycle de vie », des données sur les procédés sont nécessaires. L'obtention de ces données prend beaucoup de temps et nécessite des expériences coûteuses. Il est également possible d'obtenir des données à partir de publications scientifiques et d'autres sources, avec l'utilisation automatique ou manuelle des données massives. Cette figure met en avant les difficultés d'interaction des trois domaines qui ont chacun un expert différent. Nos travaux, à travers le cadre méthodologique développé dans ce chapitre, ont pour but de définir une proposition pour affronter ces difficultés. Ce cadre a été conçu pour un chercheur ou un ingénieur R&D qui voudrait exécuter une première analyse des trajectoires technologiques industrielles répondant à un besoin et à un contexte. Il peut être déployé pendant une phase d'étude préliminaire et être utilisé pour aider à une première décision pour le développement d'un pilote de laboratoire ou semi-industriel. Se voulant générique, il peut tout à fait s'adapter à des phases plus industrielles avec certaines adaptations dues à des problèmes de mise à l'échelle. En effet, les données industrielles et les données issues d'articles scientifiques ne sont pas, en général, à des échelles identiques.

En synthèse, ce cadre méthodologique va permettre :

- (i) Aux non-experts de la durabilité de pouvoir interpréter les résultats d'une AdCV de manière simple et rapide sans toutefois perdre de l'information et ainsi de prendre des décisions pour concevoir des trajectoires technologiques industrielles plus durables.
- (ii) Aux non-experts de l'IA et des données massives d'utiliser des outils automatiques ou semi-automatiques pour extraire et analyser les données nécessaires à leur étude.
- (iii) De placer les données et les résultats, et par extension l'IA et les données massives, au centre

de la réflexion sur le DD dans les trajectoires technologiques industrielles afin de bénéficier du capital informationnel.

- (iv) D'être plus critique sur les données et résultats par l'intégration de méthodes dédiées.
- (v) D'avoir une visualisation et une analyse des données et résultats plus interactives et automatisées pour aider les décideurs.

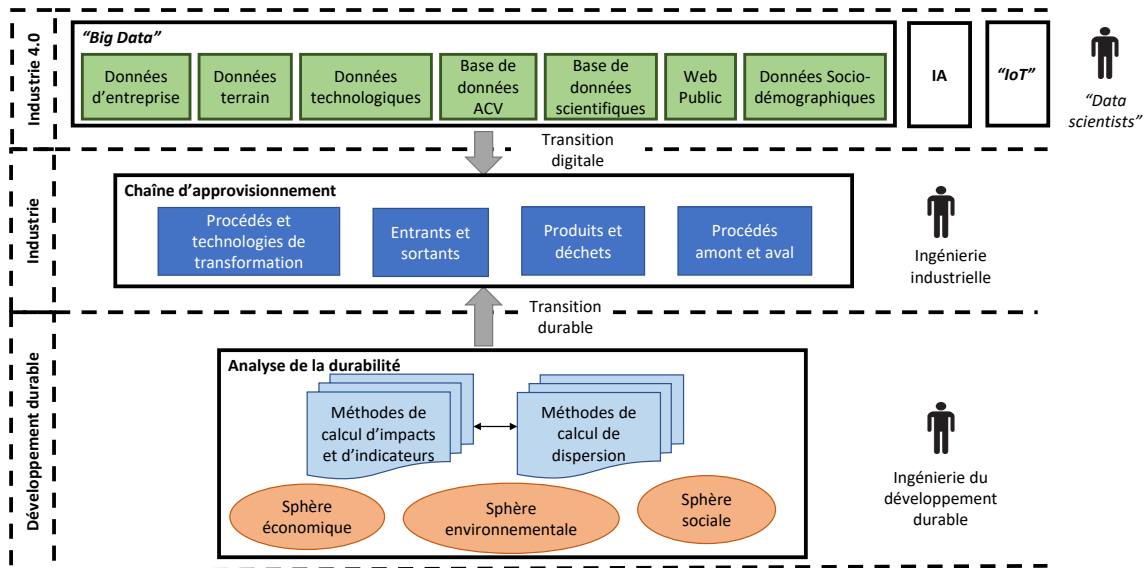


FIGURE 3.2 – Les trois domaines d'intérêt du cadre méthodologique.

Le cadre méthodologique se décompose en cinq étapes principales (figure 3.3) :

1. Le but et les frontières de l'étude
2. L'architecture des données
3. L'analyse de la durabilité
4. L'analyse et la visualisation des résultats
5. La décision

Ces étapes sont détaillées dans la suite de ce chapitre.

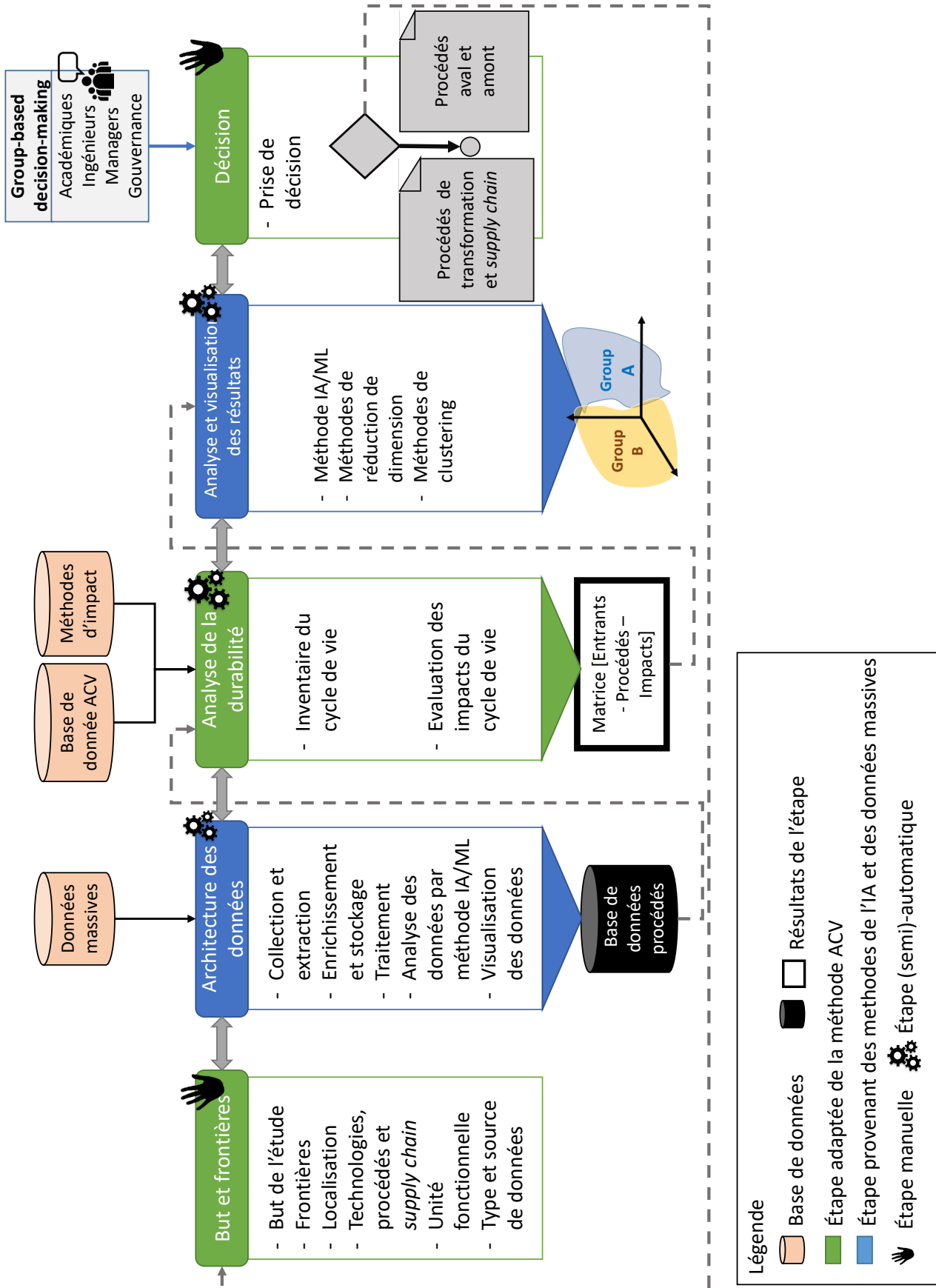


FIGURE 3.3 – Schéma général du cadre méthodologique.

3.2 But et frontières de l'étude

La première étape de notre approche est la définition, le but et les frontières de l'étude qui peut être divisée en plusieurs sous-étapes différentes. La première sous-étape consiste à définir précisément l'étude, son but et la personne ou le groupe de personnes pour lesquelles l'étude est réalisée, aussi appelé les décideurs. Dans notre approche, les décideurs sont les chercheurs ou les ingénieurs R&D qui utilisent l'approche. Cette approche est particulièrement adaptée à la recherche en raison des limites du système et de la nature des articles publiés qui sont à l'échelle laboratoire. Elle peut facilement être étendue à l'industrie [SHIBASAKI et al., 2007]. Le fait de détailler précisément le but de l'étude permet de savoir où on se situe. A-t-on affaire à une comparaison de voies de construction d'un produit dans des appareils préexistants ou à la construction de pilote de 0? Le but et la fonction de l'étude à travers le produit d'intérêt détaillent aussi l'étude : est-ce qu'on compare les procédés pour un produit sortant qui est identique ou équivalent? Sommes-nous en train de faire du *reengineering* de la conception ou de l'innovation? On répond alors aux questions : pourquoi cette étude? Quel est l'intérêt de cette étude? A-t-elle du sens?

La deuxième sous-étape est la définition des limites du système. Il est recommandé d'utiliser la PCV pour les définir. En effet, cette pensée et l'EC encouragent une approche « du berceau à la tombe » ou « du berceau à la porte » (voire « du berceau au berceau ») [KOPNINA, 2019]. Cependant, suivant l'étude, il peut être difficile d'intégrer des éléments en aval, ce qui conduit à préférer dans ce cas une approche "du berceau à la porte". Les scientifiques et les ingénieurs peuvent se retrouver dans l'impossibilité de prendre en compte le comportement de l'utilisateur final ou du consommateur dans leurs modèles. Ces limites doivent être définies avec précision, car elles ont des effets très importants sur l'analyse environnementale. Par exemple, il convient de préciser si la chaîne d'approvisionnement en amont est prise en compte. Il est aussi très important de bien préciser où se situe l'étude (le choix de la méthode d'impacts peut en dépendre). En plus, des limites du système, il convient aussi de décrire précisément le système : les procédés, les technologies, les chaînes logistiques.

La troisième sous-étape consiste à définir l'unité fonctionnelle. L'unité fonctionnelle dépend de l'objectif de l'étude et du type de comparaison de procédés que le chercheur souhaite. Le plus souvent, elle est soit sur le produit d'intérêt soit sur un des entrants limitants. Le choix de cette unité fonctionnelle est très important car il peut modifier l'étude de façon drastique si cette unité est mal définie [14040 :2006, 2006]. Il est possible de faire une étude en prenant plusieurs unités fonctionnelles et en comparant les résultats obtenus. Pour cela, il suffit d'itérer les étapes.

La dernière sous-étape porte sur le choix du type de sources de données. Le type de sources dépend de l'étude, elles peuvent être issues de bases de données privées ou du web public par exemple. Les types de données possibles sont décrits dans la section 3.3.1 et la figure 3.2.

Cette première étape contraint et guide la création d'ontologies de l'étape suivante. En effet, les définitions fournies structurent les connaissances et encadrent la sélection des articles scientifiques

et des données nécessaires à l'étude. Cette étape est réalisée manuellement par des chercheurs et des ingénieurs R&D (spécialiste de leur domaine industriel).

3.3 Architecture des données

Cette étape est issue de l'IA et des données massives. Les données à utiliser doivent d'abord être choisies. Cela nécessite de répondre à plusieurs questions telles que :

- Où puis-je trouver les données?
- Quel type de données (forme, volumétrie)?
- Quelles sont les incertitudes sur les données?
- Quelles données ai-je déjà en ma possession?
- De quel degré d'automatisation des données ai-je besoin?

La liste des différentes données dans le paragraphe suivant n'est pas exhaustive, et d'autres types de données peuvent être ajoutés selon l'objectif du projet mais elle fournit un tour d'horizon des données possibles. Une fois le choix fait il va être possible de dérouler cette étape qui comporte cinq sous-étapes. Notre cadre méthodologique étant tourné "pensée cycle de vie", deux types de données se distinguent : les données issues du système de premier plan (dites données de premier plan) et celles issues du système d'arrière-plan (dites données d'arrière-plan). Les données peuvent être obtenues directement, par des mesures sur site (données primaires), ou indirectement, à partir d'articles scientifiques publiés, de modèles et de bases de données (données secondaires). L'inventaire des flux du système de premier plan est généralement basé sur des données primaires, tandis que celui du système d'arrière-plan s'appuie sur des sources de données secondaires [GUINÉE, 2002].



Système de premier plan : ensemble des procédés et des flux directement affectés par l'étude.

Système d'arrière-plan : système qui fournit l'énergie et les matériaux au système de premier plan généralement par le biais d'un marché homogénéisé à une certaine échelle (région, pays, continent) [ELGHALI et al., 2007]

3.3.1 Données disponibles

Les différentes données utilisées et disponibles sont énumérées dans la figure 3.2. On va répartir ces données en trois grands groupes : les données structurées, les données non structurées facilement exploitables et les données non structurées difficilement exploitables.

Données structurées :

- Les **données d'entreprise** sont des données obtenues auprès des entreprises à n'importe quel stade de leur activité (de la création de l'entreprise jusqu'à sa fermeture).

- Les **données terrain** sont des données décrivant la qualité, ainsi que toutes les paramètres d'influence pour décrire les matières premières et les entrants d'intérêt.
- Les **données technologiques** concernent les technologies utilisées dans les procédés industriels étudiés : les types de technologie, les paramètres opératoires et les valeurs énergétiques.
- Les **bases de données ACV** sont des bases de données d'arrière-plan largement utilisées en ACV comme EcoInvent[®] [FRISCHKNECHT et al., 2005] et Gabi[®] [SPATARI et al., 2001]

Données non structurées facilement exploitables :

- Les **données du web public** sont accessibles à toute personne disposant d'un navigateur web et comprennent les données météorologiques, les prix mondiaux des matières premières et les facteurs d'impact des méthodes d'ACV (la méthode **ReCiPe** par exemple). Notons que ce groupe ne prend pas en compte les données issues des bases de données scientifiques (mises à part et contenues dans le point suivant).
- Les **données sociodémographiques**, servant aux analyses sociales, vont avoir deux origines ; d'une part, les données du web public qui correspondent aux données disponibles publiquement sur les populations (nombre d'habitants, PIB ...) et, d'autre part, les données issues de questionnaires effectués spécifiquement pour une analyse sociale. Celles-ci sont parfois intégrées dans des bases de données structurées d'entreprise qui en font alors des données qui rentrent dans la catégorie des données structurées.

Données non structurées difficilement exploitables :

- Les **bases de données scientifiques** sont des bases de données d'articles scientifiques comme *Web of Science*[®] et *Science Direct*[®]. Même si l'actualité tend vers le mouvement de la science ouverte pour partager les données de manière plus rigoureuse et plus ouverte (ouverte dans le sens de l'*Open Data*), les données des articles scientifiques sont principalement contenues dans les *Portable Document Format* (PDF) des articles scientifiques et détenues par les différents éditeurs scientifiques. Ces PDF contenant des tables, des graphiques (donc des images) et du texte sont difficilement exploitables par les machines.

3.3.2 Présentation générale des sous-étapes de l'étape d'Architecture des données

Nous décrivons ici succinctement les sous-étapes de l'Architecture des données. Le chapitre 4 va permettre de rentrer plus en détail sur celles-ci et apporter un descriptif détaillé des méthodes possibles afin d'en extraire les plus adaptées.

L'étape d'architecture des données comporte cinq sous-étapes :

1. **Collection et extraction des données** : c'est l'extraction et la collection des données provenant des différentes bases de données utilisées, qu'elles soient structurées (l'extraction est simple grâce à un langage de requêtes) ou non structurées. Pour ces dernières, l'extraction est plus complexe et nécessite des méthodes adaptées. L'IC intervient à travers les ontologies. Le chapitre 4 présente les ontologies et montre en quoi cela va aider l'extraction des données. Une méthode pour les utiliser sera aussi proposée.

2. **Enrichissement et stockage des données** : toutes les données extraites sont stockées soit dans des systèmes de gestion de bases de données relationnelles (SGBD) soit d'autres types des bases de données suivant les outils disponibles. L'enrichissement est le processus d'ajout de données dans les SGBD. Ces données ajoutées proviennent d'experts du domaine d'étude ou de modèles spécifiques. Les modèles sont des simulations numériques empiriques, de la thermodynamique ou de l'énergie.
3. **Traitement des données** : elle consiste à nettoyer, ajouter et supprimer des données pour la gestion. Ce traitement ne conserve que les données de valeur et ainsi obtenir un volume de données contrôlé. Elle est faite par l'expert domaine d'une manière manuelle ou semi-automatique. Après ce traitement, une nouvelle base de données, plus précise et plus accessible, est générée. Cette nouvelle base de données permet d'effectuer les prochaines sous-étapes sans toucher aux données extraites pré-traitement. Un gain de temps (grâce à une volumétrie moindre) et une diminution des erreurs dues à des données erronées sont la raison de cette base.
4. **Analyse des données brutes** : elle dépend de l'objectif, des domaines de données et des décideurs. Plusieurs types d'analyse sont possibles : l'analyse descriptive (que s'est-il passé?), l'analyse diagnostique (pourquoi cela s'est-il passé?), l'analyse prédictive (que se passera-t-il?) et l'analyse prescriptive (comment faire en sorte que cela se passe?). Les analyses descriptives et diagnostiques font appel à un certain nombre de méthodes et d'algorithmes différents, telles que le résumé, l'écart-type, les dépendances linéaires ou non linéaires, l'analyse factorielle et les méthodes de classification (induction d'arbres de décision, réseaux bayésiens, classificateurs k-voisin le plus proche). Certaines de ces méthodes sont visuelles, et la visualisation des données est donc incluse dans la sous-étape d'analyse des données
5. **Visualisation des données brutes** : elle est également réalisée en traçant les données brutes ou les données analysées sous forme de graphiques simples ou interactifs.

À la fin de cette étape, on se retrouve avec une base de données de premier plan qui va servir à l'analyse de la durabilité.

3.4 Analyse de la durabilité

Cette étape a pour but d'évaluer les impacts de la durabilité du système étudié à partir des données issues des étapes précédentes. Suivant la méthode choisie pour évaluer les impacts cette étape a quelques variations mais, en général, il y a deux grandes étapes : l'inventaire du cycle de vie (ICV) et l'évaluation des impacts du cycle de vie (EICV). Nous prenons deux hypothèses pour la suite de cette étape : une seule et même unité fonctionnelle et des limites du système qui sont les mêmes pour n'importe quel pôle du DD. Le chapitre 2 nous avait orienté vers une AdCV. Avec ces deux hypothèses l'AdCV se découpe en 3 ACV : l'AeCV, l'AcCV et l'AsCV. Cela se traduit par un inventaire pour chaque pôle ainsi que pour l'évaluation des impacts (figure 3.4).

le même format de données. Les grandes exigences concernant l'analyse des inventaires sont :

- tout ce qui a de l'importance doit être inclus
- toutes les données doivent être documentées (ainsi que leurs incertitudes)
- le traitement des données manquantes doit être documenté

Ecoinvent[®] est reconnu comme le premier fournisseur mondial de données d'inventaire du cycle de vie cohérentes et transparentes. La base de données contient plus de 3 500 processus industriels largement documentés. Sa dernière version 3.9 (publiée en 2022) comprend des mises à jour et de nouvelles données pour le secteur de l'électricité, les chaînes d'approvisionnement en gaz naturel, les produits chimiques et le recyclage des plastiques, les batteries, les déchets et l'agriculture. Il est possible aussi de trouver dans les bases de données ACV des données économiques et sociales même si cela reste assez rare. Il existe trois grands modèles dans Ecoinvent[®] :

Le modèle de système d'allocation par défaut (Alloc, Def) Il est également appelé "allocation au point de substitution". Il est basé sur deux décisions méthodologiques :

- Il utilise l'offre moyenne (offre non contrainte) des produits, telle que décrite dans les ensembles de données sur l'activité du marché.
- Il utilise le partitionnement (allocation) pour convertir les ensembles de données multiproduits en ensembles de données monoproduits.

L'offre moyenne signifie que le modèle n'utilise pas les contraintes dues aux marchés et à la technologie. Dans un tel modèle, le produit, le matériau ou le service est toujours fourni, même s'il est rare. Cela diffère du modèle consécutif, qui utilise une offre limitée de produits. Le modèle alloue les produits et sous-produits de référence de manière économique, c'est-à-dire en fonction de la valeur marchande des produits [SIMAPRO, 2022].

Le modèle du système d'allocation, contenu recyclé (Alloc, Rec) Il est aussi appelé *Allocation, cut-off by classification* par le Centre EcoInvent depuis la version 8.5 de SimaPro. Il utilise les deux mêmes décisions concernant l'offre moyenne et l'allocation que le modèle du système **Alloc, Def**. Cependant, **Alloc, Rec** ne prend pas en compte les avantages liés au recyclage d'un matériau. Dans ce modèle, les matériaux recyclables sont disponibles sans charge pour les processus de recyclage, ce qui signifie que les matériaux secondaires (recyclés) ne supportent que les impacts des processus de recyclage. Le modèle n'accorde aucun crédit aux producteurs de déchets pour le recyclage ou la réutilisation de produits issus d'un quelconque traitement des déchets. Contrairement au modèle du système **Alloc, Def** qui attribue les déchets (et les sous-produits) au point de substitution, le bénéfice du recyclage des matériaux est attribué aux processus de marché qui fournissent les matériaux secondaires ou les sous-produits (tels que la chaleur ou l'électricité provenant de l'incinération des déchets).

Le modèle du système consécutif Il traite différemment les deux décisions méthodologiques utilisées dans les deux modèles de systèmes d'allocation :

- Il utilise une offre contrainte de produits, basée sur les données d'activité du marché et sur les informations relatives au niveau technologique.
- Il utilise la substitution (expansion du système) pour convertir les ensembles de données multiproduits en ensembles de données monoproduits.

Dans un marché contraint, un changement de la demande n'entraîne pas un changement correspondant de l'offre, mais plutôt un changement de la consommation ailleurs. Par exemple, les marchés de sous-produits sont contraints parce que leurs volumes de production dépendent des volumes de production des produits de référence, et c'est la demande du produit de référence qui entraîne la production. Ce modèle de système vise à refléter les conséquences des décisions à petite échelle et à long terme en tenant compte des contraintes qui s'appliquent à cette échelle et à cet horizon temporel. En raison de l'offre limitée de produits, l'allocation des sous-produits est évitée en utilisant la substitution, et analyse un produit qui remplace ou substitue le sous-produit du produit de référence. Les émissions de ce produit de remplacement sont ensuite soustraites des émissions du produit de référence [SIMAPRO, 2022].

Chaque modèle a des avantages spécifiques, le choix du modèle dépend de l'objectif et du contexte de l'étude à réaliser.

3.4.2 Évaluation des impacts du cycle de vie

Cette étape concerne l'évaluation des impacts économiques, sociaux et environnementaux. Elle vise à transformer les résultats de l'inventaire en indicateurs (également appelés impacts). Cette transformation se fait grâce à des méthodes d'impacts spécifiques. Pour l'AcCV, les flux sont directement des flux monétaires. Il suffit alors de les agréger suivant les catégories comme par exemple maintenance des machines, coût des ressources humaines, productivité, rendement brut. C'est l'évaluation la plus simple. Pour l'AsCV, la littérature est encore jeune et souvent très similaire à l'AeCV. Il faut alors utiliser les mêmes principes que ceux que l'on va décrire ici.

3.4.2.1 Évaluation des impacts environnementaux du cycle de vie

L'Évaluation des impacts du cycle de vie environnementale (EICVe) traduit l'inventaire du cycle de vie en impacts environnementaux c'est-à-dire la quantification des échanges entre le système de produits et l'environnement. Selon le champ d'étude, l'information obtenue sera plus ou moins importante (des centaines de flux de ressources naturelles et d'émissions dans l'environnement sont quantifiés) et son utilisation pratique s'avère difficile.

L'ensemble des indicateurs proposés par la méthodologie ACV est réparti en trois niveaux d'agrégation, au fur et à mesure desquels on avance dans la chaîne de causalité. La figure 3.5 nous donne ces trois niveaux et des exemples d'impacts correspondants dans chaque niveau.

Le premier niveau concerne les effets physiques, chimiques et biologiques quantifiables des flux de matières et d'énergie entre le système étudié et son environnement. Les indicateurs correspondants sont des impacts orientés problèmes ou impacts *midpoint*. Les résultats de l'ICV sont initialement classés par catégorie d'impacts *midpoint*. Le calcul des scores d'impacts est alors obtenu en multipliant les résultats de l'ICV par des facteurs d'impact. Les catégories d'impacts ainsi que les facteurs d'impacts sont issus de diverses méthodes d'évaluation, telles que **ReCiPe 2016 Midpoint** et **ILCD 2011 Midpoint**.

Le deuxième niveau concerne les catégories orientées dommages (aussi appelé dommages ou impacts *endpoint*) causées par les effets des impacts *midpoint* sur plusieurs sphères de protection. De nombreuses méthodes d'évaluation, dont **ReCiPe 2016 Midpoint** et **ILCD 2011 Midpoint**,

prennent en compte trois catégories de dommages : la santé humaine, la qualité des écosystèmes et l'épuisement des ressources. Les facteurs de dommages sont issus des différentes méthodes d'évaluation. Cependant, certaines méthodes, comme la méthode **ReCiPe EndPoint**, ne proposent pas d'évaluation *midpoint* et utilisent directement les résultats de l'ICV pour calculer le score de dommages.

Un score unique (troisième niveau d'agrégation) est obtenu en normalisant les scores de dommages avec des facteurs de normalisation. Ces facteurs de normalisation ou de pondération sont fournis par les différentes méthodes.

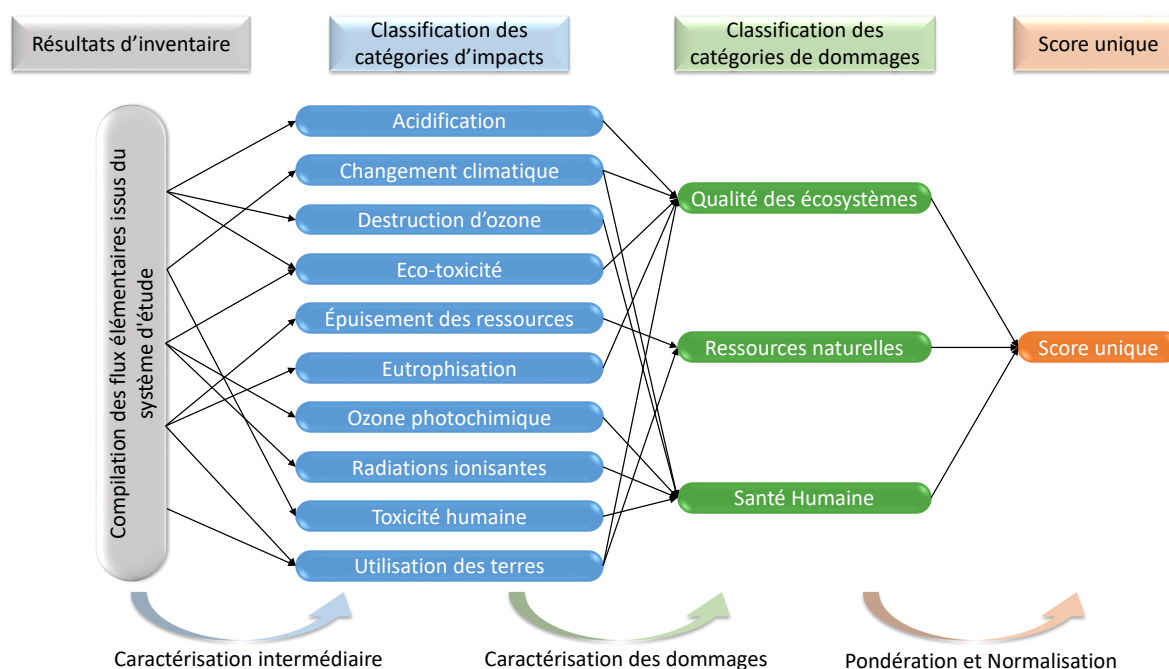


FIGURE 3.5 – Les trois niveaux d'agrégation des impacts environnementaux.

L'EICVe est décomposée en quatre étapes.

Classification C'est une étape qualitative au cours de laquelle une liste pertinente de catégorie d'impacts relatifs à des problèmes environnementaux. Les émissions et extractions obtenues lors de l'inventaire sont assignées à ces catégories, chacune pouvant contribuer à plusieurs catégories. Par exemple, les composés organiques volatils (COV) contribuent à la fois au réchauffement planétaire et à l'appauvrissement de la couche d'ozone. Selon les méthodes de caractérisation, la classification des impacts est donc différente et comporte plus ou moins de catégories d'impacts. Il est possible de procéder à une classification des dommages (*endpoint*) sur différents sujets de protection [JOLLIET et al., 2010].

Caractérisation C'est une étape quantitative consistant à calculer les impacts environnementaux. Un facteur de caractérisation représente la contribution par quantité d'un flux élémentaire à une catégorie d'impact. Concrètement, les masses des substances extraites et émises répertoriées dans l'inventaire des extractions et émissions sont multipliées par ces facteurs et sommées dans chacune des catégories permettant d'obtenir le score de la catégorie d'impact. Ce score

est imprimé dans une unité qui est propre à la catégorie. Par exemple, le CO₂ est un gaz de référence pour déterminer le potentiel de réchauffement planétaire d'autres gaz connexes, comme le CH₄, le NO₂ et d'autres COV [JOLLIET et al., 2010]. La modélisation de ces facteurs implique l'utilisation de différents modèles et paramètres et est généralement réalisée par des experts pour une catégorie d'impact particulière [HAUSCHILD et al., 2018]. La caractérisation des dommages ou du score final agrège les catégories d'impact en catégories de dommages ou en score final. Le score final doit être évité, car il y a beaucoup de pertes d'information.

Normalisation Elle est facultative dans la norme ISO 14044 [14044 :2006, 2006]. Elle présente la contribution du produit étudié comme une fraction de l'impact global dans une catégorie d'impact donnée [JOLLIET et al., 2010]. C'est-à-dire que les scores des indicateurs pour toutes les catégories d'impact sont exprimés dans une métrique commune, généralement les contributions annuelles aux impacts environnementaux totaux d'une personne moyenne. Cela sert principalement trois objectifs : 1) permettre aux décideurs de mieux comprendre l'ampleur des résultats caractérisés en les rapportant à une référence commune familière et externe, 2) vérifier l'absence d'erreurs dans l'évaluation entraînant des résultats normalisés déraisonnablement bas ou élevés et 3) ouvrir la voie à la pondération [HAUSCHILD et al., 2018]. Il y a aussi ici une perte d'information ce qui nous laisse entendre que dans notre cadre méthodologique, il est préférable de ne prendre en compte que les impacts *midpoint*.

Pondération C'est l'étape la plus subjective de l'évaluation des impacts. Elle met en lumière l'importance relative des différents impacts qui sont pondérés de manière à pouvoir les comparer entre eux. En conséquence, les différents impacts environnementaux sont réduits à une seule fonction d'impact environnemental, ou score unique, comme mesure de la performance environnementale. Plusieurs techniques ont été proposées pour être utilisées dans l'évaluation. Elles sont principalement basées sur l'expression des préférences des décideurs, des "experts" ou du public.

3.4.2.2 Les logiciels en ACV

Les résultats de l'ACV peuvent être calculés sans logiciel spécifique à l'ACV (par exemple Excel[®] ou MatLab[®]) et cela est plus éducatif et rigoureux. Cependant, le calcul de l'ACV devient généralement assez élaboré, et l'utilisation d'un logiciel spécialisé est donc recommandée. Le logiciel d'ACV le plus répandu est SimaPro[®]. Il a été conçu pour être une source d'informations scientifiques, offrant une transparence totale et évitant les processus de type boîte noire [SIMAPRO, 2022]. D'autres outils commerciaux pour les logiciels d'ACV existent comme GaBi[®] [SPATARI et al., 2001]) et Umberto[®]. Il existe également des outils gratuits tels que *OpenLCA* et *CMLCA*. Le logiciel *OpenLCA* est produit par *Green-DeltaTC GmbH*, soutenu par plusieurs organisations leaders en matière de durabilité. *CMLCA* est un outil du CML (*Centrum voor Milieuwetenschappen* : Institut des sciences de l'environnement) de l'Université de Leiden qui prend en charge le calcul de l'ACV, y compris l'analyse sociale du cycle de vie (AsCV) et l'analyse de la durabilité du cycle de vie (AdCV) [CMLCA.EU, 2022]. Cependant, l'importation d'EcoInvent[®] ou d'autres bases de données d'ICV sera nécessaire.

3.4.2.3 Qualité des données et analyse de la sensibilité

Différents types d'incertitudes apparaissent dans une étude ACV, comme les incertitudes dans les paramètres d'entrée, les incertitudes dans le modèle lui-même, les incertitudes dans les hypothèses et les choix, et les incertitudes dans la variabilité des données (spatiale, temporelle, technologique). Jusqu'à présent, peu d'études ACV ont estimé quantitativement la qualité de leurs données. Pourtant, elle doit être sérieusement prise en compte pour assurer la valeur et la crédibilité de l'ACV. En effet, si l'incertitude totale de l'ACV est plus grande que la différence entre les impacts finaux des différents scénarios, cela pourrait conduire à un jugement erroné (lors de la comparaison des scénarios). En raison des nombreuses interprétations effectuées dans une ACV, l'analyse de l'incertitude et de la sensibilité concernant divers paramètres est essentielle pour comprendre la robustesse des résultats de l'étude. Selon [JOLLIET et al., 2010], la méthode appropriée pour évaluer les incertitudes des paramètres est l'analyse Monte Carlo, lorsqu'elle est appliquée correctement, elle détermine également l'importance d'une différence entre deux scénarios.

L'analyse de Monte Carlo identifie d'abord chaque paramètre d'entrée du modèle, ainsi que le type de distribution de probabilité (par exemple, normale, log-normale ou uniforme), en utilisant des indicateurs de qualité des données. La sortie du modèle est ensuite calculée en sélectionnant aléatoirement une valeur pour chaque paramètre d'entrée en fonction de sa distribution de probabilité. Dans le contexte spécifique de l'ACV, la distribution log-normale est souvent appliquée par défaut, car les valeurs des paramètres varient parfois sur plusieurs ordres de grandeur et ont l'avantage d'exclure automatiquement plusieurs scénarios impossibles, tels que des émissions négatives [JOLLIET et al., 2010].

Dans EcoInvent[®], l'ICV du produit est décrit par des chiffres uniques par flux d'entrée ou de sortie. Ces chiffres contiennent un niveau d'incertitude, par exemple en raison d'approximations temporelles ou spatiales. Souvent, l'étendue de l'incertitude n'est pas déduite directement des informations disponibles. Pour ces cas, une procédure standard simplifiée a été développée pour dériver les facteurs d'incertitude à partir d'une évaluation qualitative des données [SIMAPRO, 2022]. Cette procédure est basée sur une matrice de Pedigree composée d'indicateurs de qualité des données. En ce qui concerne les indicateurs de qualité des données, plusieurs ont été développés [WEIDEMA et WESNÆS, 1996]. Chaque donnée collectée est décrite selon des critères de caractérisation. Ces derniers sont par la suite évalués selon une matrice qualitative où, à chaque situation observée, est attribuée une note comprise entre 1 et 5 : 1 correspondant à la situation la plus favorable et 5 à la situation la plus défavorable ou bien à l'incapacité à noter la donnée, ses métadonnées étant indisponibles. Un exemple de cette matrice Pedigree est présenté par le tableau 3.2. Les résultats de la cotation sont également présentés sous forme matricielle et caractérisent la qualité de l'information [FRISCHKNECHT et al., 2005; WEIDEMA et WESNÆS, 1996]. Les auteurs ont défini une grille présentant l'indice de dispersion (facteur d'incertitude par défaut) totale pour chaque critère évalué par la matrice de pedigree et pour chaque score correspondant (Tableau 3.1).

Compte tenu de ces propriétés, il est supposé que pour chaque degré de qualité correspond un intervalle dans lequel se trouve la valeur vraie, cet intervalle étant une fonction de l'écart-type. Ainsi, plus une donnée est de mauvaise qualité et plus l'intervalle comprenant cette donnée sera grand.

TABLEAU 3.1 – Indices de dispersion selon les différents scores de la matrice de Pedigree (selon la base EcoInvent®)

Score	1	2	3	4	5
Fiabilité	1,00	1,05	1,10	1,20	1,50
Complétude	1,00	1,02	1,05	1,10	1,20
Corrélation temporelle	1,00	1,03	1,10	1,20	1,50
Corrélation géographique	1,00	1,01	1,02	-	1,10
Corrélation technologique	1,00	-	1,20	1,50	2,00
Taille de l'échantillon	1,00	1,02	1,05	1,10	1,20

L'intervalle est élargi de sorte à couvrir la donnée de faible qualité et donc incertaine. Cette approche pallie les difficultés d'obtention ou de détermination des distributions statistiques des grandeurs utilisées pour la modélisation. Elle fournit une alternative d'alimentation du système d'information. Une fois l'ensemble des distributions définies, les incertitudes sont propagées à l'aide de simulations de Monte Carlo.

3.4.3 Conclusion sur l'étape d'analyse de la durabilité

À la fin de cette étape, on se retrouve avec une matrice [*Intrants – Procédés – Impacts*]. Les impacts *midpoint* sont au nombre de 13 à 18 (suivant les méthodes utilisées) et les procédés dépendent des technologies sélectionnées. Pour les intrants, ils vont de pair avec le but de l'étude. Si celle-ci est la comparaison des intrants, leur nombre peut être élevé. Si on utilise les articles scientifiques comme base de données, il n'est pas impossible d'identifier de nombreux procédés et de nombreux intrants. La matrice est de grande dimension et difficile à interpréter.

TABLEAU 3.2 – Matrice de Pedigree d’après la base EcoInvent® [FRISCHKNECHT et al., 2005]

Indicateur	1	2	3	4	5	Remarques
Fiabilité	Données vérifiées basées sur des mesures	Données vérifiées en partie basées sur des hypothèses ou données non-vérifiées obtenues par mesures	Données non-vérifiées basées en partie sur des estimations	Estimations (par experts), données dérivées d'information théoriques	Estimations non-qualifiées	Vérifié sous-entend : publié dans des rapports publiques, statistiques officielles... Non vérifié sous-entend : information personnelle obtenu par lettre, fax ou mail
Complétude	Données représentatives de tous les sites pertinents pour le marché considéré sur une période adéquate pour uniformiser les fluctuations normales	Données représentatives issues de plus de 50% des sites pertinents pour le marché considéré sur une période adéquate pour uniformiser les fluctuations normales	Données représentatives de quelques sites (<50%) pertinents pour le marché ou >50% des sites mais sur une période plus courte	Données représentatives seulement d'un site pertinent pour le marché considéré ou sur une période plus courte	Représentativité inconnue ou données issues d'un petit nombre de sites et sur une période plus petite	La période d'investigation adéquate dépend du procédé/technologie
Corrélation temporelle	Moins de 3 ans de différence avec l'année de référence	Moins de 6 ans de différence avec l'année de référence	Moins de 10 ans de différence avec l'année de référence	Moins de 15 ans de différence avec l'année de référence	Plus de 15 ans de différence avec l'année de référence ou l'âge de la donnée inconnu	Score pour les procédés dont le cycle d'investissement est inférieur à 10 ans pour les autres un ajustement peut être apporté La similarité est exprimée en terme législatif : Amérique du Nord, Australie, Union Européenne, Chine, ...
Corrélation géographique	Données issues de la zone d'étude	Données moyennes sur une zone plus large comprenant la zone d'étude	Données issues d'une zone plus petite que la zone d'étude ou provenant d'une zone similaire	Données représentatives de procédés ou matériaux approchant mais de technologie différente ou données représentatives du stade de laboratoire ou de technologie	Données issues d'une zone inconnue ou significativement différente	
Corrélation technologique	Données de l'entreprise, procédé ou matériau étudié (technologie identique)		Données représentatives de procédés approchant mais de même technologie ou données de procédés ou matériaux à l'étude mais de technologie différente		Données représentatives de procédés ou matériaux approchant mais au stade laboratoire et de technologie différente	Exemple de technologies différentes : turbine à vapeur au lieu d'un moteur à propulsion dans un navire
Taille de l'échantillon	> 100, mesures continues, bilan des produits achetés	> 20	> 10, données agrégées dans un rapport	≥ 3	Inconnue	La taille de l'échantillon représentative d'une figure dans une source d'information

3.5 Analyse et visualisation des résultats

L'étape consiste à inclure des méthodes dérivées de l'intelligence artificielle, et plus précisément de l'apprentissage automatique, pour faciliter l'analyse des impacts environnementaux. Certaines études ont examiné les modèles spatio-temporels de l'AeCV en se basant sur une évaluation pluri-annuelle sur une grande région géographique, et en améliorant la résolution spatiale de l'EICV [LEE et al., 2020]. Dans le même contexte, [ROMEIKO et al., 2020] ont utilisé des modèles de régression arborescents renforcés pour identifier les facteurs les plus influents parmi le sol, le climat et les pratiques agricoles, qui déterminent l'hétérogénéité spatiale et temporelle des impacts environnementaux du cycle de vie. Les résultats de cette étude ont montré que la teneur en matières organiques du sol et le taux d'application d'azote étaient les principaux facteurs d'influence pour l'eutrophisation et l'acidification du cycle de vie. La texture du sol, les taux d'application d'azote et la température du mois de mars sont les facteurs qui influencent le plus les effets du réchauffement climatique sur le cycle de vie. En utilisant un modèle d'arbre de régression boosté avec une distribution gaussienne, la même équipe de chercheurs a mené une autre étude basée sur la prédiction des impacts environnementaux du cycle de vie de la production de maïs en fonction de la variabilité et des scénarios climatiques.

À notre connaissance, de nombreuses méthodes d'apprentissage automatique de l'EICV et de l'évaluation de la durabilité en général sont "supervisées", ce qui signifie que les algorithmes d'apprentissage sont réalisés à partir de données d'entrée-sortie [ABDELLA et al., 2020]. Cependant, dans de nombreux cas, il est nécessaire d'extraire des structures de données latentes basées sur les observations et la nature mixte de la protection de la zone d'extrémité. Pour cette raison, nos objectifs de recherche reposent sur l'utilisation de techniques de réduction de dimension pour l'évaluation de la durabilité à travers un ensemble d'impacts (changement climatique, santé humaine, appauvrissement de l'ozone, toxicité humaine ...) et de processus sous-jacents. Plusieurs stratégies ont été explorées pour atteindre cet objectif en utilisant l'apprentissage "non supervisé" qui comprend la réduction de dimension et le regroupement de données.

En partant de la littérature statistique, les techniques traditionnelles de réduction de dimension (DR) et de regroupement non supervisé sont combinées pour extraire des informations des impacts environnementaux. Plus précisément, cette approche hybride est basée sur le positionnement multidimensionnel (MDS) utilisant la distance de Canberra et les k-means. L'objectif est de rechercher des structures "cachées" dans des données multidimensionnelles et de faciliter l'interprétation de la matrice. L'avantage de cette approche est que les méthodes basées sur les données nécessitent très peu de connaissances sur les procédés et les entrants pour être exécutées. Cette étape est décrite plus précisément dans le chapitre 5.

Puis vient la visualisation des résultats qui est faite par des graphiques simples ou interactifs après l'analyse des résultats. On remarquera que la plupart des méthodes de *clustering* sont visuelles et donc intègrent directement la visualisation des résultats.

3.6 Décision

Les décisions sont prises manuellement ou avec l'aide d'un outil d'aide à la décision utilisant des modèles multicritères, tel que des modèles de tri (ELECTRE, PROMETHEE) ou des outils d'agrégation [BOUYSSOU, 2006] Des méthodes de prise de décision en groupe mixte, telle que Delphi-SWOT, peuvent aussi être mises en place [TAVANA et al., 2012]. Cette étape n'est pas expérimentée par le cas d'étude (chapitre 6).

3.7 Conclusion du chapitre 3

Dans ce chapitre, le cadre méthodologique général a été proposé et spécifié. Ce cadre répond à l'objectif principal de cette thèse : proposer une approche centrée données massives et intelligence artificielle pour l'évaluation des systèmes industriels durables et circulaires à travers un cadre méthodologique de cinq étapes. La première étape spécifie les frontières et le but de l'étude ainsi que les différentes sources de données. Ce dernier point dû à l'ajout de méthodes d'IA dans les étapes suivantes constitue un point central à notre cadre méthodologique. Les données sont au coeur de celui-ci. Dès le début de l'étude le type et la volumétrie des données ainsi que la raison de l'utilisation de ces données sont prises en compte. Les données qu'il est possible d'utiliser dans ce cadre ont été spécifiées pour comprendre leur hétérogénéité et donc leur difficulté de traitement.

La seconde étape, Architecture des données, constitue les bases de données nécessaires à l'étude puis les traite et les analyse. Cette étape apporte la première contribution pour répondre à la question : **Q1 - Comment exploiter et tirer bénéfice des données massives pour aider les analyses et les évaluations du développement durable?** Elle intègre de l'IC (à travers les ontologies) pour extraire les données de la littérature scientifique et de l'IA pour l'analyse de ces données. Constituée de cinq sous-étapes, elle est détaillée dans le chapitre 4.

La troisième étape, Analyse de la durabilité, spécifie les impacts de la durabilité. Après avoir fait le choix de l'AdCV, nous avons décrit les deux sous-étapes principales que sont l'ICV et l'EICV.

La quatrième étape intègre l'apprentissage profond pour analyser les résultats des analyses de la durabilité. Cette étape apporte la seconde contribution pour répondre à la question : **Q2 - Comment bénéficier des méthodes et outils issus de l'intelligence artificielle pour aider les analyses et les évaluations de la durabilité?**

Le chapitre 5 précise les méthodes disponibles et décrit la méthode utilisée pour notre cas d'étude.

Enfin la cinquième étape est l'étape de décision.

Le cadre méthodologique général (figure 3.6) est conçu pour un chercheur ou un ingénieur R&D. Les données utilisées provenant de la littérature scientifique nous donnent un *Technology Readiness Level* (TRL) 1/2. Il est possible d'utiliser des données industrielles disponibles pour les industries, mais ces données sont sensibles et les industries ne les rendent pas publiques. Il est possible cependant d'utiliser des méthodes de changement d'échelle pour augmenter le TRL du cadre proposé. De plus, comme les procédés ne sont pas continus dans les laboratoires, les données doivent être traitées de manière à ce que les entrées d'une étape des procédés correspondent aux sorties de l'étape précédente. Le traitement des données est amélioré, par exemple en ajoutant les changements d'échelle

effectués dans l'ingénierie des procédés. Une autre limite concerne l'abondance et la qualité des données issues des articles scientifiques. En effet, celles-ci sont faibles pour certains procédés innovants.

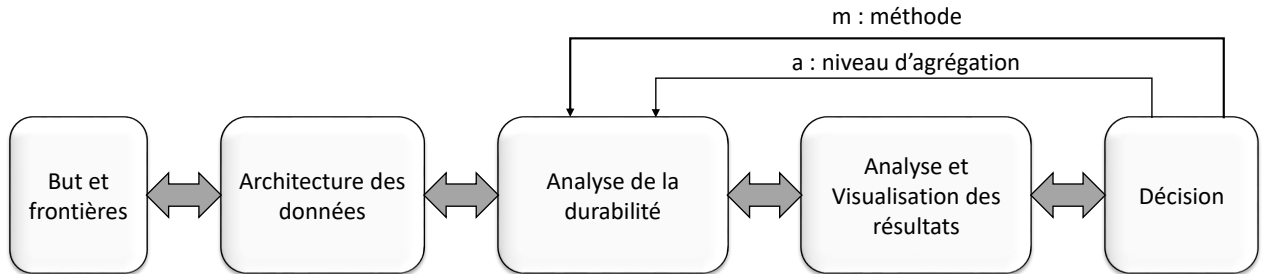


FIGURE 3.6 – Étapes du cadre méthodologique et chapitre du manuscrit associé.

Chapitre 4

Architecture des données

A basic principle of data processing teaches the folly of trying to maintain independent files in synchronism.

F. P. Brooks Jr. - *The Mythical Man-Month*, 1995

Ce chapitre (4.1) précise la deuxième étape nommée "Architecture des données" en détaillant ses quatre sous-étapes. Cette étape a deux grands objectifs, extraire les données pertinentes qui vont servir à l'étape suivante d'analyse de la durabilité, et traiter et analyser ces données. Le chapitre est organisé en cinq sections. Après une introduction, la section 4.2 revient sur la définition des concepts utilisés dans la sous-étape d'extraction et de collection des données provenant de bases de données structurées ou non. Cette section apporte des méthodes et outils de l'ingénierie des connaissances pour extraire de manière semi-automatisée les informations pertinentes contenues dans des bases de données non structurées. Les sections 4.3 et 4.4 présentent respectivement la sous-étape d'enrichissement et de stockage des données extraites dans la sous-étape précédente et la sous-étape de traitement de ces données. La section 4.5 se penche sur la description de l'étape d'analyse des données qui va intégrer des outils venant du "Machine Learning" notamment les réseaux de neurones. Enfin, la section 4.6 discute de la sous-étape de visualisation des données.

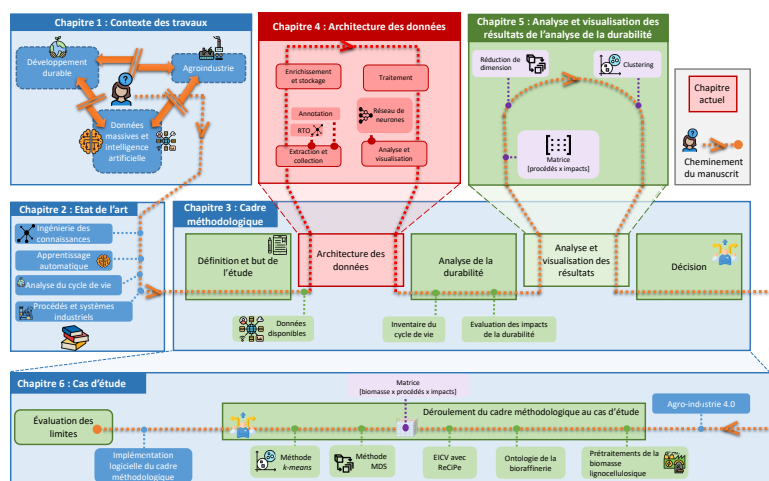


FIGURE 4.1 – Positionnement du chapitre 4 dans le manuscrit.

Sommaire

4.1 Introduction du chapitre 4	77
4.2 Collection et extraction des données	78
4.2.1 Définitions des différents concepts	79
4.2.1.1 Ontologies	79
4.2.1.2 Ressource ontologique et terminologique (RTO)	81
4.2.1.3 Méthodes d'extraction automatique des données issues d'articles scientifiques	82
4.2.2 Description de l'étape de collection et d'extraction des données	83
4.2.2.1 Sélection des documents	84
4.2.2.2 Création d'une ressource ontologique et terminologique	84
4.2.2.3 Extraction des tableaux	85
4.2.2.4 Annotation	85
4.2.2.5 Stockage dans un <i>Resource Description Framework</i>	86
4.2.2.6 Score de fiabilité à partir des métadonnées	87
4.3 Enrichissement et stockage des données extraites	89
4.4 Traitement des données extraites	90
4.5 Analyse des données	90
4.5.1 Réseaux de neurones	91
4.5.2 Interprétabilité des réseaux de neurones	94
4.5.3 Méthodes PDP et ICE	95
4.5.4 Conclusion sur l'apport des réseaux de neurones	98
4.6 Visualisation des données	98
4.7 Conclusion du chapitre 4	98

4.1 Introduction du chapitre 4

La deuxième étape du cadre méthodologique, Architecture des données, constitue un cadre méthodologique en soi. Pour le différencier du cadre méthodologique général, nous l'appellerons par son synonyme anglais *framework*. Ce *framework* est directement inspiré du cycle de vie d'une donnée présentée dans la figure 4.2, de la Business Intelligence [CHEN et al., 2012] et de la construction de l'architecture des données massives (à laquelle elle doit d'ailleurs son nom). Il est composé de cinq sous-étapes :

1. Collection et extraction des données
2. Enrichissement et stockage des données extraites
3. Traitement des données extraites
4. Analyses des données
5. Visualisation des données

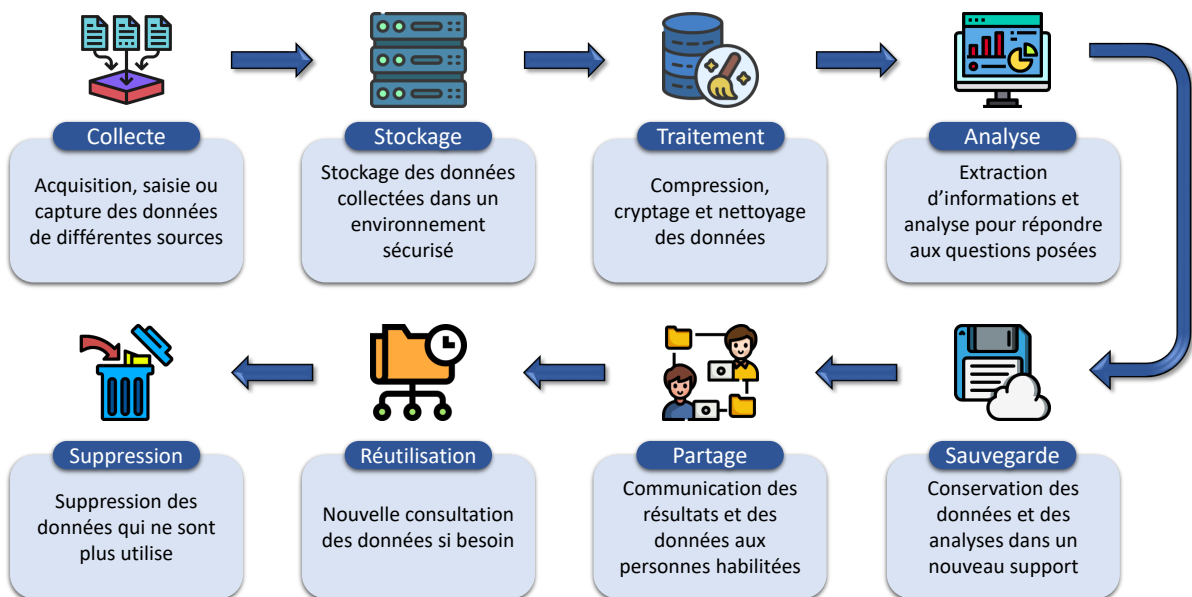


FIGURE 4.2 – Cycle de vie de la donnée.

Dans les sections suivantes, nous explicitons les méthodes et outils retenus pour compléter l'étape 2 du cadre méthodologique.

4.2 Collection et extraction des données

Les données à utiliser doivent d'abord être choisies. Cela nécessite de se poser plusieurs questions : où pouvons-nous trouver les données ? Quel type de données ? Quelles sont les incertitudes sur les données ? Quelles données avons-nous déjà en notre possession ? De quel degré d'automatisation des données avons-nous besoin ? La collecte et l'extraction de données à partir de bases de données structurées nécessitent des outils spécifiques, tels que les requêtes de données (requêtes SQL) ou le traitement analytique en ligne (OLAP). Cette sous-étape est plus compliquée pour les données non structurées provenant du web. Les articles scientifiques qui font partie des sources importantes pour nos travaux constituent des données non structurées. Les métadonnées associées aux pages web peuvent être utilisées pour leur classification et pour donner accès à leur contenu. Par exemple, la définition du format MARC au début des années 1960 a normalisé les métadonnées des ressources documentaires. Grâce au développement du web des données et notamment du *Ressource Description Framework* (RDF), une norme du W3C (*Consortium World Wide Web*), il est possible d'utiliser SPARQL (*SPARQL Protocol and RDF Query Language*) pour requêter le RDF. Le processus d'extraction génère un tableau structuré et diffère selon le format des pages web : *Application Programming Interface* (API), *HyperText Markup Language* (HTML) ou PDF. L'extraction peut être automatique, semi-automatique ou manuelle. Les données peuvent être extraites automatiquement des pages web avec un format API. Cependant, l'extraction de données pertinentes à partir d'articles scientifiques nécessite l'aide de méthode issue de l'IC pour ne pas perdre d'informations lors de l'extraction et de n'extraire que les données d'intérêt.

Cette sous-étape peut se découper en deux parties suivant le type de données qu'il y a à extraire. Si on reprend la typographie des données que l'on a détaillée dans le chapitre 3 avec les trois types de données qui sont mises en avant :

- **Les données structurées** : données d'entreprise, données terrain, données technologiques et base de données ACV. Ces données sont facilement extraites et collectées de façon automatique par des requêtes *Structured Query Language* (SQL) ou autres suivant le type de base de données (relationnelle ou non par exemple) pour les intégrer à notre base de données.
- **Les données non structurées facilement exploitables** : comme les données du web public et sociodémographique. L'extraction et la collection de ces données peuvent être extraites de façon automatique. Des requêtes d'extraction automatique (via API ou HTML par exemple) vont permettre de générer un tableau structuré à partir des différentes données non structurées. Ici, chaque type de données aura une requête spécifique et proviendra d'une page web spécifique. Les métadonnées de cette page web permettront d'avoir de l'information sur les données extraites. Si on prend par exemple les prix d'une matière première, les métadonnées donneront le nom de la matière première, la date de l'extraction et la source des données (nom du site) ainsi le prix sera associé à la matière première à un temps T suivant un fournisseur particulier (ou le site où le prix du marché a été extrait pour les matières qui sont soumises à des marchés).
- **Les données non structurées difficilement exploitables** : ce sont les données contenues dans les articles scientifiques, par exemple. Ces données extraites des PDF des articles scientifiques proviennent de tables, figures et textes qui sont difficilement exploitables pour des machines

pour produire des informations à valeur ajoutée.

Les deux premiers types de données sont aisément exploitables. Une méthode est créée pour extraire les données du dernier type c'est-à-dire les données contenues dans les articles scientifiques. Cette extraction est semi-automatique : elle utilise des méthodes de l'IC pour automatiser certaines parties, mais l'expertise d'un ingénieur du domaine est nécessaire ainsi qu'une personne pour annoter les données grâce à une grille de lecture. Le caractère semi-automatique de cette extraction avec une validation manuelle à chaque étape est indispensable pour garantir un enrichissement de la base avec des données pertinentes et qualitatives.

4.2.1 Définitions des différents concepts

4.2.1.1 Ontologies

Le concept d'ontologie désigne initialement, de manière très large, toute transformation simplifiée du monde en un groupe de concepts accompagnés des relations qui donnent de la substance à ces derniers. Il s'agit d'un concept relativement vaste, car sa définition n'est pas normée : il est ainsi possible de trouver différentes définitions suivant le domaine étudié. La plupart des auteurs donnent la paternité du terme et de sa formalisation à [GRUBER, 1993] à travers cette définition de l'ontologie :



« Une ontologie est une spécification explicite d'une conceptualisation. [...] Lorsque la connaissance d'un domaine est représentée dans un formalisme déclaratif, l'ensemble des objets qui peuvent être représentés est appelé l'univers du discours. Cet ensemble d'objets, et les relations formalisées entre eux, sont reflétés dans le vocabulaire de représentation avec lequel un programme basé sur les connaissances représente les connaissances. »

[GRUBER, 1993]

On peut aussi trouver la définition de [STUDER et al., 1998] : « Une ontologie est une spécification formelle et explicite d'une conceptualisation partagée. »

D'un point de vue formel, l'ontologie est définie comme un ensemble d'axiomes terminologiques qui décrivent des ensembles de concepts pour un domaine donné traduisant ainsi la connaissance acquise dans ce domaine. L'ontologie est ainsi restreinte à un domaine d'intérêt défini à l'avance [USCHOLD et GRUNINGER, 2009]. On gardera à l'esprit qu'une ontologie est un objet qui fournit une description d'un domaine par un ensemble de primitives de représentation et qui a une portée applicative. Les primitives peuvent être des classes (ou des ensembles), des attributs (ou des propriétés) et des relations entre les membres des classes [GUARINO et al., 2009].

Exemple d'ontologie simple On peut illustrer le principe d'une ontologie en prenant comme exemple la modélisation des relations entre différents capteurs industriels. Construisons notre

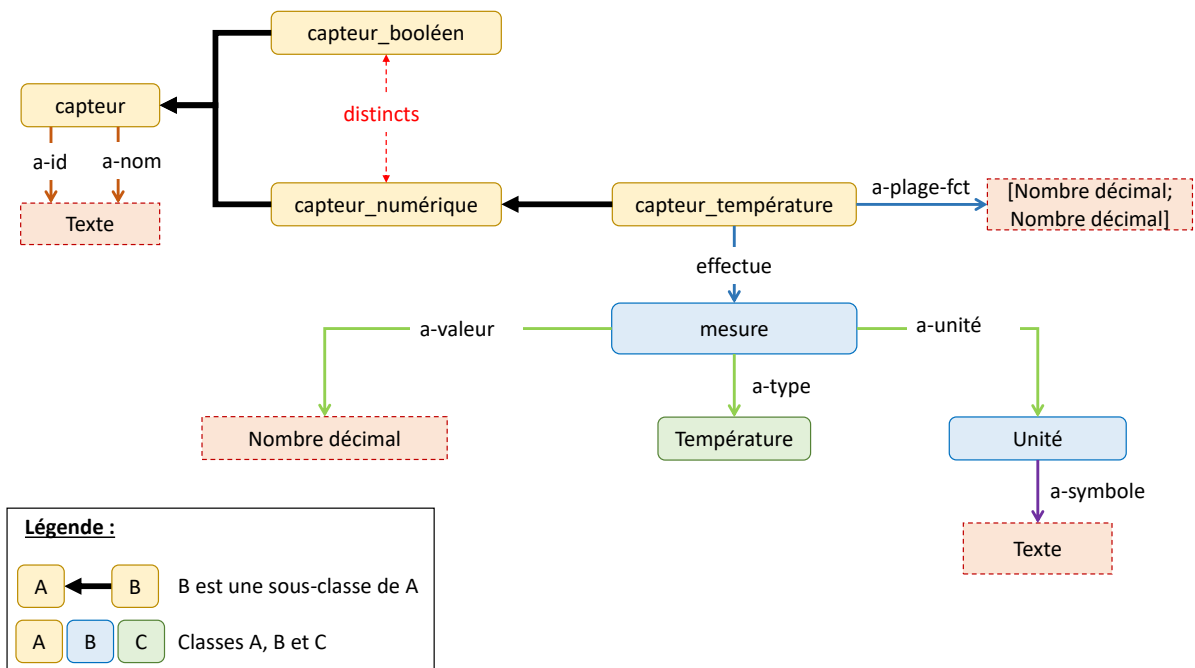


FIGURE 4.3 – Représentation de relations entre concepts et de l’instanciation de ces relations entre instances.

ontologie qui est présentée graphiquement dans la figure 4.3 Les classes principales sont capteur, capteur_booléen, capteur_numérique. Deux propriétés sont associées à la classe capteur : id_capteur (l’identifiant du capteur) et nom_capteur (le nom du capteur). capteur_booléen et capteur_numérique sont des sous-classes de capteur. Elles correspondent à des capteurs capturant respectivement des données vrai/faux et des données numériques. Elles sont définies comme disjointes, grâce à un axiome : aucune instance ne peut appartenir à la fois à la classe capteur_booléen et capteur_numérique. Le capteur de température (capteur_température), une sous-classe de capteur_numérique, effectue une mesure et a une plage de fonctionnement (a-plage-fct) définie par un intervalle [nombre décimal; nombre décimal]. La mesure est de type température (relation a-type), a une valeur numérique de type nombre décimal (relation a-valeur) et a une unité (a-unité) qui a elle-même un symbole (a-symbole) de type texte.

Imaginons deux instances qui sont respectivement des capteurs de température de l’étape de soudage numéro 1 et numéro 2 et appelons-les temp_soud1 et temp_soud2. Le capteur temp_soud1 est relié par la relation effectue à mesure11 et mesure12 qui sont deux instances du concept mesure. De même, temp_soud2 est relié à mesure21, instance de mesure. Le détail de ces instances est détaillé dans la figure 4.4. Une relation entre deux instances au sein d’une base de connaissances sera toujours exprimée avec le même formalisme que la relation mère définie entre deux classes de l’ontologie.

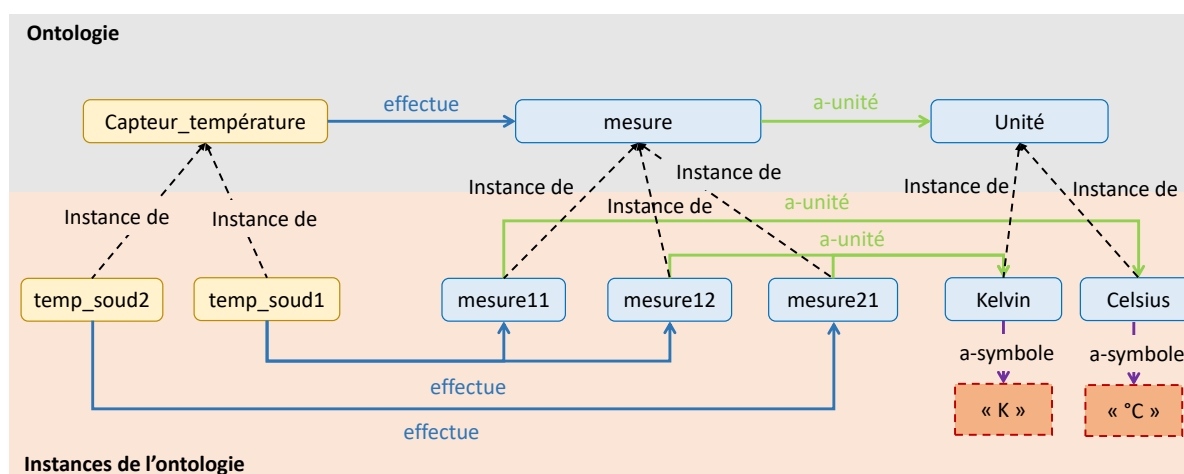


FIGURE 4.4 – Représentation de relations entre concepts et de l’instanciation de ces relations entre instances.

Famille d’ontologies Dans la littérature, il est souvent décrit quatre grandes familles d’ontologies suivant leur niveau d’abstraction [STUDER et al., 1998] :

Ontologies de représentation C’est les ontologies avec un niveau d’abstraction le plus haut et correspondent à aucun domaine. On retrouve dans ces ontologies les langages ou les métastructures de représentation. Il est possible de citer la *Basic Formal Ontology* [ARP et al., 2015] ou la *Frame Ontology*

Ontologies génériques D’un niveau d’abstraction inférieur elles sont spécifiques à un certain nombre de domaines d’intérêt. Par exemple, dans le domaine de la culture des plantes, une ontologie de référence **Ref-TO**, a été proposée [ARNAUD et al., 2012]. Cette ontologie incorpore trois autres ontologies utilisées par des experts de différentes cultures des plantes.

Ontologies de domaine Elles se limitent à la description d’un domaine spécifique

Ontologies applicatives Elles sont dans un domaine spécifique dédiées à des applications précises.

4.2.1.2 Ressource ontologique et terminologique (RTO)

La notion de Ressource Ontologique et Terminologique (RTO) provient des travaux de [REYMONNET et al., 2007] et de [ROCHE et al., 2009]. Ils proposent d’associer une partie terminologique et/ou linguistique aux ontologies afin d’établir une distinction claire entre la manifestation linguistique (le terme) et la notion qu’elle dénote (le concept). *Simple Knowledge Organisation System* (SKOS), proposé dans les travaux de [TOUHAMI et al., 2011], est une recommandation du *World Wide Web Consortium* (W3C) permettant de représenter les thésaurus, classifications et autres vocabulaires contrôlés. Il s’appuie sur le langage RDF afin de permettre la publication facile de vocabulaires structurés pour leur utilisation dans le cadre du web sémantique. La représentation s’établit sous forme de triplets RDF.

4.2.1.3 Méthodes d'extraction automatique des données issues d'articles scientifiques

En général, les données expérimentales pertinentes contenues dans les articles scientifiques sont dispersées dans différentes parties du document et exprimées sous différents formats. Dans ce contexte, l'automatisation de l'extraction et de l'annotation d'informations à partir de textes et de tableaux est une question majeure qui doit être discutée. Il y a principalement deux types d'extraction automatique dans le texte d'éléments avec une information pertinente.

D'une part, il y a l'extraction de relations binaires. Les premières approches pour découvrir les relations entre les entités se concentraient sur un contexte linguistique limité et reposaient sur la découverte de cooccurrences et sur la correspondance manuelle de modèles [HUANG et al., 2004]. Des techniques basées sur des règles manuelles, comme des expressions régulières dans les textes ou des balises, ont été utilisées pour construire des modèles linguistiques ou syntaxiques [HAO et al., 2005; HAWIZY et al., 2011; RAJA et al., 2013]. Cependant, les règles définies manuellement nécessitent un effort humain important. Plus tard, les approches basées sur l'apprentissage automatique, par exemple les machines à vecteurs de support [VAPNIK, 2000] ont été largement utilisées pour résoudre des tâches de classification [ROSARIO et HEARST, 2004; VAN LANDEGHEM et al., 2009]. Ces méthodes ont montré leur utilité mais nécessitent une grande quantité de données annotées pour l'apprentissage, ce qui nécessite généralement d'énormes efforts humains pour y parvenir [MINARD et al., 2011]. Étant basées sur des modèles numériques, elles peuvent ne pas être directement compréhensibles par l'utilisateur final.

D'autre part, l'extraction de relations n-aires (c'est-à-dire de relations ayant plus de deux arguments) est une tâche plus complexe, bien qu'elle soit nécessaire dans notre contexte. Des travaux ont été menés en divisant l'extraction de relations n-aires en trois étapes principales. La première étape consiste à identifier les entités (ou arguments) en utilisant des ressources telles que des ontologies ou des dictionnaires. La deuxième consiste à identifier le mot déclencheur de la relation (mot-clé propre à la relation) en utilisant des méthodes basées sur des dictionnaires ou des approches basées sur des règles. Elles permettent de construire des modèles à partir des résultats de l'analyse syntaxique des dépendances [LE MINH et al., 2011], ou de méthodes d'apprentissage automatique [BUI et SLOOT, 2011]. Enfin, dans la troisième étape, des relations binaires sont construites en utilisant le mot déclencheur. Puis des méthodes d'apprentissage automatique sont utilisées pour classer si les relations binaires appartiennent ou non à la relation n-aire recherchée, mais avec une perte substantielle de précision. Les méthodes d'extraction de relations sont en général basées sur ces trois étapes indépendantes. Dans notre contexte, les arguments de la relation n-aire peuvent être implicitement exprimés dans le texte et apparaissent généralement dans plusieurs phrases. Par conséquent, ces méthodes, qui font l'hypothèse de la présence d'un mot déclencheur, ne sont pas directement utilisables.

Abordons maintenant l'extraction automatique d'informations pertinentes dans les tableaux. Les méthodes et outils actuels partent du principe que les tableaux de données sont organisés de la même manière que dans les bases de données relationnelles : un tableau de données est composé de colonnes verticales (chaque colonne correspondant à une seule caractéristique) [BUCHE et al., 2013b]. Cette hypothèse n'est pas toujours valable pour les tableaux de données publiés dans des articles

scientifiques. Plusieurs caractéristiques peuvent être présentes dans la même colonne ou les tableaux peuvent avoir deux entrées. Un prétraitement automatique robuste des tableaux de données doit être conçu et validé afin d'appliquer les outils aux tableaux de données extraits de documents scientifiques. C'est pourquoi, nous considérons la méthode développée dans le paragraphe 4.2.2 comme une méthode semi-automatique, l'annotation des données est faite manuellement avec l'ontologie utilisée comme un guide (et une aide) pour l'annotation. En IC, l'extraction automatique d'informations pertinentes à partir du texte et des tableaux d'articles scientifiques est un domaine de recherche actif. Il n'existe pas encore d'outils prêts à l'emploi, mais un nombre croissant d'ontologies émergent pour l'organisation et le partage des connaissances dans des domaines particuliers, et de telles tâches d'extraction sont réalisées dans diverses applications. Par exemple, une ontologie française [MS]²O, regroupe les données relatives aux procédés de transformation dans les sciences alimentaires [DIBIE et al., 2016]. Cette ontologie permet à différentes équipes de travailler sur le même sujet et de regrouper leurs données dans une seule base de données, ce qui permet de comparer différents scénarios de production. Un autre exemple est *Rosanne*, développé comme un "plug-in" Excel[®] et construit à partir d'une ontologie des quantités et des unités de mesure [RIJGERSBERG et al., 2011].

4.2.2 Description de l'étape de collection et d'extraction des données

La figure 4.5 illustre les différentes étapes de cette extraction : (i) la sélection des articles (ii) la création d'une RTO (iii) l'extraction automatique des tables de l'article et la validation (iv) l'annotation sémantique des autres données et (v) le stockage. Avec ces données extraites, il peut aussi être pertinent d'avoir un score de fiabilité pour chaque article pour qualifier ces données. Nous décrirons à la suite des sous-étapes, un exemple de score de fiabilité qui est développé pour notre cas d'étude.

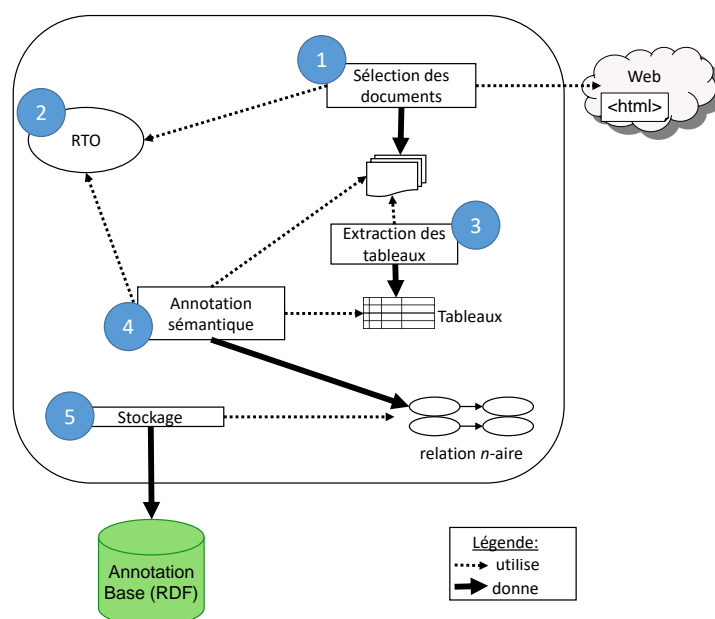


FIGURE 4.5 – Schéma de l'extraction des données d'intérêt des articles scientifiques (adapté de [LOUSTEAU-CAZALET et al., 2016]).

4.2.2.1 Sélection des documents

La première sous-étape est la sélection des documents pertinents décrivant les différents procédés ou systèmes industriels qui sont comparés dans l'étude. Les experts (en particulier les chercheurs ou ingénieurs spécialisés dans le domaine spécifique étudié) identifient tous les articles publiés grâce à différents mots clés dans les bases de données d'articles scientifiques comme Web of Science[®] ou Science Direct[®] grâce à des outils de recherche bibliographiques dédiés. Une fois sélectionné, le logiciel développé avec la RTO doit permettre de stocker les références bibliographiques des articles ainsi que leur texte intégral au format HTML ou PDF. Pour actualiser ces documents sélectionnés, l'expert doit explorer régulièrement les sites ou il peut accéder au "web caché" via l'abonnement à des systèmes d'alerte tels que les flux RSS sur les sites des bases de données scientifiques nommés précédemment.

4.2.2.2 Création d'une ressource ontologique et terminologique

La deuxième sous-étape est la création d'une RTO pour faciliter l'utilisation de données provenant de sources hétérogènes et pour guider l'annotation des données scientifiques. Cette RTO fait la distinction entre le concept (composante conceptuelle contenue dans une ontologie) et son expression linguistique dans différentes langues (composante terminologique) [TOUHAMI et al., 2011]. Dans la composante conceptuelle, la représentation d'une expérience est donnée comme une relation n-aire entre un résultat donné et plusieurs paramètres expérimentaux. Ces relations n-aires sont utilisées pour créer des tableaux annotés. Les colonnes du tableau annoté qui décrit cette opération unitaire correspondent aux arguments de la relation n-aire.

Il est important de noter que la RTO que l'on construit ici n'est pas construite pour annoter les tableaux de n'importe quels documents du web. La RTO appartient au domaine de l'étude et est créée juste pour ce domaine. Il est possible de rendre plus génériques ces ontologies, mais c'est un domaine qui est encore très récent et en pleine recherche [CHASSERAY et al., 2021]. Trois facteurs influencent la modélisation d'une RTO : le domaine d'intérêt, la tâche à réaliser et l'application [REYMONET et al., 2007]. Le domaine d'intérêt est l'étude de données expérimentales quantitatives contenues dans les articles sélectionnés et dépendant du domaine de l'étude. À noter que les données quantitatives expérimentales extraites requièrent une gestion des quantités avec leurs unités de mesure par la RTO. La tâche à réaliser est l'annotation et l'interrogation de tableaux de données par des relations n-aires. Un tableau peut être représenté par une ou plusieurs relations n-aires, au sens des bases de données relationnelles. Les colonnes du tableau, qui correspondent aux différents arguments de la ou des relations peuvent être reliées par une ou plusieurs relations sémantiques. La RTO a été définie pour représenter ces relations sémantiques. Finalement, l'application est la construction d'un entrepôt de données contenant les données des articles scientifiques sélectionnés.

La description détaillée de la RTO est présentée en annexe B et provient de travaux antérieurs. Nous décrivons succinctement les différentes composantes de cette RTO c'est-à-dire la composante conceptuelle et la composante terminologique.

Composante conceptuelle

La composante conceptuelle de la RTO est composée de deux parties :

- Une ontologie noyau qui permet de représenter des relations n-aires décomposées en deux sous-parties :
 - L'ontologie noyau supérieur permettant de représenter des relations n-aires comme n'importe quels arguments
 - L'ontologie noyau inférieur permettant de représenter les relations n-aires entre les données quantitatives expérimentales.
- Une ontologie de domaine qui permet de représenter les concepts spécifiques à un domaine donné.

L'ontologie noyau est assez générique et peut être utilisée pour différents domaines : elle appartient à la famille des ontologies génériques décrites dans le paragraphe 4.2.1.1. C'est dans sa partie supérieure que vont être retrouvés les concepts génériques *Relations* et *Arguments* permettant de représenter les relations n-aires et leurs arguments respectivement. Dans sa partie inférieure, les concepts *Dimension*, *Unit* et *Quantity* permettent de gérer les quantités et leurs dimensions ainsi que les unités associées à ces dimensions. Il est aussi important d'ajouter dans cette partie un concept qui va permettre de gérer les conversions entre les unités de mesure. Et enfin, un concept symbolique pour les arguments non numériques c'est-à-dire les données exprimées de manière qualitative et les arguments textuels comme le nom des produits par exemple. C'est l'ontologie de domaine qui va contenir les concepts spécifiques au domaine particulier de l'utilisateur de notre cadre méthodologique.

Composante terminologique

La composante terminologique de la RTO contient l'ensemble des termes du domaine étudié. Les concepts génériques *Relation*, *Symbolic*, *Unit* et *Quantity* sont chacun dénotés par un ou plusieurs termes de la composante terminologique. Cette composante permet de gérer grâce aux différents synonymes ou abréviations d'un même terme grâce à des labels. Par exemple, le terme anglais *Carbon dioxide* et le terme symbolique CO_2 dénotent un concept symbolique nommé *Carbon_dioxide*.

4.2.2.3 Extraction des tableaux

La troisième sous-étape est l'extraction des tableaux de données des documents choisis. Les tableaux de données contenus dans les documents sont extraits automatiquement de la version HTML en utilisant l'analyse des balises. Une fois cette extraction faite, l'expert peut valider les tableaux trouvés.

4.2.2.4 Annotation

C'est l'annotation sémantique manuelle des données guidée par la RTO précédemment créée. L'annotateur sélectionne, parmi les concepts des relations n-aires définies dans la RTO, ceux qui sont

pertinents pour décrire les tableaux extraits dans l'étape précédente ou pour annoter les données contenues dans le texte de l'article scientifique et qui ne sont pas contenues dans les tableaux précédemment extraits. La sélection du concept relation va permettre de guider l'expert dans sa saisie notamment de renseigner les arguments obligatoires et les importants. L'expert va être aussi aidé dans sa saisie des différentes unités et quantités en ne permettant que les unités de mesure possibles pour l'argument et en proposant des conversions d'unités. L'ontologie doit bien sûr permettre de renseigner des intervalles. Au fur et à mesure de l'annotation, l'expert doit valider celles-ci.

Il est possible d'automatiser encore plus cette étape. Nous utilisons dans ce manuscrit la méthode d'annotation sémantique d'un tableau par des concepts relations décrite dans [BUCHE et al., 2013a]. Cette méthode consiste à associer de manière semi-automatique des annotations floues à chaque instance de concepts relations identifiées dans le tableau, une ligne du tableau pouvant être annotée par plusieurs concepts relations. Ces instances de concepts relations permettent d'annoter les données contenues dans les différentes lignes des tableaux avec le même vocabulaire que celui utilisé dans les bases locales et ainsi d'enrichir les bases locales avec de nouvelles données pertinentes, qui sont directement exploitables par les experts du domaine. Les différentes étapes de cette méthode sont représentées dans la figure 4.6. Pour l'automatisation des données contenues dans le corps du texte, des méthodes issues du traitement automatique du langage sont dédiées au traitement des données textuelles. À la fin, toutes les données de l'article, pertinentes pour l'étude, sont contenues dans les tables annotées.

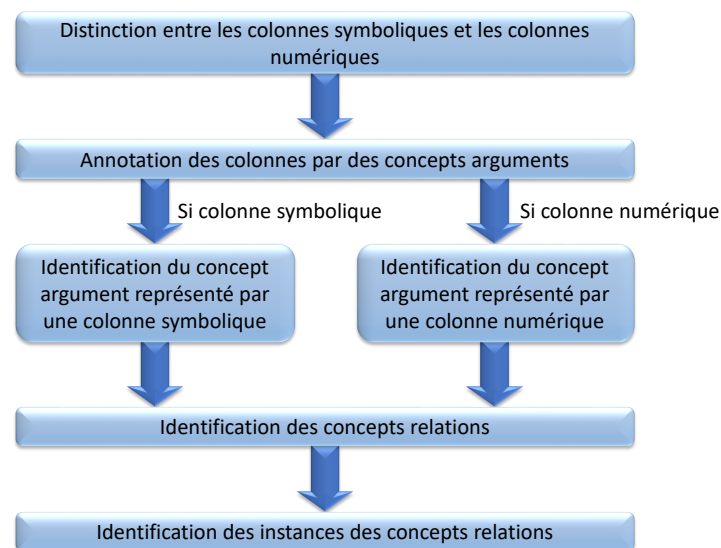


FIGURE 4.6 – Les principales étapes de l'annotation sémantique d'un tableau guidée par la RTO (adapté de [BUCHE et al., 2013a]).

4.2.2.5 Stockage dans un *Resource Description Framework*

Toutes les tables annotées sont stockées dans une base de type triplet RDF. RDF est un mode standard d'échange de données via Internet. Il est utilisé comme interface entre les utilisateurs et la RTO, permettant aux utilisateurs d'interroger la RTO de différentes façons. Une description de res-

sources réalisée par RDF comme c'est le cas ici, exprime des énoncés relatifs aux ressources web sous la forme de triplets :

$$\langle \text{subject} \rangle \langle \text{predicate} \rangle \langle \text{object} \rangle \quad (4.1)$$

Avec comme définitions des termes susmentionnés dans le triplet :

- *subject* : ressource à décrire
- *predicate* : type de propriété applicable à la ressource
- *object* : valeur de la propriété

Les utilisateurs peuvent exploiter un outil « *query* » pour classer les données dans un ordre spécifique, par exemple en sélectionnant un type de données spécifiques. Au final, cette base de données contenant les tables d'annotations contient les données de premier plan nécessaires à l'étape 3 Analyse de la durabilité de notre cadre méthodologique.

4.2.2.6 Score de fiabilité à partir des métadonnées

Lors de la collecte de données à partir de divers documents, la question se pose quant à la fiabilité de ces données ou de ces documents. Il existe des outils d'estimation de la fiabilité. Nous nous basons sur l'outil détaillé dans [DESTERCKE et al., 2013] dont on rappelle rapidement les bases et qui a été développé pour notre cas d'étude. L'outil vise à fournir une estimation automatique et a priori (c'est-à-dire en évitant un examen spécifique) de la fiabilité du document et des données à partir d'un ensemble de méta-informations liées aux données et au document.

Pour établir ce score de fiabilité, S groupes $A_1; \dots; A_S$ de méta-informations importantes sont d'abord établis en accord avec les parties prenantes. Un groupe A_i prend en compte C_i valeurs $a_{i1}; \dots; a_{iC_i}$.

Différents types de méta-informations et métadonnées peuvent être considérés :

- Les méta-informations sur la source de données elle-même, par exemple le type de source (publication scientifique, rapport technique), la réputation de la source, les données de citation, la date de publication...
- Les méta-informations liées aux moyens utilisés pour produire les données. Dans les articles basés sur des expériences, ces informations sont généralement incluses dans une section qui décrit en détail le protocole expérimental et le matériel. Certaines méthodes peuvent être connues pour être moins précises que d'autres, mais sont néanmoins choisies pour des considérations pratiques;
- Les méta-informations liées aux procédures statistiques : présence de répétitions, quantification de l'incertitude (c'est-à-dire variance, intervalle de confiance), élaboration d'un plan expérimental.

En pratique, les groupes sont constitués de manière à ce que leur impact sur la fiabilité puisse être estimé indépendamment, un groupe A_i pouvant contenir plusieurs critères (par exemple, le nombre de citations et la date de publication). Pour chaque valeur possible de chaque groupe, la méthode consiste alors à évaluer la fiabilité d'un document ou d'une donnée ayant cette valeur

particulière. Une fois les groupes formés, pour chaque valeur $a_{ij}, i = 1, \dots, S; j = 1, \dots, C_i$, un expert du domaine donne son avis sur la fiabilité des données dont la méta-information est a_{ij} . Cette opinion est exprimée linguistiquement et est choisie parmi un ensemble de modalités limitées (ou des combinaisons de celles-ci), par exemple très peu fiable, légèrement peu fiable, neutre, légèrement fiable, très fiable et inconnu.

À chaque document o , sont ensuite associées S évaluations linguistiques (en fonction de la valeur prise par la méta-information correspondante). Une valeur manquante dans la méta-information est simplement traitée comme une évaluation linguistique « inconnue » en termes de fiabilité. Afin de traiter des informations potentiellement conflictuelles (car certaines méta-informations peuvent indiquer un document peu fiable, tandis que d'autres peuvent désigner un document plutôt fiable), les évaluations linguistiques sont traduites dans un format numérique, en utilisant la notion d'ensembles flous, qui sont des modèles numériques adéquats pour des valeurs linguistiques. Afin d'avoir suffisamment de flexibilité, ces ensembles flous sont définis sur un espace de fiabilité fini ordonné $\Theta = \{\theta_1, \dots, \theta_5\}$ de 5 éléments, θ_1 étant la valeur de fiabilité la plus basse, θ_5 la plus haute. Le nombre d'éléments pourrait être supérieur à 5, mais c'est un choix raisonnable. En effet, ce nombre doit rester impair afin d'avoir un élément neutre, pas trop faible pour que les ensembles flous correspondant à des termes différents puissent être numériquement bien distingués et pas trop élevés pour que des problèmes de calcul ne se posent pas. Chaque modalité est ensuite transformée en un ensemble flou sur Θ et présenté dans la figure 4.7.

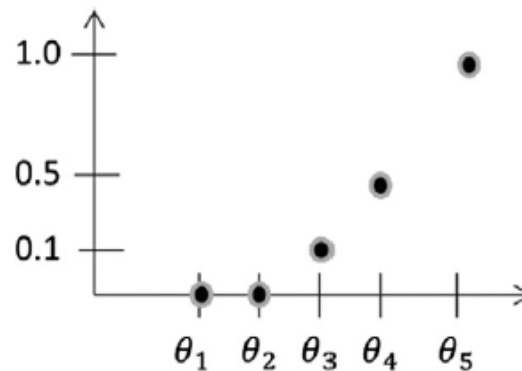


FIGURE 4.7 – Ensemble flou correspondant au terme "très fiable" défini sur Θ avec 5 éléments issu de [LOUSTEAU-CAZALET et al., 2016].

Les S ensembles flous $\mu_{a_1^o}, \dots, \mu_{a_S^o}$, correspondant à la fiabilité de groupe du document o sont ensuite fusionnés en utilisant la théorie de l'évidence (ou théorie de Dempster-Shafer [DEMPSTER, 1967]) et une approche de sous-ensemble cohérent maximal permettant de prendre en compte les données contradictoires. Une telle approche vise à réconcilier toutes les sources tout en conservant le plus d'informations possibles. Ce qui signifie que les informations manquantes (c'est-à-dire la présence de la modalité inconnue) n'ont pas d'impact sur le résultat. Le conflit potentiel dans les méta-informations (c'est-à-dire l'évaluation d'une fiabilité élevée pour un aspect, mais de faibles fiabilités pour un autre) se reflète dans l'imprécision du modèle final. La présence d'un conflit se traduira par une estimation assez imprécise de la fiabilité, alors que son absence donnera lieu à

une estimation assez précise. Le résultat de cette fusion est une fonction de masse (ou *Basic Belief Assignment*) $m_O : 2^\Theta \rightarrow [0, 1]$ qui reflète la fiabilité globale du document o .

Le masse m_O fournit une synthèse précise des différentes contributions des méta-informations, et son analyse pourrait permettre d'identifier automatiquement des sous-ensembles de méta-informations contradictoires et cohérents. Cependant, elle est encore trop complexe pour être analysée d'un seul coup d'œil. Pour cette raison, une autre synthèse est fournie, dans laquelle l'estimation de la fiabilité sous la forme d'un intervalle $[\underline{E}_O; \overline{E}_O]$ est proposée. Le score final \underline{E}_O est calculé à l'aide de la formule suivante :

$$\underline{E}_O = \sum_{E \subset \Theta} m_o(E) \inf_{\theta_i \in E} f(\theta_i) \quad (4.2)$$

Où $f(\theta_i) = i$, c'est-à-dire que chaque θ_i est remplacé par son rang correspondant.

\overline{E}_O est obtenu avec la même formule en remplaçant la borne inférieure par la supérieure. La longueur de l'intervalle $[\underline{E}_O; \overline{E}_O]$ reflète dans quelle mesure les différents éléments de la méta-information sont cohérents : \overline{E}_O fournit en quelque sorte la "pire" fiabilité possible, tandis que \underline{E}_O correspond à la "meilleure" fiabilité possible ainsi plus l'intervalle entre les deux est faible plus l'information est cohérente.

Pour conclure, cet intervalle de fiabilité peut être entré dans la base de données pour fournir un nouveau paramètre d'analyse des données.

4.3 Enrichissement et stockage des données extraites

Toutes les données de premier plan nécessaires à la poursuite de notre cadre méthodologique sont maintenant stockées dans des bases de données structurées, faciles à exploiter. Il s'agit alors de stocker toutes les données dont on a besoin dans une seule base de données structurée. C'est à partir de cette base qu'on va effectuer l'enrichissement et les étapes de traitement, d'analyse et de visualisation des données. Ceci nous permet d'avoir des bases de données brutes sans aucun traitement auxquelles il est toujours possible de se référer. La figure 4.8 détaille les enchaînements des différentes bases de données au cours de ce chapitre. L'enrichissement des données est le processus d'ajout des données provenant d'experts ou de modèles dans le domaine étudié. Suivant le domaine étudié, les modèles peuvent être des simulations numériques empiriques, des simulations thermodynamiques ou énergétiques, des fonctions de transfert de matière ou de transformation voir des réactions chimiques. Cet enrichissement dépend des parties prenantes de l'étude et des modèles disponibles dans le domaine de l'étude.

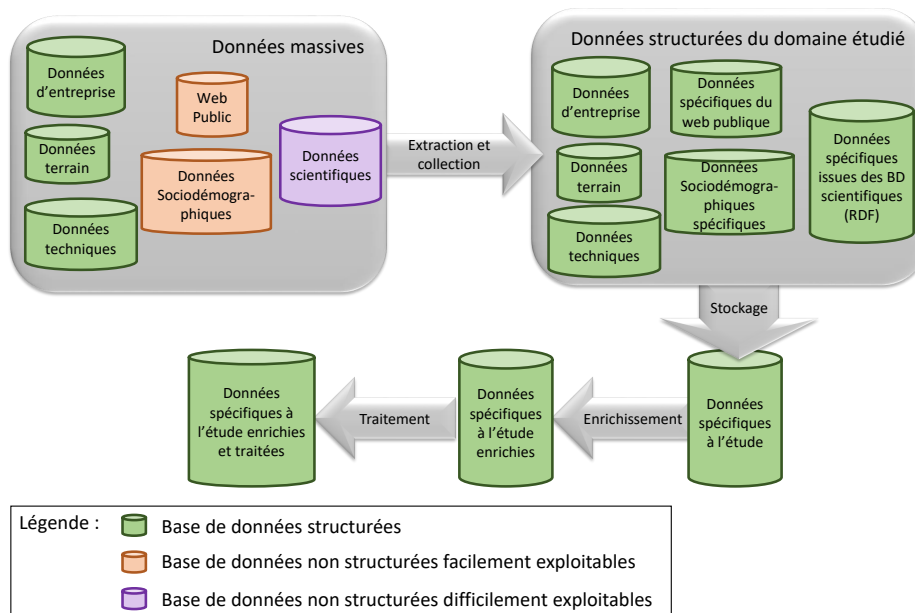


FIGURE 4.8 – Les différentes bases de données lors de l'étape d'Architecture des données.

4.4 Traitement des données extraites

La sous-étape de traitement des données implique le nettoyage, l'ajout et la suppression de données pour la gestion du volume et de la valeur de la base de données étudiée. Après ce traitement, une dernière base de données, plus précise et plus accessible, peut être générée (voir figure 4.8). Le traitement permet de gagner du temps dans les étapes suivantes et évite une interprétation incorrecte lors de la sous-étape d'analyse des données brutes et l'étape d'analyse de la durabilité. Cette étape est faite par l'expert d'une manière manuelle ou semi-automatique.

4.5 Analyse des données

Les données de premier plan ou données procédés sont brutes après l'étape précédente de traitement. Il est alors possible de passer directement à l'étape « Analyse des impacts de la durabilité » ou d'analyser ces données pour produire de l'information pertinente. Cette étape peut servir à plusieurs objectifs :

- Un apport de connaissances sur la base de données constituée;
- Une analyse des données pour trouver des incohérences ou compléter les données;
- Une analyse brute des données.

L'analyse brute des données dépend de l'objectif, du domaine et des parties prenantes de l'étude. Plusieurs types d'analyse sont possibles : l'analyse descriptive (que s'est-il passé?), l'analyse diagnostique (pourquoi cela s'est-il produit?), l'analyse prédictive (que va-t-il se passer?) et l'analyse prescriptive (comment pouvons-nous faire en sorte que cela se produise?).

Analyses descriptives et diagnostiques Elles utilisent un certain nombre de méthodes et d'algorithmes différents, tels que la compression, l'écart-type, l'analyse de dépendances non

linéaires ou linéaires, l'analyse factorielle et les méthodes de classification (réseaux de neurones, arbres de décision, réseaux bayésiens ...). Certaines de ces méthodes sont visuelles, et la visualisation des données peut donc être incluse dans cette sous-étape. Elle peut également être réalisée en traçant les données brutes sous la forme d'un graphique simple ou interactif. Les outils d'analyses descriptives et diagnostiques sont nombreux. Trois acteurs se détachent particulièrement : Microsoft[®] (avec PowerBi[®]), Tableau[®] et Qlik[®].

Analyse prédictive C'est un terme utilisé pour décrire une série de techniques analytiques et statistiques permettant de prédire des actions ou des comportements futurs. Afin de pouvoir faire des prédictions sur la base d'un ensemble de données spécifiques, on utilise une ou plusieurs variables prédictives pour prédire une variable réponse. Vu la complexité des données dont ont besoin les analyses prédictives, c'est les méthodes de ML qui sont utilisées : l'apprentissage supervisé et l'apprentissage non supervisé.

Analyse prescriptive Elle utilise les deux concepts d'analyse descriptive et prédictive, dans un but principal : Proposer des voies d'optimisations aux utilisateurs, de l'aide à la décision, leur permettant ainsi de réagir plus rapidement, et de façon la plus appropriée, à une situation en devenir. C'est sa capacité à donner à l'utilisateur les moyens de prendre des décisions rapidement, chose qu'il n'aurait pas pu forcément faire en raison du nombre d'indicateurs à prendre en compte et de leur complexité, qui rend le prescriptif intéressant.

Dans le cadre méthodologique, nous nous intéressons plus spécifiquement à l'analyse prédictive qui va permettre de répondre aux objectifs de cette partie. La prédiction va nous permettre d'apporter une certaine connaissance à nos données brutes. Le choix est fait d'utiliser une méthode d'apprentissage supervisé, plus précisément les réseaux de neurones artificiels.

4.5.1 Réseaux de neurones

Les RN est un outil d'apprentissage automatique largement utilisé dans de nombreuses applications industrielles telles que l'analyse de la maintenance des machines, le contrôle des processus de fabrication ou la planification [HAGAN et al., 2014]. Les RN s'inspirent de la biologie et notamment les tentatives d'imitation du traitement des informations dans le cerveau humain. Un intérêt pour le neurone biologique qui est le composant principal du traitement de l'information s'est alors développé. Il est constitué de quatre parties distinctes :

- un corps cellulaire qui est la structure fondamentale des neurones abrite le noyau où l'information est traitée. L'influx nerveux, ou signal électrique, qui émane de ce corps cellulaire reflète l'état d'activité du neurone [BOULARAS et DJALAB, 2020];
- les dendrites, captant les informations provenant des autres neurones, représentent les entrées principales du neurone;
- les synapses véhiculent l'information vers le neurone en moyennant des poids synaptiques;
- l'axone représente la sortie du neurone.

Les réseaux de neurones artificiels ont été construits par analogie. Le composant unitaire de ce réseau est le neurone artificiel illustré par la figure 4.9. La sortie du réseau représente toute(s) valeur(s) quantitative(s) d'intérêt pour l'étude. Cela peut donc être suivant les études soit des réseaux

de neurones à sortie unique ou à sortie multiple. On va expliciter rapidement le principe d'un réseau de neurones à sortie unique.

Le traitement d'un vecteur de données d'entrée $\langle X_1, X_2, \dots, X_n \rangle$ dans un RN s'effectue en deux étapes : la somme pondérée des entrées provenant des neurones connectés, en utilisant les poids synaptiques des liaisons, et ajoute un biais θ . Le biais agit comme un neurone supplémentaire dont la valeur est toujours de 1 et qui est lié aux autres neurones par un poids. En d'autres termes, il ajoute une somme de poids supplémentaires pour effectuer une transformation affine sur la somme pondérée des entrées. La fonction d'agrégation est alors exprimée sous la forme $\sum_{i=1}^n x_i \times w_i + \theta$, où w_i représente le poids synaptique associé à l'entrée i . Le résultat de cette somme pondérée est ensuite passé en paramètre d'une fonction d'activation ϕ associée au noeud. La sortie du neurone correspond alors à la valeur de $\phi(\sum_{i=1}^n x_i \times w_i + \theta)$. Lorsqu'une entrée n'est pas connectée au neurone, le poids synaptique lié à cette entrée prend la valeur de 0.

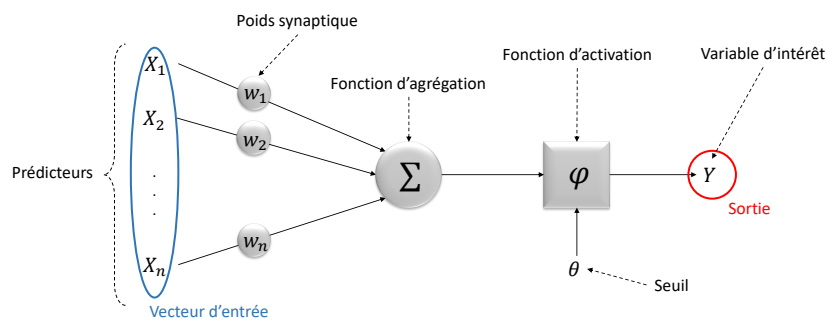


FIGURE 4.9 – Schéma général d'un neurone artificiel formel (issu de [CHEKROUN, 2009]).

La fonction d'activation ϕ , également appelée fonction de transfert, est généralement non linéaire et reflète le potentiel d'activation d'un neurone. Elle joue un rôle similaire au fonctionnement biologique en déterminant si l'information doit être transmise ou non, en fonction de l'atteinte ou non du seuil de stimulation. Son utilisation est intéressante car elle permet de capturer et représenter les dépendances non linéaires. Si l'on se limite uniquement à la fonction d'agrégation, qui est une combinaison linéaire, le réseau ne serait efficace que dans les cas de dépendances linéaires. Il existe de nombreuses fonctions d'activation courantes, et le choix dépend du problème spécifique à résoudre. Elles sont souvent bornées entre 0 et 1, ou entre -1 et 1, pour éviter les calculs impliquant de grands nombres et une utilisation excessive de la mémoire. Cependant, certaines fonctions d'activation ne sont pas bornées et peuvent être utiles, par exemple, lors des entraînements lents pour éviter une saturation potentielle due à l'approximation progressive de 0. De plus, toutes les fonctions d'activation ne sont pas nécessairement non linéaires. Le choix d'une fonction d'activation dépend donc des cas d'utilisation et des caractéristiques souhaitées pour ces fonctions. Historiquement, l'étude des fonctions d'activation a conduit à la formulation du théorème d'approximation universelle, démontrant que les réseaux de neurones peuvent approximer n'importe quelle fonction continue avec une précision arbitraire, à condition d'avoir une fonction d'activation non linéaire [CYBENKO, 1989]. [HORNİK, 1991] a démontré que ce n'est pas tant le choix particulier de la fonction

d'activation qui confère aux RN le potentiel d'être des approximateurs universels, mais plutôt l'utilisation d'une architecture multi-couches.



Un réseau de neurones acyclique peut approximer n'importe quelle fonction non linéaire.

[CYBENKO, 1989; HORNIK, 1991]

La figure 4.10 présente un exemple de fonction d'activation couramment utilisée dans les réseaux de neurones : la fonction sigmoïde. Cette fonction est appréciée pour plusieurs raisons, notamment sa facilité de dérivation, qui accélère le calcul de la dérivée et réduit ainsi le temps d'entraînement. De plus, ses bornes entre 0 et 1 permettent une interprétation probabiliste de l'activation du neurone.

La fonction sigmoïde est définie comme suit : $\phi(z) = \frac{1}{1 + e^{-z}}$, où z est le résultat de la fonction d'agrégation. Pour un neurone avec n entrées connectées chacune au neurone avec un poids w_i , l'activation de la sortie du neurone est donnée par $\phi(\sum_{i=1}^n x_i \times w_i + \theta) = \frac{1}{1 + e^{-\sum_{i=1}^n x_i \times w_i + \theta}}$. Il est important de noter que la normalisation des données d'entrée est nécessaire pour éviter que les variables ayant de très grandes valeurs n'écrasent celles qui prennent de petites valeurs. Lorsqu'on utilise la fonction sigmoïde pour une classification, l'apprentissage ajuste les poids w_i et le biais θ de manière à ce que $\phi(\sum_{i=1}^n x_i \times w_i + \theta)$ prenne une grande valeur (positive) pour les observations $\langle X_1, X_2, \dots, X_n \rangle$ dont la sortie attendue est 1, et une petite valeur (négative) pour les observations dont la sortie attendue est 0. Cela permet d'effectuer une séparation claire entre les deux classes lors de la classification.

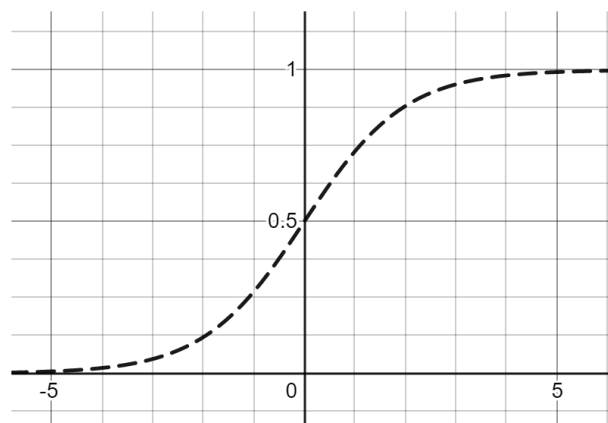


FIGURE 4.10 – Graphique présentant la fonction sigmoïde dans l'intervalle [-5; 5].

L'atout majeur des RN et qui dans notre contexte est ce qui nous a poussés à nous intéresser à ceux-là pour l'analyse de nos données brutes est leur capacité de traitement de relations complexes. La capacité des RN à devenir des approximateurs universels réside dans leur capacité à permettre la circulation de l'information entre plusieurs neurones interconnectés, disposés en une ou plusieurs couches intermédiaires avant d'atteindre la sortie finale. Bien que les fonctions d'activation utilisées

fréquemment soient relativement simples et monotones, leur combinaison permet d'approcher n'importe quelle fonction non linéaire. Cette caractéristique est particulièrement présente dans les réseaux de neurones de type "perceptron multi-couches" (MLP), où l'information circule dans un sens direct, c'est-à-dire des entrées vers la sortie en traversant une ou plusieurs couches cachées. Cette organisation en couches intermédiaires, associée à la non-linéarité des fonctions d'activation, offre une grande puissance de représentation et d'adaptation, permettant aux réseaux de neurones de résoudre une grande variété de problèmes complexes.

Ces dernières années avec la montée en puissance des puissances de calcul, les RN connaissent de fortes performances prédictives. Néanmoins, ces très bons résultats s'accompagnent généralement d'une difficulté à interpréter d'une part le processus de génération du modèle et d'autre part le résultat appris. C'est pour cela que les RN sont connus comme des « modèles boîte noire ». Dans l'industrie, des contraintes viennent se mettre en place qui fait que cet effet « boîte noire » n'est pas le bienvenu. Par exemple à travers la mise en place d'une collaboration homme-machine dans un contexte I4.0, la prise en main par un opérateur humain des algorithmes de ML est plus facile avec une bonne compréhension de ceux-là. Plus précisément, il s'agit d'extraire les informations nécessaires pour interpréter les résultats de la prédiction fournis par les algorithmes IA. Il est aussi possible de citer l'exigence de certification des modèles dans l'industrie ou l'explication comme aide au diagnostic. L'ajout des sciences de l'ingénieur en ML tente de proposer des outils et/ou méthodes afin de rendre les algorithmes interprétables. Cette exigence permet d'aller vers une hybridation des modèles boîtes noires avec les modèles de connaissances de type analyse de bilan ou modèles à base d'équations.

4.5.2 Interprétabilité des réseaux de neurones

Il existe plusieurs définitions pour l'interprétabilité en ML. Nous gardons une des définitions non mathématiques du terme celle de [DOSHI-VELEZ et KIM, 2017] qui dit que l'interprétabilité est la capacité d'expliquer ou de présenter une décision ou une prédiction en termes compréhensibles à un humain. En termes clairs, plus un modèle est interprétable, plus il est facile de comprendre comment les décisions ou les prédictions ont été faites.



L'apprentissage automatique interprétable est un terme générique utile qui capture l'«*extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model*» [MURDOCH et al., 2019].

Modèles d'interprétation par l'exemple L'idée principale derrière ces modèles est l'explication d'un algorithme par raisonnement analogique. En d'autres termes, le résultat de l'algorithme d'un entrant est expliqué par analogie avec un autre entrant soit généré soit issu de la base d'apprentissage. Il existe plusieurs modèles d'interprétation par l'exemple comme les prototypes [KIM et al., 2016], les explications par l'exemple contre-factuel [WACHTER et al., 2018], par l'exemple contradictoire, les instances influentes ou les modèles de k plus proches voisins.

Méthodes d'interprétation spécifique aux modèles Ou méthode dépendante du modèle. Ce sont

les modèles de type régression linéaire et logistique, ou d'arbres de décision qui sont interprétables par nature étant donnée leur faible complexité et leurs structures. Par exemple, les réseaux bayésiens sont interprétables de par leurs structures probabilistique et graphique.

Méthodes agnostiques aux modèles Appelées aussi méthodes d'interprétation indépendantes du modèle, ces méthodes présentent l'avantage d'être totalement découplées du modèle d'apprentissage automatique utilisé. Les trois grands avantages de ces méthodes sont (Ribeiro et al., 2016) :

- La flexibilité du modèle : la méthode d'interprétation peut fonctionner avec n'importe quel modèle d'apprentissage automatique.
- La flexibilité d'explication : l'explication n'est pas limitée à une certaine forme suivant les cas elle peut être un graphique ou une formule linéaire par exemple.
- La flexibilité de représentation : le système d'explication doit pouvoir utiliser une représentation des caractéristiques différentes du modèle d'apprentissage automatique explicité.

Ces méthodes sont souvent séparées en deux groupes les méthodes globales et les méthodes locales. Sans être exhaustif, on nommera dans les méthodes globales : les diagrammes *Partial Dependence Plot* (PDP) ou un de ses dérivés la méthode des diagrammes *Accumulated Local Effects* (ALE), la méthode de la décomposition fonctionnelle ou les modèles de substitution globale (*global surrogate model*). Dans les méthodes locales, il est possible de citer l'algorithme *Local Interpretable Model-agnostic Explanations* (LIME), la méthode *Shapley* ou sa variante la méthode *SHAP* (*SHapley Additive exPlanations*), la méthode *Anchors* ou les méthodes *Individual Conditional Expectation* (ICE). Toutes les méthodes d'interprétabilité énoncées sont utilisables pour les RN et pour toutes méthodes issues du ML supervisé. Il est aussi possible de mixer ces méthodes voire de créer de nouvelles méthodes spécifiquement développées pour les RN.

4.5.3 Méthodes PDP et ICE

La dépendance partielle aide à comprendre l'effet marginal d'un prédicteur (ou d'un sous-ensemble de celui-ci) sur le résultat prévu. Essentiellement, cela nous permet de comprendre comment la variable d'intérêt change lorsque nous modifions la valeur d'un prédicteur tout en tenant compte de l'effet moyen de toutes les autres prédicteurs du modèle.

La procédure suivante suit la méthodologie traditionnelle documentée dans [FRIEDMAN, 2001]. L'algorithme de la méthode PDP (algorithme 1) divise la variable d'intérêt en j valeurs équidistantes. L'algorithme crée d'abord une grille régulièrement espacée composée de j valeurs à travers la distribution d'un des prédicteurs. Puis il crée j copies des données d'apprentissage d'origine (une copie pour chaque valeur). Il fixe alors le prédicteur choisi pour toutes les observations de la première copie à la première valeur de la grille, puis à la deuxième valeur de la grille pour la deuxième copie, ... et ainsi jusqu'à la $j^{\text{ème}}$ valeur de la grille pour la $j^{\text{ème}}$ copie. L'algorithme prédit ensuite le résultat pour chaque observation dans chacune des j copies, puis calcule la moyenne des valeurs prédites

pour chaque ensemble. Ces valeurs prédites moyennes sont appelées valeurs de dépendance partielle et sont tracées par rapport aux j valeurs régulièrement espacées. L'algorithme peut être assez gourmand en calcul puisqu'il implique j appels à la fonction de prédiction (fonction d'agrégation et d'activation). Heureusement, l'algorithme peut être parallélisé assez facilement [GREENWELL, 2017]. Si nous traçons les valeurs de dépendance partielle par rapport aux valeurs de la grille (les 20 valeurs de température dans l'exemple ci-dessous) nous obtenons ce qu'on appelle le diagramme de dépendance partiel (PDP). Dans ce diagramme (figure 4.11), la ligne en bleu sur le graphique représente la valeur moyenne prédite pour toutes les observations à chacune des j valeurs du prédicteur choisi. Dans la figure 4.12, les PDP sont les deux lignes rouges.

🔍 Prenons l'exemple d'un prédicteur nommé température dont les valeurs sont comprises entre 0 et 100. Pour $j = 11$, la grille de température sera $[0, 10, 20, 30, \dots, 100]$. L'algorithme fixe alors le prédicteur choisi température pour toutes les observations de la première copie à 0, puis à 10 pour la deuxième copie, 20 pour la troisième, ... et enfin 100 pour la 11^{ème} copie. Il prédit ensuite le résultat pour chaque observation dans chacune des 11 exemplaires puis calcule la moyenne des valeurs prédites pour chaque exemplaire.

Algorithme 1 : PDP – Partial Dependence Plot

Params : Liste de Valeurs X : le prédicteur étudié

Entier j : nombre de valeurs équidistantes utilisé.

Résultat : graphique en deux dimensions

pour X **faire**

Construire une grille de j valeurs uniformément espacées à travers la distribution de X : $\{X_1, X_2, \dots, X_j\}$

pour $i \in \{1, \dots, j\}$ **faire**

Copier les données d'apprentissage et remplacer les valeurs originales de X par la constante X_i

Appliquer le modèle ML donné pour chaque observation // c.-à-d.

obtenir un vecteur de prédictions

Moyenner les prédictions

Tracer les prédictions moyennes en fonction de $\{X_1, X_2, \dots, X_j\}$

Cette méthode a plusieurs avantages : la création des diagrammes PDP est assez intuitive et facile à mettre en œuvre, pour des variables non corrélées l'interprétation est très claire. De plus, l'interprétation est causale : pendant l'exécution de l'algorithme, nous n'intervenons que sur une variable et mesurons les changements dans la prédiction ainsi c'est l'analyse causale entre la variable et la prédiction [ZHAO et HASTIE, 2017]. Il y a plusieurs limites à cette méthode. La plus importante et qui peut perturber le plus les interprétations dans nos travaux est l'hypothèse de l'indépendance. L'autre limitation principale est que les effets hétérogènes peuvent être masqués, car les diagrammes PDP ne montrent que les effets marginaux moyens. Par exemple, supposons que pour un prédicteur, pour toutes les copies nous trouvons la même prédiction ($moyenne(\hat{Y})$). Or cette prédiction n'est pas forcément la même pour chaque observation. La courbe construite par le PDP est alors horizontale. La conclusion serait alors que ce prédicteur n'a aucune influence sur la prédiction, ce qui peut être

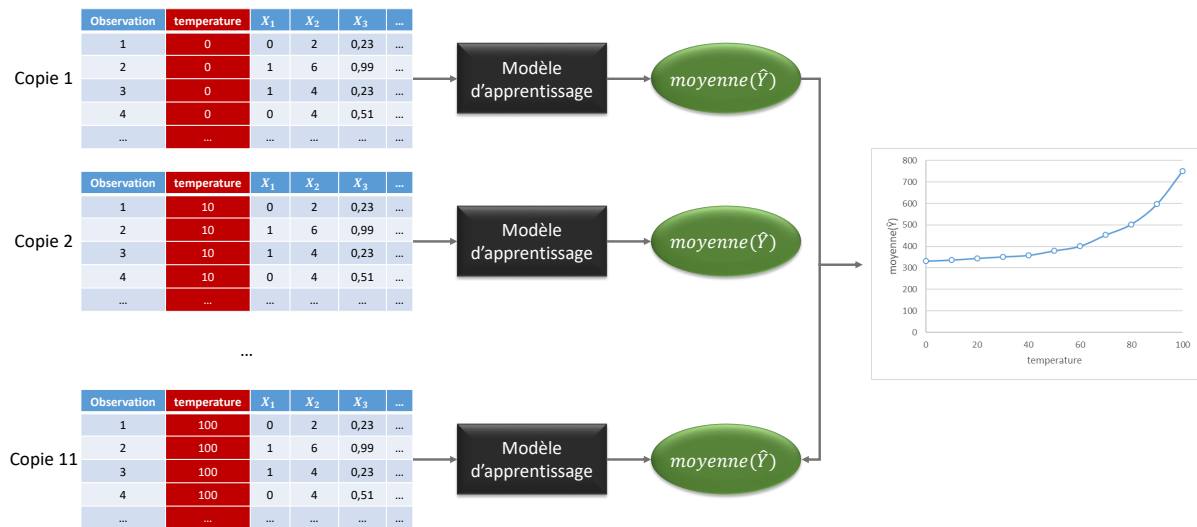


FIGURE 4.11 – Illustration du processus de la dépendance partielle jusqu’au PDP (adapté de [BOEHMKE et GREENWELL, 2019]).

faux. La méthode ICE, non globale, mais locale permet d’éviter cette lacune et c’est pour cela que dans ces travaux et dans le cas d’étude, il a été décidé d’utiliser les deux.

La méthode ICE, au lieu de faire la moyenne des valeurs prédites pour toutes les observations, trace les prédictions de chaque copie pour chaque observation. Ainsi, un tracé ICE visualise la dépendance de la réponse prédite sur un prédicteur pour chaque observation séparément. Ce qui se traduit par plusieurs lignes, une pour chaque observation, par rapport à une ligne pour le tracé PDP. Une courbe PDP est la moyenne de toutes les courbes d’un tracé ICE. L’algorithme ICE est donc le même que celui du PDP (algorithme 1) en supprimant la ligne qui moyenne.

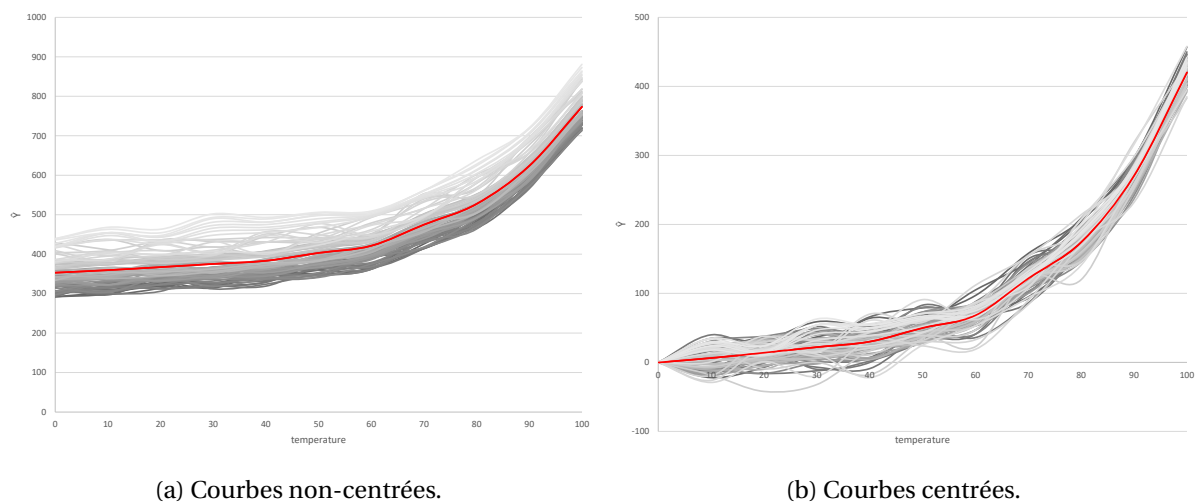


FIGURE 4.12 – Courbes ICE centrées (a) et non-centrées (b) (en noir) avec la courbe PDP (en rouge).

Précisons que souvent les différences entre les courbes ICE ne peuvent être identifiées qu'en centrant la fonction. Par exemple, la figure 4.12 affiche les courbes ICE de notre exemple précédent. La figure 4.12a montre que toutes les observations ont des effets très similaires sur toutes les valeurs de température. Cependant, la figure 4.12b montre des courbes ICE centrées, ce qui aide à mettre en évidence plus clairement l'hétérogénéité et attire également davantage l'attention sur les observations qui s'écartent du schéma général.

4.5.4 Conclusion sur l'apport des réseaux de neurones

Les RN et les méthodes d'interprétabilité utilisées apportent une certaine forme de connaissance sur notre base de données de premier plan. Premièrement, il est possible de générer des données en simulant le modèle de prédiction, permettant ainsi à l'utilisateur de ne pas faire de nouvelles expériences pour compléter la base de données. Deuxièmement, il est possible à travers des méthodes d'interprétabilité, en analysant l'impact des entrées sur la sortie du modèle, de détecter l'incohérence de certaines données. Si toutes ces données incohérentes proviennent d'un même article scientifique, l'utilisateur pourrait faire le choix de supprimer cet article. Finalement, l'utilisateur va potentiellement découvrir des relations entre des variables de sa base de données de premier plan qu'il n'avait pas envisagées entraînant un surplus de connaissances sur sa base de données.

4.6 Visualisation des données

La dernière sous-étape est celle de la visualisation des données. Elle peut être incluse dans la précédente comme avec les méthodes d'interprétabilité ou elle peut être distincte. Il est ainsi possible d'utiliser les outils de visualisation des données (*data visualization*) interactifs. La visualisation interactive est souvent utilisée dans ce type d'analyse dans le domaine de la *Business Intelligence*. Ces outils permettent de collecter des données dans diverses sources structurées, de les traiter et de créer des tableaux de bord pour visualiser les données et produire une analyse descriptive.

4.7 Conclusion du chapitre 4

Ce chapitre détaille la deuxième étape du cadre méthodologique. Cette étape comporte plusieurs sous-étapes avec pour chacune d'entre elles les apports effectués par différentes méthodes et outils issus de domaines différents.

La première contribution réside dans l'enchaînement des sous-étapes. L'étape est générique et issue du *Business Intelligence* et de l'architecture des données massives. Elle permet de répondre à notre question Q1 : Comment exploiter et tirer bénéfice des données massives pour aider les analyses et les évaluations du développement durable? Le chapitre met en valeur les différentes méthodes utilisées qui proviennent du domaine de l'intelligence artificielle et d'un de ses sous-domaines l'ingénierie des connaissances. En effet, on a de nombreuses données provenant de sources hétérogènes structurées ou non et en fin d'étape est produite une base de données structurées traitées, enrichies et analysées prête à être exploitée dans des analyses de la durabilité.

La deuxième contribution est l'ajout d'outils de l'ingénierie des connaissances et notamment l'utilisation d'une ressource ontologique et terminologique pour l'extraction et la collecte des données. Cette ressource permet l'extraction, la collecte et la structuration de données issues des articles scientifiques sans perte de connaissance sur ces données. Un score de fiabilité est défini et apporte un ajout de connaissances.

La troisième contribution est l'interprétation de méthodes d'intelligence artificielle et notamment les réseaux de neurones et leur interprétabilité pour l'analyse des données.

Chapitre 5

Analyse et visualisation des résultats de l'évaluation de la durabilité

"*Nam et ipsa scientia potestas est*"
En effet le savoir lui-même est pouvoir

Sir Francis Bacon - 1597

Ce chapitre (figure 5.1) a pour objectif de développer l'étape 4 du cadre méthodologique : Analyse et visualisation des résultats. Cette étape a pour point d'entrée la matrice de l'évaluation de la durabilité. Cette matrice [intrants × procédés × impacts] est de très grande dimension, et difficile à analyser. Ce chapitre va explorer les méthodes issues de l'apprentissage automatique pour analyser et visualiser les résultats pour aider les décideurs à faire un premier classement des procédés industriels les plus durables pour le produit d'intérêt de l'étude.

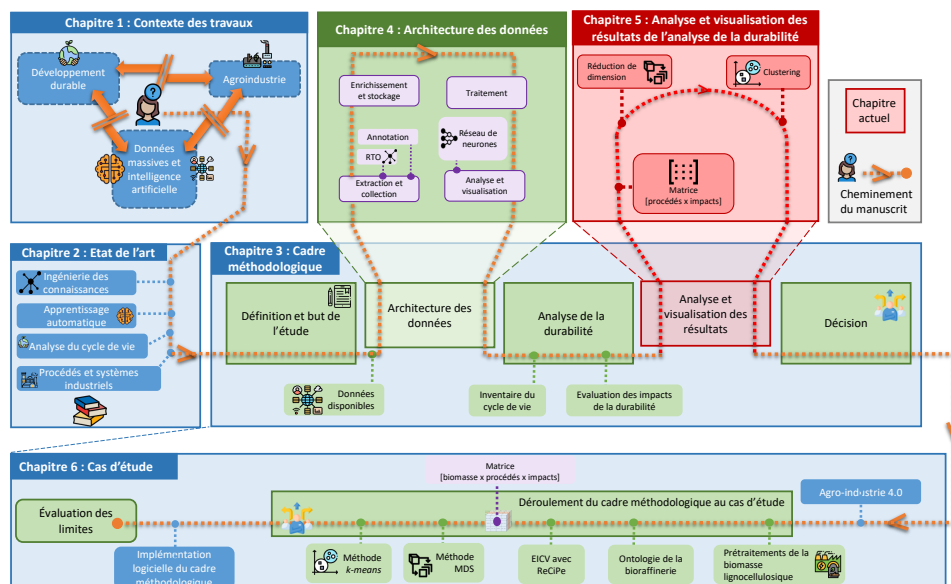


FIGURE 5.1 – Positionnement du chapitre 5 dans le manuscrit.

Sommaire

5.1	Introduction du chapitre 5	103
5.2	Catégories d'impacts et méthodes	103
5.3	Apprentissage non-supervisé pour l'analyse des résultats de l'évaluation de la durabilité	106
5.3.1	Méthodes de réduction de la dimension	107
5.3.1.1	Approches par sélection de caractéristiques	108
5.3.1.2	Approches par extraction de caractéristiques	108
5.3.1.2.1	Méthodes linéaires	108
5.3.1.2.2	Méthodes non linéaires	109
5.3.1.3	Conclusion sur les méthodes de réduction de dimension	110
5.3.2	Méthodes de <i>clustering</i>	110
5.3.2.1	<i>Clustering</i> hiérarchique	110
5.3.2.2	<i>Clustering</i> par partitionnement	111
5.3.2.3	<i>Clustering</i> basé sur la densité de probabilité	112
5.3.3	Conclusion sur les méthodes d'apprentissage	112
5.4	Positionnement multidimensionnel et k-moyennes	112
5.5	Conclusion sur le chapitre 5	114

5.1 Introduction du chapitre 5

Ce chapitre répond à la question Q2 de cette thèse : « **Comment bénéficier des méthodes et outils issus de l'intelligence artificielle pour aider les analyses et les évaluations de la durabilité?** » La matrice à analyser de dimension [nombre d'intrants × nombre de procédés × nombre d'impacts] est difficile à analyser à cause de deux grands problèmes :

- La grande dimension
- L'hétérogénéité des impacts : ils n'ont ni les mêmes dimensions, ni les mêmes informations et ni les mêmes unités.

Dans la première partie de ce chapitre, nous signifiions les impacts environnementaux pour mettre en lumière la difficulté d'interpréter les résultats. Puis nous présentons des méthodes d'apprentissage automatique qui aident à l'analyse : la réduction de dimension et le *clustering*.

5.2 Catégories d'impacts et méthodes

Comme décrit dans le chapitre 3, l'évaluation environnementale des impacts du cycle de vie (EICVe) convertit tous les flux inventoriés en impact sur l'environnement. Trois niveaux d'agrégation des impacts sont possibles : les impacts *midpoint* et les impacts *endpoint* (aussi appelés dommage) et le score unique. La figure présente les caractérisations entre ces impacts. Les facteurs de caractérisation des impacts *endpoint* (**FCe**) sont dérivés des facteurs de caractérisation des impacts *midpoint* (**FCm**) grâce à la formule 5.1 adaptée de [HUIJBREGTS et al., 2017] :

$$FCe_{x,a} = FCm_{x,a} \times F_{m \rightarrow e,a} \quad (5.1)$$

a : une des trois catégories de dommage
 où : *x* : l'impact concerné
 $F_{m \rightarrow e,a}$: facteur de conversion *midpoint* – *endpoint* pour une catégorie de dommage *a*

Pour démonstration, la figure 5.2 représente la chaîne de causalité de la méthode **ReCiPe 2016** entre les catégories d'impacts *midpoint* et les trois dommages. Cette méthode est la méthode de calcul des impacts et dommages que nous retenons dans le chapitre 6.

Il existe de nombreuses méthodes de calculs pour ces impacts. Par exemple pour les méthodes européennes on pourra citer **ReCiPe** [HUIJBREGTS et al., 2017], **CML**, **IPCC** ou **ILCD** (European Commission, 2010). Les méthodes sont catégorisées en deux classes suivant leur orientation : les méthodes orientées problèmes appelées aussi méthodes *midpoint* et les méthodes orientées dommages appelées méthodes *endpoint* (voir Tableau 5.1). Les méthodes *midpoint* et les méthodes *endpoint* permettent de calculer les trois niveaux d'agrégation des impacts. C'est-à-dire qu'avec une méthode *midpoint* on peut calculer les impacts *midpoint* et les impacts *endpoint*. Les approches *midpoint* et *endpoint* sont complémentaires, mais la caractérisation *midpoint* a une relation plus étroite avec les flux environnementaux de l'inventaire, ce qui lui confère une incertitude plus faible [JOLLIET et al., 2010].

Même si la méthode dépend surtout de l'étude, **ReCiPe 2016 MidPoint** est la méthode la plus utilisée car elle est scientifiquement la plus solide et utilise des facteurs de caractérisation qui

TABLEAU 5.1 – Liste (non exhaustive) de méthodes de calculs des impacts suivant leur orientation

Méthodes orientées problèmes	Méthodes orientées dommages
EDIP	Eco-Indicator99
IMPACT 2002+	Impact 2002+
TRACI	CML 2002
UCAS	LIME
ReCiPe Midpoint	EPS2000
	ReCiPe Endpoint

ont récemment été mis à jour. **ReCiPe 2016**, nouvelle version de **ReCiPe 2008 Midpoint**, est une fusion de deux méthodes : elle prend les indicateurs *midpoint* de la méthode **CML** et les indicateurs endpoint de l'**Eco-indicateur 99**. Elle comporte 18 catégories d'impacts *midpoint* et 3 catégories de dommages (impacts *endpoint*) (figure 5.2). Chaque catégorie a une unité qui lui est propre et qui est détaillée dans le tableau 5.2. On retrouve bien dans cette méthode des unités très différentes et donc difficiles à interpréter pour des non-experts. Les unités sont plus homogènes dans la méthode **ReCiPe 2016 Endpoint** avec seulement trois unités qui correspondent aux unités des catégories de dommages. Le DALY (*disability adjusted life years*) est l'unité du dommage « Santé humaine » et représente les années perdues par une personne (mort ou handicapé) en raison d'une maladie ou d'un accident. L'unité du dommage « qualité des écosystèmes » correspond à la perte relative locale d'espèces dans les écosystèmes terrestres, d'eau douce et marine intégrée dans l'espace et le temps (fraction potentiellement d'espèces disparues en $espèces \times an$). Le troisième et dernier dommage est la « rareté des ressources » (couramment appelé « ressources naturelles ») qui représente les coûts supplémentaires liés à la future extraction des ressources minérales et fossiles. Il est exprimé en dollars (\$).

Pour conclure les catégories d'impact qu'elles soient *midpoint* ou *endpoint* sont très différentes avec des unités hétérogènes. Pour les décideurs il est difficile de choisir quel sera le procédé le plus durable surtout si certaines catégories sont contradictoires sur des procédés. Pour faciliter la décision, une possibilité est de passer directement au score unique, mais ce score ajoute une très grande perte d'information. Pour pallier à cette perte, le choix a été fait de se servir de méthodes issues de l'IA. Le chapitre 2 a permis de mettre en lumière l'apprentissage automatique (ML). Nous nous intéressons au ML pour analyser la matrice [*intrants* × *procédés* × *impacts*]. Pour *p* procédés, *i* intrants et l'utilisation de la méthode **ReCiPe 2016 EndPoint**, cette matrice est de dimension [*i* × *p* × 18]

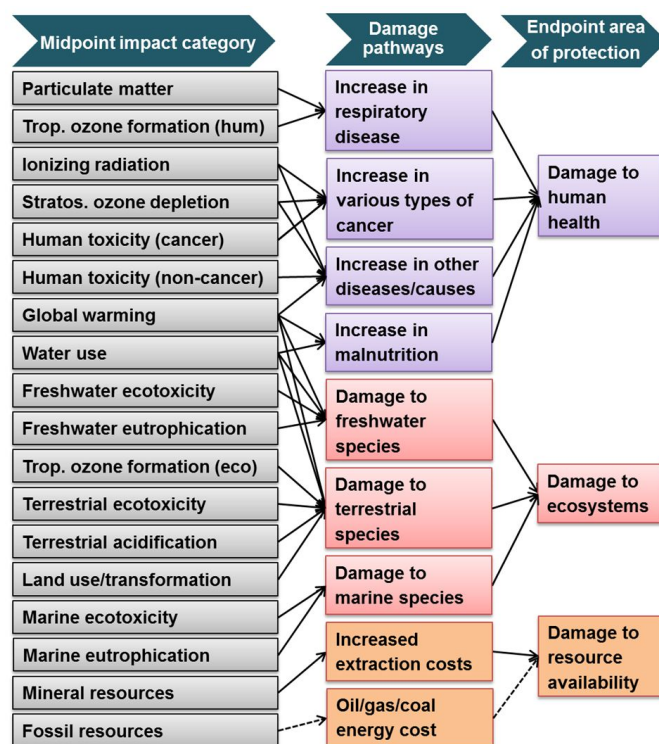


FIGURE 5.2 – Aperçu des catégories d'impact couvertes par la méthode **ReCiPe 2016 Midpoint** et de leur relation avec les zones de protection issu de [HUIJBREGTS et al., 2017]. La ligne en pointillé signifie qu'il n'y a pas de facteur constant entre le milieu et la fin pour les ressources fossiles.

TABEAU 5.2 – Liste des catégories d'impacts *midpoint* de la méthode **ReCiPe 2016** selon [HUIJBREGTS et al., 2017]

Catégorie d'impact midpoint	Midpoint impact category	Unité
Acidification terrestre	<i>Terrestrial acidification</i>	<i>kg SO₂ – eq to air</i>
Appauvrissement de l'ozone stratosphérique	<i>Appauvrissement de l'ozone stratosphérique</i>	<i>kg CFC – 11 – eq to air</i>
Changement climatique	<i>Global warming</i>	<i>kg CO₂ – eq to air</i>
Eutrophisation marine	<i>Marine eutrophication</i>	<i>kg P – eq to marinewater</i>
Eutrophisation d'eau douce	<i>Freshwater eutrophication</i>	<i>kg P – eq to freshwater</i>
Ecotoxicité marine	<i>Marine ecotoxicity</i>	<i>kg 1,4 – DBC – eq to marinewater</i>
Ecotoxicité terrestre	<i>Terrestrial ecotoxicity</i>	<i>kg 1,4 – DBC – eq to industrial soil</i>
Ecotoxicité d'eau douce	<i>Freshwater ecotoxicity</i>	<i>kg 1,4 – DBC – eq to freshwater</i>
Rareté des ressources minérales	<i>Mineral resource scarcity</i>	<i>kg Cu – eq</i>
Rareté des combustibles fossiles	<i>Fossil fuel scarcity</i>	<i>kg oil – eq</i>
Utilisation de l'eau	<i>Water use</i>	<i>m³ water – eq consumed</i>
Utilisation des sol	<i>Land use</i>	<i>m² × yr annual cropland – eq</i>
Formation de particules fines	<i>Fine particulate matter formation</i>	<i>m² × yr annual cropland – eq</i>
Formation d'ozone troposphérique (Santé humaine)	<i>Tropospheric ozone formation (Human health)</i>	<i>kg NO_x – eq to air</i>
Formation d'ozone troposphérique (Écosystème)	<i>Tropospheric ozone formation (Ecosystem)</i>	<i>kg NO_x – eq to air</i>
Rayonnement ionisant	<i>Ionizing radiation</i>	<i>kBq Co – 60 – eq to air</i>
Toxicité humaine (cancer)	<i>Human toxicity (cancer)</i>	<i>kg 1,4 – DBC – eq to urban air</i>
Toxicité humaine (non cancer)	<i>Human toxicity (non-cancer)</i>	<i>kg 1,4 – DBC – eq to urban air</i>

5.3 Apprentissage non-supervisé pour l'analyse des résultats de l'évaluation de la durabilité

D'après le chapitre 3, l'apprentissage non-supervisé permet d'identifier des classes ou des groupes d'individus présentant des caractéristiques communes. La qualité de ce type de méthode est mesurée par sa capacité à découvrir certains ou l'ensemble des motifs cachés. Dans l'étape 4 de notre cadre méthodologique, nous voulons analyser la matrice [*intrants* × *procédés* × *impacts*] et identifier des groupes dans les différents procédés ou identifier des motifs cachés et intéressants. Ce sont les méthodes de *clustering* (ou partitionnement des données) qui permettent d'extraire différents groupes homogènes appelés *clusters*. Cependant avant d'utiliser ces méthodes, il convient d'utiliser une méthode de réduction de la dimension pour réduire les dimensions de la matrice analysée (voir figure 5.3). Cette figure résume le traitement d'une matrice [*intrants* × *procédés* × *impacts*]. Tout d'abord, les techniques de DR projettent les données brutes des procédés dans un espace de dimension inférieure (2 ou 3). Après une projection des données par une technique de DR, l'approche de *clustering* est ensuite appliquée pour considérer les impacts et les procédés similaires dans l'espace de dimension inférieure. Enfin, l'utilisateur (expert) analyse les points regroupés en *clusters* pour les relier aux procédés et impacts significatifs. La visualisation des *clusters* de données aidera le chercheur ou les ingénieurs R&D dans la décision finale suivant l'objectif de leur étude.

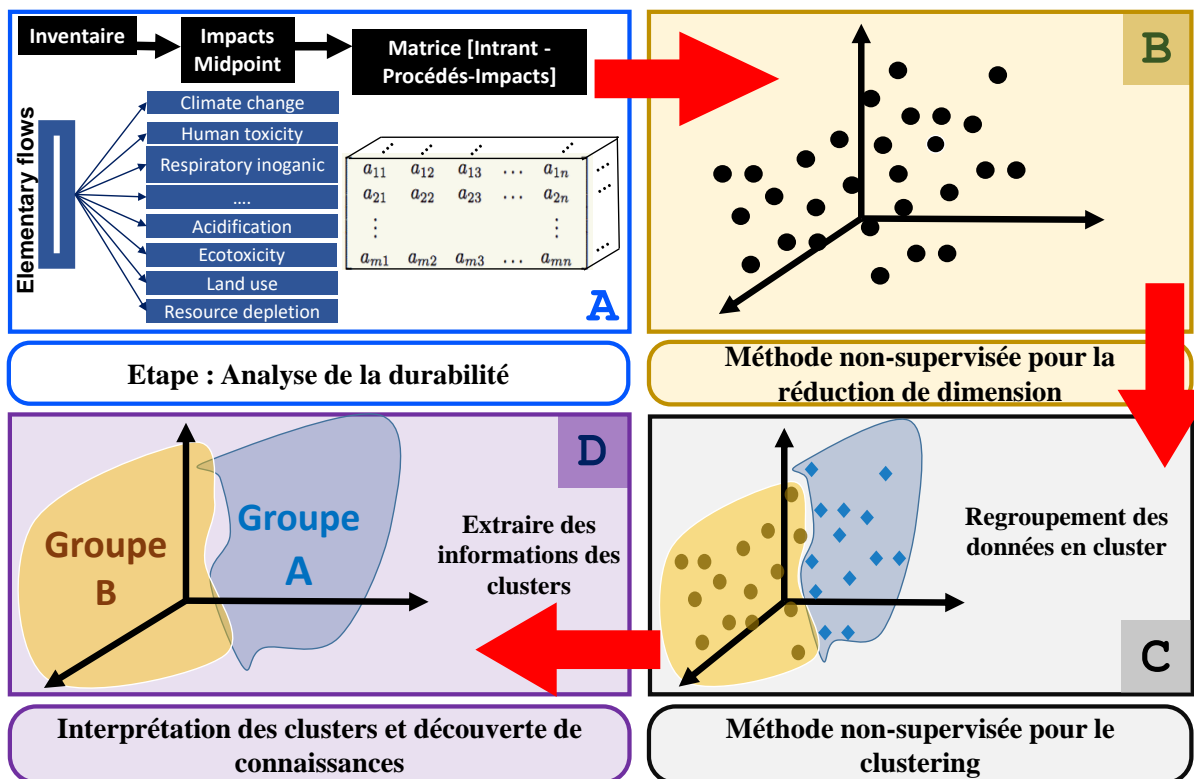


FIGURE 5.3 – Étape d'Analyse des résultats avec l'utilisation des méthodes de réduction de dimension et de *clustering*.

5.3.1 Méthodes de réduction de la dimension

Dans de nombreuses études, le nombre de variables utilisé pour représenter les données est très élevé. Par exemple, dans le traitement d'images hautes-résolutions, chaque pixel peut être représenté par plusieurs variable. Bien qu'une représentation des données qui contient plus de données est plus riche en information, il est plus difficile pour un modèle d'apprendre. La réduction de dimension est alors plébiscitée en ML pour pallier à cette difficulté. La réduction de dimension permet de :

- Visualiser les données : il est plus facile de visualiser des données en deux ou trois dimensions qu'en p dimensions.
- Réduire les coûts algorithmiques : c'est-à-dire le temps d'apprentissage sur les données, l'espace que les données prennent en mémoire mais aussi le coût d'acquisition des données si certaines variables sont redondantes.
- Améliorer la qualité des modèles : en évitant le surapprentissage des algorithmes et le fléau de la dimension.



Méthode de réduction de la dimension : toute méthode permettant de projeter des données issues d'un espace de grande dimension dans un espace de plus petite dimension. C'est-à-dire toute méthode permettant de transformer une représentation $X \in \mathbb{R}^{n \times p}$ des données en une représentation $X^* \in \mathbb{R}^{n \times m}$ où $m \ll p$



La représentation X peut être modélisée comme un tableau où une ligne correspond à une observation (n étant le nombre d'observations) et une colonne correspond à une caractéristique (p étant le nombre de caractéristiques).

Le fléau de la dimension (*curse of dimensionality* en anglais) est très courant en ML [FERNÁNDEZ-MARTÍNEZ et FERNÁNDEZ-MUÑIZ, 2020]. En effet, en grandes dimensions, il y a un besoin de beaucoup plus de points pour couvrir tout l'espace il est donc difficile de trouver ce qu'ils ont de commun ou de différent. Le fait de réduire la dimension permet de construire des modèles avec moins de paramètres ce qui permet d'éviter le surapprentissage. Les méthodes qui marchent en petite dimension (méthodes de fouilles de données) ne marchent pas forcément en grande dimension. Pour limiter le surapprentissage, on pourra citer aussi la technique de régulation qui consiste à contrôler simultanément l'erreur du modèle et sa complexité.

Deux types d'approches principales existent pour réduire la dimension d'un espace $\mathbb{R}^{n \times p}$ à un espace $\mathbb{R}^{n \times m}$:

- **Les approches par sélection de caractéristiques :** m caractéristiques sont conservée parmi les p caractéristiques.


$$(x_1, x_2, \dots, x_p) \xrightarrow{\text{sélection}} (x_{i_1}, x_{i_2}, \dots, x_{i_m})$$

- **Les approches par extraction de caractéristiques :** m nouveaux axes sont créés à partir des axes originaux.

$$(x_1, x_2, \dots, x_p) \xrightarrow{\text{extraction}} (y_1, y_2, \dots, y_m)$$

5.3.1.1 Approches par sélection de caractéristiques

Appelé *feature selection* en anglais, ce sont les approches les plus connues et les plus simples parmi celles de réduction de dimensionnalité. Des méthodes statistiques ou de notation sélectionnent les colonnes pertinentes à conserver de façon automatique ou manuelle. Les approches par sélection peuvent être séparées en trois catégories : les méthodes par filtrage, les méthodes de conteneur et les méthodes embarquées.

 Prenons l'exemple d'une sélection manuelle. Considérons un modèle qui prédit le poids d'un téléphone. Une colonne décrivant la couleur du téléphone n'aide pas à prédire le poids, tandis qu'une colonne décrivant le modèle peut être pertinente pour la prédiction. Ainsi on va donc supprimer la caractéristique couleur.

Méthodes par filtrage La sélection de caractéristiques par filtrage consiste à appliquer un critère de sélection indépendant à chacune des p variables. La pertinence de la p -ième variable du jeu de données est quantifiée par rapport à la prédiction.

Méthodes de conteneur Appelées *wrapper methods* en anglais, elles consistent à déterminer le meilleur sous-ensemble de variables pour un algorithme d'apprentissage donné. Ce sont souvent des méthodes heuristiques car il est généralement impossible de déterminer la performance d'un algorithme sur un jeu de données sans l'avoir entraîné.

Méthodes embarquées Appelées *embedded approaches* en anglais elles apprennent en même temps que le modèle les caractéristiques à y inclure. Au fur et à mesure de l'apprentissage des caractéristiques sont éliminées du modèle jusqu'à en avoir plus que m . La méthode LASSO (*least absolute shrinkage and selection operator*) [TIBSHIRANI, 1996] est une des plus connues.

Ces méthodes sont assez simples mais assez limitantes dans des cas réels soit par le temps de calcul soit parce qu'elles saturent soit parce qu'elles obtiennent des résultats inexacts. Par exemple, la méthode LASSO sature en très grande dimension et fausse la prédiction quand des caractéristiques sont très corrélées entre elles.

5.3.1.2 Approches par extraction de caractéristiques

L'objectif de l'extraction de caractéristiques est de représenter le plus fidèlement possible les échantillons dans un espace de plus faible dimension. Les méthodes linéaires et les méthodes non linéaires sont les deux grandes familles.

5.3.1.2.1 Méthodes linéaires Les techniques de projection linéaire des données en sous-espace de dimension inférieure sont une étape préliminaire dans la réduction de dimension. Elle permet de

réduire la dimension de l'espace original et rendre l'ensemble des données plus représentatif du problème. La nouvelle représentation des données est une combinaison linéaire des données originales, c'est-à-dire :

$$\mathbf{Y} = \mathbf{\Omega}\mathbf{X} \quad (5.2)$$

Avec $\mathbf{X} = \{x_{ij}\}_{i=1..n, j=1..m}$, $\mathbf{Y} = \{y_{ki}\}_{k=1..r, i=1..n}$, $\mathbf{\Omega} = \{\omega_{kj}\}_{k=1..r, j=1..m}$ la matrice de transformation linéaire et $1 < k < r, 1 < j < m$.

Les deux méthodes principales trouvées dans la littérature sont [KALIA et al., 2020] :

L'analyse en composante principale (ACP) [DAZY et al., 1996] : C'est une méthode projective non supervisée et linéaire. Le critère à maximiser est la variance dans les données projetées. L'hypothèse faite dans cette méthode est que les données de départ se trouvent dans un hyperplan et que l'on peut les exprimer au moyen des vecteurs qui définissent cet hyperplan. Si l'hypothèse se révèle vraie, on trouve la dimensionnalité intrinsèque des données. Cette méthode produit les nouvelles caractéristiques, appelées composantes principales (CP), en déterminant la variance maximale des données. Pendant la transformation, la première CP a la variance la plus élevée, et les CP suivantes ont des variances décroissantes. L'ACP permet de sélectionner un nombre de dimensions inférieur ayant une variance maximale. Assez souvent, on sélectionne un nombre de CP ayant un pourcentage de 75-80 % de la variance totale. Elle permet de réduire l'information en un nombre de composantes plus limité que le nombre initial de variables. Cependant, la limite principale est que la transformation est linéaire et ne capte pas les relations qui peuvent être plus complexes.

Le positionnement multidimensionnel (MDS)

L'algorithme MDS est basé sur une recherche de valeurs propres. MDS permet de construire une configuration de n points dans R^m à partir des distances entre n objets. Elle transforme une matrice de distance en un ensemble de coordonnées de telle sorte que les distances dérivées de ces coordonnées se rapprochent le plus possible des distances originales.

5.3.1.2.2 Méthodes non linéaires L'objectif des méthodes non linéaires de réduction de la dimension est toujours le même : trouver une représentation significative à basse dimension pour un ensemble de points de dimension élevée. Toutefois, à l'opposé des méthodes précédentes, les méthodes non-linéaires proposent de mieux approximer la structure géométrique réelle de l'ensemble des données. Le problème vient du fait que deux points peuvent être à proximité selon la distance euclidienne, mais très éloignés si on mesure la distance sur la surface définie par les points. La solution à ce problème est de ne plus considérer les distances euclidiennes, mais plutôt d'utiliser les distances géodésiques. Or, c'est là le problème pour les méthodes linéaires, elles ne reposent que sur la distance euclidienne. Dans les méthodes non linéaires, *Isometric Feature Mapping* (Isomap) reprend l'idée du MDS mais en utilisant une table de distance géodésiques et la méthode de l'**Insertion linéaire locale** réduit la dimension tout en préservant la structure globale non linéaire. D'autres méthodes existent comme ACP Kernelisée, cartes propre Laplacienne, Cartes de Sammon, Cartes de diffusion.

5.3.1.3 Conclusion sur les méthodes de réduction de dimension

Les méthodes sont nombreuses et dépendantes des données d'entrées. Dans nos travaux, les méthodes les plus utilisées sont les méthodes d'apprentissage non supervisé linéaire et non linéaire. Suivant l'étude, certaines méthodes peuvent ne pas donner de résultats il faut alors tester plusieurs méthodes pour trouver la plus pertinente c'est-à-dire celle qui donne le plus d'informations. Ceci n'est pas compliqué étant donné que la plupart des algorithmes se trouvent dans des bases de données. Par exemple, en Python, la bibliothèque `sklearn` possède la plupart de ces algorithmes.

5.3.2 Méthodes de *clustering*

Les méthodes de *clustering* (ou *data clustering* en anglais) sont des méthodes en analyse des données. Elles visent à regrouper un ensemble de données en différents groupes homogènes nommés "*cluster*". Ces derniers sont des sous-ensembles dont les données partagent des caractéristiques communes. Ces caractéristiques correspondent le plus souvent à des critères de proximité (similarité au sens mathématique) que l'on définit en introduisant des mesures et classes de distance entre objets [HINTON et SEJNOWSKI, 1999]. Pour obtenir un bon *clustering*, il convient à la fois de minimiser l'inertie intra-classe pour garantir des *clusters* les plus homogènes possibles, et de maximiser l'inertie interclasse afin de garantir des sous-ensembles bien différenciés. Le centre de gravité d'un *cluster* est nommé centroïde (figure 5.4).

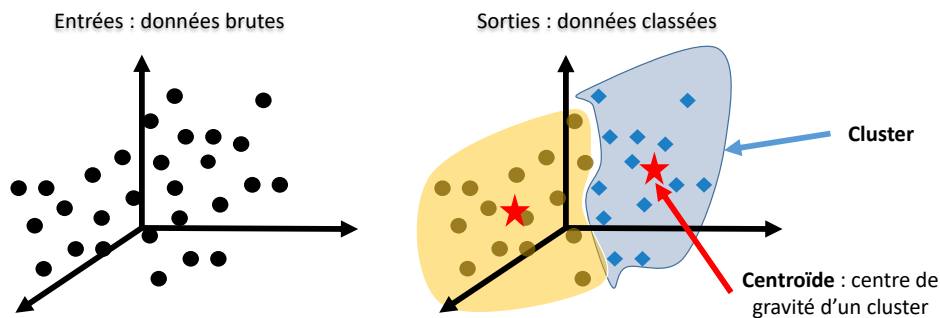


FIGURE 5.4 – Schéma explicatif du fonctionnement du *clustering*, de sa définition et de la définition du centroïde.

Les algorithmes de *clustering* peuvent être catégorisés en fonction de leur "modèle de *cluster*". La figure 5.5 montre les deux catégories usuelles. Les chercheurs utilisent différents modèles de *clusters*, et pour chacun de ces modèles de *clusters*, des algorithmes différents sont proposés. Tous ne fournissent pas des modèles pour leurs *clusters* et ne peuvent donc pas être facilement catégorisés. La notion de *cluster*, telle qu'elle est définie par différents algorithmes, varie considérablement dans ses propriétés. Nous définissons les trois catégories de méthodes de *clustering* : hiérarchiques, par partitionnement, basées sur la densité de probabilité.

5.3.2.1 *Clustering* hiérarchique

Le *clustering* hiérarchique est une méthode d'analyse de cluster qui cherche à construire une hiérarchie de *clusters* par récursivité. Il existe deux principales méthodes de *clustering* hiérarchique :

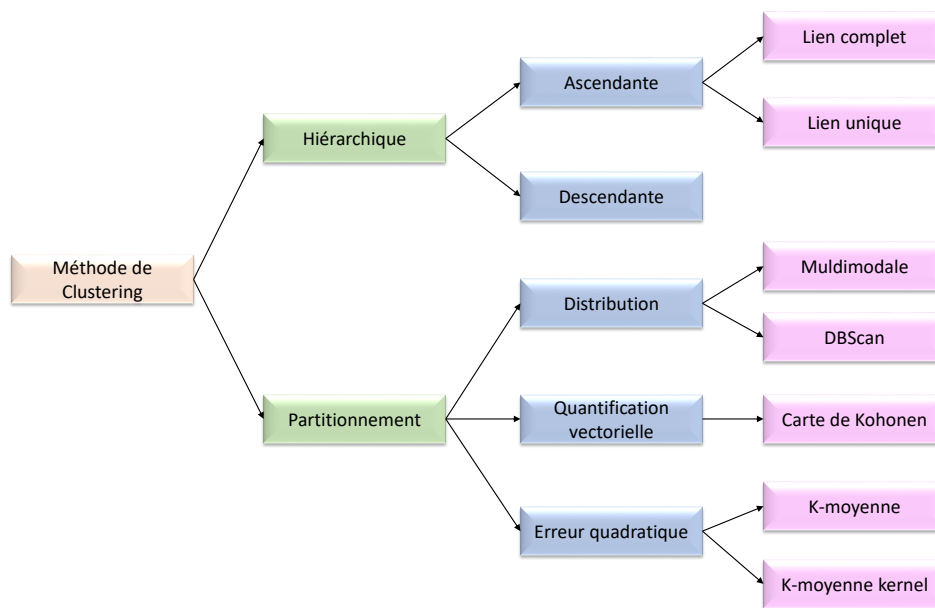


FIGURE 5.5 – Arbre des algorithmes usuels de *clustering*.

- Les méthodes ascendantes (ou agglomératives) : au début, chaque observation constitue un *cluster* puis des paires de *clusters* sont fusionnées et ainsi de suite quand la hiérarchie est remontée.
- Les méthodes descendantes (ou divisives) : toutes les observations commencent dans un *cluster* puis sont fractionnées de façon récurrente au fur et à mesure que l'on descend dans la hiérarchie.

Le résultat des méthodes hiérarchiques est un dendrogramme, représentant le regroupement imbriqué d'objets et les niveaux de similarité auxquels les groupes changent. Le résultat du *clustering* des objets est obtenu en coupant le dendrogramme au niveau de similarité désirée.

5.3.2.2 *Clustering* par partitionnement

Les méthodes de partitionnement sont des méthodes de *clustering* utilisées pour regrouper des données issues d'un ensemble en plusieurs *clusters* en fonction de leur similarité. Le fonctionnement de ces méthodes repose principalement sur le déplacement des centroïdes des *clusters*, d'un *cluster* à l'autre, à partir d'un positionnement initial. De telles méthodes exigent généralement que le nombre de *clusters* soit prédéfini par l'utilisateur [GOKILA et al., 2016]. **K-Means** ou **k-moyennes** est la méthode de *clustering* par partitionnement la plus populaire. C'est une méthode de quantification vectorielle, issue du traitement du signal. **k-moyennes** vise à partitionner n éléments en k *clusters* dans lesquelles chaque observation appartient au cluster ayant la moyenne la plus proche, servant de centroïde du cluster. **k-moyennes** vise à minimiser les variances à l'intérieur des *clusters* en utilisant la distance euclidienne pour comparer les éléments. Nous avons retenu cette méthode pour effectuer la plupart des *clustering* présentés dans ces travaux.

5.3.2.3 Clustering basé sur la densité de probabilité

Dans les méthodes de *clustering* liées à la densité de probabilité, les *clusters* sont définis comme des zones de plus forte densité que le reste de l'ensemble de données. Les éléments de ces zones clairsemées, qui sont nécessaires pour séparer les *clusters*, sont généralement considérés comme du bruit et des points frontaliers [ESTER et al., 1996]. L'algorithme **DBSCAN** pour *Density-Based Spatial Clustering of Application with Noise* [ESTER et al., 1996] et l'algorithme **Espérance-Maximisation** [DIFFEY et al., 2017] sont les plus utilisés.

5.3.3 Conclusion sur les méthodes d'apprentissage

Sont discutés ici les méthodes de réduction de dimension et de *clustering* qu'il est possible d'utiliser dans la quatrième étape de notre approche pour analyser et visualiser les résultats c'est-à-dire la matrice [*intrants* × *procédés* × *impacts*]. Nous retenons une méthode hybride qui consiste en une méthode de réduction puis en une méthode de *clustering*. Dans le paragraphe suivant, nous décrivons la méthode hybride. C'est une méthode linéaire par extraction de caractéristiques comme méthode de réduction de dimension et une méthode de *clustering* par partitionnement. Plus précisément c'est la méthode MDS et la méthode **k-moyennes**.

5.4 Positionnement multidimensionnel et k-moyennes

L'approche hybride est basée sur le positionnement multidimensionnel (MDS) et les **k-moyennes**. Nous avons utilisé la distance de Canberra pour construire la matrice de proximité, puis les données sont mises en correspondance avec une représentation spatiale de plus faible dimension (deux ou trois dimensions). Ces méthodes permettent de réduire l'espace dimensionnel (l'ensemble des variables) tout en préservant le maximum d'informations. L'objectif est de rechercher des structures "cachées" dans les données multidimensionnelles et d'aider à interpréter la zone groupée des points finaux de la matrice d'évaluation EICV. L'avantage de cette approche est que les méthodes basées sur les données nécessitent une connaissance minimale des procédés pour effectuer cette tâche.

Avant d'arriver à cette méthode hybride, plusieurs autres méthodes de réduction de dimension ont été testées mais aucune (sauf MDS) a fourni un *clustering* acceptable.

Le positionnement multidimensionnel (MDS) classique, appelé aussi *Principal Coordinates Analysis* (PCoA) est une méthode de visualisation des similarités (ou dissemblance) entre des objets dans un espace de dimension réduite. Elle est conçue pour permettre de comprendre les structures de proximités et d'opposition. Partant des informations sur les ressemblances mutuelles de n objets, souvent avec la matrice de similarité où $\Delta = \delta(i, j)_{0 \leq i \leq n; 0 \leq j \leq n}$, on cherche une configuration de n points (dans \mathbb{R}^2 en général) qui est telle que si l'objet i ressemble à l'objet j plus que l'objet l à l'objet k , on ait : $\delta(i, j) < \delta(l, k)$. Pour n'importe quelle matrice de distances de taille $n \times n$, la MDS permet de trouver, un ensemble de n points repérés par leurs coordonnées dont la matrice de similarité est égale ou très proche des données. Dans notre approche, nous utilisons la distance de Canberra [LANCE et

WILLIAMS, 1966], une version pondérée de la distance de Manhattan. Elle se calcule comme suit :

$$\Delta^{\text{Canberra}} : (x, y) \mapsto \sum_{i=1}^p \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad (5.3)$$

Soient \mathbf{x}_r ($r = 1, \dots, n$) les coordonnées de n points dans un espace euclidien de p dimensions données où $\mathbf{x}_r = (x_{r1}, x_{r2}, \dots, x_{rp})^\top$ et définissons $[\mathbf{B}]_{rs} = b_{rs} = \mathbf{x}_r^\top \mathbf{x}_s$. Pour une distance euclidienne $\Delta = \delta_{rs}$, à partir d'une matrice \mathbf{A} d'éléments $[\mathbf{A}]_{rs} = a_{rs} = -\frac{1}{2}\delta_{rs}^2$, déduite de la décomposition $\mathbf{x}_r^\top \mathbf{x}_s$, la matrice \mathbf{B} est obtenue en utilisant la relation suivante 5.4 :

$$\mathbf{B} = \mathbf{H}\mathbf{A}\mathbf{H} \quad (5.4)$$

où \mathbf{H} est la matrice centrage : $\mathbf{H} = \mathbf{I} - n^{-1}\mathbb{1} \cdot \mathbb{1}^\top$ avec $\mathbb{1} = (1, 1, \dots, n)^\top$. Les éléments de \mathbf{A} sont définis comme suit :

$$a_{r\bullet} = n^{-1} \sum_s a_{rs}, a_{\bullet s} = n^{-1} \sum_r a_{rs}, a_{\bullet\bullet} = n^{-1} \sum_r \sum_s a_{rs}.$$

La procédure algorithmique du positionnement multidimensionnel est résumée, comme illustré dans [COX et COX, 2001] par les étapes suivantes données dans l'algorithme 2.

Algorithme 2 : MDS – Positionnement Multidimensionnel

Params : \mathbf{x}_r ($r = 1, \dots, n$) : les coordonnées de n points dans un espace euclidien de p dimensions

Résultat : $x_{ri} = v_{ir}$ ($r = 1, \dots, n; i = 1, \dots, \ell$) : les coordonnées des n points dans ℓ

1. Obtenir la matrice de proximité $\Delta = \delta_{rs}$
 2. Trouver la matrice $\mathbf{A} = [-\frac{1}{2}\delta_{rs}^2]$
 3. Trouver la matrice $\mathbf{B} = [a_{rs} - a_{r\bullet} - a_{\bullet s} + a_{\bullet\bullet}]$
 4. Trouver les valeurs propres $\lambda_1, \lambda_2, \dots, \lambda_{n-1}$ et les vecteurs propres $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{n-1}$ si \mathbf{B} est semi-définie positive (certaines valeurs propres sont négatives),
 - (a) soit on ignore les valeurs négatives et on continue,
 - (b) soit on ajoute une constante appropriée c aux (dis)similitudes.
 5. Choisissez une taille appropriée ℓ , éventuellement en utilisant $\frac{\sum_1^\ell \lambda_i}{\sum(\text{positive eigenvalues})}$. Les coordonnées des n points dans l'espace de dimension euclidienne ℓ sont données par $x_{ri} = v_{ir}$ ($r = 1, \dots, n; i = 1, \dots, \ell$).
-

Une fois que cette réduction de dimension est effectuée, nous utilisons l'algorithme des **k-moyennes** pour trouver des *clusters* dans ces données. On cherche à partitionner nos n points trouvés à la fin de l'algorithme précédent en k ensembles $\mathbf{S} = \{\mathbf{S}_k\}_{1 \leq k \leq n}$ en minimisant la distance entre les points à l'intérieur de chaque partition :

$$\text{args}(\min \sum_{i=1}^k \sum_{j=1}^n \|x_j - \mu_i\|^2) \quad (5.5)$$

où μ_i est le barycentre des points dans \mathbf{S}_i .

L'algorithme classique des **k-moyennes** est donné dans l'algorithme 3. Précisons que la convergence peut être due à n nombre d'itérations fixé à l'avance, dans ce cas, l'algorithme effectuera les itérations et s'arrêtera, peu importe la forme de clusters composés ou une stabilisation des centroïdes.

Algorithme 3 : *k* – means – k-moyennes

1. Choisir k points représentant la position moyenne des partition m_1^1, \dots, m_k^1
 2. Répéter jusqu'à convergence les étapes suivantes :
 - (a) Affecter à chaque observation la partition la plus proche :
 $S_i^t = \{x_j \mid \|x_j - m_i^t\| \leq \|x_j - m_{i^*}^t\|, \forall i^* = 1, \dots, k\}$
 - (b) Mettre à jour la moyenne de chaque cluster : $m_i^{t+1} = \frac{1}{|S_i^t|} \sum_{x_j \in S_i^t} x_j$
-

5.5 Conclusion sur le chapitre 5

Ce chapitre a permis de détailler la quatrième étape du cadre méthodologique développée dans ce manuscrit. Cette étape comporte deux grandes sous-étapes : la réduction de la dimension et le *clustering*. Nous avons mis en place une méthode hybride contenant ces deux étapes avec les méthodes MDS et **k-moyennes**. À la fin de ce dernier algorithme, on se retrouve avec des *clusters* qui nous permettent de faire une analyse de nos résultats soit en regardant directement les valeurs de l'algorithme soit en les projetant sur des graphiques. Les graphiques peuvent être simples ou interactifs en utilisant par exemple des logiciels comme PowerBi®. Il est d'ailleurs possible de pratiquer cette étape de façon itérative : on trouve des premiers *clusters* à la première itération puis on relance l'étape pour un *cluster* particulier et ainsi de suite. Les graphiques interactifs permettraient alors de bien représenter cette répétition et de pouvoir prendre des décisions avec toutes les informations. On peut revenir jusqu'au procédé et aux impacts voire aux flux élémentaires et donc analyser d'où viennent les similarités ou les dissimilarités. La contribution de ce chapitre est de répondre à l'objectif 2 de nos travaux et donc de fournir des pistes de réflexion pour analyser les résultats des analyses de la durabilité. C'est en exploitant les méthodes d'apprentissage non supervisé, notamment les méthodes de réductions de dimension et de *clustering*, que nous avons pu répondre à cet objectif.

Chapitre 6

Cas d'étude : valorisation des déchets lignocellulosiques

*Tout homme s'enrichit quand abonde l'esprit.
(Luna Lovegood)*

J.K Rowling - *Harry Potter et l'Ordre du Phénix*, 2003

Au cours des travaux, la mise en application du cadre méthodologie est traduite par l'implémentation d'une preuve de concept. Ce chapitre (figure 6.1) présente un cas d'étude et l'implémentation d'un outil. Dans un premier temps, la biomasse et les bioraffineries, coeurs du cas d'étude sont présentés. Dans un second temps, le cadre méthodologique est déroulé. Finalement, l'outil est discuté avant de mettre en perspectives les limites de l'application.

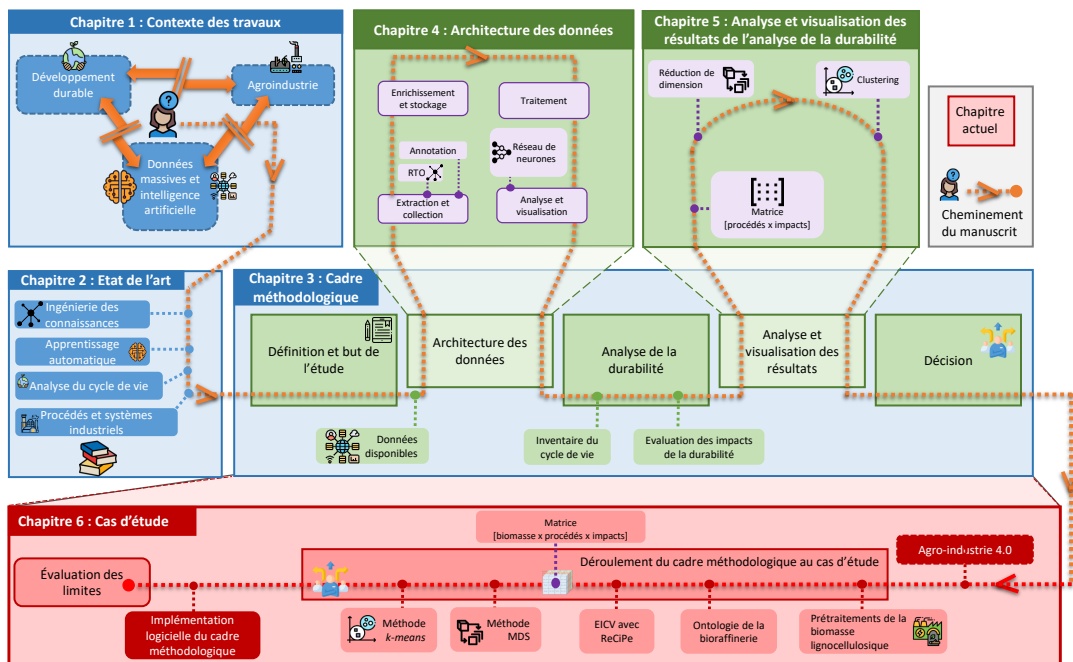


FIGURE 6.1 – Positionnement du chapitre 6 dans le manuscrit.

Sommaire

6.1 Introduction du chapitre 6	117
6.2 Industrie agricole 4.0	118
6.2.1 <i>Supply chain</i> agricole 4.0	118
6.2.2 Les données massives dans l'agro-industrie	118
6.2.3 L'intelligence artificielle dans l'agro-industrie	120
6.3 Présentation du cas d'étude	121
6.3.1 La biomasse	121
6.3.2 Les bioraffineries	123
6.4 Cadre méthodologique appliqué au cas d'étude	124
6.4.1 Étape 1 : But et frontière de l'étude	124
6.4.2 Étape 2 : Architecture des données	126
6.4.2.1 Collection et extraction des données	126
6.4.2.2 Enrichissement et stockage des données extraites	129
6.4.2.3 Traitement des données extraites	130
6.4.2.4 Analyses et visualisation des données	131
6.4.3 Étape 3 : Évaluation de la durabilité	134
6.4.3.1 Inventaire du cycle de vie	134
6.4.3.2 Évaluation environnementale du cycle de vie	135
6.4.4 Étape 4 : Analyse et visualisation des résultats	138
6.4.4.1 Analyse de la dispersion des impacts	140
6.4.4.2 Analyse de la dispersion des procédés	140
6.4.4.3 Conclusions de la partie analyse et visualisation	142
6.5 Création d'un outil-prototype	142
6.6 Limites du cas d'étude	147
6.7 Conclusion du chapitre 6	147

6.1 Introduction du chapitre 6

La population mondiale génère deux milliards de tonnes de déchets agricoles par an [MILLATI et al., 2019] et l'accroissement estimé de la population augmentera cette production de déchets. La quantité de terres dédiée à l'agriculture diminue avec l'augmentation des activités humaines entraînant inévitablement des répercussions sur les systèmes agricoles. [GARNETT et al., 2013] estiment que la meilleure approche pour l'avenir de l'agriculture est celle des nouvelles technologies agricoles supportant une intensification durable. Néanmoins, cette intensification entraînera un plus grand gaspillage de produits et de ressources [WEST et al., 2014]. Deux catégories de déchets peuvent être identifiées : les déchets provenant des intrants, tels que les engrais ou l'eau, et les déchets issus des procédés de traitement. Ces derniers proviennent de la conversion incomplète de la biomasse ou de la transformation des matériaux dans la chaîne d'approvisionnement qui va de la production agricole à la consommation alimentaire et qui est principalement composée de sous-produits lignocellulosiques.

La biomasse lignocellulosique est l'une des ressources renouvelables les plus abondantes et les moins onéreuses. La production de biomatériaux, de biomolécules et de bioénergie est basée sur la bioconversion de la biomasse lignocellulosique, qui implique l'hydrolyse enzymatique de la biomasse pour libérer du glucose. Cependant, une ressource renouvelable n'implique pas que les procédés de transformation de cette biomasse soient soutenables. C'est pourquoi de plus en plus de procédés agroalimentaires intègrent des analyses de la durabilité [COMMISSION, 2012; RAYMOND, 2012]. Les procédés de transformation de la biomasse impliquent des prétraitements pour générer de bons rendements en glucose. Au cours des trente dernières années, de nombreux procédés de prétraitement et leurs technologies associées ont été étudiés et publiés [DAVIS et al., 2017]. Divers facteurs ont été exploités pour comparer les performances, l'efficacité ou les impacts environnementaux du prétraitement [JOGLEKAR et al., 2019]. Les facteurs environnementaux, la consommation d'énergie et l'efficacité énergétique peuvent être considérés comme des facteurs classiques [BARAKAT et al., 2013; CHUETOR et al., 2015; ZHU et al., 2010]. Malgré tout, il y a un manque certain de critères pour guider le choix et différencier ces procédés.

Les chapitres précédents ont présenté les concepts, méthodes et le cadre méthodologique développés dans ces travaux et ont mis en exergue leurs potentialités. Ce chapitre illustre le cadre méthodologique développé pour une analyse des différentes voies technologiques des procédés de prétraitement et des différentes biomasses dans un contexte d'EC. La première partie aborde les méthodes d'IA et de données massives dans l'agro-industrie. La deuxième partie présente notre cas d'étude et la troisième partie l'exécution de notre cadre méthodologique pour ce cas. La quatrième et la dernière partie présentent respectivement le développement d'un prototype logiciel et les limites et perspectives de ce cas.

6.2 Industrie agricole 4.0

6.2.1 Supply chain agricole 4.0

La chaîne d'approvisionnement est un élément indispensable dans l'agriculture. À l'ère des données, une amélioration peut être apportée grâce à la *Data Science*. Avec la digitalisation de la chaîne d'approvisionnement, le suivi de l'état des flux de matières en temps réel devient possible. Cela permet aux entreprises et aux agriculteurs de détecter et de s'adapter aux risques, et de prendre une décision en temps réel ou des plans prospectifs. La résilience de la chaîne d'approvisionnement est améliorée. Il est complexe de parvenir à des chaînes d'approvisionnement agroalimentaires robustes, résilientes et durables, car elles sont confrontées à davantage de sources d'incertitude et de risques en comparaison aux autres chaînes d'approvisionnement, suscitant des questions et préoccupations quant à leurs performances durables [LEZOCHÉ et al., 2020]. Avant la consommation du produit final, la chaîne d'approvisionnement agricole comprend plusieurs opérations principales : la production, le stockage, la transformation et la distribution [BORODIN et al., 2016]. Les problèmes qui surviennent dans l'une ou l'autre de ces phases peuvent entraîner de graves difficultés. Afin de faire face à ces problèmes, quatre types d'incertitudes sont identifiés : le produit (durée de conservation, taux de détérioration, manque d'homogénéité, qualité des aliments et sécurité alimentaire), les processus (rendement de la récolte, délai d'approvisionnement, besoins en ressources, production), le marché (demande, valeur marchande) et l'environnement (conditions météorologiques, parasites, maladies et réglementations) [ESTESO et al., 2018; LEZOCHÉ et al., 2020]. Des conséquences indésirables surviennent lorsque ces incertitudes sont mal gérées. Une gestion plus précise de ces incertitudes procure une chaîne d'approvisionnement plus stable et une agriculture durable.

Afin de contrôler de manière appropriée la connaissance incomplète des données, plusieurs techniques de recherche opérationnelle sont mises en œuvre pour aider les difficultés de prise de décision dans le secteur agricole : la programmation stochastique, la programmation robuste et la programmation par simulation [BORODIN et al., 2016].

6.2.2 Les données massives dans l'agro-industrie

Comme les technologies de l'information s'introduisent progressivement dans le domaine agricole, les données massives représentent un nouveau facteur de production qui génère une nouvelle réalité dans l'agriculture pour faire face à toutes sortes de difficultés et améliorer l'efficacité de la production agricole. La première difficulté des données massives est de surmonter l'énorme quantité de données multidimensionnelles générées chaque année par l'industrie agricole. Ces données comprennent les données générales spatio-temporelles, les paramètres agricoles, les attributs environnementaux et les attributs géographiques [HIRA et DESHPANDE, 2015]. Analyser correctement ces données et générer des informations utiles est une tâche plutôt difficile. Des idées et des technologies novatrices sont alors nécessaires. [HIRA et DESHPANDE, 2015] proposent d'appliquer la MDS, l'analyse statistique et les techniques d'exploration de données pour analyser les données. Ils effectuent une analyse pour juger des différents impacts des paramètres agricoles sur la productivité agricole. Dans le même ordre d'idées, [SABARINA et PRIYA, 2015] proposent un modèle de réduction des caractéristiques basé sur l'algorithme IHOSVD (*Incremental High-order*

Singular Value Decomposition) pour décomposer puis extraire les données essentielles à partir de la grande quantité de données recueillies dans l'agriculture.

La disponibilité de données agricoles de qualité est une condition préalable à la réussite de l'analyse des données massives. Un système de transfert de données agricoles automatisé et intégré est nécessaire pour les transferts entre différentes organisations tout en assurant la confidentialité et la sécurité de ces données. À cette fin, [ISLAM SARKER et al., 2019] suggèrent un modèle conceptuel pour l'agriculture intelligente, qui comprend sept étapes : source de données, implication technologique, unité de traitement, paramètres institutionnels, processus de gouvernance, indicateurs et résultats. Les données massives sont mises en oeuvre avec de multiples autres technologies pour réaliser des améliorations dans divers domaines agricoles. Si nous divisons l'industrie agricole en trois niveaux d'activité (stratégique, opérationnel et tactique), nous constatons que la technologie issue des données massives joue un rôle important à chaque niveau.

Au niveau stratégique, les données massives sont utilisées pour soutenir les systèmes de prise de décision en aidant les agriculteurs à faire des choix plus rentables et durables. [ZHAO et GUO, 2018] créent un système de décision intelligent agricole basé sur les données massives.

Au niveau tactique, les données massives se présentent avec des objectifs spécifiques tels que la prévision des risques météorologiques, la prévision des rendements et l'analyse de la durabilité de l'agriculture. [BENDRE et al., 2015] proposent un modèle de prédiction météorologique qui utilise *MapReduce* pour le traitement des données massives et la régression linéaire pour la prédiction des données. [CHARVAT et al., 2018] présentent leur utilisation de techniques de visualisation avancées dans un projet de bioéconomie piloté par les données. Ils appliquent des techniques de visualisation et une analyse issue des données massives pour identifier la zone productive de rendement à partir de vastes zones. Pour cela ils calculent la productivité du rendement de différentes zones à l'aide des indices de végétation collectés par satellite et des données de rendement des cultures enregistrées lors de la récolte précédente.

Au niveau opérationnel, les technologies issues des données massives peuvent être utilisées pour promouvoir l'agriculture de précision et surveiller la croissance des cultures. L'irrigation et la fertilisation sont responsables d'une énorme proportion de la pollution environnementale générée par les processus agricoles. Par conséquent, l'agriculture de précision est promue pour irriguer et fertiliser les cultures à différents endroits en fonction de leurs caractéristiques spécifiques et de leur variabilité spatiale. [ZHANG et al., 2017] construisent un système d'irrigation intelligent basé sur les données massives.

Des améliorations peuvent également être apportées aux données massives elles-mêmes tout en les mettant en oeuvre dans l'agriculture. Les données massives, malgré leur potentiel à changer révolutionnairement l'agriculture, sont susceptibles de produire des risques en raison du déséquilibre de pouvoir entre les petites et les grandes exploitations. Les grandes exploitations, qui s'appuient sur leurs bases plus importantes de sources de données et de capitaux, sont plus susceptibles de générer plus de bénéfices à partir des données massives, créant ainsi un environnement défavorable au développement des autres petites exploitations [LIOUTAS et CHARATSARI, 2020]. Certaines propositions comme une architecture orientée services cloud pour l'agriculture qui peut être utilisée pour

résoudre le problème de stockage des données et de capacité de calcul et promouvoir le partage des informations agricoles sont étudiées [SRINIVASULU et al., 2017].

6.2.3 L'intelligence artificielle dans l'agro-industrie

L'IA augmente l'efficacité de la production agricole. [SUN et al., 2018] appliquent un système d'IA (vision par ordinateur) pour identifier la qualité des longes de porc en analysant leur couleur et leur marbrage à grande vitesse. [YOSSY et al., 2017] combinent le réseau neuronal artificiel et le système de vision par ordinateur pour déterminer la maturité des mangues en fonction de leurs couleurs. [STREICH et al., 2020] voient un bel avenir pour les systèmes d'IA dans le domaine de la bio-ingénierie. [ELAHI et al., 2019] emploient une méthode de réseau neuronal artificiel pour estimer le gaspillage de produits agrochimiques, l'économie des ressources agricoles et le potentiel de réduction de l'utilisation des pesticides tout en maintenant le rendement.

L'IA aide les agriculteurs et les entreprises agro-industrielles à prendre de meilleures décisions et à accroître les avantages économiques. [CHARANIA et LI, 2020] développent une application d'agriculture intelligente pour aider les agriculteurs à prendre des décisions et à gérer leurs exploitations. Il s'agit d'un logiciel en ligne doté d'algorithmes d'analyse prédictive et d'apprentissage automatique. Les instruments utilisés comprennent des caméras multispectrales montées sur drone, des images satellites et des capteurs. Par exemple, ils utilisent des moniteurs au sol et des images satellites pour collecter des données sur les conditions météorologiques, la santé des cultures et la qualité des sols, afin de prendre des décisions fondées sur des données concernant la planification, la culture et la récolte des cultures. Ils sont convaincus que ces technologies s'avéreront très utiles pour faire face aux impacts environnementaux potentiels des changements climatiques. L'IA fait des prévisions liées à l'agriculture, comme la saison des sécheresses ou la dégradation des aliments. [VERDOUW et al., 2014] proposent une conception basée sur l'Internet du futur pour des systèmes d'information logistiques agroalimentaires intelligents comprenant un système expert pour prédire des résultats tels que la prévision de la dégradation de la qualité et pour fournir des recommandations et des résultats respectifs. De nombreuses recherches sont également menées pour améliorer l'efficacité et la précision des techniques d'IA utilisées en agriculture. [PATRÍCIO et RIEDER, 2018] mènent une enquête concernant l'utilisation graphique de l'IA dans les études agricoles.

L'agriculture se veut de plus en plus durable et digitale. D'une part sa digitalisation est due à l'arrivée de nombreuses méthodes pour aider l'agriculture en termes de rendement puis en général. D'autre part, l'agriculture et les industries agricoles sont novatrices en termes de durabilité comme le montre le concept de bioéconomie. Dans un contexte d'EC, l'agro-industrie recycle ses déchets en les transformant en biocarburant, biomolécule ou biomatériau. Le cas d'étude choisi se voulait entrant dans ce contexte d'EC et l'agroindustrie s'est imposé comme un bon candidat. Pour rester dans ce contexte, les procédés de transformation des déchets agricoles en produit de valorisation doivent être eux-mêmes le plus durables possible. Pour valoriser ces déchets, les prétraitements sont indispensables et portent un pourcentage important des impacts. Notre choix de cas d'étude s'est donc porté sur la comparaison de ces prétraitements et des biomasses qu'on y traitera.

6.3 Présentation du cas d'étude

Le projet 3BCar de l'institut Carnot IC2AdCV impliquant une équipe de l'IATE de Montpellier, du Laboratoire de Chimie Agro-industrielle et de notre laboratoire a étudié les procédés de prétraitement des déchets lignocellulosiques. Le cas d'étude s'appuie sur ce projet. Des données et un outil (@Web) sont réutilisées et complétées. Les déchets lignocellulosiques agricoles sont aussi appelés biomasse lignocellulosique. Avant d'exécuter notre cadre méthodologique sur ce cas d'étude, la biomasse, le choix de la biomasse lignocellulosique et les bioraffineries sont succinctement présentés.

6.3.1 La biomasse

D'après le ministère de l'Écologie, du Développement durable et de l'Énergie, la biomasse est la « fraction biodégradable des produits, des déchets et des résidus d'origine biologique provenant de l'agriculture, y compris les substances végétales et animales, de la sylviculture et des industries connexes, y compris la pêche et l'aquaculture, ainsi que la fraction biodégradable des déchets, notamment les déchets industriels ainsi que les déchets ménagers et assimilés lorsqu'ils sont d'origine biologique» [MINISTÈRE DE L'ÉCOLOGIE, 2021]. Quatre grands type de biomasse sont distingués et sont représentés dans le tableau 6.1.

TABLEAU 6.1 – Classement de la biomasse et sa génération adapté de [PTASINSKI, 2016].)

Génération	Type de biomasse		Exemples
Première	cultures alimentaires et fourragères	cultures sucrières	cannes à sucre, betterave à sucre, sorgho doux
		plantes oléagineuses	maïs, blé, pomme de terre, seigle, manioc, triticale
		plantes amylacées	soja, palmier à huile, graine de coton, arachide, tournesol
		fourrage	luzerne, graminée
Deuxième	cultures ligno-cellulosiques	ligneux	cultures ligneuses à courte rotation : saule, peuplier, acacia, eucalyptus
		herbacée	miscanthus, Phalaris faux roseau, panic érigé
Troisième	cultures aquatiques	micro-algue	chlorella, spiruline
		macro-algue	algue marine, laminaire
		plantes aquatiques	jacinthe d'eau, végétaux de marais salant
Déchet	naturel	agricole	résidus d'origine végétale et animale, résidus de terrain
		forestier	Résidus de coupe et d'entretien des forêts
	artificiel	municipales	déchets solides municipaux, boues d'épuration, bois de démolition, déchets alimentaires
		industriel	liqueur noire de l'industrie de la pâte de papier, boues

Le choix de ces travaux a été de se restreindre à la biomasse lignocellulosique. Comme toute ma-

tière organique, la composition chimique élémentaire de la biomasse lignocellulosique consiste principalement de carbone, d'hydrogène, d'oxygène et en quantités mineurs d'azote. La composition élémentaire massique varie peu selon le type de biomasse, mais elle se présente généralement comme suit : 50% de carbone, 6% d'hydrogène, 44% d'oxygène et entre 0,1 et 0,5% d'azote [TSOUMIS, 1968]. La biomasse lignocellulosique se compose de 3 polymères naturels, la cellulose, les hémicelluloses et la lignine (Figure 6.2). Elle comporte aussi des fractions minoritaires : sucres solubles, pectines, protéines et minéraux. La composition des trois majeures fractions varie selon les espèces, la maturation de la plante, les tissus et les organes. Ces 3 polymères sont liés avec des liaisons covalentes et non covalentes. En général, les fractions sont les suivantes [MCKENDRY, 2002] :

- la cellulose : 30-50%
- les hémicelluloses : 20-40%
- la lignine : 15-25% .

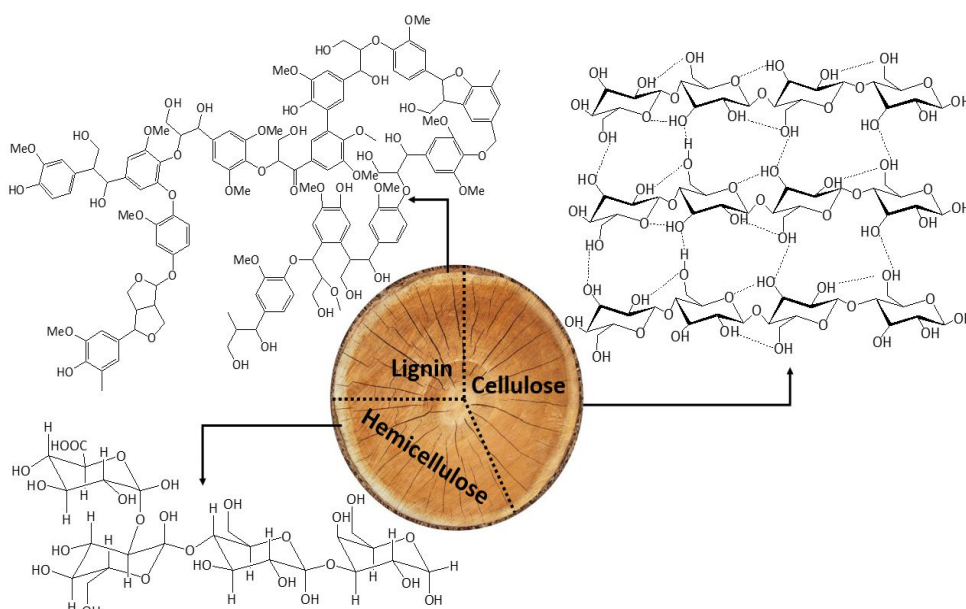


FIGURE 6.2 – Présentation de la composition globale de la biomasse lignocellulosique issu de [ALPER et al., 2020]

La cellulose est un homopolymère polysaccharide linéaire de sous-unités de D-glucose composées d'unité de cellobiose reliées par des liaisons glycosidiques β -(1→4). Le motif de répétition de cette structure est appelé cellobiose [PEREZ et MAZEAU, 2005]. Les interactions entre les chaînes de celluloses confèrent une structure fibrillaire. Ces microfibrilles sont attachées les uns aux autres par des hémicelluloses et encrassées par la lignine. L'organisation de ces fibrilles peut prendre deux formes une partie majeure cristalline et l'autre amorphe. La partie amorphe de la cellulose est plus sensible à la dégradation enzymatique et la partie cristalline est plus résistante.

Contrairement à la cellulose, les hémicelluloses sont des hétéropolymères polysaccharides qui ont une structure ramifiée et amorphe. Ils sont présents dans la quasi-totalité des parois de cellules végétales [BAJPAI, 2016]). La composition et la structure des hémicelluloses des feuillus et des résineux sont différentes. De plus, il existe des différences considérables en teneur et en composition des

hémicelluloses entre la tige, les branches, les racines et l'écorce [SJOSTROM, 1993]. Les hémicelluloses ont des ramifications avec de courtes chaînes latérales qui se composent de différents monosaccharides de type xylose, mannose, galactose et arabinose. Les hémicelluloses peuvent être facilement hydrolysés par des acides dilués ou par des bases ainsi que par de nombreuses enzymes hémicellulases (xylanase).

La lignine est le polymère le plus abondant après la cellulose et l'hémicellulose, présent dans les parois des cellules végétales, elle confère le support structural, l'imperméabilité et la résistance aux attaques microbiennes. La lignine est un hétéropolymère aromatique tridimensionnel totalement amorphe et composé d'unités monomériques de phénylpropane [SHEN et al., 2010]. La composition en lignine varie selon les espèces et le stade de maturation de la plante. La lignine est insoluble dans l'eau et optiquement inactive ce qui rend sa dégradation très difficile. Elle limite la biodégradabilité des polysaccharides au sein de la matrice lignocellulosique.

6.3.2 Les bioraffineries

La bioraffinerie est définie par l'Agence Internationale de l'Énergie (*International Energy Agency – IEA*) comme un processus qui transforme les produits agricoles afin de les valoriser en énergie, par exemple la biomasse est transformée en biocarburant (Figure 6.3). L'étape de prétraitement a pour objectif de rendre la cellulose plus accessible aux enzymes qui vont la transformer en glucose. Comme expliqué précédemment la cellulose est entourée par l'hémicellulose et la lignine, comme première étape on vise à les éliminer de la matrice. La taille des particules, la teneur en lignine, la cristallinité de la cellulose, la composition en hémicelluloses, la porosité et la surface accessible sont toutes des caractéristiques physiques qui peuvent influencer l'action enzymatique. En traitant la biomasse avant l'hydrolyse enzymatique, on cherche à contrôler ces différentes caractéristiques pour optimiser l'hydrolyse. Il existe de nombreux types de prétraitements, qui peuvent être classés en différentes catégories : physiques, chimiques, physico-chimiques et biologiques.

Les raffineries s'inscrivent dans l'EC si et seulement si les procédés utilisés et en particulier les procédés de prétraitements sont durables. C'est l'objectif de cette étude : aider les chercheurs ou les ingénieurs R&D à classer ou choisir un procédé et un type de biomasse selon des critères de durabilité.

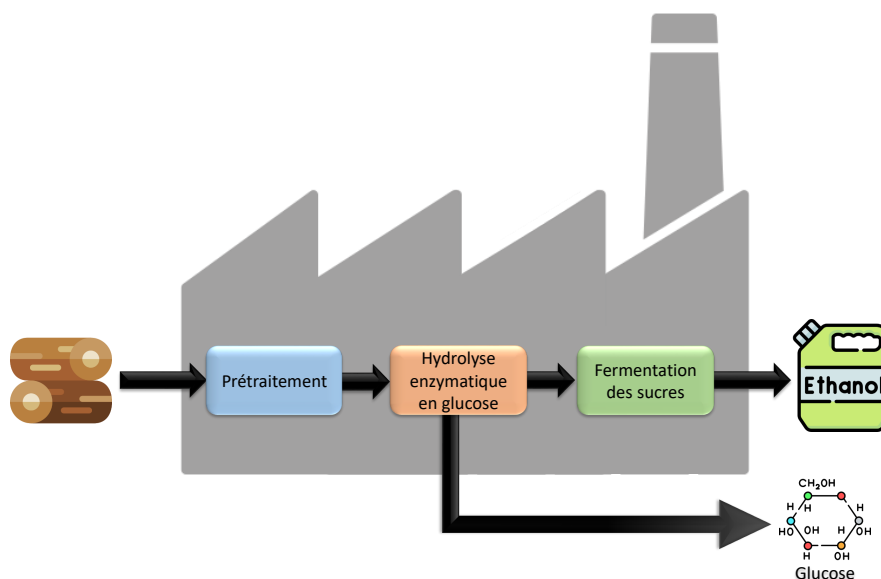


FIGURE 6.3 – Schéma général d'une bioraffinerie de production de glucose ou d'éthanol à partir de la biomasse lignocellulosique

6.4 Cadre méthodologique appliqué au cas d'étude

Cette section détaille le cadre méthodologique (figure 6.4) appliqué aux prétraitements de la biomasse lignocellulosique pour une production de glucose. Les étapes utilisent les méthodes et outils décrits dans les chapitres précédents (3, 4 et 5).

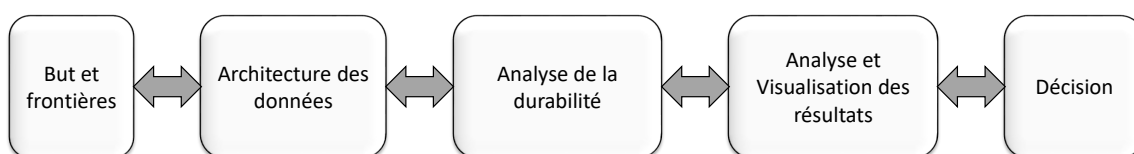


FIGURE 6.4 – Cadre méthodologique général.

6.4.1 Étape 1 : But et frontière de l'étude

La première étape de notre cadre méthodologique spécifie le but et les frontières du cas d'étude. L'objectif général de ce cas d'étude est :

1. de soutenir la sélection d'un enchaînement d'opérations unitaires de prétraitement ou d'une biomasse grâce à des critères environnementaux;
2. de tirer bénéfice des données massives pour accroître la connaissance et réduire le nombre d'expériences à développer en exploitant celles des chercheurs du monde entier via les articles scientifiques;
3. de diminuer le temps d'analyse des articles et des résultats de la durabilité.

La fonction du cas d'étude est la sélection d'un (ou un groupe d') enchaînement d'opérations unitaires de prétraitement et d'une biomasse pour la réalisation d'un prototype de laboratoire de production de glucose à partir de biomasse lignocellulosique. Le système est représenté dans la figure 6.5. On se retrouve dans une approche de la « porte-à-la-porte » si on considère tout le cycle de vie. La biomasse lignocellulosique étant un déchet agricole, tous les impacts environnementaux et économiques (le déchet est gratuit) sont sur le produit final de l'agriculture, nous nous retrouvons alors avec une approche « du berceau à la porte » pour la production de glucose à partir des déchets lignocellulosiques.

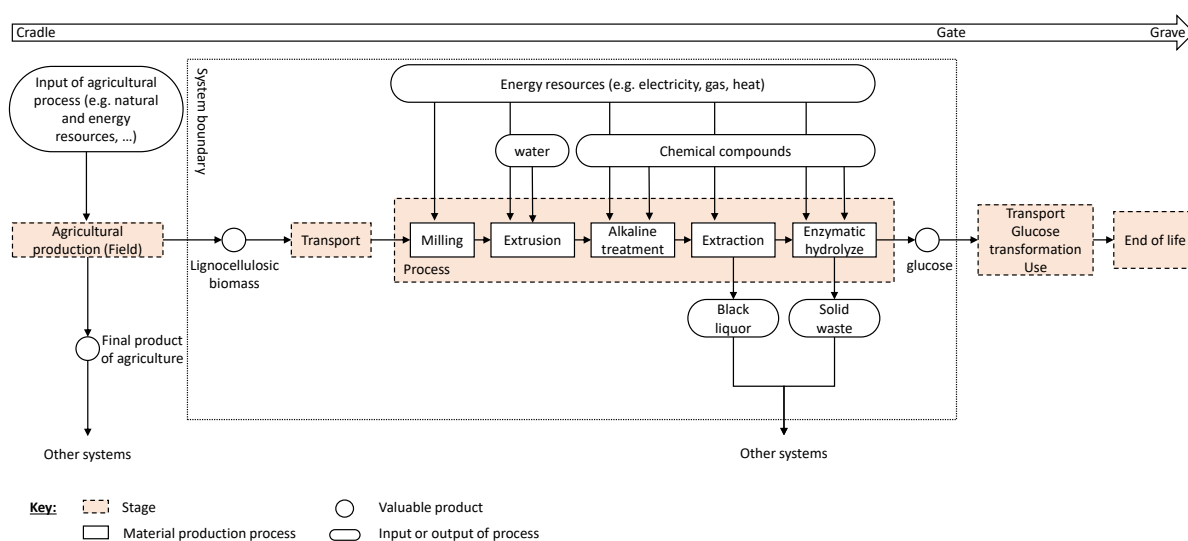


FIGURE 6.5 – Schéma des frontières du système pour le cycle de vie d'une bioraffinerie ayant pour entrée la biomasse lignocellulosique [PRIOUX et al., 2021]

Les données relatives aux procédés en tant qu'intrants se trouvent dans des documents scientifiques et le résultat final est un classement de ces procédés sur la base d'indicateurs de la durabilité. Il existe différents types de procédés de prétraitement qui ont chacun leur propre particularité. Dans la suite, nous avons classé les procédés de prétraitement suivant différentes catégories qui correspondent aux différentes opérations unitaires qui se succèdent (Figure 6.6). La production des autres intrants est incluse. La bioraffinerie est considérée relativement à proximité du champ, l'étape de transport étant alors négligeable. L'énergie nécessaire au procédé est un paramètre clé de l'ACV, mais les données relatives à ce paramètre sont presque toujours manquantes. Cela peut s'expliquer par l'échelle laboratoire des études sélectionnées et par le type d'article dans lequel les procédés sont décrits. Des efforts sont faits pour compenser ce problème en estimant l'énergie de broyage par l'application de modèles statistiques basés sur la régression classique à des données complémentaires extraites d'études publiées [BARAKAT et al., 2013]. Les sorties solides et liquides sont considérées comme n'ayant aucun impact, car elles contiennent des molécules qui peuvent être valorisées, et ne sont donc pas gaspillées. L'unité fonctionnelle est la "production de 1 g de glucose". Tous les résultats sont exprimés en fonction de cette unité fonctionnelle. Nous supposons aussi que la bioraffinerie est en France et qu'elle est proche du champ. Cette hypothèse nous permet de comparer les prétraitements entre eux. De plus, l'hypothèse de la France impose que l'électricité soit le mix énergétique

français et que les intrants soient soit français soit européens.

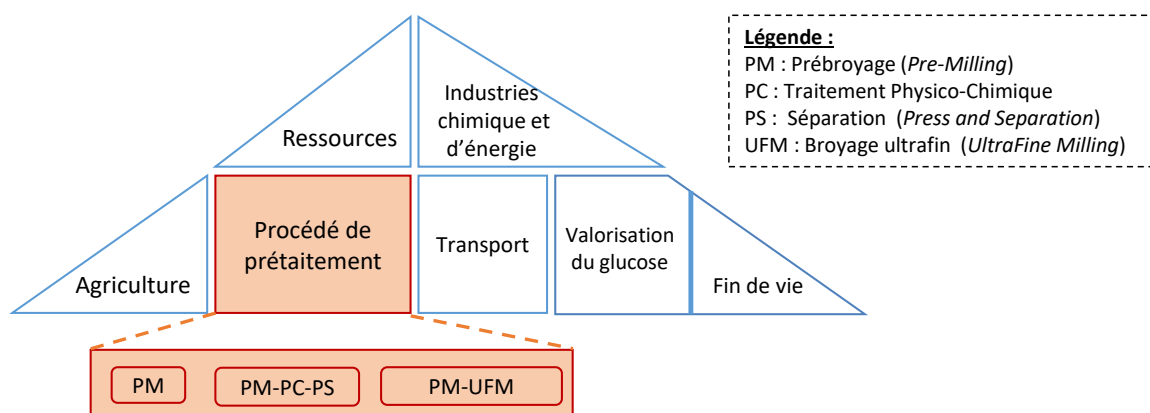


FIGURE 6.6 – Étapes du cycle de vie du glucose avec le détail des différents procédés de prétraitement étudiés.

Pour effectuer les analyses de la durabilité nécessaires à la sélection de l'enchaînement des opérations unitaires ou de la biomasse, deux types de données sont nécessaires :

- les données d'arrière-plan contenant les processus d'arrière-plan de fabrications des intrants de nos procédés,
- les données de premier-plan qu'on appellera aussi données procédés contenant toutes les données relatives aux procédés étudiés.

6.4.2 Étape 2 : Architecture des données

Les données ACV issues de la base EcoInvent [INVENTORY, 2022] pour l'inventaire des données d'arrière-plan sont extraites directement. Les bases de données EcoInvent possèdent les données d'arrière-plan et aussi les incertitudes relatives à ces données.

6.4.2.1 Collection et extraction des données

Pour les données des procédés se trouvant dans les articles scientifiques, nous utilisons le *framework* développé dans le paragraphe 4.2 et notamment la RTO définie.

La première activité consiste à sélectionner les documents décrivant les procédés de prétraitement de la biomasse lignocellulosique. Nous avons identifié tous les articles publiés correspondant à des mots clés : « biomasse lignocellulosique », « prétraitement », « glucose » et « hydrolyse enzymatique » dans les bases de données scientifiques (en langue anglaise), telles que Web of Science ou Science Direct. Il a été identifié 57 articles scientifiques pertinents répartis entre 6 types de procédés décrits dans la figure 6.7 et 9 biomasses (tableau 6.2). La liste complète des articles sélectionnés se trouve en Annexe C.

TABLEAU 6.2 – Classement de la biomasse se la génération adapté de [PTASINSKI, 2016]

Biomasse (terme anglais)	Biomasse (terme français)	Description
<i>Bagasse</i>	Bagasse	Résidu fibreux de la canne à sucre qu'on a passé par le moulin pour en extraire le suc
<i>Birch chips</i>	Copeaux de bouleau	-
<i>Corn stover</i>	Fane (ou paille) de maïs	Résidus de la culture de maïs (feuilles + tiges + épis) abandonnés après la récolte des grains
<i>Rice straw</i>	Paille de riz	-
<i>Spruce chips</i>	Copeaux d'Épicéa	-
<i>Spruce saw dust</i>	Sciure d'Épicéa	-
<i>Sugarcane leaf</i>	Feuille de canne à sucre	-
<i>Switchgrass</i>	Panic érigé ou millet vivace (herbe à sanglier)	Espèce de plantes herbacées originaire d'Amérique du Nord
<i>Wheat straw</i>	Paille de blé	-

Les acronymes utilisés pour les types d'enchaînements d'opérations unitaires qu'on appelle par abus de langage procédés viennent des noms des étapes en anglais :

- PM : *Pre-Milling*
- UFM : *UltraFine Milling*
- PS : *Press and Separation*
- PC : *Physico-Chemical treatment*
- Ex : *Extrusion*

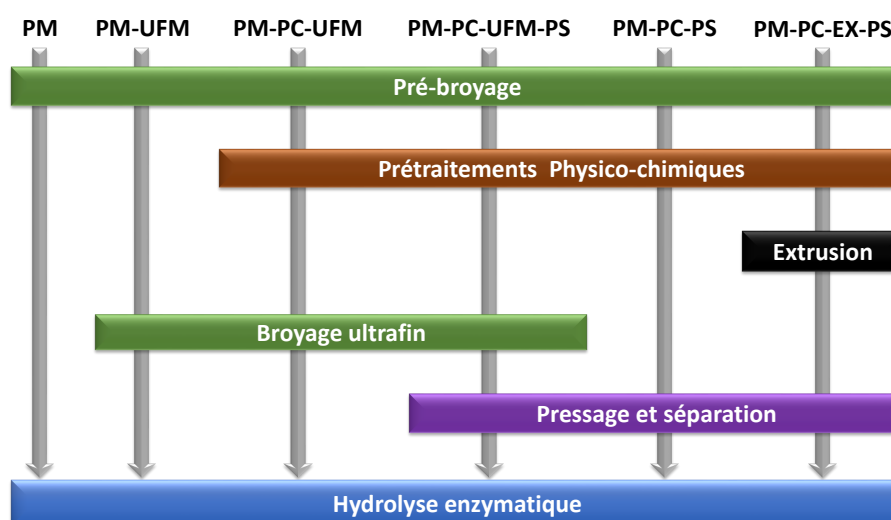


FIGURE 6.7 – Descriptif des six types d'enchaînement d'opérations unitaires.

La deuxième sous-étape est la création de la ressource terminologique et ontologique (RTO) appelée *Biorefinery TOR* et décrite plus précisément dans [BREYSSE et al., 2017]. La partie conceptuelle de la RTO est constituée d'une ontologie cœur - contenant les relations *n-aire* et leurs arguments,

la gestion des quantités et leurs unités associées - et d'une ontologie du domaine représentant les concepts spécifiques à la bioraffinerie. La partie terminologique de la RTO contient l'ensemble des termes liés au domaine utilisé pour annoter les tableaux de données. Ces relations sont caractérisées par plusieurs types de paramètres : paramètres d'entrée, de sortie et de contrôle. Un exemple de relation n-aire est fourni dans la figure 6.10. La RTO a été implémentée dans l'outil **@Web** [IATE, 2022] développé par l'INRAE ICO IATE. **@Web** (pour *Annotated Tables from the web*) est une plateforme collaborative partageant des documents avec des tableaux annotés [LOUSTEAU-CAZALET et al., 2016]. Elle permet de créer la RTO et d'effectuer les sous-étapes d'extraction des tableaux, d'annotation et de stockage dans une base RDF que nous avons explicité dans le chapitre 4 (figure 4.5). La RTO est détaillée dans l'annexe B.

Prenons par exemple l'article de [LIU et al., 2013] dont le procédé est présenté dans la figure 6.8.

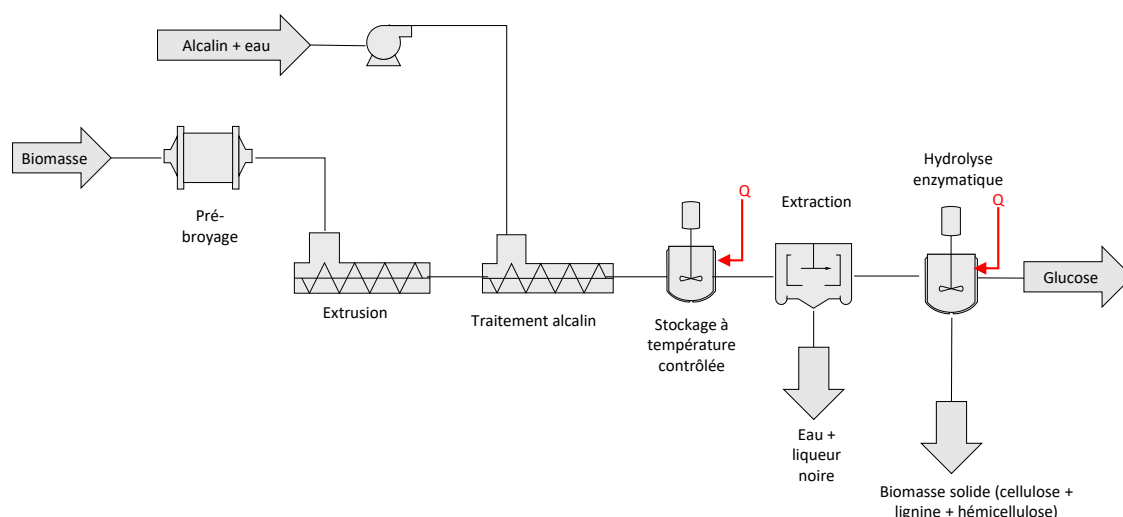


FIGURE 6.8 – Procédé de prétraitement issu de l'article scientifique de [LIU et al., 2013]

Les tableaux contenus dans l'article sont extraits automatiquement par **@Web** puis validés par expertise. Ceci enrichit la RTO. Par exemple, le tableau 3 de l'article intitulé « *Chemical composition of corn stover* » permet d'enrichir la relation *biomass_composition* de l'ontologie et d'extraire ainsi toutes les données de la composition de la tige de maïs. Une fois tous les tableaux traités, le chercheur parcourt le document en l'annotant grâce à l'ontologie qui fournit une grille de lecture. La relation n-aire *Quantité de constituant solide issue d'un broyage* - présentée dans la figure 6.10 - requiert plusieurs arguments d'entrée (quantité des intrants, type des intrants, type de broyage), de contrôle (température, pression, durée) et de sortie (la quantité de solide en sortie). Une fois les données extraites du document, elles sont visibles sous forme de trois tableaux : composition de la biomasse, cocktail d'enzymes et description du procédé. On présente un extrait du tableau « description du procédé » de l'article de Liu dans la figure 6.9. Toutes les données de premier plan nécessaires pour l'établissement de l'inventaire du cycle de vie sont fournies dans le tableau de description des procédés.

Process Description (table 1) [en](#) [export](#) / [view info](#)

n°	Output solid constituent size Unit : cm	Treatment	Experience number Unit : 1	Process step number Unit : 1	Biomass	Biomass quantity Unit : g	Total pretreatment energy Unit : kW.h.kg-1	Water quantity Unit : l	Rotation speed Unit : min-1	Treatment duration Unit : s	Output solid constituent quantity Unit : g	Salt	Salt quantity Unit : l	Acid	Acid quantity Unit : l	Alkali1	Alkali quantity1 Unit : g	Temperature Unit : oC
1	[2.000e+00 ; 5.000e+00]	Cutting milling	1.000e+00	1.000e+00	Corn stover	1.000e+03	[0.000e+00 ; inf]	0.000e+00	[0.000e+00 ; inf]	[0.000e+00 ; inf]	1.000e+03							
2		Extrusion treatment	1.000e+00	2.000e+00	Corn stover	1.000e+03	2.400e-01	2.000e+00	3.250e+02	3.000e+01	3.000e+03	Salt	0.000e+00	Acid	0.000e+00	Sodium hydroxide	4.000e+01	9.900e+01
3		Alkali treatment	1.000e+00	3.000e+00	Corn stover	3.000e+03	0.000e+00	0.000e+00	0.000e+00	0.000e+00	3.000e+03	Salt	0.000e+00	Acid	0.000e+00	Sodium hydroxide	0.000e+00	[-inf ; 9.900e+01]
4		Washing and filtration	1.000e+00	4.000e+00	Corn stover	3.000e+03	[0.000e+00 ; inf]	[0.000e+00 ; inf]	0.000e+00	[0.000e+00 ; inf]	1.000e+03	Salt	0.000e+00	Acid	0.000e+00	Alkali	0.000e+00	[-inf ; inf]
5		Enzymatic hydrolysis treatment	1.000e+00	5.000e+00	Corn stover	4.000e-01			1.000e+02	[0.000e+00 ; inf]	3.579e-01							5.000e+01
6	[2.000e+00 ; 5.000e+00]	Cutting milling	2.000e+00	1.000e+00	Corn stover	1.000e+03	[0.000e+00 ; inf]	0.000e+00	[0.000e+00 ; inf]	[0.000e+00 ; inf]	1.000e+03							

FIGURE 6.9 – Capture d'écran du tableau « description du procédé » d'@Web de l'article de [LIU et al., 2013]

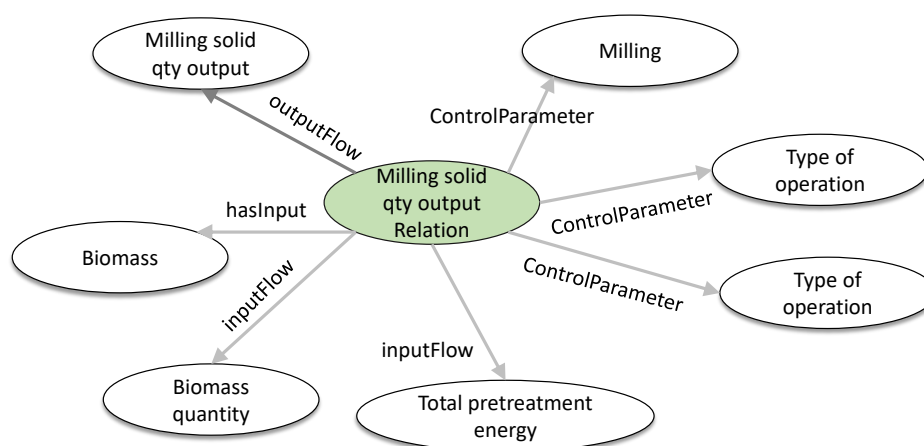


FIGURE 6.10 – Relation *n-aire* "Quantité de constituant solide issue d'un broyage"

6.4.2.2 Enrichissement et stockage des données extraites

L'opération de l'unité de broyage a été identifiée pour chaque article, mais les données énergétiques étaient manquantes, ce qui dessert l'analyse de durabilité. Des modèles empiriques internes créés par des experts du prétraitement de la paille de maïs étaient disponibles et ont été utilisés pour enrichir les données. Ces données proviennent de différentes études faites par le laboratoire IATE de SupAgro Montpellier [TABOUBI, 2019]. Elles nous ont permis de créer un modèle d'expert qui nous permet de compléter les données énergétiques de l'étape de broyage pour tous les articles. L'énergie de broyage moyenne est calculée par la formule 6.1 :

$$E_{milling} = -0,0125 \times taille_{particule} + 0,0605 \quad (6.1)$$

où $E_{milling}$ est l'énergie moyenne de broyage en kW.h/kg de biomasse et $taille_{particule}$ la taille de la particule en mm. Les données de ce modèle sont disponibles dans l'annexe D et l'énergie moyenne

est une simple régression linéaire suivant la taille des particules en sortie de broyage. Après cet enrichissement, nous avons stocké les données dans une base de données CSV pour nous permettre de les traiter dans 6.4.2.3.

6.4.2.3 Traitement des données extraites

Les données procédés sont traitées à la recherche d'incohérence. La première vérification repose sur une vérification par bilan matière et sur le calcul du rendement en glucose. Si les sorties sont supérieures aux entrées (en massique) ou si le rendement de glucose après la dernière étape (l'hydrolyse enzymatique) est supérieur au rendement théorique, alors les articles contenant ces incohérences sont alors supprimés. De plus, quand le bilan matière ne peut pas être vérifié ou calculé, nous considérons que la qualité de ces articles ne correspond pas à nos attentes et les articles sont supprimés. En effet, dans certains articles de prétraitement, les auteurs ne précisent pas suffisamment les étapes du procédé de transformation de la matière et les flux correspondants. Quatre articles possèdent ces incohérences et sont écartés. Cette analyse est faite par les ingénieurs procédés sur la base des connaissances extraites de @Web.

Pour permettre une analyse plus rapide et plus cohérente des procédés, nous décidons ici d'utiliser les articles scientifiques présentant les biomasses et les procédés les plus représentés. Nous nous retrouvons avec 4 biomasses et 3 types de procédés (*Pre-milling* (PM), *Pre-milling and UltraFine Milling* (PM-UFM) et *Pre-milling, Physico-Chimical treatment, and Press and Separation* (PM-PC-PS)) contenus dans 51 articles contenant 456 expériences comptabilisant 30 705 données. Pour la suite, nous utilisons 3 niveaux de granularité sur ces données :

- **Niveau 1** : Chaque expérience décrite dans les articles est utilisée soit les 456 expériences.
- **Niveau 2** : Dans chacun des articles, nous avons pris un ou plusieurs expériences-types sous la base de plusieurs critères. Une expérience-type doit avoir le même cocktail enzymatique (pour enlever la variance sur le type de cocktail) et le même nombre d'étapes. Il y a 63 expériences types.
- **Niveau 3** : Une expérience-type est prise par article. Il y a 51 expériences-types.

Le détail des niveau 1 et 3 par types de procédés et type de biomasses est donné dans le tableau 6.3.

TABLEAU 6.3 – Nombre d'expériences-types suivant le niveau de granularité de l'étude, le type de procédé et la biomasse

Niveau	Type de procédés	Biomasse				Total
		Bagasse	Corn Stover	Rice Straw	Wheat Straw	
Niveau 1	PM-PC-PS	127	135	26	93	378
	PM	6	13	3	7	23
	PM-UFM	25	13	9	2	49
	Total	158	161	35	102	456
Niveau 3	PM-PC-PS	6	10	6	8	30
	PM	2	4	3	3	12
	PM-UFM	2	3	2	2	9
	Total	10	17	11	13	51

Les articles présentent des procédés discontinus et se focalisent majoritairement sur certaines étapes de ceux-ci. Beaucoup de ces articles ne contiennent donc pas la donnée sur l'entrée de la bio-

masse dans l'étape de prébroyage, la première étape des procédés. Cette donnée est trouvée dans l'étape suivante. L'hypothèse prise est que l'entrée du procédé est la même que l'entrée de la première étape où l'on trouve l'information. Ceci sous-entend qu'il n'y a pas de perte sur cette étape de prébroyage (ce qui est en réalité faux) et c'est donc une limite. Il est possible d'améliorer cela en mettant un pourcentage de pertes qu'il faudra évaluer à dire d'expert.

6.4.2.4 Analyses et visualisation des données

Les parties analyse et visualisation des données sont effectuées en même temps. Dans le chapitre 2, l'apport du ML et notamment de l'apprentissage supervisé est considéré pour faire cette analyse. Le chapitre 4 présente les réseaux de neurones, méthode de régression, choisis pour ces analyses ainsi que les méthodes PDP et ICE pour les interpréter. Les données de niveau 3 de granularité de l'étape précédente sont envoyées dans l'outil Orange pour effectuer la régression de ces données puis dans R pour calculer les graphes PDP et ICE. La figure 6.11 schématise toutes les liaisons entre les divers outils utilisés dans ce cadre.

Pour la régression, Orange permet de faire la succession des étapes de façon visuelle et automatique (figure 6.11). Pour simplifier le discours, seules les expériences de niveau 1 des procédés PM et PM-UFM sont prises en compte dans la suite soient 78 expériences. Après que les données soient extraites dans le logiciel puis normalisées, la variable d'intérêt est spécifiée ainsi que les variables d'entrée. La normalisation est la transformations des valeurs en valeurs centrées et réduites. Si x_i la valeur réelle alors la valeur centrée-réduite z_i est égale à :

$$z_i = \frac{x_i - \text{moyenne}(x_i)}{\sigma} \quad (6.2)$$

Avec σ est l'écart type de x_i .

La variable d'intérêt est le rendement en glucose (*glucose_yield*). Seules les variables numériques d'entrée sont prises en compte : Taux de glucose théorique possible, Activité enzymatique, Taille minimale du constituant solide de sortie des étapes de broyage, Température (maximale), Quantité de biomasse à l'entrée, Quantité de liquide tampon et la durée de l'étape d'hydrolyse enzymatique.

Les variables sont séparées en deux groupes :

- le groupe d'apprentissage : ce sont les données qui vont permettre d'entraîner le modèle ML.
- le groupe de test : ce sont les données qui vont permettre d'évaluer la fiabilité du modèle en comparant la valeur prédite par le modèle à la valeur correspondante à ce groupe.

Le groupe d'apprentissage représente 80% des données soit 63 expériences et le groupe de test 20% soit 15 expériences. Cinq modèles sont testés sur Orange et c'est le modèle RN qui est le meilleur avec un $R^2 = 0,974$. Le RN est constitué de 4 couches de 100 neurones chacune. La visualisation des données prédites en fonction des données réelles est représentée sur la figure 6.12. Les détails des autres modèles et les visualisations sont présentés en Annexe E.

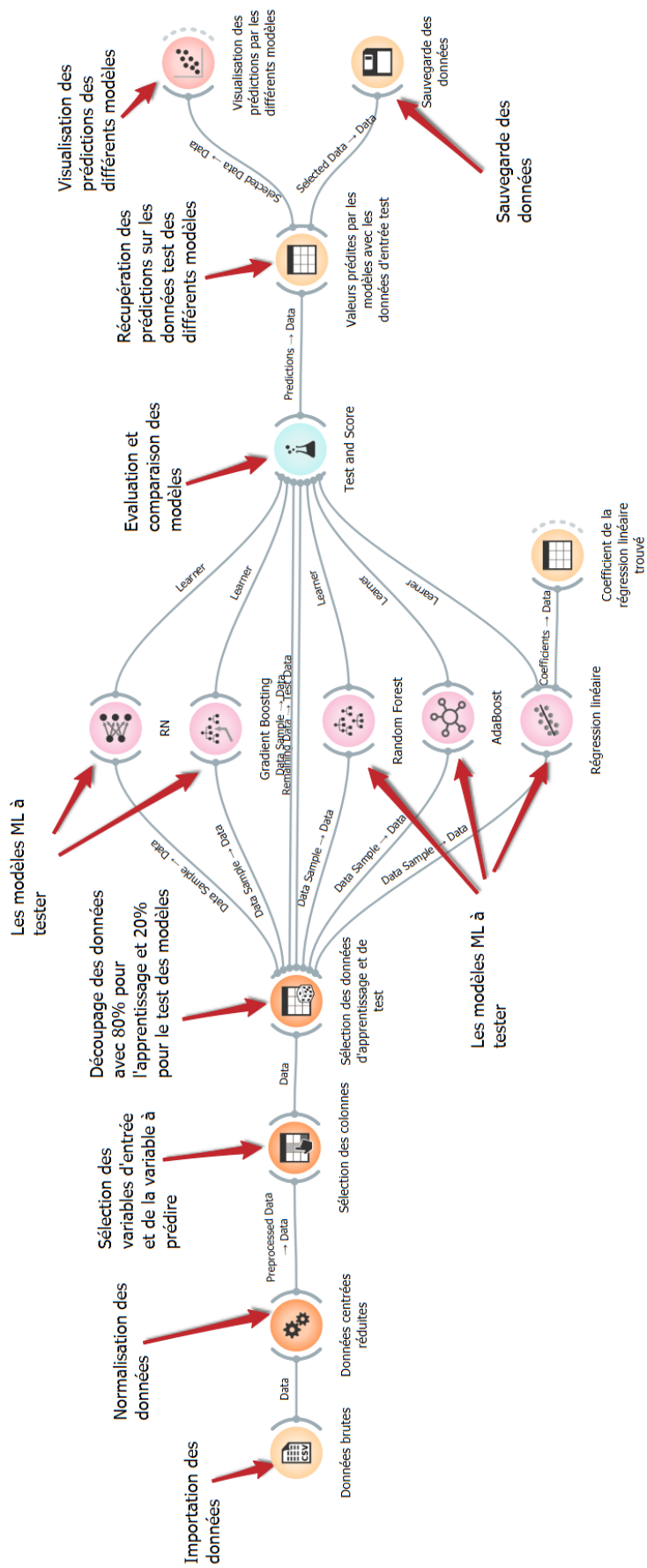


FIGURE 6.11 – Schéma général de la régression sous le logiciel Orange.

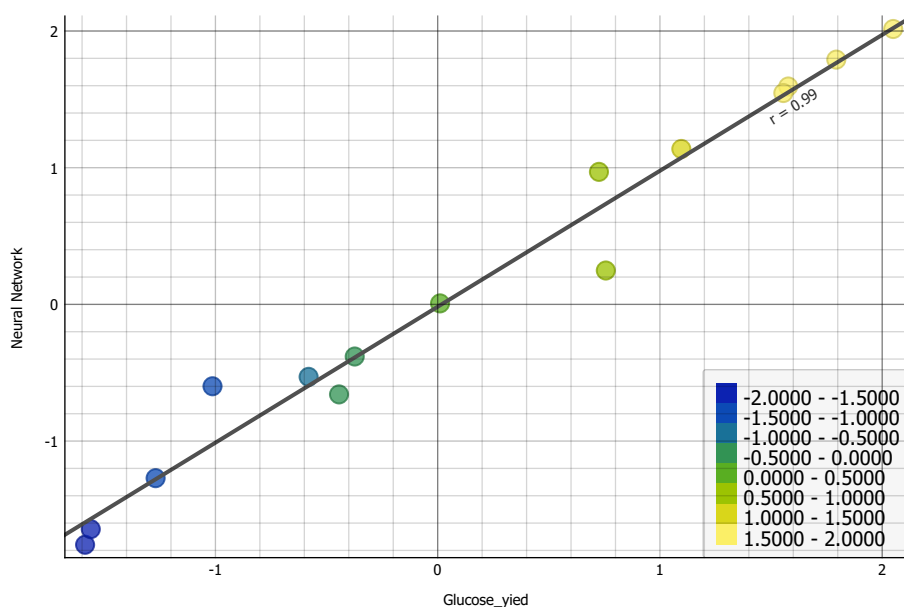


FIGURE 6.12 – Visualisation des rendements en glucose prédits par le modèle RN en fonction des rendements en glucose réel (valeurs normalisées)

Une fois la régression faite, il est possible de reprendre le modèle pour "simuler" des expériences sans avoir à faire des expériences réelles. Il est recommandé de rester dans les intervalles des variables d'entrée. Pour visualiser la répartition des différents variables d'entrée, les graphes PDP et ICE sont tracés (Figure 6.13). On remarque que les courbes ICE sont globalement semblables pour chaque graphique, c'est à dire que les expériences évoluent de la même façon si on fait varier les variables une à une. Les graphiques 2 et 3 de la première ligne, nous font revenir sur les différences entre PDP et ICE. Si une courbe PDP est horizontale sur ces graphiques, on pourrait en conclure que la variable d'entrée correspondante au graphique n'influence pas sur le rendement. Or, nous voyons très bien par les courbes ICE que cela est faux. Par exemple, l'activité enzymatique et la taille des particules à la sortie des broyeurs influencent le rendement. Il est à noter que certaines expériences nous montrent que plus la durée de l'hydrolyse augmente plus le rendement en glucose (*glucose yield* diminue). Or, la majorité des expériences montre le contraire (deuxième graphique de la deuxième ligne).

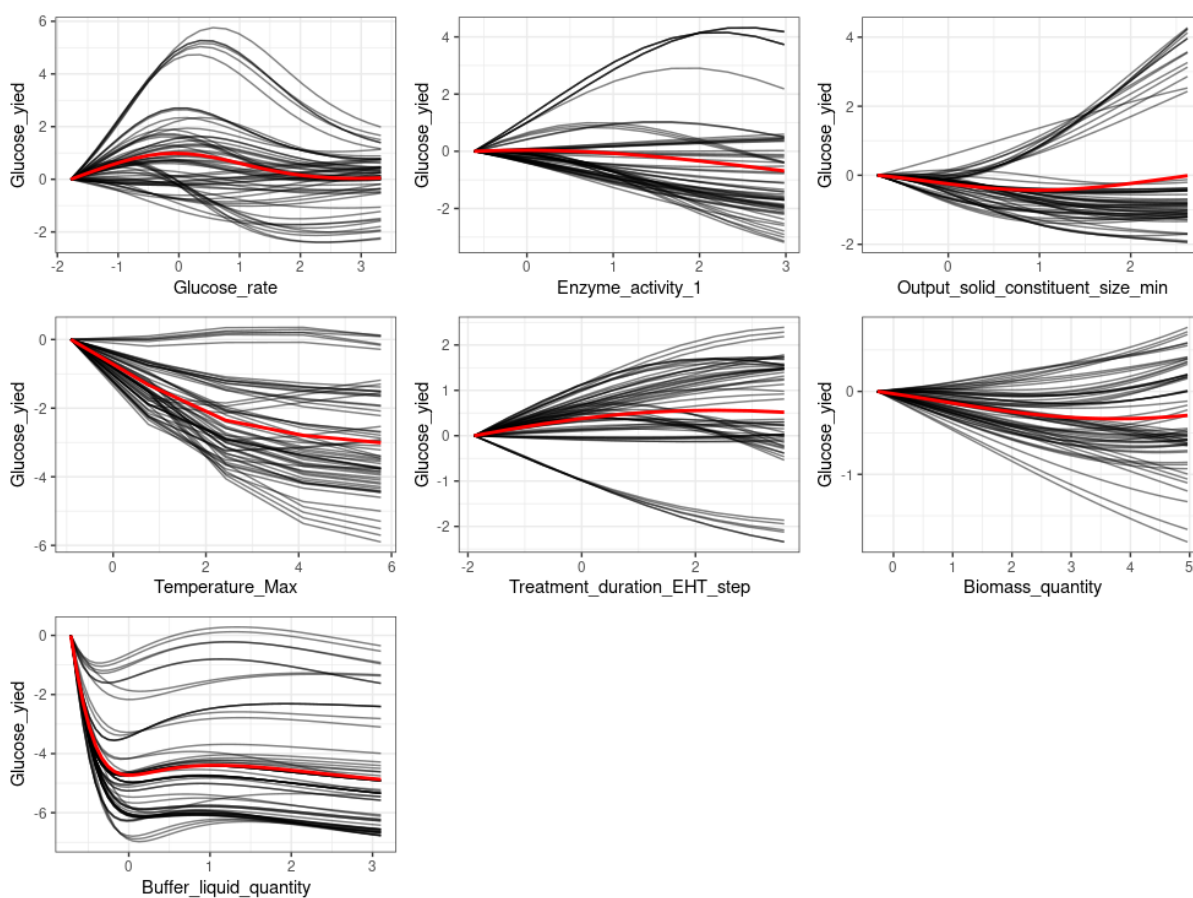


FIGURE 6.13 – Graphiques centrés de dépendance partielle (en rouge) et d'espérance conditionnelle individuelle (en noir).

6.4.3 Étape 3 : Évaluation de la durabilité

Suite à l'étape 2, les données procédés sont traitées et disponibles pour exécuter une analyse de la durabilité. Nous nous concentrons sur une analyse environnementale pour ce cas d'étude.

6.4.3.1 Inventaire du cycle de vie

L'inventaire du cycle de vie est réalisé selon la méthode ACV comme discuté au chapitre 3. La quantification des flux directs (entrées et sorties) tels que l'électricité ou l'acide est effectuée. Les données de premier plan issues des publications sur les prétraitements sont utilisées pour cette quantification. Ces données sont regroupées dans la base d'annotation traitée issue de l'étape 2. Une fois extraites, les données de premier plan sont exprimées par unité fonctionnelle. Une base de données qui contient les données d'arrière-plan est nécessaire pour compléter les données de premier plan selon la pensée cycle de vie. La base de données EcoInvent v3 est retenue dans cette étude.

6.4.3.2 Évaluation environnementale du cycle de vie

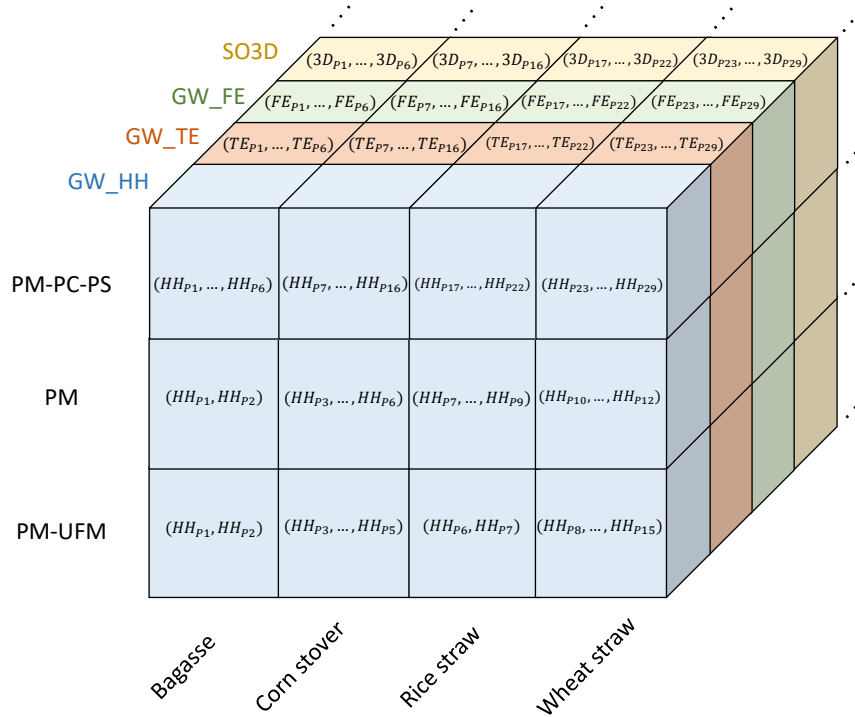
Pour cette étude, la méthode **ReCiPe 2016 Endpoint Hierarchical** est sélectionnée. Cette méthode calcule les indicateurs du premier niveau d'agrégation (impacts *midpoint*) jusqu'au troisième niveau (score unique) (voir section 5.2. Le tableau 6.4 présente les 18 impacts *midpoint* de cette méthode.

TABLEAU 6.4 – Liste des impacts *midpoint* de la méthode **ReCiPe 2016 Endpoint Hierarchical**, abréviations et unité

Impacts (en anglais)	Abréviation	Unité
Global warming, Human health	GW_HH	DALY
Global warming, Terrestrial ecosystems	GW_TE	species.yr
Global warming, Freshwater ecosystems	GW_FE	species.yr
Stratospheric ozone depletion	SO3D	DALY
Ionizing radiation	IR	DALY
Ozone formation, Human health	O3_HH	DALY
Fine particulate matter formation	FPM	DALY
Ozone formation, Terrestrial ecosystems	O3_TE	species.yr
Terrestrial acidification	TA	species.yr
Freshwater eutrophication	FE	species.yr
Marine eutrophication	ME	species.yr
Terrestrial ecotoxicity	TEco	species.yr
Freshwater ecotoxicity	FEco	species.yr
Marine ecotoxicity	Meco	species.yr
Human carcinogenic toxicity	HcT	DALY
Human non-carcinogenic toxicity	HncT	DALY
Land use	LU	species.yr
Mineral resource scarcity	MS	USD2013
Fossil resource scarcity	FS	USD2013

Attardons-nous sur les données disponibles à la sortie de cette étape et toute la complexité de ces données pour l'analyse des résultats. Reprenons les niveaux 1 et 3 définis précédemment (section 6.4.2.3). Pour chacun de ces niveaux nous calculons les 18 impacts *midpoint* de **ReCiPe 2016 Endpoint Hierarchical** générant une matrice en trois dimensions [*type de procédés* × *biomasse* × *impacts*]. Les cellules de cette matrice seront représentées par des *k* – *uplets* où *k* dépend du nombre d'expériences-type par biomasse : ils sont donc différents pour chaque cellule. La figure 6.14 schématise la matrice au niveau 3 de granularité.

Niveau 1 de granularité La matrice est de dimension [3 × 4 × 18]. Pour la cellule [PM – PC – PS × *bagasse* × GW_HH], c'est un 127 – *uplet* conformément au tableau 6.3 et de même, on retrouve un 127 – *uplet* pour tous les autres impacts de la ligne [PM – PC – PS × *bagasse*]. Pour la ligne [PM × *bagasse*] ce sont des 6-uplets et ainsi de suite pour toutes les cellules de la matrice. Cela nous conduit à un nombre de données égal à 8 208 données. Chaque donnée correspond à l'impact d'une expérience donnée pour une biomasse donnée et un type de procédé donnée. Mathématiquement, si *b* représente la biomasse, *p* le type de procédé et $n_{p,b}$ le nombre d'expérience pour un *p* et un *b* fixé, on obtient le nombre de données pour ce niveau de granularité nbd_{g1} :



X_{pi} : impact X du procédé Pi

X représente les 2 dernières lettres des abréviations des impacts (tableau 6.4)

FIGURE 6.14 – Représentation de la matrice [type de procédés × biomasse × impacts] pour les impacts *endpoint* de la méthode ReCiPe 2016 Endpoint Hierarchical

$$nbd_{g1} = \left(\sum_p \sum_b \sum_{i=1}^{n_{p,b}} E_{p,b,i} \right) \times imp_{mp} \quad (6.3)$$

Où $E_{p,b,i}$ est l'expérience i pour un p et un b fixés, imp_{mp} le nombre d'impacts *midpoint* et $g1$ le niveau 1 de granularité.

Niveau 3 de granularité La matrice est toujours de dimension [3 × 4 × 18]. Pour la cellule [PM – PC – PS × bagasse × GW_HH], c'est cette fois-ci un 6 – *uplet* conformément au tableau 6.3 et de même, on retrouve un 6 – *uplet* pour tous les autres impacts de la ligne [PM – PC – PS × bagasse]. Pour la ligne [PM × bagasse] ce sont des 2-uplets et ainsi de suite pour toutes les cellules de la matrice. Ce qui nous donne à la fin un nombre de données qui est égal à 918 données. L'équation mathématique est la même que l'équation 6.3 en remplaçant le nombre de données du niveau 1 par celui de niveau 2 nbd_{g2} .

Généralisons l'équation. Le nombre d'expérience $n_{p,b}$ ne varie plus seulement en fonction de p et b comme l'équation 6.3 mais aussi en fonction du niveau de granularité qu'on nomme g . Alors la

somme totale des données de tous les niveaux de granularité est :

$$nbd = \left(\sum_{g=1}^3 \sum_p \sum_b \sum_{i=1}^{n_{p,b}} E_{p,b,i} \right) \times imp_{mp} \quad (6.4)$$

Pour généraliser cette équation on prend en compte les trois niveaux d'agrégation de la méthode **ReCiPe 2016 Endpoint Hierarchical**. Et généralement, dans une approche ACV, plusieurs méthodes sont utilisées entraînant une boucle et une augmentation des données. Ces boucles avec l'agrégation et les méthodes sont illustrées dans la conclusion du chapitre 3 avec la figure 3.6. L'équation globale est alors : si m représente la méthode de calcul d'impacts, a le niveau d'agrégation des impacts, b la biomasse, p le type de procédé et $n_{p,b}$ le nombre d'expérience pour un p et un b fixé :

$$nbd_{total} = \sum_m \left[\left(\sum_{g=1}^3 \sum_p \sum_b \sum_{i=1}^{n_{p,b}} E_{p,b,i} \right) \times \sum_{a=1}^{max_m(a)} imp_{a,m} \right] \quad (6.5)$$

Avec $max_m(a)$ le nombre maximal d'agrégation suivant la méthode m

Le nombre de données à analyser et visualiser est donc élevé et particulièrement d'une dimension impossible à visualiser. La figure 6.15 présente toutes les boucles prises en compte dans le cadre méthodologique (et dans l'équation 6.5) et le nombre de données à la sorties des étapes 2 et 4. Á la sortie de la deuxième étape, nous retrouvons nos 30 705 données. Á la sortie de l'étape 4, suivant le niveau de granularité pris en compte, le nombre de données varie entre 8 208 et 918 pour $m = \text{ReCiPe 2016 Endpoint Hierarchical}$ et $a = 1(\text{midpoint})$. Ce qui est impossible à visualiser car de dimension trop importante. C'est pour cela que l'étape suivante utilise des méthodes du ML. Tous les impacts, calculés avec notre outil interne conçu pour ces travaux et présenté dans l'annexe F pour plus de lisibilité, ont été vérifié à l'aide du logiciel SimaPro®.

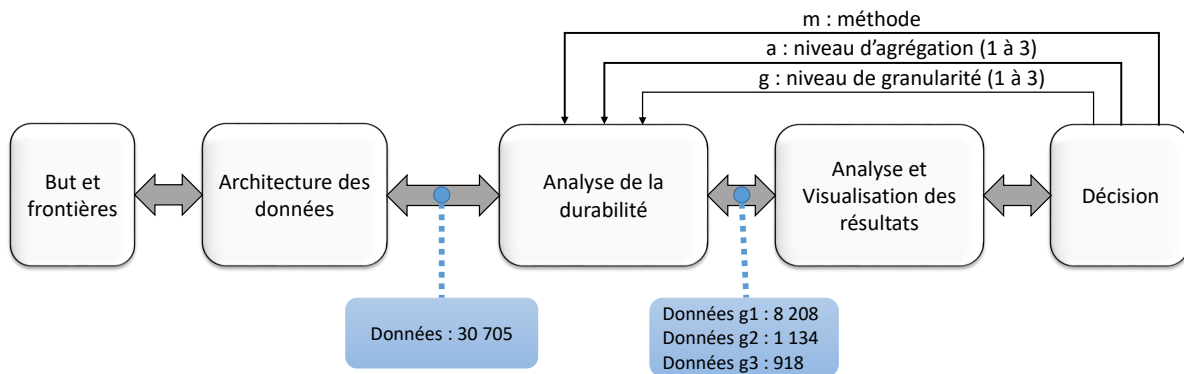


FIGURE 6.15 – Étapes du cadre méthodologique avec les boucles d'itération et la volumétrie des données

Une représentation de nos résultats dans une forme graphique usuelle de résultats d'ACV (figure 6.16) présente les impacts *endpoint* pour 13 prétraitements pour deux des biomasses étudiées : six prétraitements concernent la paille de riz et sept la tige de maïs. Une première analyse des résultats est ensuite effectuée.

Premièrement, le prétraitement PCS6 a un impact très fort par rapport aux autres et peut-être mis à part. Ce résultat peut être surprenant et il est nécessaire de vérifier les résultats en remontant les

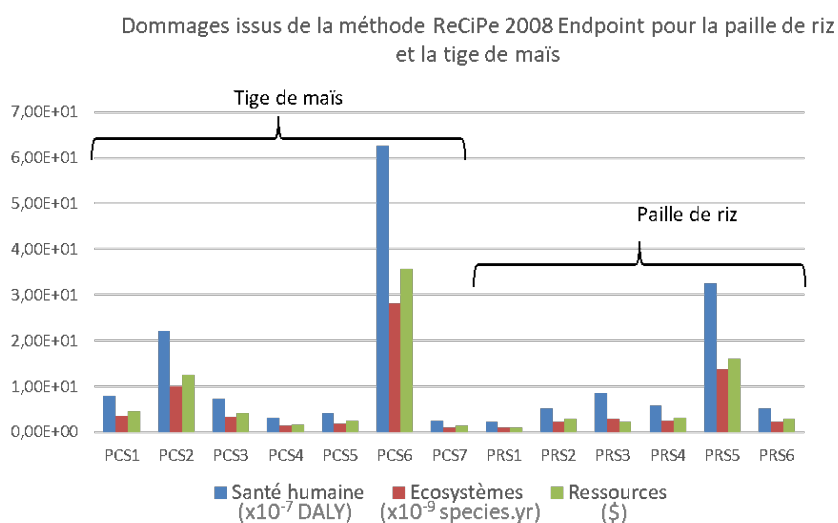


FIGURE 6.16 – Indicateurs de dommages des procédés de prétraitement calculés par la méthode ReCiPe 2016 Endpoint

calculs jusqu'aux données de premier plan. Cette vérification met en évidence une utilisation importante d'eau (presque 10 fois plus que les autres procédés) et d'alcalin (ici du sulfate d'ammonium). Rappelons que tous ces résultats sont exprimés par unité fonctionnelle et peuvent donc être comparés.

Ensuite, le procédé PCS7 qui est le seul procédé purement mécanique traitant la tige de maïs est celui qui a le moins d'impacts.

Troisièmement, PRS5 a des impacts plus importants que les autres procédés traitant la paille de riz et même de tous les autres procédés (hors PCS6). Une quantité d'eau élevée est à l'origine de ces impacts. Les impacts des autres procédés sont sensiblement les mêmes et nous ne pouvons pas mettre en évidence d'autres regroupements significatifs dans cette première itération de l'approche.

Cette première analyse qui s'est effectuée sur des impacts *endpoint* pour seulement deux biomasses et 13 procédés reste très sommaire. Si nous voulons avoir des informations plus avancées, il convient de multiplier les graphiques. Notre cadre méthodologique embarque des méthodes de ML pour aider à cette analyse.

6.4.4 Étape 4 : Analyse et visualisation des résultats

Nous venons de voir qu'il est difficile d'analyser les résultats de l'analyse environnementale. Nous développons dans cette étape les méthodes choisies dans le chapitre 5 c'est-à-dire la méthode MDS (pour la réduction de la dimension) et la méthode k-means pour le clustering. Les résultats sont présentés sous forme de graphiques. Pareillement à l'étape d'analyse des données de l'étape 2, seules les visualisations des données pour les procédés purement mécaniques c'est-à-dire les procédés PM et PM-UFM sont exploitées pour le niveau 1 de granularité. Les 78 expériences avec pour chacune 18 impacts sont analysées. Pour faire le choix de la méthode MDS, plusieurs méthodes ont été testées sur les données du niveau 3 de granularité : les méthodes *Principal Component Analysis* (PCA), *Independent component analysis* (ICA) *Multidimensional scaling* (MDS), *t-distributed stochastic*

neighbor embedding (t-SNE). Parmi ces quatre méthodes seule la méthode MDS donne des résultats pertinents.

Les cartes des procédés et des impacts agrégées en deux dimensions de la méthode MDS et *k-means* sont présentées respectivement dans les figures 6.17 et 6.18. La représentativité de chaque couple de dimensions est précisée. Elle est le pourcentage d'information contenue dans chaque dimension. Les cartes des quatre premières dimensions des procédés et des impacts sont dans l'annexe G. Pour la projection sur les impacts en quatre dimensions, la représentativité est de 66% et pour celle sur les procédés elle est de plus de 97%. La représentativité cumulée jusqu'à la dimension huit pour les deux projections est représentée dans l'annexe G. L'interprétation d'un résultat de MDS est la même que pour toute autre réduction de dimension; les objets qui sont plus proches sur le nuage de points sont plus semblables que ceux qui sont plus éloignés. C'est-à-dire que les points projetés sont disposés de telle manière que les points groupés (petite distance géométrique entre eux) reflètent les relations originales dans les données. Cependant, des informations supplémentaires sont nécessaires pour rendre la projection plus informative. L'algorithme de *clustering* (ou de regroupement), ici la méthode *k-means*, met en évidence les objets les plus similaires. Pour déterminer le nombre de *clusters* de chaque représentation, des méthodes automatiques sont utilisées qui maximisent la variance inter-groupe (la distance la plus lointaine entre groupes) et minimisent la variance intra-groupe. À la fois pour les procédés et pour les impacts, le nombre de *clusters* optimal est de trois.

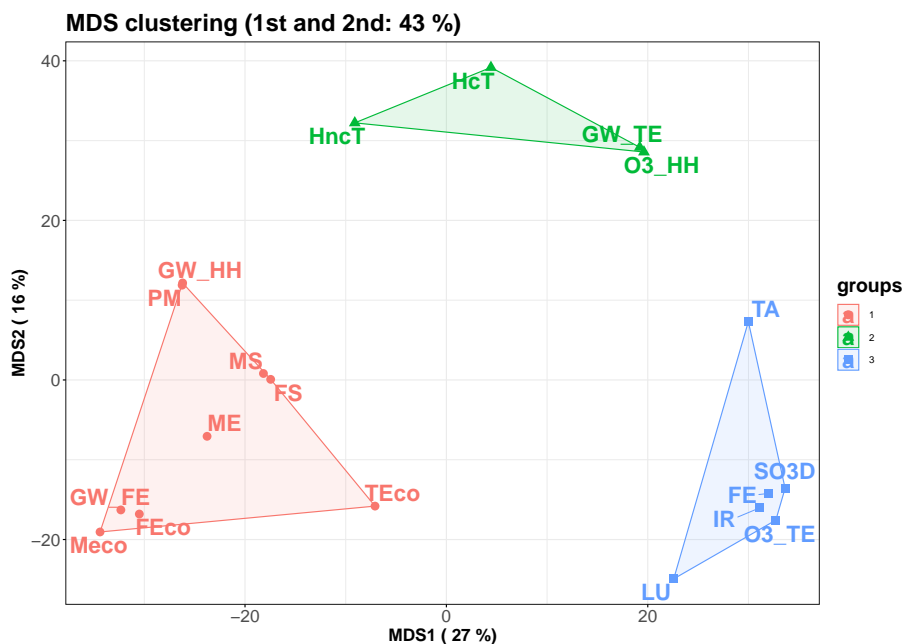


FIGURE 6.17 – Diagramme de dispersion de la projection MDS (deux dimensions) et du *clustering k-means* basé sur la matrice de distance des impacts avec la représentativité entre chaque dimension

6.4.4.1 Analyse de la dispersion des impacts

Pour le diagramme de dispersion des impacts (6.17), nous pouvons clairement distinguer trois groupes à l'aide des k-means :

Groupe 1 *Global warming (Human health), Global warming (Freshwater ecosystems), Fine particulate matter formation, Marine eutrophication, Terrestrial ecotoxicity, Freshwater ecotoxicity, Marine ecotoxicity, Mineral resource scarcity et Fossil resource scarcity.* La première remarque est la présence des deux impacts de déplétion (MS et FS) qui sont très proches les uns des autres, ce qui est cohérent. La deuxième remarque est la présence des trois impacts qui concernent l'écotoxicité (TEco, FEco et MEco) ce qui nous paraît aussi cohérent. De plus, les impacts sur le changement climatique lié à la santé humaine (GW_HH) et aux écosystèmes d'eau douce (GW_FE) se retrouvent dans ce groupe contrairement à celui lié aux écosystèmes terrestres (GW_TE) qui est dans le groupe 2. Nous pouvons en conclure que pour nos procédés les intrants ont des impacts similaires sur la santé humaine et les écosystèmes d'eau douce. La formation de fines particules de matière (FPM) dans ce groupe est explicitée par la présence de GW_HH, ces deux impacts sont très liés dans la méthode ReCiPe. Ils sont dans la représentation sur le même point mettant en lumière ce lien fort.

Groupe 2 *Global warming (Terrestrial ecosystems), Ozone formation (Human health), Human carcinogenic toxicity et Human non-carcinogenic toxicity.* Les impacts sur la toxicité humaine cancérigène et non (HcT et HncT) sont similaires et sont dans ce groupe comme la formation d'ozone lié à la santé humaine. Cette présence dans le même groupe est cohérente en vue de la méthode ReCiPe. Cependant les deux autres impacts ne sont pas similaires dans la méthode. Nous pouvons en conclure que c'est les intrants de nos procédés qui sont la cause de cette présence.

Groupe 3 *Stratospheric ozone depletion, Ionizing radiation, Ozone formation (Terrestrial ecosystems), Terrestrial acidification, Freshwater eutrophication et Land use.* Ce groupe comporte le reste des impacts qui ne sont pas similaires aux autres groupes. Notamment deux impacts liés à l'ozone (SO3D et O3_TE) et la radiation ionisante (IR) très liés dans la méthode ReCiPe et sont très proches sur la représentation.

Cette description est assez cohérente avec le calcul de la méthode ReCiPe et valide l'utilisation du MDS.

6.4.4.2 Analyse de la dispersion des procédés

Passons maintenant à l'analyse des différents procédés (Figure 6.18).

Contrairement à la représentation projetée sur les impacts, celle sur les procédés comporte des groupes très étroits avec une représentativité beaucoup plus grande (près du double). Analysons les trois groupes pour comprendre pourquoi la méthode MDS a généré ces groupes-ci. Pour ne pas surcharger le graphique toutes les étiquettes des procédés n'ont pas été écrites mais un tableau donnant le groupe par procédé se trouve dans l'annexe G.

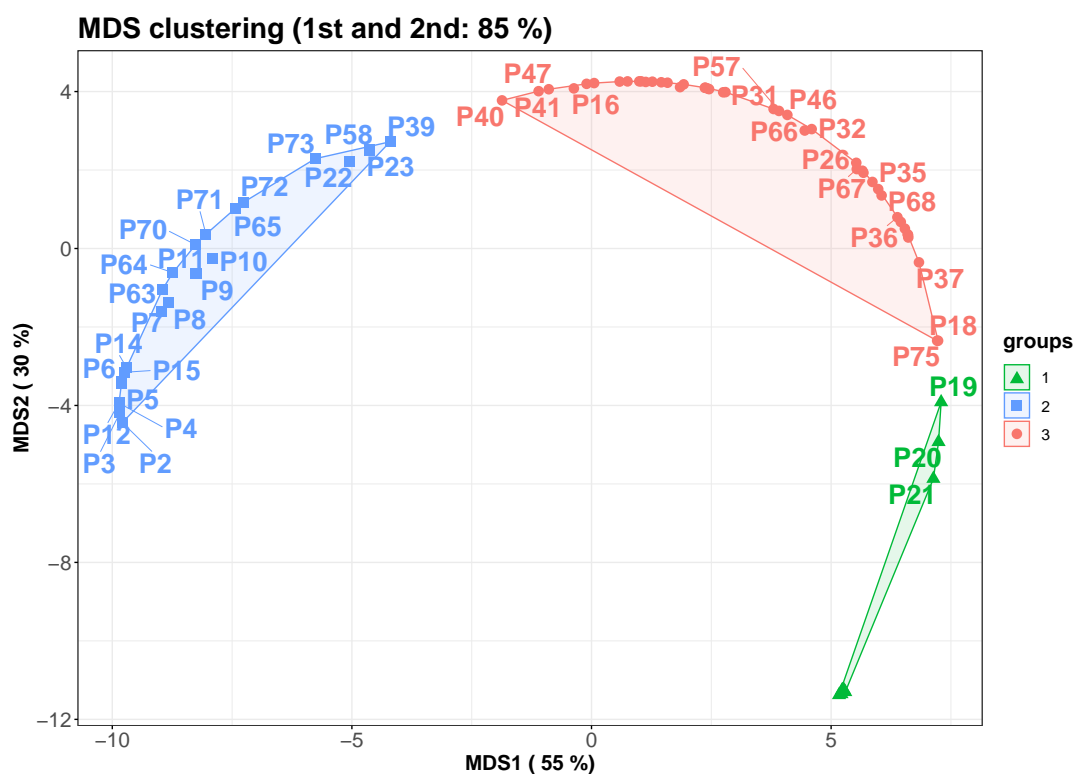


FIGURE 6.18 – Diagramme de dispersion de la projection MDS (deux dimensions) et du clustering *k-means* basé sur la matrice de distance des procédés avec la représentativité entre chaque dimension

Groupe 1 Il contient 10 expériences. Dans ce groupe, on retrouve les expériences P19, P20, P21 qui proviennent du même article. L'expérience P18 qui provient aussi de cette article est dans le groupe 3. Si on remonte aux impacts, on remarque qu'ils sont plus élevés pour P18 que pour les trois autres. En remontant jusqu'aux données, c'est le rendement en glucose de P18 qui est très inférieur aux autres (presque la moitié de P21) entraînant, par conversion à l'unité fonctionnelle, des intrants plus importants. Cet écart est dû à la durée de traitement très courte pour ce procédé et donc une conversion de la cellulose en glucose moins importante. Dans ce groupe, il y a aussi les procédés P27, P28 et P29 qui constituent les trois expériences d'un même article ; les procédés P76, P77 et P78, celles d'un autre article et P74 l'expérience unique d'un article. En remontant aux impacts, les procédés du groupe 1 sont ceux qui ont les impacts les plus faibles. Ce groupe est composé des deux types de procédés étudiés (PM et PM-UFM).

Groupe 2 Il contient 25 expériences. Dans ce groupe, se trouvent les expériences P8 à P15 (appartenant au même article) ainsi que les expériences de P2 à P6, l'expérience P7, les expériences P22 et P23, l'expérience P58 toutes appartenant à un article particulier. Mettons en exergue les expériences P59 à P65 qui appartiennent au même article (annexe G.1). Les P59 à P62 appartiennent au groupe 3 contrairement aux P63, P64 et P65 qui sont dans ce groupe. La même observation est faite pour les expériences de P66 à P73 : P70 à P73 appartiennent au groupe 2 et P66 à P69 au groupe 3. En remontant aux impacts, on remarque que les procédés du groupe 2 ont

des impacts plus importants que ceux du groupe 3. Les données procédés fournissent l'explication de cette différence (de groupe et d'impacts) : les expériences du groupe 2 nécessitent de l'eau en entrée contrairement à ceux du groupe 3.

Groupe 3 Ce groupe contient le plus d'expériences (43) et notamment le plus d'expériences de type PM-UFM.

6.4.4.3 Conclusions de la partie analyse et visualisation

A la suite de cette analyse, les décideurs s'intéressent au groupe 1 contenant les impacts les plus faibles. Il est possible de refaire tourner la méthode MDS avec seulement les données de ce groupe pour trouver des sous-groupes et ainsi affiner l'analyse. L'analyse grâce à la méthode MDS est facilitée. Les différents groupes d'expériences mis en lumière auraient été difficiles à remarquer sans ça. Pour faciliter l'analyse et la visualiser, un logiciel de visualisation interactive (PowerBi®) permet d'aller facilement des données de la méthode MDS, aux impacts *midpoint* puis aux données procédés. A la suite, un graphique avec le rendement en glucose peut être tracé pour évaluer les procédés les uns par rapport aux autres et compléter l'analyse.

Le niveau 1 de granularité permet d'obtenir les résultats les plus détaillés mais est aussi le plus difficile à mettre en place à cause du nombre de données. Pour avoir des résultats plus simples à mettre en place, le niveau 2 et 3 sont recommandés. Cependant, le niveau 3 en faisant une moyenne de toutes les expériences par article ne permet pas de différencier les expériences d'un même article entre elles. Par exemple, grâce au niveau 1, nous avons distingué deux groupes dans les expériences P59 à P65 qui appartiennent au même article (PM-UFM5). De même, dans cet exemple, le niveau 2 ne fait pas cette distinction car ces expériences ont le même cocktail enzymatique et le même nombre d'étapes.

Si le niveau 1 est le « meilleur » pour l'analyse, il a un temps de calcul plus important notamment le calcul des impacts. Pour réduire ce temps, la solution est de commencer par le niveau 2, de choisir le groupe avec les impacts les moins importants et d'itérer le calcul des impacts avec le niveau 1 en ne gardant que les expériences contenues dans le groupe choisi. Cela économise le temps de calcul des expériences provenant d'articles qui décrivent des procédés moins durables.

6.5 Création d'un outil-prototype

Pour confronter notre cadre méthodologique, un outil combinant Excel® (notamment Excel-VBA), Orange, Python ou R est conçu et développé. La figure 6.19 montre l'architecture générale de cet outil et les bases de données utilisées. La première étape est de mettre dans une base structurée pour l'ACV les tables extraites de @Web, d'homogénéiser les dimensions et de passer à l'unité fonctionnelle les données extraites. La deuxième étape est l'enrichissement avec les données à titre d'expert et le traitement de ces données notamment en vérifier les bilans matières (le logiciel ProSim® fait office ici de validation pour cette partie). Puis ces données sont stockées dans des fichiers CSV. La troisième étape est l'analyse de ces données sur Orange ou Python grâce aux réseaux

de neurones. La quatrième étape est l'analyse de la durabilité. La validation des analyses environnementales se fait grâce au logiciel Simapro®. Puis les résultats de ces analyses sont, pour la dernière étape, envoyés via un fichier CSV dans Python ou R pour utiliser les méthodes MDS et *k-means*. Pour la partie analyse et visualisation des résultats et aider les décideurs, un logiciel de visualisation interactif peut être utilisé en exploitant toutes les données précédentes.

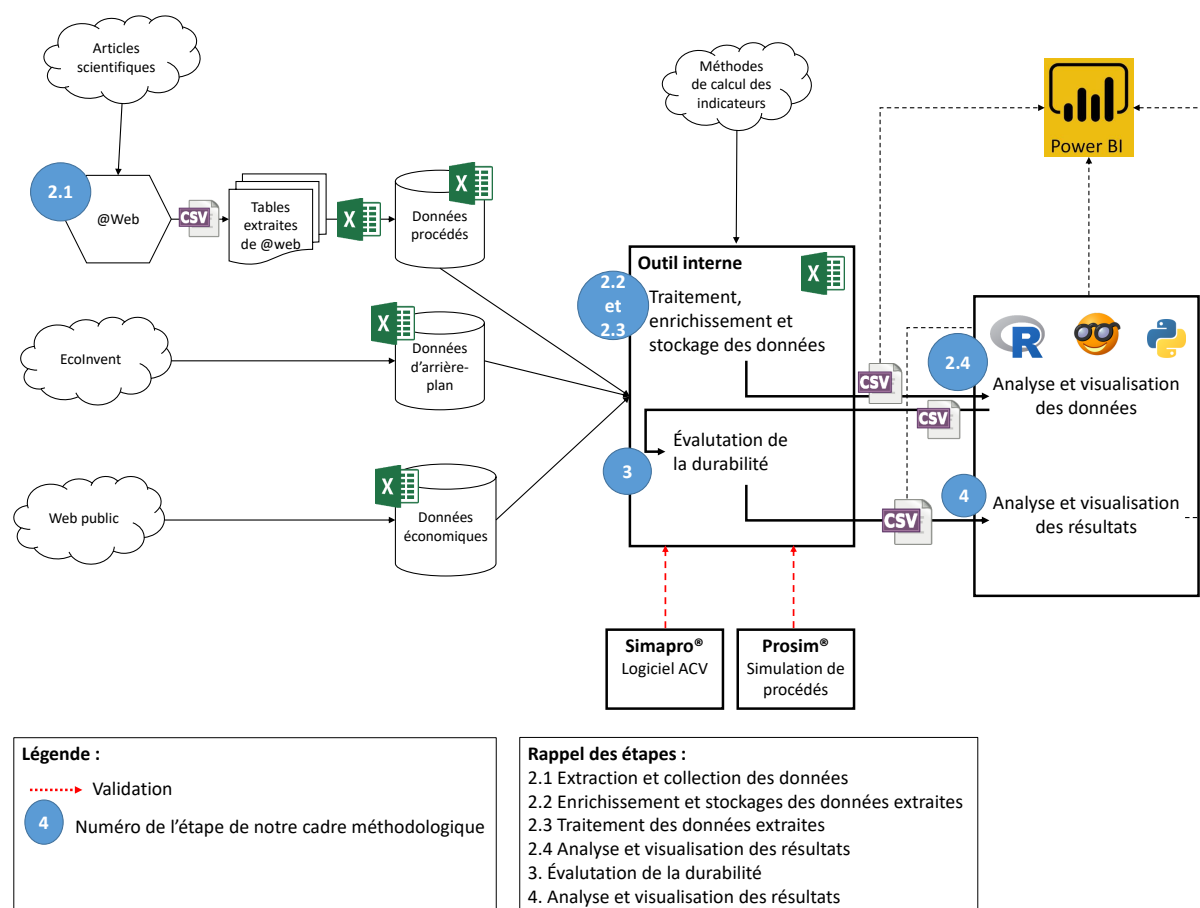


FIGURE 6.19 – Schéma de fonctionnement de l'outil interne créé pour valider le cadre méthodologique

L'outil tel un prototype logiciel est une aide pour prouver le concept du cadre méthodologique. Du fait de l'intégration de différents logiciels, il est impossible d'avoir une automatisation totale du cadre. Un prototype logiciel plus intégré permettrait d'avoir une automatisation du cadre ainsi qu'une généricité de l'outil. Cela constitue une perspective à long terme. Les différents modules de l'outil VBA-Excel sont spécifiés dans la figure 6.20 ainsi que leur interaction. Des visualisations de l'outil sont fournies en figure 6.21, 6.22, 6.23 et 6.24. Le guide d'utilisation simple se trouve en annexe H.

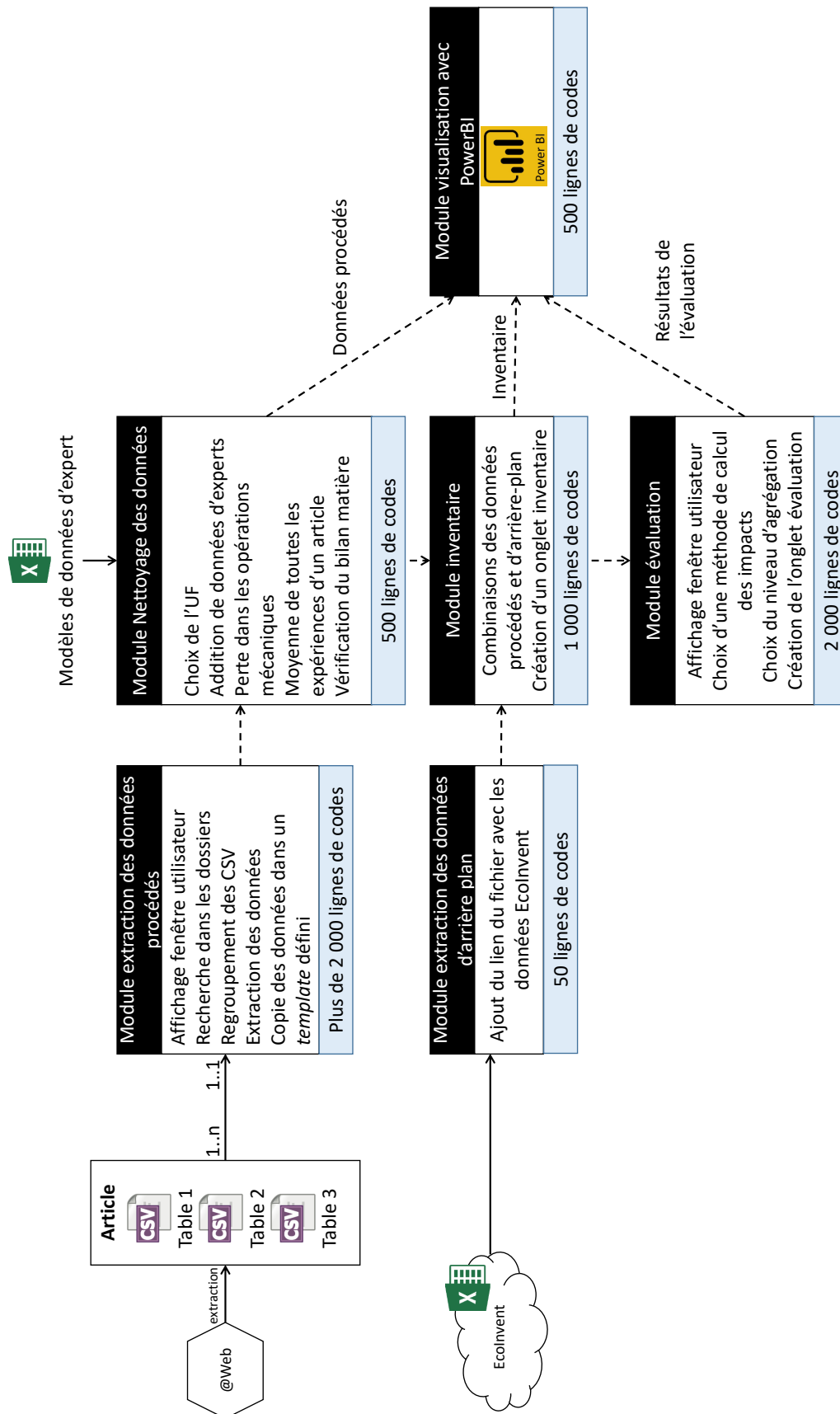


FIGURE 6.20 – Architecture des modules VBA-Excel.

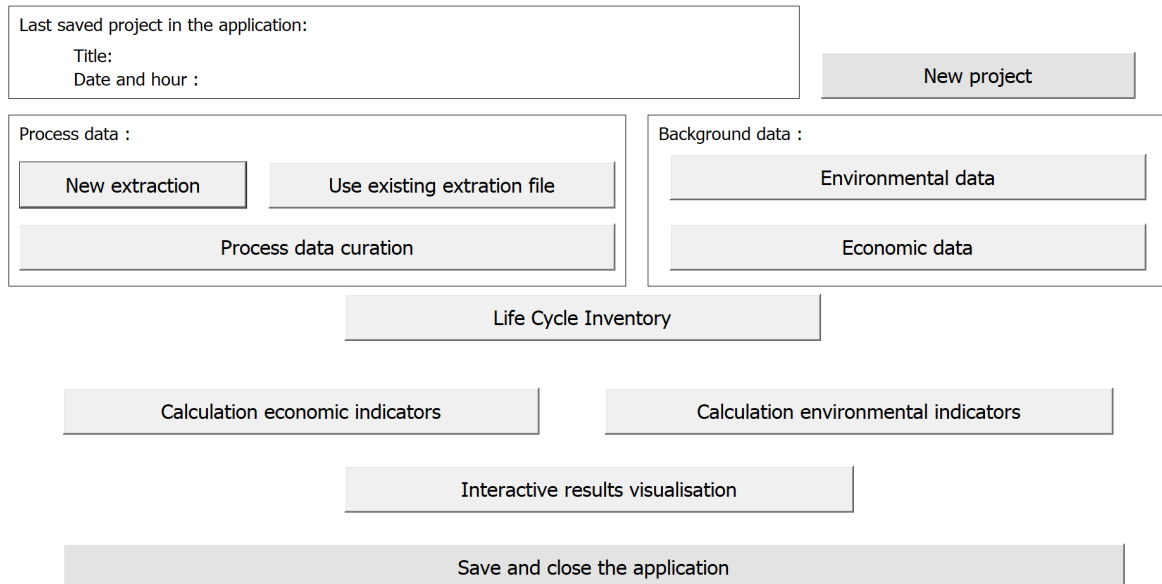


FIGURE 6.21 – Menu principal de l’outil Excel.

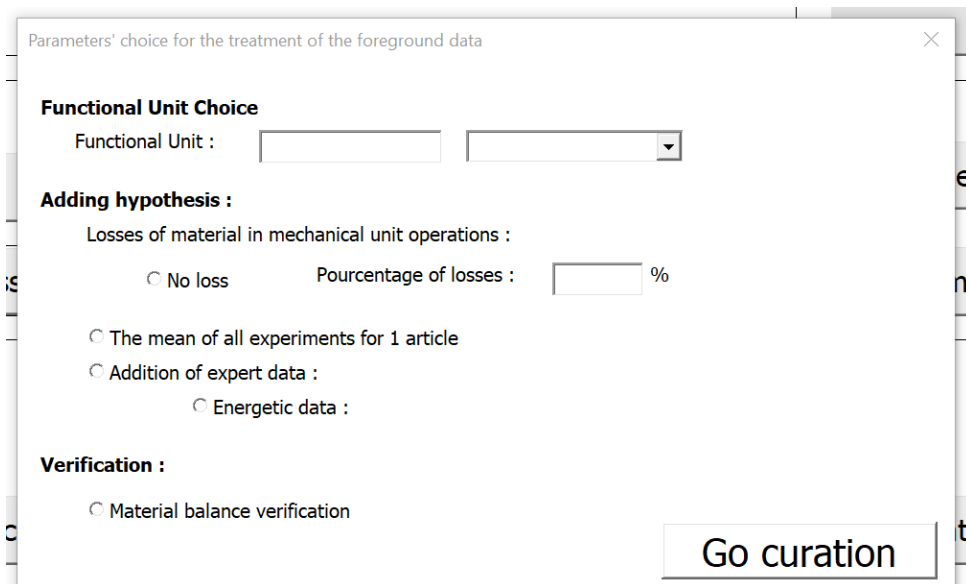


FIGURE 6.22 – Fenêtre du nettoyage des données procédés (ou données de premier-plan).

Impact category	Unit	PM-PC-EX-P5	PM-PC-EX-P5	PM-PC-PS14	PM-PC-PS20	PM-PC-PS23	PM-PC-PS5	PM-UFM1
Climate change Human Health	DALY	5,46E-07	1,53E-06	5,02E-07	2,09E-07	2,94E-07	4,35E-06	1,69E-07
Ozone depletion	DALY	1,86E-10	5,22E-10	1,71E-10	7,14E-11	1,00E-10	1,48E-09	5,76E-11
Human toxicity	DALY	1,21E-07	3,38E-07	1,11E-07	4,62E-08	6,38E-08	9,59E-07	3,73E-08
Photochemical oxidant formation	DALY	5,50E-11	1,54E-10	5,05E-11	2,11E-11	2,92E-11	4,37E-10	1,70E-11
Particulate matter formation	DALY	1,19E-07	3,34E-07	1,09E-07	4,63E-08	6,39E-08	9,48E-07	3,68E-08
Ionising radiation	DALY	2,42E-09	6,77E-09	2,22E-09	9,25E-10	1,27E-09	1,92E-08	7,47E-10
Climate change Ecosystems	species.yr	3,09E-09	8,67E-09	2,84E-09	1,19E-09	1,67E-09	2,46E-08	9,57E-10
Terrestrial acidification	species.yr	8,25E-12	2,31E-11	7,58E-12	3,23E-12	4,41E-12	6,56E-11	2,55E-12
Freshwater eutrophication	species.yr	9,18E-12	2,57E-11	8,43E-12	3,52E-12	4,82E-12	7,29E-11	2,84E-12
Terrestrial ecotoxicity	species.yr	1,51E-11	4,24E-11	1,39E-11	5,79E-12	8,05E-12	1,20E-10	4,68E-12
Freshwater ecotoxicity	species.yr	9,35E-13	2,62E-12	8,58E-13	3,58E-13	4,92E-13	7,42E-12	2,89E-13
Marine ecotoxicity	species.yr	2,86E-15	8,01E-15	2,62E-15	1,09E-15	1,51E-15	2,27E-14	8,83E-16
Agricultural land occupation	species.yr	7,71E-11	2,16E-10	7,08E-11	2,96E-11	4,07E-11	6,13E-10	2,38E-11
Urban land occupation	species.yr	3,86E-11	1,08E-10	3,54E-11	1,48E-11	2,04E-11	3,07E-10	1,19E-11
Natural land transformation	species.yr	3,08E-10	8,65E-10	2,83E-10	1,18E-10	1,64E-10	2,45E-09	9,54E-11
Metal depletion	\$	0,00185769	0,0052076	0,0017065	0,00071383	0,00098591	0,01478681	0,00057443
Fossil depletion	\$	4,4708218	12,551151	4,1128136	1,7148928	2,386151	35,598624	1,3845246

FIGURE 6.23 – Vue d'un onglet de résultats de l'évaluation environnementale.

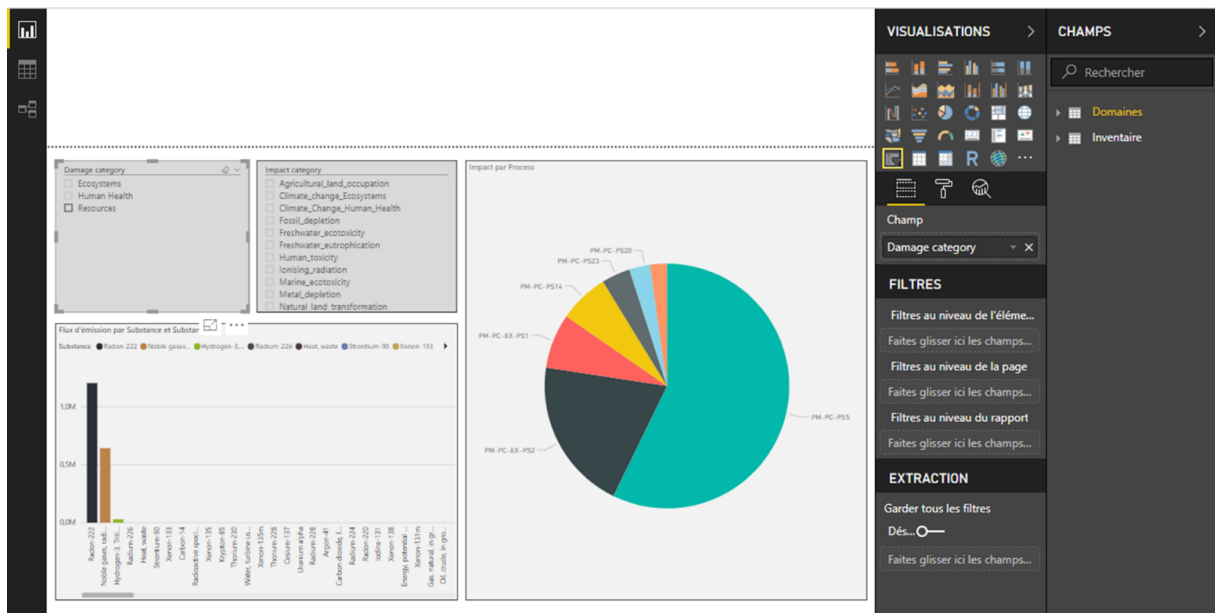


FIGURE 6.24 – Vue d'une visualisation interactive des résultats de l'évaluation environnementale.

6.6 Limites du cas d'étude

Ce cas d'étude présente différentes limites. Les données sont issues d'articles scientifiques générés par des laboratoires. L'ACV est réalisée à une échelle TRL de recherche (échelle TRL 1 ou 2) ce qui peut conduire à un changement d'échelle si nous voulons envisager un pilote industriel. De plus, les procédés n'étant pas continus dans les laboratoires, les données ont dû être traitées pour que les entrées d'une étape du procédé correspondent aux sorties de l'étape précédente. Le traitement des données peut être amélioré, par exemple en ajoutant les changements d'échelle effectués dans l'ingénierie des procédés. Une autre limite des données concerne l'abondance et la qualité qui peuvent être très faibles pour certains procédés innovants et nouveaux tels que les procédés de prétraitement de la biomasse. Il serait intéressant pour les études futures de changer l'unité fonctionnelle et d'élargir les limites de l'étude pour prendre en compte la logistique globale ou pour considérer le recyclage des effluents. Tous les impacts sont liés au produit final (glucose) et aucun recyclage des effluents n'est pris en compte. Pourtant, le recyclage pourrait réduire les impacts de certains procédés plus que d'autres. Cette approche sert à aider à décider quels procédés ou biomasses ne doivent pas être utilisés ou quelles voies pourraient être empruntées. Une dernière limitation provient du principal intrant, à savoir la biomasse, un déchet de l'agriculture. Elle peut être de qualité variable en fonction du temps et du stockage. En outre, il est nécessaire d'évaluer sa durabilité et son approvisionnement, qui peuvent également varier dans le temps. Elle est aussi considérée comme un déchet avec un impact nul (tous les impacts sont mis sur le produit agricole final). Il faudrait faire une pondération de ces impacts entre le produit final et la biomasse.

6.7 Conclusion du chapitre 6

Ce chapitre a présenté l'application de notre cadre méthodologique au travers du cas d'étude du prétraitement de la biomasse pour la production de glucose. Le cas d'étude a été présenté en définissant la biomasse lignocellulosique et le concept de bioraffinerie. Le cadre méthodologique est exécuté avec les données de notre cas d'étude en se servant d'un outil développé grâce à une RTO (@Web) et le développement d'un outil interne.

La première contribution de ce chapitre réside dans une réponse à la question Q3 de nos travaux. Le cas d'étude rentre dans le contexte d'EC en exploitant les déchets agricoles pour en faire un produit. Cette valorisation de biomasses lignocellulosiques en glucose se fait à travers des procédés de transformation. Le cadre méthodologique appliqué au cas d'étude permet de valoriser les données de la littérature scientifique pour aider les chercheurs à prendre une décision sur le choix d'un enchaînement d'opérations unitaires et de biomasses le plus durable possible.

La deuxième contribution est la mise en exergue du nombre de données traitées et de la complexité de traitement. Les méthodes issues du ML que ce soient les méthodes d'apprentissage supervisé comme les RN pour le traitement des données procédés ou les méthodes de réduction de la dimension et de clustering pour l'analyse et la visualisation des résultats des analyses de la durabilité diminuent cette complexité.

La dernière contribution est une vision sur les limites du cas d'étude et de notre cadre pour des procédés innovants et nouveaux. Ces limites donnent lieu à des perspectives pour améliorer notre cadre et diversifier les cas d'étude possibles. Des cas d'études avec des données temps réelles, plus simple à trouver et en de nombreuses quantités sont des perspectives pour tester l'automatisation de notre cadre et sa flexibilité.

Conclusion générale et perspectives des travaux

Ces dernières années ont vu l'émergence de plans d'action pour le climat, le développement durable ou l'environnement. L'un des principaux plans d'action propose d'utiliser le modèle d'économie circulaire pour simuler l'utilisation de modèles durables. Ces travaux, labélisés par le conseil scientifique de l'EUR BioEco, entre dans ce contexte d'économie circulaire et de bioéconomie. La bioéconomie se définit comme la production de ressources biologiques renouvelables et leur conversion en denrées alimentaires, aliments pour animaux, produits biosourcés et bioénergie. Spécialement, le cas d'étude sur les prétraitements des déchets lignocellulosiques pour une bioraffinerie s'inscrit dans ces contextes scientifiques et sociétaux.

Il ne fait plus de doute que l'évolution de l'industrie grâce à l'intelligence artificielle et aux données massives entrent dans une nouvelle ère croisant transformation digitale et transition écologique. Pour suivre cette évolution, les systèmes et les procédés industriels se doivent d'être éco-conçus et durables. Au-delà de l'amélioration des analyses de la durabilité des systèmes et procédés industriels, c'est **toute une approche centrée sur les données massives et l'intelligence artificielle pour l'aide à la décision des systèmes et procédés industriels durables et circulaires qui est recherchée dans les travaux présentés tout au long de ce manuscrit**. La figure 6.25 reprend la figure 1.10 présentant les questions scientifiques de la thèse et l'enrichit des contributions et des chapitres dans lesquelles elles sont explicitées ainsi que les actions de dissémination avec actes référencés (annexe Production Scientifique). La figure 6.26 reprend ces contributions et les positionne sur la feuille de route du manuscrit.

L'étude de la littérature dans le domaine des données massives et de l'intelligence artificielle appliquée à l'industrie amène à considérer l'ingénierie des connaissances (ontologies), l'apprentissage automatique (supervisé et non supervisé) et le bénéfice des données massives. Les travaux s'appliquent à la définition de l'approche et la création d'un cadre méthodologique centré sur les données massives et l'intelligence artificielle pour des systèmes industriels durables et circulaires. L'approche et le cadre sont décrits dans cinq chapitres et sont organisés autour des deux grandes questions (Q1 et Q2). Les méthodes - notamment la RTO - de l'ingénierie des connaissances sont appliquées pour extraire des données massives issues des articles scientifiques. Ces données, organisées sémantiquement, permettent aux chercheurs et ingénieurs d'enrichir leurs connaissances et projets. Elles constituent les données de premier plan pour les analyses de la durabilité. C'est une contribution de ces travaux l'**adaptation des méthodes d'ingénierie des connaissances pour**

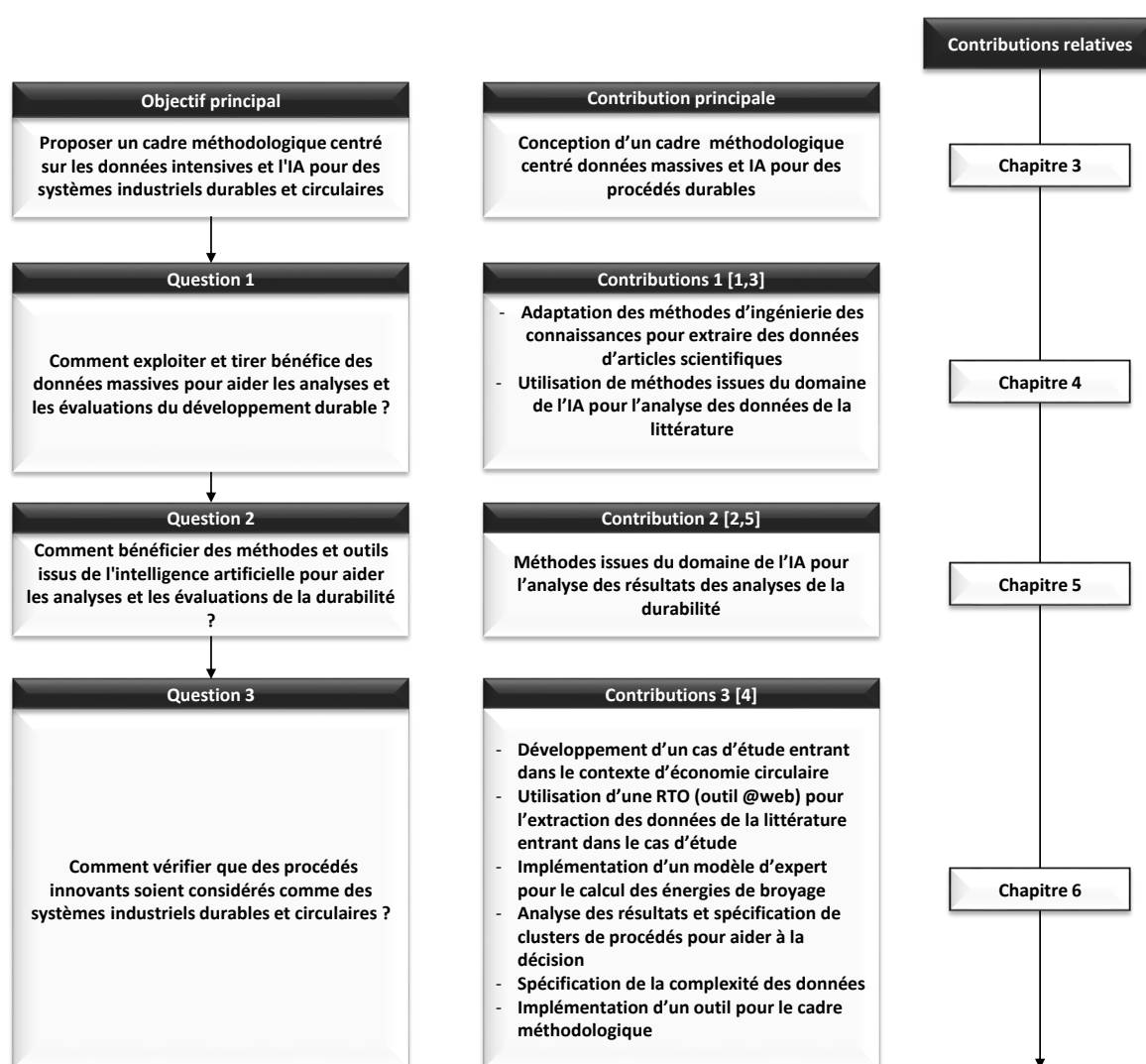


FIGURE 6.25 – Questions scientifiques, contributions et chapitres associés.

extraire des données d'articles scientifiques et des données massives.

Avant d'utiliser la base de données créée dans les analyses de la durabilité, un traitement de ces données est nécessaire. Un enrichissement grâce à des modèles issus d'experts est pris en compte. Les méthodes d'apprentissage automatique tels que les RN assurent une très bonne prédiction du *glucose yield*. Ces modèles de régression captent la relation non-linéaire entre le rendement de glucose et les variables (comme le taux de glucose de la biomasse, la température maximale, ...). Des outils de sensibilité globale des modèles ML, tels que PDP et ICE, sont intégrés afin d'enrichir le pouvoir interprétatif du modèle RN. **L'utilisation de ces outils appliqués dans le contexte de la durabilité des procédés permet d'interpréter les données procédés avant d'effectuer une analyse de la durabilité** et constitue une contribution de ces travaux. Une **spécification de la complexité des données** est développée et l'ajout de trois niveaux de granularité de ces données pour simplifier les calculs et leur temps d'exécution est proposé. Les résultats des analyses de la durabilité constituent une matrice trois dimensions dont tous les éléments sont des k-uplets différents. Les

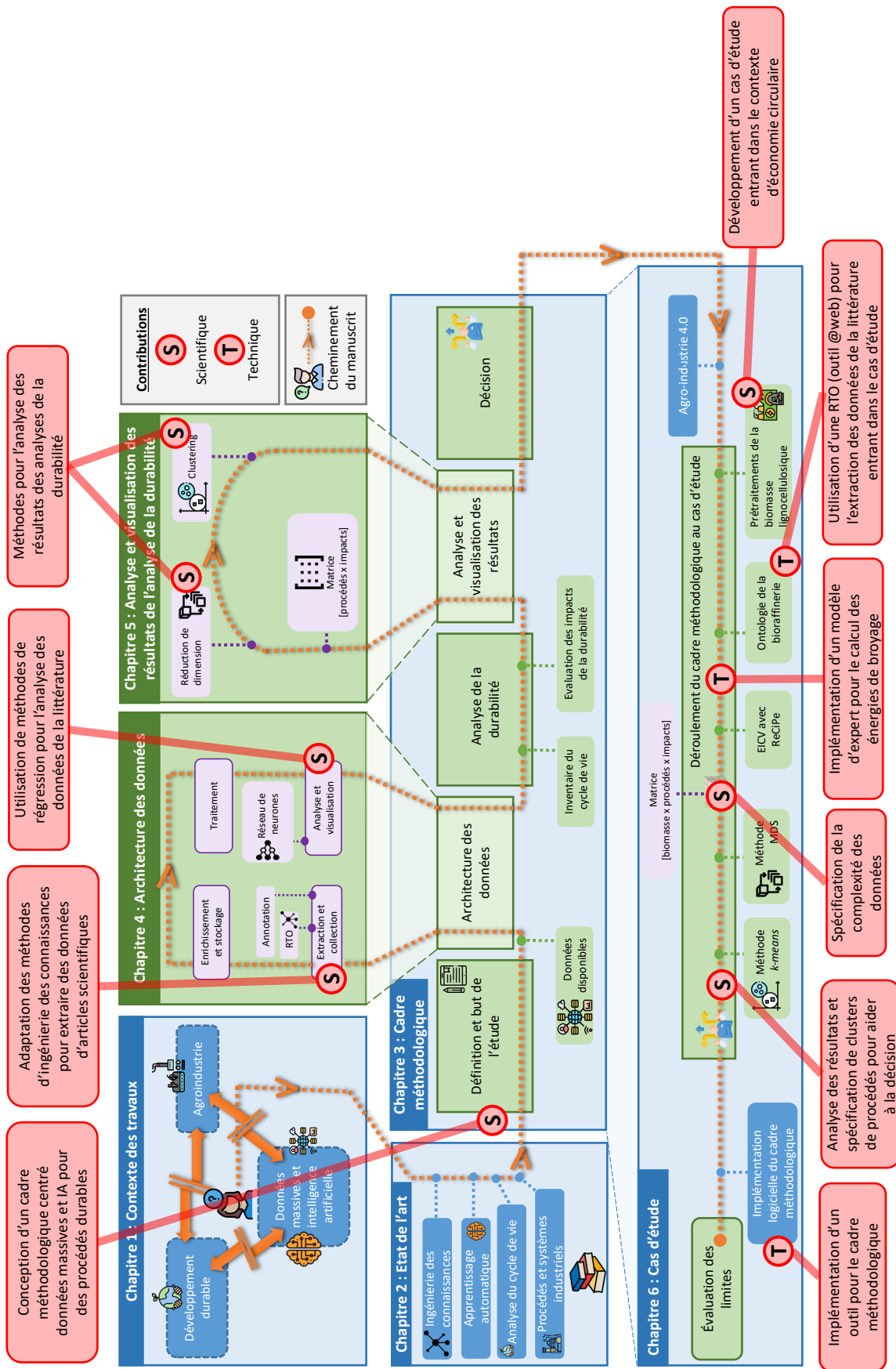


FIGURE 6.2.6 – Feuille de route du manuscrit, augmentée des contributions scientifiques (S) et techniques (T) apportées par les travaux de thèse.

trois dimensions étant le type de procédés, le type de biomasse (dans notre cas d'étude) et le type d'impacts (dépendant de la méthode de calcul des impacts et du niveau d'agrégation), cela engendre une matrice de grande dimension difficilement analysable. Une méthode hybride couplant des méthodes de réduction de dimension et de *clustering* permet de **visualiser des clusters sur ces résultats et faciliter l'analyse et la prise de décision**. Cela constitue une contribution de nos travaux. Le cadre méthodologique propose plusieurs boucles d'itération : le niveau de granularité, le niveau d'agrégation des méthodes de calcul des impacts et le nombre de méthodes utilisées.

Dans l'objectif de tester le cadre présenté, un cas d'étude concernant la comparaison de procédés de prétraitement de différentes biomasses lignocellulosiques pour produire du glucose est développé. Ce cas d'étude met en avant certaines limites au cadre méthodologique. Les améliorations du cadre répondant à ces limites sont explicitées dans les perspectives. De plus, une version logicielle du cadre a été proposée en fin de manuscrit. L'outil développé, s'appuie sur différents outils comme @web, Excel-VBA, R, Orange et PowerBI. Il est important d'insister sur le fait que l'outil n'est qu'une aide de preuve de concept du cadre méthodologique. L'automatisation du cadre n'est alors pas possible avec cet outil. Un prototype logiciel plus intégré permettrait d'avoir une automatisation du cadre ainsi qu'une généricité de l'outil. Cela fait partie des perspectives à long termes de ces travaux.

Perspectives à court terme

P1 – Apport de la sphère économique dans le cas d'étude

Dans le cas d'étude, seuls les impacts environnementaux sont pris en compte. Les travaux (immatures à ce jour) sur l'intégration de la dimension économique ont commencé mais ne sont pas présentés ici. Plusieurs niveaux d'intégration de cette sphère sont à prendre en compte.

Collection des données Contrairement aux données procédés extraites des articles scientifiques, les données permettant de faire une analyse des coûts du cycle de vie sont extraites de différents types de sources et de différentes natures plus ou moins faciles à identifier et manipuler. Certains coûts à prendre en compte sont présentés dans la figure 6.27. Par exemple, les coûts des matières premières sont extraits soit des données de marchés soit de divers fournisseurs. Les coûts de maintenance dépendent des types de machines et de gestion. Suivant les hypothèses, les coûts de construction sont aussi à estimer. Des méthodes provenant du génie des procédés permettent d'estimer ces coûts.

Impact ou impacts économiques L'AcCV permet de calculer un coût global. Comme avec les méthodes de calcul des impacts environnementaux, il est possible de fixer des coûts intermédiaires. Dans notre cas d'étude, les données restent insuffisantes pour une telle analyse.

Intégration des impacts économiques aux impacts environnementaux Selon le chapitre 2, l'AdCV pris en compte dans notre cadre méthodologique revient à « sommer » les impacts économiques et environnementaux. Or cette « somme » peut être pondérée ou non suivant l'importance attribuée à la sphère économique et environnementale. Cette pondération intervient avant l'analyse des résultats d'impacts ce qui induit une boucle d'itération supplémentaire.

P2 – Approfondissement du cas d'étude

Changement d'unité fonctionnelle Le changement d'unité fonctionnelle en "g de biomasse" au lieu de "g de glucose" permet une autre vision. Ceci ajoute aussi une boucle à notre cas d'étude augmentant encore la complexité de l'analyse.

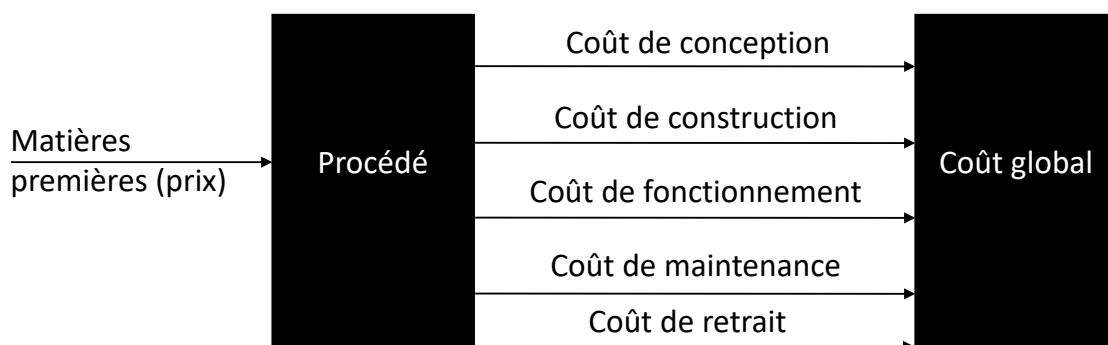


FIGURE 6.27 – Différents coûts nécessaires à une analyse des coûts du cycle de vie

Prise en compte de l'échelle logistique globale Des étapes de transport suivant l'emplacement des bioraffineries et des champs sont ajoutés. L'étape amont (étape agricole) est décrite dans le paragraphe suivant. Les étapes avalées comme la valorisation du glucose en différents matériaux et la fin de vie de la bioraffinerie sont aussi à envisager tout comme des étapes de recyclage.

Pondération des impacts environnementaux de la biomasse La biomasse est considérée comme un déchet de l'agriculture avec des impacts environnementaux nuls. Tous les impacts sont attribués au produit agricole final. Une pondération de ces impacts entre le produit final et la biomasse-déchet permet de faire une différenciation entre les différentes biomasses comparées et d'avoir une analyse étendue.

Amélioration de l'outil logiciel Des méthodes calculant la fiabilité des articles sont mises en place, mais leur exploitation dans l'outil n'est pas optimal. L'ajout d'un impact de fiabilité est envisagé pour compléter l'analyse. La qualité des données des bases de données d'arrière-plan a besoin d'être exploitée par l'outil.

P3 – Ajout de méthodes automatiques de prise de décision

Les décisions sont prises manuellement dans notre cas d'étude et ne sont pas intégrées à notre outil. L'aide d'un outil d'aide à la décision utilisant des modèles multicritères, tel que des modèles de tri (ELECTRE, PROMETHEE) ou des outils d'agrégation [BOUYSSOU, 2006] ou des méthodes de prise de décision en groupe mixte, telle que Delphi-SWOT, peuvent aussi être mises en place [TAVANA et al., 2012] dans notre outil et notre cas d'étude.

Perspectives à moyen terme

P4 – Apport d'un deuxième cas d'étude

Un deuxième cas d'étude traitant un procédé industriel autre que la bioraffinerie ou un domaine différent du génie des procédés testerait le cadre méthodologique.

P5 – Changement d'échelle

Les données sont issues d'articles scientifiques générés par des laboratoires. L'ACV est réalisée à une échelle TRL 1 ou 2. Ceci conduit à un changement d'échelle pour accroître le niveau de maturité technologique. Les procédés n'étant pas continus dans les laboratoires, les données ont dû être traitées pour que les entrées d'une étape du procédé correspondent aux sorties de l'étape précédente. Le traitement des données peut être amélioré, par exemple en ajoutant les changements d'échelle effectués dans l'ingénierie des procédés.

P6 – Ajout d'autres types d'analyse environnementale

Dans notre cadre méthodologique, seule une AeCV attributionnelle est effectuée. Une AeCV conséquentielle, dynamique ou territoriale peuvent venir compléter l'approche et le cadre.

Perspectives à long terme

P7 – Propagation des erreurs et des incertitudes dans le cadre méthodologique

La littérature sur la propagation des erreurs et des incertitudes dans les analyses de la durabilité est encore jeune. Des méthodes existent et peuvent être intégrées dans notre cadre méthodologique [CELLURA et al., 2011] . La prise en compte de ces incertitudes améliore l'analyse des résultats et influence la prise de décision.

P8 – Création d'un outil intégré

La création d'un outil totalement intégré augmente l'automatisation de notre cadre méthodologique et sa généralité. Un outil totalement générique n'est pas envisageable car il y a trop de domaines différents dans les systèmes industriels. Un outil modulable suivant le besoin avec des parties plus ou moins génériques est possible.

Annexes

Annexe A

Descriptions des principales évaluations environnementales.

Cette annexe fournit un descriptif des analyses environnementales spécifiées dans le tableau 2.1, afin de relever les avantages et les inconvénients de chaque méthode.

A.1 Analyse environnementale du cycle de vie

A.1.1 Description et points clés :

- Identifier et quantifier les flux de processus et les systèmes qui contribuent fortement à la dégradation de l'environnement. Le champ d'application de l'évaluation englobe l'extraction et le traitement des matières premières, les processus de fabrication et d'assemblage, la distribution des produits, leur utilisation, leur réutilisation, l'entretien, le recyclage et l'élimination finale.
- Régie et guidée par la norme ISO.
- Elle comprend quatre étapes principales : définition de l'objectif et définition du but et de la portée, analyse de l'inventaire, évaluation des impacts et l'interprétation.

A.1.2 Avantages :

- Prendre en compte une série de catégories d'impact environnemental.
- Une approche des impacts environnementaux non spécifique à un site est nécessaire.
- Planification stratégique à long terme
- Identifier les domaines d'amélioration qui auront la plus grande influence sur les impacts totaux du cycle de vie.

A.1.3 Inconvénients :

- Le manque de données peut limiter les conclusions que l'on peut tirer d'une étude spécifique.
- Les risques intrinsèques des processus ne sont pas abordés.
- Les données de l'inventaire du cycle de vie doivent être utilisées avec précaution, car les processus de production diffèrent d'un pays à l'autre ou d'un fournisseur à l'autre.

Sources : [FINNVEDEN et al., 2009; GUINÉE et al., 2011; PENNINGTON et al., 2004]

A.2 Analyse des flux de matières

A.2.1 Description et points clés :

- Fournit un ensemble complet et cohérent d'informations sur tous les flux et stocks d'un matériau particulier au sein d'un système.
- Se base sur deux principes fondamentaux : l'approche système et le bilan massique.
- Permet de créer une liste des quantités des différents flux.
- Comprend généralement quatre étapes : définition des objectifs et des systèmes, analyse de la chaîne de processus, comptabilisation et équilibrage, modélisation et évaluation.

A.2.2 Avantages :

- L'analyse peut être appliquée à l'échelle industrielle, nationale ou mondiale.
- Elle permet de connaître l'historique de la pollution et de la consommation des ressources dans une zone ou une région.
- Le bilan massique permet de connaître le stock de matière ignoré ou sous-estimé.

A.2.3 Inconvénients :

- Se concentre sur un seul matériau.
- Pour déterminer les incidences sur l'environnement, une ACV ou une analyse des exergies est nécessaire.

Sources : [AYRES et AYRES, 2002; BRUNNER et RECHBERGER, 2004]

A.3 Analyse des impacts environnementaux

A.3.1 Description et points clés :

- Analyser et évaluer les impacts que les activités humaines peuvent avoir sur l'environnement.
- Leurs principaux objectifs sont d'envisager toutes les questions environnementales et socio-économiques possibles associées au projet proposé, de manière qualitative et quantitative, et de fournir aux décideurs les informations qui en résultent.
- Cette approche utilise des indices environnementaux.
- Comporte trois grandes phases : la sélection et la délimitation du projet, l'évaluation des incidences sur l'environnement, et la prise de décision et l'examen.

A.3.2 Avantages :

- Les dommages environnementaux du projet peuvent alors être minimisés et les avantages environnementaux éventuels identifiés.
- C'est un processus systématique.

A.3.3 Inconvénients :

- Il faut définir un site et un moment précis pour estimer les impacts environnementaux.
- Seuls les impacts directs qui se situent dans les limites du système étudié sont analysés.
- Une analyse rigoureuse et quantitative des données est souvent nécessaire pour donner un sens aux grandes quantités de données.
- Il s'agit davantage d'une procédure légale que d'un outil d'évaluation environnementale détaillée.

Sources : [BURGESS et BRENNAN, 2001; MANUILOVA et al., 2009]

A.4 Analyse des risques environnementaux

A.4.1 Description et points clés :

- Implique l'estimation et l'évaluation du risque pour l'environnement causé par la relation entre le niveau d'application d'un contaminant ou d'une activité et l'apparition d'un effet ou d'un événement indésirable.
- L'impact peut concerner les humains, la flore ou la faune, mais aussi les bâtiments, le sol, l'eau ou l'air.
- Comprend les étapes suivantes : formulation du problème, caractérisation du danger, caractérisation de l'exposition, évaluation du risque.

A.4.2 Avantages :

- Les législations sur la protection de l'environnement sont motivées par cette méthode.
- Facilite la prise de décision grâce à l'identification des effets non désirés.
- Prend en compte le contexte du système étudié.

A.4.3 Inconvénients :

- Le temps n'est pas explicitement inclus.
- Le montant des impacts environnementaux générés par un processus ou un service n'est pas calculé.
- La différence entre risque et situation dangereuse doit être claire.
- Difficulté d'établir une échelle de mesure du risque.

Sources : [AYRES et AYRES, 2002; BURGESS et BRENNAN, 2001; CALOW, 1998]

A.5 Analyse des exergies

A.5.1 Description et points clés :

Mesurer la quantité maximale de travail qui peut être obtenu en amenant une ressource en équilibre avec son environnement par un processus réversible.

A.5.2 Avantages :

Fournir des informations pour identifier l'emplacement, les ressources et causes des problèmes de déviation du système idéal en équilibre.

A.5.3 Inconvénients :

- Se concentre uniquement sur un point de vue thermodynamique.
- Complexité dans la définition de l'état de référence (système idéal en équilibre).

Sources : [BRUNNER et RECHBERGER, 2004; SZARGUT, 2005]

A.6 Apport de matière par unité de service

A.6.1 Description et points clés :

- Mesure le flux massique total de matière causé par la production, la consommation et l'élimination des déchets d'une unité de service définie (fabrication d'une cuisine, cycle de lavage d'un lave-vaisselle...) ou d'un produit.
- Utilise uniquement les flux d'entrée (considérant que les flux d'entrée sont égaux aux flux de sortie).
- L'apport de matières est calculé en cinq catégories : matières premières abiotiques, matières premières biotiques, eau, érosion et air.
- Toute la consommation de matériaux pendant la fabrication, l'utilisation et le recyclage ou l'élimination est ramenée à la consommation de ressources à l'aide de facteurs de calcul simples exprimés en kg ou en tonne.
- L'objectif est de réduire le montant total de l'addition de tous les flux de matière et d'énergie exprimés en kg ou en tonne.

A.6.2 Avantages :

- Permet de comparer la consommation de ressources de différentes solutions pour produire le même service.
- Révèle l'ampleur de l'utilisation des ressources tout au long du cycle de vie et permet de concentrer les efforts sur les phases les plus significatives pour réduire la charge environnementale du produit.
- Mesurer la matière et l'énergie dans la même unité constitue un excellent outil de communication pour identifier les principaux problèmes.

A.6.3 Inconvénients :

Ne tient pas compte de l'écotoxicité des matériaux ou de la biodiversité.

Sources : [BRUNNER et RECHBERGER, 2004; RITTHOFF et al., 2002]

Annexe B

Descriptif détaillé de la RTO utilisée pour extraire les données de la littérature.

Cette annexe fournit un descriptif plus détaillé de la RTO définie dans les paragraphes 4.2.2.2 et 6.4.2.1. Une explication vient accompagnée la figure B.1 qui présente la RTO de la bioraffinerie.

La RTO est composée en deux parties : une partie conceptuelle et une partie terminologique. La figure B.1 présente ses deux parties, leur détail ainsi que des exemples issus de la RTO *biorefinery* de notre cas d'étude.

Composante ontologique Deux ontologies principales la composent : l'ontologie noyau qui représente les relations n-aires et l'ontologie de domaine qui représente les concepts spécifique à la bioraffinerie.

L'ontologie noyau est elle même composée deux parties, le noyau supérieur et le noyau inférieur. Ce découpage de l'ontologie noyau assure la généricité de la RTO. Elle peut être utilisée pour des données quantitatives expérimentales, mais aussi pour représenter des relations n-aires entre n'importe quels arguments. Dans le noyau supérieur, on retrouve les concepts de *Relation* et *Argument* qui représentent comme leur nom l'indique les relations n-aires et leurs arguments. Dans l'ontologie noyau inférieur, on trouve les concepts génériques suivants :

- *Dimension* : il contient les dimensions qui classifient les quantités et leurs unités de mesures associées. Par exemple, le concept quantité *rotation speed* et ses unités de mesure sont rattachés à la dimension *speed*.
- *UM_Concept* : il gère les conversions entre unités de mesure.
- *Unit* : il gère les concepts d'unités de mesure. Il possède 4 sous-concepts *Singular_Unit*, *Unit_Division_Or_Multiplication*, *Unit_Multiple_Or_Submultiple* et *Unit_Exponentiation*. Les unités sont des instances de ces sous-concepts et chaque instance est associée à une partie terminologique. Par exemple, m^3 dénote l'instance d'unité *volume_Meter* qui appartient au sous-concept *Unit_Exponentiation*.
- *Quantity_Concept* : il gère les arguments de type quantitatif et se subdivise en sous-concepts en fonction du domaine étudié. Ces sous-concepts représentent les arguments quantitatifs des re-

lations n-aires représentées dans la RTO. Par exemple, si on reprend la relation de la figure *Milling solid qty output Relation* et dont on retrouve la relation n-aire dans la figure 6.10, deux sous-concepts sont utilisés *BiomassQty* et *PretreatmentEnergy* signifiant respectivement la quantité de biomasse et l'énergie du prétraitement. Ces sous-concepts sont retrouvés dans d'autres relations.

- *Symbolic_Concept* : il gère les arguments non numériques de la relation n-aire. Toujours avec la même relation, on retrouve deux sous-concepts le traitement (*treatment*) et la biomasse (*biomasse*) qui sont des argument à notre relation.

L'ontologie de domaine contient les concepts spécifiques à la bioraffinerie. Nos exemples précédents concernant la relation *Milling solid qty output Relation* font partie de cette ontologie. Tous les concepts spécifiques sont modélisés comme des sous-concepts des concepts génériques de l'ontologie noyau. Ainsi, le traitement (*treatment*) et la biomasse (*biomasse*) sont des sous-concepts de *Symbolic-concept*. Ces sous-concepts sont organisés en hiérarchie pour être le plus précis possible et apporter une certaine connaissance au domaine étudié.

Composante terminologique Elle contient l'ensemble des termes du domaine étudié. Cette composante est très importante car elle complète la terminologie du domaine d'informations jugées pertinentes. Elle contient notamment tous les synonymes, les abréviations d'un terme ainsi que ses équivalents dans d'autres langues.

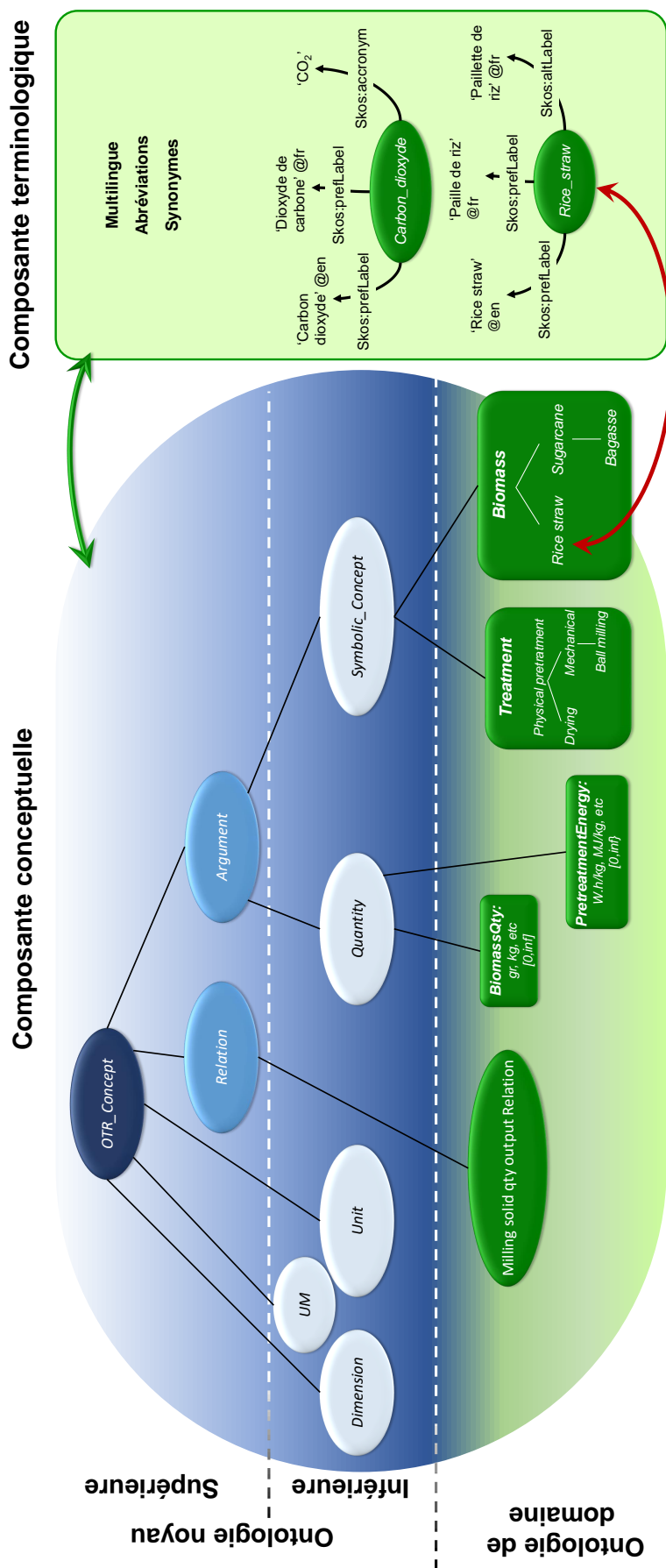


FIGURE B.1 – Descriptif de la ressource ontologique et terminologique spécialisée dans la bioraffinerie

Annexe C

Liste des articles sélectionnés à titre d'experts avant traitement.

Cette annexe fournit la liste complète des articles scientifiques sélectionnés à titre d'expert avant le traitement des données et le choix des biomasses et voies technologiques.

TABLEAU C.1 – Liste des articles sélectionnés à titre d'experts avant traitement des procédés de type PM.

Référence type procédés	Auteurs	Titre de l'article	Année	Journal
PM1	Ilgook Kima, Bomi Leea, Ji-Yeon Parkb, Sun-A. Choib, Jong-In Han	Effect of nitric acid on pretreatment and fermentation for enhancing ethanol production of rice straw	2014	Carbohydrate Polymers
PM2	Zhang, Hongdan, Wu, Shubin	Enhanced enzymatic cellulose hydrolysis by subcritical carbon dioxide pretreatment of sugarcane bagasse	2014	Bioresour Technol
PM3	Poornejad, N, Karimi, K, Behzad, T	Improvement of saccharification and ethanol production from rice straw by NMMO and [BMIM][OAc] pretreatments	2013	Industrial Crops and Products
PM4	Meijuan Zeng, Nathan S. Mosier, Chia-Ping Huang, Debra M. Sherman, Michael R. Ladisch	Microscopic Examination of Changes of Plant Cell Structure in Corn Stover Due to Hot Water Pretreatment and Enzymatic Hydrolysis	2007	Biotechnology and Bioengineering
PM5	Hiroyuki Inoue, Takashi Endo, Ayla Sant'Ana da Silva, Shinichi Yano, Elba P S Bon	Milling pretreatment of sugarcane bagasse and straw for enzymatic hydrolysis and ethanol fermentation	2010	Bioresour Technol
PM6	Hamid Amiri, Keikhosro Karimi, Hamid Zilouei	Organosolv pretreatment of rice straw for efficient acetone, butanol, and ethanol production	2014	Bioresour Technol
PM7	Xiao-le Liu, Chengyu Dong, Shao-Yuan Leu, Zhen Fang, Zheng-diao Miao	Efficient saccharification of wheat straw pretreated by solid particle-assisted ball milling with waste washing liquor recycling	2022	Bioresour Technol
PM8	Yong Cheol Park, Tae Hyun Kim, Jun Seok Kim	Effect of Organosolv Pretreatment on Mechanically Pretreated Biomass by Use of Concentrated Ethanol as the Solvent	2017	Biotechnology and Bioprocess Engineering
PM9	Tianjiao Qu, Ximing Zhang, Xingwei Gu, Lujia Han, Guanya Ji, Xueli Chen, Weihua Xiao	Ball Milling for Biomass Fractionation and Pretreatment with Aqueous Hydroxide Solutions	2017	ACS Sustainable Chemistry & Engineering
PM10	Lianyue Shao, Huan Chen, Yanli Li, Jiannan Li, Guang Chen, Gang Wang	Pretreatment of corn stover via sodium hydroxide-urea solutions to improve the glucose yield	2020	Biotechnology and Bioprocess Engineering
PM11	Haiyan Zhang, Longjian Chen, Junbao Li, Minsheng Lu, Lujia Han	Quantitative characterization of enzyme adsorption and hydrolytic performance for ultrafine grinding pretreated corn stover	2017	Bioresour Technol

TABLEAU C.2 – Liste des articles sélectionnés à titre d'experts avant traitement des procédés de type PM-UFM.

Référence type procédés	Auteurs	Titre de l'article	Année	Journal
PM-UFM1	Lin, Zengxiang Huang, He Zhang, Hongman Zhang, Lin Yan, Lishi Chen, Jingwen	Ball Milling Pretreatment of Corn Stover for Enhancing the Efficiency of Enzymatic Hydrolysis	2010	Applied Biochemistry and Biotechnology
PM-UFM2	A. Licari, F. Monlau, A. Solhy, P. Buche and A. Barakat	Comparison of various milling modes combined to the enzymatic hydrolysis of lignocellulosic biomass for bioenergy production : Glucose yield and energy efficiency	2015	Energy Journal
PM-UFM3	Abdellatif Barakat, Santi Chuetor, Florian Monlau, Abderrahim Solhy, Xavier Rouau	Eco-friendly dry chemo-mechanical pretreatments of lignocellulosic biomass : Impact on energy and yield of the enzymatic hydrolysis	2014	Applied Energy
PM-UFM4	Mominul Islam Sheikh, Chul-Hwan Kim, Hyeon-Jin Park, Sung-Ho Kim, Gyeong-Chul Kim, Ji-Young Lee, Sung-Woong Sima and Jae Won Kimb	Effect of torrefaction for the pretreatment of rice straw for ethanol production	2013	Science of food and agriculture
PM-UFM5	Hiroyuki Inoue, Takashi Endo, Ayla Sant'Ana da Silva, Shimichi Yano, Elba P S Bon Akhiro Hidenou, Hiroyuki Inoue, Kenichiro Tsukahara, Shinji Fujimoto, Tomoaki Minowa, Seiichi Inoue, Takashi Endo, Shigeki Sawayama	Milling pretreatment of sugarcane bagasse and straw for enzymatic hydrolysis and ethanol fermentation	2010	Bioresource Technology
PM-UFM6		Wet disk milling pretreatment without sulfuric acid for enzymatic hydrolysis of rice straw	2009	Bioresource Technology
PM-UFM7	Junbao Li, Haiyan Zhang, Minsheng Lu, Lujia Han	Comparison and intrinsic correlation analysis based on composition, microstructure and enzymatic hydrolysis of corn stover after different types of pretreatments	2019	Bioresource Technology
PM-UFM8	Tianjiao Qu, Ximing Zhang, Xingwei Gu, Lujia Han, Guanya Ji, Xueli Chen, Weihua Xiao	Ball Milling for Biomass Fractionation and Pretreatment with Aqueous Hydroxide Solutions	2017	ACS Sustainable Chemistry & Engineering
PM-UFM9	Haiyan Zhang, Longjian Chen, Junbao Li, Minsheng Lu, Lujia Han	Quantitative characterization of enzyme adsorption and hydrolytic performance for ultrafine grinding pretreated corn stover	2017	Bioresource technology

TABLEAU C.3 – Liste des articles sélectionnés à titre d'experts avant traitement des procédés de type PM-PC-PS (1/3).

Référence type procédés	Auteurs	Titre de l'article	Année	Journal
PM-PC-PS1	A Kallioinen, M Hakola, T Riekkola, T Repo, M Leskela, N von Weymarn, M Siika-aho	A novel alkaline oxidation pretreatment for spruce, birch and sugar cane bagasse	2013	Bioresource Technology
PM-PC-PS2	Krishnan, Chandraraj Sousa, Leonardo da Costa Jin, Mingjie Chang, Linpei Dale, Bruce E. Balan, Venkates	Alkali-based AFEX pretreatment for the conversion of sugarcane bagasse and cane leaf residues to ethanol	2010	Biotechnology and Bioengineering
PM-PC-PS3	Liu, Chen-Guang Liu, Li-Yang Zi, Li-Han Zhao, Xin-Qing Xu, You-Hai Bai, Feng-Wu Magnus Wiman, Dora Dienes, Mads A. T. Hansen, Torbjørn van der Meulen, Guido Zacchi, Gunnar Lid	Assessment and regression analysis on instant catapult steam explosion pretreatment of corn stover	2014	Bioresource technology
PM-PC-PS4	Zacchi, Gunnar Lid	Cellulose accessibility determines the rate of enzymatic hydrolysis of steam-pretreated spruce	2012	Bioresource Technology
PM-PC-PS5	Zhou, Wei Yang, Maohua Wang, Caixia Liu, Jianfei Xing, Jianmin	Changes in plant cell-wall structure of corn stover due to hot compressed water pretreatment and enhanced enzymic hydrolysis	2014	World Journal of Microbiology and Biotechnology
PM-PC-PS6	Hiroyuki Inoue, Takashi Endo, Akihiro Hiden, Takashi Yanagida, Kenichiro Tsukahara, Shigeki Sawayama	Combination of hot compressed water treatment and wet disk milling for high sugar recovery yield in enzymatic hydrolysis of rice straw	2012	Bioresource Technology
PM-PC-PS7	Rosgaard, Lisa Pedersen, Sven Meyer, AnneS	Comparison of Different Pretreatment Strategies for Enzymatic Hydrolysis of Wheat and Barley Straw	2007	Applied Biochemistry and Biotechnology
PM-PC-PS8	A. Maarten J. Kootstra, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, A. Hendrik H. Beftink, Elinor L. Scott, a, Johan P.M. Sanders	Comparison of dilute mineral and organic acid pretreatment for enzymatic hydrolysis of wheat straw	2009	Biochemical Engineering Journal
PM-PC-PS9	Bondesson, Pia-Maria Galbe, Mats Zacchi, Guido	Comparison of energy potentials from combined ethanol and methane production using steam-pretreated corn stover impregnated with acetic acid	2014	Biomass and Bioenergy
PM-PC-PS10	Y N Guragain, J De Coninck, F Husson, A Durand, S K Rakshit	Comparison of some new pretreatment methods for second generation bioethanol production from wheat straw and water hyacinth	2011	Bioresource Technology
PM-PC-PS11	Saha, Badal C. Iten, Loren B. Cotta, Michael A. Wu, Y. Victor	Dilute acid pretreatment, enzymatic saccharification and fermentation of wheat straw to ethanol	2005	Process Biochemistry
PM-PC-PS12	Ilgook Kima, Bomi Leea, Ji-Yeon Parkb, Sun-A. Choib, Jong-In Han	Effect of nitric acid on pretreatment and fermentation for enhancing ethanol production of rice straw	2014	Carbohydrate Polymers

TABLEAU C.4 – Liste des articles sélectionnés à titre d'experts avant traitement des procédés de type PM-PC-PS (2/3).

Référence type procédés	Auteurs	Titre de l'article	Année	Journal
PM-PC-PS13	Mominul Islam Sheikh, Chul-Hwan Kim, Hyeon-Jin Park, Sung-Ho Kim, Gyeong-Chul Kim, Ji-Young Lee, Sung-Woong Sima and Jae Won Kimb	Effect of torrefaction for the pretreatment of rice straw for ethanol production	2013	Science of food and agriculture
PM-PC-PS14	Zhi-Hua Liu, Lei Qin, Feng Pang, Ming-Jie Jin, Bing-Zhi Li, Yong Kang, Bruce E. Dale, Ying-Jin Yuan	Effects of biomass particle size on steam explosion pretreatment performance for improving the enzyme digestibility of corn stover	2013	Industrial Crops and Products
PM-PC-PS15	Gao, Yueshu Xu, Jingliang Zhang, Yu Yu, Qiang Yuan, Zhenhong Liu, Yunyun	Effects of different pretreatment methods on chemical composition of sugarcane bagasse and enzymatic hydrolysis	2013	Bioresource technology
PM-PC-PS16	Hongdan Zhang, Shubin Wu	Enhanced enzymatic cellulose hydrolysis by subcritical carbon dioxide pretreatment of sugarcane bagasse	2014	Bioresource technology
PM-PC-PS17	Saha, Badal C. Cotta, Michael A.	Enzymatic hydrolysis and fermentation of lime pretreated wheat straw to ethanol	2007	Journal of Chemical Technology & Biotechnology
PM-PC-PS19	Saha, Badal C. Cotta, Michael A.	Ethanol Production from Alkaline Peroxide Pretreated Enzymatically Saccharified Wheat Straw	2006	Biotechnology Progress
PM-PC-PS20	Chao Tai, Deepak Keshwani	Impact of pretreatment with dilute sulfuric acid under moderate temperature on hydrolysis of corn stover with two enzyme systems	2014	Applied Biochemistry and Biotechnology
PM-PC-PS21	Poornejad, N, Karimi, K, Behzad, T	Improvement of saccharification and ethanol production from rice straw by NMMO and [BMIM][OAc] pretreatments	2013	Industrial Crops and Products
PM-PC-PS22	Joseph A. Rollin, Zhiguang Zhu, Noppadon Sathitsuksanoh, Y.-H. Percival Zhang	Increasing Cellulose Accessibility Is More Important Than Removing Lignin : A Comparison of Cellulose Solvent-Based Lignocellulose Fractionation and Soaking in Aqueous Ammonia	2011	Biotechnology and Bioengineering
PM-PC-PS23	Bryan D. Bals & Farzaneh Teymouri & Tim Campbell & Mingjie Jin & Bruce E. Dale	Low Temperature and Long Residence Time AFEX Pretreatment of Corn Stover	2011	BIOENERGY RESEARCH
PM-PC-PS24	Meijuan Zeng, Nathan S. Mosier, Chia-Ping Huang, Debra M. Sherman, Michael R. Ladisch	Microscopic Examination of Changes of Plant Cell Structure in Corn Stover Due to Hot Water Pretreatment and Enzymatic Hydrolysis	2007	Biotechnology and Bioengineering
PM-PC-PS25	Hamid Amiri, Keikhosro Karimi, Hamid Zilouei	Organosolv pretreatment of rice straw for efficient acetone, butanol, and ethanol production	2014	Bioresource technology

TABLEAU C.5 – Liste des articles sélectionnés à titre d'experts avant traitement des procédés de type PM-PC-PS (3/3).

Référence type procédés	Auteurs	Titre de l'article	Année	Journal
PM-PC-PS26	Mesa, L., González, E., Cara, C., González, M., Castro, E., Mussatto, S. I.	The effect of organosolv pretreatment variables on enzymatic hydrolysis of sugarcane bagasse	2011	Chemical Engineering Journal
PM-PC-PS27	Akihiro Hiden, Hiroyuki Inoue, Kenichiro Tsukahara, Shinji Fujimoto, Tomoaki Minowa, Seiichi Inoue, Takashi Endo, Shigeki Sawayama	Wet disk milling pretreatment without sulfuric acid for enzymatic hydrolysis of rice straw	2009	Bioresource Technology
PM-PC-PS28	Biswas, Rajib Uellendahl, H. Ahring, B. K.	Wet explosion pretreatment of sugarcane bagasse for enhanced enzymatic hydrolysis	2014	Biomass and Bioenergy
PM-PC-PS29	Junbao Li, Haiyan Zhang, Minsheng Lu, Lujia Han	Comparison and intrinsic correlation analysis based on composition, microstructure and enzymatic hydrolysis of corn stover after different types of pretreatments	2019	Bioresource Technology
PM-PC-PS30	Xiao-le Liu, Chengyu Dong, Shao-Yuan Leu, Zhen Fang, Zheng-diao Miao	Efficient saccharification of wheat straw pretreated by solid particle-assisted ball milling with waste washing liquor recycling	2022	Bioresource Technology
PM-PC-PS31	Yong Cheol Park, Tae Hyun Kim, Jun Seok Kim	Effect of Organosolv Pretreatment on Mechanically Pretreated Biomass by Use of Concentrated Ethanol as the Solvent	2017	Biotechnology and Bioprocess Engineering
PM-PC-PS32	Tianjiao Qu, Ximing Zhang, Xingwei Gu, Lujia Han, Guanya Ji, Xueli Chen, Weihua Xiao	Ball Milling for Biomass Fractionation and Pretreatment with Aqueous Hydroxide Solutions	2017	ACS Sustainable Chemistry & Engineering
PM-PC-PS33	Lianyue Shao, Huan Chen, Yanli Li, Jiannan Li, Guang Chen, Gang Wang	Pretreatment of corn stover via sodium hydroxide-urea solutions to improve the glucose yield	2020	Biotechnology and Bioprocess Engineering

TABLEAU C.6 – Liste des articles sélectionnés à titre d'experts avant traitement des procédés des type PM-PC-EX-PS, PM-PC-UFM et PM-PC-UFM-PS (non sélectionnés).

Référence type procédés	Auteurs	Titre de l'article	Année	Journal
PM-PC-EX-PS1	Chao Liu, Evert van der Heide, Haisong Wang, Bin Li, Guang Yu, Xindong Mu	Alkaline twin-screw extrusion pretreatment for fermentable sugar production	2013	Biotechnology for Biofuels
PM-PC-EX-PS2	Chen, JW Zhang, WG Zhang, HM Zhang, QX Huang, H	Screw extrude steam explosion : A promising pretreatment of corn stover to enhance enzymatic hydrolysis	2014	Bioresource Technology
PM-PC-UFM1	Abdellatif Barakat, Santi Chuetor, Florian Monlau, Abderrahim Solhy, Xavier Rouau	Eco-friendly dry chemo-mechanical pretreatments of ligno-cellulosic biomass : Impact on energy and yield of the enzymatic hydrolysis	2014	Applied Energy
PM-PC-UFM-PS1	Hiroyuki Inoue, Takashi Endo, Akihiro Hiden, Takashi Yanagida, Kenichiro Tsukahara, Shigeki Sawayama	Combination of hot compressed water treatment and wet disk milling for high sugar recovery yield in enzymatic hydrolysis of rice straw	2012	Bioresource Technology

Annexe D

Données expérimentales pour le modèle à titre d'expert pour l'énergie d'un broyage.

Cette annexe fournit un tableau des données expérimentales utilisées pour créer le modèle à titre d'expert du calcul de l'énergie d'un broyage en fonction de la taille de sortie de la biomasse. Les expériences effectuées concernent plusieurs types de broyage qui sont pris en compte dans les énergies maximales et minimales. Elles ont été effectuées sur deux types de biomasse la paille de maïs et la paille de riz. A titre d'expert il est considéré que les énergies sont équivalentes pour les deux autres types de biomasses [TABOUBI, 2019].

TABLEAU D.1 – Énergies de broyage moyenne, minimale et maximale en fonction de la taille de la biomasse en sortie.

Taille de la biomasse (mm)	Énergie moyenne (10^{-2} kwh/kg de biomasse)	Énergie minimale (10^{-2} kwh/kg de biomasse)	Énergie maximale (10^{-2} kwh/kg de biomasse)
$8,0 \times 10^{-1}$	5,05	3,76	6,35
$9,0 \times 10^{-1}$	4,93	3,65	6,20
1,0	4,80	3,54	6,06
1,1	4,68	3,43	5,92
1,2	4,55	3,32	5,78
1,3	4,43	3,21	5,64
1,4	4,30	3,10	5,50
1,5	4,18	2,99	5,36
1,6	4,05	2,87	5,23
1,7	3,93	2,76	5,09
1,8	3,80	2,64	4,96
1,9	3,67	2,52	4,83
2,0	3,55	2,40	4,70
2,1	3,42	2,28	4,57
2,2	3,30	2,16	4,44
2,3	3,17	2,04	4,31
2,4	3,05	1,91	4,19
2,5	2,92	1,79	4,06
2,6	2,80	1,66	3,94
2,7	2,67	1,53	3,82
2,8	2,55	1,40	3,69
2,9	2,42	1,27	3,57
3,0	2,30	1,14	3,46
3,1	2,17	1,01	3,34
3,2	2,05	$8,73 \times 10^{-1}$	3,22
3,3	1,92	$7,37 \times 10^{-1}$	3,11
3,4	1,80	$6,01 \times 10^{-1}$	3,00
3,5	1,67	$4,63 \times 10^{-1}$	2,88
3,6	1,55	$3,24 \times 10^{-1}$	2,77
3,7	1,42	$1,84 \times 10^{-1}$	2,66
3,8	1,30	$4,33 \times 10^{-2}$	2,55
3,9	1,17	0,00	2,44

Annexe E

Visualisations des rendements en glucose prédits par différents modèles.

Cette annexe fournit les paramètres et les visualisations des rendements en glucose prédits en fonction des rendements en glucose réel par le modèle AdaBoost, Gradient Boosting, Random Forest et Régression Linéaire

Les paramètres pour le modèle *AdaBoost* :

- Estimateur de base : arbre
- Nombre d'estimateurs : 1
- Algorithme de classification : SAMMER.R
- Fonction de perte par régression : linéaire

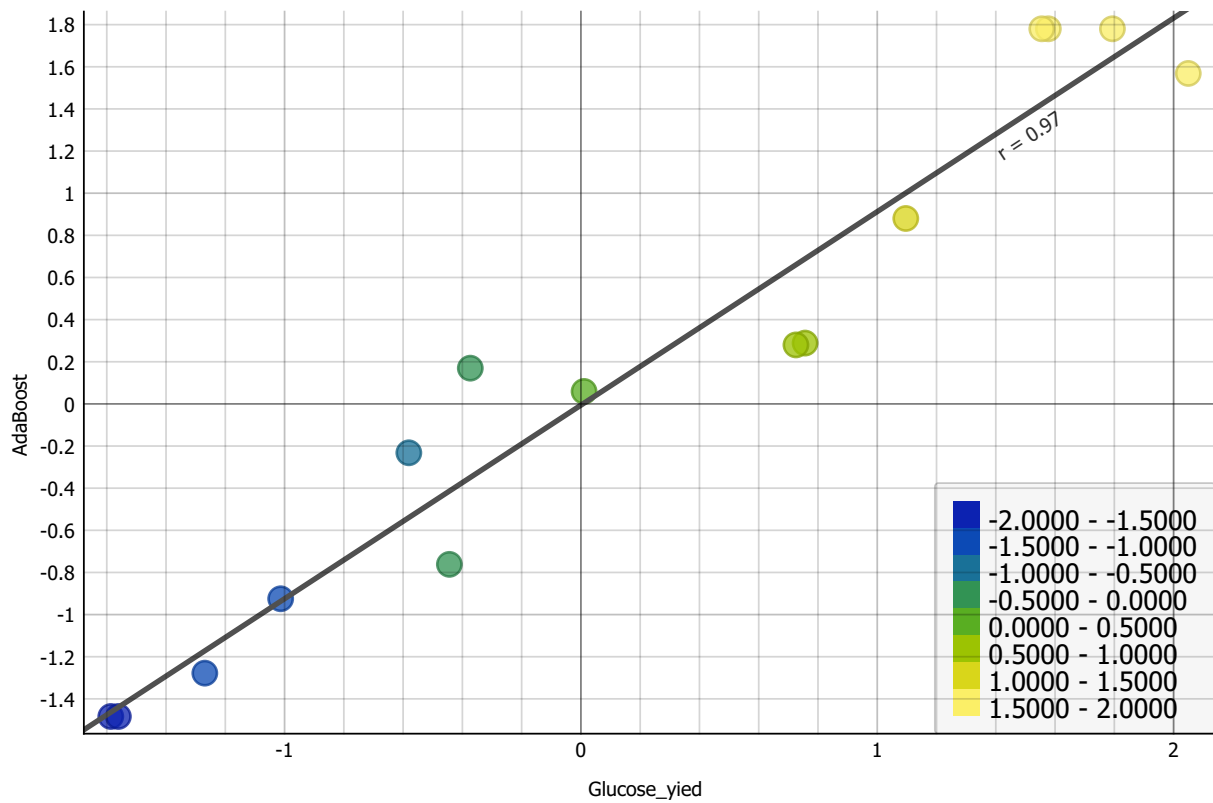


FIGURE E.1 – Visualisation des rendements en glucose prédits par le modèle AdaBoost en fonction des rendements en glucose réel (valeurs normalisées).

Les paramètres pour le modèle *Gradient Boosting* :

- Méthode venant de scikit-learn
- Nombre d'arbres : 100
- Taux d'apprentissage : 0,100

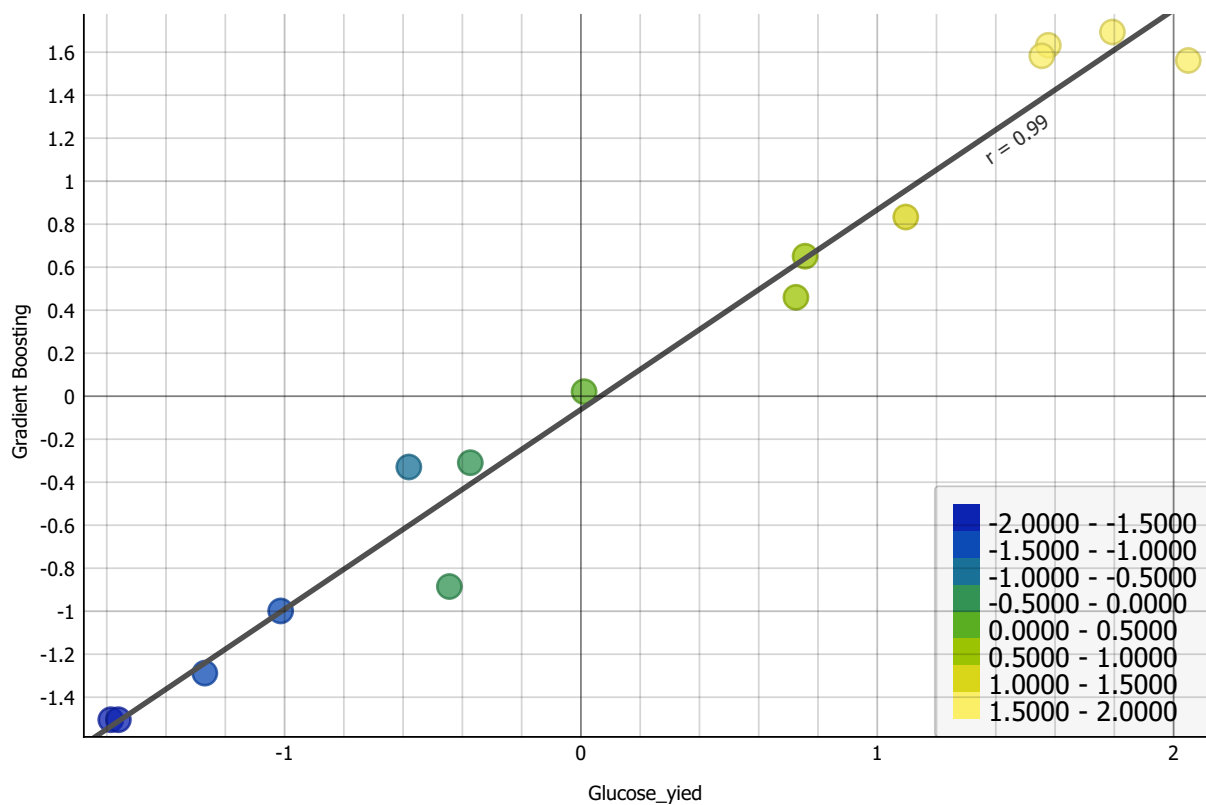


FIGURE E.2 – Visualisation des rendements en glucose prédits par le modèle *Gradient Boosting* en fonction des rendements en glucose réel (valeurs normalisées).

Les paramètres pour le modèle *Random Forest* :

— Nombre d'arbres : 10

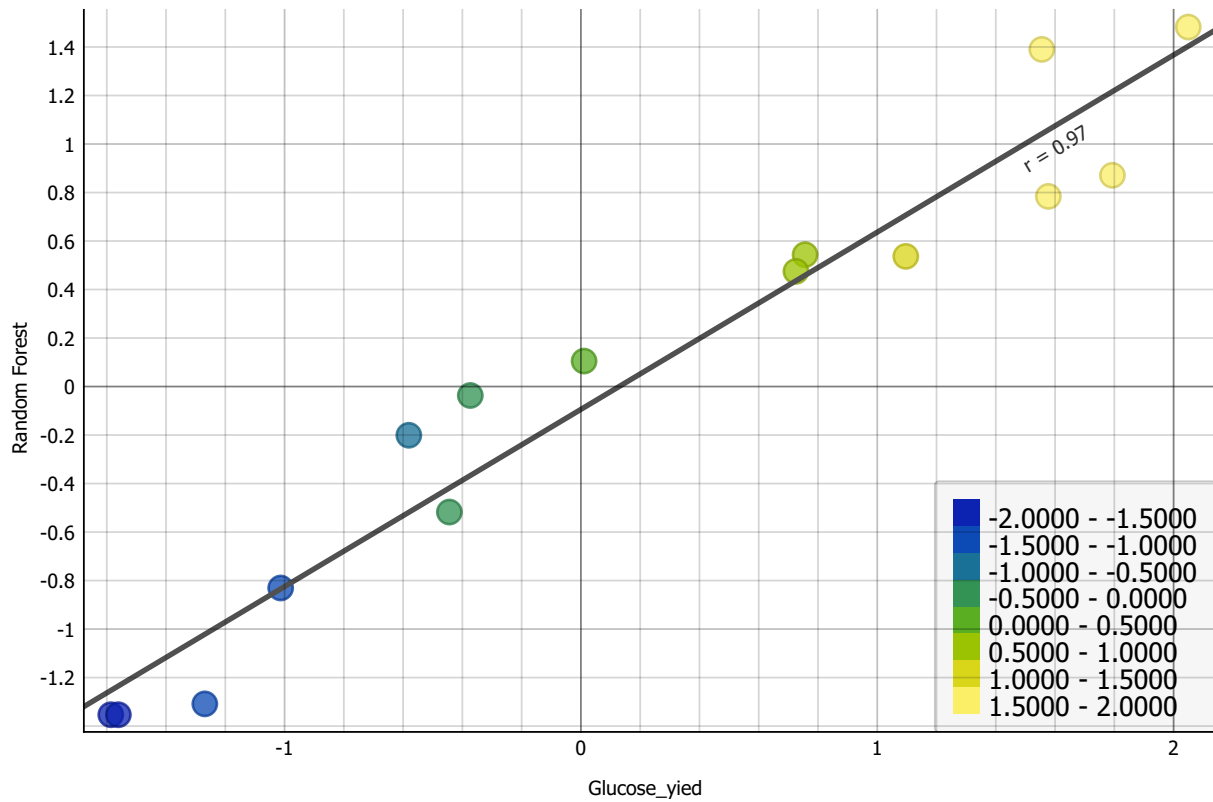


FIGURE E.3 – Visualisation des rendements en glucose prédits par le modèle *Random Forest* en fonction des rendements en glucose réel (valeurs normalisées).

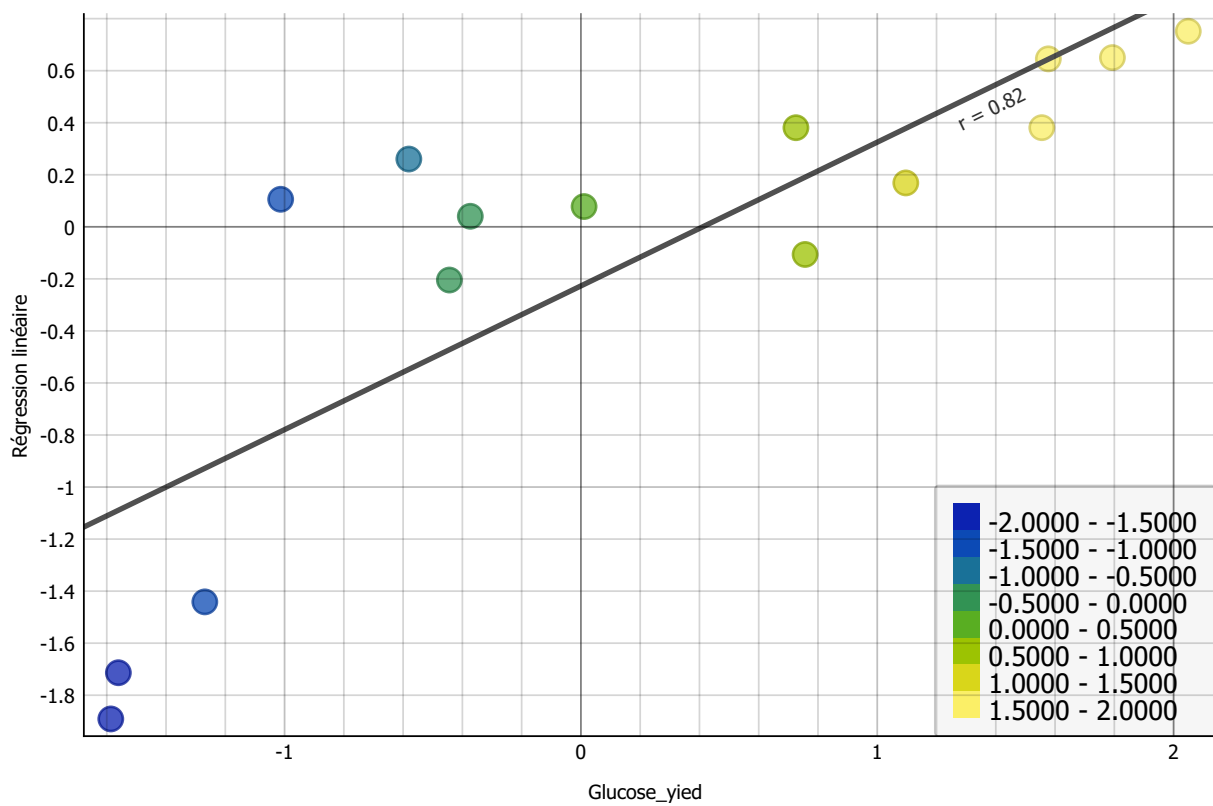


FIGURE E.4 – Visualisation des rendements en glucose prédits par le modèle de régression linéaire en fonction des rendements en glucose réel (valeurs normalisées).

Annexe F

Résultats des évaluations des impacts du cycle de vie méthode *ReCiPe 2016 Endpoint Hierarchical*.

Cette annexe fournit les résultats des évaluations des impacts du cycle de vie de la méthode ReCiPe 2016 Endpoint Hierarchical. Pour ne pas surcharger les tableaux de données, le tableau F.1 est un récapitulatif des catégories d'impacts, leur abréviation et leur unité.

TABLEAU F.1 – Catégories d’impacts *midpoint* avec la méthode *ReCiPe Endpoint*, leur abréviations et leur unité.

Catégories d’impacts	Abréviations	Unités
Global warming, Human health	GW_HH	10 ⁻⁷ DALY
Global warming, Terrestrial ecosystems	GW_TE	10 ⁻⁹ species.yr
Global warming, Freshwater ecosystems	GW_FE	10 ⁻¹⁴ species.yr
Stratospheric ozone depletion	SO3D	10 ⁻¹⁰ DALY
Ionizing radiation	IR	10 ⁻¹⁰ DALY
Ozone formation, Human health	O3_HH	10 ⁻⁹ DALY
Fine particulate matter formation	FPM	10 ⁻⁶ DALY
Ozone formation, Terrestrial ecosystems	O3_TE	10 ⁻¹⁰ species.yr
Terrestrial acidification	TA	10 ⁻⁹ species.yr
Freshwater eutrophication	FE	10 ⁻¹⁰ species.yr
Marine eutrophication	ME	10 ⁻¹⁵ species.yr
Terrestrial ecotoxicity	TEco	10 ⁻¹¹ species.yr
Freshwater ecotoxicity	FEco	10 ⁻¹² species.yr
Marine ecotoxicity	Meco	10 ⁻¹³ species.yr
Human carcinogenic toxicity	HcT	10 ⁻⁹ DALY
Human non-carcinogenic toxicity	HncT	10 ⁻⁸ DALY
Land use	LU	10 ⁻¹⁰ species.yr
Mineral resource scarcity	MS	10 ⁻⁵ USD2013
Fossil resource scarcity	FS	10 ⁻¹ USD2013

TABLEAU F.2 – Évaluation des impacts *midpoint* du cycle de vie avec la méthode *ReCiPe 2016 Endpoint Hierarchical* (P1 à P14).

Catégorie d'impacts	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14
GW_HH	1,166	12,727	11,757	10,911	9,559	9,559	6,361	6,056	5,255	4,869	5,252	10,876	9,420	8,648
GW_TE	0,352	3,841	3,548	3,293	2,885	2,885	1,920	1,828	1,586	1,469	1,585	3,282	2,843	2,610
GW_FE	0,961	10,491	9,691	8,994	7,879	7,879	5,244	4,992	4,331	4,014	4,329	8,964	7,765	7,128
SO3D	0,632	6,850	6,328	5,874	5,146	5,146	3,427	3,263	2,832	2,624	2,830	5,854	5,072	4,657
IR	0,370	2,975	2,757	2,568	2,265	2,265	1,550	1,493	1,313	1,216	1,302	2,562	2,236	2,063
O3_HH	0,363	4,004	3,698	3,432	3,006	3,006	1,999	1,902	1,650	1,529	1,649	3,420	2,962	2,719
FPM	0,150	1,660	1,533	1,423	1,246	1,246	0,828	0,788	0,684	0,633	0,683	1,418	1,228	1,127
O3_TE	0,552	6,093	5,628	5,222	4,574	4,574	3,041	2,894	2,510	2,326	2,509	5,205	4,507	4,137
TA	0,155	1,710	1,579	1,466	1,284	1,284	0,853	0,812	0,704	0,653	0,704	1,461	1,265	1,161
FE	0,711	7,961	7,352	6,822	5,973	5,973	3,967	3,774	3,271	3,031	3,271	6,799	5,886	5,401
ME	0,447	2,936	2,729	2,549	2,260	2,260	1,583	1,538	1,367	1,266	1,347	2,546	2,236	2,071
TEco	0,297	3,285	3,034	2,816	2,466	2,466	1,639	1,560	1,353	1,253	1,352	2,806	2,430	2,230
FEco	0,105	1,168	1,079	1,001	0,877	0,877	0,583	0,554	0,481	0,445	0,481	0,998	0,864	0,793
Meco	0,351	3,891	3,594	3,335	2,921	2,921	1,942	1,848	1,602	1,485	1,602	3,324	2,878	2,642
HcT	0,903	9,761	9,017	8,370	7,333	7,333	4,884	4,651	4,037	3,741	4,034	8,342	7,227	6,636
HncT	0,307	3,291	3,041	2,823	2,474	2,474	1,649	1,570	1,364	1,264	1,362	2,813	2,438	2,239
LU	0,191	2,043	1,888	1,752	1,536	1,536	1,024	0,975	0,847	0,785	0,846	1,747	1,514	1,390
MS	1,111	6,711	6,247	5,842	5,195	5,195	3,677	3,587	3,204	2,965	3,148	5,839	5,142	4,773
FS	0,582	6,487	5,991	5,559	4,868	4,868	3,234	3,077	2,667	2,472	2,667	5,540	4,797	4,402

TABLEAU E.3 – Évaluation des impacts *midpoint* du cycle de vie avec la méthode *ReCiPe 2016 Endpoint Hierarchical* (P15 à P30).

Catégorie d'impacts	P15	P16	P17	P18	P19	P20	P21	P22	P23	P24	P25	P26	P27	P28	P29	P30
GW_HH	8,891	1,585	1,269	0,281	0,222	0,189	0,163	2,977	2,795	1,053	1,004	0,601	0,046	0,044	0,043	1,048
GW_TE	2,683	0,478	0,383	0,085	0,067	0,057	0,049	0,898	0,844	0,318	0,303	0,181	0,014	0,013	0,013	0,316
GW_FE	7,328	1,307	1,046	0,232	0,183	0,156	0,134	2,454	2,304	0,868	0,828	0,495	0,038	0,037	0,036	0,864
SO3D	4,787	0,857	0,689	0,157	0,125	0,108	0,093	1,605	1,508	0,570	0,544	0,317	0,031	0,030	0,029	0,569
IR	2,118	0,464	0,416	0,193	0,179	0,172	0,166	0,801	0,764	0,366	0,352	0,264	0,136	0,135	0,135	0,364
O3_HH	2,795	0,495	0,395	0,084	0,065	0,055	0,046	0,931	0,874	0,325	0,310	0,176	0,010	0,009	0,009	0,325
FPM	1,159	0,205	0,163	0,034	0,026	0,022	0,019	0,386	0,362	0,135	0,128	0,074	0,004	0,003	0,003	0,134
O3_TE	4,254	0,753	0,600	0,127	0,098	0,083	0,070	1,416	1,329	0,494	0,471	0,266	0,014	0,013	0,013	0,494
TA	1,194	0,211	0,168	0,035	0,027	0,023	0,019	0,397	0,373	0,138	0,132	0,075	0,004	0,003	0,003	0,138
FE	5,554	0,974	0,771	0,152	0,115	0,094	0,078	1,837	1,722	0,631	0,601	0,319	0,005	0,004	0,003	0,633
ME	2,123	0,536	0,511	0,297	0,284	0,277	0,271	0,888	0,855	0,470	0,453	0,418	0,239	0,238	0,238	0,460
TEco	2,293	0,405	0,322	0,067	0,052	0,043	0,036	0,762	0,715	0,265	0,253	0,139	0,006	0,006	0,006	0,265
FEco	0,815	0,144	0,114	0,023	0,018	0,015	0,013	0,271	0,254	0,094	0,089	0,049	0,002	0,002	0,002	0,094
Meco	2,716	0,480	0,382	0,079	0,061	0,051	0,043	0,902	0,846	0,313	0,299	0,163	0,007	0,007	0,006	0,314
HcT	6,822	1,224	0,984	0,227	0,181	0,156	0,136	2,298	2,158	0,822	0,785	0,501	0,046	0,045	0,044	0,814
HncT	2,301	0,415	0,335	0,080	0,065	0,056	0,049	0,778	0,731	0,281	0,268	0,173	0,019	0,019	0,018	0,278
LU	1,429	0,258	0,209	0,050	0,041	0,036	0,031	0,483	0,454	0,174	0,166	0,104	0,013	0,012	0,012	0,173
MS	4,889	1,311	1,278	0,795	0,767	0,751	0,739	2,108	2,039	1,170	1,128	0,965	0,661	0,660	0,660	1,162
FS	4,526	0,796	0,631	0,127	0,096	0,080	0,066	1,499	1,406	0,517	0,492	0,262	0,007	0,006	0,005	0,518

TABLEAU F.4 – Évaluation des impacts *midpoint* du cycle de vie avec la méthode *ReCiPe 2016 Endpoint Hierarchical* (P31 à P46).

Catégorie d'impacts	P31	P32	P33	P34	P35	P36	P37	P38	P39	P40	P41	P42	P43	P44	P45	P46
GW_HH	0,842	0,726	0,543	0,574	0,525	0,462	0,383	0,424	2,639	1,913	1,720	1,366	1,290	1,054	1,057	0,795
GW_TE	0,254	0,219	0,164	0,173	0,158	0,139	0,115	0,128	0,796	0,577	0,519	0,412	0,389	0,318	0,319	0,240
GW_FE	0,694	0,599	0,447	0,473	0,433	0,381	0,315	0,349	2,175	1,577	1,418	1,126	1,063	0,869	0,871	0,656
SO3D	0,459	0,396	0,298	0,314	0,288	0,254	0,212	0,234	1,425	1,035	0,931	0,741	0,700	0,573	0,575	0,434
IR	0,318	0,292	0,251	0,258	0,247	0,233	0,215	0,225	0,721	0,558	0,521	0,442	0,425	0,372	0,372	0,314
O3_HH	0,260	0,224	0,166	0,176	0,160	0,141	0,116	0,129	0,826	0,598	0,537	0,425	0,401	0,327	0,328	0,245
FPM	0,107	0,092	0,068	0,072	0,066	0,058	0,047	0,053	0,342	0,247	0,222	0,176	0,166	0,135	0,135	0,101
O3_TE	0,395	0,340	0,252	0,267	0,243	0,213	0,175	0,195	1,257	0,909	0,816	0,646	0,610	0,497	0,498	0,373
TA	0,111	0,095	0,070	0,074	0,068	0,059	0,049	0,054	0,352	0,255	0,229	0,181	0,171	0,139	0,139	0,104
FE	0,503	0,431	0,316	0,335	0,305	0,265	0,216	0,241	1,631	1,176	1,054	0,831	0,783	0,636	0,638	0,473
ME	0,416	0,392	0,353	0,359	0,349	0,335	0,318	0,327	0,799	0,644	0,614	0,539	0,523	0,473	0,472	0,417
TEco	0,212	0,182	0,135	0,143	0,130	0,114	0,093	0,104	0,677	0,489	0,439	0,347	0,327	0,267	0,267	0,200
FEco	0,075	0,064	0,048	0,050	0,046	0,040	0,033	0,037	0,240	0,174	0,156	0,123	0,116	0,094	0,095	0,071
Meco	0,251	0,215	0,159	0,169	0,154	0,135	0,110	0,123	0,801	0,579	0,520	0,411	0,388	0,316	0,316	0,236
HcT	0,656	0,568	0,427	0,451	0,413	0,365	0,304	0,336	2,033	1,477	1,330	1,058	1,000	0,819	0,821	0,621
HncT	0,225	0,195	0,147	0,155	0,143	0,127	0,106	0,117	0,688	0,501	0,452	0,360	0,340	0,280	0,280	0,213
LU	0,140	0,122	0,092	0,097	0,089	0,079	0,067	0,073	0,428	0,312	0,281	0,224	0,212	0,174	0,175	0,133
MS	1,064	1,009	0,921	0,935	0,912	0,882	0,844	0,864	1,922	1,575	1,514	1,344	1,310	1,198	1,195	1,070
FS	0,413	0,354	0,260	0,276	0,251	0,219	0,178	0,199	1,331	0,960	0,861	0,680	0,641	0,521	0,522	0,388

TABLEAU F.5 – Évaluation des impacts *midpoint* du cycle de vie avec la méthode *ReCiPe 2016 Endpoint Hierarchical* (P47 à P62).

Catégorie d'impacts	P47	P48	P49	P50	P51	P52	P53	P54	P55	P56	P57	P58	P59	P60	P61	P62
GW_HH	1,672	1,335	1,240	1,129	1,282	0,988	1,502	1,209	1,470	1,186	0,821	2,789	0,587	0,467	0,442	0,443
GW_TE	0,504	0,403	0,374	0,341	0,387	0,298	0,453	0,365	0,443	0,358	0,248	0,842	0,177	0,141	0,133	0,134
GW_FE	1,378	1,100	1,022	0,930	1,057	0,815	1,238	0,996	1,211	0,977	0,677	2,299	0,484	0,385	0,364	0,366
SO3D	0,905	0,724	0,673	0,613	0,696	0,538	0,814	0,656	0,797	0,644	0,448	1,495	0,321	0,256	0,243	0,243
IR	0,510	0,435	0,414	0,389	0,424	0,358	0,472	0,406	0,465	0,402	0,319	0,766	0,240	0,213	0,207	0,208
O3_HH	0,521	0,415	0,386	0,350	0,399	0,306	0,468	0,376	0,458	0,368	0,253	0,865	0,181	0,143	0,135	0,136
FPM	0,216	0,172	0,159	0,145	0,165	0,126	0,193	0,155	0,189	0,152	0,105	0,360	0,075	0,059	0,056	0,056
O3_TE	0,793	0,631	0,586	0,532	0,606	0,465	0,711	0,571	0,696	0,560	0,385	1,315	0,275	0,217	0,205	0,206
TA	0,222	0,177	0,164	0,149	0,170	0,130	0,199	0,160	0,195	0,157	0,108	0,369	0,077	0,061	0,057	0,057
FE	1,023	0,812	0,753	0,682	0,779	0,594	0,917	0,733	0,896	0,718	0,490	1,691	0,348	0,273	0,257	0,258
ME	0,604	0,532	0,513	0,489	0,522	0,459	0,568	0,505	0,561	0,501	0,422	0,903	0,323	0,298	0,292	0,292
TEco	0,426	0,339	0,315	0,286	0,326	0,250	0,382	0,307	0,374	0,301	0,206	0,704	0,147	0,116	0,110	0,110
FEco	0,151	0,120	0,112	0,101	0,115	0,088	0,136	0,109	0,133	0,107	0,073	0,250	0,052	0,041	0,039	0,039
Meco	0,505	0,402	0,373	0,338	0,385	0,295	0,453	0,363	0,443	0,356	0,244	0,833	0,174	0,137	0,130	0,130
HcT	1,292	1,034	0,962	0,876	0,994	0,769	1,162	0,938	1,138	0,920	0,641	2,178	0,459	0,367	0,348	0,349
HncT	0,439	0,352	0,328	0,299	0,339	0,263	0,395	0,320	0,387	0,314	0,220	0,738	0,158	0,127	0,120	0,121
LU	0,273	0,219	0,204	0,186	0,211	0,164	0,246	0,199	0,241	0,195	0,137	0,455	0,098	0,079	0,075	0,075
MS	1,491	1,329	1,288	1,234	1,308	1,167	1,409	1,269	1,395	1,259	1,081	2,067	0,833	0,776	0,764	0,765
FS	0,836	0,664	0,616	0,559	0,637	0,487	0,749	0,600	0,733	0,588	0,402	1,379	0,286	0,225	0,212	0,212

TABLEAU F.6 – Évaluation des impacts *midpoint* du cycle de vie avec la méthode *ReCiPe 2016 Endpoint Hierarchical* (P63 à P78).

Catégorie d'impacts	P63	P64	P65	P66	P67	P68	P69	P70	P71	P72	P73	P74	P75	P76	P77	P78
GW_HH	9,323	8,516	5,970	0,787	0,627	0,549	0,470	7,311	6,901	5,766	4,368	0,045	0,279	0,046	0,045	0,045
GW_TE	2,813	2,570	1,801	0,237	0,189	0,166	0,142	2,206	2,082	1,740	1,318	0,014	0,084	0,014	0,014	0,013
GW_FE	7,685	7,019	4,921	0,649	0,517	0,452	0,387	6,026	5,688	4,752	3,601	0,037	0,230	0,038	0,037	0,037
SO3D	3,469	3,170	2,225	0,427	0,341	0,299	0,256	2,720	2,568	2,147	1,628	0,030	0,156	0,031	0,030	0,030
IR	2,832	2,596	1,851	0,257	0,221	0,203	0,185	2,216	2,096	1,764	1,356	0,146	0,198	0,146	0,146	0,145
O3_HH	1,797	1,642	1,151	0,245	0,195	0,170	0,145	1,409	1,330	1,112	0,842	0,009	0,083	0,009	0,009	0,009
FPM	0,977	0,892	0,625	0,101	0,080	0,070	0,060	0,766	0,723	0,604	0,458	0,003	0,034	0,003	0,003	0,003
O3_TE	2,600	2,375	1,665	0,372	0,296	0,258	0,220	2,039	1,925	1,608	1,218	0,013	0,125	0,013	0,013	0,013
TA	0,815	0,744	0,521	0,104	0,083	0,072	0,062	0,639	0,603	0,504	0,382	0,003	0,035	0,003	0,003	0,003
FE	1,298	1,185	0,830	0,479	0,378	0,329	0,280	1,018	0,961	0,802	0,607	0,003	0,149	0,003	0,003	0,003
ME	10,351	9,469	6,689	0,314	0,280	0,263	0,247	8,110	7,662	6,422	4,896	0,258	0,307	0,257	0,257	0,257
TEco	0,930	0,850	0,596	0,200	0,159	0,138	0,118	0,730	0,689	0,575	0,436	0,006	0,066	0,006	0,006	0,006
FEco	0,278	0,254	0,178	0,071	0,056	0,049	0,042	0,218	0,206	0,172	0,130	0,002	0,023	0,002	0,002	0,002
Meco	1,019	0,931	0,653	0,237	0,188	0,164	0,140	0,799	0,754	0,630	0,478	0,006	0,078	0,007	0,007	0,006
HcT	11,161	10,194	7,145	0,610	0,487	0,427	0,367	8,753	8,262	6,902	5,228	0,046	0,225	0,047	0,046	0,046
HncT	3,832	3,501	2,454	0,208	0,166	0,146	0,126	3,005	2,837	2,370	1,796	0,019	0,080	0,019	0,019	0,019
LU	1,878	1,716	1,204	0,129	0,104	0,091	0,079	1,473	1,390	1,162	0,881	0,013	0,050	0,013	0,013	0,013
MS	10,496	9,633	6,913	0,785	0,709	0,671	0,633	8,208	7,769	6,556	5,063	0,716	0,825	0,713	0,712	0,712
FS	0,979	0,894	0,626	0,391	0,310	0,270	0,229	0,767	0,724	0,605	0,458	0,005	0,124	0,005	0,005	0,005

Annexe G

Cartes MDS des quatre dimensions des procédés et des impacts.

Cette annexe fournit les cartes des quatre premières dimensions projetées sur les impacts et les procédés. Le descriptif détaillé des groupes pour les procédés est dans le tableau G.1 . Un graphique présente la représentativité cumulée des dimensions pour la projection sur les impacts et sur les procédés pour les 8 premières dimensions.

TABLEAU G.1 – Numéro des groupes issus de la méthode hybride MDS - *k-means* par procédé et référence bibliographique.

Référence bibliographique	Numéro du procédé	Groupe k-means	Référence bibliographique	Numéro du procédé	Groupe k-means
PM1	P1	3	PM-UFM1	P30	3
	P2	2		P31	3
	P3	2		P32	3
PM2	P4	2		P33	3
	P5	2		P34	3
PM3	P6	2		P35	3
	P7	2		P36	3
PM4	P8	2		P37	3
	P9	2		P38	3
	P10	2		P39	2
	P11	2	P40	3	
	P12	2	P41	3	
	P13	2	P42	3	
PM5	P14	2	P43	3	
	P15	2	P44	3	
	P16	3	P45	3	
PM6	P17	3	PM-UFM2	P46	3
	P18	3		P47	3
PM7	P19	1		P48	3
	P20	1		P49	3
	P21	1		P50	3
PM8	P22	2		P51	3
	P23	2		P52	3
PM9	P24	3		P53	3
	P25	3		P54	3
PM10	P26	3		P55	3
	P27	1	P56	3	
PM11	P28	1	PM-UFM3	P57	3
	P29	1	PM-UFM4	P58	2
			PM-UFM5	P59	3
				P60	3
				P61	3
				P62	3
				P63	2
			P64	2	
			P65	2	
			PM-UFM6	P66	3
				P67	3
				P68	3
				P69	3
				P70	2
				P71	2
			P72	2	
			P73	2	
			PM-UFM7	P74	1
			PM-UFM8	P75	3
			PM-UFM9	P76	1
				P77	1
				P78	1

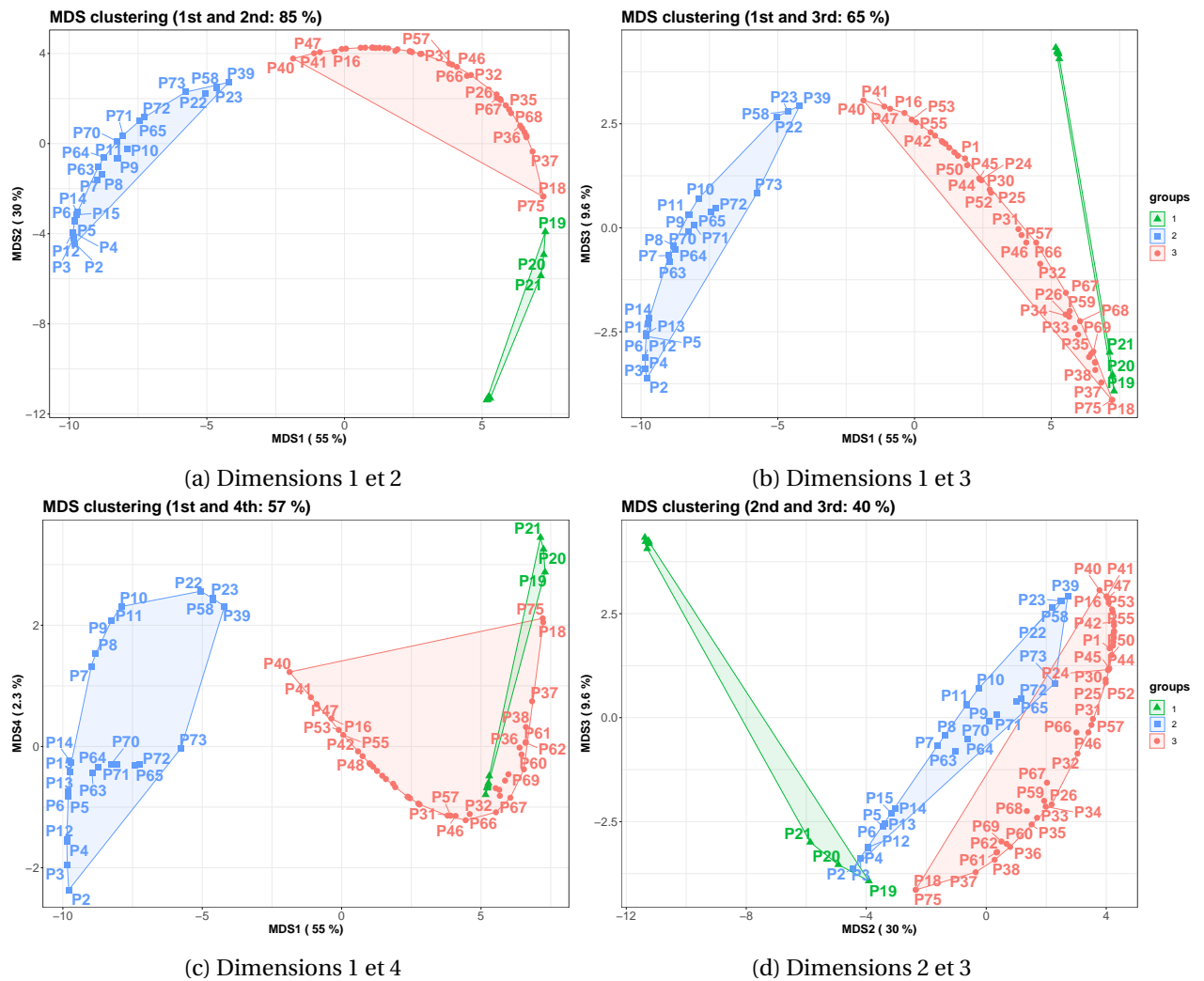


FIGURE G.1 – Cartes MDS des quatre dimensions projetées sur les procédés

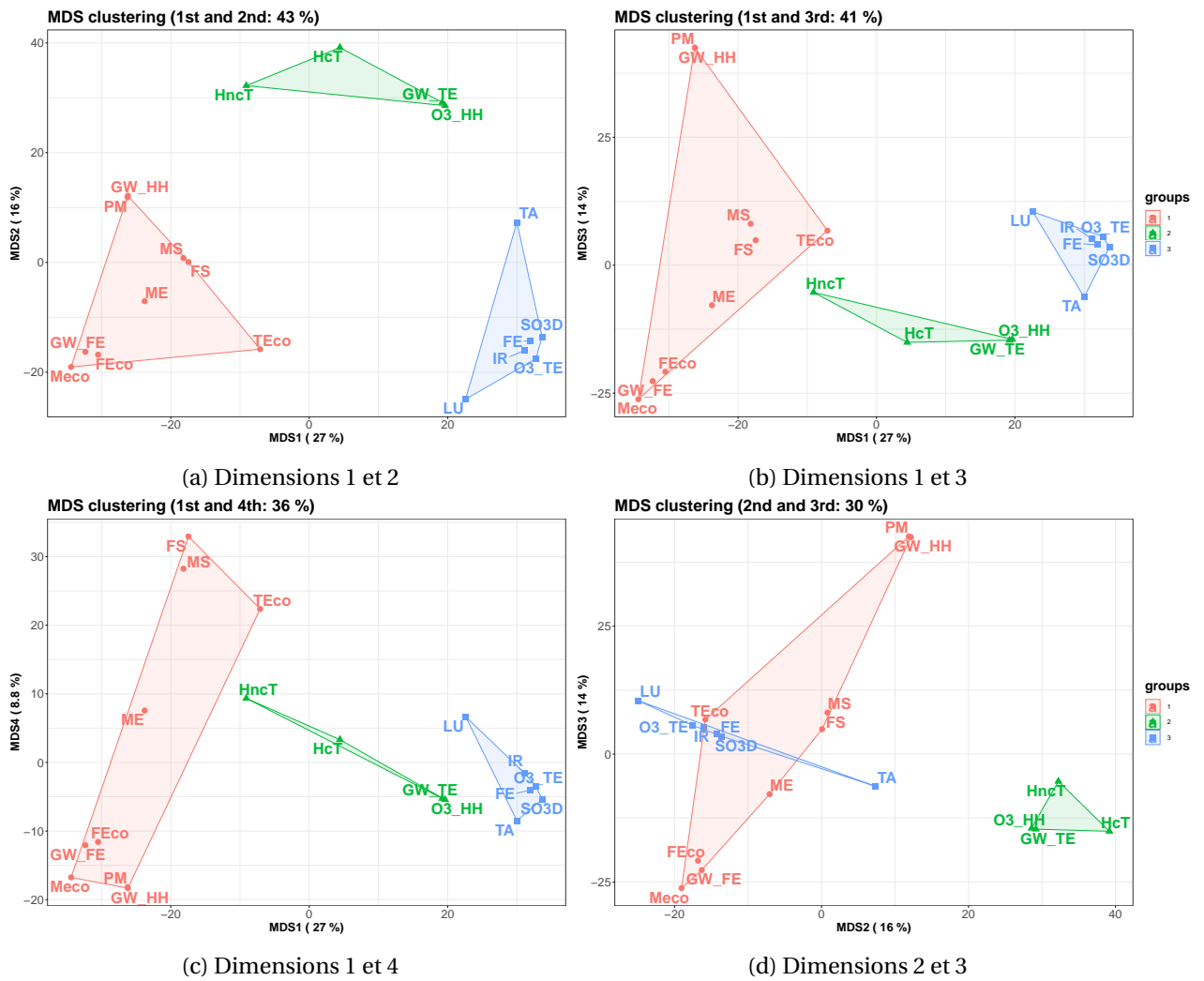


FIGURE G.2 – Cartes MDS des quatre dimensions projetées sur les impacts

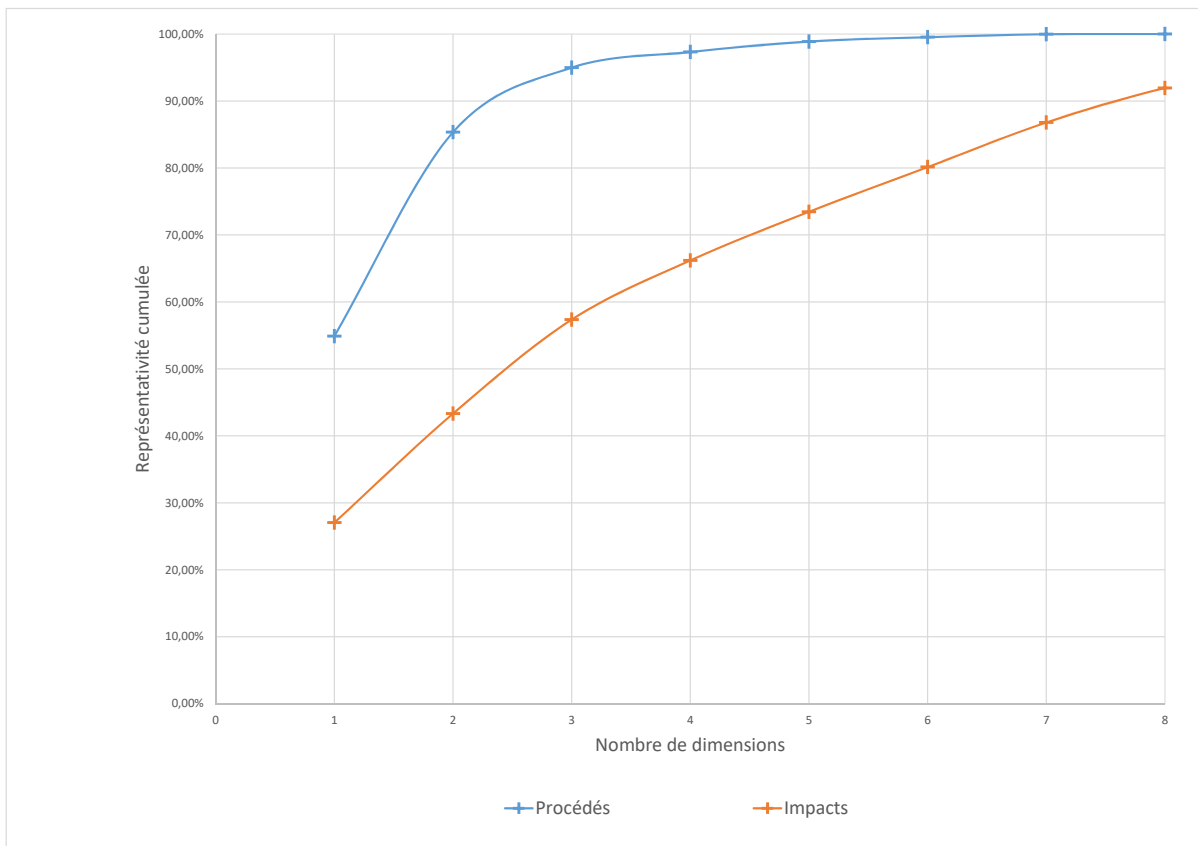


FIGURE G.3 – Représentativité cumulée des huit premières dimensions pour les projections sur les impacts et les procédés de la méthode MDS

Annexe H

Guide d'utilisation simple de l'outil.

Cette annexe fournit un guide d'utilisation simple de l'outil développé lors de nos travaux.

H.1 Menu principal

À l'ouverture de l'Excel, le menu principal s'ouvre. Il contient différentes informations et l'accès aux différents sous-menu servant à faire les évaluations environnementales et économiques.

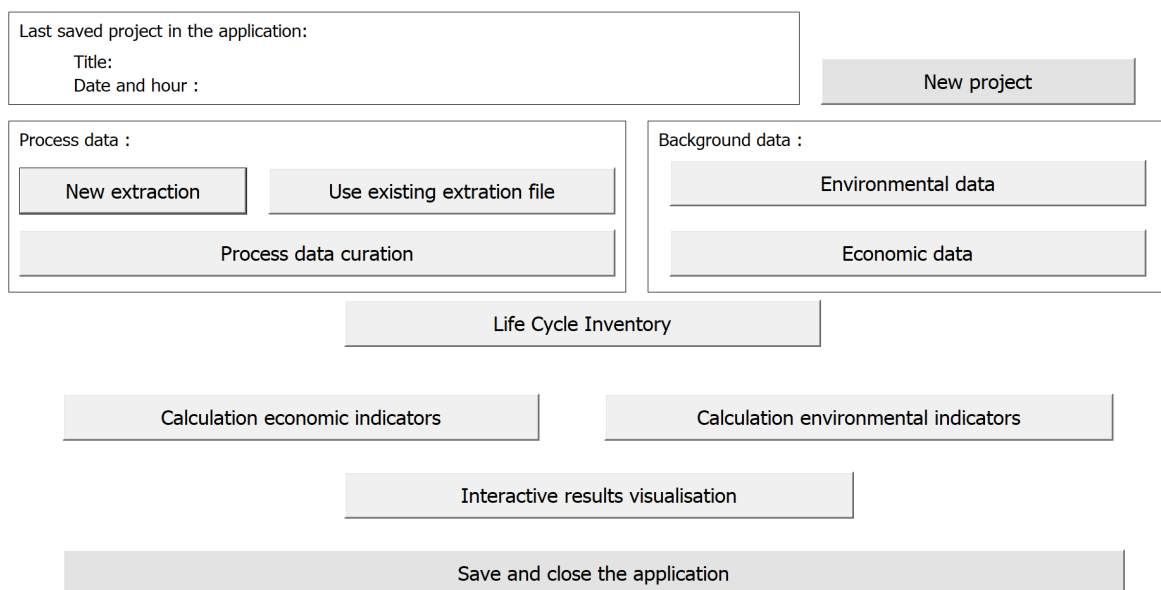


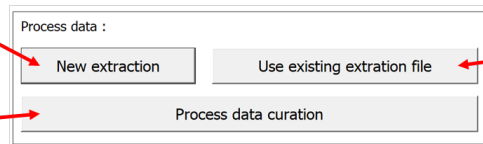
FIGURE H.1 – Menu principal de l'outil Excel.

H.2 Extraction des données

Cette partie permet d'extraire les données procédés des fichiers CSV issus d'@Web et de les nettoyer avant de les stocker dans l'Excel.

Nouvelle extraction : fusion et triage des données contenues dans différents fichiers CSV dans un fichier CSV type

Nettoyage des données : nettoyage des données procédés par l'Excel suivant les hypothèses émises par l'utilisateur



Utilisation d'un fichier pré-existant : Utilisation d'un fichier CSV type précédemment créé

FIGURE H.2 – Fenêtre de la partie "Extraction des données procédés" du menu principal.

H.2.1 Nouvelle extraction

En appuyant sur le bouton « Nouvelle extraction », l'utilisateur ouvre une nouvelle fenêtre.

Choix des fichiers: Ouverture d'une boîte de dialogue pour le choix des différents fichiers CSV qui seront utilisés

Triage: Triage alphabétique des différents fichiers sélectionnés

Importation des données: Extraction des données des fichiers sélectionnés pour les intégrer dans un seul fichier CSV type. Apparition d'une boîte de dialogue pour entrer le nom de ce fichier type.

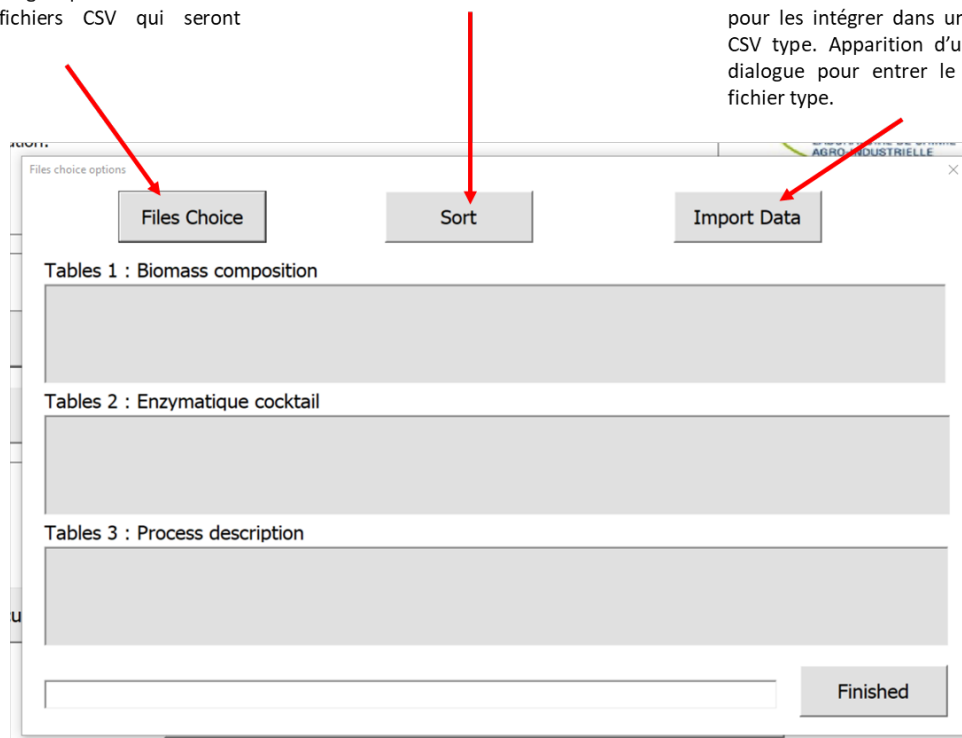


FIGURE H.3 – Fenêtre de choix des fichiers pour une nouvelle extraction des données procédés.

Une fois le choix des fichiers CVS sources utilisés, les noms entiers de ceux-ci se retrouvent écrits dans les parties grisées. Ces fichiers qui sont extraits de @Web sont triés suivant leur contenu :

1. Fichiers contenant les tables de composition de la biomasse
2. Fichiers contenant les détails sur les cocktails enzymatiques
3. Fichiers contenant la description complète des procédés.

Dans @Web ces fichiers correspondent aux 3 différentes tables décrivant un même article. Il faut donc bien par article 3 fichiers dont 1 dans chaque catégorie. L'application vérifie que c'est bien le cas et prévient l'utilisateur si ce n'est pas le cas avant l'importation. Il est important de bien trier les noms des fichiers avant l'import. En appuyant sur l'importation, une boîte de chargement apparaît pendant que l'application crée le fichier CSV type avec les données extraites des fichiers sélectionnés puis une boîte de dialogue demandant le nom de ce fichier avant l'enregistrement apparaît. Le fichier est enregistré directement dans le dossier source du fichier Excel de l'application.

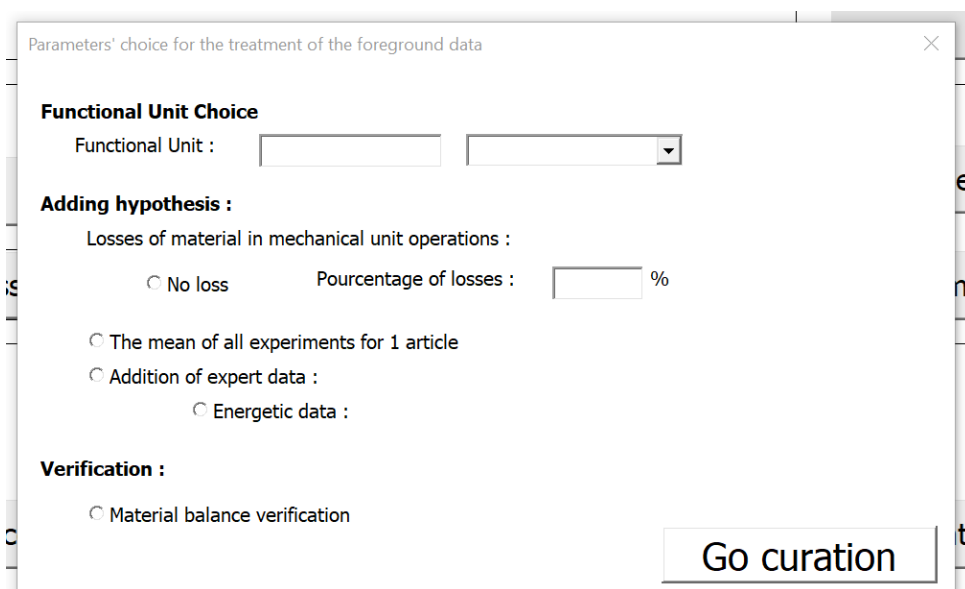
H.2.2 Utilisation d'un fichier préexistant

Si l'utilisateur possède déjà un fichier CSV préexistant contenant les données sources dont il a besoin il peut directement le sélectionner. Le nom du fichier sera alors enregistré dans le projet.

H.2.3 Nettoyage des données

Une nouvelle fenêtre s'ouvre pour entrer les paramètres et les différentes hypothèses. Il est possible aussi de lancer une vérification du bilan matière. Les différents paramètres possibles sont :

- Le choix de l'unité fonctionnelle
- L'addition de données d'experts (comme les données énergétiques)
- Le pourcentage de perte dans les opérations unitaires mécaniques
- La moyenne de toutes les expériences d'un article



The image shows a dialog box titled "Parameters' choice for the treatment of the foreground data". It contains several sections for user input:

- Functional Unit Choice:** A label "Functional Unit :" followed by a text input field and a dropdown menu.
- Adding hypothesis :**
 - A label "Losses of material in mechanical unit operations :" followed by two radio buttons: "No loss" and "The mean of all experiments for 1 article".
 - A label "Percentage of losses :" followed by a text input field and a "%" symbol.
 - A radio button "Addition of expert data :" followed by a sub-label "Energetic data :" and a radio button.
- Verification :** A radio button "Material balance verification".

A "Go curation" button is located at the bottom right of the dialog box.

FIGURE H.4 – Fenêtre du nettoyage des données procédés (ou données de premier-plan).

Une fois le nettoyage effectué, les données sont stockées dans un nouvel onglet créé. Cet onglet appelé "*Foreground data*" contient un résumé des hypothèses sélectionnées ainsi que la date du nettoyage. C'est à partir de cet onglet que sont effectuées les évaluations environnementales.

A	B	C	D	E	F	G	H	I	J	K	L	M	N
Tab name:		Foreground data											
Tab description:		Curation results of foreground data											
Date:		23/3/20 16:55											
Process/experience number:		7											
Functional unit:		1 g de glucose											
Adding hypothesis:		No loss of material in mechanical unit operations Mean of all experiment for 1 article Adding of expert data : NO											
Material balance verification		OK											
Biblio ref	Biomass	Biomass state	Experience number	Glucose rate	Cocktail number	Enzymatic cocktail	Enzyme activity1	Enzyme activity2	Enzyme activity3	Experience number	Process step number	Biomass	Treatment
			Unit : 1	Unit : %	Unit : 1		Unit : FPU/g	Unit : FPU/g	Unit : FPU/g	Unit : 1	Unit : 1		
PM-PC-PS5	Corn stover	Untreated bi	1,00E+00	3,43E+01	1,00E+00	Cellulase	1,50E+01	0,00E+00	0,00E+00	1	1,00E+00	Corn stover	Cutting milling
					1,00E+00	Beta-glucosid	0,00E+00	6,50E+01	0,00E+00	1	2,00E+00	Corn stover	Drying
										1	3,00E+00	Corn stover	Hot water treatment
										1	4,00E+00	Corn stover	Washing and separation
										1	5,00E+00	Corn stover	Enzymatic hydrolysis treatment
PM-PC-PS9	Corn stover	Untreated bi	1,00E+00	3,73E+01	1,00E+00	Cellic C-Tec	7,50E+00	0,00E+00	0,00E+00	1	1,00E+00	Corn stover	Cutting milling
										1	2,00E+00	Corn stover	Impregnation
										1	3,00E+00	Corn stover	Washing and press
										1	4,00E+00	Corn stover	Steam treatment

FIGURE H.5 – Vue d'un onglet "Données de premier plan" contenant les données de premier-plan nettoyées.

H.3 Données d'arrière-plan (*Background data*)

L'utilisateur choisit les fichiers CSV contenant les données d'arrière-plan (environnementales et économiques) à travers l'ouverture d'une boîte de dialogue de recherche Windows. Une fois choisi le chemin des fichiers est enregistré dans le projet.

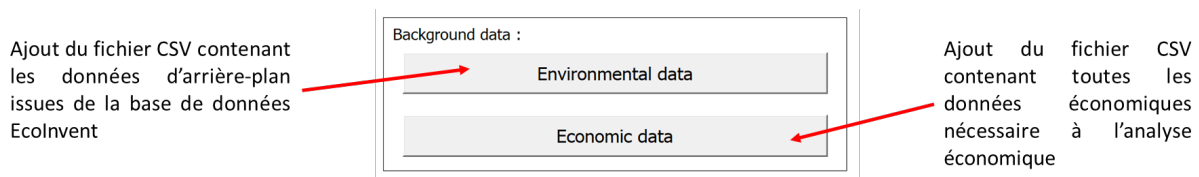


FIGURE H.6 – Description de la partie de données d'arrière-plan du menu principal.

H.4 Évaluation des impacts environnementaux et économiques

La dernière partie du menu principal concerne l'évaluation des impacts environnementaux et économiques.

L'utilisateur doit dans un premier temps faire l'inventaire du cycle de vie. Un onglet « Inventaire » apparait alors contenant l'inventaire de tous les procédés étudiés. Suivant les hypothèses précédentes sélectionnées, on retrouve un inventaire par article scientifique ou par expérience.

Une fois l'inventaire effectué, l'utilisateur peut effectuer l'analyse environnementale. Une fenêtre s'ouvre alors pour que l'utilisateur puisse choisir la méthode de calcul utilisée et le niveau d'agrégation des calculs.

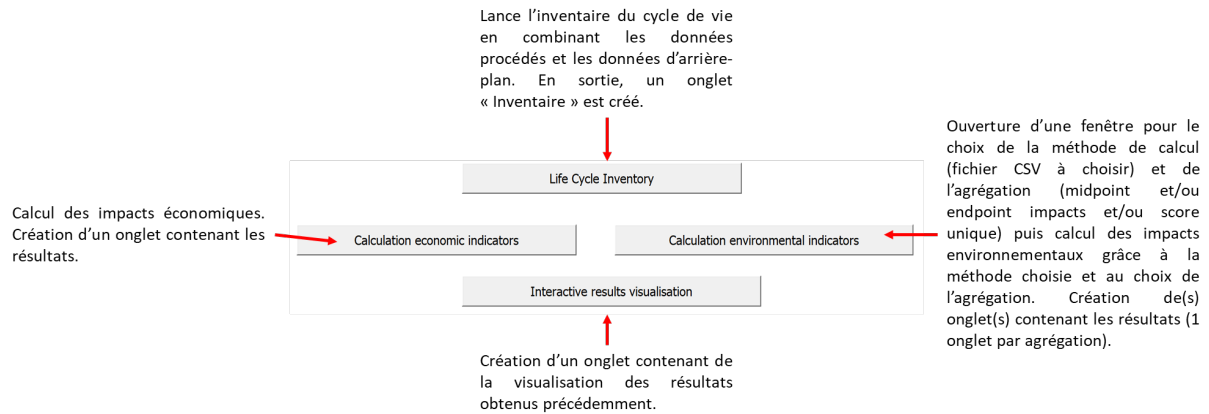


FIGURE H.7 – Description de la partie "évaluation des impacts économiques et environnementaux" du menu principal.

No	Substance	Compartment	Sub-compart	Unit	PM-PC-EX-PS1	PM-PC-EX-PS2	PM-PC-PS14	PM-PC-PS20	PM-PC-PS23	PM-PC-PS5	PM-UFM1
1	1-Butanol	Air	high. pop.	pg	18,308382	51,325516	16,818559	7,0372427	9,7288722	145,75076	5,6617373
2	1-Butanol	Water	river	µg	1,4594869	4,0905316	1,3404016	0,56111069	0,77420141	11,617847	0,45122835
3	1-Pentanol	Air	high. pop.	pg	9,6349171	26,984318	8,84244	3,6941808	5,0964929	76,58715	2,9766735
4	1-Pentanol	Water	river	pg	23,123829	64,762442	21,221882	8,8660446	12,231598	183,80938	7,1440251
5	1-Pentene	Air	high. pop.	pg	7,2809199	20,391526	6,6820604	2,7916207	3,8513208	57,875423	2,2494144
6	1-Pentene	Water	river	pg	17,474192	48,939618	16,036931	6,6998839	9,2431618	138,90089	5,3985899
7	1-Propanol	Air	high. pop.	ng	5,2026159	14,563125	4,7721101	1,9901554	2,7257062	41,29441	1,6064595
8	1-Propanol	Water	river	pg	46,816025	131,15795	42,978788	17,961463	24,797709	372,33094	14,468141
9	1,4-Butanediol	Air	high. pop.	ng	4,6573641	13,053339	4,2773699	1,7905373	2,4706197	37,073598	1,4399196
10	1,4-Butanediol	Water	river	ng	1,8629591	5,2213735	1,7109604	0,71622012	0,98825501	14,829547	0,57597203
11	2-Aminopropanol	Air	high. pop.	pg	1,4092096	3,9502306	1,294434	0,54130019	0,74828235	11,218118	0,43575033
12	2-Aminopropanol	Water	river	pg	3,5301418	9,8955487	3,2426295	1,3559873	1,8745016	28,102047	1,0915789
13	2-Rutene, 2-methyl-	Air	high. pop.	ng	0,001614995	0,004523083	0,001482161	0,00061921	0,00085427	0,01283766	0,00049895

FIGURE H.8 – Vue de l'onglet "Inventaire" contenant les résultats de l'inventaire.

	A	B	C	D	E	F	G	H	I	J	K
1	Tab name:		Impacts								
2	Tab description:		Environmental impacts results								
3	Functional unit:		1 g de glucose								
4	Ecoinvent version:		v2.1								
5	Process/experience number:		7								
6	Agregation:		Characterization (Midpoint impacts)								
7	Environmental analysis method:		ReCiPe Endpoint (H) V1.06 / Europe ReCiPe H/A								
8	Impact categories number:		17								
9											
10	Impact category	Unit	PM-PC-EX-P5	PM-PC-EX-P5	PM-PC-PS14	PM-PC-PS20	PM-PC-PS23	PM-PC-PS5	PM-UFM1		
11	Climate change Human Health	DALY	5,46E-07	1,53E-06	5,02E-07	2,09E-07	2,94E-07	4,35E-06	1,69E-07		
12	Ozone depletion	DALY	1,86E-10	5,22E-10	1,71E-10	7,14E-11	1,00E-10	1,48E-09	5,76E-11		
13	Human toxicity	DALY	1,21E-07	3,38E-07	1,11E-07	4,62E-08	6,38E-08	9,59E-07	3,73E-08		
14	Photochemical oxidant formation	DALY	5,50E-11	1,54E-10	5,05E-11	2,11E-11	2,92E-11	4,37E-10	1,70E-11		
15	Particulate matter formation	DALY	1,19E-07	3,34E-07	1,09E-07	4,63E-08	6,39E-08	9,48E-07	3,68E-08		
16	Ionising radiation	DALY	2,42E-09	6,77E-09	2,22E-09	9,25E-10	1,27E-09	1,92E-08	7,47E-10		
17	Climate change Ecosystems	species.yr	3,09E-09	8,67E-09	2,84E-09	1,19E-09	1,67E-09	2,46E-08	9,57E-10		
18	Terrestrial acidification	species.yr	8,25E-12	2,31E-11	7,58E-12	3,23E-12	4,41E-12	6,56E-11	2,55E-12		
19	Freshwater eutrophication	species.yr	9,18E-12	2,57E-11	8,43E-12	3,52E-12	4,82E-12	7,29E-11	2,84E-12		
20	Terrestrial ecotoxicity	species.yr	1,51E-11	4,24E-11	1,39E-11	5,79E-12	8,05E-12	1,20E-10	4,68E-12		
21	Freshwater ecotoxicity	species.yr	9,35E-13	2,62E-12	8,58E-13	3,58E-13	4,92E-13	7,42E-12	2,89E-13		
22	Marine ecotoxicity	species.yr	2,86E-15	8,01E-15	2,62E-15	1,09E-15	1,51E-15	2,27E-14	8,83E-16		
23	Agricultural land occupation	species.yr	7,71E-11	2,16E-10	7,08E-11	2,96E-11	4,07E-11	6,13E-10	2,38E-11		
24	Urban land occupation	species.yr	3,86E-11	1,08E-10	3,54E-11	1,48E-11	2,04E-11	3,07E-10	1,19E-11		
25	Natural land transformation	species.yr	3,08E-10	8,65E-10	2,83E-10	1,18E-10	1,64E-10	2,45E-09	9,54E-11		
26	Metal depletion	\$	0,00185769	0,0052076	0,0017065	0,00071383	0,00098591	0,01478681	0,00057443		
27	Fossil depletion	\$	4,4708218	12,551151	4,1128136	1,7148928	2,386151	35,598624	1,3845246		
28											
29											
30											

FIGURE H.9 – Vue d'un onglet de résultats de l'évaluation environnementale.

H.5 Visualisation des résultats

Après le calcul des impacts, il est possible de visualiser les résultats de façon interactive grâce à PowerBI.

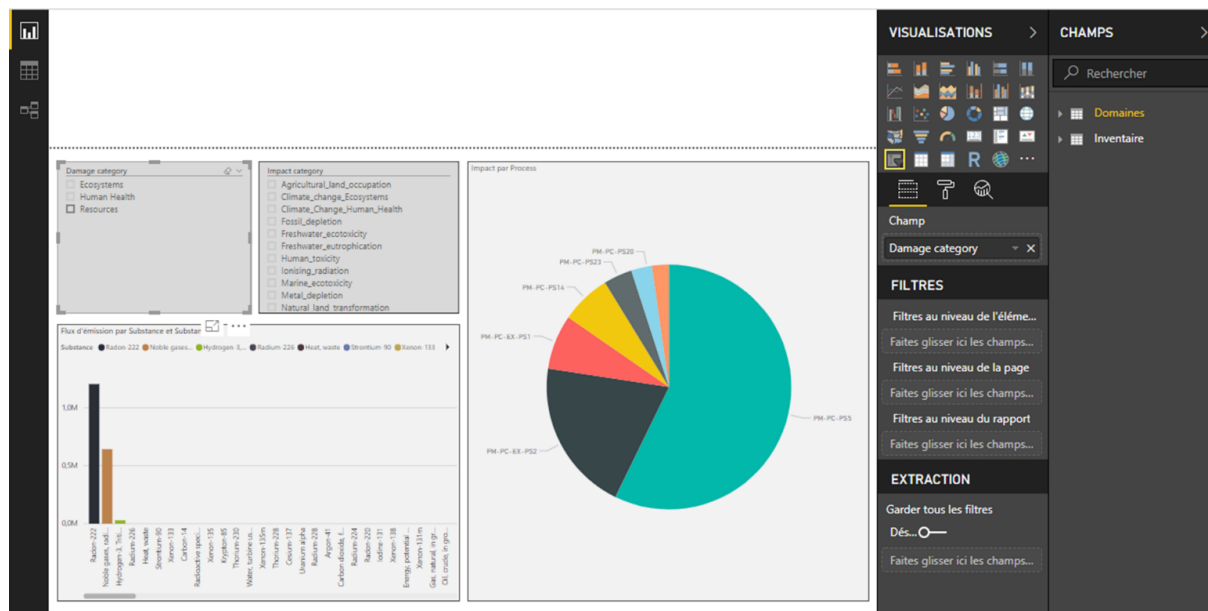


FIGURE H.10 – Vue d'une visualisation interactive des résultats de l'évaluation environnementale.

Annexe I

Liste des acronymes

- AcCV** Analyse des Coûts du Cycle de Vie. 45, 46, 63, 66, 153
- ACP** analyse en composante principale. 109
- ACV** Analyse du Cycle de Vie. 9, 33, 42–48, 50, 52, 53, 62, 63, 65, 66, 68, 69, 78, 134, 137, 142, 147, 154
- AdCV** Analyse de la durabilité du Cycle de Vie. 46, 47, 63, 68, 153
- ADEME** Agence Française de l'Environnement et de la Maîtrise de l'Énergie. 11
- ADMC** Analyse Décisionnelle MultiCritères. 48
- AeCV** Analyse environnementale du Cycle de Vie. 44, 46, 63, 66, 72, 154
- ALE** *Accumulated Local Effects*. 95
- API** *Application Programming Interface*. 78
- AsCV** Analyse sociale du Cycle de Vie. 46, 63, 66, 68
- CP** composantes principales. 109
- CPS** *Cyber Physical Systems*. 16–19, 33
- DD** développement durable. 7–9, 12–16, 19, 21, 27, 33, 41, 44, 47, 53, 57, 58, 63
- DR** réduction de dimension. 72, 106
- EC** économie circulaire. 9–11, 19, 21, 22, 41, 52, 53, 60, 117, 120, 123, 147
- EICV** Évaluation des impacts du cycle de vie. 40, 44, 48, 63, 72, 73, 112
- EICVe** Évaluation des impacts du cycle de vie environnementale. 66, 67, 103
- ERP** *Enterprise Resource Planning*. 17
- GIEC** Groupe d'experts Intergouvernemental sur l'Évolution du Climat. 7, 8
- HTML** *HyperText Markup Language*. 78, 85
- I4.0** Industrie 4.0. 16, 17, 19, 20, 40, 57, 94
- IA** Intelligence Artificielle. 14, 18, 19, 21, 23, 24, 27, 33, 34, 38–40, 47–50, 53, 57, 61, 73, 94, 104, 120
- IC** Ingénierie des Connaissances. 38, 41, 47, 50, 53, 73, 78, 79, 83

- ICE** *Individual Conditional Expectation*. 95, 97, 98, 131, 133, 150
- ICV** Inventaire du Cycle de Vie. 44, 48, 50, 63, 64, 66–69, 73
- IoE** *Internet of Everything*. 18, 19
- IoT** *Internet of Things*. 16–19, 32, 33
- IS** Ingénierie Système. 19–21
- ISBM** Ingénierie Système Basée sur les Modèles. 21
- LCT** *Life Cycle Thinking*. 41
- LIME** *Local Interpretable Model-agnostic Explanations*. 95
- MBSE** *Model-Based Systems Engineering*. 21
- MDS** positionnement multidimensionnel. 72, 109, 112, 114, 118, 138–140, 142, 143
- ML** *Machine Learning*. 34–36, 41, 47, 50, 94, 95, 104, 107, 131, 137, 150
- NLP** traitement du langage naturel. 37
- ODD** Objectifs du Développement Durable. 9, 15
- ONG** Organisation non gouvernementale. 8
- ONU** Organisation des Nations Unies. 8, 14, 47
- PCV** pensée cycle de vie. 41, 42, 53, 60
- PDF** *Portable Document Format*. 62, 78
- PDP** *Partial Dependence Plot*. 95–97, 131, 133, 150
- PGI** Progiciel de Gestion Intégré. 17
- PM** *Pre-milling*. 130, 138, 141
- PM-PC-PS** *Pre-milling, Physico-Chimical treatment, and Press and Separation*. 130
- PM-UFM** *Pre-milling and UltraFine Milling*. 130, 138, 141
- PSI** procédés et systèmes industriels. 21, 27, 50–53
- R&D** Recherche et Développement. 24, 46, 57, 60, 61, 73, 106, 123
- RDF** *Resource Description Framework*. 78, 81, 86, 87, 128
- RN** Réseaux de Neurones Artificiels. xv, 91–95, 98, 131, 133, 147, 150
- RTO** Ressource Ontologique et Terminologique. 81, 83–86, 126–128, 147, 149, VII, VIII
- SGBD** systèmes de gestion de bases de données relationnelles. 63
- SKOS** *Simple Knowledge Organisation System*. 81
- SQL** *Structured Query Language*. 78
- TRL** *Technology Readiness Level*. 73, 147, 154
- UE** Union Européenne. 8
- W3C** *World Wide Web Consortium*. 81

Production scientifique

Articles publiés

- [1] BELAUD, J.-P., N. PRIOUX, C. VIALLE ET C. SABLAYROLLES. 2019, «Big data for agri-food 4.0 : Application to sustainability management for by-products supply chain», *Computers in Industry*, vol. 111, doi :10.1016/j.compind.2019.06.006, p. 41–50, ISSN 01663615. URL <https://linkinghub.elsevier.com/retrieve/pii/S0166361518306419>.
- [2] PRIOUX, N., R. OUARET, G. HETREUX ET J.-P. BELAUD. 2022, «Environmental assessment coupled with machine learning for circular economy», *Clean Technologies and Environmental Policy*, doi :10.1007/s10098-022-02275-4. <https://link.springer.com/article/10.1007/s10098-022-02275-4>
- [3] PRIOUX, N., J.-P. BELAUD ET G. HETREUX. 2022, «Intensive Data and Knowledge-based Approach for Sustainable and Circular Industrial Systems», *Insight*, vol. 24, n°4, doi : 10.1002/inst.12363, p. 43-46. <https://incose.onlinelibrary.wiley.com/doi/10.1002/inst.12363>

Congrès avec actes

- [4] PRIOUX, N., J.-P. BELAUD, R. OUARET ET G. HÉTREUX. 2021, «Data and environment based approach for process systems engineering», *Computer Aided Chemical Engineering*, vol. 50, doi :10.1016/B978-0-323-88506-5.50154-6, p. 999-1004, ISSN 1570-7946. <https://www.sciencedirect.com/science/article/pii/B9780323885065501546>
- [5] PRIOUX, N., R. OUARET, Z. CHEN, G. HETREUX ET J.-P. BELAUD. 2020, «Évaluation Environnementale Couplée à l'Analyse Multidimensionnelle des Données pour l'Économie Circulaire», *13ème Conférence internationale de modélisation, optimisation et simulation (MOSIM2020)*, 12-14 Nov 2020, AGADIR (virtual), Maroc. hal-3192823. <https://hal.archives-ouvertes.fr/hal-03192823/>.

Autres communications

- [6] PRIOUX, N., G. HETREUX, ET J.-P. BELAUD. 2020, «Cadre méthodologique centré données et connaissances intensives pour l'analyse des procédés industriels durables et circulaires», *Journées STP – GDR MACS*, 12-13 Février 2019, Nantes, France.
- [7] PRIOUX, N., G. HETREUX, ET J.-P. BELAUD. (2020). «Intensive data and knowledge based approach for sustainable and circular industrial systems», Séminaire Doctoral AFIS (UTC), 10 December 2019, Compiègne, France.
- [8] PRIOUX, N., R. OUARET, G. HETREUX ET J.-P. BELAUD 2021, «Data oriented sustainability analysis for agricultural processes», *5th CIGR International Conference "Integrating agriculture and society*

through engineering", Quebec City, Canada.

- [9] PRIoux, N., R. OUARET, G. HÉTREUX ET J.-P. BELAUD. 2022, «Digitalization in sustainability assessment for supporting agri-food waste circularity», *32th congress Escape, ESCAPE 32*, June 2022, Toulouse, France.
- [10] PRIoux, N., R. OUARET, J.-P. BELAUD. 2022. «Bridging machine learning and sustainability assessment : application to biomass circularity», *26th International Congress of Chemical and Process Engineering*, August 2022, Prague, Czech Republic.
- [11] PRIoux, N., G. HÉTREUX ET J.-P. BELAUD. 2022, «Approche centrée données massives et intelligence artificielle pour les systèmes industriels durables et circulaires : application aux trajectoires de valorisation des déchets lignocellulosiques», *Les Avant-Premières du 18ème congrès de la SFGP : Systèmes, éco-systèmes industriels & économie circulaire*, 26 Septembre 2022, France.
- [12] PRIoux, N., G. HÉTREUX ET J.-P. BELAUD. 2022, «Couplage de l'analyse environnementale et de l'apprentissage automatique pour les procédés de valorisation des déchets lignocellulosiques», *18ème Congrès de la SFGP : Sciences et Solutions technologiques pour la Transition (SFGP 2022)*, 7-10 Novembre 2022, Toulouse, France.

Références bibliographiques

- 14040 :2006, I. 2006, «Environmental management - Life cycle Assessment - Principles and Framework», cahier de recherche, International Organization for Standardization, Geneva, Switzerland. 43, 44, 45, 46, 60
- 14044 :2006, I. 2006, «Environmental management - Life cycle Assessment - Requirements and guidelines», cahier de recherche, International Organization for Standardization, Geneva, Switzerland. 43, 68
- ABDELLA, G. M., M. KUCUKVAR, N. C. ONAT, H. M. AL-YAFAY et M. E. BULAK. 2020, «Sustainability Assessment and Modeling Based on Supervised Machine Learning Techniques : The Case for Food Consumption», *Journal of Cleaner Production*, vol. 251, doi :10.1016/j.jclepro.2019.119661., p. 119661. URL <https://doi.org/10.1016/j.jclepro.2019.119661>. 49, 72
- ACCORSI, R., S. CHOLETTE, R. MANZINI et A. TUFANO. 2018, «A Hierarchical Data Architecture for Sustainable Food Supply Chain Management and Planning», *Journal of Cleaner Production*, vol. 203, n° December, doi :10.1016/j.jclepro.2018.08.275., p. 1039–54. URL <https://doi.org/10.1016/j.jclepro.2018.08.275>. 49
- ADEME. 2014, «L'économie circulaire», URL <https://expertises.ademe.fr/expertises/economie-circulaire/leconomie-circulaire>. xiii, 11
- AFIS. 2004, «Glossaire de base de l'Ingénierie Système», cahier de recherche, Association Française de l'Ingénierie Système. xiii, 19, 20
- AHN, B., J. KIM et B. CHOI. 2019, «Artificial intelligence-based machine learning considering flow and temperature of the pipeline for leak early detection using acoustic emission», *Engineering Fracture Mechanics*, vol. 210, doi :10.1016/j.engfracmech.2018.03.010, p. 381–392, ISSN 0013-7944. URL <https://www.sciencedirect.com/science/article/pii/S0013794417312742>. 40
- AL-FUQAHA, A., M. GUIZANI, M. MOHAMMADI, M. ALEDHARI et M. AYYASH. 2015, «Internet of Things : A Survey on Enabling Technologies, Protocols, and Applications», *IEEE Communications Surveys & Tutorials*, vol. 17, n° 4, doi :10.1109/COMST.2015.2444095, p. 2347–2376, ISSN 1553-877X. Conference Name : IEEE Communications Surveys & Tutorials. 32
- ALEXOPOULOS, K., S. MAKRIS, V. XANTHAKIS, K. SIPSAS et G. CHRYSSOLOURIS. 2016, «A concept for context-aware computing in manufacturing : the white goods case», *International Journal of Computer Integrated Manufacturing*, vol. 29, n° 8, doi :10.1080/0951192X.2015.1130257, p. 839–849, ISSN 0951-192X. URL <https://doi.org/10.1080/0951192X.2015.1130257>, publisher : Taylor & Francis_eprint : <https://doi.org/10.1080/0951192X.2015.1130257>. 17
- ALPAYDIN, E. 2010, *Introduction to Machine Learning*, the mit press, 2nd edition éd.. 34
- ALPER, K., K. TEKIN, S. KARAGÖZ et A. J. RAGAUSKAS. 2020, «Sustainable energy and fuels from biomass : a review focusing on hydrothermal biomass processing», *Sustainable Energy & Fuels*, vol. 4, n° 9, p. 4390–4414. Publisher : Royal Society of Chemistry. xiv, 122

- ARNAUD, E., L. COOPER, R. SHRESTHA, N. MENDA, R. NELSON, L. MATTEIS, M. SKOFIC, R. BASTOW, P. JAISWAL, L. MUELLER et G. MCLAREN. 2012, «Towards a Reference Plant Trait Ontology for Modeling Knowledge of Plant Traits and Phenotypes», Barcelona, Spain, p. 220–225. URL <https://www.scitepress.org/Link.aspx?doi=10.5220/0004138302200225>. 81
- ARP, R., B. SMITH et A. D. SPEAR. 2015, *Building Ontologies with Basic Formal Ontology*, MIT Press, ISBN 978-0-262-52781-1. Google-Books-ID : AUxQCgAAQBAJ. 81
- AUSSENAC-GILLES, N., J. CHARLET et C. REYNAUD-DELAÎTRE. 2012, «Chapitre 7 - les enjeux de l'ingénierie des connaissances», dans *Information- Interaction-Intelligence : le point sur le I3*, Toulouse, Cépaduès, p. 244–266. 38
- AYRES, R. U. et L. W. AYRES. 2002, *A Handbook of Industrial Ecology*, Elgaronline. IV, V
- AZAPAGIC, A. et R. CLIFT. 1999, «The application of life cycle assessment to process optimisation», *Computers & Chemical Engineering*, vol. 23, n° 10, doi :10.1016/S0098-1354(99)00308-7, p. 1509–1526, ISSN 0098-1354. URL <https://www.sciencedirect.com/science/article/pii/S0098135499003087>. 52
- BABAR, M., F. ARIF, M. A. JAN, Z. TAN et F. KHAN. 2019, «Urban data management system : Towards Big Data analytics for Internet of Things based smart urban environment using customized Hadoop», *Future Generation Computer Systems*, vol. 96, doi :10.1016/j.future.2019.02.035, p. 398–409, ISSN 0167-739X. URL <https://www.sciencedirect.com/science/article/pii/S0167739X18321095>. 32
- BABICEANU, R. et R. SEKR. 2016, «Big data and virtualization for manufacturing cyber-physical systems : a survey of the current status and future outlook», *Comput. Ind.*, vol. 81, p. 128–137. 16
- BAG, S., L. C. WOOD, L. XU, P. DHAMIJA et Y. KAYIKCI. 2020, «Big Data Analytics as an Operational Excellence Approach to Enhance Sustainable Supply Chain Performance», *Resources, Conservation and Recycling*, vol. 153, n° February, doi :10.1016/j.resconrec.2019.104559., p. 104 559. URL <https://doi.org/10.1016/j.resconrec.2019.104559>. 49
- BAJPAI, P. 2016, «Pretreatment of Lignocellulosic Biomass», dans *Pretreatment of Lignocellulosic Biomass for Biofuel Production*, édité par P. Bajpai, SpringerBriefs in Molecular Science, Springer, Singapore, ISBN 978-981-10-0687-6, p. 17–70, doi :10.1007/978-981-10-0687-6_4. URL https://doi.org/10.1007/978-981-10-0687-6_4. 122
- BAKSHI, B. R. et J. FIKSEL. 2003, «The quest for sustainability : Challenges for process systems engineering», *AIChE Journal*, vol. 49, n° 6, p. 1350, ISSN 0001-1541. URL https://www.academia.edu/1910346/The_quest_for_sustainability_Challenges_for_process_systems_engineering. 51
- BARAKAT, A., H. DE VRIES et X. ROUAU. 2013, «Dry fractionation process as an important step in current and future lignocellulose biorefineries : A review», *Bioresource Technology*, vol. 134, doi :10.1016/j.biortech.2013.01.169, p. 362–373, ISSN 09608524. URL <https://linkinghub.elsevier.com/retrieve/pii/S0960852413002010>. 117, 125

- BELAUD, J.-P., C. ADOUE, C. SABLAYROLLES, C. VIALLE et A. CHORRO. 2017, «Decision Making Approach for Industrial Ecology : Layout and Commercialization of an Industrial Park», *Chemical Engineering Transactions*, vol. 57, doi :10.3303/CET1757261, p. 1561. URL <https://hal.archives-ouvertes.fr/hal-02407082>. 52
- BELAUD, J.-P., C. ADOUE, C. VIALLE, A. CHORRO et C. SABLAYROLLES. 2019a, «A circular economy and industrial ecology toolbox for developing an eco-industrial park : perspectives from French policy», *Clean Technologies and Environmental Policy*, vol. 21, n° 5, doi :10.1007/s10098-019-01677-1, p. 967–985, ISSN 1618-9558. URL <https://doi.org/10.1007/s10098-019-01677-1>. 11, 23
- BELAUD, J.-P., N. PRIoux, C. VIALLE et C. SABLAYROLLES. 2019b, «Big data for agri-food 4.0 : Application to sustainability management for by-products supply chain», *Computers in Industry*, vol. 111, doi :10.1016/j.compind.2019.06.006, p. 41–50, ISSN 01663615. URL <https://linkinghub.elsevier.com/retrieve/pii/S0166361518306419>. 33
- BENDRE, M. R., R. C. THOOL et V. R. THOOL. 2015, «Big data in precision agriculture : Weather forecasting for future farming», dans *2015 1st International Conference on Next Generation Computing Technologies (NGCT)*, p. 744–750, doi :10.1109/NGCT.2015.7375220. 119
- BERNSTEIN P. A., MADHAVAN J. et RAHM E. 2011, «Generic schema matching, ten years later», *PVLDB*, vol. 4, n° 11, p. 695–701. 38
- BERRY, M. J. A. et G. LINOFF. 1997, *Data mining techniques: for marketing, sales, and customer support*, Wiley, New York, ISBN 978-0-471-17980-1. OCLC : 36755821. 31
- BHINGE, R., A. SRINIVASAN, S. ROBINSON et D. DORNFELD. 2015, «Data-intensive Life Cycle Assessment (DILCA) for Deteriorating Products», *Procedia CIRP*, vol. 29, doi :10.1016/j.procir.2015.02.192, p. 396–401, ISSN 22128271. URL <https://linkinghub.elsevier.com/retrieve/pii/S2212827115005132>. 50
- BOEHMKE, B. et B. M. GREENWELL. 2019, *Hands-On Machine Learning with R*, Chapman and Hall/-CRC, New York, ISBN 978-0-367-81637-7, doi :10.1201/9780367816377. xiv, 97
- BORODIN, V., J. BOURTEMBourg, F. HNAIEN et N. LABADIE. 2016, «Handling uncertainty in agricultural supply chain management : A state of the art», *European Journal of Operational Research*, vol. 254, n° 2, doi :10.1016/j.ejor.2016.03.057, p. 348–359, ISSN 0377-2217. URL <https://www.sciencedirect.com/science/article/pii/S037722171630203X>. 118
- BOULARAS, N. et C. DJALAB. 2020, *Système de la prévision mensuelle de la consommation d'énergie électrique basée sur les réseaux de neurones artificielle-etude cas la ville de m'sila*, thèse de doctorat. 91
- BOUYSSOU, D., éd.. 2006, *Evaluation and decision models with multiple criteria : stepping stones for the analyst*, n° 86 dans *International series in operations research & management science*, Springer, New York, ISBN 978-0-387-31098-5 978-0-387-31099-2. 73, 154

- BRETTEL, M., N. FRIEDERICHSEN, M. KELLER et M. ROSENBERG. 2014, «How virtualization, decentralization and network building change the manufacturing landscape : an Industry 4.0 perspective», *Int. J. Mech. Ind. Sci. Eng.*, vol. 8, n° 1, p. 37–44. 17
- BREYSSE, A., A. BARAKAT, H. ANGELLIER-COUSSY, C. MAYER-LAIGLE et P. BUCHE. 2017, «A termino-ontological resource to compare ligno-cellulosic biomass and agro-waste valorisation routes», URL <https://hal-lirmm.ccsd.cnrs.fr/lirmm-01651974>. 127
- BRUNNER, P. H. et H. RECHBERGER. 2004, «Practical handbook of material flow analysis», *The International Journal of Life Cycle Assessment*, vol. 9, n° 5, doi :10.1007/BF02979426, p. 337–338, ISSN 1614-7502. URL <https://doi.org/10.1007/BF02979426>. IV, VI
- BUCHE, P., S. DERVAUX, J. DIBIE-BARTHELEMY, L. IBANESCU, L. SOLER et R. TOUHAMI. 2013a, «Intégration de données hétérogènes et imprécises guidée par une ressource termino-ontologique», *Revue des Sciences et Technologies de l'Information - Série RIA : Revue d'Intelligence Artificielle*, vol. 27, n° 4-5, doi :10.3166/ria27.539-568, p. 539–568. xiv, 86
- BUCHE, P., J. DIBIE-BARTHELEMY, L. IBANESCU et L. SOLER. 2013b, «Fuzzy Web Data Tables Integration Guided by an Ontological and Terminological Resource», *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, n° 4, doi :10.1109/TKDE.2011.245, p. 805–819, ISSN 1558-2191. Conference Name : IEEE Transactions on Knowledge and Data Engineering. 38, 82
- BUI, Q.-C. et P. M. SLOOT. 2011, «Extracting Biological Events from Text Using Simple Syntactic Patterns», Portland, Oregon, USA. 82
- BURGER, J., V. PAPAIOANNOU, S. GOPINATH, G. JACKSON, A. GALINDO et C. S. ADJIMAN. 2015, «A hierarchical method to integrated solvent and process design of physical CO₂ absorption using the SAFT- γ Mie approach», *AIChE Journal*, vol. 61, n° 10, doi :10.1002/aic.14838, p. 3249–3269, ISSN 1547-5905. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/aic.14838>, _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/aic.14838>. 51
- BURGESS, A. et D. BRENNAN. 2001, «Application of life cycle assessment to chemical processes», *Chemical Engineering Science*, vol. 56, n° 8, doi :10.1016/S0009-2509(00)00511-X, p. 2589–2604, ISSN 00092509. URL <https://linkinghub.elsevier.com/retrieve/pii/S000925090000511X>. V
- BUSSET, G., C. C. SABLAYROLLES, M. MONTREJAUD-VIGNOLES, C. VIALLE et J.-P. BELAUD. 2015, «Computer Aided Process Engineering for Sustainability Analysis of Food Production», AIDIC Servizi, ISBN 978-88-95608-34-1, p. 6, doi :10.3303/CET1543224. URL <https://hal.inrae.fr/hal-02744084>. 53
- CAIARDI, F., J.-P. BELAUD, C. VIALLE, F. MONLAU, S. TAYIBI, A. BARAKAT, A. OUKARROUM, Y. ZEROUAL et C. SABLAYROLLES. 2022, «Waste-to-energy innovative system : Assessment of integrating anaerobic digestion and pyrolysis technologies», *Sustainable Production and Consumption*, doi : 10.1016/j.spc.2022.03.021, ISSN 2352-5509. URL <https://www.sciencedirect.com/science/article/pii/S2352550922000744>. 20, 45, 52

- CALOW, P. P. 1998, *Handbook of Environmental Risk Assessment and Management*, John Wiley 1 Sons. V
- CALVO-SERRANO, R., M. GONZÁLEZ-MIQUEL, S. PAPADOKONSTANTAKIS et G. GUILLÉN-GOSÁLBEZ. 2018, «Predicting the cradle-to-gate environmental impact of chemicals from molecular descriptors and thermodynamic properties via mixed-integer programming», *Computers & Chemical Engineering*, vol. 108, doi :10.1016/j.compchemeng.2017.09.010, p. 179–193, ISSN 0098-1354. URL <https://www.sciencedirect.com/science/article/pii/S0098135417303150>. 51, 53
- CALZOLARI, T., A. GENOVESE et A. BRINT. 2022, «Circular Economy indicators for supply chains : A systematic literature review», *Environmental and Sustainability Indicators*, vol. 13, doi :10.1016/j.indic.2021.100160, p. 100 160, ISSN 2665-9727. URL <https://www.sciencedirect.com/science/article/pii/S2665972721000611>. 12
- CARUS, M. 2016, «The bioeconomy is much more than a circular economy», *Blickwinkel - BRAIN Biotech AG.*, n° 3, p. 12–14. URL <https://www.brain-biotech.com/blickwinkel/circular/the-bioeconomy-is-much-more-than-a-circular-economy/>. 22
- CELLURA, M., S. LONGO et M. MISTRETTA. 2011, «Sensitivity analysis to quantify uncertainty in Life Cycle Assessment : The case study of an Italian tile», *Renewable and Sustainable Energy Reviews*, vol. 15, n° 9, doi :10.1016/j.rser.2011.07.082, p. 4697–4705, ISSN 1364-0321. URL <https://www.sciencedirect.com/science/article/pii/S1364032111003273>. 155
- CHARANIA, I. et X. LI. 2020, «Smart farming : Agriculture’s shift from a labor intensive to technology native industry», *Internet of Things*, vol. 9, doi :10.1016/j.iot.2019.100142, p. 100 142, ISSN 2542-6605. URL <https://www.sciencedirect.com/science/article/pii/S2542660519302471>. 120
- CHARVAT, K., K. C. JUNIOR, T. REZNIK, V. LUKAS, K. JEDLICKA, R. PALMA et R. BERZINS. 2018, «Advanced Visualisation of Big Data for Agriculture as Part of Databio Development», dans *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, p. 415–418, doi : 10.1109/IGARSS.2018.8517556. ISSN : 2153-7003. 119
- CHASSERAY, Y., A.-M. BARTHE-DELANOË, J.-M. LE LANN et S. NEGNY. 2021, «Extraction generique de connaissances à partir de donnees textuelles et mesure de la performance des systèmes d’extraction de relations dans un contexte non supervisee.», Grenoble (à distance), France, p. 660–668. 84
- CHEKROUN, S. 2009, *Commande neuro-floue sans capteur de vitesse d’une machine asynchrone triphasée*, thèse de doctorat. xiv, 92
- CHEN, C. et C.-Y. ZHANG. 2014, «Data-intensive applications, challenges, techniques and technologies : a survey on big data», *Inf. Sci*, vol. 275, p. 314–347. 19, 32
- CHEN, H., R. H. L. CHIANG et V. C. STOREY. 2012, «Business Intelligence and Analytics : From Big Data to Big Impact», *MIS Quarterly*, vol. 36, n° 4, p. 1165–1188. 77

- CHOU, J.-S., Y. TAI et L.-J. CHANG. 2010, «Predicting the development cost of TFT-LCD manufacturing equipment with artificial intelligence models», *International Journal of Production Economics*, vol. 128, n° 1, doi :10.1016/j.ijpe.2010.07.031, p. 339–350, ISSN 0925-5273. URL <https://www.sciencedirect.com/science/article/pii/S0925527310002744>. 40
- CHUETOR, S., R. LUQUE, C. BARRON, A. SOLHY, X. ROUAU et A. BARAKAT. 2015, «Innovative combined dry fractionation technologies for rice straw valorization to biofuels», *Green Chemistry*, vol. 17, n° 2, doi :10.1039/C4GC01718H, p. 926–936, ISSN 1463-9262, 1463-9270. URL <http://xlink.rsc.org/?DOI=C4GC01718H>. 117
- CIMIANO P., BUITELAAR P. et VÖLKER J. 2010, «Ontology construction», dans *Handbook of natural language processing*, édité par N. Indurkha et E.J. Damerou, 2^e éd., CRC Press, Taylor and Francis Group, Boca Raton, FL, ISBN 978-1-4200-8592-1, p. 577–604. 38
- CLIFT, R., R. FRISCHNECHT, G. HUPPES, A.-M. TILLMAN et B. WLIDEMA. 1998, «Toward a Coherent Approach to Life Cycle Inventory Analysis», cahier de recherche, Report of the Working Group on Inventory Enhancement, Brussels. 44
- CMLCA.EU. 2022, «Logiciel CMLCA», URL cmlca.eu. 68
- CNRTL. 2022, «Centre National de Ressources Textuelles et Lexicales», URL <http://www.cnrtl.fr/>. 12
- COBO, S., A. DOMINGUEZ-RAMOS et A. IRABIEN. 2018, «From linear to circular integrated waste management systems : A review of methodological approaches», *Resources, Conservation and Recycling*, vol. 135, doi :10.1016/j.resconrec.2017.08.003, p. 279–295, ISSN 0921-3449. URL <https://www.sciencedirect.com/science/article/pii/S0921344917302422>. 52
- COMMENGE, J.-M. 2020, *Big Data et Intelligence Artificielle pour le Génie des Procédés*, école nationale supérieure des industries chimiques éd., France. 39
- COMMENGE, J.-M., L. FALK et J. F. PORTHA. 2017, «Process Intensification for successful energetic transition.», . 51
- COMMISSION, E. 2010, «International Reference Life Cycle Data System (ILCD) Handbook - General guide for Life Cycle Assessment - Provisions and Action Steps», Text. URL <https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/international-reference-life-cycle-data-system-ilcd-handbook-general-guide-life-cycle>. 64
- COMMISSION, F. S. R. T. E. 2012, «Continous Environmental Improvement - Final Report», cahier de recherche. 117
- COOPER, J., M. NOON, C. JONES, E. KAHN et P. ARBUCKLE. 2013, «Big Data in Life Cycle Assessment», *Journal of Industrial Ecology*, vol. 17, n° 6, doi :10.1111/jiec.12069, p. 796–799, ISSN 1530-9290. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/jiec.12069>, _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jiec.12069>. 33, 50

- COP25. 2019, «La conférence de l'ONU sur le Changement climatique», URL <https://unfccc.int/fr/cop25>. 9
- CORBY O., DIENG-KUNTZ R., FARON-ZUCKER C. et GANDON F. L. 2006, «Searching the SemanticWeb : Approximate Query Processing Based on Ontologies», *IEEE Intelligent Systems*, vol. 21, n° 1, p. 20–27. 38
- CORNUÉJOLS, A., L. MICLET, Y. KODRATOFF et T. MITCHELL. 2002, «Apprentissage artificiel : Concepts et algorithmes (EYROLLES)», *Eyrolles*. 36
- CORTES, C. et V. VAPNIK. 1995, «Support-vector networks», dans *Machine Learning*, p. 273–297,. 36
- COULLET-DEMAIZIÈRE, C. 2010, *Au-delà de l'écologie : développement durable*, Afnor, ISBN 978-2-12-465252-5. Google-Books-ID : en6JSQAACAAJ. 8
- COUNCIL, N. R. 2003, *Beyond the Molecular Frontier : Challenges for Chemistry and Chemical Engineering*, ISBN 978-0-309-08477-2, doi :10.17226/10633. URL <https://nap.nationalacademies.org/catalog/10633/beyond-the-molecular-frontier-challenges-for-chemistry-and-chemical-engineering>. 51
- COX, T. F. et M. A. A. COX. 2001, *Multidimensional scaling*, Chapman & Hall/CRC, Boca Raton, ISBN 978-1-58488-094-3 978-1-4200-3612-1 978-0-367-80170-0. URL <http://www.crcnetbase.com/isbn/9781584880943>, oCLC : 152674155. 113
- CRICELLI, L. et M. GRIMALDI. 2008, «A dynamic view of knowledge and information : a stock and flow based methodology», *International Journal of Management and Decision Making*, vol. 9, n° 6, doi :10.1504/IJMDM.2008.021221, p. 686–698, ISSN 1462-4621. URL <https://www.inderscienceonline.com/doi/abs/10.1504/IJMDM.2008.021221>, publisher : Inderscience Publishers. 30
- CUÉLLAR-FRANCA, R. M. et A. AZAPAGIC. 2015, «Carbon capture, storage and utilisation technologies : A critical analysis and comparison of their life cycle environmental impacts», *Journal of CO2 Utilization*, vol. 9, doi :10.1016/j.jcou.2014.12.001, p. 82–102, ISSN 2212-9820. URL <https://www.sciencedirect.com/science/article/pii/S2212982014000626>. 52
- CYBENKO, G. 1989, «Approximation by superpositions of a sigmoidal function», *Mathematics of Control, Signals and Systems*, vol. 2, n° 4, doi :10.1007/BF02551274, p. 303–314, ISSN 1435-568X. URL <https://doi.org/10.1007/BF02551274>. 92, 93
- DATABASE, U. L. C. I. 2022, «U.S. Life Cycle Inventory Database», URL <https://www.nrel.gov/lci/>. 64
- DAVIS, C. B., G. AID et B. ZHU. 2017, «Secondary Resources in the Bio-Based Economy : A Computer Assisted Survey of Value Pathways in Academic Literature», *Waste and Biomass Valorization*, vol. 8, n° 7, doi :10.1007/s12649-017-9975-0, p. 2229–2246, ISSN 1877-2641, 1877-265X. URL <http://link.springer.com/10.1007/s12649-017-9975-0>. 117

- DAZY, F., J.-F. L. BARZIC, G. SAPORTA et F. LAVALLARD. 1996, *L'analyse des données évolutives : méthodes et applications*, Editions Technip, ISBN 978-2-7108-0700-1. Google-Books-ID : fSvvAAAA-MAAJ. 109
- DE MAURO, A., M. GRECO et M. GRIMALDI. 2016, «A formal definition of Big Data based on its essential features», *Library Review*, vol. 65, n° 3, doi :10.1108/LR-06-2015-0061, p. 122–135, ISSN 0024-2535. URL <https://www.emerald.com/insight/content/doi/10.1108/LR-06-2015-0061/full/html>. 30, 31
- DEMPSTER, A. P. 1967, «Upper and Lower Probabilities Induced by a Multivalued Mapping», *The Annals of Mathematical Statistics*, vol. 38, n° 2, doi :10.1214/aoms/1177698950, p. 325–339, ISSN 0003-4851, 2168-8990. URL <https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-38/issue-2/Upper-and-Lower-Probabilities-Induced-by-a-Multivalued-Mapping/10.1214/aoms/1177698950.full>, publisher : Institute of Mathematical Statistics. 88
- DESTERCKE, S., P. BUCHE et B. CHARNOMORDIC. 2013, «Evaluating Data Reliability : An Evidential Answer with Application to a Web-Enabled Data Warehouse», *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, n° 1, doi:10.1109/TKDE.2011.179, p. 92–105, ISSN 1041-4347. URL <http://ieeexplore.ieee.org/document/5989808/>. 38, 87
- DIBIE, J., S. DERVAUX, E. DORIOT, L. IBANESCU et C. PÉNICAUD. 2016, «[MS]²O – A Multi-scale and Multi-step Ontology for Transformation Processes : Application to Micro-Organisms», dans *Graph-Based Representation and Reasoning*, vol. 9717, édité par O. Haemmerlé, G. Stapleton et C. Faron Zucker, Springer International Publishing, Cham, ISBN 978-3-319-40984-9 978-3-319-40985-6, p. 163–176, doi :10.1007/978-3-319-40985-6_13. URL http://link.springer.com/10.1007/978-3-319-40985-6_13. 83
- DIFFEY, S. M., A. B. SMITH, A. H. WELSH et B. R. CULLIS. 2017, «A new REML (parameter expanded) EM algorithm for linear mixed models», *Australian & New Zealand Journal of Statistics*, vol. 59, n° 4, doi :10.1111/anzs.12208, p. 433–448, ISSN 1467-842X. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/anzs.12208>, _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1111/anzs.12208>. 112
- DIWEKAR, U. 2005, «Green process design, industrial ecology, and sustainability : A systems analysis perspective», *Resources, Conservation and Recycling*, vol. 44, n° 3, doi:10.1016/j.resconrec.2005.01.007, p. 215–235, ISSN 0921-3449. URL <https://www.sciencedirect.com/science/article/pii/S0921344905000212>. 51
- DOAN, A., A. HALEVY et Z. IVES. 2012, *Principles of Data Integration*, Morgan Kaufmann Publishers. 38
- DOSHI-VELEZ, F. et B. KIM. 2017, «Towards a rigorous science of interpretable machine learning», *arXiv preprint arXiv :1702.08608*. 94

- DURAN, M. A. et I. E. GROSSMANN. 1986, «Simultaneous optimization and heat integration of chemical processes», *AIChE Journal*, vol. 32, n° 1, doi :10.1002/aic.690320114, p. 123–138, ISSN 1547-5905. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/aic.690320114>, _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/aic.690320114>. 51
- EARLES, J. M. et A. HALOG. 2011, «Consequential life cycle assessment : a review», *The International Journal of Life Cycle Assessment*, vol. 16, n° 5, doi :10.1007/s11367-011-0275-9, p. 445–453, ISSN 1614-7502. URL <https://doi.org/10.1007/s11367-011-0275-9>. 43
- EEA. 2018, «The circular economy and the bioeconomy — Partners in sustainability», cahier de recherche 8/2018, European Environment Agency. 22
- EL-HALWAGI, M. M. 1997, *Pollution Prevention through Process Integration : Systematic Design Tools*, Elsevier, ISBN 978-0-08-051418-5. 51
- EL ZANT, C., Q. CHARRIER, K. BENFRIHA et P. LE MEN. 2021, «Enhanced Manufacturing Execution System “MES” Through a Smart Vision System», dans *Advances on Mechanics, Design Engineering and Manufacturing III*, édité par L. Roucoules, M. Paredes, B. Eynard, P. Morer Camo et C. Rizzi, Lecture Notes in Mechanical Engineering, Springer International Publishing, Cham, ISBN 978-3-030-70566-4, p. 329–334, doi :10.1007/978-3-030-70566-4_52. 16
- ELAHI, E., C. WEIJUN, H. ZHANG et M. NAZEER. 2019, «Agricultural intensification and damages to human health in relation to agrochemicals : Application of artificial intelligence», *Land Use Policy*, vol. 83, doi :10.1016/j.landusepol.2019.02.023, p. 461–474, ISSN 0264-8377. URL <https://www.sciencedirect.com/science/article/pii/S0264837718308548>. 120
- ELGHALI, L., R. CLIFT, P. SINCLAIR, C. PANOUTSOU et A. BAUEN. 2007, «Developing a sustainability framework for the assessment of bioenergy systems», *Energy Policy*, vol. 35, n° 12, doi :10.1016/j.enpol.2007.08.036, p. 6075–6083, ISSN 03014215. URL <https://linkinghub.elsevier.com/retrieve/pii/S030142150700362X>. 44, 61
- ER-RAKI. 2021, *Etudes des impacts environnementaux pour l'évaluation dynamique des filières industrielles : application au déploiement des scénarios de la transition énergétique*, thèse de doctorat, Toulouse INPT. 47
- ESTER, M., H.-P. KRIEGEL, J. SANDER et X. XU. 1996, «A density-based algorithm for discovering clusters in large spatial databases with noise», p. 226–231. URL <https://dl.acm.org/doi/10.5555/3001460.3001507>. 112
- ESTESO, A., M. ALEMANY et A. ORTIZ. 2018, «Conceptual framework for designing agri-food supply chains under uncertainty by mathematical programming models», *International Journal of Production Research*, vol. 56, n° 13, doi :10.1080/00207543.2018.1447706, p. 4418–4446, ISSN 0020-7543. URL <https://doi.org/10.1080/00207543.2018.1447706>, publisher : Taylor & Francis _eprint : <https://doi.org/10.1080/00207543.2018.1447706>. 118
- EUGENE, E. A., W. A. PHILLIP et A. W. DOWLING. 2019, «Data Science-Enabled Molecular-to-Systems Engineering for Sustainable Water Treatment», *Current Opinion in Chemical Engineering, Energy*,

- Environment & Sustainability : Sustainability Modeling Reaction engineering and catalysis : Green Reaction Engineering*, vol. 26, n° December, doi :10.1016/j.coche.2019.10.002., p. 122–30. URL <https://doi.org/10.1016/j.coche.2019.10.002>. 49
- EUROPEAN COMMISSION. 2019, «The European Green Deal», URL https://ec.europa.eu/info/sites/info/files/european-green-deal-communication_en.pdf. 9
- FERNANDEZ RIVAS, D., D. C. BOFFITO, J. FARIA-ALBANESE, J. GLASSEY, N. AFRAZ, H. AKSE, K. V. K. BOODHOO, R. BOS, J. CANTIN, Y. W. (EMILY) CHIANG, J.-M. COMMENGE, J.-L. DUBOIS, F. GALLI, J. P. G. DE MUSSY, J. HARMSSEN, S. KALRA, F. J. KEIL, R. MORALES-MENENDEZ, F. J. NAVARRO-BRULL, T. NOËL, K. OGDEN, G. S. PATIENCE, D. REAY, R. M. SANTOS, A. SMITH-SCHOETTKER, A. I. STANKIEWICZ, H. VAN DEN BERG, T. VAN GERVEN, J. VAN GESTEL, M. VAN DER STELT, M. VAN DE VEN et R. S. WEBER. 2020, «Process intensification education contributes to sustainable development goals. Part 1», *Education for Chemical Engineers*, vol. 32, doi :10.1016/j.ece.2020.04.003, p. 1–14, ISSN 1749-7728. URL <https://www.sciencedirect.com/science/article/pii/S1749772820300257>. 51
- FERNÁNDEZ-MARTÍNEZ, J. L. et Z. FERNÁNDEZ-MUÑIZ. 2020, «The curse of dimensionality in inverse problems», *Journal of Computational and Applied Mathematics*, vol. 369, doi :10.1016/j.cam.2019.112571, p. 112571, ISSN 0377-0427. URL <https://www.sciencedirect.com/science/article/pii/S037704271930576X>. 107
- FESCHET, P. 2014, *Analyse du cycle de vie sociale : Pour un nouveau cadre conceptuel et theorique*, thèse de doctorat, Université Montpellier 1. 46
- FINKBEINER, M., E. M. SCHAU, A. LEHMANN et M. TRAVERSO. 2010, «Towards Life Cycle Sustainability Assessment», *Sustainability*, vol. 2, n° 10, p. 3309–3322. 47
- FINNVEDEN, G., M. Z. HAUSCHILD, T. EKVAL, J. GUINÉE, R. HEIJUNGS, S. HELLWEG, A. KOEHLER, D. PENNINGTON et S. SUH. 2009, «Recent developments in Life Cycle Assessment», *Journal of Environmental Management*, vol. 91, n° 1, doi :10.1016/j.jenvman.2009.06.018, p. 1–21, ISSN 0301-4797. URL <https://www.sciencedirect.com/science/article/pii/S0301479709002345>. III
- FOUNDATION, E. M. 2015, «Towards a circular economy : business rationale for an accelerated transition.», cahier de recherche, Ellen Macarthur Foundation. 10
- FRIEDMAN, J. H. 2001, «Greedy function approximation : A gradient boosting machine», *Annals of statistics*, p. 1189–1232. 95
- FRISCHKNECHT, R., N. JUNGBLUTH, H.-J. ALTHAUS, G. DOKA, R. DONES, T. HECK, S. HELLWEG, R. HIRSCHIER, T. NEMECEK, G. REBITZER et M. SPIELMANN. 2005, «Theecoinvent Database : Overview and Methodological Framework (7 pp)», *The International Journal of Life Cycle Assessment*, vol. 10, n° 1, doi :10.1065/lca2004.10.181.1, p. 3–9, ISSN 1614-7502. URL <https://doi.org/10.1065/lca2004.10.181.1>. xvii, 62, 64, 69, 71
- GANDOMI, A. et M. HAIDER. 2015, «Beyond the hype : Big data concepts, methods, and analytics», *International Journal of Information Management*, vol. 35, n° 2, doi :10.1016/j.ijinfomgt.2014.10.

- 007, p. 137–144, ISSN 0268-4012. URL <https://www.sciencedirect.com/science/article/pii/S0268401214001066>. 32
- GARCIA, J., D. MILLET et P. TONNELIER. 2015, «A tool to evaluate the impacts of an innovation on a product's recyclability rate by adopting a modular approach : automotive sector application», *International Journal of Vehicle Design*, vol. 2015, n° 1, p. 1. URL <https://hal.archives-ouvertes.fr/hal-01064949>. 52
- GARNETT, T., M. C. APPLEBY, A. BALMFORD, I. J. BATEMAN, T. G. BENTON, P. BLOOMER, B. BURLINGAME, M. DAWKINS, L. DOLAN, D. FRASER, M. HERRERO, I. HOFFMANN, P. SMITH, P. K. THORNTON, C. TOULMIN, S. J. VERMEULEN et H. C. J. GODFRAY. 2013, «Sustainable Intensification in Agriculture : Premises and Policies», *Science*, vol. 341, n° 6141, doi :10.1126/science.1234485, p. 33–34, ISSN 0036-8075, 1095-9203. URL <http://www.sciencemag.org/lookup/doi/10.1126/science.1234485>. 117
- GEISSDOERFER, M., P. SAVAGET, N. M. P. BOCKEN et E. J. HULTINK. 2017, «The Circular Economy – A new sustainability paradigm?», *Journal of Cleaner Production*, vol. 143, doi :10.1016/j.jclepro.2016.12.048, p. 757–768, ISSN 0959-6526. URL <https://www.sciencedirect.com/science/article/pii/S0959652616321023>. 10
- GHISELLINI, P., C. CIALANI et S. ULGIATI. 2016, «A review on circular economy : the expected transition to a balanced interplay of environmental and economic systems», *Journal of Cleaner Production*, vol. 114, doi :10.1016/j.jclepro.2015.09.007, p. 11–32, ISSN 09596526. URL <https://linkinghub.elsevier.com/retrieve/pii/S0959652615012287>. 9
- GHOLIZADEH, H., H. FAZLOLLAHTABAR et M. KHALILZADEH. 2020, «A Robust Fuzzy Stochastic Programming for Sustainable Procurement and Logistics under Hybrid Uncertainty Using Big Data», *Journal of Cleaner Production*, vol. 258, n° June, doi :10.1016/j.jclepro.2020.120640., p. 120 640. URL <https://doi.org/10.1016/j.jclepro.2020.120640>. 49
- GIBBS, A. L. et F. E. SU. 2002, «On choosing and bounding probability metrics», cahier de recherche arXiv :math/0209021, arXiv. URL <http://arxiv.org/abs/math/0209021>, arXiv :math/0209021 type : article. 36
- GIEC. 2014, «Changements climatiques 2014 : Rapport de synthèse. Contribution des Groupes de travail I, II et III au cinquième Rapport d'évaluation du Groupe d'experts intergouvernemental sur l'évolution du climat», cahier de recherche, GIEC, Genève, Suisse. 7
- GIEC. 2018, «2018 : Résumé à l'intention des décideurs, Réchauffement planétaire de 1,5 °C, Rapport spécial du GIEC sur les conséquences d'un réchauffement planétaire de 1,5 °C par rapport aux niveaux préindustriels et les trajectoires associées d'émissions mondiales de gaz à effet de serre, dans le contexte du renforcement de la parade mondiale au changement climatique, du développement durable et de la lutte contre la pauvreté», cahier de recherche, Genève, Suisse. 8
- GILLANI, S. T., J.-P. BELAUD, C. SABLAYROLLES, M. VIGNOLES et J.-M. L. LANN. 2010, «Review of Life Cycle Assessment in Agro-Chemical Processes», *Chemical Product and Process Modeling*,

- vol. 5, n° 1, doi :10.2202/1934-2659.1496, ISSN 1934-2659. URL <https://www.degruyter.com/document/doi/10.2202/1934-2659.1496/html>, publisher : De Gruyter. 52
- GLIGOR, A., C.-D. DUMITRU et H.-S. GRIF. 2018, «Artificial intelligence solution for managing a photovoltaic energy production unit», *Procedia Manufacturing*, vol. 22, doi :10.1016/j.promfg.2018.03.091, p. 626–633, ISSN 2351-9789. URL <https://www.sciencedirect.com/science/article/pii/S2351978918303871>. 40
- GOKILA, S., K. A. KUMAR et A. BHARATHI. 2016, «Different versions of K-Mean Clustering in complete set of numerical data points», vol. 2, n° 7, p. 6. 111
- GONG, Y., J. LIU et J. ZHU. 2019, «When to Increase Firms' Sustainable Operations for Efficiency? A Data Envelopment Analysis in the Retailing Industry», *European Journal of Operational Research*, vol. 277, n° 3, doi :10.1016/j.ejor.2019.03.019., p. 1010–26. URL <https://doi.org/10.1016/j.ejor.2019.03.019>. 49, 53
- GOOSENS, J. et P. RICHARD. 2013, *Handbook of cyber-physical systems., Multiprocessor Real-Time Scheduling*, vol. 250127. URL <http://hdl.handle.net/2013/ULB-DIPOT:oai:dipot.ulb.ac.be,250127>. 18
- GREENWELL, B. M. 2017, «pdp : An R Package for Constructing Partial Dependence Plots», *greenwellpdp*. 96
- GROSSMANN, I. 2005, «Enterprise-wide optimization : A new frontier in process systems engineering», *AIChE Journal*, vol. 51, n° 7, doi :10.1002/aic.10617, p. 1846–1857, ISSN 1547-5905. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/aic.10617>, _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/aic.10617>. 52
- GROSSMANN, I. E., J. A. CABALLERO et H. YEOMANS. 1999, «Mathematical programming approaches to the synthesis of chemical process systems», *Korean Journal of Chemical Engineering*, vol. 16, n° 4, p. 407, ISSN 0256-1115. URL https://www.academia.edu/81730911/Mathematical_programming_approaches_to_the_synthesis_of_chemical_process_systems. 51
- GROSSMANN, I. E. et I. HARJUNKOSKI. 2019, «Process Systems Engineering : Academic and industrial perspectives», *Computers & Chemical Engineering*, vol. 126, doi :10.1016/j.compchemeng.2019.04.028, p. 474–484, ISSN 0098-1354. URL <https://www.sciencedirect.com/science/article/pii/S009813541930078X>. 51
- GROSSMANN, I. E. et A. W. WESTERBERG. 2000, «Research challenges in process systems engineering», *AIChE Journal*, vol. 46, n° 9, doi :10.1002/aic.690460902, p. 1700–1703, ISSN 1547-5905. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/aic.690460902>, _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/aic.690460902>. 51
- GRUBER, T. R. 1993, «A translation approach to portable ontology specifications», *Knowledge acquisition*, vol. 5, n° 2, p. 199–220. 79

- GUARINO, N., D. OBERLE et S. STAAB. 2009, «What Is an Ontology?», dans *Handbook on Ontologies*, édité par S. Staab et R. Studer, International Handbooks on Information Systems, Springer, Berlin, Heidelberg, ISBN 978-3-540-92673-3, p. 1–17, doi :10.1007/978-3-540-92673-3_0. URL https://doi.org/10.1007/978-3-540-92673-3_0. 79
- GUINÉE, J. B., éd.. 2002, *Handbook on life cycle assessment : operational guide to the ISO standards*, n° v. 7 dans *Eco-efficiency in industry and science*, Kluwer Academic Publishers, Dordrecht; Boston, ISBN 978-1-4020-0228-1. 61
- GUINÉE, J. B., R. HEIJUNGS, G. HUPPES, A. ZAMAGNI, P. MASONI, R. BUONAMICI, T. EKVALL et T. RYDBERG. 2011, «Life Cycle Assessment : Past, Present, and Future», *Environmental Science & Technology*, vol. 45, n° 1, doi :10.1021/es101316v, p. 90–96, ISSN 0013-936X. URL <https://pubs.acs.org/doi/10.1021/es101316v>, publisher : American Chemical Society. 42, 46, III
- VAN HAASTER, B., A. CIROTH, J. FONTES, R. WOOD et A. RAMIREZ. 2017, «Development of a methodological framework for social life-cycle assessment of novel technologies», *The International Journal of Life Cycle Assessment*, vol. 22, n° 3, doi :10.1007/s11367-016-1162-1, p. 423–440, ISSN 1614-7502. URL <https://doi.org/10.1007/s11367-016-1162-1>. 46
- HACKING, T. et P. GUTHRIE. 2008, «A framework for clarifying the meaning of Triple Bottom-Line, Integrated, and Sustainability Assessment», *Environmental Impact Assessment Review*, vol. 28, n° 2, doi :10.1016/j.eiar.2007.03.002, p. 73–89, ISSN 0195-9255. URL <https://www.sciencedirect.com/science/article/pii/S0195925507000297>. 14
- HAGAN, M. T., H. B. DEMUTH, M. HUDSON BEALE et O. DE JESUS. 2014, *Neural network design - 2nd edition*. 91
- HAGEMANN, S., A. SÜNNETCIOGLU et R. STARK. 2019, «Hybrid Artificial Intelligence System for the Design of Highly-Automated Production Systems», *Procedia Manufacturing*, vol. 28, doi : 10.1016/j.promfg.2018.12.026, p. 160–166, ISSN 2351-9789. URL <https://www.sciencedirect.com/science/article/pii/S2351978918313684>. 40
- HAO, Y., X. ZHU, M. HUANG et M. LI. 2005, «Discovering patterns to extract protein–protein interactions from the literature : Part II», *Bioinformatics*, vol. 21, n° 15, doi :10.1093/bioinformatics/bti493, p. 3294–3300, ISSN 1367-4803. URL <https://doi.org/10.1093/bioinformatics/bti493>. 82
- HAUSCHILD, M., M. GOEDKOOP, J. GUINÉE, R. HEIJUNGS, M. HUIJBREGTS, O. JOLLIET, M. MARGNI, A. DE SCHRYVER, S. HUMBERT, A. LAURENT, S. SALA et R. PANT. 2013, «Identifying best existing practice for characterization modeling in life cycle impact assessment», *The International Journal of Life Cycle Assessment*, vol. 18, n° 3, doi :10.1007/s11367-012-0489-5, p. 683–697, ISSN 0948-3349. 45
- HAUSCHILD, M. Z., R. K. ROSENBAUM et S. I. OLSEN. 2018, *Life Cycle Assessment : Theory and Practice*, springer international publishing éd.. 45, 64, 68
- HAWIZY, L., D. M. JESSOP, N. ADAMS et P. MURRAY-RUST. 2011, «ChemicalTagger : A tool for semantic text-mining in chemistry», *J Cheminform*, vol. 3. 82

- HEALTH, G. A., D. D. HSU, D. INMAN, A. ADEN et M. K. MANN. 2009, «Life Cycle Assessment of the Energy Independence and Security Act of 2007 : Ethanol - Global Warming Potential and Environmental Emissions», URL <https://www.osti.gov/biblio/964200-life-cycle-assessment-energy-independence-security-act-ethanol-global-warming-potential> 50
- HEATH T. et BIZER C. 2011, *Linked Data : Evolving the Web into a Global Data Space*, vol. 1, Morgan & Claypool. Issue : 1. 38
- HEINTZ, J., J.-P. BELAUD et V. GERBAUD. 2014, «Chemical enterprise model and decision-making framework for sustainable chemical product design», *Computers in Industry*, vol. 65, n° 3, doi : 10.1016/j.compind.2014.01.010, p. 505–520, ISSN 0166-3615. URL <https://www.sciencedirect.com/science/article/pii/S0166361514000268>. 32, 52
- HELLWEG, S. et L. MILÀ I CANALS. 2014, «Emerging approaches, challenges and opportunities in life cycle assessment», *Science*, vol. 344, n° 6188, doi :10.1126/science.1248361, p. 1109–1113. URL <https://www.science.org/doi/abs/10.1126/science.1248361>, publisher : American Association for the Advancement of Science. 52
- HINTON, G. et T. J. SEJNOWSKI. 1999, *Unsupervised Learning : Foundations of Neural Computation*, MIT Press, ISBN 978-0-262-58168-4. Google-Books-ID : yj04Y0lje4cC. 36, 110
- HIRA, S. et P. S. DESHPANDE. 2015, «Data Analysis using Multidimensional Modeling, Statistical Analysis and Data Mining on Agriculture Parameters», *Procedia Computer Science*, vol. 54, doi : 10.1016/j.procs.2015.06.050, p. 431–439, ISSN 1877-0509. URL <https://www.sciencedirect.com/science/article/pii/S1877050915013745>. 118
- HORNIK, K. 1991, «Approximation capabilities of multilayer feedforward networks», *Neural Networks*, vol. 4, n° 2, doi :10.1016/0893-6080(91)90009-T, p. 251–257, ISSN 0893-6080. URL <https://www.sciencedirect.com/science/article/pii/089360809190009T>. 92, 93
- HOU, P., O. JOLLIET, J. ZHU et M. XU. 2020, «Estimate ecotoxicity characterization factors for chemicals in life cycle assessment using machine learning models», *Environment International*, vol. 135, doi :10.1016/j.envint.2019.105393, p. 105393, ISSN 0160-4120. URL <https://www.sciencedirect.com/science/article/pii/S0160412019314412>. 40
- HUANG, M., X. ZHU, D. G. PAYAN, K. QU et M. LI. 2004, «Discovering patterns to extract protein-protein interactions from full biomedical texts», p. 22–28. URL <https://dl.acm.org/doi/10.5555/1567594.1567599>. 82
- HUANG, Y. L. et L. T. FAN. 1993, «Artificial intelligence for waste minimization in the process industry», *Computers in Industry*, vol. 22, n° 2, doi :10.1016/0166-3615(93)90059-A, p. 117–128, ISSN 0166-3615. URL <https://www.sciencedirect.com/science/article/pii/S016636159390059A>. 39

- HUIJBREGTS, M. A. J., Z. J. N. STEINMANN, P. M. F. ELSHOUT, G. STAM, F. VERONES, M. VIEIRA, M. ZIJP, A. HOLLANDER et R. VAN ZELM. 2017, «ReCiPe2016 : a harmonised life cycle impact assessment method at midpoint and endpoint level», *The International Journal of Life Cycle Assessment*, vol. 22, n° 2, doi :10.1007/s11367-016-1246-y, p. 138–147, ISSN 1614-7502. URL <https://doi.org/10.1007/s11367-016-1246-y>. xiv, xvii, 103, 105
- IATE, I. I. 2022, «@Web», URL <https://ico.iate.inra.fr/atWeb/>. 128
- INVENTORY, E. L. C. 2022, «EcoInvent Life Cycle Inventory Database», URL <http://www.ecoinvent.org/>. 126
- ISLAM SARKER, M. N., M. WU, B. CHANTHAMITH, S. YUSUFZADA, D. LI et J. ZHANG. 2019, «Big Data Driven Smart Agriculture : Pathway for Sustainable Development», dans *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, p. 60–65, doi :10.1109/ICAIBD.2019.8836982. 119
- ISO 14050 :2020. 2020, «Management environnemental — Vocabulaire», cahier de recherche, International Organization for Standardization, Geneva, Switzerland. URL <https://www.iso.org/cms/render/live/fr/sites/isoorg/contents/data/standard/07/53/75300.html>. 8, 9, 11
- ISO/IEC/IEEE 15288 :2015. 2015, «Ingénierie des systèmes et du logiciel — Processus du cycle de vie du système», URL <https://www.iso.org/cms/render/live/fr/sites/isoorg/contents/data/standard/06/37/63711.html>. 9, 20, 21
- JACQUEMIN, L., P.-Y. PONTALIER et C. SABLAYROLLES. 2012, «Life cycle assessment (LCA) applied to the process industry : a review», *Int. J. Life Cycle Assess*, vol. 17, p. 1028–1041. 43, 52
- JOGLEKAR, S. N., A. P. TANDULJE, S. A. MANDAVGANE et B. D. KULKARNI. 2019, «Environmental Impact Study of Bagasse Valorization Routes», *Waste and Biomass Valorization*, vol. 10, n° 7, doi :10.1007/s12649-018-0198-9, p. 2067–2078, ISSN 1877-2641, 1877-265X. URL <http://link.springer.com/10.1007/s12649-018-0198-9>. 117
- JOLLIET, O., M. SAADÉ et P. CRETZAZ. 2010, *Analyse du cycle de vie : comprendre et réaliser un écobilan*, PUR Presses polytechniques, ISBN 978-2-88074-886-9. Google-Books-ID : g9S55CklsOoC. xvii, 43, 44, 46, 64, 67, 68, 69, 103
- JULIO, R., J. ALBET, C. VIALLE, C. VACA-GARCIA et C. SABLAYROLLES. 2017, «Sustainable design of biorefinery processes : existing practices and new methodology», *Biofuels, Bioproducts and Biorefining*, vol. 11, n° 2, doi :10.1002/bbb.1749, p. 373–395, ISSN 1932-1031. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/bbb.1749>, eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/bbb.1749>. 52, 53
- JØRGENSEN, A., M. FINKBEINER, M. S. JØRGENSEN et M. Z. HAUSCHILD. 2010, «Defining the baseline in social life cycle assessment», *The International Journal of Life Cycle Assessment*, vol. 15, n° 4, doi :10.1007/s11367-010-0176-3, p. 376–384, ISSN 1614-7502. URL <https://doi.org/10.1007/s11367-010-0176-3>. 46

- KAAB, A., M. SHARIFI, H. MOBLI, A. NABAVI-PELESARAEI et K.-W. CHAU. 2019, «Combined Life Cycle Assessment and Artificial Intelligence for Prediction of Output Energy and Environmental Impacts of Sugarcane Production», *Science of The Total Environment*, vol. 664, n° May, doi :10.1016/j.scitotenv.2019.02.004., p. 1005–19. URL <https://doi.org/10.1016/j.scitotenv.2019.02.004>. 48
- KAGERMANN, H., W. WAHLSTER et J. HELBIG. 2013, «Recommendations for Implementing the Strategic Initiative Industry 4.0 - Final Report of the Industry 4.0 Working Group», cahier de recherche. Place : Acatech, München. 18
- KALIA, V., D. I. WALKER, K. M. KRASNODEMSKI, D. P. JONES, G. W. MILLER et M.-A. KIOUMOURTZOGLOU. 2020, «Unsupervised dimensionality reduction for exposome research», *Current Opinion in Environmental Science & Health*, vol. 15, doi :10.1016/j.coesh.2020.05.001, p. 32–38, ISSN 2468-5844. URL <https://www.sciencedirect.com/science/article/pii/S2468584420300349>. 109
- KAPLAN, A. et M. HAENLEIN. 2019, «Siri, Siri, in my hand : Who’s the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence», *Business Horizons*, vol. 62, n° 1, doi:10.1016/j.bushor.2018.08.004, p. 15–25, ISSN 0007-6813. URL <https://www.sciencedirect.com/science/article/pii/S0007681318301393>. xiii, 34
- KARUPPIAH, R. et I. E. GROSSMANN. 2006, «Global optimization for the synthesis of integrated water systems in chemical processes», *Computers & Chemical Engineering*, vol. 30, n° 4, doi :10.1016/j.compchemeng.2005.11.005, p. 650–673, ISSN 0098-1354. URL <https://www.sciencedirect.com/science/article/pii/S0098135405002991>. 51
- KIM, B., R. KHANNA et O. O. KOYEJO. 2016, «Examples are not enough, learn to criticize! Criticism for Interpretability», dans *Advances in Neural Information Processing Systems*, vol. 29, Curran Associates, Inc. URL <https://papers.nips.cc/paper/2016/hash/5680522b8e2bb01943234bce7bf84534-Abstract.html>. 94
- KIRCHHERR, J. 2022, «Circular economy and growth : A critical review of “post-growth” circularity and a plea for a circular economy that grows», *Resources, Conservation and Recycling*, vol. 179, doi :10.1016/j.resconrec.2021.106033, p. 106 033, ISSN 0921-3449. URL <https://www.sciencedirect.com/science/article/pii/S0921344921006418>. 10
- KLASHANOV, F. 2016, «Artificial Intelligence and Organizing Decision in Construction», *Procedia Engineering*, vol. 165, doi :10.1016/j.proeng.2016.11.813, p. 1016–1020, ISSN 1877-7058. URL <https://www.sciencedirect.com/science/article/pii/S1877705816341741>. 40
- KLATT, K.-U. et W. MARQUARDT. 2009, «Perspectives for process systems engineering—Personal views from academia and industry», *Computers & Chemical Engineering*, vol. 33, n° 3, doi:10.1016/j.compchemeng.2008.09.002, p. 536–550, ISSN 0098-1354. URL <https://www.sciencedirect.com/science/article/pii/S0098135408001737>. 50

- KLOEPFFER, W. 2008, «Life cycle sustainability assessment of products», *The International Journal of Life Cycle Assessment*, vol. 13, n° 2, doi :10.1065/lca2008.02.376, p. 89, ISSN 1614-7502. URL <https://doi.org/10.1065/lca2008.02.376>. 46, 47
- KOPNINA, H. 2019, «Green-washing or best case practices? Using circular economy and Cradle to Cradle case studies in business education», *Journal of Cleaner Production*, vol. 219, doi : 10.1016/j.jclepro.2019.02.005, p. 613–621, ISSN 09596526. URL <https://linkinghub.elsevier.com/retrieve/pii/S0959652619303841>. 60
- KRICHEN, M. 2021, «Détection des anomalies environnementales via les capteurs des smartphones», Research Report, Sfax University - ReDCAD Laboratory. URL <https://hal.archives-ouvertes.fr/hal-03214225>. 15
- KULIN, M., T. KAZAZ, E. DE POORTER et I. MOERMAN. 2021, «A Survey on Machine Learning-Based Performance Improvement of Wireless Networks : PHY, MAC and Network Layer», *Electronics*, vol. 10, n° 3, doi :10.3390/electronics10030318, p. 318. xiii, 29
- KUMAR, L. S. P. 2017, «State of The Art-Intense Review on Artificial Intelligence Systems Application in Process Planning and Manufacturing», *Engineering Applications of Artificial Intelligence*, vol. 65, doi :10.1016/j.engappai.2017.08.005, p. 294–329, ISSN 0952-1976. URL <https://www.sciencedirect.com/science/article/pii/S0952197617301896>. 34
- LANCE, G. N. et W. T. WILLIAMS. 1966, «Computer programs for hierarchical polythetic classification ("similarity analyses").», *The Computer Journal*, vol. 9, n° 1, p. 60–64. 112
- LE, S. T. K., M. MONGO, A. TANGUY et V. LAFOREST. 2022, «European innovation policies towards resource-constrained innovations In favour of circular economy», URL <https://hal-emse.ccsd.cnrs.fr/emse-03663954>. 11
- LE MINH, Q., S. N. TRUONG et Q. H. BAO. 2011, «A Pattern Approach for Biomedical Event Annotation», Portland, Oregon, USA. 82
- LEE, E. K., X. ZHANG, P. R. ADLER, G. S. KLEPPEL et X. X. ROMEIKO. 2020, «Spatially and temporally explicit life cycle global warming, eutrophication, and acidification impacts from corn production in the U.S. Midwest», *Journal of Cleaner Production*, vol. 242, doi :10.1016/j.jclepro.2019.118465, p. 118 465, ISSN 09596526. URL <https://linkinghub.elsevier.com/retrieve/pii/S0959652619333359>. 72
- LEGAT, C. et B. VOGEL-HEUSER. 2017, «A configurable partial-order planning approach for field level operation strategies of PLC-based industry 4.0 automated manufacturing systems», *Engineering Applications of Artificial Intelligence*, vol. 66, doi :10.1016/j.engappai.2017.06.014, p. 128–144, ISSN 0952-1976. URL <https://www.sciencedirect.com/science/article/pii/S0952197617301331>. 40
- LEHMANN, A., E. ZSCHIESCHANG, M. TRAVERSO, M. FINKBEINER et L. SCHEBEK. 2013, «Social aspects for sustainability assessment of technologies—challenges for social life cycle assessment (SLCA)»,

- The International Journal of Life Cycle Assessment*, vol. 18, n° 8, doi :10.1007/s11367-013-0594-0, p. 1581–1592, ISSN 1614-7502. URL <https://doi.org/10.1007/s11367-013-0594-0>. 46
- LEZOUCHE, M., J. E. HERNANDEZ, M. D. M. E. ALEMANY DÍAZ, H. PANETTO et J. KACPRZYK. 2020, «Agri-food 4.0 : A survey of the supply chains and technologies for the future agriculture», *Computers in Industry*, vol. 117, doi :10.1016/j.compind.2020.103187, p. 103 187, ISSN 0166-3615. URL <http://www.sciencedirect.com/science/article/pii/S0166361519307584>. 118
- LI, X., D. LI, J. WAN, A. V. VASILAKOS, C.-F. LAI et S. WANG. 2017, «A review of industrial wireless networks in the context of Industry 4.0», *Wireless Networks*, vol. 23, n° 1, doi :10.1007/s11276-015-1133-7, p. 23–41, ISSN 1572-8196. URL <https://doi.org/10.1007/s11276-015-1133-7>. 17
- LINKE, B. S., D. R. GARCIA, A. KAMATH et I. C. GARRETSON. 2019, «Data-Driven Sustainability in Manufacturing : Selected Examples», *Procedia Manufacturing*, vol. 33, doi :10.1016/j.promfg.2019.04.075., p. 602–9. URL <https://doi.org/10.1016/j.promfg.2019.04.075>. 48
- LINNHOFF, B. 1993, «Pinch analysis - a state-of-the-art overview», *Chemical Engineering Research and Design; (United Kingdom)*, vol. 71 :A5. URL <https://www.osti.gov/etdeweb/biblio/5876781>. 52
- LIOUTAS, E. D. et C. CHARATSARI. 2020, «Big data in agriculture : Does the new oil lead to sustainability?», *Geoforum*, vol. 109, doi :10.1016/j.geoforum.2019.12.019, p. 1–3, ISSN 0016-7185. URL <https://www.sciencedirect.com/science/article/pii/S0016718519303562>. 119
- LIU, C., E. VAN DER HEIDE, H. WANG, B. LI, G. YU et X. MU. 2013, «Alkaline twin-screw extrusion pretreatment for fermentable sugar production», *Biotechnology for Biofuels*, vol. 6, n° 1, doi :10.1186/1754-6834-6-97, p. 97, ISSN 1754-6834. URL <http://biotechnologyforbiofuels.biomedcentral.com/articles/10.1186/1754-6834-6-97>. xv, 128, 129
- LOISEAU, E. 2014, *Elaboration d'une démarche d'évaluation environnementale d'un territoire basée sur le cadre méthodologique de l'Analyse du Cycle de vie (ACV)*, thèse de doctorat, Centre international d'études supérieures en sciences agronomiques, Montpellier. 47, 52
- LOUSTEAU-CAZALET, C., A. BARAKAT, J.-P. BELAUD, P. BUCHE, G. BUSSET, B. CHARNOMORDIC, S. DERVAUX, S. DESTERCKE, J. DIBIE, C. SABLAYROLLES et C. VIALLE. 2016, «A decision support system for eco-efficient biorefinery process comparison using a semantic approach», *Computers and Electronics in Agriculture*, vol. 127, doi :10.1016/j.compag.2016.06.020, p. 351–367, ISSN 01681699. URL <https://linkinghub.elsevier.com/retrieve/pii/S0168169916304136>. xiv, 33, 83, 88, 128
- LÓPEZ-DÍAZ, D. C., L. F. LIRA-BARRAGÁN, E. RUBIO-CASTRO, M. SERNA-GONZÁLEZ, M. M. EL-HALWAGI et J. M. PONCE-ORTEGA. 2018, «Optimization of biofuels production via a water–energy–food nexus framework», *Clean Technologies and Environmental Policy*, vol. 20, n° 7, doi :10.1007/s10098-017-1395-0, p. 1443–1466, ISSN 1618-9558. URL <https://doi.org/10.1007/s10098-017-1395-0>. 52

- MINISTÈRE DE L'ÉCOLOGIE, D. D. D. E. D. L. F. 2021, «Article L211-2», . 121
- MACMILLAN, M., C. A. MURPHY et M. D. BAZILIAN. 2022, «Exploring acute weather resilience : Meeting resilience and renewable goals», *Renewable and Sustainable Energy Reviews*, vol. 168, doi : 10.1016/j.rser.2022.112841, p. 112 841, ISSN 1364-0321. URL <https://www.sciencedirect.com/science/article/pii/S1364032122007249>. 10
- MALLIER, L., G. HÉTREUX, R. THERY-HÉTREUX et P. BAUDET. 2021, «A modelling framework for energy system planning : Application to CHP plants participating in the electricity market», *Energy*, vol. 214, doi :10.1016/j.energy.2020.118976, p. 118976, ISSN 0360-5442. URL <https://www.sciencedirect.com/science/article/pii/S0360544220320831>. 52
- MANCENO, F. 2010, *Le développement durable*, U, Armand Colin, Paris, ISBN 978-2-200-24411-8. URL <https://www.cairn.info/le-developpement-durable--9782200244118.htm>. 8
- MANGINA, E., P. K. NARASIMHAN, M. SAFFARI et I. VLACHOS. 2020, «Data Analytics for Sustainable Global Supply Chains», *Journal of Cleaner Production*, vol. 255, n° May, doi :10.1016/j.jclepro.2020.120300., p. 120 300. URL <https://doi.org/10.1016/j.jclepro.2020.120300>. 50
- MANUILOVA, A., J. SUEBSIRI et M. WILSON. 2009, «Should Life Cycle Assessment be part of the Environmental Impact Assessment? Case study : EIA of CO2 Capture and Storage in Canada», *Energy Procedia*, vol. 1, n° 1, doi :10.1016/j.egypro.2009.02.269, p. 4511–4518, ISSN 1876-6102. URL <https://www.sciencedirect.com/science/article/pii/S1876610209009126>. V
- MAO, S., B. WANG, Y. TANG et F. QIAN. 2019, «Opportunities and Challenges of Artificial Intelligence for Green Manufacturing in the Process Industry», *Engineering*, vol. 5, n° 6, doi :10.1016/j.eng.2019.08.013, p. 995–1002, ISSN 2095-8099. URL <https://www.sciencedirect.com/science/article/pii/S2095809919300074>. 39
- MAYER, A. L. 2008, «Strengths and weaknesses of common sustainability indices for multidimensional systems», *Environment International*, vol. 34, n° 2, doi :10.1016/j.envint.2007.09.004, p. 277–291, ISSN 0160-4120. URL <https://www.sciencedirect.com/science/article/pii/S0160412007001833>. 14
- MCCARTHY, J. 2007, «From here to human-level AI», *Artificial Intelligence*, vol. 171, n° 18, doi : 10.1016/j.artint.2007.10.009, p. 1174–1182, ISSN 0004-3702. URL <https://www.sciencedirect.com/science/article/pii/S0004370207001476>. 33
- MCCULLOCH, W. S. et W. PITTS. 1943, «A logical calculus of the ideas immanent in nervous activity», *The bulletin of mathematical biophysics*, vol. 5, n° 4, p. 115–133,. 36
- MCKENDRY, P. 2002, «Energy production from biomass (part 1) : overview of biomass», *Bioresource Technology*, vol. 83, n° 1, doi :10.1016/S0960-8524(01)00118-3, p. 37–46, ISSN 0960-8524. URL <https://www.sciencedirect.com/science/article/pii/S0960852401001183>. 122
- MEINADIER, J.-P. 2003, *Le métier d'intégration de systèmes*, Hermès science publications, ISBN 978-2-7462-0596-3. Google-Books-ID : CiBLPQAACAAJ. 19

- MENDOZA, G. A., P. MACOUN, R. PRABHU, D. SUKADRI, H. PURNOMO et H. HARTANTO. 2000, «Application de l'analyse multicritère à l'évaluation des critères et indicateurs», cahier de recherche 9, CIRAD-Foret and CIFOR, Montpellier, France. xiii, 13
- MENTEN, F., B. CHÈZE, L. PATOUILLARD et F. BOUVART. 2013, «A Review of LCA Greenhouse Gas Emissions Results for Advanced Biofuels : The Use of Meta-Regression Analysis», *Renewable and Sustainable Energy Reviews*, vol. 26, n° October, doi :10.1016/j.rser.2013.04.021., p. 108–34. URL <https://doi.org/10.1016/j.rser.2013.04.021>. 48
- MEZGHANI, E., E. EXPOSITO, K. DRIRA, M. DA SILVEIRA et C. PRUSKI. 2015, «A Semantic Big Data Platform for Integrating Heterogeneous Wearable Data in Healthcare», *Journal of Medical Systems*, vol. 39, n° 12, doi :10.1007/s10916-015-0344-x, p. 185, ISSN 1573-689X. URL <https://doi.org/10.1007/s10916-015-0344-x>. 32
- MILLATI, R., R. B. CAHYONO, T. ARIYANTO, I. N. AZZAHrani, R. U. PUTRI et M. J. TAHERZADEH. 2019, «Agricultural, Industrial, Municipal, and Forest Wastes», dans *Sustainable Resource Recovery and Zero Waste Approaches*, Elsevier, ISBN 978-0-444-64200-4, p. 1–22, doi :10.1016/B978-0-444-64200-4.00001-3. URL <https://linkinghub.elsevier.com/retrieve/pii/B9780444642004000013>. 117
- MILLET, D., N. PERRY, S. LEPOCHAT, S. ZINCK et Y. LEROY. 2016, «Efficient and successful design methods and tools for sustainable industrial systems», *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 10, n° 3, doi :10.1007/s12008-016-0315-6, p. 211–212, ISSN 1955-2505. URL <https://doi.org/10.1007/s12008-016-0315-6>. 51
- MINARD, A.-L., A.-L. LIGOZAT et B. GRAU. 2011, «Multi-class SVM for Relation Extraction from Clinical Reports», Association for Computational Linguistic, Hissar, Bulgaria, p. 604–609. 82
- MING, X., H. CAI et S. LIANG. 2015, «Big Data and Industrial Ecology», *Journal of Industrial Ecology*, vol. 19, n° 2, doi :10.1111/jiec.12241, p. 205–210, ISSN 10881980. URL <http://doi.wiley.com/10.1111/jiec.12241>. 23, 32
- MITCHELL, T. M. 1997, *Machine Learning*, mcgraw-hill science/engineering/math éd., MacG. 36
- MOHEBALI, S., S. MAGHSOUDY et F. D. ARDEJANI. 2020, «Application of Data Envelopment Analysis in Environmental Impact Assessment of a Coal Washing Plant : A New Sustainable Approach», *Environmental Impact Assessment Review*, vol. 83, n° July, doi :10.1016/j.eiar.2020.106389. URL <https://doi.org/10.1016/j.eiar.2020.106389>. 48
- MORALES-MENDOZA, L. F., C. AZZARO-PANTEL, J.-P. BELAUD et A. OUATTARA. 2018, «Coupling life cycle assessment with process simulation for ecodesign of chemical processes», *Environmental Progress & Sustainable Energy*, vol. 37, n° 2, doi :10.1002/ep.12723, p. 777–796, ISSN 1944-7450. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/ep.12723>, _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/ep.12723>. 11, 51
- MURDOCH, W. J., C. SINGH, K. KUMBIER, R. ABBASI-ASL et B. YU. 2019, «Definitions, methods, and applications in interpretable machine learning», *Proceedings of the National Academy of Sciences*,

- vol. 116, n° 44, doi :10.1073/pnas.1900654116, p. 22 071–22 080. URL <https://www.pnas.org/doi/10.1073/pnas.1900654116>, publisher : Proceedings of the National Academy of Sciences. 94
- MURRAY, A., K. SKENE et K. HAYNES. 2017, «The Circular Economy : An Interdisciplinary Exploration of the Concept and Application in a Global Context», *Journal of Business Ethics*, vol. 140, n° 3, doi :10.1007/s10551-015-2693-2, p. 369–380, ISSN 1573-0697. URL <https://doi.org/10.1007/s10551-015-2693-2>. 24
- NATIONS, U. 2022, «UN unveils data partnership with Twitter in support of Global Goals», URL <https://www.un.org/en/academic-impact/un-unveils-data-partnership-twitter-support-global-goals>. 9, 15
- NEGNY, S., J. P. BELAUD, G. CORTES ROBLES, E. ROLDAN REYES et J. B. FERRER. 2012, «Toward an eco-innovative method based on a better use of resources : application to chemical process preliminary design», *Journal of Cleaner Production*, vol. 32, doi :10.1016/j.jclepro.2012.03.023, p. 101–113, ISSN 0959-6526. URL <https://www.sciencedirect.com/science/article/pii/S0959652612001709>. 11, 51
- NEGRA, C., R. REMANS, S. ATTWOOD, S. JONES, F. WERNECK et A. SMITH. 2020, «Sustainable Agri-Food Investments Require Multi-Sector Co-Development of Decision Tools», *Ecological Indicators*, vol. 110, n° March, doi :10.1016/j.ecolind.2019.105851., p. 105 851. URL <https://doi.org/10.1016/j.ecolind.2019.105851>. 49
- NESS, B., E. URBEL-PIRSALU, S. ANDERBERG et L. OLSSON. 2007, «Categorising tools for sustainability assessment», *Ecological Economics*, vol. 60, n° 3, doi :10.1016/j.ecolecon.2006.07.023, p. 498–508, ISSN 0921-8009. URL <https://www.sciencedirect.com/science/article/pii/S0921800906003636>. 12, 14
- NOY, N. F. 2004, «Semantic integration : a survey of ontology-based approaches», *ACM SIGMOD Record*, vol. 33, n° 4, doi :10.1145/1041410.1041421, p. 65, ISSN 01635808. URL <http://portal.acm.org/citation.cfm?doid=1041410.1041421>. 38
- NY, H., J. P. MACDONALD, G. BROMAN, R. YAMAMOTO et K.-H. ROBÉRT. 2006, «Sustainability Constraints as System Boundaries : An Approach to Making Life-Cycle Management Strategic», *Journal of Industrial Ecology*, vol. 10, n° 1-2, doi :10.1162/108819806775545349, p. 61–77, ISSN 1530-9290. URL <https://onlinelibrary.wiley.com/doi/abs/10.1162/108819806775545349>, _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1162/108819806775545349>. 12
- OECD. 2003, «OECD environmental indicators - development, measurement and use», cahier de recherche, Paris. 14
- ONU. 2022, «Les mégadonnées au service du développement | Nations Unies», URL <https://www.un.org/fr/global-issues/big-data-for-sustainable-development>, publisher : United Nations. 14

- PANG, W. Y., J. J. QING, Q. L. LIU et G. Z. NONG. 2020, «Developing an Artificial Intelligence (AI) System to Patch Plywood Defects in Manufacture», *Procedia Computer Science*, vol. 166, doi :10.1016/j.procs.2020.02.036, p. 139–143, ISSN 1877-0509. URL <https://www.sciencedirect.com/science/article/pii/S1877050920301587>. 39
- PARK, J.-H. et K.-K. SEO. 2003, «Approximate life cycle assessment of product concepts using multiple regression analysis and artificial neural networks», *KSME International Journal*, vol. 17, n° 12, p. 1969–1976. 50
- PATRÍCIO, D. I. et R. RIEDER. 2018, «Computer vision and artificial intelligence in precision agriculture for grain crops : A systematic review», *Computers and Electronics in Agriculture*, vol. 153, doi :10.1016/j.compag.2018.08.001, p. 69–81, ISSN 0168-1699. URL <https://www.sciencedirect.com/science/article/pii/S0168169918305829>. 120
- PENNINGTON, D. W., J. POTTING, G. FINNVEDEN, E. LINDEIJER, O. JOLLIET, T. RYDBERG et G. REBITZER. 2004, «Life cycle assessment Part 2 : Current impact assessment practice», *Environment International*, vol. 30, n° 5, doi :10.1016/j.envint.2003.12.009, p. 721–739, ISSN 0160-4120. URL <https://www.sciencedirect.com/science/article/pii/S0160412003002551>. III
- PEREZ, S. et K. MAZEAU. 2005, *2 Conformations, Structures, and Morphologies of Celluloses*, polysaccharides (2nd edition) éd., New York. 122
- POWER, D. J. 2002, «Ask Dan! - What is the "true story" about data mining, beer and diapers?», *DSS News*, 3(23). URL <http://www.dssresources.com/newsletters/66.php>. 31
- PRIOUX, N., J.-P. BELAUD et G. HETREUX. 2021, «Intensive Data and Knowledge-based Approach for Sustainable and Circular Industrial Systems», *INSIGHT*, vol. 24, n° 4, doi :10.1002/inst.12363, p. 43–46, ISSN 2156-4868. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/inst.12363>, _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/inst.12363>. xv, 125
- PTASINSKI, K. J. 2016, *Efficiency of Biomass Energy : An Exergy Approach to Biofuels, Power, and Bio-refineries*, General & Introductory Chemical Engineering, Wiley, ISBN 978-1-118-70210-9. xvii, 121, 127
- PURVIS, B., Y. MAO et D. ROBINSON. 2019, «Three pillars of sustainability : in search of conceptual origins», *Sustainability Science*, vol. 14, n° 3, doi :10.1007/s11625-018-0627-5, p. 681–695, ISSN 1862-4057. URL <https://doi.org/10.1007/s11625-018-0627-5>. xiii, 7
- QUINLAN, J. 1986, «Induction of decision trees», *Machine Learning*, vol. 1, n° 1, p. 81–106,. 36
- RAJA, K., S. SUBRAMANI et J. NATARAJAN. 2013, «PPInterFinder—a mining tool for extracting causal relations on human proteins from literature», *Database*, vol. 2013, doi :10.1093/database/bas052, p. bas052, ISSN 1758-0463. URL <https://doi.org/10.1093/database/bas052>. 82
- RAUT, R. D., S. K. MANGLA, V. S. NARWANE, B. B. GARDAS, P. PRIYADARSHINEE et B. E. NARKHEDE. 2019, «Linking Big Data Analytics and Operational Sustainability Practices for Sustainable Business Management», *Journal of Cleaner Production*, vol. 224, n° July, doi :10.1016/j.jclepro.2019.03.181., p. 10–24. URL <https://doi.org/10.1016/j.jclepro.2019.03.181>. 49

- RAYMOND, R. 2012, «Improving food systems for sustainable diets in a green economy», *Greening the Economy with Agriculture*. 117
- REICHMAN, O. J., M. B. JONES et M. P. SCHILDHAUER. 2011, «Challenges and Opportunities of Open Data in Ecology», *Science*, vol. 331, n° 6018, doi :10.1126/science.1197962, p. 703–705, ISSN 0036-8075. Place : Washington Publisher : Amer Assoc Advancement Science WOS :000287205700050. 15
- REINSEL, D., J. GANTZ et J. RYDNING. 2018, «Data Age 2025 : The digitization of the world from Edge to Core», cahier de recherche, An IDC White Paper US44413318. 16, 31
- REYMONET, A., J. THOMAS et N. AUSSENAC-GILLES. 2007, «Modelling ontological and terminological resources in OWL DL», Busan, South-Korea. 81, 84
- RIJGERSBERG, H., M. WIGHAM et J. TOP. 2011, «How semantics can improve engineering processes : A case of units of measure and quantities», *Advanced Engineering Informatics*, vol. 25, n° 2, doi :10.1016/j.aei.2010.07.008, p. 276–287, ISSN 14740346. URL <https://linkinghub.elsevier.com/retrieve/pii/S1474034610000753>. 83
- RITTHOFF, M., H. ROHN et C. LIEDTKE. 2002, «Calculating MIPS : resource productivity of products and services», cahier de recherche, Wuppertal Institute for Climate, Environment and Energy, Wuppertal. ISBN : 3-929944-56-1e Series : Wuppertal Spezial Volume : 27e. VI
- ROCHE, C., M. CALBERG-CHALLOT, L. DAMAS et P. ROUARD. 2009, «ONTOTERMINOLOGY - A New Paradigm for Terminology», Funchal - Madeira, Portugal, p. 321–326. URL <https://www.scitepress.org/Link.aspx?doi=10.5220/0002330803210326>. 81
- ÁLVAREZ RODRÍGUEZ, C., M. MARTÍN-GAMBOA et D. IRIBARREN. 2020, «Sustainability-Oriented Efficiency of Retail Supply Chains : A Combination of Life Cycle Assessment and Dynamic Network Data Envelopment Analysis», dans *Science of The Total Environment* 705, p. 135977, doi :10.1016/j.scitotenv.2019.135977. URL <https://doi.org/10.1016/j.scitotenv.2019.135977>. 48
- ROJKO, A. 2017, «Industry 4.0 concept : Background and overview», *International journal of interactive mobile technologies*, vol. 11, n° 5. 17
- ROMEIKO, X. X., E. K. LEE, Y. SORUNMU et X. ZHANG. 2020, «Spatially and Temporally Explicit Life Cycle Environmental Impacts of Soybean Production in the U.S. Midwest», *Environmental Science & Technology*, vol. 54, n° 8, doi :10.1021/acs.est.9b06874, p. 4758–4768, ISSN 0013-936X, 1520-5851. URL <https://pubs.acs.org/doi/abs/10.1021/acs.est.9b06874>. 72
- ROSARIO, B. et M. A. HEARST. 2004, «Classifying Semantic Relations in Bioscience Texts», p. 430–437. 82
- ROWLEY, J. 2007, «The wisdom hierarchy : representations of the DIKW hierarchy», *Journal of Information Science*, vol. 33, n° 2, doi :10.1177/0165551506070706, p. 163–180. xiii, 30

- SABADASH, A. 2022, «Sustainable development in the European Union — Overview of progress towards the SDGs in an EU context — 2022 edition», cahier de recherche, Eurostat. URL <https://ec.europa.eu/eurostat/fr/web/products-catalogues/-/ks-06-22-017>. 15
- SABARINA, K. et N. PRIYA. 2015, «Lowering Data Dimensionality in Big Data for the Benefit of Precision Agriculture», *Procedia Computer Science*, vol. 48, doi :10.1016/j.procs.2015.04.134, p. 548–554, ISSN 1877-0509. URL <https://www.sciencedirect.com/science/article/pii/S1877050915006432>. 118
- SALA, S., B. CIUFFO et P. NIJKAMP. 2015, «A systemic framework for sustainability assessment», *Ecological Economics*, vol. 119, doi :10.1016/j.ecolecon.2015.09.015, p. 314–325, ISSN 0921-8009. URL <https://www.sciencedirect.com/science/article/pii/S0921800915003821>. 14
- SANTOS, M. Y., J. OLIVEIRA E SÁ, C. COSTA, J. GALVÃO, C. ANDRADE, B. MARTINHO, F. V. LIMA et E. COSTA. 2017, «A Big Data Analytics Architecture for Industry 4.0», dans *Recent Advances in Information Systems and Technologies*, édité par Rocha, A. M. Correia, H. Adeli, L. P. Reis et S. Costanzo, Advances in Intelligent Systems and Computing, Springer International Publishing, Cham, ISBN 978-3-319-56538-5, p. 175–184, doi :10.1007/978-3-319-56538-5_19. 16, 17
- SCHORR, S., M. MÖLLER, J. HEIB, S. FANG et D. BÄHRE. 2020, «Quality Prediction of Reamed Bores Based on Process Data and Machine Learning Algorithm : A Contribution to a More Sustainable Manufacturing», *Procedia Manufacturing, Sustainable Manufacturing - Hand in Hand to Sustainability on Globe : Proceedings of the 17th Global Conference on Sustainable Manufacturing*, vol. 43, n° January, doi :10.1016/j.promfg.2020.02.180., p. 519–26. URL <https://doi.org/10.1016/j.promfg.2020.02.180>. 48
- SHEN, D. K., S. GU, K. H. LUO, S. R. WANG et M. X. FANG. 2010, «The pyrolytic degradation of wood-derived lignin from pulping process», *Bioresource Technology*, vol. 101, n° 15, doi :10.1016/j.biortech.2010.02.078, p. 6136–6146, ISSN 0960-8524. URL <https://www.sciencedirect.com/science/article/pii/S0960852410003895>. 123
- SHIBASAKI, M., M. FISCHER et L. BARTHEL. 2007, «Effects on Life Cycle Assessment — Scale Up of Processes», dans *Advances in Life Cycle Engineering for Sustainable Manufacturing Businesses*, édité par S. Takata et Y. Umeda, Springer London, London, ISBN 978-1-84628-934-7, p. 377–381, doi :10.1007/978-1-84628-935-4_65. URL http://link.springer.com/10.1007/978-1-84628-935-4_65. 60
- SHILA, D. M., W. SHEN, Y. CHENG, X. TIAN, SHEN et X. SHERMAN. 2017, «AMCloud : Toward a Secure Autonomic Mobile Ad Hoc Cloud Computing System», *IEEE Wireless Communications*, vol. 24, n° 2, doi :10.1109/MWC.2016.1500119RP, p. 74–81, ISSN 1558-0687. Conference Name : IEEE Wireless Communications. 18
- SHROUF, F., J. ORDIERES-MERÉ et G. MIRAGLIOTTA. 2014, «Smart Factories in Industry 4.0 : A Review of the Concept and of Energy Management Approached in Production Based on the Internet of Things Paradigm», Selangor Darul Ehsan, Malaysia. URL https://www.researchgate.net/publication/274249276_Smart_Factories_in_Industry_

40_A_Review_of_the_Concept_and_of_Energy_Management_Approached_in_Production_Based_on_the_Internet_of_Things_Paradigm. 16

SHVAIKO P. et EUZENAT J. 2013, «Ontology matching : State of the art and future challenges», *IEEE Trans. Knowl. Data Eng.*, vol. 25, n° 1, p. 158–176. 38

SIKDAR, S. K. 2003, «Sustainable development and sustainability metrics», *AIChE Journal*, vol. 49, n° 8, doi :10.1002/aic.690490802, p. 1928–1932, ISSN 1547-5905. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/aic.690490802>, _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/aic.690490802>. 51

SIMAPRO. 2022, «SimaPro | The world’s leading LCA software», URL <https://simapro.com/>. 65, 66, 68, 69

SINGH, R. K., H. R. MURTY, S. K. GUPTA et A. K. DIKSHIT. 2009, «An overview of sustainability assessment methodologies», *Ecological Indicators*, vol. 9, n° 2, doi :10.1016/j.ecolind.2008.05.011, p. 189–212, ISSN 1470-160X. URL <https://www.sciencedirect.com/science/article/pii/S1470160X08000678>. 12, 14

SJOSTROM, E. 1993, *Wood Chemistry : Fundamentals and Applications*, Gulf Professional Publishing, ISBN 978-0-12-647481-7. Google-Books-ID : Sv3xcS6eS5QC. 123

SLAPNIK, M., D. ISTENIČ, M. PINTAR et A. UDOVČ. 2015, «Extending life cycle assessment normalization factors and use of machine learning – A Slovenian case study», *Ecological Indicators*, vol. 50, doi :10.1016/j.ecolind.2014.10.028, p. 161–172, ISSN 1470-160X. URL <https://www.sciencedirect.com/science/article/pii/S1470160X14005214>. 50

SNIJDERS, C., U. MATZAT et U.-D. REIPS. 2012, «“Big Data” : Big Gaps of Knowledge in the Field of Internet Science», *International Journal of Internet Science*, vol. 7 (1), n° 1, p. 1–5. 32

SOUSA, I. et D. WALLACE. 2006, «Product classification to support approximate life-cycle assessment of design concepts», *Technological Forecasting and Social Change*, vol. 73, n° 3, doi : 10.1016/j.techfore.2004.03.007, p. 228–249, ISSN 0040-1625. URL <https://www.sciencedirect.com/science/article/pii/S0040162505000934>. 50

SPATARI, S., M. BETZ, H. FLORIN, M. BAITZ et M. FALTENBACHER. 2001, «Using GaBi 3 to perform life cycle assessment and life cycle engineering», *The International Journal of Life Cycle Assessment*, vol. 6, n° 2, doi :10.1007/BF02977842, p. 81, ISSN 1614-7502. URL <https://doi.org/10.1007/BF02977842>. 62, 68

SRINIVASULU, P., M. S. BABU, R. VENKAT et K. RAJESH. 2017, «Cloud service oriented architecture (CSOA) for agriculture through internet of things (IoT) and big data», dans *2017 IEEE International Conference on Electrical, Instrumentation and Communication Engineering (ICEICE)*, p. 1–6, doi : 10.1109/ICEICE.2017.8191906. 120

STAAB S. et STUDER R, éd.. 2009, *Handbook on ontologies*, Springer, Berlin Heidelberg. 38

- STREICH, J., J. ROMERO, J. G. F. M. GAZOLLA, D. KAINER, A. CLIFF, E. T. PRATES, J. B. BROWN, S. KHOURY, G. A. TUSKAN, M. GARVIN, D. JACOBSON et A. L. HARFOUCHE. 2020, «Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the United Nations sustainable development goals?», *Current Opinion in Biotechnology*, vol. 61, doi : 10.1016/j.copbio.2020.01.010, p. 217–225, ISSN 0958-1669. URL <https://www.sciencedirect.com/science/article/pii/S0958166920300100>. 120
- STROZZI, F., C. COLICCHIA, A. CREAZZA et C. NOÈ. 2017, «Literature review on the ‘Smart Factory’ concept using bibliometric tools», *International Journal of Production Research*, vol. 55, n° 22, doi :10.1080/00207543.2017.1326643, p. 6572–6591, ISSN 0020-7543. URL <https://doi.org/10.1080/00207543.2017.1326643>, publisher : Taylor & Francis _eprint : <https://doi.org/10.1080/00207543.2017.1326643>. 17
- STUDER, R., V. R. BENJAMINS et D. FENSEL. 1998, «Knowledge engineering : principles and methods», *Data & knowledge engineering*, vol. 25, n° 1-2, p. 161–197. Publisher : Elsevier. 79, 81
- SUN, X., J. YOUNG, J.-H. LIU et D. NEWMAN. 2018, «Prediction of pork loin quality using on-line computer vision system and artificial intelligence model», *Meat Science*, vol. 140, doi :10.1016/j.meatsci.2018.03.005, p. 72–77, ISSN 0309-1740. URL <https://www.sciencedirect.com/science/article/pii/S0309174017314481>. 120
- SWARR, T. E., D. HUNKELER, W. KLOPFER, H.-L. PESONEN, A. CIROTH, A. C. BRENT et R. PAGAN. 2011, *Environmental Life Cycle Costing : A Code of Practice*, ociety of Environmental Toxicology and Chemistry. 46
- SWARTWOOD, J. D. 2013, «Wisdom as an Expert Skill», *Ethical Theory and Moral Practice*, vol. 16, n° 3, doi :10.1007/s10677-012-9367-2, p. 511–528, ISSN 1572-8447. URL <https://doi.org/10.1007/s10677-012-9367-2>. 30
- SZARGUT, J. 2005, *Exergy method : technical and ecological applications*, Wit press. VI
- SÁNCHEZ, A. et M. MARTÍN. 2018, «Optimal renewable production of ammonia from water and air», *Journal of Cleaner Production*, vol. 178, doi :10.1016/j.jclepro.2017.12.279, p. 325–342, ISSN 0959-6526. URL <https://www.sciencedirect.com/science/article/pii/S0959652617332730>. 51
- TABOUBI, F. 2019, «Développement d’un schéma de bioraffinerie de la biomasse lignocellulosique eco-compatible», Rapport de Stage M2, Institut National des Sciences Appliquées et de Technologie - Université de Carthage. 129, XIX
- TAVANA, M., M. PIRDASHTI, D. KENNEDY, J.-P. BELAUD et M. BEHZADIAN. 2012, «A hybrid Delphi-SWOT paradigm for oil and gas pipeline strategic planning in Caspian Sea basin», *Energy Policy*, vol. 40, p. 345–360. 73, 154
- TERZIYAN, V., S. GRYSHKO et M. GOLOVIANKO. 2018, «Patented intelligence : Cloning human decision models for Industry 4.0», *Journal of Manufacturing Systems*, vol. 48, doi :10.1016/j.jmsy.2018.04.

- 019, p. 204–217, ISSN 0278-6125. URL <https://www.sciencedirect.com/science/article/pii/S0278612518300608>. 40
- THABREW, L., A. WIEK et R. RIES. 2009, «Environmental decision making in multi-stakeholder contexts : applicability of life cycle thinking in development planning and implementation», *Journal of Cleaner Production*, vol. 17, n° 1, doi :10.1016/j.jclepro.2008.03.008, p. 67–76, ISSN 0959-6526. URL <https://www.sciencedirect.com/science/article/pii/S0959652608000528>. 14
- TIBSHIRANI, R. 1996, «Regression Shrinkage and Selection Via the Lasso», *Journal of the Royal Statistical Society : Series B (Methodological)*, vol. 58, n° 1, doi :10.1111/j.2517-6161.1996.tb02080.x, p. 267–288, ISSN 2517-6161. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x>, _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1996.tb02080.x>. 108
- TOUHAMI, R., P. BUCHE, J. DIBIE-BARTHÉLEMY et L. IBĂNESCU. 2011, «An Ontological and Terminological Resource for N-ary Relation Annotation in Web Data Tables», dans *Proceedings of the 2011th Confederated International Conference on On the Move to Meaningful Internet Systems - Volume Part II*, OTM'11, Springer-Verlag, Berlin, Heidelberg, ISBN 978-3-642-25105-4, p. 662–679. URL <http://dl.acm.org/citation.cfm?id=2075764.2075791>. 81, 84
- TSOUMIS, G. 1968, *Wood as Raw Material : Source, Structure, Chemical Composition, Growth, Degradation and Identification*, Elsevier Science. 122
- TURING, A. M. 1950, «Computing machinery and intelligence», *Computing Machinery and Intelligence*, vol. 49, p. 433–460. 33
- TURKSON, C., A. ACQUAYE, W. LIU et T. PAPADOPOULOS. 2020, «Sustainability assessment of energy production : A critical review of methods, measures and issues», *Journal of Environmental Management*, vol. 264, doi :10.1016/j.jenvman.2020.110464, p. 110 464, ISSN 0301-4797. URL <https://www.sciencedirect.com/science/article/pii/S0301479720303984>. 12
- UNEP, U. N. E. P. 2011, *Towards a Life Cycle Sustainability Assessment : Making informed choices on products*, catherine benoit & bernard mazijn, montréal ghent éd.. 46
- USCHOLD, M. et M. GRUNINGER. 2009, «Ontologies : principles, methods and applications», *The Knowledge Engineering Review*, vol. 11, n° 2, doi :10.1017/S0269888900007797, p. 93–136. 79
- VAN LANDEGHEM, S., Y. SAEYS, B. DE BAETS et Y. VAN DE PEER. 2009, «Analyzing text in search of bio-molecular events : : a high-precision machine learning framework», p. 128–136. URL <https://dl.acm.org/doi/10.5555/1572340.1572362>. 82
- VAPNIK, V. N. 2000, *The Nature of Statistical Learning Theory*, Springer, New York, NY, ISBN 978-1-4419-3160-3 978-1-4757-3264-1, doi :10.1007/978-1-4757-3264-1. URL <http://link.springer.com/10.1007/978-1-4757-3264-1>. 82

- VENTURA, A. 2022, «Transition life cycle assessment : A new method to face ecological transition», *Frontiers in Sustainability*, vol. 3, ISSN 2673-4524. URL <https://www.frontiersin.org/articles/10.3389/frsus.2022.801668>. 47
- VERDOUW, C., V. NICOLA, H. SENDMAEKER et A. BEULENS. 2014, «Future Internet as a Driver for Virtualization, Connectivity and Intelligence of Agri-Food Supply Chain Networks», *International Journal on Food System Dynamics*, vol. 4, n° 4, p. 1–12. URL <https://ideas.repec.org/a/ags/ijofsd/164812.html>. 120
- WACHTER, S., B. MITTELSTADT et C. RUSSELL. 2018, «Counterfactual explanations without opening the black box : automated decisions and the GDPR», *Harvard Journal of Law & Technology*, vol. 31, n° 2, p. 52. 94
- WALLER, M. A. et S. E. FAWCETT. 2013, «Data Science, Predictive Analytics, and Big Data : A Revolution That Will Transform Supply Chain Design and Management», *Journal of Business Logistics*, vol. 34, n° 2, doi :10.1111/jbl.12010, p. 77–84, ISSN 2158-1592. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/jbl.12010>, eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1111/jbl.12010>. 29
- WANG, Y. P. et R. SMITH. 1994, «Wastewater minimisation», *Chemical Engineering Science*, vol. 49, n° 7, doi :10.1016/0009-2509(94)80006-5, p. 981–1006, ISSN 0009-2509. URL <https://www.sciencedirect.com/science/article/pii/0009250994800065>. 52
- WARING, J., C. LINDVALL et R. UMETON. 2020, «Automated machine learning : Review of the state-of-the-art and opportunities for healthcare», *Artificial Intelligence in Medicine*, vol. 104, doi : 10.1016/j.artmed.2020.101822, p. 101 822, ISSN 0933-3657. URL <https://www.sciencedirect.com/science/article/pii/S0933365719310437>. 35
- WCED. 1987, «World commission on environment and development. Our common future.», *Oxford University Press*, vol. 17, n° 1, p. 1–91. 7
- WEIDEMA, B. P. 2006, «The Integration of Economic and Social Aspects in Life Cycle Impact Assessment», *The International Journal of Life Cycle Assessment*, vol. 11, n° 1, doi:10.1065/lca2006.04.016, p. 89–96, ISSN 1614-7502. URL <https://doi.org/10.1065/lca2006.04.016>. 46
- WEIDEMA, B. P. et M. S. WESNÆS. 1996, «Data quality management for life cycle inventories—an example of using data quality indicators», *Journal of Cleaner Production*, vol. 4, n° 3, doi :10.1016/S0959-6526(96)00043-1, p. 167–174, ISSN 0959-6526. URL <https://www.sciencedirect.com/science/article/pii/S0959652696000431>. 69
- WERNET, G., S. HELLWEG, U. FISCHER, S. PAPADOKONSTANTAKIS et K. HUNGERBÜHLER. 2008, «Molecular-Structure-Based Models of Chemical Inventories using Neural Networks», *Environmental Science & Technology*, vol. 42, n° 17, doi:10.1021/es7022362, p. 6717–6722, ISSN 0013-936X. URL <https://pubs.acs.org/doi/10.1021/es7022362>, publisher : American Chemical Society.

- WEST, P. C., J. S. GERBER, P. M. ENGSTROM, N. D. MUELLER, K. A. BRAUMAN, K. M. CARLSON, E. S. CASSIDY, M. JOHNSTON, G. K. MACDONALD, D. K. RAY et S. SIEBERT. 2014, «Leverage points for improving global food security and the environment», *Science*, vol. 345, n° 6194, doi : 10.1126/science.1246067, p. 325–328, ISSN 0036-8075, 1095-9203. URL <http://www.sciencemag.org/cgi/doi/10.1126/science.1246067>. 117
- WU, B., F. TIAN, M. ZHANG, H. ZENG et Y. ZENG. 2020, «Cloud Services with Big Data Provide a Solution for Monitoring and Tracking Sustainable Development Goals», *Geography and Sustainability*, vol. 1, n° 1, doi :10.1016/j.geosus.2020.03.006., p. 25–32. URL <https://doi.org/10.1016/j.geosus.2020.03.006>. 49
- XU, D. 2020, «Sustainability Prioritization of Energy Systems by Developing an Integrated Decision Support Framework with Hybrid-Data Consideration», *Sustainable Energy Technologies and Assessments*, vol. 39, n° June, doi :10.1016/j.seta.2020.100719., p. 100 719. URL <https://doi.org/10.1016/j.seta.2020.100719>. 48
- YANG, L. T., B. DI MARTINO et Q. ZHANG. 2017, «Internet of Everything», *Mobile Information Systems*, vol. 2017, doi :10.1155/2017/8035421, p. 8035 421, ISSN 1574-017X. URL <https://doi.org/10.1155/2017/8035421>, publisher : Hindawi. 18
- YE, Y., X. LU et T. LU. 2020, «Examining the Spillover Effect of Sustainable Consumption on Microloan Repayment : A Big Data-Based Research», *Information & Management, February*, vol. 103288, doi : 10.1016/j.im.2020.103288. URL <https://doi.org/10.1016/j.im.2020.103288>. 49
- YOSSY, E. H., J. PRANATA, T. WIJAYA, H. HERMAWAN et W. BUDIHARTO. 2017, «Mango Fruit Sortation System using Neural Network and Computer Vision», *Procedia Computer Science*, vol. 116, doi :10.1016/j.procs.2017.10.013, p. 596–603, ISSN 1877-0509. URL <https://www.sciencedirect.com/science/article/pii/S1877050917320501>. 120
- YOU, F. et B. WANG. 2011, «Life Cycle Optimization of Biomass-to-Liquid Supply Chains with Distributed–Centralized Processing Networks», *Industrial & Engineering Chemistry Research*, vol. 50, n° 17, doi :10.1021/ie200850t, p. 10 102–10 127, ISSN 0888-5885. URL <https://pubs.acs.org/doi/10.1021/ie200850t>, publisher : American Chemical Society. 52
- ZAMAGNI, A., J. B. GUINÉE, R. HEIJUNGS, P. MASONI et A. RAGGI. 2012, «Lights and shadows in consequential LCA», *The International Journal of Life Cycle Assessment*, vol. 17, n° 7, doi :10.1007/s11367-012-0423-x, p. 904–918, ISSN 0948-3349. URL <http://www.scopus.com/inward/record.url?scp=84868201490&partnerID=8YFLogxK>. 43
- ZANERO, S. 2017, «Cyber-Physical Systems», *Computer*, vol. 50, n° 4, doi :10.1109/MC.2017.105, p. 14–16, ISSN 1558-0814. Conference Name : Computer. 18
- ZANGHELINI, G. M., E. CHERUBINI et S. R. SOARES. 2018, «How Multi-Criteria Decision Analysis (MCDA) Is Aiding Life Cycle Assessment (LCA) in Results Interpretation», *Journal of Cleaner Production*, vol. 172, n° January, doi :10.1016/j.jclepro.2017.10.230., p. 609–22. URL <https://doi.org/10.1016/j.jclepro.2017.10.230>. 48

- ZHANG, P., Q. ZHANG, F. LIU, J. LI, N. CAO et C. SONG. 2017, «The Construction of the Integration of Water and Fertilizer Smart Water Saving Irrigation System Based on Big Data», dans *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, vol. 2, p. 392–397, doi : 10.1109/CSE-EUC.2017.258. 119
- ZHANG, Y., X. LUO, J. BUIS et J. SUTHERLAND. 2015, «LCA-oriented semantic representation for the product life cycle», *J. Clean. Prod.*, vol. 86, p. 146–162. 50
- ZHAO, J.-C. et J.-X. GUO. 2018, «Big data analysis technology application in agricultural intelligence decision system», dans *2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, p. 209–212, doi :10.1109/ICCCBDA.2018.8386513. 119
- ZHAO, Q. et T. HASTIE. 2017, «Causal interpretation of black-box models», , p. 16. 96
- ZHU, J. Y., X. PAN et R. S. ZALESNY. 2010, «Pretreatment of woody biomass for biofuel production : energy efficiency, technologies, and recalcitrance», *Applied Microbiology and Biotechnology*, vol. 87, n° 3, doi :10.1007/s00253-010-2654-8, p. 847–857, ISSN 0175-7598, 1432-0614. URL <http://link.springer.com/10.1007/s00253-010-2654-8>. 117
- ČUČEK, L., J. J. KLEMEŠ et Z. KRAVANJA. 2012, «A Review of Footprint analysis tools for monitoring impacts on sustainability», *Journal of Cleaner Production*, vol. 34, doi :10.1016/j.jclepro.2012.02.036, p. 9–20, ISSN 0959-6526. URL <https://www.sciencedirect.com/science/article/pii/S0959652612001126>. 51

Crédits Tous les icônes utilisés vient de flatico.com et les contributeurs sont : DinosoftLabs, Freepik, blackwoods, pongsakornRed, Becris, Eucalyp, popo2021, Smashicons, Nixxdsgn, Vitaly Gorbachev, orvipixel, aficons studio, raftelDesign, juicy_fish, payungkead, dDara

Résumé

Ces dernières années ont vu l'émergence de plans d'action pour le climat, le développement durable ou l'environnement comme le plan européen « *Green Deal* » de janvier 2020. L'un des principaux plans d'action, intitulé "Plan d'action pour l'économie circulaire", propose d'utiliser le modèle d'économie circulaire pour simuler l'utilisation de modèles durables. Dans ce contexte, les procédés et systèmes industriels se veulent plus durables et circulaires. Pour évaluer cette durabilité, de nombreuses approches existent et nécessitent un besoin de données, d'informations et de connaissances croissant. Par la transition digitale et l'émergence des données massives, de nombreuses données sont disponibles pour les industries. Qu'elles soient structurées ou non, sous différents formats (PDF, XML, HTML), et de différents types (Web, bases de données, articles scientifiques), les sources de données sont potentiellement des mines de connaissances qui peuvent compléter les analyses de la durabilité. Allant souvent de pair avec les données massives, l'intelligence artificielle apporte des méthodes pour extraire, traiter, analyser et visualiser ces données. De nombreux outils et méthodes sont désormais disponibles à cette fin grâce aux techniques du « *Big Data* » ou de l'intelligence artificielle comme l'apprentissage automatique. L'évaluation des systèmes industriels durables et circulaires peut bénéficier de ces technologies. Évaluer la durabilité d'un système ou d'un procédé regroupe maintenant un ensemble de méthodes en interaction les unes avec les autres. Cela s'explique par la quantité des données disponibles, la complexité grandissante des procédés industriels et des analyses de la durabilité plus globales.

Les travaux de cette thèse s'intéressent à définir une approche centrée données massives et intelligence artificielle pour l'évaluation des systèmes industriels durables et circulaires. Un cadre méthodologique est développé afin de profiter des articles scientifiques et du web public, d'analyser ces données et de fournir un support pour aider à l'interprétation des résultats. Cette approche induit le couplage de quatre domaines d'intérêt : données massives, intelligence artificielle, analyses de la durabilité et procédés et systèmes industriels. La spécification de ce cadre permet de reconsidérer la façon d'exécuter les analyses de la durabilité. Il donne lieu à son application dans l'agro-industrie. Différentes trajectoires de valorisation des déchets lignocellulosiques sont étudiées selon deux principales dimensions, les intrants de type biomasse et les technologies de procédés. Ce cas d'étude soutient l'ingénierie circulaire et met à l'épreuve le cadre méthodologique proposé.

Mots-clés : Procédés Industriels, Données massives, Développement Durable, Economie Circulaire, Pensée Cycle de Vie, Apprentissage automatique

Abstract

The last years have seen the emergence of action plans for climate, sustainable development or the environment such as the European "Green Deal" plan of January 2020. One of the main action plans, entitled "Circular Economy Action Plan", proposes to use the circular economy model to simulate the use of sustainable models. In this context, industrial processes and systems are intended to be more sustainable and circular. To assess this sustainability, many approaches exist and require an increasing need for data, information and knowledge. With the digital transition and the emergence of big data, a lot of data is available for industries. Whether structured or not, in different formats (PDF, XML, HTML), and of different types (Web, databases, scientific articles), data sources are potentially wealth of knowledge that can complement sustainability analyses. Often going hand-in-hand with massive data, artificial intelligence brings methods to extract, process, analyze and visualize these data. Many tools and methods are now available for this purpose thanks to "Big Data" techniques or artificial intelligence such as machine learning. Sustainability evaluation and circular industrial systems can benefit from these technologies.

Assessing the sustainability of a system or process now brings together a set of methods that interact. This is due to the amount of data available, the increasing complexity of industrial processes, and more comprehensive sustainability analyses.

The work of this thesis is interested in defining a big data and artificial approach driven approach for the evaluation of sustainable and circular industrial systems. A framework is developed in order to take advantage of scientific articles and the public web, to analyze these data and to provide support for the interpretation of the results. This approach induces the coupling of four domains of interest : massive data, artificial intelligence, sustainability analysis and industrial processes and systems. The specification of this framework allows to reconsider the way to perform sustainability analyses. It gives rise to its application in agribusiness. Different trajectories of lignocellulosic waste valorization are studied along two main dimensions, biomass inputs and process technologies. This case study supports circular engineering and tests the proposed methodological framework.

Keywords : Industrial Engineering, Big Data, Sustainability, Circular Economy, Life Cycle Thinking, Machine Learning