Creation of a Biodiversity Severity Index to evaluate the risks of accidental pollutions in the industry : a multi-criteria sorting approach

Tom Denat

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences et Lettres
PSL Research University

Préparée à l'Université Paris-Dauphine

Creation of a Biodiversity Severity Index to evaluate the risks of accidental pollutions in the industry:A multi-criteria sorting approach

École Doctorale de Dauphine — ED 543

Spécialité **Informatique**

COMPOSITION DU JURY :

Meltem ÖZTÜRK
Université Paris-Dauphine
Co-Directrice de thèse

Alexis TSOUKIAS
Université Paris-Dauphine
Président du jury

Vincent MOUSSEAU
CentraleSupélec
Rapporteur

Marc PIRLOT
Université de Mons
Rapporteur

Soutenue le 05.07.2017
par Tom DENAT

Dirigée par **Denis BOUYSSOU**

DAUPHINE UNIVERSITÉ PARIS | PSL RESEARCH UNIVERSITY PARIS

Remerciements

J'avais initialement voulu faire un remerciement que j'aurais nommé "remerciement de Russell" ou encore "remerciement du barbier" et qui se serait exprimée ainsi: "Toute personne lisant ces lignes est directement concernée par mes remerciements si et seulement si elle ne se croit pas concernée par ceux-ci". Cette version m'aurait épargné quelques minutes de travail ainsi que l'embarras probable d'avoir oublié quelqu'un. Néanmoins, cette formulation ambigüe pourrait semer le trouble chez certains lecteurs or il est des personnes à qui je tiens à affirmer sans équivoque ma reconnaissance. J'espère que ceux que j'aurai oubliés (car il y en aura) me connaissent assez bien pour ne pas le prendre personnellement mais plutôt comme une manifestation supplémentaire de mon étourderie.

Cette thèse a été menée au LAMSADE à l'université Paris Dauphine. Ce fût pour moi un lieu propice à la recherche mais j'ai également eu la chance d'y faire des rencontres exceptionnelles (et je n'exagère rien).

Mes pensées vont en premier lieu à Meltem. Merci Meltem d'avoir cru en moi. Merci pour le temps et l'énergie que tu as consacré à ma thèse. Merci pour les conseils et l'aide que tu m'as donnée, pour ta bonne humeur et ta générosité. Merci de m'avoir soutenu lorsqu'en fin de thèse j'ai lancé cette idée saugrenue du DBMC. Ce fût une chance et un honneur pour moi de travailler avec toi. Merci...

Je dois un remerciement particulier à Denis, pour avoir accepté de diriger cette thèse et pour son apport en rigueur dans mon travail.

Je tiens également à remercier les enseignants chercheurs au LAMSADE, en particulier Vangelis mon ami. Alexis, je te remercie pour ta gestion du LAMSADE comme pour ta personnalité joviale. Je remercie l'équipe administrative du LAMSADE sans qui rien n'est possible.

Je tiens à remercier tous les doctorants du LAMSADE pour leurs discussions enflammées, leurs joies, leurs soutiens dans les moments pénibles: Nath, Rafa, tous les Amines possibles, notre président bien aimé (gloire à lui), Dalal, Anaëlle, Olivier (qui une fois encore s'incruste parmi les doctorants), Renaud, Justin (un homme fort sympathique bien qu'anglais), Sat, Ian Christopher Ternier, Lydia, Linda, Maude, Marcel, Khalil, Fabien, Youcef, Meriem, Amal, Marek, Ioannis, Pedro, Sami, Saeed, Yassine, Diana et tant d'autres.

Cette thèse étant en coopération avec l'INERIS, j'y ai également fréquenté des personnes agréables et stimulantes. Bien entendu en premier lieu vient Chabane. Je crois

que je ne connais toujours pas le visage de Chabane sans un sourire. Mis à part sa bonne humeur constante, je dois aussi beaucoup à Chabane pour son support notamment dans l'étude de la gestion de risques. Je tiens également à remercier Christophe qui a participé à ce travail. Pierre Roux est notre expert en toxicologie très souvent mentionné dans ce document. Merci Pierre.

J'ai également eu la joie de collaborer avec les chercheurs du Muséum National d'Histoire Naturelle. Je les en remercie, en particulier Alexandre, ils ont contribué de manière considérable à ce travail.

Mes pensées vont également à Michaela Saisana du Joint research center à Ispra pour son invitation et son accueil chaleureux.

Je remercie également mes rapporteurs, Vincent et Marc. Je sais par avance que leurs remarques seront une fois encore constructives et profitables.

Viennent ensuite mes proches sans qui j'aurais peut-être jeté l'éponge. Ma famille; mes parents et mon frère bien sûr, mais également mes cousins et mes grand mères. Ils sont un pilier dans ma vie, je leur dois tellement. Ma reconnaissance va aussi à mes amis qui sont un repère, un appui (je n'en cite aucun car si je les cite tous ça fait un chapitre de thèse et si j'en cite certains et d'autre pas ça créera un malentendu, ils sauront se reconnaitre). Je remercie Cristina qui m'a soutenu durant une partie de cette thèse et à qui je dois beaucoup.

Finalement, Misha. Merci pour tout. Merci pour ta patience, pour ta tendresse, pour avoir rendu cette épreuve plus supportable. Ça y est, c'est fini cette "p***** de thèse" je vais pouvoir vous consacrer plus de temps, à toi et à Noah. Noah, tu ne pourras pas lire ces mots tout de suite. Ton vocabulaire est pour l'instant constitué de 7 ou 8 mots relatifs à des besoins primitifs. Néanmoins quand tu le pourras, j'aimerais te dire que ce sont tes rires, tes caprices et tes câlins qui m'ont fait tenir. Oups, j'en oubliais un. Un ou une d'ailleurs on ne sait pas encore...

Résumé en français

The following pages are an abstract written in French. However, the entire text, written in English, follows.

Création d'un Indice de Gravité sur la Biodiversité pour évaluer les risques de pollutions accidentelles

Introduction

La révolution industrielle fût un tournant majeur de l'histoire de l'humanité. Le niveau de vie du plus grand nombre commença à s'améliorer considérablement pour la deuxième fois depuis l'apparition de l'homme (la première fois étant due au développement de l'agriculture il y a entre 15 000 et 20 000 ans) et aujourd'hui presque tous les aspects de nos vies sont impactés d'une manière ou d'une autre par des activités industriels. Néanmoins, les procédés industriels impliquent parfois des transformations chimiques complexes et des matériaux potentiellement dangereux qui peuvent représenter une menace pour les travailleurs, les usagers, les habitants des environs ou l'environnement. Alors que les activités industrielles furent tout d'abord principalement vues sous un angle positif, à partir de la deuxième moitié du vingtième siècle la société fût progressivement sensibilisée au risques et aux impactes qu'elles entrainent. De nos jours, limiter les risques liés aux activités industrielles est devenu un enjeux important pour les gérants de ces activités comme pour les acteurs publics, associatifs ou tout citoyen intéressé par les questions relatives à la santé, la sécurité des personnes ou la protection de l'environnement.

Pour ce faire, la principale approche consiste à identifier chaque scénario d'accident possible, à évaluer pour chacun leur probabilité d'occurrence et la gravité potentielle de leurs conséquences afin de juger de l'acceptabilité des risques induits par chaque scénario par le biais d'une matrice de risques (comme celle donnée en exemple sur la figure 1). En France, cette procédure existe concernant les risques d'atteinte à la vie humaine et elle est encadrée par plusieurs textes, en particulier l'arrêté du 29 septembre 2005 et la circulaire du 10 mai 2010.

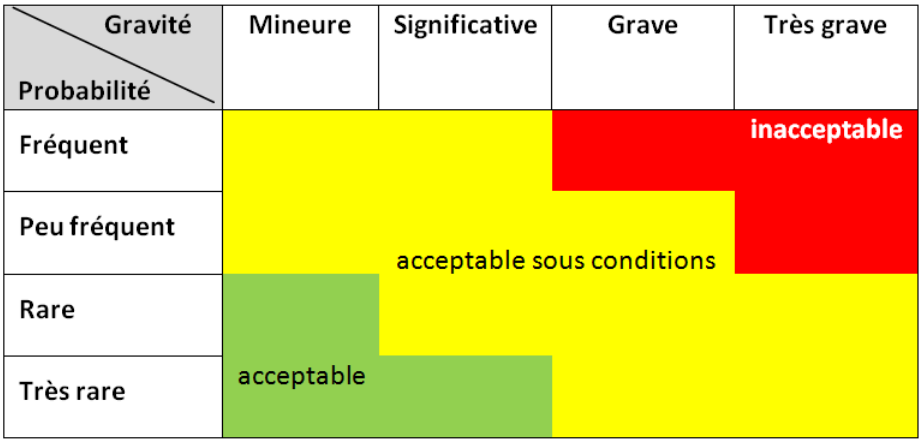

Figure 1: Exemple d'une matrice de risques

Mais cette procédure ne prend pas en compte les risques d'accidents industriels ayant un impact sur l'environnement en général et sur la biodiversité en particulier. L'un des objectifs de cette thèse est d'aider à la création d'un méthodologie pour prendre en compte ces risques dans les études de risques. L'arrêté du 29 septembre 2005 et la circulaire du 10 mai 2010 fournissent des outils permettant d'évaluer la probabilité d'occurrence d'un scénario d'accident industriel et étant donné la probabilité et la gravité d'un scénario d'accident d'en évaluer l'acceptabilité du risque. Ces méthodologies peuvent être utilisées telles quelles dans le cadre d'une évaluation concernant les risques d'atteinte à l'environnement. Par contre la méthodologie utilisée pour évaluer la gravité d'un scénario doit être modifiée afin de représenter la dimension environnementale. Pour cela nous créons dans ce document un indicateur que nous nommons Indice de Gravité sur la Biodiversité sensé répondre à la question "si le scénario S se produisait, quelle serait la gravité de son impact sur la biodiversité environnante?".

Afin de proposer une méthodologie fiable pour donner une valeur à l'Indice de Gravité sur la Biodiversité, le présent travail s'appuie sur une discipline nommé l'Aide Multi-Critère à la Décision. Cette discipline est une branche de l'informatique décisionnelle qui propose un ensemble de méthodes et de calculs permettant de choisir la meilleure solution, un ensemble de bonnes solutions parmi tout un ensemble de solutions réalisables ou encore d'évaluer chaque solution réalisable indépendamment. Une méthode d'aide à la décision classique et très largement utilisée est la somme pondérée mais un grand nombre de méthodes (Méthode de Choquet, ELECTRE TRI, UTA...) sont également disponibles, permettant de prendre en compte les préférences d'un ou plusieurs décideurs et les interactions possibles entre plusieurs critères. Dans le cadre de la création de

l'Indice de Gravité sur la Biodiversité, cet indicateur sera obtenu par l'agrégation de plusieurs données relatives au scénario d'accident étudié. L'aide multi-critère à la décision fourni de nombreuses méthodes pour agréger les valeurs d'un objet (ici un scénario de pollution accidentelle) sur plusieurs critères afin d'obtenir un critère unique de synthèse. Chacune de ces méthodes d'agrégation doit être paramétrée afin de s'adapter au préférences du décideur ou au contexte et afin de trouver ces paramètres, plusieurs méthodes d'élicitation existent. Nous avons étudié un grand nombre de ces méthodes afin de trouver celles qui étaient adaptés à notre problématique et ce faisant nous en avons proposé une nouvelle méthode d'élicitation des préférences. La méthode proposé se nome Dominance Based Monte Carlo et constitue le deuxième axe de recherche de cette thèse.

Création de l'Indice de Gravité sur la Biodiversité

L'Indice de Gravité sur la Biodiversité pour un scénario s a pour but de répondre à la question: "Si le scenario s se produisait quelle serait la gravité de ses conséquences sur la biodiversité environnante?". Afin de répondre à cette question le problème fût structuré à l'aide d'une hiérarchie de critères. Afin de pouvoir rendre l'utilisation de cet indicateur plus facilement compatible avec la législation déjà existante sur les risques accidentels nous décidons d'utiliser une échelle discrête de 5 échelons pour représenter l'Indice de Gravité sur la Biodiversité.

Les hiérarchies de critères

Par hierarchie de critères nous entendons ici une arborescence dans laquelle chaque nœud est un critère. Pour chaque critère i, les critères situés directement en dessous (les fils du nœud représentant i dans l'arborescence) seront appelés les sous-critères de i. Les valeurs des critères représentés par des feuilles dans l'arborescence devront être fournis par l'utilisateur de la méthode. Nous les appelerons les critères "entrées". La valeur de tout autre critère i sera obtenue par l'agrégation des sous critères de i. Chacune de ces agrégations peut être considérée comme un problème de TRI multi-critère, nous les appellerons des sous-problème.

La hiérarchie de critères utilisée pour l'Indice de Gravité sur la Biodiversité

La hierarchie de critères correspondant à l'Indice de Gravité sur la Biodiversité a été obtenue par l'interaction avec des experts, en particulier:

- Un expert en toxicologie de l'INERIS.
- Plusieurs experts en écologie du Muséum National d'Histoire Naturelle (MNHN).

Nous avons interagi avec ces experts en nous inspirant de la méthodologie fournie par [Keeney, 1992]. Décrivons de haut en bas l'arborescence illustrée en Figure . L'Indice de Gravité sur la Biodiversité se décompose en deux indices: l'Indice de Gravité sur la Biodiversité sur les sols et l'Indice de Gravité sur la Biodiversité sur les eaux de surfaces.

Ensuite, l'espace pouvant être impacté par le scénario étudié est divisé en plusieurs cibles, chacune de ces cibles devant être relativement homogène du point de vue de la biodiversité et étant potentiellement similairement impacté si le scénario se produisait. Dans ce travail, nous nous sommes principalement intéressés à la gravité sur les eaux de surface. L'Indice de Gravité sur la Biodiversité sur les sols et l'Indice de Gravité sur la Biodiversité sur les eaux de surfaces sont chacun divisés en autant d'indices qu'il y a de cibles afin d'évaluer la gravité attendue des conséquences du scénario étudié à un niveau local: les Indices Locaux de Gravité sur la Biodiversité.

Chaque Indice Local de Gravité sur la Biodiversité est divisé en trois critères: le *potentiel destructeur*, la *valeur de l'environnement* sur la cible et la *vulnérabilité* de la cible. Nous décidons d'utiliser une échelle discrête de 5 échelons pour représenter l'Indice de Gravité sur la Biodiversité, l'Indice de Gravité sur la Biodiversité sur les eaux de surfaces, l'Indice de Gravité sur la Biodiversité sur les sols, les Indices locaux de Gravité sur la Biodiversité, la vulnérabilité des cibles et les valeurs environnementales des cibles. Le potentiel destructeur a pour but de représenter la capacité d'un scénario de pollution accidentel à impacter un écosystème. Le potentiel destructeur est obtenu en agrégeant trois critères: la concentration attendue du produit dans l'environnement après l'occurrence de la pollution (exprimée en gramme par litre), le temps de résidence du produit (exprimée sous forme binaire, "court", "long") et la toxicité du liquide (exprimée par sa concentration admissible c'est à dire en gramme par litre). Cet arborescence ainsi que les échelles qui ont été choisies l'ont été grâce à des interactions avec les experts de l'INERIS et du Muséum National d'Histoire Naturelle. De même pour chaque critère "entrée" des points de références ont été fournis afin d'aider les utilisateurs de la méthode à fournir ces données entrée. Le tableau 1 donne les significations des différents échelons des échelles utilisées pour représenter les critères. Il est à noter que le potentiel destructeur est exprimé en "puissance attendue de l'impact sur une cible moyennement vulnérable".

	Échelon C1	**Échelon C2**	**Échelon C3**	**Échelon C4**	**Échelon C5**
Potentiel destructeur	Aucun impact sur la biodiversité	Faible impact sur la biodiversité	Impact relativement fort sur la biodiversité	Fort impact sur la biodiversité	Destruction totale de la biodiversité
Valeur de l'environnement	Très faible valeur	Plutôt faible valeur	Valeur moyenne	Valeur plutôt forte	Très forte valeur
Vulnérabilité	Très faible vulnérabilité	Plutôt faible vulnérabilité	Vulnérabilité moyenne	Vulnérabilité plutôt forte	Très forte vulnérabilité
Indice Local de Gravité sur la Biodiversité	Aucun impact ou pollution négligeable	Faible pollution	Pollution moyenne	Pollution importante	Désastre écologique

Table 1: Table définissant les différentes échelles utilisées pour représenter les critères: *Potentiel destructeur*, *Valeur de l'environnement*, *Vulnérabilité* et l'*Indice Local de Gravité sur la Biodiversité*

Travaux antérieurs et méthodologie de construction de l'arborescence de critères

Avant le début de mes travaux sur la construction de l'Indice de Gravité sur la Biodiversité un travail avait déjà été fait à l'INERIS pour apporter une première proposition. Cette proposition consistait à créer pour chaque binôme scénario cible, trois scores appelés "module source", "module transfert" et "module cible". Le but du "module source" est de représenter le potentiel destructeur du scénario. Le "module transfert" est sensé représenter les obstacles permettant d'empêcher le scénario étudié d'impacter la cible ou d'en diminuer l'impact. Le "module cible" représente l'importance écologique et économique de la cible étudiée. Le score du "module source" est diminué en fonction de la valeur du "module transfert". Ensuite le score obtenu est multiplié par la valeur du "module cible". Enfin le produit obtenu permet d'obtenir un indice sur une échelle de cinq échelons grâce à des valeurs seuils délimitant les échelons.

Ce document est à ma connaissance le premier à aborder de sujet de la création d'un indicateur pour évaluer la gravité attendue d'un scénario de pollution accidentelle. Néanmoins aussi bien les trois scores des modules que l'agrégation finale semblent avoir été obtenus intuitivement (ce sont principalement des puissances de 10). Il n'est pas non plus fait mention d'interaction avec des experts. Nous avons donc choisi de

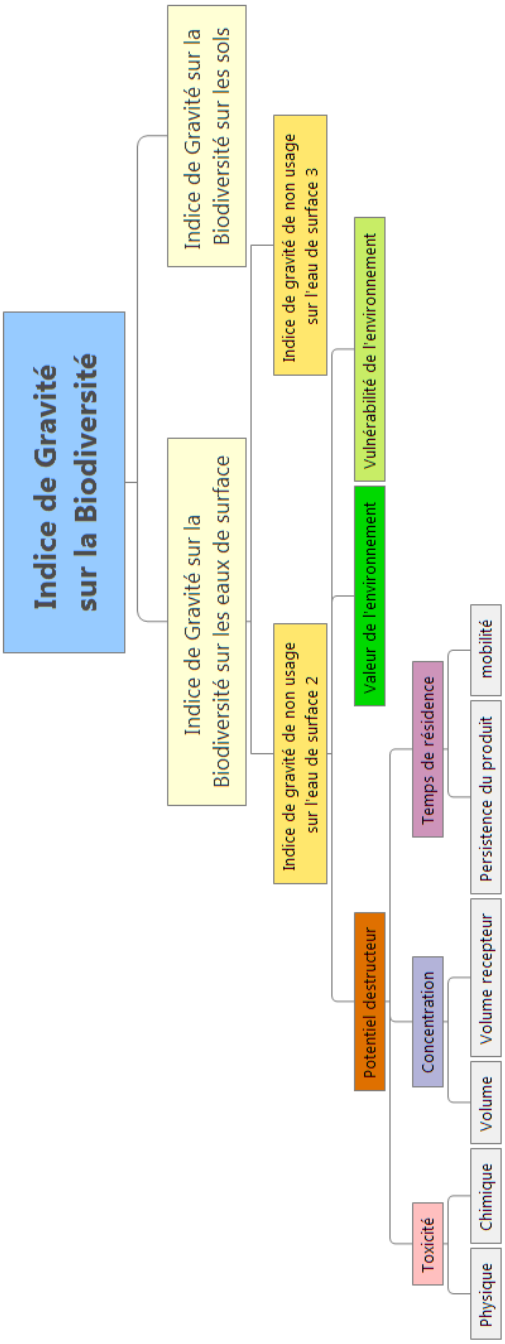

Figure 2: Illustration de la hiérarchie de critères de l'Indice de Gravité sur la Biodiversité.

reprendre ce travail en y ajoutant des interactions avec plusieurs experts en toxicologie et en écologie ainsi que l'utilisation de méthodes formelles d'aide la décision.

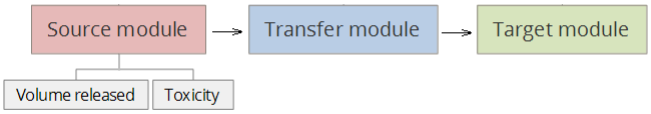

Figure 3: Représentation graphique de la méthode décrite dans le document nommé *Méthode d'évaluation de la gravitéé des conséquences environnementales d'un accident industriel*

Afin d'obtenir la hiérarchie de critères présentée plus haut, nous nous sommes inspirés du livre Value Focused Thinking [Keeney, 1992]. Dans ce document, une méthodologie d'interaction avec le décideur est proposée afin de construire une arborescence de critères. Nous avons partiellement suivi cette méthodologie lors de nos interactions avec différents experts à l'INERIS et au Muséum National d'Histoire Naturelle.

Méthodes d'agrégation pour les différents nœuds de l'arborescence

L'agrégation pour le potentiel destructeur

Comme nous l'avons expliqué précédemment à chaque nœud de l'arborescence excepté les feuilles, une agrégation doit être faite entre les critères qui peut s'apparenter à un problème de TRI multi-critère.

L'agrégation de la *concentration* de produit polluant dans la cible étudié, du *temps de résidence du produit* et de la *toxicité du produit polluant* pour obtenir le *potentiel destructeur* a été obtenue par des interactions avec un expert en toxicologie à l'INERIS. Lors de cette interaction, nous avons compris que son raisonnement était principalement fondé sur le rapport entre la *concentration de produit polluant dans la cible étudié* et la concentration admissible du produit polluant. Nous avons choisi la concentration admissible des produits comme mesure de la *toxicité du produit polluant*. La méthode choisie fût donc basée sur le rapport entre la *concentration de produit polluant dans la cible étudié* et la concentration admissible (combien de fois la concentration admissible serait elle présente après le scénario?), les échelons du *potentiel destructeur* étant délimités par des valeurs seuils. Le tableau suivant présente les valeurs seuils obtenues.

$\frac{Cons}{MATC}$	Temps de résidence court	Temps de résidence long
≤ 1	$C1$	$C1$
$]1, 10]$	$C1$	$C2$
$]10, 50]$	$C2$	$C3$
$]50, 100]$	$C3$	$C4$
$]100, 1000]$	$C4$	$C5$
> 1000	$C5$	$C5$

Table 2: Tableau représentant le potentiel destructeur d'une fuite en fonction du rapport concentration par concentration admissible.

Un modèle MR-SORT pour l'agrégation pour l'Indice local de gravité sur la Biodiversité

L'agrégation du *potentiel destructeur*, de la *valeur de l'environnement* sur la cible et de la *vulnérabilité* de la cible pour obtenir l'*Indice Local de Gravité sur la Biodiversité* a été trouvé par l'interaction avec plusieurs experts du Muséum National d'Histoire Naturelle. Le modèle qui fût choisi est un modèle MR-SORT avec véto.

Le modèle MR-SORT est une méthode de TRI multi-critère qui fonctionne comme suit. Chaque critère se voit attribuer un poids. Une méthode de surclassement permet de comparer toute paire d'objets évalués sur les critères. Afin de comparer un objet a à un object b on regarde quels sont les critères pour lesquels a est meilleur que b. Si la somme des poids des critères pour lesquels a est meilleur que b est plus haute qu'un seuil s prédéfini alors il est dit que a surclasse b noté aSb.

Ensuite pour classer un objet a, il est comparé tour à tour à des objets multi-critères fictifs appelé des profils. Ces profils auront le rôle de seuils pour délimiter les différentes catégories. On compare donc a au profil le plus bas b_2. S'il ne le surclasse pas alors a est affecté à la catégorie la plus basse $C1$, sinon on le compare au profil b_3 et ainsi de suite. Si a surclasse le profil le plus haut alors il sera affecté à la catégorie la plus haute. Dans le cas où un phénomène de véto est intégré, pour accéder à la catégorie 2 un objet doit surclasser le profil b_2 et dominer (être au moins aussi bon sur tous les critères) un profil de véto v_2 et ainsi de suite pour toutes les catégories.

Afin de trouver les paramètres du modèle MR-SORT les plus adaptés pour la création de l'Indice Local de Gravité sur la Biodiversité, nous avons interrogé un expert du Muséum National d'Histoire Naturelle en lui demandant d'attribuer à plusieurs scénarios de pollutions accidentelles des valeurs pour l'Indice Local de Gravité sur la Biodiver-

sité, ces valeurs allant de $C1$ (pollution négligeable) à $C5$ (désastre environnemental). Nous avons ensuite utilisé un algorithme d'élicitation des préférences basé sur de la programmation mathématique pour trouver des paramètres MR-SORT adaptés aux réponses obtenues. Les poids des critères sont tous égaux à 1/3 et le seuil est de 1/2, ce qui signifie qu'il faut dépasser un profile sur deux critères au moins pour le surclasser. Le tableau 3 représente les profils obtenus tandis que le tableau 4 représente les profils de véto.

	Dest Pot	Val Env	Vuln
Profile b_5	$C5$	$C5$	$C4$
Profile b_4	$C3$	$C5$	$C3$
Profile b_3	$C3$	$C3$	$C3$
Profile b_2	$C3$	$C3$	$C3$

Table 3: Profils delimitant les catgories.

	Dest Pot	Val Env	Vuln
Veto profile v_5	$C2$	$C4$	$C2$
Veto profile v_4	$C2$	$C4$	$C2$
Veto profile v_3	$C1$	$C3$	$C1$

Table 4: Profils de véto délimitant les catégories

L'agrégation des différents Indices Locaux de Gravité sur la Biodiversité en un Indice de Gravité sur la Biodiversité

Afin d'agréger les différents les différents Indices Locaux de Gravité sur la Biodiversité en un Indice de Gravité sur la Biodiversité nous avons proposé d'utiliser une version modifiée du max qui prendrait en compte l'étendue d'un impact environnemental. L'idée de ce max modifié est de donner à l'Indice de Gravité sur la Biodiversité la valeur la plus haute parmi les Indices Locaux de Gravité sur la Biodiversité qu'au dessus d'une certaine superficie l'Indice de Gravité sur la Biodiversité peut être augmenté d'un échelon. Néanmoins cette partie du travail n'a pas pu faire l'objet d'une validation par les experts ce qui constitue encore aujourd'hui un travail en perspective.

Illustration d'une évaluation d'un scénario de fuite de liquide toxique

Nous illustrons maintenant comment cet indicateur pourrait fonctionner dans le cadre d'une étude de risque. Le cas que nous décrivons ici est totalement fictif et n'est pas

fondé sur l'étude d'une installation industrielle. Assumons qu'une usine de fabrication de conservateurs industriels à Créteil dans le Val de Marne (94000) soit sujette à une étude de risques qui prenne en compte les risques de pollutions accidentelles. Une liste de scénarios serait alors définie. Parmi ces scénarios le scénario a représente une fuite de biphenyl. Si la fuite a se produisait elle rejoindrait le lac de Créteil. Etudions donc l'Indice Local de Gravité sur la Biodiversité de la fuite a sur la cible "le lac de Créteil".

Étudions d'abord le potentiel destructeur de cette fuite sur le lac de Créteil. Admettons que les experts qui participent à l'étude estiment la concentration en biphenyl dans le lac à $33 \pm 5 \mu g l^{-1}$ si la fuite se produisait. L'eau du lac étant static et le biphenyl étant un produit persistant, le temps de résidence du produit serait alors défini comme long. La concentration acceptable du biphenyl étant de $4 \mu g l^{-1}$, selon la règle de création du potentiel destructeur, pour le scénario a et la cible "le lac de Créteil", le potentiel destructeur serait égal à $C3$.

Bien que le lac de Créteil ne soit pas une zone classée natura 2000, selon les cartes @d le niveau de biodiversité ordinaire est considéré comme haut et le niveau de biodiversité remarquable est considéré comme non nul ce qui est relativement rare dans une zone urbaine proche de Paris. Admettons que les experts en biodiversité participant à cette étude de risque ont décidé de donner à la valeur de l'environnement sur cette cible la valeur $C3$. De plus étant donné que le lac de Créteil n'est connecté biologiquement à aucune autre cible sa vulnérabilité pourrait être estimée par les experts participant à l'étude comme ayant pour valeur $C4$. Alors utilisant ces valeurs et le modèle MR-SORT créé pour l'agrégation de l'Indice Local de Gravité sur la Biodiversité, cet indice est donc fixé pour le scénario a et la cible "le lac de Créteil" à la valeur $C3$ c'est-à-dire pollution moyenne.

Le principal apport de mon travail sur la création de l'Indice de Gravité sur la Biodiversité est d'avoir été créé en coopération avec des experts et d'être basée sur une méthodologie formelle d'aide multi-critre à la décision. Ainsi la hiérarchie de critères obtenue, les méthodes d'agrégations choisies à chaque nœud de l'arborescence ainsi que les paramètres retenus pour chaque méthode d'agrégation ont été obtenus par le biais d'interactions avec des experts en toxicologie et en écologie.

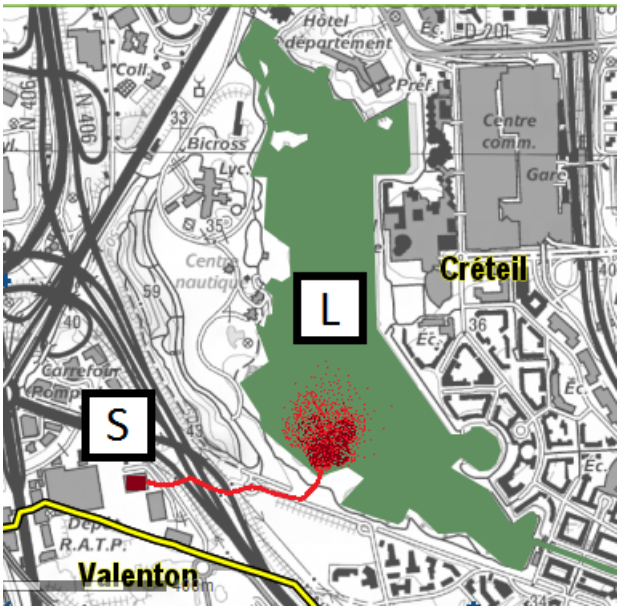

Figure 4: Illustration du scénario d'une fuite de biphenyl dans le lac de Créteil. **S** représente la source de la fuite et **L** représente le lac de Créteil.

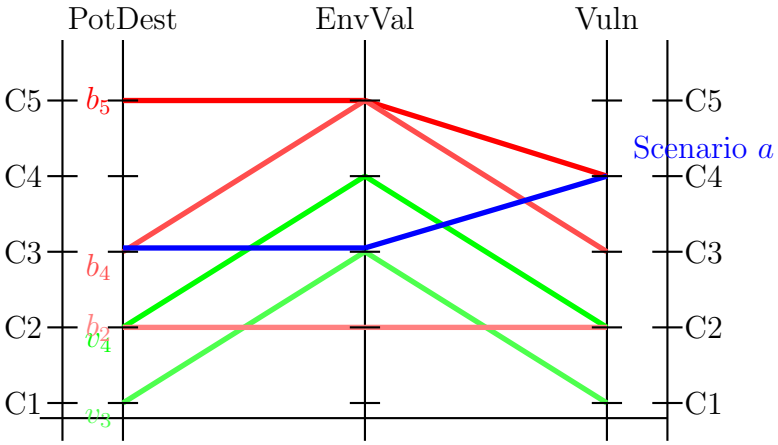

Figure 5: Représentation graphique de l'évaluation de l'Indice Local de Gravité Environnementale par le modèle MR-SORT. Les profils sont représentés par des lignes rouges tandis que les profils de véto sont représentés par des lignes vertes. Les scénarii décrit dans ce paragraphe est représenté par une ligne bleue. Nous pouvons donc constater ici que le scénario étudié doit être classé en catégorie $C3$.

L'algorithme Dominance Based Monte Carlo

Durant cette thèse, j'ai étudié un grand nombre d'algorithmes d'élicitation des préférences en particulier pour le problème du tri multi-critère. Ce faisant j'en ai proposé un nouveau nommé Dominance Based Monte Carlo (DBMC).

Le problème du tri multi-critère consiste à affecter des objets à des catégories parmi un ensemble pré-défini de catégories en se basant sur les évaluations de ces objets sur un ensemble des critères prédéfinis. Pour ce faire, notre méthode se base sur un échantillon d'objets triés par le décideur (le learning set). Notre méthode réuni deux propriétés plutôt rares en aide multi-critère à la décision : l'absence de modèle et le fonctionnement aléatoire. Ici l'"absence de modèle" signifie qu'il n'est pas supposé que le raisonnement du décideur soit fondé sur un ensemble de règles connues de l'analyste. Comme tout algorithme de monte carlo son fonctionnement est non déterministe. La monotonie (améliorer un objet sur un critère ne peut pas le rendre globalement moins bon) et le fait de retourner systématiquement chaque objet du learning set dans la catégorie à laquelle le décideur l'a affecté, considérés dans les autres méthodes comme de bonnes propriétés, constituent les uniques contraintes qui encadrent celles ci.

Notre méthode prend en paramètres : un ensemble de critères N exprimés sur des échelles finies et discrètes v_i, un ensemble d'objets A (ici chaque combinaison de valeurs sur les critères forme un objet) et un ensemble de catégories C.

Il fonctionne comme suit et décrit par l'algorithme 1. Le décideur fournit un learning set L (il affecte certains objets à des catégories). Ensuite un objet est choisi au hasard puis on l'affecte à une catégorie au hasard parmi les catégories auxquelles cet objet peut appartenir sans violer la monotonie. Cette affectation génère de nouvelles contraintes sur les classifications possibles des autres objets. Les contraintes dues au respect du learning set et de la monotonie sont illustrés en Figure 6. Un autre objet est ensuite choisi au hasard et affecté à une catégorie et ce jusqu'à ce que chaque objet soit affecté à une catégorie. Cette classification que nous appelons un lancer est hautement aléatoire. Afin de corriger cette propriété, on effectue T lancers. A la fin du processus chaque objet est affecté dans une classification globale, la catégorie médiane des catégories auxquelles il a été affecté au cour des T lancers.

	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	[1,2]	[1,2]	[1,3]	[1,3]	[1,3]	[1,3]
2	1	1	1	1	[1,2]	[1,2]	[1,3]	[1,3]	[1,3]	[1,3]
3	1	1	1	1	[1,2]	2	[2,3]	[2,3]	[2,3]	[2,3]
4	[1,2]	[1,2]	[1,2]	[1,2]	[1,3]	[2,3]	[2,3]	[2,3]	[2,3]	[2,3]
5	[1,2]	[1,2]	[1,2]	[1,2]	[1,3]	[2,3]	[2,3]	[2,3]	[2,3]	[2,3]
6	[1,2]	[1,2]	[1,2]	2	[2,3]	[2,3]	[2,3]	[2,3]	[2,3]	[2,3]
7	[1,3]	[1,3]	[1,3]	[2,3]	[2,3]	[2,3]	[2,3]	[2,3]	3	3
8	[1,3]	[1,3]	[1,3]	[2,3]	[2,3]	3	3	3	3	3
9	[1,3]	[1,3]	[1,3]	[2,3]	[2,3]	3	3	3	3	3
10	[1,3]	[1,3]	[1,3]	[2,3]	[2,3]	3	3	3	3	3

Figure 6: Illustration des contraintes de classifications liées au respect de la monotonie et du learning set. Sur ce graphique on a un problème de tri avec deux critères, dix échelons sur chaque critère, trois catégories et cinq objets dans le learning set (ceux encadrés en bleu).

Algorithm 1: Algorithme DBMC

Data: Problème de tri $S =< N, V, A, C, L >$

Result: Classification $f_t : A \to C$ monotone stable et compatible avec le learning set L

1 **for** *i de 1 à T* **do**

2 $S' \leftarrow S$

3 Compléter aléatoirement S' par l'algorithme 2.

4 Affecter chaque objet à la médiane des valeurs qu'il a pris au cours des T lancers.

Distribution de probabilités des lancers

Comme nous avons pu le voir le résultat d'un lancer est aléatoire. Si nous pouvions connaitre la distribution de probabilité de l'affectation de chaque objet, alors l'affectation médiane pourrait être calculée sans qu'il ne soit nécessaire d'appliquer l'algorithme. Malheureusement, cette distribution semble très difficile à calculer. Nous avons néanmoins prouvé que chaque classification monotone compatible avec le learning set peut être obtenu avec une probabilité strictement positive bien que cette distribu-

tion ne soit pas uniforme parmi toutes les classifications monotones compatibles avec le learning set.

Algorithm 2: Complétion aléatoire - lancer

Data: Problème de tri $S =< N, V, A, C, L >$
Result: Classification $f_t : A \to C$ monotone et compatible avec le learning set L

while *Un objet qui n'est pas affecté à une catégorie* **do**
  Choisir aléatoirement un objet χ uniformément parmi A;
  Choisir aléatoirement une catégorie Δ uniformément $\gamma_{min}(\chi)$ and $\gamma_{max}(\chi)$;
  Ajouter l'information $< \chi, \Delta >$ au learning set;
  $\gamma_{max}(\chi) \leftarrow \Delta$;
  $\gamma_{min}(\chi) \leftarrow \Delta$;
  for *Tout* $a^- \in D^-(\chi)$ **do**
    $\gamma_{max}(a^-) \leftarrow min\{\Delta, \gamma_{max}(a^-)\}$
  for *Tout* $a^+ \in D^+(\chi)$ **do**
    $\gamma_{min}(a^+) \leftarrow max\{\Delta, \gamma_{min}(a^+)\}$
for *Tout* $a \in A$ **do**
  $f_t(a) \leftarrow \gamma_{max}(a)$ (or $\gamma_{min}(a)$ les deux sont égaux à ce moment)

Propriétés théoriques

Nous avons étudié les propriétés théoriques de notre algorithme. Chaque lancer respecte la monotonie et le learning set. De ce fait, le résultat de l'algorithme respecte également ses deux propriétés. Le résultat de l'algorithme est théoriquement sujet à l'aléatoire mais nous avons prouvé qu'il converge presque surement lorsque le nombre de lancers tend vers $+\infty$. De plus nos tests appliqués ont montré qu'à partir de 100 lancers les résultats sont relativement stables. La complexité algorithmique de l'algorithme DBMC est de l'ordre de $O(m^2T)$ et il tourne en une heure avec 100.000 objets et 100 lancers.

Validations pratiques

Afin de tester les performances de prédiction de l'algorithme DBMC nous avons appliqué un teste nommé 2-fold validation. Il consiste à utiliser un learning set réel que l'on divise aléatoirement en deux jeux de données de taille identique. Ensuite l'algorithme est appliqué pour apprendre sur une moitié des données, tenter de prédire l'autre moitié et comparer cette prédiction avec l'affectation réelle. Les trois learning sets utilisés, car evaluation (CEV), lecture evaluation (LEV) et breast cancert

(BCC) ont également été testé par [Sobrie et al., 2013] sur deux autres algorithmes d'élicitation pour le tri :Dominance Based Rough Set Approach (DRSA), UTADIS, non-compensatory sorting et MR-Sort.

	DRSA	NCS	MR-Sort	UTADIS	DBMC
CEV	$4.91 \pm 0.41\%$	$12.6 \pm 2.63\%$	$13.9 \pm 1.19\%$	$6.9 \pm 0.71\%$	$3.72 \pm 0.28\%$
LEV	$18.76 \pm 0.35\%$	$14.92 \pm 1.88\%$	$15.92 \pm 1.22\%$	$15.01 \pm 1.31\%$	$18.67 \pm 1.12\%$
BCC	$25.95 \pm 1.33\%$	$26.72 \pm 3.45\%$	$27.5 \pm 3.79\%$	$28.70 \pm 1.11\%$	$25.92 \pm 0.63\%$

Table 5: Résultat du teste de 2-fold validation. On peut voir le pourcentage d'erreur dans la prédiction acompagné de son écart type.

Nous proposons une méthode d'élicitation qui présente de bonnes propriétés théoriques et des résultats pratiques comparables aux autres algorithmes d'élicitation pour le Tri. Les résultats obtenus avec la méthode Dominance Based Rough Set Approach (DRSA) sont particulièrement proches de ceux obtenus par la méthode du Dominance Based Monte Carlo (DBMC). L'absence de modèle à proprement parler et le fait d'être fondé sur la monotonie sont deux points communs de ces deux méthodes, ce qui pourrait expliquer ces similarités.

L'absence de modèle peut être vue comme une bonne ou une mauvaise propriété selon le contexte. Dans le cadre d'une décision publique un modèle peut aider à justifier la décision. Dans le cas de l'étude d'un jeu de données représentant des préférences humaines il se peut qu'aucun modèle ne puisse a priori être choisi pour les représenter. Notre algorithme doit être appliqué lorsque les échelles utilisées sont finies et discrètes. Dans le cas inverse une discrétisation peut être envisagées sous certaines précautions.

Conclusion

Cette thèse est basée sur deux axes principaux. L'un appliqué traite de la création d'un indicateur multi-critère. J'ai pu appliquer des méthodes mathématiques d'aide multi-critère à la décision à un contexte réel. J'ai animé une interaction avec des experts de différents domaines afin de valoriser au mieux leur expertise dans la création de l'Indice de Gravité sur la Biodiversité.

L'autre axe de ma thèse est plus théorique bien que l'algorithme Dominance Based Monte carlo soit opérationnel. J'ai créé un outil d'éliciation des préférences, étudié ses propriété théoriques et testé ses performances pratiques pour les comparer à des algorithmes dont l'objectif est similaire. Ses performances sont proches de celles d'autre

algorithmes particulièrement de celles de DRSA. Ses propriétés, en particulier le fait d'être considéré comme une boite noir, le rendent peu attractif dans des contextes décisionnels nécessitant des justifications mais peuvent lui être bénéfiques lorsqu'aucun modèle connu ne semble approprié pour représenter un jugement.

Introduction

The Industrial Revolution marked a major turning point in history; the standard of living for the general population began to increase consistently for the second time since men existed [Fitzgerald, 2015] (the first occurred 15 000–20 000 years ago during the Neolithic Revolution) and today almost every aspect of our daily life is influenced in some way by industrial activities. Nevertheless, these processes sometimes involve complex chemical transformations and unstable or toxic products that may represent a threat either to workers, neighbours, users or to the environment. While industrial activities were at first only seen from a positive perspective, in the second part of the twentieth century, society gradually became aware of these issues [Ellul, 1967]. Thus today, limiting the risks associated to these processes is a major issue for both industrial managers, public institutions and many associations or citizens that feel concerned about matters such as health, safety and protection of the environment.

To do so, the most common approach consists in identifying all the scenarios of accident that could happen on the studied industrial facility, evaluating their probabilities to happen, the severity of their expected consequences and, based on these two elements, assess the acceptability of the risk induced by the studied industrial facility. The INERIS, as a public actor in industrial risks management, already provides such a process concerning human consequences of scenarios of accident in a working context that is well defined and framed by the French and European legislations. However, this type of evaluation process does not exist concerning other types of consequences such as possible socio-economical impacts of scenarios, impacts on the biodiversity etc. In particular, risk of accidental pollutions becomes an important concern as the population understand the urgent need for action on environmental issues and recent events such as the massive red mud pollution in Brazil[1] or the sulphuric acid leak in Australia both happening in 2015 [2] (among others) corroborate this emergency.

[1]Read more about the red mud pollution in Brazil in 2015 at http://www.ibtimes.co.uk/brazil-dam-disaster-toxic-red-mud-threatens-endangered-wildlife-rio-doce-basin-1530236

[2]Read more about the sulphuric acid leak in Australia in 2015 at https://www.theguardian.com/australia-news/2015/dec/27/

The aim of this thesis is part of an effort that has been done in the INERIS to provide an evaluation of the consequences of possible scenarios of a toxic leak on the surrounding biodiversity through the creation of the Biodiversity Severity Index (BSI). This indicator must take into consideration several criteria such as the toxicity of the liquid that is lost, the vulnerability of the surrounding environment etc.

This evaluation process requires an evaluation of the severity expressed as a single indicator so that the study does not only have a descriptive objective but might also be later considered while deciding whether or not allowing an industrial manager to start or continue some industrial activities as it is proposed or asking for additional protection measures. Therefore, all the data that are taken into account in this assessment must be aggregated together to get one single indicator.

The severity is a human concept dealing with human values and people's subjective feeling of what is important and what is not (or less). Thus, aggregating together the data to create such indicator requires to consider, survey, represent and use the system of values of experts or concerned individuals, considering the different criteria that may be taken into account by them. The concept of severity of an impact on biodiversity and the task of forecasting this impact involves several scientific fields such as toxicology and biology and requires to consider several criteria and data. For these two reasons it seemed appropriate to use a multi-criteria decision aiding approach to represent as well as possible the judgement of decision makers on this topic and classify according to these preferences the scenarios of accident in categories of severity. This approach and the scientific field that is associated to it (multi-criteria decision aiding) are central in this thesis. I interacted repeatedly with several experts with different scientific backgrounds so as to understand and model which criteria matter in this context, why they matter and how they impact the global judgement on the evaluation of the expected strength of a pollution. I particularly focused my attention on preference elicitation, studying various existing methods in order to choose in my application those that seem the most adapted to the context.

Doing so, I proposed a new preference elicitation method for the sorting problem, the Dominance Based Monte Carlo algorithm (DBMC), that deals with the sorting problem. This algorithm has two main specificities that are generally not combined by other methods. At first, we do not assume that the decision maker's reasoning follows some well known and explicitly described rules or logic system. The human judgement is probably too complex to be described by simple rules and may not be totally deterministic. The second specificity of the DBMC is its stochastic functioning. This algorithm is tested in some part of the Biodiversity Severity Index problem and

queensland-flash-flooding-hampers-rescue-crews-at-freight-train-derailment

the results that were obtained is compared to other decision aiding method used in the same context.

Public decision making being more compatible with methods based on formal decision rules, this method was finally not used in the creation of the Biodiversity Severity Index. However, we tested its performances in this context.

The first chapter aims at presenting the problem that we are facing. We will describe the context that makes risk of accidental pollution become an important issue, introducing the scientific fields that are related to our topic, namely, risk management and valuation of the environment and draw the legal frame of risks management in France that will bound our work. We show how this context and the associated constraints led us to opt for specific choices on methodological approaches.

The second chapter of this thesis introduces multi-criteria decision aiding that is a major issue of this work. We will provide to the reader some important definitions, notions and notations about multi-criteria decision aiding. Within this discipline, two topics seem particularly important and thus will be presented in details: The construction of a hierarchy of criteria and preference elicitation for criteria aggregation.

The third chapter will describe the Biodiversity Severity Index and the sorting method that we proposed in order to obtain its value for a scenario of accidental pollution. This method was based on a hierarchy of criteria. We describe this hierarchy, the criteria aggregation methods that we used as well as the reasoning that led us to this model, including the interaction with various experts.

Finally, a more theoretical work will be exposed, dealing with a proposition of multi-criteria elicitation method for sorting based on the dominance principle (also called monotonicity) and Monte Carlo principle. We describe the functioning of this algorithm and we make a comparison with other similar algorithms. We show some of its theoretical properties and compare its practical performances to those of other elicitation algorithms for the sorting problem. Then, we will conclude and give the pros, the cons of this method and some possible context in which it could be used.

Contents

Introduction 1

table of contents 5

1 Evaluating the risk of accidental pollution 7

1.1 The INERIS and its needs . . . 10

1.1.1 The INERIS . . . 10

1.1.2 About risk of accidental pollution . . . 11

1.1.3 The ministry's demand . . . 11

1.2 Major accidental risks management . . . 12

1.2.1 About risks . . . 12

1.2.2 Introduction to risk governance frameworks . . . 15

1.2.3 How to adapt this framework to an environmental dimension? . 28

1.3 The environmental dimension of risk . . . 28

1.3.1 Definition of biodiversity . . . 28

1.3.2 Valuing environmental losses . . . 29

1.4 Definition and characterization of our problem . . . 33

1.4.1 Definition of our problem . . . 33

1.4.2 Dividing resources and biodiversity . . . 33

1.4.3 Using a monetarist approach for assessing the severity on resources 34

1.4.4 The Biodiversity Severity Index . . . 34

1.4.5 How scientific issues and challenges led to precise methodological choices . . . 36

2 Multi-criteria decision aiding 39

2.1 Structuring a multi-criteria decision aiding problem . . . 42

2.1.1 Introduction to decision aiding . . . 42

2.1.2 Actors involved in a decision aiding process . . . 47
2.1.3 Defining the problem formulation: what type of output is expected 48
2.1.4 Choosing an approach . . . 49
2.1.5 Choosing the variables of the problem . . . 50
2.1.6 Value focused thinking and the hierarchies of criteria . . . 53
2.2 Multi-criteria Aggregation procedures . . . 59
2.2.1 Basic notions of criteria aggregation . . . 60
2.2.2 Outranking methods . . . 61
2.2.3 Synthetic criteria and utility . . . 69
2.2.4 Rule based methods . . . 70
2.2.5 SMAA: A stochastic methods to deal with uncertainty or imprecise informations or group decisions . . . 72
2.2.6 Choosing the appropriate Multi-criteria aggregation procedure . 73
2.3 Elicitation methods . . . 76
2.3.1 Aggregation approach . . . 77
2.3.2 Disaggregation approach . . . 77
2.3.3 Expression of the preferences for the sorting problem with a disaggregation approach . . . 78
2.3.4 UTA methods: a disaggregation approach algorithm based on utility . . . 79
2.3.5 A mixed integer programming algorithm for disaggregation elicitation with MR-Sort . . . 81
2.3.6 An heuristic algorithm for disaggregation elicitation with MR-Sort 83
2.3.7 An heuristic algorithm for disaggregation elicitation with 2-additive NCS model . . . 86
2.3.8 ORCLASS: A tool to interact with the decision maker . . . 87
2.3.9 DRSA: A disaggregation elicitation algorithm based on rule based principle . . . 89
2.3.10 Logistic and choquistic regression . . . 92

3 Biodiversity Severity Index 95
3.1 Final hierarchy of the Biodiversity Severity Index . . . 98
3.1.1 Local Biodiversity Severity Indices for surface water targets . . 98
3.1.2 Destructive potential . . . 100
3.1.3 Distinction between facts and values in our model . . . 102
3.2 Scales on the criteria . . . 103

3.2.1 Standard scales . . . 103
3.2.2 Boolean values and scales . . . 104
3.2.3 Semantically defined ordinal scales . . . 104
3.2.4 A five value levels scale for both the *destructive potential*, the *value of the environment*, the *vulnerability of the target* and the *local biodiversity severity index* . . . 107
3.2.5 Reference points for the value levels . . . 108
3.2.6 Adaptation of this methodology to the impact on ground targets 110
3.2.7 Conclusion on the hierarchy of criteria . . . 110
3.3 Construction process for the hierarchy of criteria . . . 111
3.3.1 Adapting the Value Focused Thinking approach to our problem 111
3.3.2 First proposition from the INERIS . . . 113
3.3.3 First modifications of the hierarchy of criteria . . . 117
3.3.4 Interacting with experts . . . 118
3.3.5 Second modification based on the *Muséum National d'Histoire Naturelle*'s expertise . . . 121
3.3.6 Validation of the previously obtained model with a toxicologist . 122
3.3.7 Validating the destructive potential as a criterion . . . 123
3.3.8 Including the *residence time* and the concentration as new criteria 124
3.3.9 Adding the *resistance* criterion by interacting with the *Muséum National d'Histoire Naturelle* . . . 126
3.3.10 From an exhaustive model to a "user friendly" one . . . 126
3.3.11 Why we stopped the evolution and accepted the hierarchy as it is. 127
3.4 Aggregation to get the destructive potential . . . 129
3.4.1 Finding possible dependences between sub-criteria of the destructive potential . . . 130
3.4.2 About additive methods for the destructive potential . . . 131
3.4.3 Defining the aggregation method for the destructive potential . 132
3.5 Aggregation to get the Local Biodiversity Severity Indices . . . 135
3.5.1 Frame of this sub-problem . . . 136
3.5.2 The questionnaire . . . 138
3.5.3 Processing the data . . . 140
3.5.4 Aggregating the criteria with a MR-Sort model . . . 145
3.5.5 A MR-Sort model with vetoes . . . 148
3.6 Aggregation of the Biodiversity Severity Index . . . 155

3.6.1 Description of the aggregation procedure . . . 156
3.6.2 Practical example of the evaluation of a scenario of accidental pollution . . . 157

4 Dominance Based Monte Carlo algorithm 163
4.1 Basic ideas about the Dominance Based Monte Carlo algorithm . . . 166
4.1.1 Using the monotonicity and the learning set as a frame . . . 166
4.1.2 A model free approach . . . 167
4.1.3 A stochastic functioning . . . 167
4.1.4 From the existing methods in the literature to the Dominance Based Monte Carlo algorithm . . . 168
4.2 Description of the algorithm . . . 169
4.2.1 Theoretical basis and notation . . . 170
4.2.2 Process of a trial . . . 172
4.2.3 Collecting and using the trial's information . . . 178
4.2.4 Aggregating the DBMC Vector . . . 179
4.2.5 Expected properties . . . 181
4.2.6 Summary of the theoretical properties . . . 185
4.2.7 Theoretical properties of the AverageDBMC and the ModeDBMC 185
4.3 Experimental validations and comparison to other sorting algorithms . 187
4.3.1 Stability . . . 187
4.3.2 Presentation of the k-fold cross validation . . . 190
4.3.3 Comparison of the DBMC algorithm with other elicitation algorithms through a k-fold validation . . . 192
4.4 What are the problems to which this algorithm can be adapted? . . . 201
4.5 Perspectives . . . 202
4.5.1 Modified idiosyncrasy: An other experimental test for the efficiency of the algorithm . . . 202
4.5.2 Possible applications of the Dominance based Monte Carlo . . . 203

Appendices 211
A Description of the DOMLEM algorithm . . . 211
B Reference points for the criteria of the Biodiversity Severity Index . . . 213
B.1 Reference points of the *Local Biodiversity Severity Indices* . . . 214
B.2 Reference points of the *Value of the environment* . . . 216

C Elicitation of the preferences of the experts of the MNHN for Local Biodiversity Severity indices 218

Chapter 1

Evaluating the risk of accidental pollution

Contents

1.1 The INERIS and its needs 10

1.1.1 The INERIS 10

1.1.2 About risk of accidental pollution 11

1.1.3 The ministry's demand 11

1.2 Major accidental risks management 12

1.2.1 About risks 12

1.2.2 Introduction to risk governance frameworks 15

1.2.3 How to adapt this framework to an environmental dimension? 28

1.3 The environmental dimension of risk 28

1.3.1 Definition of biodiversity 28

1.3.2 Valuing environmental losses 29

1.4 Definition and characterization of our problem 33

1.4.1 Definition of our problem 33

1.4.2 Dividing resources and biodiversity 33

1.4.3 Using a monetarist approach for assessing the severity on resources 34

1.4.4 The Biodiversity Severity Index 34

1.4.5 How scientific issues and challenges led to precise methodological choices 36

"Only after the last tree has been cut down. Only after the last river has been poisoned. Only after the last fish has been caught. Only then we will find that money cannot be eaten.", Cree Indian Prophecy.

This chapter aims at explaining the context in which our risk problem happens to be concerning and introduces risk management to the reader. We will first present the INERIS and the request that was made by them. Then, we will give some elements of risk management and environmental valuing. Finally, we will give a proper definition of our problem, highlight the challenges that are associated with it and explain some methodological choices that we made to deal with these challenges.

1.1 The INERIS and its needs

1.1.1 The INERIS

The INERIS (Institut national de l'environnement industriel et des risques) is a French public institution under the authority of the Ministry of Ecology, Sustainable Development and Energy. It aims at contributing to the protection of both workers, citizens, goods and the environment from eventual negative impacts of industrial and mining activities. Its surveys are technical supports to public authorities for the development and implementation of regulations, standards, reference methods and certification systems. The INERIS mainly has three missions:

- It conducts research programs aiming at a better understanding of phenomena that may affect the environment, public health and at improving its expertise capacity on risk prevention.
- It supports the French Ministry of Ecology, Sustainable Development and Energy making researches about safety and sustainability in the industry.
- The INERIS also provides surveys on the previously mentioned topics for various actors such as industries or local authorities.

One of the INERIS activities consists in performing risk studies (defined in 1.2.2) on industrial plants and product storages. The goal of such a process is to determine whether or not the risk of an eventual scenario of accident is acceptable and to propose measures to improve the safety.

1.1.2 About risk of accidental pollution

On 4 October 2010, in Ajkai Timfldgyr alumina plant in Ajka [1][2], in western Hungary, a part of the dam of a reservoir collapsed (as illustrated in Figure 1.1), freeing approximately one million cubic meters of an extremely basic liquid (with a pH value of 13). Ten people died, and 150 people were injured. The waste extinguished all life in the Marcal river, and reached the Danube. About 40 square kilometres of land were affected. This is only one example of accidental pollutions due to industrial activities that happened in the last decades, we could also mention the ecological catastrophe of Baia Mare in Romania, 2000, or the accident of Algona in the USA in 2001 both with very serious consequences on the environment, resources and public health. Since the industrial revolution, during the ninteenth century, the impact of man on the environment has continued to rise from both chronic and accidental pollution. In the second part of the twentith century these impacts became more and more visible since scientists and activists alarmed the public [Aspe, 1989]. In 1992, the United Nations Conference on Environment and Development mentioned three main objectives:

- Preservation of biodiversity
- Sustainable use of the biological resources
- The fair and equitable sharing of benefits arising from the use of resources

Thus, the reduction of human footprint, whether on natural resources available, on biodiversity or on our own health, became a major issue for both public opinion and governments.

Today, while applying the current risk studies in France, there is no regulatory requirement for industrial operators to assess and prevent environmental consequences of eventual accidental pollution. Indeed, in its current state, only scenarios involving explosions, fire and toxic gas releasing are treated.

1.1.3 The ministry's demand

Within the mission of supporting the ministry's decisions the INERIS was mandated to create a tool to prevent society from risk of accidental pollutions. The aim of this

[1]Find information about the accident in Ajka at http://www.theguardian.com/world/2010/oct/05/hungary-sludge-disaster-state-of-emergency

[2]Find information about the accident in Ajka at http://www.bbc.co.uk/news/world-europe-11475361

work is to propose a method that would take into account the environmental impact of scenarios of accident during the risk analysis. This mission did not have a formal scientific formulation. Indeed, the term of environmental consequences covers many different types of consequences on both human uses of natural resources or intrinsically ecological consequences. The means to get to this consideration of the environment in risk studies were not clear.

In order to explain more clearly how this informal problem gave birth to a formal problem, we will introduce some notions of environmental valuation, risk and risk management in the industry.

Figure 1.1: Ajka accident, Hungary, 2010. Source: Io9 blog[3]

1.2 Major accidental risks management

1.2.1 About risks

This thesis will mainly be anchored in existing accidental governance frameworks. Therefore, first of all, it is necessary to introduce the reader to the risk concept and associated working framework developed and used by the risk community.

Definition of risks

Risk is a word that has been given to numerous definitions across the literature according to different sources from different schools of thought and perspectives [Aven, 2012].

[3]Url: http://io9.gizmodo.com/5663280/hungarys-river-of-death-as-seen-from-space

Etymologically, most dictionaries consider that the word "risk" comes from the Italian word "rischio", itself derived from the Latin word "rescecum" meaning "which sharps". From this root, "risk" is associated to a painful or unwanted outcome. A few dictionaries relate that the word risk comes from the Arabic *rizq* that may be translated by "that which God allots". This possible root underlines the unpredictable and uncontrollable nature of risks. Although we will obviously not discuss in this thesis which of these two roots is the real one, we can observe that they both refer to one of the two main elements of risk: "unwanted outcome" and "uncertainty of the future events".

Let us now mention some of the current definitions that are being given to risk depending on the field that is considered.

- The Oxford English dictionary [Stevenson, 2010] defines risk as "the possibility of loss, injury, or other adverse or unwelcome circumstance; a chance or situation involving such a possibility".
- The Health and Safety Executive [Jan Duijm et al., 2008] [4] defines risk as "the combination of the likelihood and the strength of the possible scenarios".
- On *Investopedia*[5] risk is defined as "The possibility that a company will have lower than anticipated profits, or that it will experience a loss rather than a profit".
- In the OHSAS [6] risk is defined as "a combination of the likelihood of an occurrence of a hazardous event or exposure(s) and the severity of injury or ill health that can be caused by the event or exposure(s)".
- Risk is understood by the IRGC[7] in *the white paper on risk governance* [Ortwin, 2005] as "an uncertain consequence of an event or an activity with respect to something that humans value".

[4]The Health and Safety Executive (HSE) is a non-departmental public body of the United Kingdom. It is the body responsible for the encouragement, regulation and enforcement of workplace health, safety and welfare, and for research into occupational risks in England and Wales and Scotland.

[5]Link available at: http://www.investopedia.com/terms/b/businessrisk.asp

[6]OHSAS 18001, Occupational Health and Safety Management SystemsRequirements (officially BS OHSAS 18001) is an internationally applied British Standard for occupational health and safety management systems. It exists to help all kinds of organizations put in place demonstrably sound occupational health and safety performance.

[7]The International Risk Governance Council (IRGC) is an independent non-profit organisation, based at École Polytechnique Fédérale de Lausanne (EPFL) in Lausanne, Switzerland.

This thesis being about hazard risk assessment, we will thereafter choose the definition given by organisms that deals with such problems and thus the definitions of the IRGC's retained our attention. Furthermore, the fact that the IRGC's definition refers to consequences on things that human values seems encompassing, potentially including environmental issue and particularly adapted to a public issue such as this thesis, at the center of which human matter should be placed.

Accidental and chronicle risks

Risks are generally divided into two distinct categories: accidental and chronicle. Chronicle risks refers to consequences of the impact of the normal and expected functioning of the studied plant. For instance, we will consider as a chronicle risk the potential consequences of the daily dropping of 50 liters of polluted water by an industrial plant. By accidental risk, we refer to the possible consequences of an unexpected and uncontrolled functioning of a human process, tool or facility (an explosion for example). INERIS divided the study of these two fields in two directorates; the DRA (Direction des Risques Accidentels) that studies accidental risks and the DRC (Direction des Risques Chroniques) that studies the chronicle risks. In this thesis, due to the ministry's request, we will only be dealing with accidental risks.

A gradual awareness of accidental risks in the industry

In August 31, 1794 at 7:15 am, in Grenelle powder factory (Grenelle has since become a district of Paris), between 30 and 150 tons of black powder (also called gunpowder) exploded in an urban area, killing more than a thousand people, workers and neighbours [Le Roux, 2011] (see Figure 1.2). This tragedy was the first eye opener on the risks induced by industrial activities in France and strongly influenced the imperial decree of 1810 on dangerous, unhealthy and inconvenient facilities. More recently, Toulouse's accident in 2001 considerably impacted risk perception and management in France. The importance of the needed actions implies to reason on the long term.

The industrial sector is an important part of French and European economy. According to the INSEE, in 2014, industrial activities represented 12.4 % of French gross domestic product and employed 3.1 million people in France [8]. Nevertheless, industrial activities must often involve some materials, products and processes that are susceptible to create unwanted impacts or accidents. Between 1992 and 2005, the ARIA

[8] http://www.gouvernement.fr/partage/3813-l-industrie-en-france

database on industrial accidents registered in France 21 601 accidents related to industrial activities that led to the death of 625 persons, 15 168 wounded, numerous pollutions and economical losses[9]. Thus, management of the risks induced by the industry has become an important topic for either populations, authorities and industrial managers.

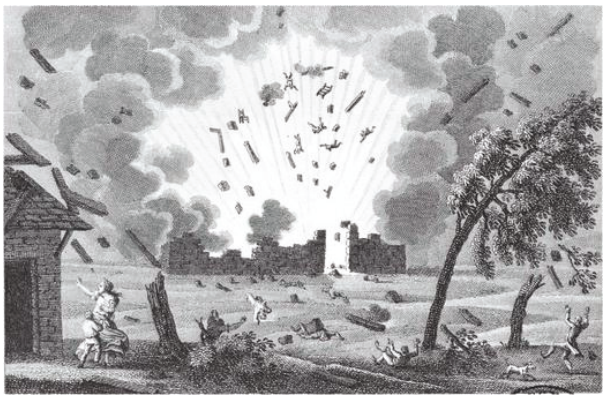

Figure 1.2: Explosion of Grenelle powder factory, August 31, 1794. Source:[Le Roux, 2011]

1.2.2 Introduction to risk governance frameworks

In this subsection, we will introduce some basic notions on the way risk is generally managed by the different stakeholders that are in contact with this problem, what are the methodological and scientific tools that were created to deal with this issue and how risk management is framed and regulated by the legislation and the authorities.

When do we practice risk studies in the industry?

The accident in the town of Seveso in Italy, during which a significant toxic cloud was released in 1976, prompted European states to adopt a common policy on the prevention of major industrial risks. From June 24 1982, the Seveso directive (now replaced by Seveso 3 directive[10]) requires states and companies to identify the risks associated with some listed hazardous industrial activities and take the necessary steps to address it. Today in France, based on the amount of hazardous products present in

[9] http://www.aria.developpement-durable.gouv.fr/

[10] Directive of the European parliament and of the council of 4 July 2012 on the control of major-accident hazards involving dangerous substances

the establishment, public and private facilities can be classified into different categories of *Classified facilities for the protection of the environment* (ICPE). According to its category, a facility will be subject to more or less frequent risk studies by a certified organism to be authorized by the *Regional Directorates of Environment, Planning and Housing* (DREAL) to act on French soil. The main purpose of this thesis relates to these risk studies.

In order to understand how these risks studies work, we are about to introduce the *IRGC framework* which is a recognized methodological base for risk management and risk studies. It must be noticed that this framework is very similar to other frameworks proposed on the same topic by other organizations such as the HSE framework [Jan Duijm et al., 2008], the UK cabinet approach [UKC, 2015], and the All Hazards Risk Assessment Methodology Guidelines [Haz, 2012]. The IRGC framework being the most recent and quite encompassing, we thereafter use it as an important base for the further works although we will also refer to other frameworks in some circumstances.

The IRGC framework

The International Risk Governance Council (IRGC) is an independent non-profit organization, based at École Polytechnique Fédérale de Lausanne (EPFL) in Lausanne, Switzerland. The IRGC aims at improving the understanding, management and governance of emerging systemic risks that may represent a threat either to human health, the environment, the economy or society.

In this purpose, they developed a framework [Ortwin, 2005] for risk governance which includes 5 elements (described in Figure 1.3)

1) Risk Pre-Assessment: early warning and framing the risk in order to provide a structured definition of the problem, of how it is framed by different stakeholders, and of how it may best be handled.

2) Risk Appraisal: combining a scientific risk assessment (of the hazard and its probability) with a systematic concern assessment (of public concerns and perceptions) to provide the knowledge base for subsequent decisions

3) Characterization and Evaluation: in which the scientific data and a thorough understanding of societal values affected by the risk are used to evaluate the risk as acceptable, tolerable (requiring mitigation), or intolerable (unacceptable)

4) Risk Management: the actions and remedies needed to avoid, reduce transfer or retain the risk

5) Risk Communication: how stakeholders and civil society understand the risk and participate in the risk governance process

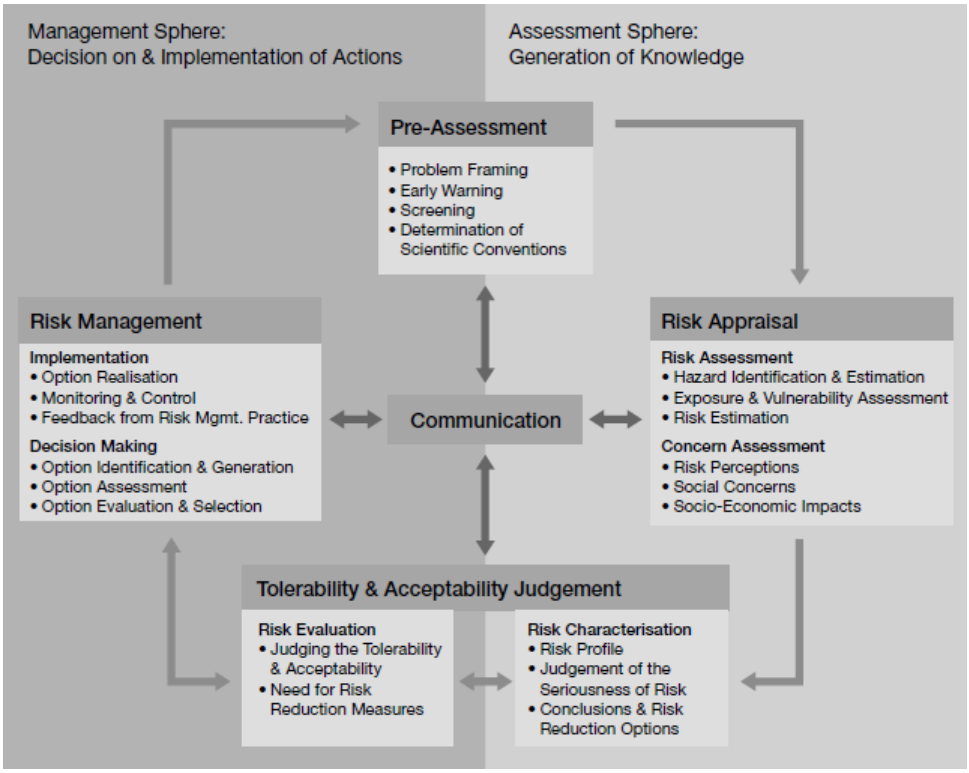

Figure 1.3: IRGC Risk Governance Framework. Source: [Ortwin, 2005]

The first three steps of this framework are directly related to the present work and we believe that it is important to introduce the reader to some basic notions concerning them so that our problem and our approach to deal with it can be properly introduced. However, the topics of the two last parts are quite out of both the problems frame and my competences. Thus we will not treat these parts in this thesis.

Risk pre-assessment

The purpose of the pre-assessment phase is to capture both the variety of issues that stakeholders and society may associate with a certain risk as well as existing indicators, routines, and conventions that may prematurely narrow down, or act as a filter for, what is going to be addressed as risk [Ortwin, 2005]. This includes among other things the identification of the laws, conventions that frame risk assessment, the identification of the control institutions and of the experts that might help in the process. This work

deals with risk of accidental pollution. Thus, in France, several laws and conventions frame the present work among which the three following worth being mentioned:

- The "Arrêté du 29 septembre 2005" frames the operating mode of a risk study. It specifies the rules of risks studies concerning the evaluation of the probability, the kinetic and the severity of the effects of all the scenarios that may happen and affect the interests covered by the article L. 511-1 of the "code de l'environnement" (the convenience in the neighbourhood, public health and safety, protection of the environment, conservation of sites, monuments and archaeological heritage).
- The "Circular of 10/05/2010" gives additional tools to conduct of a risk survey. The part about the methodological rules for drafting safety reports, the longest text, frames the measure of the likelihood in each case, consequences and the kinetics of scenarios (atmospheric dispersion, the hazards associated with liquefied petroleum gas in storage facilities, the hazards associated liquefied flammable gas in storage facilities ...). This section contains a large number of technical data specific to evaluate severity and likelihood in each case.
- The "European Directive" of 21 April 2004 (DRE) creates a system of environmental responsibility. The Directive's centrepiece idea is to prevent and remedy environmental damage, mainly from industrial sources, applying the polluter-pays principle. Indeed, the operator of a professional activity to which the Directive is now held financially responsible for repairing the damage it causes to the environment. The Directive also has a goal of under imminent threat of damage prevention: these operators are obliged to take necessary measures so that damage does not occur.

Risk Appraisal

This thesis will specifically focus on the evaluation of one dimension of the consequences of scenarios of accident: consequences on the environment. The evaluation of the consequences of a scenario is a part the "Risk appraisal" step. Risk appraisal is defined as the combination of two processes scientific risk assessment and a concern assessment.

Risk assessment

According to the white paper on the IRGC [Ortwin, 2005] "The purpose of risk assessment is the generation of knowledge linking specific risk agents with uncertain but

possible consequences [Lave, 1987][Graham and Rhomberg, 1996]. The final product of risk assessment is an estimation of the risk in terms of a probability distribution of the modelled consequences (drawing on either discrete events or continuous loss functions)". It refers to factual, physical and measurable characteristics of risk. Risk appraisal intends to produce the best possible scientific estimate of the physical, economic and social consequences of a risk source. As it will be explained later, multi-criteria decision aiding seems particularly appropriate to deal with this approach given that a large place is made for human perception and values. Risk assessment is a step that, among other things, deals with the necessity mentioned in every framework to determine the scenarios that are worth being taken into account, evaluate the likelihood (or probability) and the severity of every scenarios of accident. As mentioned before, this last step will constitute the main topic of this thesis.

As the following subsections will show, while evaluating either the likelihood of occurrence of scenarios or the severity of their consequences, the *Arrêté du 29 septembre 2005* and the *Circulaire du 10/05/2010* may be consider as both a scientific tool and a legal framework.

Concern assessment

Risk concern assessment refers to a study of how risk are perceived by society, what are the socio-economic impacts of risk. While studying public issues (risk issues in particular) it is important to figure out who are the stakeholders that could interact with the process of risk management, why they feel concerned about risks, what are their objectives and what are their possibility to impact this process [Ackermann and Eden, 2011]. Freeman et al. [Freeman, 2010] define stakeholders as "any group of individual that can affect or is affected by the achievement of an organization's purpose". Those affected are usually referred to as the claimants whereas those who affect are the influencers [Mitchell et al., 1997][Kaler, 2002]. The identification of the stakeholders and their inclusion into the decision process offers mainly two advantages:

- It increases the legitimacy of the decision process. Indeed, putting them aside, voluntarily or not, may generate outrage and suspicion both among participants and external observers.
- Then, the stakeholder's knowledge and opinion is often seen as a resource that benefits to the decision process.

A common approach [Eden and Ackermann, 1998][Enserink et al., 2010] in stakeholders analysis consists in classifying the stakeholders in a matrix power/interest according

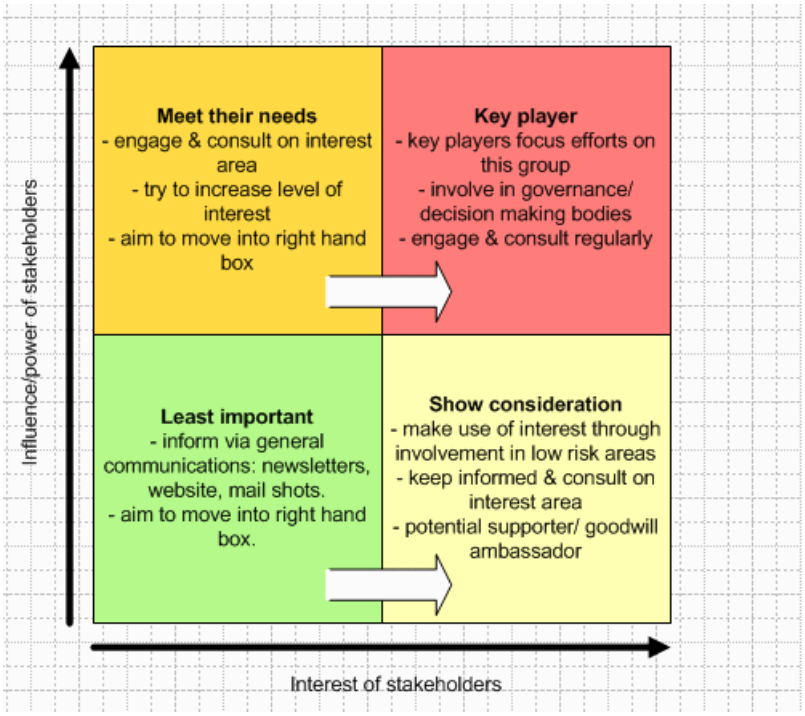

Figure 1.4: Power versus interest grid adapted from Eden and Ackermann [Eden and Ackermann, 1998]

to their interest and their capacity to have an impact on the situation (described in Figure 1.4). For each category of stakeholders this matrix provides the user advices on how to interact with the stakeholders, how important they are and the main goals that the user should follow while interacting with them.

Evaluating the likelihood of occurrence of scenarios

According to the *Arrêté du 29 septembre 2005*, the evaluation of the probability of occurrence must rely on methods whose relevance is demonstrated. It is recommended to rely on internationally recognized databases. It is also advisable when possible to determine the probability of occurrence of the triggers. The data used to represent the frequency of occurrence should be quantitative, semi-quantitative (fork) or qualitative (Article 3). The appendix 1 of the *Arrêté du 29 septembre 2005* determines the different classes of qualitative likelihoods, defines them and links them to probabilities (described in Figure 1.5).

The Circular of 10/05/2010 provides a large number of technical information to evaluate the likelihood of occurrence of the scenarios in various specific cases. The two

Probability class	**E**	**D**	**C**	**B**	**A**
Qualitative	Possible but extremely unlikely event	Very unlikely event	Unlikely event	likely event	frequent event
Description of the class	Is not impossible but did not happen worldwide on a large period of time	A similar event happened worldwide in this activity area but was subject to modifications significantly reducing its likelihood	A similar event happened worldwide in this activity area with no guarantee that the likelihood was significantly reduced since then	Likely to happen in the considered plant's lifetime	Already happened in the past or likely to happen several times in the considered plant's lifetime
Quantitative (unity per year)	10^{-5}	10^{-4}	10^{-3}	10^{-2}	

Figure 1.5: Table defining the likelihood classes from the appendix 1 of the *Arrêté du 29 septembre 2005*. Personnal translation made by myself

most common methods to evaluate the likelihood of occurrence of a scenario are the "bow-tie" methodology (to learn about the boe-tie methodology the reader may refer to [Khakzad et al., 2012][De Dianous and Fiévez, 2006]) and the study of the frequency of similar scenarios in similar circumstances.

Evaluating the expected severity of scenarios

The aim of our work is to provide a scientific methodological tool to evaluate the expected severity of scenarios of accident from the environmental perspective. For now, while assessing the severity of scenarios considered, only short term damages on humans due to toxic effects of pressure, thermal effects and effects associated with the impact of projectiles for men are considered.

The severity of a scenario is obtained by first evaluating the strength of the effects on the surrounding of the scenario's source and then counting how many people could

be present in these impacted areas. To do so, the *Arrêté du 29 septembre 2005* defines three types of danger areas (SEL, SELS, SEI[11]) for each scenario regarding the expected conditions of temperature and pressure in case the scenario happen (described on Table 1.1). These projections are generally made with the help of software modelling of physical phenomena such as EFFEX, PROJEX, MISSILE or EXORIS for the effects of projection and pressure conditions and the software PHAST 6.5.3.1 for thermal effects.

The appendix 3 of the decree provides a severity scale based on the number of persons in a SELS area, the number of people located in a SEL area and the number of people located in a SEI area. It is noteworthy that the decree also provides the values of these thresholds either in the case of a thermal danger of a toxic hazard or danger of overpressure (see table 1.6).

Effect on humans

	Overpressure effect	Threshold for thermal effects
	Mbar	kWm^{-2}
Indirect effects	20	
Irreversible effects	50	3
First lethal effects	140	5
likely lethal effects	200	8

Table 1.1: Table defining the danger areas. Source: Appendix 3 of the *Arrêté du 29 septembre 2005*

It is to be noticed that the severity of the consequences of scenario are generally considered as one single criterion. This criterion might possibly be an aggregation of several criteria as we will see later in this thesis.

Tolerability and acceptability judjement

Determining whether a risk is acceptable is a major goal in risk management and it is quite a sensitive issue due to the facts that this decision can have a major impact on people's lives and that this task can be seen as subjectively evaluated. Indeed, each

[11] area of significant lethal effects (SELS), area of lethal effects thresholds (SEL) area of irreversible effects thresholds (SEI)

Severity of the consequences	**SELS**	**SEL**	**SEI**
Disastrous	More than 10 persons exposed	More than 100 persons exposed	More than 1000 persons exposed
Catastrophic	Between 2 and 10 persons exposed	Between 11 and 100 persons exposed	Between 101 and 1000 persons exposed
Important	One person exposed	Between 2 and 10 persons exposed	Between 11 and 101 persons exposed
Serious		One person exposed	Between 2 and 10 persons exposed
Moderate			One person exposed

Figure 1.6: Table of evaluation of the human severity of scenarios. Source: Appendix 3 of the *Arrêté du 29 septembre 2005*

individual has her own willingness to accept risks and her own perception of how risk is important combining the probability that the scenario happen and its strength. This evaluation being a social concern, in order to make it fair and scientifically justified several formal methods are available and these methodological tools are strictly framed by the law. Although this topic is not directly part of the present work, the evaluation of severity aims at being used to evaluate acceptability of risk. Thus, it seems useful to introduce some basic notions of "risk acceptability" and of the methods that are used to evaluate it.

On the limits of zero tolerance principle

An easy answer to the risk issue would consist in saying "Risk is not acceptable in society. Let us highlight every accident that is theoretically possible, i.e., that has a probability strictly higher than 0 to occur, and let us propose a plan to make this accident impossible". This is called zero tolerance principle. Applied to our daily life, this principle would lead to make swimming, driving or climbing stairs illegal which, obviously, is not a good solution. In accidental risks prevention also this approach could seem excessive, some accidents being theoretically possible but in reality extremely unlikely to happen. Then, we admit that some risk should be considered acceptable. Conversely, as we saw before, some other risks should not be accepted by the society. The main issue in risk governance consists in finding the happy medium between a

too restrictive zero tolerance principle and a dangerous anarchy, a policy that would protect what humans care about from risks without avoiding innovation and useful human activities.

Social and individual risk evaluation

While assessing risk acceptability on human beings, two ways of measuring it, distinct and complementary, are generally considered: Social and individual risk acceptability judgement (Illustrated in Figure 1.7).

By individual risk we mean the annual probability for a human being present on a given point of the space to be killed by any accident happening in the studied industrial plant. According to the HSE framework [Jan Duijm et al., 2008] "societal risk is defined as the relationship between frequency and the number of people suffering from a specified level of harm in a given population from the realisation of specified hazards". In other words we could says that social risk acceptability judgement covers the consideration risk induced by one, several or all the possible scenarios on the whole society while the individual risk acceptability judgement covers the evaluation of the risks induced by all the possible scenarios on one given point of the space.

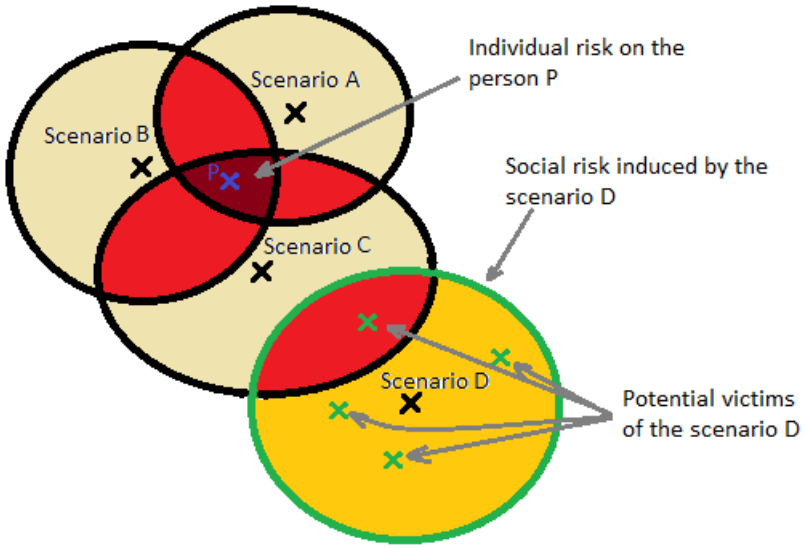

Figure 1.7: Illustration of the concepts of social and individual risk.

In the French legislation societal risk assessment is used to decide whether or not risks are acceptable on an industrial plant while individual risk is used to create risk maps that will help public authorities to decide where new infrastructure will be allowed to

be built. The purpose of our work is about the creation of a risk assessment method that could be defined as both social and individual given that the assessment will be made on one specific scenario and one specific target. We will now introduce the two most used methods to evaluate social risk acceptability: risk exposure table and FN curve. The following methods aim to deal with a problem in which we consider that:

- The list of all the possible scenarios of accident is already known
- An evaluation of the likelihoods or probabilities of these scenarios has been performed
- An evaluation of the severity of each scenario regarding human consequences has been performed

Risk matrices

In France, the use of a risk exposure table is imposed by the *Circulaire du 10/05/2010*. The concept is based on a two dimensions table in which the likelihood and consequence level of the scenario are then cross-tabulated to give a risk exposure rating. This determines whether a risk is categorised as acceptable, moderate, or unacceptable in other cases. Using this method, each scenario is assessed individually and, in most of the uses, there is no aggregation method to combine them with each other. From a multi-criteria decision aiding point of view, this method might be considered as a rule based method or clustering (considering the probability of a scenario as a criterion). Although this method suffers from weak points as claimed by [Cox, 2008] (namely poor resolution, errors, suboptimal resource allocation and ambiguous inputs and outputs) and the specific matrix imposed by the *Circulaire du 10/05/2010* could be seen as having too wide classes of both likelihood and severity, this method is easy to use and to understand and thus accepted by public opinion and decision makers. An example of a risk matrix is given in Figure 1.8.

Likelihood	Potential Consequence				
	Negligible	Minor	Moderate	Major	Extreme
Almost Certain	Medium	High	High	Very high	Very high
Likely	Medium	Medium	High	High	Very high
Possible	Low	Medium	Medium	High	High
Unlikely	Low	Low	Medium	Medium	High
Rare	Low	Low	Low	Medium	Medium

Figure 1.8: Example of a risk matrix

FN curve

Societal risk can be represented by FN curves [Wiley, 2009], which are plots of the cumulative frequency (F) of various accident scenarios against the number (N) of casualties associated with the modelled accidents. The plot is cumulative in the sense that, for each number of casualties N, F is the sum of the probabilities of the scenarios that would cause N casualties or more. Often "casualties" are defined in a risk assessment as fatal injuries, in which case N is the number of people that could be killed by the accidents. Then the curve corresponding to the risks induced by an industrial plant will be compared to reference curves (also referred to as tolerance criteria) to assess the acceptability of the risk on the studied plant[Baybutt, 2012]. An example of an FN curve is given in Figure 1.9.

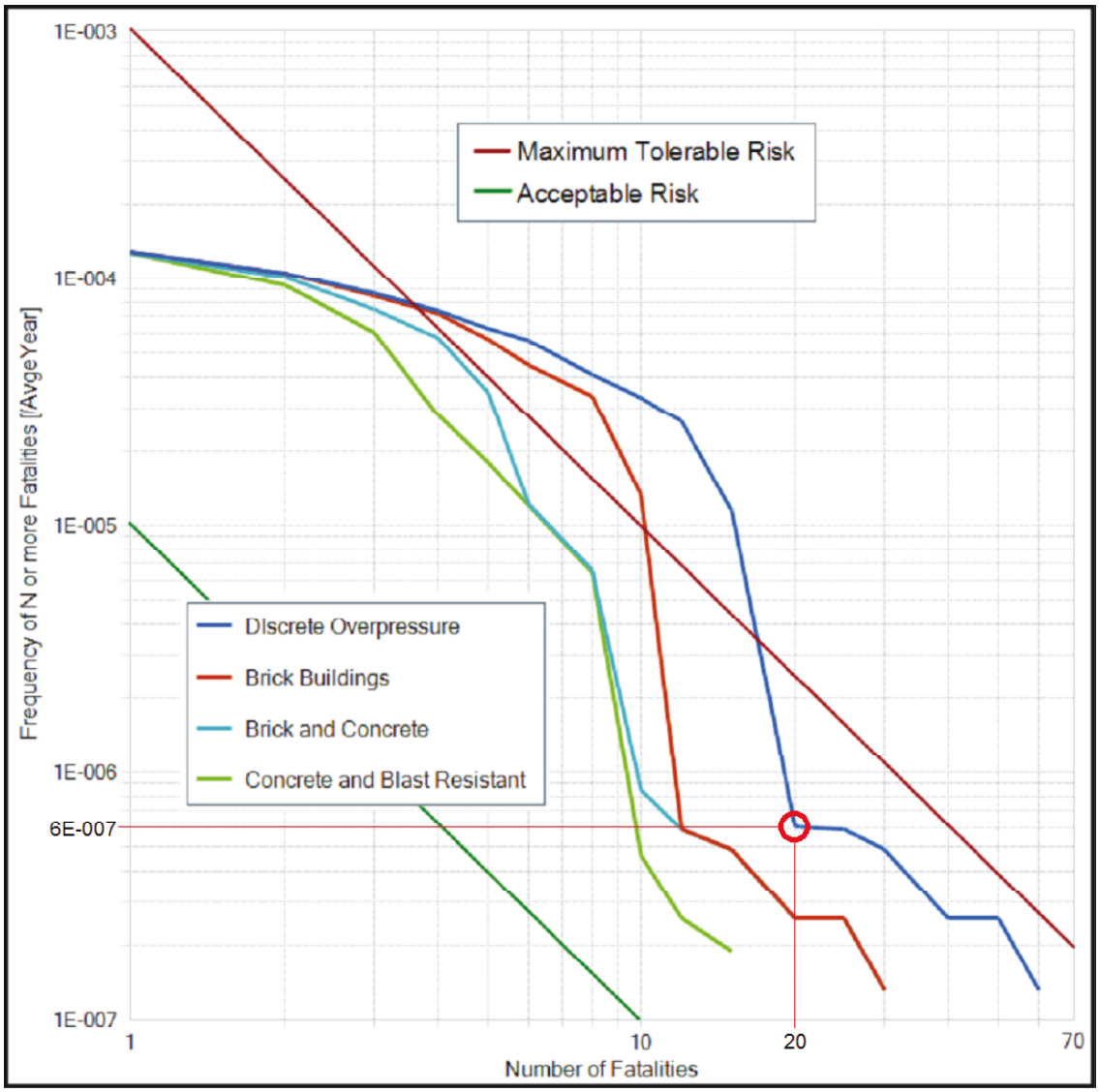

Figure 1.9: An example of Societal Risk [13], FN Curve with Risk tolerance criteria displayed. We can see on this figure that with the option "discrete overpressure" there is every year a probability of 6×10^{-7} that an accident kills 20 persons or more.

I will not here mention the pros and the cons of these methods nor will i make a choice between them, this part of the risk management being out of the frame of this thesis. Nevertheless, it worth mentioning that while the risk matrices are now imposed by the legislation, the use of the indicator that we created in this thesis with a FN curve would probably not need any modification.

[13] from DNV GL risk management website http://blogs.dnvgl.com/software/2016/01/good-safety-practices-equal-good-business

1.2.3 How to adapt this framework to an environmental dimension?

The goal of this work is to help the INERIS in the study of the possibilities to include an environmental dimension in risk studies. It seems to us that the framework previously described could be a good basis for an environmental evaluation of risks. Indeed, is was developed, approved and used by a large part of the risk community and thus it is more likely to be accepted in this new context than a new one. Furthermore, when may notice that, while applying this process to human risks, the list of possible scenario and the evaluation of their likelihood are already provided for an eventual environmental risk study. The main work that remains to do consist in evaluating the severity of the consequences on the environment. This will be the main purpose of this thesis.

1.3 The environmental dimension of risk

As we stated earlier, the currently used methodology considers the strength of consequences of a scenario through a short term human perspective and there is a social and political demand for the consideration of other criteria to evaluate it. The criterion that we will focus on across this thesis is the environmental impact of accidental pollution. These impact being due to accidental event instead of chronicle risk and impact being specifically located, this impact will be considered by most of the ecology community as *disturbance* which refers to a constraint, variable in space and time rather than *stress* which refers to a permanent (or in all places, or both) reduction in the average environmental quality [Lorrillière et al., 2012]. Obviously, measuring the severity of the consequences of an accident on something implies to understand the initial value of this thing (before any accident occurs), how impacted it could be by the accident and thus how fragile it is before any accident occurs.

Thus, it will be useful to introduce in this section some fundamental ideas of environmental valuing and environmental vulnerability.

1.3.1 Definition of biodiversity

According to the Rio 1992 Convention on Biological Diversity “ “Biological diversity” means the variability among living organisms from all sources including, inter alia, terrestrial, marine and other aquatic ecosystems and the ecological complexes of which

they are part: this includes diversity within species, between species and of ecosystems. "Biological resources" includes genetic resources, organisms or parts thereof, populations, or any other biotic component of ecosystems with actual or potential use or value for humanity". We will thereafter consider this definition while mentioning the conservation of biodiversity.

1.3.2 Valuing environmental losses

In order to improve the management of risks linked to accidental pollution it is important to understand "how severe would the environmental consequences of a given accident be if it would happen?". This question clearly leads to question ourselves on how important are the services that are given by the environment, either through monetary value (the value of global crop pollination services is estimated to 153 billion euros per year in the world i.e., 9.5% of the value of the global food production value [Gallai et al., 2011]) and non monetary value.

The value that is given to environment is generally decomposed into two main categories: the use-value, representing the value of the benefits that man can obtain, directly or indirectly, from the environment and the non-use value representing the intrinsic sake of minimizing the impact of human on the environment [Jochem, 2006][Pascual et al., 2010]. Each of these two categories are also subdivided in sub-categories as described in Table 1.2. Let us define each of these categories:

- **Direct use value:** This term refers to a direct and obvious benefice for man (monetary or not) obtained through human exploitation of natural resources. For example the pleasure to hike in a forest, the monetary benefice of fishing in the sea is part is considered as part of the direct use value. The value that we give to these services directly depends on the benefice that is associated to it.
- **Indirect use value:** Indirect use value refers to services provided to human by the environment without a human extraction or use activity. We could mention the role of pollination bees in agriculture, the role of mangrove swamp as a protection from erosion and storms or the impact of vultures in reducing the spread of diseases. The value that we give to this service depends of the value of other activities that may be impact by the loss of this service.
- **Option values:** The value that people set for having and retaining the option to use a product or a service if its need increases.

- **Bequest values:** The bequest value is related to the satisfaction to know that the future generations will benefit of the natural goods.
- **Existence value:** The existence value relates to the fact to know that some particular species or ecosystems exist. For instance many people care about the fact that the rain forest is being destroyed although they do not plan to spend holidays there nor to get any personal benefits from it.

Total economical value				
Use value			**Non-use economical value**	
Direct Use	Indirect Use	Option Value	Bequest Value	Existence Value
Outputs directly consumable	Functional benefits	Future direct and indirect values	use and non-use value of environmental legacy	value from knowledge of continued existence

Table 1.2: Decomposition of environmental value [Jochem, 2006]

Several economical tools have been used to value environmental losses on use value such as market based method or travel cost method [Ott et al., 2008]. In this document we will focus on non-use evaluation. Thus, we will here present methods that deal with non-use evaluation:

- Contingent Valuation Method
- Hedonistic Price Method
- Restoration Costs

1.3.2.1 Contingent Valuation Method

The main principle in monetary valuation is to get the willingness to pay (WTP) of the affected individuals (i.e., the price that they would agree to pay) to avoid a negative impact, i.e., to prevent biodiversity loss or to improve the environmental situation on a given location. The idea of contingent valuation method is that values should be based on individual preferences that are being elicited by direct questionnaires through the WTP method. In order to do so, the elicited person get proposed several scenarios of

improvement or restoration of the environment and get asked the highest price that she would agree to pay for these scenario. As an example [Katossky and Marical, 2011] applied this methodology to make an evaluation of a scenario of accident on the "Marais du Cotentin" (a French National Park located in Normandy). The authors presented successively to the elicited persons two scenarios of accidents in the "Marais du Cotentin" involving a truck containing a large amount of a very toxic liquid. Some precise descriptions of the scenarios were provided to the subjects (with among other things the illustration 1.10) and it was considered that these scenarios would result in an important degradation of the biodiversity of respectively 30 000 and 90 000 acres of protected area. Then they proposed to the subjects three possibilities: leaving the place in its current polluted state, a restoration plan resulting in a complete restoration of the impacted area to its initial state and an intermediate solution including a partial restoration of the impacted area. Given that these measures would have a cost that would finally impact everybody's taxes the elicited person get asked what would be the highest price that they would agree to pay for each of these possibilities. The average price that people are willing to pay for the total restoration may be considered as severity of the scenario of accident.

1.3.2.2 Hedonistic Price Method

The basic premise of the hedonistic pricing method is that the price of a marketed good is related to its characteristics, or the services it provides. For example, the price of a laptop reflects the characteristics of that laptop: performance, quality of the screen, quality of the sound, etc. Thus, this method aims to find the WTP for environmental goods as known in related markets. This technique seeks to elicit preferences from actual, observed market based information. Preferences for the environmental good are revealed indirectly. For example to evaluate the value of a given forest to man applying the hedonistic price method would consist in looking at the price of the houses around this forest and comparing these prices with the prices of other houses with similar characteristics (space , surrounding services etc) excepted that they are not located near a forest. The difference of price will be attributed to the value that is given by man to this forest.

1.3.2.3 Restoration Costs

The restoration cost approach is based on the cost of replacing or restoring a damaged asset to its original state and uses this cost as a measure of the benefit of restoration.

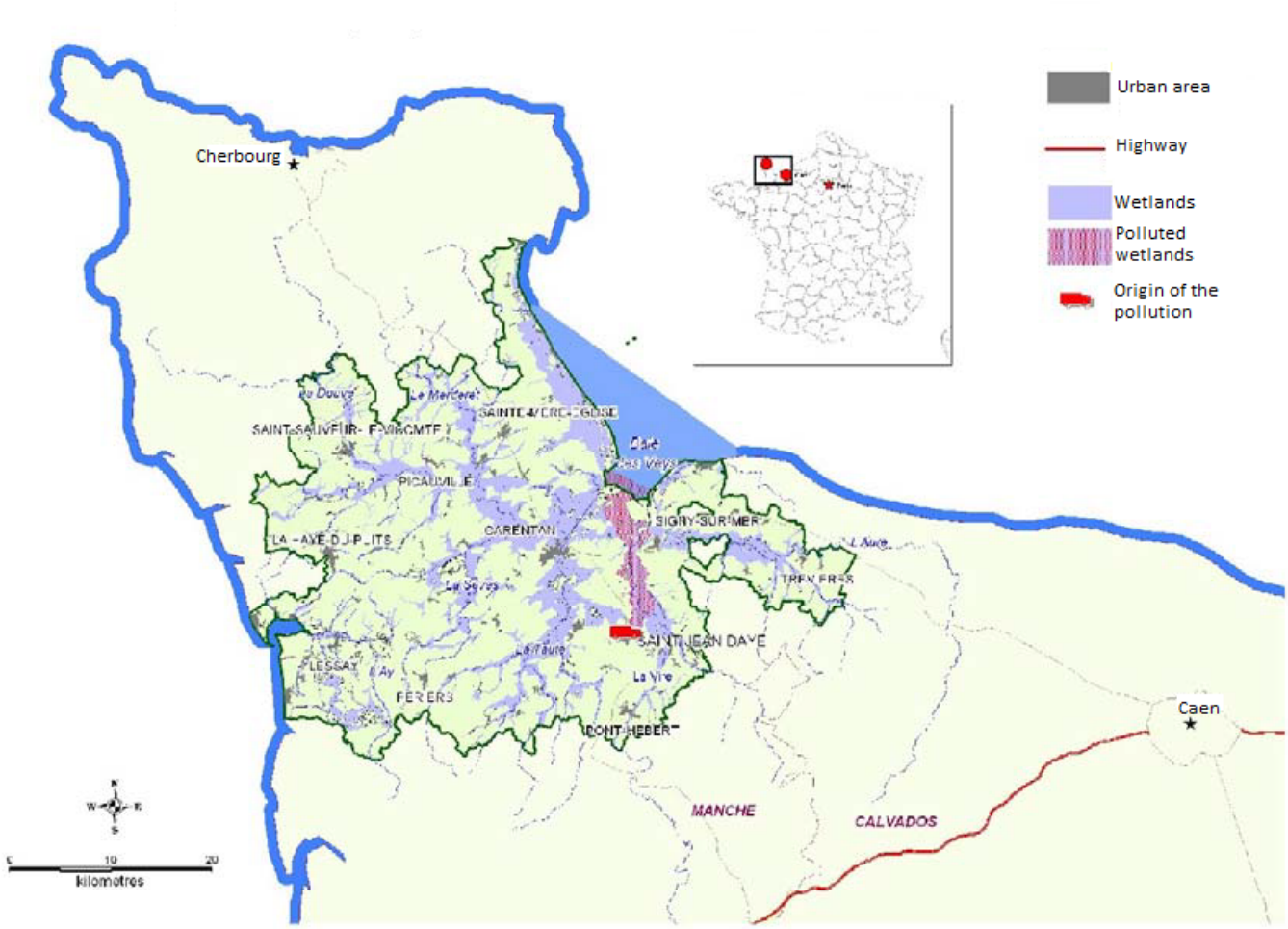

Figure 1.10: Illustration of the scenario of accidental pollution of 30 000 acres at the "marais du Cotentin". Source:[Katossky and Marical, 2011]

Restoration costs are the investment expenditures required to offset any damage done to the environment by any human activity.

As we can observe in this section, valuation of the environment is generally divided into two types of value (which can also be divided in sub-categories): use and non use value. This division seems appropriate in our context and as we will see later, we use it in this thesis. We can also note that the evaluation of the environment is generally expressed as an economical value. The reader will see in the next section that a different was used to deal with our problem.

In this chapter we saw that that the INERIS is interested in a method to evaluate

risks of accidental pollution and we presented some basic idea of risk governance and environmental valuation. We are now about to formulate explicitly the problem that is given to us.

1.4 Definition and characterization of our problem

Now that the context was introduced to the reader, we will give an explicit definition of the problem, highlight the various challenges that are raised by it and present the methodological choice that we made in order to deal with them.

1.4.1 Definition of our problem

So to sum up, we stated earlier in this document that the ministry is looking for a way to include environmental damages into the risk management methodology. Doing this implies to perform a risk assessment. As we just saw, most of the risk assessment are made up of three main ingredients; a list of all the scenarios of accidents, an evaluation of the probability of each scenario that was obtained and an evaluation of the severity of each scenario. The listing of the scenarios is already done is the current operating mode of risk studies as well as the evaluation of their likelihood. The assessment of severity is also made in the current functioning of risk studies in France but from the only perspective of short term damages on human beings. Thus, the main lacking part in an environmental risk assessment is the evaluation of the severity of scenarios.

1.4.2 Dividing resources and biodiversity

The severity of the impact of an accidental pollution on biodiversity and on the resources used by humans are two very different topics that we should first treat separately. Indeed, the interest in avoiding negative impact on resources is motivated by the willingness to be able to use these resources. Therefore, it responds to a short or medium term political and economic motivation. On the other hand, the interest in preserving the biodiversity is an almost intrinsic philosophical, ethic matter driven by the idea that biodiversity is a priceless good that man should preserve for the next generations. Beside that, while studying the impact on the environment or the *value of the environment* dividing resources and biodiversity is a common process as shown in

Subsection 1.3.2. As we will later see, this choice was validated during the interviews with the experts.

Following this reasoning, we think that the environmental consequences of a scenario should be represented through two indices: the *Use Severity Index* that will describe the global expected severity due to the loss of resources and the *Biodiversity Severity Index* that will describe the global severity of all the losses on biodiversity.

1.4.3 Using a monetarist approach for assessing the severity on resources

As we explained earlier in Subsection 1.4.2, the meaning of the use severity index is to represent the harm caused to humans by the loss of goods and services that are provided by the environment. In history, money was the most used tool to value goods and services and one of the three functions that were attributed to it by Aristotle [Aristotle, 50BC] is the function of "unit of account". Besides being appropriate to represent the "severity on resources", monetarist approaches have the advantage of being a well developed scientific field (economical valuation) for non-market goods [List et al., 2006][Drake, 1992][Bateman et al., 2002] and that economic value of market goods are directly observable on the market. Furthermore, the aggregation between several monetary values is generally the sum of these values. Thus it seems appropriate to use a monetarist approach to deal with this part of our problem. The INERIS employs economists that are already working on this topic. Thus, although we advocate for a monetarist approach regarding the use severity index, the calculation of this index will not be treated in this document. In addition to this monetarist approach, a careful consideration should be given to the risk that an accidental pollution deprives man of a resource of which the lack could be a risk. As an example we could imagine an oil spill that would prevent a nuclear plant to use water for the cooling. In case of such a chain reaction, the severity would not be based on the price of the resource, the second accident should be analysed separately with the proper scientific tool and with the competent authorities.

1.4.4 The Biodiversity Severity Index

The goal of this work is to create a *Biodiversity Severity Index* (BSI) for scenarios of accidental pollution. The meaning of this indicator is to answer the following question:

If the scenario S would happen in the current conditions regarding the functioning of the studied industrial plant and its surrounding, how important to society would be its expected consequences on the surrounding environment?

In order to to use this indicator in a risk exposure table similar to the one used in risk management on human consequences, the *biodiversity severity index* should be expressed on a finite discrete scale. We think that a five value levels scale with similar definitions of the categories to the risk exposure table imposed by the *circulaire du 10/05/2010* ("negligible", "minor", "moderate", "major", "extreme") although the description should be adapted to the language relevant to the pollution topic. One of the major concerns of the future users will be to minimize the necessary resources for the application of a risk study. The term "resources" here might refer to either monetary resources, resources in working time, resort to experts etc. The method should then require as little of these resources as possible and use only information that are available to the INERIS. Our main goal is to obtain a methodology that returns an indicator by aggregating the values on a set of criteria to be defined. Our methodology should return an indicator for any scenario that has values on the criteria. We believe that the value of any scenario s_0 should not be influenced by the existence or not of the any scenario s_1 in the set of scenarios to be assessed. We should mention that this particularity is generally considered while facing a sorting problem.

Focusing on the toxic leaks

While looking at the ARIA database [14] we realized that a huge majority of the accidental pollutions that happened in France in the last 40 years were caused by leaks of toxic liquids (the few remaining were caused by burning toxic product). While studying accidental pollution, toxic leak have several specificities: they are characterized by a volume and a toxicity as we will later see in Section 3.1.2 and it is possible to know what environmental targets will be impacted. These characteristics make the evaluation of risk of accidental pollution by a toxic leak quite different to the evaluation of other accidental pollutions. Furthermore, the risk induced by an emission of a toxic gas is already taken into account through its consequences on humans.

Moreover, the main consequences of toxic leaks are generally observed on ground water and surface water, and the study of the impact on water is different from the

[14]The ARIA database (Analysis, Research and Information on Accidents) identify incidents or accidents that have, or could have caused damages to health or public safety, agriculture, nature and the environment. Essentially, these events result from the activity of factories, workshops, warehouses, construction sites, quarries, farms ... classified under the legislation on classified installations.

impact on the ground due to the fact that we cannot use the concentration of liquid in the environment.

Therefore, the BSI will for now only be applied to the risk of toxic leaks and we will focus on its consequences on liquid environments.

1.4.5 How scientific issues and challenges led to precise methodological choices

We will now present some important characteristics of this problem that can be seen as scientific challenges and argue on how these properties influenced us in doing some methodological choices.

1) **The problem's complexity**: Estimating the severity of the possible impact of a scenario of accident is a quite wide issue. It requires to predict as accurately as possible the strength of the impact on the environment, involving several scientific disciplines, and make a judgement about the importance of these damages. For instance, evaluating the destructive potential of a toxic leak (described Subsection in Section 3.1.2) requires the expertise of a toxicologist while evaluating the vulnerability of an environment requires the expertise of an ecology specialist. This complex feature of the problem induces several difficulties while performing this evaluation. First of all, as we will later see there is not one expert that is competent in all the scientific fields that are involved in the process. The solution that will be chosen here is to decompose this problem in several sub-problems through a hierarchy of criteria (as described in Section 2.1.6).

2) **The cognitive load and the limits of the human brain**: As stated by [Miller-George, 1956], the human brain has difficulties to understand more than 7 ± 2 criteria. A decision such as the evaluation of the severity of a toxic leak requires considering the volume of the leak, the toxicity of the product, its persistence, the mobility and the volume of water in which it will be mixed to, the biological importance of the target and its vulnerability. As its name suggests Multi-criteria decision aiding aims at dealing with that issue and the use of multiple sub-problems in a hierarchy of criteria also helps at it.

3) **The public feature of the problem**: Ostanello and Tsoukias [Ostanello and Tsoukiàs, 1993] defined a public decision process by the type of actors that are involved, at least one public actor; and by the topic, which should include at

least one public matter (a matter that may affect one or several social groups). They generally mainly deal with some non-commercial activities (in contrary to private companies) although they might also include commercial activities. These decision processes may involve several institutions and are generally framed by important legal and administrative constraints (these constraints may also bring some stability). The reader may observe that risk management includes all these properties. Indeed, the consequences of an accident could potentially affect any citizen of our country, it is managed by several public actors (the INERIS, the DREALS...) and managed by several laws (in particular in France the *arrêté du 29 Septemble 2005* and the *Circulaire du 10 mai 2010*). This is why it is generally considered as a public issue [McPhee, 2005][Bell et al., 2012]. Thus, this problem cannot be managed only with the help of experts and it resolution should be given to a democratic legitimacy by either associating directly the concerned citizen or by implementing transparent processes that can convince public decision makers, be publicly understood and justified. For that purpose, formal models used in decision aiding theories and methodologies offer several advantages[Bouyssou et al., 2012]

- They contribute to communication between the intervening parties in decision making and evaluation process by providing them a common language
- They are instruments in structuring the problems; the process of developing them forces the intervening parties to make explicit those aspect of "reality" that are important to the decision or evaluation.
- They lend themselves naturally to "what-if" types of questions thereby contributing to the development of robust and increasing the intervening parties's degree of confidence in the decision.

Multi-criteria decision aiding provides tools and methodology that gained there spurs so we can find a result that we can be convinced of and justify it. Then this approach seem particularly adapted to this kind of public issues.

4) **The subjective nature of what is to be measured**: As we stated before, different individual might have different willingness to accept risks [Hillson and Murray-Webster, 2007], evaluate differently the level of a risk according to its likelihood to happen and the severity of its potential consequences or, more simpler, evaluate differently the severity of the its potential consequences. We can then observe that the subjectivity is a part of our problem. Trying to apply a methodology that does not take into account this subjectivity might lead to shortcomings and misunderstanding. Multi-criteria decision aiding is a scientific field that specifically considers subjectivity and this issue is generally dealt with

through the use of preference parameters, in a process of which the decision maker is placed at the center.

5) **Economical valuing, a common but limited approach**: As explained in Section 1.3.2 most of the environmental assessments that are made today are based on a monetary valuation of the environmental goods[Nunes and van den Bergh, 2001][Hufschmidt et al., 1983][Jochem, 2006]. By the way the same kind of reasoning is also widely used while evaluating the value of human lives[15]. The underlying idea while performing such assessments is that any harm that could be done the either human lives or the environment can be compensated by an amount of money. [Roy and Damart, 2002] demonstrated the limitations of this approach. Indeed, following this reasoning leads us to think that any damage that one could cause on either the environment or human lives is acceptable as long as she can pay for it. Then, we could wonder "what is the price that one should pay so that it is "Ok" if humanity or life on planet earth disappears?". On the other side many multi-criteria decision aiding methods are designed to limit compensation. We will later see how their use can be adapted to our context to our context. Furthermore, the three method presented in Section 1.3.2 to evaluate the an impact on an environment suppose that the user can get as an input a prevision of the state of every environmental target after the scenario happened. Which we think may be difficult to model with precision. The restoration cost principle is based on the assumption that any damage on the environment may be cancelled with restoration measures which in practice is not always true. The contingent method in our case would require to b applied on every surrounding environmental target for every scenario which would be a very costly process. Concerning the hedonistic price method its aim is to evaluate an environmental target instead of evaluating the severity of a scenario on an environmental target. Thus, the "pollution part" is not included in this method.

In this chapter, we presented the origins of our risk problem with some basic ideas of risk management. Then we formalized the problem and identified the scientific challenges that are associated to it. These challenges led us to adopt two main methodologies: multi-criteria decision aiding (or rather multi-criteria evaluation as seen in Subsection 2.1.1) and in particular hierarchies of criteria. These concepts are two already well developed scientific fields thanks to the work of a large community of researchers. In order to describe our work and reasoning we must introduce in the next part of this document some of their fundamental ideas.

[15] the price of life was estimated to $9.4 million by [Thomson and Monje, 2015].

Chapter 2

Multi-criteria decision aiding

Contents

2.1 Structuring a multi-criteria decision aiding problem . . . 42

2.1.1 Introduction to decision aiding . . . 42

2.1.2 Actors involved in a decision aiding process . . . 47

2.1.3 Defining the problem formulation: what type of output is expected . . . 48

2.1.4 Choosing an approach . . . 49

2.1.5 Choosing the variables of the problem . . . 50

2.1.6 Value focused thinking and the hierarchies of criteria . . . 53

2.2 Multi-criteria Aggregation procedures . . . 59

2.2.1 Basic notions of criteria aggregation . . . 60

2.2.2 Outranking methods . . . 61

2.2.3 Synthetic criteria and utility . . . 69

2.2.4 Rule based methods . . . 70

2.2.5 SMAA: A stochastic methods to deal with uncertainty or imprecise informations or group decisions . . . 72

2.2.6 Choosing the appropriate Multi-criteria aggregation procedure 73

2.3 Elicitation methods . . . 76

2.3.1 Aggregation approach . . . 77

2.3.2 Disaggregation approach . . . 77

2.3.3 Expression of the preferences for the sorting problem with a disaggregation approach . . . 78

2.3.4 UTA methods: a disaggregation approach algorithm based on utility . . . 79
2.3.5 A mixed integer programming algorithm for disaggregation elicitation with MR-Sort . . . 81
2.3.6 An heuristic algorithm for disaggregation elicitation with MR-Sort . . . 83
2.3.7 An heuristic algorithm for disaggregation elicitation with 2-additive NCS model . . . 86
2.3.8 ORCLASS: A tool to interact with the decision maker . . . 87
2.3.9 DRSA: A disaggregation elicitation algorithm based on rule based principle . . . 89
2.3.10 Logistic and choquistic regression . . . 92

"Nature has placed mankind under the governance of two sovereign masters, pain and pleasure. It is for them alone to point out what we ought to do. By the principle of utility is meant that principle which approves or disapproves of every action whatsoever according to the tendency it appears to have to augment or diminish the happiness of the party whose interest is in question: or, what is the same thing in other words to promote or to oppose that happiness. I say of every action whatsoever, and therefore not only of every action of a private individual, but of every measure of government", Jeremy Bentham, An Introduction to the Principles of Morals and Legislation, 1790.

In the previous chapter, we came to the conclusion that multi-criteria decision aiding and in particular hierarchies of criteria are scientific tools that are well adapted to deal with the problem that was given to us. In this chapter we will introduce the reader to some important notions of these two concepts so that the next chapters can be properly understood. We will first present how a decision problem is built and structured, then we will give some bases of multi-criteria aggregation and finally we will show how elicitation methods may help to find some decision parameter that are adapted to a decision maker.

2.1 Structuring a multi-criteria decision aiding problem

2.1.1 Introduction to decision aiding

To many people, the term decision aiding can be seen as abstract and vague. Aiding what decision? In what way? Indeed, somehow, a physicist engineer that would be employed in a skateboard factory to evaluate the forces that will be applied on skateboards and the resistance of various materials would definitely help this factory to create better skateboards, and thus to take the decision associated to the question "how should we create skateboards?". However, this expertise will generally not be considered as what is called decision aiding.

According to [Tsoukiàs, 2008], "what characterises decision aiding, both from a scientific and a professional point of view, is its approach which I will call both "formal" and "abstract". With the first term I mean the use of formal languages, ones which reduce the ambiguity of human communication. With the second term I mean the use of languages which are independent from a specific domain of discourse. The basic idea is that the use of such an approach implies the adoption of a model of "rationality" a key concept in decision aiding". This definition of decision aiding would indeed exclude the work of the physicist engineer on the fabrication of skateboards explained above. Indeed, the task of this expert would require her to use a language based on a specific domain namely physic.

Adopting a formal and abstract approach has some disadvantages such as its costs (not necessarily economical), being less effective with respect to human communication or imposing a limiting framework on people's intuition and creativity. Nevertheless this approach also has its pros. Indeed using a formal and abstract approach allows every

participant to speak the same language, it is not affected by the biases of human reasoning that are due to education or tradition and finally it may help to avoid the common errors that are due to an informal use of formal methods.

According to [Roy and Bouyssou, 1993] "Decision aiding is the activity of the one that, based on clearly specified models but not necessarily formalized, helps to get some response elements to some problem that a stakeholder faces in a decision process, those elements being supposed to enlighten this stakeholder and to promote a behaviour that should lead to increase consistency between the evolution of the process on the one hand and the objectives and the values supported by the stakeholder on the other hand.[1]"

Here an important notion seems to be the objectives and the values. Indeed one specificity of decision aiding is that the solution will entirely depend on the values of the stakeholder that we are helping [Keeney, 1992]. Here again the work of a physicist engineer working in a skateboard factory would consist in providing the decision makers some information about the materials and the forces that would be applied on a skateboard but the objectives and values would not be taken into consideration.

Situations that require decision aiding

Why can sometime decision aiding be useful? Why would anybody need help to determine what she prefers? The complexity of some decision problems (not in an algorithmic meaning) can derive from several difficulties that can lead to several subcategories of decision aiding [Rolland, 2008].

- The uncertainty on the future events and its consequences on the result of our decision. This problem will obviously be faced while betting on a number playing the roulette or choosing an insurance service. This kind of problem is generally called decision under uncertainty.
- The different and possibly conflicting interests or preferences of several actors to be considered. This problematic occurs for instance while voting for a candidate

[1]Personal translation made by me from the original text in French "L'aide à la décision est l'activité de celui qui, prenant appui sur des modèles clairement explicités mais non nécessairement complètement formalisés, aide à obtenir des éléments de réponses aux questions que se pose un intervenant dans le processus de décision, éléments concourant à éclairer la décision et normalement à prescrire, ou simplement à favoriser un comportement de nature à accroître la cohérence entre l'évolution du processus d'une part, les objectifs et le système de valeurs au service desquels cet intervenant se trouve placé d'autre part."

or while bargaining to buy a product. It can lead either to social choice problems or to game theory depending on the specific problem.

- The presence of multiple criteria to be taken into account. This feature will lead to multi-criteria decision aiding, it will be at the core of this all thesis and will be explain in detail later.
- In public decision contexts there is generally a need for formal method. Decision aiding provides a large panel of methods that support decisions so that these decisions can be defended. It is particularly important when the stakes are high such as the location of a railway for instance.

Decision aiding can also be useful when there is a need for justification of the decision to be taken or if we want the decision to be made automatically (if we want the same kind of decision to be made a large number of times or if we want the decision to be fair and neutral). One may also use decision aiding to deal with complex problems that involve several scientific fields and require different experts to manage them.

Decision aiding is a long process that costs efforts and requires the participation of several participants. While deciding whether or not using decision aiding methodology, one should wonder if the given decision is to be taken soon and if the stakes are high enough to require these efforts and resources.

Decision aiding and subjectivity

[Belton and Stewart, 2002] named as Myth number 1 the following sentence "Multi-criteria decision aiding will give *the right* answer". Indeed, subjectivity is a common feature to many decision problems and decision disciplines [Yevseyeva, 2007][Keeney, 1992] and in many cases there is no "right answer". Subjectivity here means that, unlike what is done in most the scientific fields, we are not checking if an assumption is true or false, we judge if a solution or an object is good or bad, right or wrong, acceptable or not, from a greedy, a moral point of view or both mixed. Decision maker is a key topic in decision aiding. Indeed, it is about what people want, hope, how much they care for some issues. It is actually a pretty rare feature in science. If you ask two ornithologists what bird is singing and they do not give you the same answer, then at least one of them is wrong. Of course the choice of the expert that is chosen to help us has an influence on the result because the more qualified the expert is, the more likely it is that her answer will be correct, but these experts do not express their preferences. If two persons disagree on which house they would prefer to live in, that does not signify that one of them is wrong and the other one is right.

What would be a good decision aiding process?

To this question several answers are generally given [Roy and Bouyssou, 1993]:

- A good decision aiding process is a process that is based on a rational methodology such that its result can be vindicated with rational arguments.
- A good decision aiding process is a process that leads to a good decision, meaning a decision that will not later be regretted by the decision makers.

Thus, we can say that a good decision aiding process is a process that allow the stakeholders to overcome the main difficulties linked to the problem that they are facing, that is based on rigorous scientific bases and rational reasoning and whose conclusions suit to the decision maker's values. [Landry et al., 1983] suggest that analyst should, in order to validate their decision process:

- Evaluate the degree of relevance of the assumptions and theories underlying the conceptual model of the problem situation for the intended users and use of the model (Conceptual validation).
- Study the capacity of the formal model to describe correctly and accurately the problem situation as defined in the conceptual model (logical validation). It implies verifying whether any pertinent variable or relationship has been omitted from the formal model.
- Take into account the quality and efficiency of the solution mechanism, either algorithmic, heuristic, or experimental (experimental validation).

Multi-criteria decision aiding

Multi-criteria decision aiding (MCDA) is the branch of decision aiding whose aim is to analyse the preferences of a given decision maker in order to assess the overall attractiveness of several objects or to compare them with each other taking into account every criteria on which these objects should be judged. The understanding of the way evaluations on criteria influences on the overall assessment and the understanding of compensation between criteria are crucial issues in MCDA.

The all process of MCDA is structured by interactions between the decision maker (or the expert) and the analyst. As stated by [Dodgson et al., 2009] "The main assumption

embodied in decision theory is that decision makers wish to be coherent in taking decisions. That is, decision makers would not deliberately set out to take decisions that contradict each other.“

We will thereafter divide this process in three parts:

- Formally describing and structuring the problem.
- Choosing the appropriate Multi-Criteria Aggregation Procedure (defined in Section 2.2).
- Eliciting the preferences of the decision maker(s) in order too find the appropriate preference parameters for the chosen MCAP (defined in Section 2.3).

We will later describe these steps in more details in the following sections.

Decision and assessment

In many cases of multi-criteria sorting, the terms “decision”, “alternative” or “consequences” might not be suitable to the problem that is being faced. According to [Rousval, 2005] in decision aiding “Evaluating an action is evaluating the changes caused by the action that are part of the decision maker’s concerns. Note that in most cases these changes are in fact estimated because the consequences of actions are not necessarily known beforehand. These are often projections.”[2] Indeed, in the literature while using the word alternative we refer to the action that we can do, of which we study the attractiveness of the consequences. There are situations in which we are not studying the attractiveness of an alternative in the sense of a possible action to do. If we are assessing the global performance of a student, evaluating the Human Development Index of a country (which can be seen as multi-criteria decision aiding) what we are assessing are not actions that we could do. One could argue that an evaluation is willing to impact a decision. I do not completely agree with that, one could be interested in the Human Development Index only to have a better understanding of our world. Anyway this affirmation does not mean that what is being evaluated is an alternative. In some circumstances, what is being evaluated is an object about which a decision is to be taken and that this decision depends on its evaluation (a student could

[2]Personal translation made by the author from the original text in French “En ce sens, évaluer une action, c’est évaluer les changements d’états que provoque l’action et faisant partie des préoccupations du décideur. Remarquons que dans la plupart des cas, ces variations sont en fait estimées car les conséquences des actions ne sont pas forcément connues préalablement. Il s’agit souvent de prévisions.”

be evaluated to take the decision to make her pass or not). In our context, evaluating the severity of a scenario of accidental pollution cannot be directly associated to "taking a decision". Thus, in evaluation contexts such as the topic of this thesis, we will more often use the term "objects" rather than "alternatives" and we will prefer talking about "criteria" rather than "objectives". In order to have a unique notation and avoid confusion we will thereafter use the words "object" and "criterion" in some contexts to which the words "alternative", "action" or "objective" could also be appropriate. However, most of the remarks that were made previously about decision aiding may as well be formulated for evaluation problems.

2.1.2 Actors involved in a decision aiding process

While taking a decision either personal or collective, different types of actors can be involved. As seen in Subsection 1.2.2 several approaches exist according to the scientific field that is considered. For instance in the CATWOE analysis described by its main contributor Peter Checkland as a simple checklist that can be used to stimulate thinking about problems and solutions [Checkland and Scholes, 1990], three types of actors are identified: The customer that will be subject to the consequences of the decisions, the owner that will decide and the actors that will act to apply the decision of the owner. In decision aiding, two actors are almost always involved: the decision maker(s), the analyst. In some situations expert(s) may also be involved.

- **Decision maker(s)**: While talking about decision maker in MCDA we refer to the actor whose preferences we try to probe to build the decision model or the assessment model. According to [Bouyssou et al., 2012] the decision maker is "the actor in the decision process that the implemented model tries to enlighten".

  There can be one or several decision makers. There are two reasons why there can be several decision makers. First, it can be that different people are interested in the consequences of a decision to be taken, that all of them have a power on this decision and that they do not have the same objectives or values. In this case we might also be facing a social choice problem. We would try to get a solution that would fit as well as possible to every decision maker.

  But it could also that each decision maker is interested or competent on only one part of the problem.

  According to Ralph Keeney in his book *Value Focused Thinking* [Keeney, 1992] "the objectives for a decision situation should come from individuals interested in and knowledgeable about that situation". That person is generally called the

decision maker in decision aiding and fixing the objectives/criteria is one of her tasks.

The model that is built during the process should help the decision maker/s to understand and formalize her own value system and reproduce her judgement as well as possible.

- **The analyst**: According to [Roy and Bouyssou, 1993] (page 22), the analyst is "the person that is in charge of decision aiding. Implementing models within the decision process she contributes to guide it and transform it"[3]. It must be noticed that the decision process should be independent from the preferences or the value system of the analyst (neutrality of the analyst). Indeed, according to [Mousseau, 2003] "In a decision context, the decision maker is the only actor that should be allowed to specify a preferential information"[4]. The analyst's value system should probably even not be known by the decision maker in order to avoid influencing him. Nevertheless, the analyst should not be neutral regarding the methodology to be applied during the decision process, some of them being objectively more suitable to some decision contexts.

- **Expert(s)**: An expert is an actor that helps the decision process with her knowledge and her experience. In a decision process an expert in one field can often be useful when it comes to forecast the causal relationship between a cause (possibly an action of the decision maker) and its consequences. For example, if you are the manager of a super market and you try to evaluate the gain obtained employing one more person at the check out you could possibly ask an expert this information. Her answer could be based on precise models and calculations or she could answer instinctively based on her experience. This information will more often not be subjective in the sense that, if two experts give you two different answers to this question, then at least one of them is wrong.

2.1.3 Defining the problem formulation: what type of output is expected

While talking about multi-criteria decision aiding, three types of problem are generally considered regarding the expected form of the output [Roy and Bouyssou, 1993] (page

[3]Personal translation made by me from the original text in French "L'homme d'étude est celui qui prend en charge l'aide à la décision. Mettant en œuvre des modèles dans le cadre d'un processus de décision il contribue à l'orienter et à le transformer"

[4]Personal translation made by me from the original text in French "Dans le cadre d'une étude d'aide à la décision, seul le décideur est à même de spécifier une information préférentielle"

31)

1) **Choice problems**: The choice problem is probably the most common in decision aiding. In most of the cases the objects to be chosen are possible action to do in the problem's context and we are looking the most attractive one (or the set of the best alternatives) without being interested in how good are the others.

2) **Ranking problems**: The ranking problem consists in establishing a pre-order among the objects. The output of this problem allows the decision maker to compare every pair of objects. This information is more complete than only knowing the best object and in some circumstances (for example if some alternative could finally not be possible) which may be useful. However, this output only gives a relative judgement on the objects and does not enable us to know how attractive an object is.

3) **Sorting problems**: The sorting problem consists in assigning individually every object in a predefined ordered category $c \in C$ (the order in denoted by $\leq$) depending how globally attractive they are. The objects are assigned to categories regardless to the set of objects to be sorted. This output finally gives us an individual evaluation on how globally attractive each object is. Nevertheless inside each category we do not have any information about which is the best object and which is the worst one and the method does not provide a pairwise comparison between every pair of objects assigned to the same category. This thesis is about a multi-criteria sorting problem. Indeed as each toxic leak will have to be assigned individually to a category representing the severity level of its consequences on the environment. Thus, from now on we will focus on issues relative to the sorting problem.

2.1.4 Choosing an approach

According to [Dias and Tsoukiàs, 2004], there are 4 main ways to consider and practice decision making:

- **The normative approach**: "Normative approaches derive rationality models from norms established a priori. Such norms are postulated as necessary for rational behaviour. Deviations from these norms reflect mistakes or shortcomings of the DM who should be aided in learning to decide in a rational way". A normative approach consists in logically deduced conclusions on decisions from a

priori norms. These norms are generally related to the specific domain that we are studying.

- **The descriptive approach**: A descriptive approach in decision analysis consists in reasoning based on the observation of how decision makers make decisions. In particular, these approaches may link the way decisions are made with the quality of the outcomes. The analyst generally believes that the decision maker prefers the decision that she made to any other decision that she could have considered.

- **The prescriptive approach**: While using a prescriptive approach we consider that the preferences of the decision maker exist and that the role of the analyst is to help him to discover, understand and formalize them. Therefore, the models do not intend to be general, but only to be suitable for the contingent DM in a particular context. Indeed the DM can be in difficulty trying to reply to the analyst's questions and/or unable to provide a complete description of the problem situation and her values. Nevertheless, a prescriptive approach aims to be able to provide an answer fitting at the best the DM's information here and now.

- **The constructive approach**: The constructive approach is mainly considered at the LAMSADE, in particular by Alexis Tsoukiàs [Dias and Tsoukiàs, 2004]. The main idea of the constructive approach is that there is no pre-existing preference system in the DM's mind and that it must be constructed by the interaction between the analyst and the decision maker. Structuring and formulating a problem becomes as important as trying to "solve" it in such an approach.

In our context, constructive approach seems appropriate to us. Indeed, the problem that we are facing is complex in the sense that it involves several scientific domains, its structure was not defined initially and its construction required repeated interactions with various experts.

2.1.5 Choosing the variables of the problem

A multi-criteria decision problem is structured around mainly two types of data: the objects and the criteria. At the beginning of the decision process, the actors of the decision process must decide which objects and criteria should be considered later in the process. In multi-criteria sorting problems, one must also determine the set of categories in which the objects will be sorted.

Choosing the objects (or not yet?)

Here we consider the object whose attractiveness we are studying. This term may cover either an object that we try to evaluate or a possible action which consequence must be appreciated. We will thereafter use the notation $A = \{a, b, ...\}$ as the set of the m objects that we plan to assess.

According to [Keeney, 1992] it is common to characterize a decision problem by the alternatives available. Keeney argues that conversely we should first focus on what matters to the decision maker in order to understand her value system. The alternative could be chosen at any moment and some could be added even after the elicitation was made.

Criteria and families of criteria

As its name suggests, criteria are a central issue in multi-criteria decision aiding. A criterion is an ordered information concerning an object on which we base our judgement. It will represent the performance of this object from a specific point of view. According to [Roy and Bouyssou, 1993] (page 46) "Essentially, criteria aim to summarize, with the help of a function, the evaluation on an object on various dimensions that are related to a "same signification axis", the latter being the operational translation for a *point of view*"[5]. The set of the criteria that are considered as relevant will here be refereed to as a family of criteria. According to [Roy and Bouyssou, 1993] (page 79) there exist three properties that are considered as essential to any family of criteria. First of all, a family of criteria will be considered exhaustive if the decision maker would be indifferent between any two objects that have the same evaluations on all the criteria. A family of criteria is said non-redundant if the abduction of any criterion would make it not exhaustive. Finally, we say that a family of criteria meets the axiom of cohesion if improving an object on any criterion cannot make it become less attractive globally. A family of criteria is said coherent if it is exhaustive, non-redundant and if it meets the cohesion axiom. In a multi-criteria decision aiding process enumerating, studying and organizing the family of criteria are crucial issues. We will thereafter consider as a notation $N = \{1, 2, ..., n\}$ as the family of criteria that we are working with.

[5]Personal translation made by me from the original test in French "Pour l'essentiel un critère vise à résumer à l'aide d'une fonction les évaluations d'une action sur diverses dimensions pouvant se rattacher à un même *axe de signification*, ce dernier étant la traduction opérationelle d'un *point de vue* "

Scales on the criteria

In order to facilitate the comparison between the objects, the value of an object a on the criterion i have to be expressed in a same way for all the objects. In this document we will consider the scale only as a way to represent the different values that object can have on the criteria. We will call v_i the scale on the criterion i, i.e., the set of every evaluation that can be made of the criterion i here named value levels. In multi-criteria disciplines, evaluations on criteria are ordered, i.e., there exists an order $\leq$ on the scale v_i. The scales that are used during the process are built together by the analyst and the decision maker according to the way they want to use them. According to the case, the chosen scales can be either continuous or discrete, bounded or unbounded.

Psychologist Stanley Smith Stevens identified 4 types of scales of measure: nominal, ordinal, interval, and ratio.[Stevens, 1946][Kirch, 2008]

1) **Nominal scale**: Nominal scales refers to property more than quantity. A nominal level of measurement is simply a matter of distinguishing by name, for instance, 1 = male, 2 = female. Even though we might use the numbers 1 and 2, they do not denote quantity.

2) **Ordinal scale**: An ordinal scale indicates a direction, in addition to providing nominal information. Low/Medium/High; or Faster/Slower are examples of ordinal levels of measurement. It allows for rank order (1st, 2nd, 3rd, etc.) by which data can be sorted, but still does not allow for relative degree of difference between them. Then a score of 5 will be preferred to a score of 3 (if the variable is to be maximized) but this is all we know about it. Any transformation of the values of the objects through a strictly increasing function is acceptable in the sense that it should not change the way the preferential information will be understood. Ordinal scales may be considered as a particular case of nominal scales.

3) **Interval scale**: An interval scale contains an ordinal information but in addition it gives some information about the degree of preference between the evaluations which is relative to their mathematical difference. In other words, if v is an interval scale then, the degree of preference between x_0 and y_0 on v will be considered as equivalent to the degree of preference between x_1 and y_1 if and only if $x_0 - y_0 = x_1 - y_1$. Any transformation of the values of the objects through the addition of a real number is acceptable in the sense that it should not change the way the preferential information will be understood. Interval scales may be considered as a particular case of ordinal scales.

4) **Ratio scale**: Just as interval scales, ratio scales contain an ordinal information and some information about the degree of preference between the evaluations. With ratio scales it is relative to their mathematical ratio. In other words, if f is an interval scale then, the degree of preference of value between x_0 and y_0 will be considered as equivalent to the degree of preference between x_1 and y_1 if and only if $\frac{x_0}{y_0} = \frac{x_1}{y_1}$. Any transformation of the values of the objects through the multiplication with a strictly positive real number is acceptable in the sense that it should not change the way the preferential information will be understood. By the way, ratio scales may be considered as a particular case of ordinal scales.

2.1.6 Value focused thinking and the hierarchies of criteria

Definition of a hierarchy of criteria

What we refer to as a hierarchy of criteria is a tree of composite indices or criteria. At the lower level, these indices are information that are given as an input, we will call these criteria "Input criteria". Every other criteria are obtained by aggregating its sub-criteria. In every sub-problem we will keep the same set of objects but the MCAP (defined in Section 2.2) might be different on each of them as well as the decision maker or expert. Actually, on each of these nodes we are facing a local multi-criteria decision aiding problem that we will call a sub-problem. Since the late 90's, hierarchies of criteria have been used in many different kinds of context such as assessing the passenger's comfort in the train [Mammeri, 2013], selecting wall structures regarding to the environmental impact [O.A.B., 2004] or finding strategic implications of mobile technology [Sheng et al., 2005].

Several good properties for a hierarchy of criteria

As we do while looking for a good object, when trying to get the appropriate hierarchy of criteria, we must first have a thought about what would be the desirable features for hierarchies of criteria. Here is a non exhaustive list of properties that we would like the hierarchy of criteria to meet as far as possible:

- The criteria at the lowest level of the hierarchy should be data accessible to the decision maker.
- The highest criteria in the hierarchy (those directly located under the top node representing *global attractiveness criterion*) should be criteria that intrinsically

matter to the decision maker and could not be considered as a specification of a wider criterion.

- Every criterion must be understood in the same way by every stakeholders as well as the scales on which these criteria are expressed.
- As far as possible, we would like to limit the size of each sub-problem. Indeed, as stated by [Miller-George, 1956] the human brain has difficulties to manage to many criteria simultaneously.

We add to this list two properties that seem positive:

- Every maximal set of criteria (by inclusion) such that no criterion in the set is a sub-criterion of another criterion in the set (maximal sub-family of criteria) should be a coherent family of criteria for the decision problem that we are facing. An illustration of a maximal sub-family of criteria is provided on figure 2.1.
- As far as possible we would like, in any maximal sub-family of criteria, every criteria to be independent as we will later define in Subsection 2.2.6 to every other criteria in the maximal sub-family of criteria. Indeed, we need this feature to be able to manage each sub-problem independently.

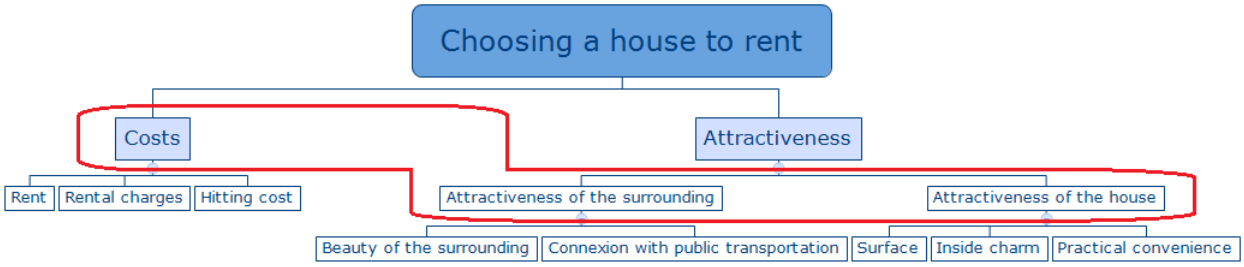

Figure 2.1: Illustration of a maximal sub-family of criteria on a hierarchy of criteria modelling the problem of choosing a house to rent.

2.1.6.1 Value Focused Thinking

One of the main and most notorious advocate for the use of the hierarchy of criteria is Ralph Keeney, with a particular contribution in his book untitled *Value Focused Thinking* [Keeney, 1992]. This book explains a wide methodology that we will not

entirely describe. We mainly based the creation of our family of criteria and the structuring of our hierarchy of criteria on *Value Focused Thinking*. In *Value Focused Thinking* [Keeney, 1992] Keeney argues that, while facing a problem, people generally first focus on finding all the alternatives (here called objects) and try afterwards to find the best one among them whereas he suggests that the first and most important thing in a decision process is to focus on the values of the decision maker. We should first understand what matters to the decision maker, why it does and how much it does. In order to do so, Keeney gives advice to obtain a first appropriate family of criteria that may evolve later in the process. Ralph Keeney then proposes to build the *fundamental objective network* and eventually the *mean-end objective network*. They should help the user to understand better both the values of the decision maker and causal relationship that are included in the problem. The reason why the decision maker is interested in improving the performance on a *mean-end objective* is that it contributes to improve the performance on a higher objective. Conversely *fundamental objectives* intrinsically matter to the decision maker. As an example we may mention the illustration of both the mean-ends objective network and the fundamental objective network proposed by Keeney to deal with the concern of the concentration of carbon dioxide in the air (see Figures 2.2 and 2.3).

Finding a first family of criteria

As was mentioned earlier, we expect the family of criteria to be coherent, i.e., to be exhaustive, non redundant and if it meets the cohesion axiom. In value focused thinking [Keeney, 1992] (chapter 3) a large number of advises is proposed so as to help both the analyst and the decision maker to find that list. They could wonder "what would be a bad object (or item) and what would be a good one?", "can you give us two objects such as one of them is clearly better than the other one?". From the answer that will be given by the decision maker, we could try to figure out "why is this object good and why this one is not?". That reasoning should probably lead us to highlight some criteria that we would add to the list. To these questionings the exhaustiveness question can also be added; "Would it be possible that we have two objects a and b with the same evaluations on every criteria but we still prefer a to b?". If the answer to this question is "yes", then we should try to understand why and we might add an other criterion. We will later see in this subsection that this family might be improved later in the structuring process.

Causal factor

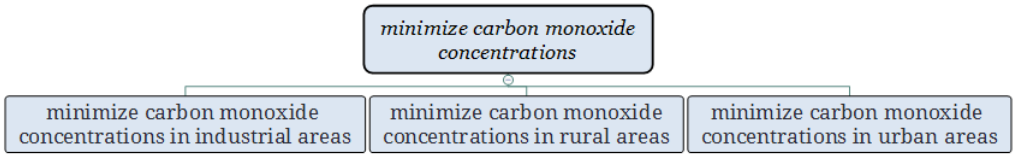

Figure 2.2: Illustration of a fundamental objective network. Source: page 88 [Keeney, 1992]

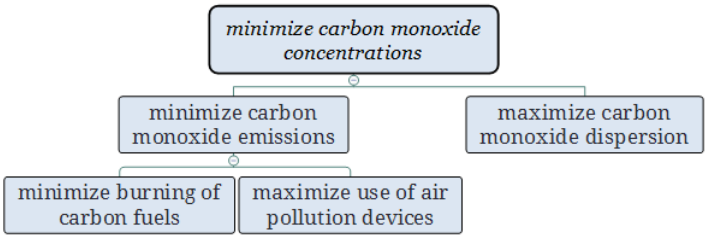

Figure 2.3: Illustration of a Mean-ends objective network. Source: page 88 [Keeney, 1992]

According to Keeney [Keeney, 1992] (page 78), the aim of *causal factor* is to predict the consequences of the happening of an alternative (or object) from an objective point of view. The causal factor is something that we could check after the decision has been taken. Causal factors will be built in the *means-ends objective* network. For example if we are interested in the influence of the rise of the number of car in circulation on the number of fatalities on the roads (illustrated in Figure 2.4) this is an objective fact that can be difficult to assess before anything happen. Two experts could make different predictions on it, but if this is the case, at the end, we can check that at least one of them was wrong. In many cases the MCAPs (defined in Section 2.2) might not come from MCDA but from a different field, closer to the specific topic. For instance in order to estimate the expected concentration of released liquid from the volume of this liquid and the water flow we should better use some methods from fluid mechanics. Due to its nature causal factor is mainly based on mean-end criteria. Thus, this kind of influence will mainly be observed at the bottom of the hierarchy of criteria.

Specifications

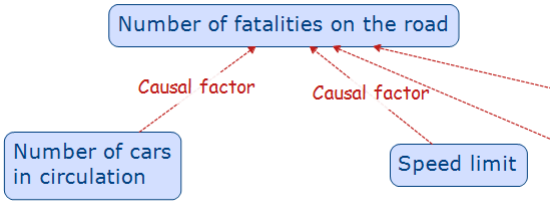

Figure 2.4: Illustration of the causal factor of the number of cars in circulation on the number of fatalities.

Keeney [Keeney, 1992] (page 78) refers to *specification* as the fact of saying that a sub-criterion is one specific part of a higher level criterion whose importance depends on the decision maker's value system. Specification will mainly be built in the fundamental objective network. For instance while choosing a house, the quality of the surrounding transport system may be considered as a specification of the location's attractiveness (illustrated in Figure 2.5). When a criterion is related to its sub-criteria through a specification, its evaluation can be seen as subjective. Indeed, we can easily assume that we would all make the same choice between two alternatives a and b if one of them would be better than the other on every criterion. However if neither a nor b is better than the other on every criterion it is very likely that two different decision makers can make different judgements on this comparison and none of them would be "wrong". In literature while talking about MCDA we more often think about this kind of influence.

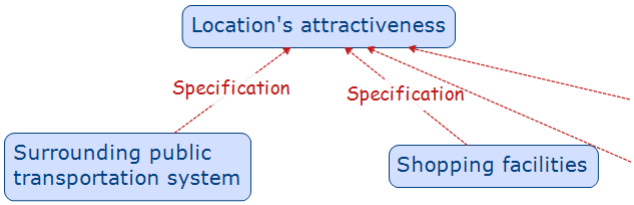

Figure 2.5: Illustration of influence of the surrounding transport system on the location's attractiveness through specification

Facts/values, means/fundamental

In Value Focused Thinking [Keeney, 1992], Keeney separates the criteria that should be taken into account between facts and values. On the one hand facts are information

that can be objectively measured and that are independent to the values the one that measures it. On the other hand, values are subjective evaluation made by the one that measures it that depends on her preference system as described in Section 2.3. Returning to the example given previously the number of fatalities on the roads are a fact while the attractiveness of a location is a value. It is to be noticed that, in the hierarchy of criteria, facts can only be subject to causal factor from facts while values can be influenced by either facts and values. Therefore, at the top of the hierarchy it is more likely to find values than facts. A distinction is also made between the objectives that intrinsically matter (the fundamental objectives) and those that matter because they will have an influence on an other objective (mean objectives).

Structuring the family of criteria into a hierarchy of criteria

Depending on the chosen approach to structure the problem, different hierarchies could arise from studied problem which would lead to different perceptions of the problem and different conclusions [Brownlow and Watson, 1987]. The *Value Focused Thinking* approach provides several recommendations to build the hierarchy of criteria. Once we have obtained a first family of criteria, we apply a bottom-up top-down methodology.

The bottom-up top-down methodology consists in iterating the following process for each criterion i. The bottom-up part of the process consists in finding the criterion that is influenced by i whereas the top-down part consists in finding the sub-criteria of i.

For each criterion i, the decision maker get asked why she is interested in the value of an alternative on this criterion. Three answers can come out from this question.

- First, the decision maker can answer that this criterion matters because it has a causal influence on an other criterion i' already named in the family of criteria. In this case, we declare that i is a sub-criterion of i' in the mean-ends objective network. Then we wonder "What could be the other criteria that would have a causal factor on i'?". This can bring us to include some other criteria to the mean-ends objective network.
- The decision maker can also answer that this objective matters because it has a causal influence on an other objective i', which has not been named already. In that case, we add i' to the mean-ends objective, we say that i is a sub-criterion of i' on the mean-ends objective network and we try to check whether or not i' has other sub-criteria in the mean-ends objective network.

- The third answer that can be given is that this objective matters... and that's all! This means that i is an end objective of the fundamental objective network. In that case we ask the decision maker an other question: "Do you think that i could be a specific part of a more general issue?" If the answer is "yes we could say that it is a specific part of the criterion i' " then we consider that i is a specification of i' (if has not been named in the family of criteria yet, we include it). In that case also, we have to find out what could be the other specific parts of the criterion i' and that could lead us add some other criteria to the fundamental objective network. If the answer is "no", we consider that the only wider issue is the problem that we are facing. Then we consider that i is a specification on the *global attractiveness criterion*.

At the end, these two networks may be connected together in a synthetic network. In our case, for reasons explained in Subsection 3.3.1 we could not exactly follow this procedure and instead we directly created a synthetic hierarchy of criteria.

2.1.6.2 Choosing the decision makers and experts for each sub-problem

As its name suggests, in most of the cases the MCDA's aim is to help a decision maker to take a decision. In these cases the choice of the decision maker whose preferences we must elicitate is immediate (the one that is facing the decision problem). In our case for example, we will see that we are mainly facing two kinds of situation in which we have to aggregate causal factors. In order to aggregate this data to find this factual information, we are looking for expertise. The legitimacy of the expert comes from her knowledge in the given purview. This is the case when we try to aggregate together the *concentration* of liquid in an environmental target the *toxicity* of the liquid and the *mobility of the water* to get the destructive potential (see Section 3.4).

2.2 Multi-criteria Aggregation procedures

Once the problem is explicitly stated we are on the right track to aggregate the criteria. But the aggregation is also a sensitive task. Several approaches can be used and it requires time and cognitive effort from both the decision maker and the analyst. As we are about to see, there are several different existing methods and we will have to find the best adapted one in our context. While talking about methods here we generally refer to mainly two things: Multiple Criteria Aggregation Procedures and elicitation methods.

According to [Tsoukiàs, 2008] an aggregation operator or Multiple Criteria Aggregation Procedures (MCAP) is an operator enabling to obtain synthetic information about the elements of A or of $A \times A$. In MCDA there are several ways to get to those synthetic information. For sure the choice of the Multi-criteria Aggregation Procedure widely depends on how the problem was structured. After we give the reader some basic notions of aggregation of criteria we will introduce the three main families of Multi-criteria Aggregation Procedure, explain how to choose the one that suits the most to the studied context and finally, we will explain how to find the preference parameters that correspond to our decision maker. Our problem being a sorting problem, thereafter we will particularly focus our attention on MCAPs dealing with the sorting problem rather than on MCAPs dealing with the choice problem or the ranking problem.

2.2.1 Basic notions of criteria aggregation

Dominance relation and monotonicity principle

In multi-criteria decision aiding, we consider that an object a weakly dominates an object b if a is at least as good as the object b on every criteria. We say that an object a strictly dominates an object b if a is at least as good as the object b on every criteria and if a is strictly better than b on at least one criterion. It is to be noticed that each object weakly dominates itself. Thereafter, when simply talking about domination we refer to weak domination. The monotonicity principle, which is widely accepted in multi-criteria decision aiding, states that if an object a weakly dominates an object b then b should not be preferred to a and in multi-criteria sorting b should not be classified in a higher category than a. In the following, we will denote by aDb the fact the object a weakly dominates the object b.

Compensation between criteria

According to [Bouyssou, 1986] "Intuitively, compensation refers to the existence of tradeoffs, i.e. the possibility of offsetting a 'disadvantage' on some attribute by a sufficiently large advantage on another attribute whereas smaller advantages would not do the same". The level of compensation between criteria is a property that may characterize either preference structure of decision makers or Multi-Criteria Aggregation Procedure. Indeed, in decision aiding there are several decision methods available that are considered either fully compensatory such as monetarist approaches less or non compensatory like ELECTRE methods. Obviously, the analyst should propose to the decision maker methods that are adapted to the decision maker's preference structure

from this point of view. We will then later speak about limiting compensation while making some judgement on some criteria impossible to compensate by any performance on the other criteria. For instance we will later see that according to the experts a toxic leak, evaluated on three criteria; the destructive potential, the value of the environment and the vulnerability; cannot be considered as "very severe" on a given target if the *value of the environment* on this target is considered as "very low", regardless of the *destructive potential* of the leak on this target and regardless of the *vulnerability* of the target (see subsection 3.5.5).

2.2.2 Outranking methods

We generally call outranking methods, methods in which binary relations between the objects are established on each criterion and then binary relations between the objects are established globally. Obviously these binary relations can be used to obtain a ranking or a choice but they can also be used in sorting problem using some profile used as limits. Among the outranking methods, we could mention PROMETHEE methods [Brans and Vincke, 1985] but the probably most famous are the ELECTRE methods [Roy, 1978]. The method named ELECTRE TRI is an ELECTRE method that deals with the sorting problem.

In all the ELECTRE methods, the outranking philosophy is quite similar, then from this outranking relation, according to the ELECTRE method that has been chosen different conclusion might arise. The outranking relation aSb in ELECTRE has the meaning "a is at least as good as b". The main idea of ELECTRE's outranking principle is that an object a outranks an object b if a weighted majority of criteria are consistent with the fact that a is at least as good as b (concordance) and if there is no criterion that is strongly consistent with the opposite assumption (non-discordance).

These methods offer three advantages:

- They allow the user to control the compensation between criteria through the use of the veto threshold later presented in subsection 2.2.6.
- They allow to deal with heterogeneous scales on the criteria (ordinal, interval or ratio).
- They are a pretty intuitive method that may suit to the way decision maker compare objects, especially by the principle of concordance and non discordance.

However, the fact that outranking relation is generally not complete through incomparability (there are some pairs of objects $\{a, b\}$ for which a does not outrank b and b does not outrank a neither) may either be seen as a weak point or as a good match with the human values that may express indifference or incapacity to compare. As well, the outranking relation may not be transitive which does not fit with the intuition that preference should be transitive. Given that in this thesis, the sorting problem is a central issue, we will thereafter mainly focus on ELECTRE TRI method.

Outranking relation in ELECTRE

According to [Roy and Bouyssou, 1993] "we say that an object a outranks an object b, also denoted by aSb if, given what is known about the preferences of the decision maker, the quality of assessments and the nature of the problem, there is enough evidence for admit that a is at least as good as b and there are no significant arguments to admit the contrary (vetoes)[6]". The outranking relation in ELECTRE method is based on two concepts:

- Concordance principle: We say that the object a outranks the object b denoted by aSb if a sufficient number of criteria (weighted by the criteria's weight) is in favour of the fact that a is at last as good as the object b.
- The non discordance principle that states that there is no criterion for which the preference for b over a is so strong that it avoids the outranking aSb.

Thus, formally the parameters used in ELECTRE TRI are the following:

- For every criterion i a preference threshold p_i. If an object a has a better evaluation than b with a difference higher than p_i then it is considered that there is a strong preference of a over b from the point of view of the criterion i.
- For every criterion i an indifference threshold q_i. If an object a has a better evaluation that b with a difference lower than q_i or if the evaluation of a is lower than b on the criterion i, then it is considered that there is no preference of a over b from the point of view the criterion i. If the difference between two objects on

[6] Personal translation made by the author for the original text in French "On dit qu'une alternative a surclasse une alternative b et on note aSb si, étant donné ce que l'on sait des préférences du décideur, de la qualité des évaluations et de la nature du problème, il y a suffisamment d'arguments pour admettre que a est au moins aussi bonne que b et qu'il n'y a pas d'arguments importants prétendant le contraire (vétos)."

the criterion i is between the indifference threshold and the preference threshold we speak about weak preference on the criterion i .

- For every criterion i a veto threshold v_i. Veto thresholds express the power attributed to a given criterion to be against the assertion "a outranks b".
- For every criterion i a voting power w_i also called its weight. The sum of the weights of all the criteria is equal to 1.
- A concordance level s which represents the limit for the global concordance indicator (presented later) above which we say that an object may outrank an other object (if there is no veto).

In order to establish a global outranking relation, a binary relation is first established from the point of view of each criterion i through a concordance indicator $c_i(a, b)$. This indicator will represent the extent to which the sentence "The object a is better or as good as the object b from the point of view of the criterion i" is true, 1 being "in complete agreement with that sentence", 0 meaning "completely disagree with it". This score of 0 to 1 is based on a fuzzy logic idea that the transition between agreement and disagreement to this sentence is not sudden. In practice the concordance indicator is defined as follows (in particular in ELECTRE TRI and ELECTRE III [Roy, 1978])(illustrated in Figure 2.6):

$$c_i(a,b) = \begin{cases} 0 & \text{if } g_i(b) \geq g_i(a) + p_i \\ 1 & \text{if } g_i(b) \leq g_i(a) + q_i \\ \frac{g_i(b) - g_i(a) - q_i}{p_i - q_i} & otherwise \end{cases}$$

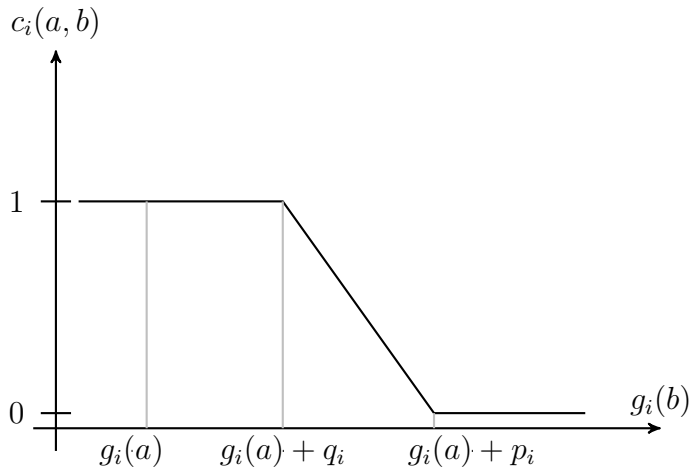

Figure 2.6: Curve representing the concordance indicator on one criterion

Then from the concordance indicators of all the criteria, a global concordance indicator is created $C(a,b)$ which is equal to the weighted sum of the concordance indicators with the weights of the criteria i.e. $C(a,b) = \sum_{i \in N} w_i c_i(a,b)$. The discordance indices are found for every criterion $d_i(a,b), \forall i \in N$ as follows (illustrated in Figure 2.7):

$$d_i(a,b) = \begin{cases} 0 & \text{if } g_i(b) \leq g_i(a) + p_i \\ 1 & \text{if } g_i(b) \geq g_i(a) + v_i \\ \frac{g_i(b) - g_i(a) - p_i}{v_i - p_i} & otherwise \end{cases}$$

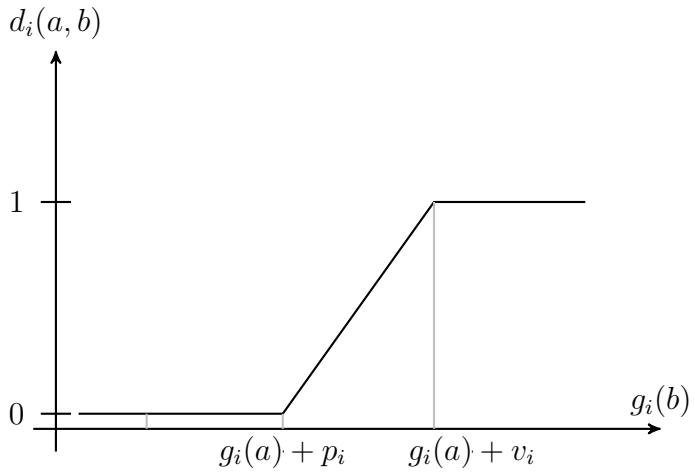

Figure 2.7: Curve representing the discordance indicator on one criterion

Then, in ELECTRE TRI, the credibility index $\sigma(a,b)$ is computed as follows:

$$\sigma(a,b) = \begin{cases} \sum_{i \in N} c_i(a,b) & \text{if } d_i(a,b) \leq c_i(a,b) \forall i \in N \\ \sum_{i \in N} c_i(a,b) \prod_{i \in \bar{N}} \frac{1-d_i(a,b)}{1-c_i(a,b)} & \text{otherwise} \end{cases}$$

Where $\bar{N} = \{i \in N : d_i(a,b) > c_i(a,b)\}$
It is considered that a outranks b if $\sigma(a,b) > s$.

The reader may notice here that if for any criterion $i \in N$, $d_i(a,b) = 1$ then automatically $\sigma(a,b) = 0$. This property is desired as a possibility to avoid a to outrank b based on a veto on the criterion i regardless to the values of these two objects on the other criteria following the non-discordance principle.

It is important to notice that two objects a and b can outrank each other. In this case we will consider that there is an indifference relation between them. Between two objects a and b it is also possible that none of them outrank the other one. In that case it will be considered that they are incomparable.

From an outranking relation to a sorting process

In ELECTRE TRI [Mousseau and Slowinski, 1998] the assignment of an object a results from the comparison of a with the profiles defining the limits of the classes. What we refer to as profiles here is a set of $r-1$ objects that will be considered as the frontiers of r categories, the category c_r being the highest one and the category c_1 being the lowest one. Each profile b_h is the upper bound of the category c_h and the lowest bound of the category c_{h+1}. It is to be mentioned that each bound b_h weakly dominate the bound b_{h-1} located above. Two assignment procedures are then available to assign a:

Pessimistic procedure:

1) Compare a successively to b_j, for $j = r$ to 2.

2) b_h being the first profile such that aSb_h, a is assigned to category c_h

3) If $\neg aSb_2$, then assign a to category c_1.

Optimistic procedure:

1) Compare a successively to b_j, for $j = 2$ to r.

2) b_h being the first profile such that $b_h S a$ and $\neg a S b_h$, a is assigned to category c_h

3) If no b_h is such that $b_h S a$ and $\neg a S b_h$ (often refereed to as $b_h P a$), then a is assigned to category c_r

MR-Sort model

MR-Sort [Bouyssou and Marchant, 2007a][Bouyssou and Marchant, 2007b] is a simplified version of ELECTRE TRI which does not involve a indifference threshold and a preference threshold. Instead it is here considered that for each criterion i the concordance indicator $c_i(a, b_h)$ is equal to 1 if and only if $g_i(a) \geq g_i(b_h)$. Some versions of MR-Sort involve a set of veto profiles V_h [Leroy et al., 2011] although it is generally not the case. In MR-Sort, each object $a \in A$ is assigned to a class as follows:

- Compare a successively to b_j ,for $j = 2$ to r.
- b_h being the first profile such that *not* $a S b_h$, a is assigned to category c_{h-1}. Here we say that $a S b_h$ if $\sigma(a, b) > s$ and, when veto profiles are considered, if a weakly dominates V_h.
- If $a S b_r$, then assign a to category c_r.

Given that this algorithm will later be used in the construction of the *Biodiversity Severity Index*, we illustrate the functioning of this method through an example so that the reader can easily understand it. Let us assume that a bank wants to create a rule for project funding acceptance. The project will be assigned to one of three categories: "rejected" c_1, "to be discussed at the next meeting" c_2 and "accepted" c_3 according to three criteria : the amount of the loan (criterion 1 expressed in €), the risk (criterion 2 expressed through five categories for the safest A to the riskiest E) and the interest rate that the client is willing to accept (criterion 3 expressed in %). We assume that being a small size bank, a small loan is preferred to a big one, that the bank is averse to risk and prefers a safe investment to a risky one and that a high interest rate is preferred to a low one. Let us assume that in order to take that decision systematically and from a neutral point of view, the bank decides to use an MR-Sort model with two profiles: the profile $b_2 = (100\,000€, C, 4\%)$ delimiting the categories c_1 and c_2 and the profile $b_3 = (50\,000€, A, 6\%)$ delimiting the categories c_2 and c_3, each criteria with a weight of 0.333 and the concordance level equal to 0.5. Then, a project $a_1 = \{50\,000€, D, 6\%\}$ would outrank the profile b_2 due to the fact that it is better than b_2 on two criteria. For the same reason it would outrank the profile b_3 and thus would be assigned to category

c_3 "accepted". Likewise the project $a_2 = \{100\,000€, B, 6\%\}$ would outrank b_2 but not b_3 and hence would be assigned to category c_2 "to be discussed at the next meeting". The bank could also claim that a loan of more than 200 000€ or with a risk rate of D or E cannot be accepted directly and thus can only be assigned to c_1 or to c_2. This decision could be modelled through the use of a veto profile $v_3 = \{200\,000€, C, -\infty\%\}$. With this change, the project a_1 will not weakly dominate the veto profile v and thus will be assigned to category c_2. This example is represented in Figure 2.8.

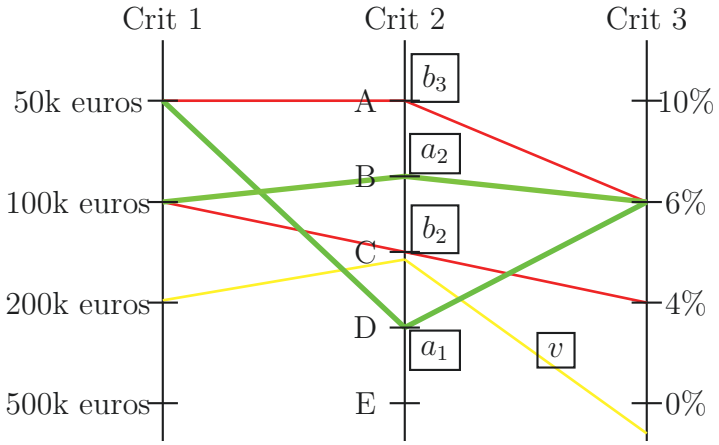

Figure 2.8: Diagram representing the MR-Sort procedure. The profiles b_2 and b_3 are represented in red, the veto profile v is represented in yellow and the projects a_1 and a_2 are represented in green. Each column represents a criterion.

Non Compensatory Sorting model

MR-Sort being a simplified model for decision is unable to return some possible preferences. As illustrated in [Sobrie et al., 2015] we could imagine an application in which a committee for a higher education program wants to chose whether or not students should be admitted on the basis of their evaluations in 4 courses: math, physics, economics and history. The committee states that, to be accepted, a student should have an evaluation above 10/20 in at least one of these coalitions: $\{history, physics\}$, $\{history, math\}$, $\{economics, math\}$ and $\{economics, physics\}$. From a practical point of view, this choice could be understood as the will to recruit students that have abilities in scientific matters and in social sciences. However, this model cannot be represented as an MR-Sort model (which cannot represent preferences with dependence between criteria as described in Subsection 2.2.6). Indeed, if an MR-Sort model had to represent this two category classification model then necessarily the profile would be equal to (10,10,10,10). Let us call w_m (resp. w_p, w_h, w_e) the weight of math (resp. physics, economics, history) in the MR-Sort model, we have:

$$\begin{cases} w_m + w_h \geq s \\ w_m + w_e \geq s \\ w_p + w_h \geq s \\ w_p + w_e \geq s \\ w_m + w_p < s \\ w_h + w_e < s \end{cases}$$

However

$$\begin{aligned} w_m + w_h > s \; and \; w_p + w_e > s \; &\Rightarrow \; w_m + w_h + w_p + w_e \geq \; 2s \\ w_m + w_p < s \; and \; w_h + w_e < s \; &\Rightarrow w_m + w_h + w_p + w_e < 2s \end{aligned} \tag{2.1}$$

From a theoretical point of view this can be seen as the result of a dependency between criteria, here a negative synergy (or redundancy) between economics and history and identically between math and physics.

In order to deal with this limitation one may use a model named Non Compensatory Sorting model (NCS model) proposed by [Bouyssou and Marchant, 2007a]. The basic idea of this model is that an object a outranks a profile b_i if and only if a is at least as good a b_i on a set of criteria $I \subseteq N$ that is considered as "sufficiently important". $\mathbb{F}$ is the set of all the "sufficiently important" sets of criteria and if $I \in \mathbb{F}$ then $J \in \mathbb{F}$ for each J such that $I \subseteq J$. Then, the objects are assigned to categories by successively comparing them to the profiles as it is done with MR-Sort model without veto.

In [Sobrie et al., 2015], a version of the Non Compensatory Sorting model was proposed where weights are associated to coalitions of criteria i.e. there exists a capacity (or weight) function $\mu : 2^N \rightarrow [0, 1]$ such that:

- $\mu(\emptyset) = 0$
- $\mu(N) = 1$
- $\mu(I) \leq \mu(J)$ if $I \subseteq J$
- $\mu(I) = \sum_{J \subseteq I} m(J)$ and $m(J) = \sum_{K \subset J} (-1)^{|J|-|K|} \mu(K)$.

Here the meaning of $m(I)$ is to represent the surplus value of the coalition I against the value of each criterion or of subsets of I.

In order to assign objects to category the principle is similar to MR-Sort method. An object a will be sorted in category c_i if the set of criteria I for which a is better than the profile b_i has a weight (or capacity) higher than the concordance level s and the set of criteria J for which a is better than the profile b_{i+1} has a weight (or capacity) lower than the concordance level s. We can say that this method is a specific case of Non Compensatory Sorting model where $\mathbb{F} = \{I \subseteq N : \mu(I) \geq s\}$. In practice, there exists 2^n different sets of criteria which may make the elicitation difficult and long if one want to find the weights of all of them. In order to make this model available in practice a solution consist in ignoring the possible interactions of more than two criteria i.e. $\forall I \in N$ such that $|I| \geq 3$, $m(H) = 0$. This method is called 2-additive NCS model.

2.2.3 Synthetic criteria and utility

Synthetic criterion methods, also referred to as scoring methods, consist in attributing a score $u(a)$ to each object $a \in A$ individually [Sarin, 2001][Keeney and Raiffa, 1994]. The synthetic criterion may then be used to establish a ranking, a choice or an assignment. This is a common approach to aggregate criteria and many methods are based on it such as UTA [Siskos et al., 2005], MACBETH [Bana e Costa et al., 2016], AHP [Saaty, 1990] or goal programming [Tamiz et al., 1998]. We can see two main advantages to synthetic criteria. First, the synthetic criteria are quite well accepted by decision makers. Indeed, from childhood, people are used to evaluate and be evaluated by scores, whether through marks in school in many countries (France, Switzerland, Spain, Romania etc.) or through many common indicators such as the Gross Domestic Product or the Human Development Index. We could also mention as a good point the fact that a synthetic criterion allows to compare any pair of objects. Nevertheless this point could also be seen as a bad property if one suggests that incomparability may exist in the given context. Indeed, utility methods are generally considered as compensatory. When the synthetic criterion of an object is understood as "proportional" to its attractiveness we generally use the term *utility*.

Utility was described by its main and first developer Jeremy Bentham [Bentham, 1780] as follows: "By utility is meant that property in any object, whereby it tends to produce benefit, advantage, pleasure, good, or happiness, (all this in the present case comes to the same thing) or (what comes again to the same thing) to prevent the happening of mischief, pain, evil, or unhappiness to the party whose interest is considered".

Multi-attribute utility methods [Mateo and Ramón San, 2012] aims at reflecting an individual's affinity for some objects through a synthetic criterion. Most of the utility

methods are said additive which means that objects has a utility on every criterion depending upon its evaluation on each criterion and the synthetic criterion is generally the sum of its utilities on every criteria although some non-additive utility methods exists, although some methods as Choquel integrals [Grabisch and Roubens, 2000] were created to deal with non-additive utility. The additive utility function u takes the following form:

$$u(a) = \sum_{i \in N} w_i u_i(g_i(a)) \tag{2.2}$$

Where u_i are non-decreasing real utility functions, named utility functions, which are normalized between 0 and 1, and w_i is the weight of the criterion i.

In order to deal with the sorting problem, a common approach [Devaud et al., 1980] consists in fixing a set of thresholds $s_k, k = 2, ..., r$ that separate the categories i.e. an object a will be assigned to category c_1 if $u(a) < s_2$, to category c_r if $u(a) \geq s_r$ and to category c_k with $2 \leq k \leq r-1$ if $s_k \leq u(a) < s_{k+1}$.

Choquet integrals

Choquet integral is a non-additive utility method that aims at representing preferences where the criteria are not considered as being independent according to the definition of independence given in Subsection 2.2.6. The idea of this method is to integrate the interactions between criteria by the creation a capacity $\kappa : 2^N \to [0, 1]$ which is a function that associates a utility to each subset of criteria $N' \subseteq N$. Then, the choquet integral of an object a according to κ is

$$\varsigma_\kappa(a) = \kappa(N) \cdot g_{\tau(1)}(a) + \sum_{i=2}^{n} g_{\tau(i)}(a) - g_{\tau(i-1)}(a) \cdot \kappa(\{\tau(i), \tau(i+1), ..., \tau(n)\}) \tag{2.3}$$

τ being a permutation over N such that $g_{\tau(1)}(a) \leq g_{\tau(2)}(a) \leq ... \leq g_{\tau(n)}(a)$. κ is monotonic i.e. $N'' \subseteq N' \Rightarrow \kappa(N'') \leq \kappa(N')$.

An illustration of the functioning of choquet integrals is given through an example with 5 criteria on figure 2.9.

2.2.4 Rule based methods

We generally call a rule based method, a sorting procedure $f : A \to C$, that consists in assigning each object to a category following a set of logical rules such as "if $g_1(a) \geq x_1$ and ... and $g_n(a) \geq x_n$ then a should be classified in category at least c i.e. $f(a) \geq c$"

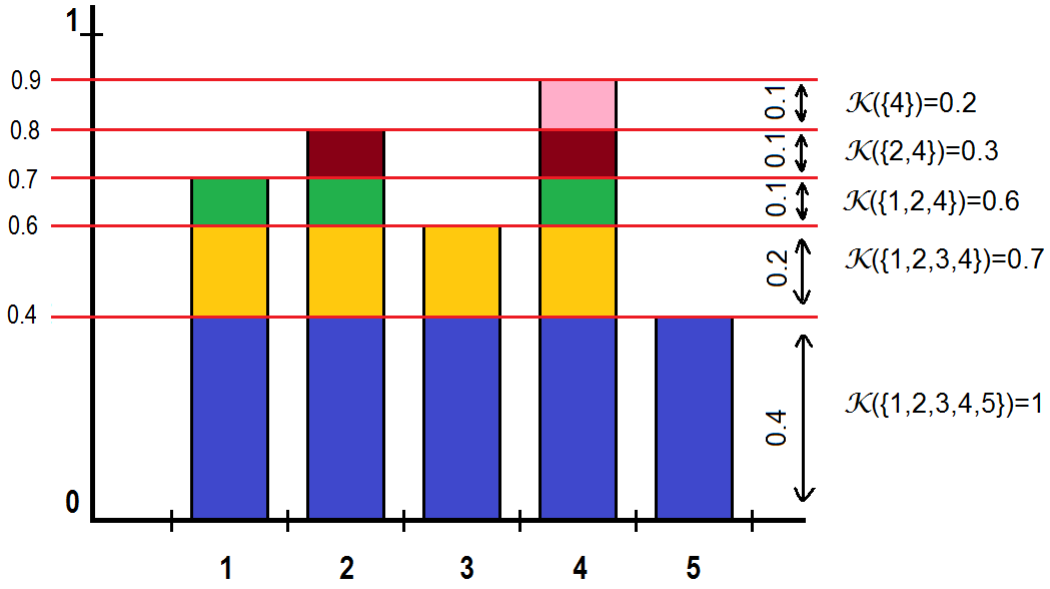

Figure 2.9: Graphic representation of the Choquet integrals operating principle. Here the columns represent the values $g_i(a)$ of the object a on the five criteria. Here the choquet integral of a is $\varsigma_\kappa(a) = 1 \cdot 0.4 + 0.7 \cdot 0.2 + 0.6 \cdot 0.1 + 0.3 \cdot 0.1 + 0.2 \cdot 0.1 = 0.65$. The capacities that are useful in this situation are shown on the right side of the figure.

with $c \in C$ and $x_i \in v_i$)[Greco et al., 2001a][Pawlak and Slowinski, 1994] (or conversely "if $g_1(a) \leq x_1$ and ... and $g_n(a) \leq x_n$ then a should be classified in category at most c"). Any object $a \in A$ should be assigned to exactly one category $c \in C$.

The advantage of this very simple approach consists in being very flexible. Any monotonic assignment based on discrete and finite scales can be obtained through a rule based method. Indeed, for any monotonic assignment $f : A \to C$, the following set of rule f' would reproduce f:

$\forall a \in A$

1) if an object a' weakly dominates a, then $f'(a') \geq f(a)$

2) if an object a' is weakly dominated by a, then $f'(a') \leq f(a)$

Thus, rule base methods can represent assignment that do not respect independence between the criteria or influenced by veto threshold as long as monotonicity is not violated.

2.2.5 SMAA: A stochastic methods to deal with uncertainty or imprecise informations or group decisions

In applied decision problems, it is common that some of the values of the objects on the criteria is to some degree subject to uncertainty or imprecision. As well the decision maker may not have the necessary understanding of the meaning of the preference parameters of a MCAP to express them exactly or could not have a categorical opinion on the appropriate preference parameters. In this context SMAA methods aims at studying how different could be the conclusion of a method with slight modifications of its input.

Stochastic multi-criteria acceptability analysis (SMAA)[Lahdelma et al., 1998][Lahdelma and Salminen, 2001] is a family of methods used for multi-criteria decision aiding in problems with uncertain, imprecise, partially missing information or for studying robustness in multi-criteria decision aiding problems. The main principle of these methods consists in exploring a studied weight and criteria space with a Monte Carlo process in which, at each step (trial), a random set of parameters and a random object a are generated.

In SMAA methods the object are not necessarily characterized by exact values on the criteria but instead they generally follow a probability distribution of values on the criteria. If it was expressed as an interval we may chose the uniform probability distribution over this interval. A MCAP is chosen that is supposed to be relevant in the decision context. At each step (trial) a random object a is generated, its values on the criteria following the probability distribution associated to each criterion. As well, at each trial a random set of preference parameters for the chosen MCAP (for instance weights in SMAA-2 which consider the weighted sum as the MCAP) is generated, following a probability distribution given as an input of the method. According to these values a ranking, an assignment or a classification can be found. Formally the generic simulation scheme for analyzing stochastic multi-criteria problems with

different variants of SMAA is presented in Algorithm 3.

Algorithm 3: Generic SMAA simulation, source: [Lahdelma and Salminen, 2010]

Data: An MCAP M, for an object a a probability distribution X over the values on criteria, a probability distribution W over the possible parameters of the MCAP M

1 **repeat**
2 Draw $< a, w >$ from their distributions X and W.
3 Rank, sort or classify a using M, and the parameters w.
4 Update statistics about a.
5 **until** k *times*;

A main issue of this document is the sorting problem while SMAA is mainly used for the ranking and nominal classification problems. The SMAA TRI [Figueira et al., 2004] methods deals with the sorting problem but only with the intention to study the robustness of an ELECTRE TRI model.

In SMAA TRI the object a is fixed and its values on the criteria are deterministic. The parameters of the ELECTRE TRI model instead (presented in Subsection 2.2.2), are distributed randomly over a space of ELECTRE TRI parameters. The category acceptability index for the category c measures the stability of the assignment of a. It can be interpreted as a probability for the object a to be assigned in category c with an ELECTRE TRI model with the parameters following the predefined distribution W. The category acceptability indices are within the range [0,1], 0 meaning that the action almost surely not be assigned to the category, and 1 indicates that it will almost surely be assigned to the category. For each action, the sum of the acceptabilities for different categories is equal to one. If the parameters are stable, the category acceptability indices for each action should be close to 1 for one category, and for the others. In this situation the assignments are said to be robust with respect to the imprecise parameters. An illustration of the principle of SMAA TRI is shown on Figure 2.10.

2.2.6 Choosing the appropriate Multi-criteria aggregation procedure

In order to determine which decision aiding method and thus which MCAP should be chosen to deal with a given problem, we should ask ourself several questions about possible properties that could be appropriate or not in the studied decision context.

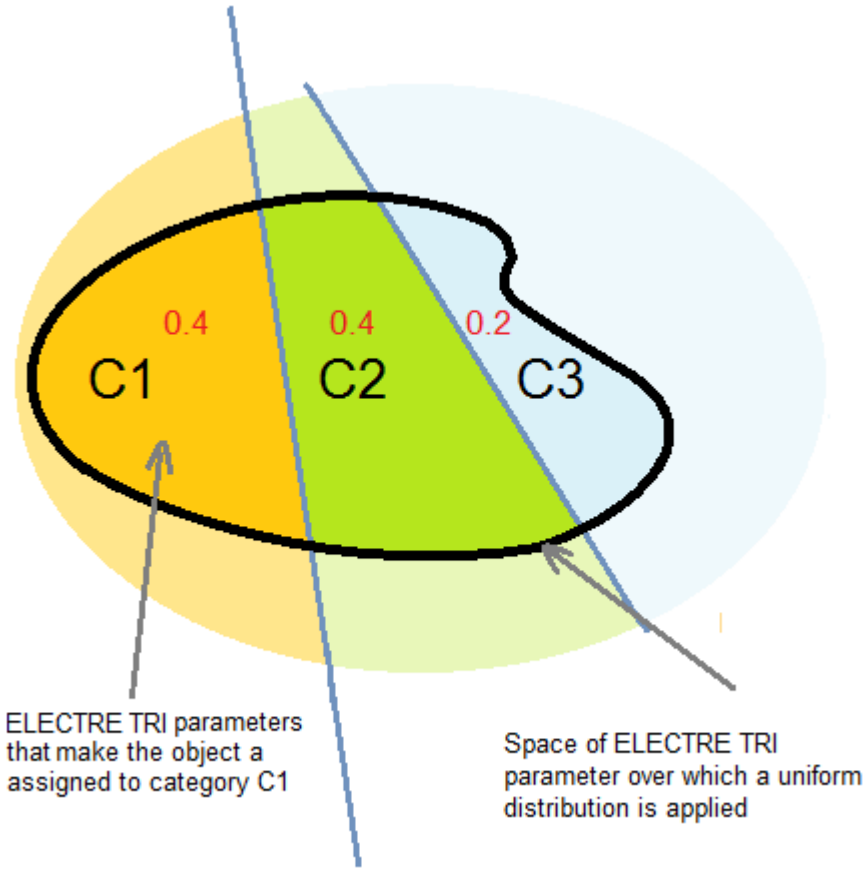

Figure 2.10: Graphic representation of the SMAA TRI operating principle with fixed values for the object a and random values for the ELECTRE TRI parameters. Here every point of the space represents an ELECTRE TRI parameter. The shape in black represents the space over which the ELECTRE TRI parameters are uniformly distributed. We can observe three areas that represent respectively the parameters that would assign the object a to category $C1$, $C2$ and $C3$. Here, we can see that the category acceptability index of a for the category $C1$ is 0.4.

- **Expected output**: Each MCAP is conceived to deal with a specific problem. In Subsection 2.1.3 we saw that the analyst should at first figure out which problem she is facing. Thus one of the MCAP that is chosen should correspond to the problem that is being faced (sorting, ranking or choice).
- **Compatibility of the original performance scales with the considered method**: Some methods cannot handle directly the evaluation criteria whose performances are expressed on nominal scales, or even numerical but purely ordinal scales. In this case, the user should check if it is possible to transform in a

meaningful way the original scales, such that the properties of scales required by the considered method are satisfied.

- **Ability of the decision maker to answer the elicitation questions**: MCAP are associated with elicitation methods as we will see in Section 2.3. Thus, while looking for an appropriate MCAP the analyst should consider whether the decision maker feels able to answer the questions that will be asked to him during this elicitation process.

- **Independence to irrelevant objects**: We say that a MCAP is dependent to irrelevant objects (sometimes mentioned as rank reversal in ranking problems) if the evaluation of an object a might depend on the presence or not of an other object b in A the set of objects. In other words, we say that a MCAP is independent to irrelevant objects if adding an object a' to A cannot change the relationship between a and b or the assignment of a (depending on the decision problem that we are facing).

- **Dependence between criteria**: Although the term dependence may cover various phenomenons in decision aiding, what we will here refer while talking about independences between criteria is what [Keeney, 1992] names preferential independence. The pair of criteria i, j is preferentially independent of the other criteria $k, ..., n$, if the preference order for consequences involving only changes in the values of i and j does not depend on the values at which criteria $k, ..., n$ are fixed. When there are dependencies between criteria, it can be due to positive synergy between criterion i and criterion j, i.e. improving the evaluation on criterion i has more impact if the evaluation on criterion j is hight than if it is low or to negative synergy (or redundancy) i.e. improving the evaluation on criterion i has less impact if the evaluation on criterion j is hight than if it is low.

  In some particular contexts the decision maker might think that such dependencies are appropriate while, in others, she could think that no such effect is desired. For instance, while choosing a house, one could think that the absence of public transport would be more problematic if there is no place to park a car. This would mean that, in the decision maker's mind, there is a negative synergy (or redundancy) between the criterion "quality of the surrounding public transport system" and the criterion "presence of a car parking place". The choquet integrals presented in Subsection 2.2.3 is particularly designed to deal with problem where this type of independence may be inappropriate.

- **Veto phenomenon**: The main idea of veto is to claim that some "too bad values" on some criteria may not be compensable by good values on other criteria.

The veto phenomenon is mainly used in ELECTRE methods as presented in Subsection 2.2.2. For instance, while comparing houses to rent, one could think that she could never tell that a house H_0 is better than a house H_1 if H_1 is cheaper than H_0 with a difference of prices higher than 2000 euros per month. The veto phenomenon is considered as an important limit to compensation in multi-criteria decision aiding. It can also lead to incomparability if several vetoes are met although incomparability can also happen without veto.

- **Imperfect or incomplete information**:

  For various reasons the informations that are used as criteria and/or the judgement made by the decision maker might be subject to imprecision or inconsistency. Some methods such as rough set [Slowinski et al., 2002] or SMAA methods (presented in Subsection 2.2.5) are adapted to this kind of noises due to their ability to deal with non-monotonic preferences.

Choosing an appropriate Multi-Criteria Aggregation procedure for the sub-problems of a hierarchy of criteria

When a problem is modeled with a hierarchy of criteria, the output (category) of each sub-problem excepted the highest one will be the input (criterion) of a higher sub-problem. Therefore, all of the sub-problems of the hierarchy must be sorting problems (minus eventually the highest sub-problem).

2.3 Elicitation methods

As was stated earlier, decision aiding in general and multi-criteria decision aiding in particular are scientific fields that deal with subjectivity. By this word we mean that there might not be a best solution or a best assignment in absolute but instead, we should look for some solutions or assignments that are adapted to the preferences of our decision makers (that could be considered as not adapted to other decision makers). Most of the MCAPs that are developed in MCDA takes into account the subjectivity of the expressed point of view through some preferences parameters (such as weights, thresholds etc) that can take several forms according to the MCAP used.

For each MCAP, there exists one or several elicitation methods i.e. methods to interact with the decision maker and find the parameters that correspond to her preferences. These elicitation methods can be categorised into two approaches [Mousseau,

2003]: Aggregation approach and disaggregation approach. Nevertheless, these two approaches are sometimes combined together.

2.3.1 Aggregation approach

The aggregation approach is the most frequently used in multi-criteria decision aiding [Mousseau, 2003]. Adopting this approach induces six steps:

1) Defining the set of studied objects
2) Defining the set of criteria on which these objects are going to be judged.
3) Choose a MCAP
4) Choose some preference parameters to this method that seem to be appropriate
5) Aggregate the data on the criteria to get the global preference

The aggregation approach assumes that the preference parameters of the chosen MCAP can be understood by the decision maker and that she is capable, with the help of the analyst, to choose those that are appropriate to her preferences.

2.3.2 Disaggregation approach

The disaggregation approach [Jacquet-Lagrèze and Siskos, 2001][Jacquet-Lagrèze, 1979] in preference elicitation consists in asking the decision maker to give a partial information about the output that we expect to obtain [Cailloux, 2012]. Then a disaggregation algorithm must be applied in order to find the preference parameters that return as well as possible the preferences expressed by the decision maker. According to Vincent Mousseau [Mousseau, 2003] adopting it implies six steps:

1) Defining the set of studied objects.
2) Defining the set of criteria on which these objects are going to be judged.
3) Choose a MCAP.
4) Ask the decision maker about her preference on a subset of objects.

5) Use a disaggregation algorithm to get the preference parameters that returns as well as possible the preferences expressed by the decision maker.

6) Apply the MCAP with the previously obtained preference parameters on every object to be assessed.

In practice these two steps are often applied alternately in what is generally called an aggregation/disaggregation process [Mousseau, 2003].

2.3.3 Expression of the preferences for the sorting problem with a disaggregation approach

As stated earlier the use of a disaggregation approach requires from the decision maker to express her preferences in term of an expected output over a $A' \subset A$. In a sorting problem, the expected output of the process is the assignment of the objects to categories. Thus the expression of the preferences of a decision maker is in this case relative to object assignment. We will present three possible ways to collect preference in such context:

1) **Exact Assignments:** In this situation the decision maker gets asked in which category each object a of $\Theta \subseteq A$ should be assigned. Formally a learning set is expressed as $L= < \Theta, f_l >$, $\Theta \subseteq A$ being the set of objects that are supposed to be assigned by the decision maker and $f_l : \Theta \rightarrow C$ being the assignment of these examples.

2) **Assignments into intervals of categories:** Here, the decision maker should express for each object $a \in \Theta \subseteq A$ an interval of categories into which it should be assigned (such as "I think that the object a should be assigned between category 3 and category 4").

3) **Argument Strength Assessment:** In this situation proposed in [Cailloux, 2012] the decision maker expresses for each object $a \in \Theta \subseteq A$ and each category c a integer score representing level of confidence for the assertion "I think that the object a should be sorted in the category c". This method is referred to as cardinal argument strength assessment when the sum of the scores is fixed. This information can be used to obtain intervals of categories.

We are now going to present some elicitation methods. Given that the main topic of this work is multi-criteria sorting problem, we will mainly present preference learning methods for the sorting problem.

2.3.4 UTA methods: a disaggregation approach algorithm based on utility

The UTA (UTilités Additives) [Jacquet-Lagreze and Siskos, 1982] method, is both a MCAP and an elicitation method proposed by Jacquet-Lagrèze and Siskos (1982) that aims at inferring the utility functions of the criteria from a given ranking on a learning set. The elicitation provided while applying UTA is a disaggregation type. The method uses piecewise linear functions as utility functions for every criteria. UTA Methods use linear programming techniques to find these functions so that the ranking or the assignment obtained through these functions is as consistent as possible with the ranking given as an input. The MCAP in UTA is assumed to be an additive utility function.

The utility function u takes the following form:

$$u(a) = \sum_{i \in N} w_i u_i(g_i(a)) \tag{2.4}$$

Where u_i are non-decreasing piecewise linear functions, named utility functions, which are normalized between 0 and 1, and w_i is the weight of the criterion i. One may observe that the weights are not necessary if the utility function are not normalized between O and 1. We will thereafter consider this situation.

UTADIS [Devaud et al., 1980][Zopounidis and Doumpos, 1997] is a multi-criteria elicitation method for the sorting problem based on additive utility principle. As with other UTA methods the utility functions are piecewise linear functions u_i and the global utility u is a weighted sum of these functions. The introduction of thresholds t_c delimiting the classes allow the method to determine for each object the category to which it should belong.

Thus, the input given by the decision maker is a partial exact assignment of a subset of objects $\Theta \subseteq A$. The output is a complete assignment of the object set A. It is obtained with the use of a linear program that aims at finding the break points of all the utility function on the criteria and the thresholds t while minimizing the difference between the partial assignment given by the decision maker and the complete assignment obtained as an output. For that purpose we add for each object a in le learning set one or several

constraints that forces the a to be assigned to the category $f_l(a)$ or if not possible that counts it as an error. For the object assigned to a category $1 < k < r$ these two constraints are the following:

$$\begin{cases} \sum_{i \in N} u(g_i(a)) - \sigma^-(a) \leq t_k - \varepsilon & \forall a \in c_k, 1 < k < r \\ \sum_{i \in N} u(g_i(a)) + \sigma^+(a) \geq t_{k+1} & \forall a \in c_k, 1 < k < r \end{cases}$$

ε is a very small real number. Here is the simplified linear program used in UTADIS when there is a break point for every value on every criterion (illustration in Figure 2.11):

$$\begin{cases} \text{minimize} & \sum_{a \in C} \sigma^+(a) + \sigma^-(a) & & \\ \text{subject to} & \sum_{i \in N} u(g_i(a)) - \sigma^-(a) \leq t_1 - \varepsilon & \forall a \in c_1 & \text{DM's preferences} \\ & \sum_{i \in N} u(g_i(a)) - \sigma^-(a) \leq t_k - \varepsilon & \forall a \in c_k, 1 < k < r & \text{DM's preferences} \\ & \sum_{i \in N} u(g_i(a)) + \sigma^+(a) \geq t_{k+1} & \forall a \in c_k, 1 < k < r & \text{DM's preferences} \\ & \sum_{i \in N} u(g_i(a)) + \sigma^+(a) \geq t_{r-1} & \forall a \in c_r, r = |C| & \text{DM's preferences} \\ & u(\alpha_i) \leq u(\alpha_{i+1}) & \forall \alpha_i, \alpha_{i+1} \in v_i, i \in N & \text{Monotonicity} \\ \textit{Variables} & u(\alpha_i) \in [0,1] & \forall \alpha_i \in v_i, i \in N & \\ & \sigma^+(a), \sigma^-(a) & \forall a \in A & \end{cases}$$

Intuitively, here we try to make the UTADIS model as far as possible to violate the learning set and if the learning set must be violated, to violate it as few as possible.

Several other versions of UTADIS exist (UTADIS I [Zopounidis and Doumpos, 1997], UTADIS II [Zopounidis and Doumpos, 2001], UTADIS III [Doumpos and Zopounidis, 2002]) dealing differently with the two objectives of minimizing the number of misclassified objects and of maximizing the distances of the correctly classified objects from the utility thresholds.

UTADIS methods are efficient methods to find the parameters of an additive utility function based on a disaggregation approach for elicitation. Some methods such as UTADIS II and UTADIS III give a guaranty to return all the objects in Θ in the correct category if there exists an additive utility function that does so. However, its use should be restricted to situations in which an additive utility method and the principle of disaggregation approach for elicitation of the preference is appropriate.

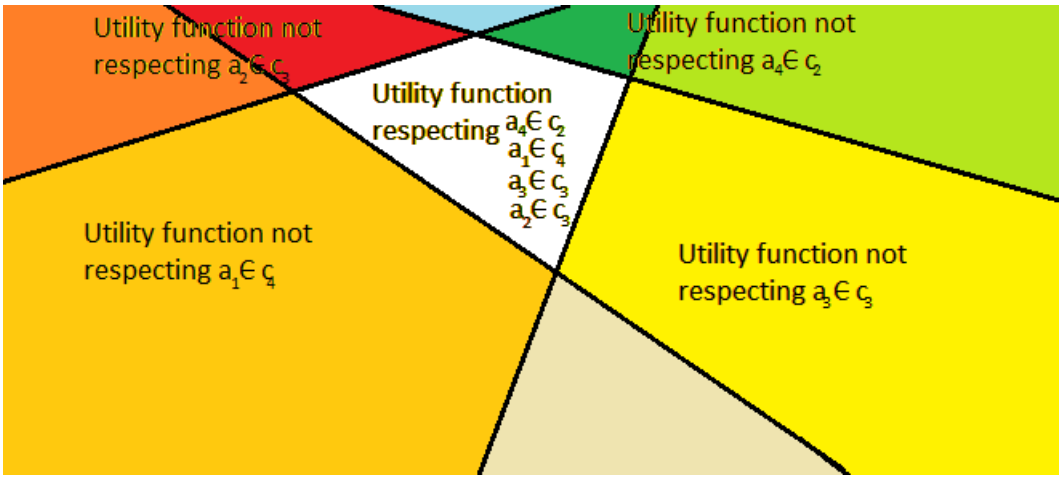

Figure 2.11: Graphic representation of the UTADIS operating principle from a given learning set

For instance, if there may be some dependence between some of the criteria then we should consider that additive utility method is not appropriate. As well, the meaning of the different utility functions on each criterion may not look obvious to the decision makers.

2.3.5 A mixed integer programming algorithm for disaggregation elicitation with MR-Sort

Several methods are available to find MR-Sort parameters. Let us present an elicitation algorithm based on a mixed integer program [Leroy et al., 2011]. As other elicitation algorithm based on mathematical programming (UTA methods presented in Subsection 2.3.4 for instance) the basic idea is to find a set of parameters for the studied MCAP (here an MR-Sort model) such that maximizes the number of objects from the learning set that are assigned with this MCAP to the category in which the decision maker chose to assign them.

The parameters to be obtained in MR-Sort (defined in Subsection 2.2.2) are the following:

- For every criterion i a weight w_i.
- $r - 1$ profiles delimiting the r categories.

- A concordance index s.

In order to find the most adapted parameters for a decision maker, she gets asked a partial exact assignment i.e. to assign a subset of objects.

For the sake of simplicity, we will first present the case in which there are two categories $C1$ the worst and $C2$ the best separated by a profile b. Let us define for each object $a_j \in \Theta$ and each criterion $i \in N$ an integer variable $\delta_{i,j} = 1 \Leftrightarrow g_i(a_j) \geq g_i(b)$ and 0 otherwise. In order to impose this value for $\delta_{i,j}$ the following constraints are added for each $\delta_{i,j}$.

$$\begin{cases} M(\delta_{i,j} - 1) \leq g_i(a_j) - g_i(b) < M(\delta_{i,j}) \\ \delta_{i,j} \in \{0, 1\} \end{cases} \tag{2.5}$$

Then from these integer variables we define the continuous variables $\xi_{i,j} = \delta_{i,j} \cdot w_i$. The aim of these variables is to be summed and represent for each a the total weight of the criteria such that $g_i(a) \geq g_i(b)$. These two variable being non fixed we cannot simply multiply them which would make this program quadratic. In order to fix this problem, we impose the following constraints for each $\xi_{i,j}$:

$$\begin{cases} \delta_{i,j} - 1 + w_i \leq \xi_{i,j} \leq \delta_{i,j} \\ 0 \leq \xi_{i,j} \leq w_i \end{cases} \tag{2.6}$$

Then, for each object in $C1$ (resp. $C2$) a constraint is added to check that for each of them the total weight of the criteria such that $g_i(a) \geq g_i(b)$ is lower (resp. lower) than s.

$$\begin{cases} \sum\limits_{i \in N} \xi_{i,j} - \alpha \leq s & \forall a_j \in C1 \\ \sum\limits_{i \in N} \xi_{i,j} + \alpha \geq s & \forall a_j \in C2 \end{cases} \tag{2.7}$$

Here α intuitively represents how far from being assigned to the good category is the most misclassified object. If α is equal to 0 or lower this means that every object are assigned in the good category (the one given by the decision maker). Obviously α must be minimized. Thus, the final linear program is the one given in (2.8)

$$\begin{cases} \textit{Minimize} & \alpha & \\ \textit{Subject to} & \sum_{i \in N} \xi_{i,j} - \alpha \leq s & \forall a_j \in C1 \\ & \sum_{i \in N} \xi_{i,j} + \alpha \geq s & \forall a_j \in C2 \\ & M(\delta_{i,j} - 1) \leq g_i(a_j) - g_i(b) & \forall i \in N, \forall a_j \in A \\ & g_i(a_j) - g_i(b) < M(\delta_{i,j}) & \forall i \in N, \forall a_j \in A \\ & \delta_{i,j} - 1 + w_i \leq \xi_{i,j} & \forall i \in N, \forall a_j \in A \\ & \xi_{i,j} \leq \delta_{i,j} & \forall i \in N, \forall a_j \in A \\ & 0 \leq \xi_{i,j} & \forall i \in N, \forall a_j \in A \\ & \xi_{i,j} \leq w_i & \forall i \in N, \forall a_j \in A \\ & \sum_{i \in N} w_i = 1 & \\ \textit{Variables} & \delta_{i,j} \in \{0, 1\} & \forall i \in N, \forall a_j \in A \\ & \xi_{i,j} \in [0, 1] & \forall i \in N, \forall a_j \in A \\ & w_i \in [0, 1] & \forall i \in N \\ & \alpha \in \mathbb{R} & \\ & g_i(b) \in \mathbb{R} & \forall i \in N \end{cases} \tag{2.8}$$

In order to adapt this method to situation with more that two categories, a simple change must be made. First $r-1$ profiles must be found (thus, $(r-1) \cdot n$ variables). For each object a belonging to the category C_h with $h \in \{2, ..., r-1\}$ two variables $\delta_{i,j}^{h-1}$ and $\delta_{i,j}^{h}$ are created such that $\delta_{i,j}^{h-1} = 0 \Leftrightarrow g_i(a_j) \geq g_i(b_{h-1})$ (similarly for h). We also add for each object a belonging to the category C_h with $h \in \{2, ..., r-1\}$ two variables $\xi_{i,j}^{h-1}$ which is equal to $\delta_{i,j}^{h-1} \cdot w_i$. Then or each object a belonging to the category C_h with $h \in \{2, ..., r-1\}$ two constraints are added (presented in 2.9) to avoid or limit potential misclassifications.

$$\begin{cases} \sum_{i \in N} \xi_{i,j}^{h-1} + \alpha \geq s & \forall a_j \in C_h \\ \sum_{i \in N} \xi_{i,j}^{h} - \alpha \leq s & \forall a_j \in C_h \end{cases} \tag{2.9}$$

2.3.6 An heuristic algorithm for disaggregation elicitation with MR-Sort

The elicitation method for MR-Sort based on a mixed integer programming (previously defined in Subsection 2.3.5) provides to the user a mathematical guaranty to return all the objects in Θ in the correct category if there exists an MR-Sort model that

does so. However the computational complexity of this method generally becomes too high when the number of criteria and the size of the learning set increase [Mousseau et al., 2001]. In order to deal with this issue, [Sobrie et al., 2015] proposed method of elicitation based on genetic algorithm for MR-Sort. The disaggregation process is an evolutionary algorithm that works iteratively as follows.

This algorithm being a genetic algorithms, the idea is to use a "population" of MR-Sort parameter sets that will evolve to become a "good population" of parameter sets. At first, some initial population of parameter sets are given with a heuristic based on a linear program that is supposed to give a "better than random" population. In order to do so, we set the profiles values independently on every criterion. For each of them, we act as it is a dictator criterion (context with only one criterion). We are looking for the profiles delimiting the categories that give the best assignment accuracy i.e. while reassigning the objects of the learning set based only on the observed criterion getting the best matching with the initial assignment. Then, the algorithm iteratively proceeds in two steps until a stopping condition is met. The algorithm may stop either because a parameter set is obtained that perfectly returns the given learning set or because a predefined maximum number of iterations is exceeded. The three steps applied at each iteration are the following:

1) **A linear programming** is applied to obtain for each parameter set from the current population, the weight and the concordance index that has the best fit with the learning set. In this step the profiles limiting the categories are fixed to their last position found.

$$
\begin{aligned}
&\text{minimize} && \sum_{a \in A} x'_a + y'_a \\
&\text{subject to} && \sum_{i:g_i(a) \geq g_i(b_{h-1})} w_i + x'_a \geq s \forall a \in c_h, h = 2, ..., p-1 \\
& && \sum_{i:g_i(a) \geq g_i(b_{h-1})} w_i - y'_a \leq s \forall a \in c_h, h = 1, ..., p-2 \\
&w_i \in [0,1] && \forall i \in N \\
&\sum_{i \in N} w_i = 1 \\
&s \in [0.5, 1]
\end{aligned}
$$

2) **A genetic method** is used to obtain a better population of profiles. At this step weights and concordance index are fixed to the values that were previously given to them. This heuristic is based on genetic algorithms. It consists in proposing several changes for the profiles. A set of changes is chosen randomly. Given that

we expect each profile b_k to dominate at least weakly the profile b_{k-1}, on each criterion i, $-\delta$ or δ will be chosen between $g_i(b_{h-1})$ and $g_i(b_{h+1})$.

Then a rating of each change "adding δ to the threshold b_h on the criterion i" is made, with a score $P^{\delta}_{h,i} \in [0,1]$ regarding to "how much this change would improve the match with the learning set".

Finally, for each profile and each criterion the change with the highest score is selected. This change may be applied with a probability equal to $P^{\delta}_{h,i}$.

3) With MR-Sort model in the population we assign the objects of given in the learning set and we calculate the classification accuracy i.e. the proportion CA that was correctly assigned i.e. $CA = \frac{Number\ of assignment\ examples\ restored}{Total\ number\ of\ assignment\ examples}$. The "worst" half of the population according to CA is reinitialized as it was made at the beginning of the algorithm while the other half is used at the next step. This selection is the fitness function of this algorithm.

The loop is stopped when the classification accuracy of a model is equal to 1 or when the number of iteration exceeds a pre-defined maximum number. The model with the highest classification accuracy is chosen as the output of this algorithm.

Formally, the algorithm may be described as follows:

Algorithm 4: Metaheuristic to learn all the parameters of an MR-Sort model. Source:[Sobrie et al., 2013]

```
Generate a population of N_model models with profiles initialized with a heuristic
repeat
    for All models M in the set do
        Learn the weights and concordance index with a linear program, using
        the current profiles.
        Adjust N_it the profiles with a metaheuristic using the current weights
        and threshold.
    Reinitialize the ⌊N_model / 2⌋ models giving the worst CA
until Stopping condition is met;
```

Algorithm 5: Randomized heuristic used for improving the profiles. Source:[Sobrie et al., 2013]

```
R_≥ = ∅
for All profile b_h do
    for all criterion i chosen in random order do
        Choose, uniformly, a set of positions in the interval [g_i(b_{h-1}), g_i(b_{h+1})]
        Select the one such that P^Δ_{h,i} is maximal.
        Draw uniformly a random number r within the interval [0, 1]
        if r < P^Δ_{h,i} then
            b_h ← b_h + Δ
            Update the object assignment
```

This algorithm based on a mixture of genetic algorithm and linear programming has the main advantage of computing in a reasonable time compared to other algorithm based on mixed integer programming. The approach of using evolutionary algorithm for preference learning is, to my knowledge innovative and seems hopeful in the sense that it seems to be adaptable to every MCAP even for ranking problems. By the way, similar algorithms are also used to find the parameters of ELECTRE TRI [Doumpos et al., 2009].

2.3.7 An heuristic algorithm for disaggregation elicitation with 2-additive NCS model

The heuristic algorithm described in Subsection 2.3.6 can be adapted to 2-additive Non Compensatory Sorting model [Sobrie et al., 2015]. In order to do so the mixed integer programming that aims at finding the weights and the concordance threshold must be enhanced to integrate the weights of the $\binom{n}{2}$ pairs of criteria as follows:

$$
\begin{array}{ll}
\text{minimize} & \sum_{a \in A} x'_a + y'_a \\
\text{subject to} & \sum_{i:a_i \geq b_{h-1,i}} (m(i) + \sum_{i':a_j \geq b_{h-1,j}} m(i,j)) - x_a + x'_a = s \forall a \in c_h, h = 1, ..., p-1 \\
 & \sum_{i:a_i \geq b_{h,i}} (m(i) + \sum_{i':a_j \geq b_{h,j}} m(i,j)) - x_a + x'_a = s \quad \forall a \in c_h, h = 1, ..., p-1 \\
m(i,j) \in [-1,1] & \forall i,j \in N \\
m(i) \in [0,1] & \forall i \in N \\
s \in [0.5,1] &
\end{array}
$$

As well, the heuristic adjusting the profiles must take into account the role of the weights of pairs of criteria while defining $P^{\delta}_{h,i}$.

2.3.8 ORCLASS: A tool to interact with the decision maker

The previously mentioned methods assume that the decision maker already answered to which category she would assign each object of a subset $\Theta \in A$. Thus the choice of which Θ is going to be presented to the decision maker is supposed to be made before this process starts. Conversely, ORCLASS [Pinheiro et al., 2014] is a multi-criteria sorting method that includes an interaction procedure to find the appropriate learning set to ask. It does not use a MCAP but instead it is based on monotonicity. The output assignment assigns an object a in the category c if a dominates and is dominated by objects assigned by the decision maker to category c. This method is generally used in sorting contexts with two categories, we will thereafter consider that it is the case here, category 0 being the bad one and category 1 good one. ORCLASS works as a step by step process in which at each step the decision maker get asked to assign a new object to a category. At each step some object are fixed to a category either because they were assigned to it by the decision maker or because they dominate (resp. are dominated) an object that was assigned by the decision maker in category 1 (resp. 0). In order to make this process as short as possible, the main idea is to propose at each step the object that guaranties regardless to the category in which it may be assigned the highest number of new fixed objects. For instance, let us assume that one is dealing with a simple sorting problem with two criteria expressed in discrete scales of 5 value levels to be maximized and two categories. Initially no object is assigned by the decision maker. On table 2.12 each cell represents an object and the number at the left (resp. right) represents the number of new fixed objects if the decision maker assigns the object to category 0 (resp. 1). As we can see it is interesting to ask the decision maker to assign the object (3,3) that regardless to the category in which it could be assigned will make 9 objects fixed. Let us assume that the decision maker assigns this object to the category 1. The situation would then be the one represented in table 2.13 and the decision maker will get asked to assign the object (2,2). The process is stopped when every object is fixed into a category.

		Crit 2				
		1	2	3	4	5
Crit 1	1	1;25	2;20	3;15	4;10	5;5
	2	2;20	4;16	6;12	8;8	10;4
	3	3;15	6;12	9;9	12;6	15;3
	4	4;10	8;8	12;6	16;4	20;2
	5	5;5	10;4	15;3	20;2	25;1

Figure 2.12: Illustration of the functioning of ORCLASS algorithm. Step 1

		Crit 2				
		1	2	3	4	5
Crit 1	1	1;16	2;11	3;6	4;4	5;2
	2	2;11	4;7	6;3	8;2	10;1
	3	3;6	6;3	Cat 1	Cat 1	Cat 1
	4	4;4	8;2	Cat 1	Cat 1	Cat 1
	5	5;2	10;1	Cat 1	Cat 1	Cat 1

Figure 2.13: Illustration of the functioning of ORCLASS algorithm. Step 2

As a good point, we can mention that the decision maker may be confident with the result given that any object is sorted either directly by him or as a direct consequence of monotonicity. Then while eliciting the preferences of the decision maker it is possible to show some eventual inconsistencies with the monotonicity principle. However this method is mainly adapted to sorting problems with two categories since with more categories at the first step it is impossible to choose the object to present to the decision maker given that, for any object, in the worst case, if the decision chooses to assign it to a category that isn't the worst nor the best one, then no other object is fixed. With a higher dimension problem this process could take a long time to be applied until every object is fixed. To illustrate this we could imagine a problem with two criteria, both represented on a scale of k value levels. At the first step the most interesting object to be presented to the decision maker is the $(\frac{k}{2}, \frac{k}{2})$ object which in any case will fix a quarter of the objects. Now let us imagine a similar problem with 5 criteria. Then again the most interesting object to present to the decision maker is the $(\frac{k}{2}, \frac{k}{2}, \frac{k}{2}, \frac{k}{2}, \frac{k}{2})$ object that will fix only a 32^{th} of the objects. Indeed, the more criteria there are the less frequent is dominance between pairs of objects. Furthermore, considering that a similar proportion of the remaining objects may be fixed at each step, the required number of step should be proportional to the logarithm of the total number of objects, thus increasing with it. From a computational complexity point of

view, it may seem very quick but one should keep in mind that an interaction with a decision maker is way longer than an automatic operation performed by a computer.

2.3.9 DRSA: A disaggregation elicitation algorithm based on rule based principle

A rough set is a formal approximation of a conventional set in terms of a pair of sets which give the lower approximation (objects that surely belong to the original set) and the upper approximation (object that may belong to it) of the original set. Rough set theory is generally use in context in which the attributes are not ordered (contrarily to criteria).

In particular, rough set theory provides a methodology to complete incomplete data sets. Indeed, it includes a methodology to create by inference, sets of rules that aims at recovering a missing attribute (generally called decision attribute).

In practice, in order to create the lower approximation (resp. the upper approximation) of each set of object $X \subseteq N$ denoted by $\underline{P}(X)$ (resp. $\overline{P}(X)$) an equivalence relation I is created such that aIb if $g_i(a) = g_i(b), \forall i \in N$ and for each object a the set of similar objects $I(a) = \{b \in A/aIb\}$. Then, the lower approximation of $X \subseteq N$ is the set of object a such that any object b equivalent to a belongs to X (formally $\underline{P}(X) = \{a \in A/I(a) \subseteq X\}$). The upper approximation of $X \subseteq N$ is the set of object a such that some object b equivalent to a belongs to X (formally $\overline{P}(X) = \{a \in A/I(a) \cap X \neq \emptyset\}$).

Dominance-based rough set approach (DRSA)[Greco et al., 2001a][Błaszczyński et al., 2009] is a multi-criteria sorting method based on rough set theory. One of the main purposes is to deal with violation of monotonicity in the learning set. The key idea of DRSA is to replace the lower and upper approximations of classes by lower and upper approximations of dominating and dominated cones. We define the dominating cone (resp. dominated cone) of an object a denoted $D^+(a)$ as the set of all the objects that dominates a (resp. are dominated by a). Let us call upward (resp. downward) unions of decision category c_t denoted as $c_t^{\geq}$ (resp. $c_t^{\leq}$) the union of all the classes higher or equal (resp. lower or equal) to c_t i.e. $c_t \cup c_{t+1} \cup ... \cup c_r$ (resp. c_t i.e. $c_t \cup c_{t-1} \cup ... \cup c_1$). The lower approximation of $c_t^{\geq}, t \in T$, denoted as $\underline{P}(c_t^{\geq})$, is defined as the set of the objects that belong to $c_t^{\geq}$ and are dominated only by objects in $c_t^{\geq}$, formally:

$$\underline{P}(c_t^{\geq}) = \{a \in A \colon D_P^+(a) \subseteq c_t^{\geq}\} \tag{2.10}$$

The upper approximation of $c_t^{\geq}, t \in T$, denoted as $\overline{P}(c_t^{\geq})$, is defined as the set of the objects that belong to $c_t^{\geq}$ or are dominated by at least one object in $c_t^{\geq}$, formally:

$$\overline{P}(c_t^{\geq}) = \{x \in U \colon D_P^-(a) \cap c_t^{\geq} \neq \emptyset\} \tag{2.11}$$

Analogously, the lower and the upper approximation of $c_t^{\leq}, t \in T$ with respect to $P \subseteq C$, denoted as $\underline{P}(c_t^{\leq})$ and $\overline{P}(c_t^{\leq})$, respectively, are defined as:

$$\underline{P}(c_t^{\leq}) = \{a \in A \colon D_P^-(a) \subseteq c_t^{\leq}\} \tag{2.12}$$

$$\overline{P}(c_t^{\leq}) = \{x \in U \colon D^+(a) \cap c_t^{\leq} \neq \emptyset\} \tag{2.13}$$

These concepts are illustrated in figure 2.14.

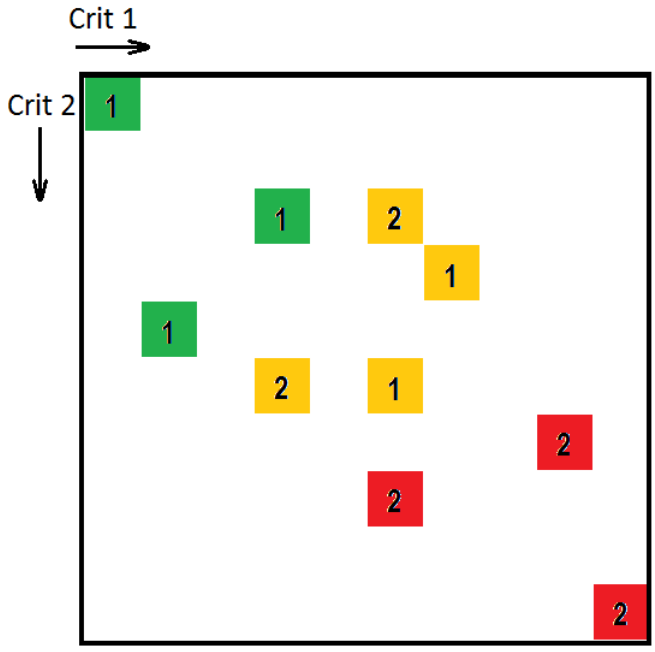

Figure 2.14: Illustration of the lower and upper approximation $c_2^{\geq}$. All the criteria are to be maximized and the category c_2 is the better than the category c_1. The numbers represent categories of the objects. In red the object that belong to $\underline{P}(c_2^{\geq})$. In yellow the objects that belong to $\overline{P}(c_2^{\geq})$. In green the objects that to not belong to $\overline{P}(c_2^{\geq})$.

The Dominance Based Rough Set Approach aims at providing a set of rules to classify all the possible object in A induced by the approximations that were obtained by

means of the dominance relations. Procedures for generating decision rules from a decision table use an inductive learning principle. We can distinguish three types of rules: certain, possible and approximate. Certain rules are generated from lower approximations of unions of categories; possible rules are generated from upper approximations of unions of categories and approximate rules are generated from boundary regions.

Certain rules has the following form:
if $g_1(a) \geq r_1$ and $g_2(a) \geq r_2$ and $\dots g_p(a) \geq r_p$ then $a \in c_t^{\geq}$, ($D_{\geq}$Decision rules).
or:
if $g_1(a) \leq r_1$ and $g_2(a) \leq r_2$ and $\dots g_p(a) \leq r_p$ then $a \in c_t^{\leq}$, ($D_{\leq}$Decision rules).

The syntax of possible rules is similar to the previous one, however the consequence part of the rule has the form: "x could belong to $c_t^{\geq}$" or the form: "x could belong to $c_t^{\leq}$". Approximate rules has the syntax can be seen as a $D_{\geq}$Decision rule and a $D_{\leq}$Decision rule.

In DRSA, DOMLEM [Greco et al., 2001b] is frequently used for rules induction. It can be seen as a greedy algorithm that iteratively add a new rule which will assign as many object as possible from the learning set in their category and a few object as possible from the learning set in the wrong category. The precise algorithm of DOMLEM is described in the Appendix A

Once a set of rules is obtained, each object a can be sorted in a category or in a set of categories representing the intersection of the approximations of the rules that cover a. For instance, if an object a is covered by two rules $E_1 =$ "if $conditions_1$ then $a \in c_2^{\geq}$ and $E_1 =$ "if $conditions_2$ then $a \in c_3^{\leq}$ then we state that $a \in c_2 \cup c_3$.

Several remarks may be done about DRSA and the DOMLEM algorithm. At first we can mention the fact that it is a model free approach in the sense that it is not really based on a MCAP. One can argue that a set of rule is a MCAP but it is a very flexible one in the sense that any assignment that respects monotonicity may be represented by a set of rules as demonstrated in Subsection 2.2.4 and thus in contexts in which dependencies between criteria or veto phenomenons may exist, it makes sense to use such methods. Then, rule based methods are a tool whose signification is easy to understand and easy to apply for decision makers and user. In contexts such as public decision making it is important that the mechanism that rule decisions are meaningful to every actor. Furthermore, the strength of DRSA is also its capacity to deal with violation of monotonicity. However, this method does not systematically assign each object to one only category and instead may assign an object to an interval of categories which in some context may be problematic. Furthermore one could argue that the will in DOMLEM to obtain a set of rules "as simple as possible" may lead to

an oversimplified set of rules that would not represent that decision makers reasoning.

2.3.10 Logistic and choquistic regression

Logistic regression [Freedman, 2009][Walker and Duncan, 1967] is a prediction technique that as other prediction techniques (rough set for example) may be used for preference elicitation. Although the following methods are based on probability theory, no randomized experience is involved in them.

These methods are based on the idea that the assignments in the learning set $\{a, f_l(a)\}$ are i.i.d (independent and identically distributed) generated by an unknown underlying probability measure $\mathbb{P}$ over $A \times C$. Let us thereafter call $f_k : A \to C$ such that $f_k(a)$ is a random variable representing the distribution of the category to which a is being assigned.

The goal here is to obtain an assignment with minimal risk, where the risk $R(f)$ of an assignment f is defined as its expected loss:

$$R(f) = \sum_{a \in A} \mathbb{E}(\varrho(f(a), f_k(a))) \tag{2.14}$$

Here ϱ is the 0/1 loss function counting the number of incorrect predictions i.e. :

$$\varrho(c_1, c_2) = \begin{cases} 0, & \text{if } c_1 = c_2 \\ 1, & \text{otherwise} \end{cases} \tag{2.15}$$

We are first going to consider the case where there are only two classes (without loss of generality we will assume that $c = 0$ aka negative category or $c = 1$ aka positive category). Then, logistic regression models the probability of the positive category and thus of the negative category as an affine function of the values of the object on the criteria. Given that a linear function does not necessarily produce values in the unit interval, the response is defined as a generalized linear model, namely in terms of the logarithm of the probability ratio:

$$log\left(\frac{\mathbb{P}(f_k(a) = 1)}{1 - (\mathbb{P}(f_k(a) = 1)}\right) = k_0 + k_1 \cdot g_1(a) + k_2 \cdot g_2(a) + ... + k_n \cdot g_n(a) \tag{2.16}$$

It is indeed a "regression" process because we want to show a dependency relationship between variable and a set of explanatory variables, although in classical regression the output data is generally continuous. We call this regression "logistic" given that the

probability distribution is modeled from a logistic distribution [Balakrishnan, 2013]. Indeed, after transformation of the above equation, we obtain:

$$\mathbb{P}(f_k(a) = 1) = \frac{1}{1 + exp(-k_0 - k_1 \cdot g_1(a) - ... - k_n \cdot g_n(a))} \tag{2.17}$$

Estimation of the parameters $k_0, k_1, ..., k_n$ is made based the learning set, by maximizing the log-likelihood function ϑ i.e. the logarithm of the probability that all the element in the learning set are well classified:

$$\vartheta(k) = \sum_{a \in \Theta} log\left(\mathbb{P}(f_k(a) = f_l(a))\right) \tag{2.18}$$

Considering the case of ordinal classification, where we are given r ordered classes idea is now to reduce the classification problem to a binary context as seen above dividing for each c_q the categories in two groups: $c_q^{\geq}$ and $c_q^{\leq}$. Thus, we state that $\pi_q(a) = \mathbb{P}(f_k(a) \leq c_q) = \frac{1}{1+exp(-k_0-k_1 \cdot g_1(a)-...-k_n \cdot g_n(a))}$.Hence:

$$\mathbb{P}(f_k(a) = c_q) = \pi_q(a) - \pi_{q-1}(a) \tag{2.19}$$

Once again the estimation of the parameters $k_0, k_1, ..., k_n$ is made based the learning set maximizing the log-likelihood function i.e. the probability that all the element in the learning set are well classified. Finally, when a new object a is given that should be sorted, we assign it to the category c for which the probability $\mathbb{P}(f_k(a) = c)$ is the highest.

This method has several variants [Tehrani et al., 2011][Labreuche et al., 2014], among which choquistic regression in which the simple weighted sum $k_0 + k_1 \cdot g_1(a) + k_2 \cdot g_2(a) + ... + k_n \cdot g_n(a)$ is replaced by a choquet integrals (presented in Subsection 2.2.3).

Conclusion

We presented in the chapter multi-criteria decision aiding as a scientific field mainly grouping a structuring of the problem and a preference elicitation and aggregation process. The creation of the *Biodiversity Severity Index* is a sorting problem. Indeed we are not interested in finding the most severe scenario of accident but instead we are attempting to characterize scenarios individually according to the expected severities of their impacts on the environment. We will call out experts in toxicology from the INERIS and in ecology from the "Muséum National d'Histoire Naturelle". The author of this document is influenced by a constructivist school of thought and the reader

may observe later in this document that during the elicitation of the experts's preferences these preferences constantly evolved during the process. We will use a hierarchy of criteria mainly based on Value Focused Thinking although we do not exactly follow Keeney's procedure. Several aggregation method will be tested in the thereafter described process but finally mainly outranking methods will be used. During the preference elicitation we will use alternately the disaggregation and the aggregation approach.

Chapter 3

Biodiversity Severity Index

Contents

3.1 Final hierarchy of the Biodiversity Severity Index 98

3.1.1 Local Biodiversity Severity Indices for surface water targets . 98

3.1.2 Destructive potential 100

3.1.3 Distinction between facts and values in our model 102

3.2 Scales on the criteria 103

3.2.1 Standard scales 103

3.2.2 Boolean values and scales 104

3.2.3 Semantically defined ordinal scales 104

3.2.4 A five value levels scale for both the *destructive potential*, the *value of the environment*, the *vulnerability of the target* and the *local biodiversity severity index* 107

3.2.5 Reference points for the value levels 108

3.2.6 Adaptation of this methodology to the impact on ground targets 110

3.2.7 Conclusion on the hierarchy of criteria 110

3.3 Construction process for the hierarchy of criteria 111

3.3.1 Adapting the Value Focused Thinking approach to our problem 111

3.3.2 First proposition from the INERIS 113

3.3.3 First modifications of the hierarchy of criteria 117

3.3.4 Interacting with experts 118

3.3.5 Second modification based on the *Muséum National d'Histoire Naturelle*'s expertise . . . 121
3.3.6 Validation of the previously obtained model with a toxicologist 122
3.3.7 Validating the destructive potential as a criterion . . . 123
3.3.8 Including the *residence time* and the concentration as new criteria . . . 124
3.3.9 Adding the *resistance* criterion by interacting with the *Muséum National d'Histoire Naturelle* . . . 126
3.3.10 From an exhaustive model to a "user friendly" one . . . 126
3.3.11 Why we stopped the evolution and accepted the hierarchy as it is. . . . 127
3.4 Aggregation to get the destructive potential . . . 129
3.4.1 Finding possible dependences between sub-criteria of the destructive potential . . . 130
3.4.2 About additive methods for the destructive potential . . . 131
3.4.3 Defining the aggregation method for the destructive potential 132
3.5 Aggregation to get the Local Biodiversity Severity Indices 135
3.5.1 Frame of this sub-problem . . . 136
3.5.2 The questionnaire . . . 138
3.5.3 Processing the data . . . 140
3.5.4 Aggregating the criteria with a MR-Sort model . . . 145
3.5.5 A MR-Sort model with vetoes . . . 148
3.6 Aggregation of the Biodiversity Severity Index . . . 155
3.6.1 Description of the aggregation procedure . . . 156
3.6.2 Practical example of the evaluation of a scenario of accidental pollution . . . 157

"Misura ciò che è misurabile, e rendi misurabile ciò che non lo è.", Galileo Galilei

An important purpose of this thesis is the construction of an index to represent the expected severity of scenarios of accidental pollution on biodiversity in our case toxic leak. This indicator should be expressed on a finite discrete scale of 5 value levels in order to make it usable in risk matrices similar to those used to deal with risks on humans as presented in Subsection 1.2.2. We previously defined the context of this problem and presented some methodological tools that we used to face this problem, namely multi-criteria decision aiding with among others a particular attention given to Bernard Roy and Ralph Keeney's works [Roy, 1978][Keeney, 1992]. We are now about to describe the process that we followed and the model that we created. We will first describe the final state of the hierarchy before explaining its construction. Indeed, we think that this last part will be easier to understand for the reader once the definition of all the criteria will be given. After the final state of the hierarchy was described as well as the way that we found it, we will describe the process that was used to find the aggregation methods to aggregate each sub-problem.

3.1 Final hierarchy of the Biodiversity Severity Index

This section aims at describing the final state of the hierarchy of criteria with few precisions on how this hierarchy was created. What we should mention yet is that this hierarchy of criteria was built with repeated interactions with two types of experts; a toxicologist from the INERIS and several experts from the Muséum National d'Histoire Naturelle (MNHN) in Paris. The methodology used to build this hierarchy, being an important point, will be treated separately in Section 3.3.

3.1.1 Local Biodiversity Severity Indices for surface water targets

The Biodiversity Severity Index will be decomposed into two indices as illustrated in Figure 3.1: the Biodiversity Severity on Surface Water Index and the Biodiversity Severity on the ground Index. We chose to make that division because the biodiversities that lie in these environments are different. Water pollution and ground pollution is generally seen as two separate topics (for instance they are treated in different books by the "Cahiers d'Habitat Natura 2000" [Bensettiti et al., 2002]).

We will later call a target a specific geographical section of the territory on which

we decide to examine the impact of the scenario that is being considered. The area surrounding the industrial plant should be divided into several connected areas globally homogeneous from a biological point of view and in which every points could be affected in a same way if the scenario would occur. In order to make each of these leaks comparable, the target should broadly cover the same surface. We first focus on assessing the impact on biodiversity on each specific target before evaluating the global severity of a leak. This choice was also validated during the interview with the experts. The division of the surrounding area into several target is not treated in this document. The application of our method require this task to be made previously.

Thus, both the *biodiversity severity on surface water index* and the *biodiversity severity on the ground index* are divided into as many indices as there are such targets that need to be taken into account. These indices will be called Local Biodiversity Severity Indices (LBSI) on either surface water targets and on ground targets. The part of the hierarchy that covers these indices is represented by Figure 3.2.

The Local Biodiversity Severity Indices must represent the severity of the impact on one given target. Each of them must answer the question "How unfortunate would the effect of that leak be for the biodiversity of this target?".

The following description will only deal with biodiversity severity on surface water. We chose not to consider the air as a target because, based on the ARIA database and on interviews with experts, we know that it is very unlikely that a toxic leak has a significant impact on the biodiversity living in the air. After this description will be made, we will see in Subsection 3.2.6 how this evaluation is adapted to calculate the *biodiversity severity on the ground indices*.

We consider that this criterion is influenced through specification by three criteria:

- The **value of the environment**: We will call *value of the environment* the importance that man gives to the biodiversity which is present on the specific target. This topic is described in more details in Section 1.3. This data should be given by a potential user of this methodology as an input of our methodology. We strongly suggest the user of this method to seek advice to an expert in ecology to provide this information. In this document we provided some reference points about the value of the environment (in Subsection 3.2.5) that were proposed with the help of an expert in ecology from the Muséum National d'Histoire Naturelle (MNHN). Therefore, we believe that it is reasonable to think that this information can be provided by the user of this method if she is supported by an expert in ecology.

- The **vulnerability of the target**: We here consider the vulnerability of the environment as the ability of a target to be impacted on the long term by a toxic leak. We consider that this criteria influenced by the resilience (ability to recover after being impacted) and the resistance of the considered target (these concepts are defined with more detail in Subsection 3.3.5). Once again, this data should be given by the user as an input of our methodology. We strongly suggest the user of this method to seek advice to an expert in ecology to provide this information. In this document we provided some reference points about vulnerability (in Subsection 3.2.5) that were proposed with the help of an expert in ecology from the MNHN. Hence, it seems reasonable to think that this information can be provided by the user of this method if she is supported by an expert in ecology.

- The **destructive potential**: This criterion is more sophisticated than the previous two. Indeed, in this model it is not an input of our methodology but instead it is obtain by an aggregation of other criteria. We are about to describe it in more details.

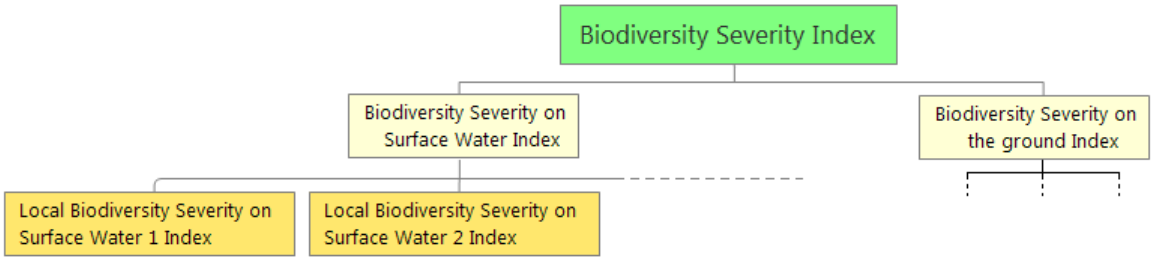

Figure 3.1: Hierarchy of criteria for the Biodiversity Severity Index (upper part of the hierarchy)

3.1.2 Destructive potential

The goal of the *destructive potential* is to represent the capacity of a toxic leak to impact a target. The *destructive potential* can be obtained aggregating three criteria:

- The **concentration** of the product in the target's water: The *concentration* is defined by the Oxford dictionaries as "The relative amount of a particular substance contained within a solution or mixture or in a particular volume of space". It is influenced by the receiving volume, i.e. the volume of water initially

contained in the target and by the volume of product that gets to the target. This criterion should be given as an input of this process. When the water is static this concentration can be easily approximated. With moving water this evaluation may be more difficult and probably requires the use of fluid mechanics or the creation of a set of rule to approximate it.

- The **toxicity** of the liquid: According to the Collins dictionary the word toxicity refers to the degree of strength of a poison. We will use this definition in the following of this thesis. The *toxicity* can either be physical or chemical. However, in practice the expert in toxicology told us that it is very unlikely for a product to be toxic both chemically and physically. Thus, no aggregation is here necessary.
- The **residence time** of the product in the environment: This criterion will here be defined as the time that the product of the leak keeps having an impact on the local biodiversity. It depends on the mobility of the water and on the persistence of the product. A product will be defined as persistent if its negative impact on the environment does not stop by itself.

Input criteria for the local biodiversity indices

Each *local biodiversity severity index* will be calculated from 6 input criteria as shown in Figure 3.2: the *toxicity* of the product, the expected *concentration* of the product, the *persistence* of the product, the *mobility* of water, the *vulnerability* of the target and the *value of the environment* in the target. Obtaining of the value of these criteria is not part of this model. The user should provide them with the use of appropriate data and methodologies.

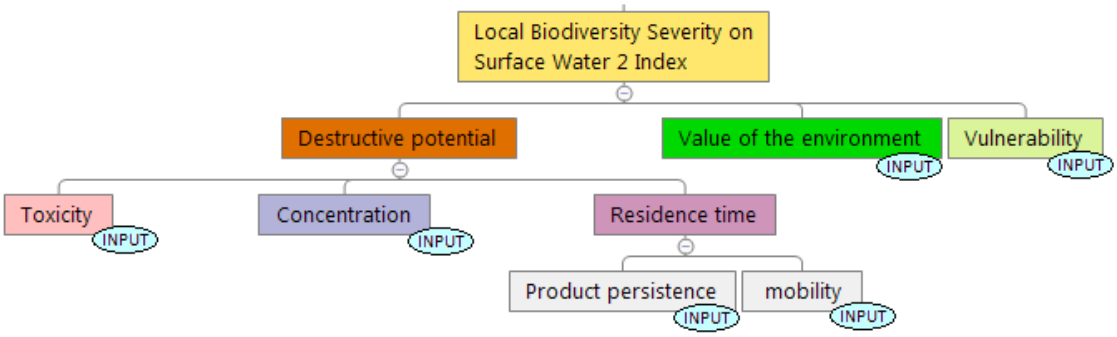

Figure 3.2: Hierarchy of criteria for the local Biodiversity Severity Index (biodiversity severity on surface water)

3.1.3 Distinction between facts and values in our model

We decided to make the distinction between facts and values because we believe that having a good understanding of what we are measuring, how univocal or subject to judgement it is, will help us in the following, in particular while choosing the scales, as we will see in Subsection 3.2.4. As explained in Subsection 3.3.4 this work is mainly based on predicting the consequences of a scenario of accidental pollution. Thus, this evaluation does more look like a fact than a value. Let us try to differentiate what are the facts and what are the values. As defined in Subsection 2.1.6.1, in this document we use Keeney's terminology [Keeney, 1992] which refers to values as an evaluation that depends on the decision makers value system (what matter to him, and how much it does). On the other hand she call a fact an expertise, a factual data or a prediction that does not depends of the expert's value system. At first we will try to determinate if the evaluation of the biodiversity severity of an accident on a target is a fact. In order to understand this, we should ask ourselves "If two experts clearly disagree on their evaluation of the severity of a scenario of accidental pollution on the biodiversity or on the comparison of such events, would that mean that at least one of them is wrong?". I think that the answer to this question is "No". Indeed, the value that we give to the environment is a subjective notion. Two experts could disagree on the importance. Thus, the value that we give to its deterioration is also subjective. All the criteria that are located over this criterion in the hierarchy of criteria must also be considered as values given that they are at least partially created by values. The same question can be formulated with the *destructive potential*. As explained later, the *destructive potential* can be considered as fact. Indeed according to the toxicologist that we interviewed, it is very unlikely that two toxicologists could have two very different judgements about the *destructive potential*.

To summarize this paragraph, as illustrated on Figure 3.3, we can say that the *destructive potential*, the *concentration*, the *product persistence*, the *residence time*, the *mobility*, the *toxicity* and the *vulnerability* are facts while the *value of the environment*, the *local biodiversity severity indices* and all the criteria that are located above are values.

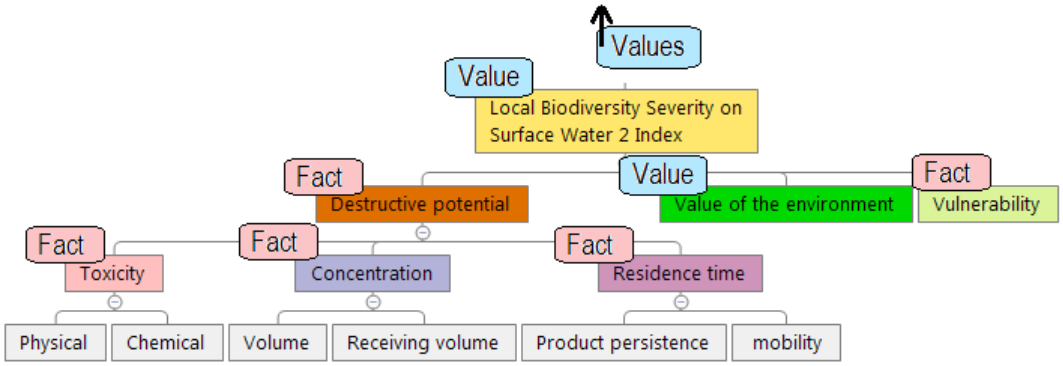

Figure 3.3: Distinction between facts and values in the hierarchy of criteria

3.2 Scales on the criteria

In a hierarchy of criteria, each criterion must be represented on a precise scale and this scale must be understood in the same way by every actor in the process. This section aims at explaining the issue of choosing the scale for each criterion, the choices that we made and the reasoning that led us to them. Thereafter we will call a scale on a criterion the mean to represent the different values that an object (here a toxic leak) can take on this criterion. It is to be mentioned that the signification of these scales is generally only ordinal in the sense that the attractiveness of a value is not supposed to be proportional to its numerical value. The experts were associated to the choice of the scales and they finally validate them.

Given that we are dealing with a hierarchy of criteria, the output of a sub-problem may be the input of an other one. Thus, in the following section, while talking about scales and value level, considering a specific sub-problem, we will refer either to the input criteria and to the output of the local sorting. In this hierarchy of criteria we can divide the scales that we will use in three categories. In the following, given that we created a hierarchy of criteria in which the output of a sub-problem can be the input of an other one, we may use the term value level (used for the input scales) where the term categories might also be appropriate.

3.2.1 Standard scales

Some criteria whose values are measurable unambigusly such as the *concentration* and the *toxicity*. These criteria are generally expressed on standardized scales such as

cubic meters or Celsius degree. These scales have the advantage of being scientifically justifiable, easily accepted by the actors and unequivocal i.e. commonly understood by the different actors. The values on these criteria should be calculated with the help of the appropriate scientific field. These measurable scales will mainly be used to measure facts rather than values. Thus, to represent the *concentration* of liquid we chose the commonly used mass concentration unit, the gramme per litre although the unit imposed by the *International System of Units* is the $\mathrm{kg\,m^{-2}}$. In order to represent the *toxicity* we chose the maximum acceptable toxicant concentration. The maximum acceptable toxicant concentration (MATC) is a measure used in chronic toxicology [Verma et al., 1981] that defines for each product the bound of concentration under which biodiversity will not be impacted. The higher the acceptable concentration is, the less toxic it is.

3.2.2 Boolean values and scales

Boolean scales could be described as ordinal scales with two value levels. There are generally used to describe the state of truth of a proposition. For instance while evaluating houses, the existence or not of the parking place can be seen as a criterion. But we could also define as a boolean scale every scale for which the possible values are approximated to two mainly obtained features. In our context, concerning the *residence time*, the *mobility* and the *persistence*, as mentioned in Subsection 3.3.8 we chose to simplify the scale to a boolean variable.

3.2.3 Semantically defined ordinal scales

Values may also be measured through what Keeney calls "constructed scales" [Keeney, 1992] (page 146). In our case it seems most appropriate to use ordinal discrete scales made of a finite set of value levels for most of the criteria in the hierarchy (example: the *destructive potential*, the *vulnerability* etc). We believe that it is important to give to each value level an intuitive definition created with semantic that the decision maker or expert feels comfortable with. Concerning the way to create an appropriate scale that may be understandable and easy to use for humans, we obtained some interesting elements from psychometrics [Hershey and Taiwo, 1999] where the creation of questionnaires is studied in which a subject has to pick an answer among a predefined set.

An important issue concerns the number of value levels in the scales. A higher number

of value levels is supposed to make the description of the object more accurate which is a good property as explained in [Warren, 1973]. Nevertheless, the higher the number of value levels is, the more likely it is that these value levels will not be distinct enough to allow the decision maker or expert to definitely distinguish the values of these value levels [Hershey and Taiwo, 1999]. Indeed, she could have difficulties to assign an object with confidence to one precise value level rather than to an adjacent one.

As argued in MACBETH [Bana e Costa et al., 2003] the existence of a neutral element can in many cases be meaningful to the decision maker or expert. In a *bipolar scale* [Öztürk and Tsoukiàs, 2008], this neutral value level will be located at the center of two areas the "good performance area" containing all the "good categories" with different levels of "goodness" and conversely the "bad performance area". The different levels are intuitively described with quantity adverbs such as "rather", "very" or "extremely". The number of these adverbs being obviously finite and relatively low, these scales will then be discrete scales. It seems quite intuitive for the two areas (the good and bad performance areas) to be symmetrically created although [Worcester and Burns, 1975] demonstrated that the same adverbs used before a positive and a negative adjectives could be evaluated differently. Furthermore [Brown et al., 1973] showed that an unbalanced scale may influence the decision maker's answer. Thus, we chose to turn to odd numbers of value levels. From a computational point of view, depending on the elicitation algorithm that will be used, a higher number of value levels in the scales might increase the computational complexity. According to the previously explained constraints we chose to focus more specifically on the discrete scales with 3, 5 and 7 value levels as recommended by [Lehmann and Hulbert, 1972].

As we will later see, these value levels being intuitively defined, their meaning can be seen as subjective and they can initially be understood differently by the actors of the process. This is why it is important to have a proper discussion with the actors to make the definition of each value level commonly interpreted. This can be done giving, for each value level on each criterion, examples of objects familiar to the decision makers or experts that have this value level on this criterion.

Three value levels scales

The type of three value levels scale that seems the most intuitive and legitimate according to the previously mentioned reasoning is the one with the *bad performance*, the *average performance* and the *good performance* value levels. The scale can be adapted to some contexts the different levels of good and bad performance are hardly discernible. According to [Hershey and Taiwo, 1999] three value levels scales are also

appropriate when evaluating low involvement topics. However, in many cases, when dealing with more important issues, the different levels of good and bad are distinguishable by the decision maker or expert, this difference matters and do have an impact on decisions or judgement. In these cases it is preferable to choose a higher number of value levels in the scales.

Five value levels scales

The type of five value levels scale that seems the most appropriate in our context are the *very bad performance*, *rather bad performance*, *neutral performance*, *rather good performance*, *very good performance* scales. These scales have the advantage of being pretty intuitive and covering quite well the possible judgements on the concerned criteria. Then, this five value levels scale, compared to scales with more value levels has the advantage that every value level is distinct so that it seems unlikely that a scenario could be assigned to two different value levels by a decision maker.

Seven value levels scales

As explained earlier a higher number of value levels allows a higher accuracy in the evaluation, it contains more information. In our context we can see two interesting types of seven value levels scales. The first one could be similar to the previously described five value levels scale, namely: the *very bad performance*, the *bad performance*, the *rather bad performance*, the *average performance*, the *rather good performance*, the *good performance* and the *very good performance*. This type of scale has the advantage to be rather intuitive to decision makers covers well the possible judgements. However in this type of scale, the value levels might not be distinct in the decision maker's mind and she could have difficulties to assign an object to one specific value level as explained by [Hershey and Taiwo, 1999]. Thus, the use of this value level would be more appropriate in an elicitation method that would allow the decision maker or expert to assign objects to intervals of value levels (for example "Object a is sorted between category C2 and C4").

The second one would be a kind of five value levels scale as defined previously with additional extreme value levels in the "good and bad performance areas". These extreme value level would not have a subjective meaning. For example in an other context, for a price criterion these criteria could be *for free* and *you cannot pay it*. In our context it is not clear what could be the meaning of such value levels nor what contribution they could bring that cannot be expressed by the *very good performance* and *very bad performance value levels*.

3.2.4 A five value levels scale for both the *destructive potential*, the *value of the environment*, the *vulnerability of the target* and the *local biodiversity severity index*

Concerning the *destructive potential* and the *vulnerability of the target*, although we consider that they are facts, to our knowledge there does not exist a standard scale to measure them. Given that most of the elicitation methods that are available for sorting problems require the decision maker to assign objects to one precise value level rather than to an interval of value levels and according to the previous observations we considered that the five value levels scales are the most appropriate for both the *destructive potential*, the *value of the environment*, the *vulnerability of the target* and the *local biodiversity severity index*. We expressed the *destructive potential* as the strength of the expected consequences of a scenario on an averagely resilient target.

Thus, the five value levels that we defined for every criteria are described in the Table 3.1.

	value level C1	**value level C2**	**value level C3**	**value level C4**	**value level C5**
Destructive potential	No impact on biodiversity	Weak impact on biodiversity	Relatively large impact on biodiversity	Large impact on biodiversity	Total annihilation of the local biodiversity
Value of the environment	Very low value	Rather low value	Medium value	Rather high value	Very high value
Vulnerability	Very low vulnerability	Rather low vulnerability	Medium vulnerability	Rather high vulnerability	Very high vulnerability
Local Biodiversity Severity Indices	No impact or negligible pollution	Low pollution	Medium pollution	Serious pollution	Ecological disaster

Table 3.1: Table defining the scales used for the *destructive potential*, the *value of the environment*, the *vulnerability of the target* and the *local biodiversity severity indices*

The reader may notice that concerning the *destructive potential* and the *local biodiversity severity indices*, the scale that is used was not created symmetrically with a neutral element at the center. Indeed, it is very unlikely for a toxic leak to have positive consequences on the environment and thus, due to there definition these two

criteria only represent some kind of *bad performance*. What could be seen as a neutral element is in both case located in value level one.

3.2.5 Reference points for the value levels

As we mentioned earlier, several criteria are expressed on discrete semantically defined ordinal scales and these scales have an intuitive meaning that may lead to some difficulties. Indeed, the experts may interpret differently some subjective degree of strength expressed through the use of adverbs like "rather" ("rather low vulnerability"), even a single expert may not have a clear understanding of their meaning. In order to face this issue we decided to fix some reference points to the value levels. What will call a reference point is a concrete example of an environmental target, a destructive potential or a scenario of accidental pollution that reach the studied value level on the studied criterion. For instance, while fixing reference points on the *value of the environment* we agreed with the experts that a target that would be located in an "important area for the conservation of birds"(*ZICO*) [1] should be considered as "rather high value" on the criterion *Value of the environment*. This way when an expert faces a scenario impacting a target considered as having a "Rather high value" she may refer a *ZICO* area. As well while applying this method, when the experts that will collect the data from the ground will have to choose the value that the value of the environment of a given target impacted by a scenario, they may also compare this target to the reference points of the *value of the environment*.

Reference points for the *Local Biodiversity Severity Indices*

In order to reach an agreement the reference points for the *Local Biodiversity Severity Indices* we proposed to the experts a set of 3 scenarios of accidental pollutions and the description of a real accidental pollution happening in 2015. We asked the experts if they could assign each of these scenarios to a value level *Local Biodiversity Severity Indices* according to the definitions given in Table 3.1. The reader will find in Appendix B.1 the assignments provided by the experts and the description of the scenarios. No reference point was found in value level "No impact or negligible pollution" as the meaning seems easy to understand.

Reference points for the *Vulnerability*

[1] Personal translation from the french "Zone importante pour la conservation des oiseaux"

In order to find some reference point for the vulnerability we interacted with the experts and it seemed that they could easily find some reference points for the value levels $C1$ and $C5$. It appeared that wastelands recover very quickly their original condition. We decided to chose them as a reference point for the value level $C1$, "Very low vulnerability". At the opposite that primary forests, coral reef and disconnected islands are generally very vulnerable areas. Thus, we decided to use them as reference points for the value level $C5$ "very high vulnerability". In the middle, the experts thought that temperate forest may be considered as a reference point for the value level $C3$ "medium vulnerability". The value levels $C2$ and $C4$ do not have a reference point. However, the experts involved in the process can evaluate at a value level $C2$ (resp. $C4$) a target that would be less vulnerable (resp. more vulnerable) than $C3$ but more vulnerable (resp. less vulnerable) that $C1$ (resp. $C5$). An idea could be to choose a 3 value level scale to represent this criterion, however having five value levels does not bring any additional difficulties and may be useful to represent intermediate values on this criterion.

Reference points for the *value of the environment*

While looking for some appropriate reference point for the criterion *Value of the environment* we used the BIOMOS indicator which was developed by the "Direction régionale de l'équipement d'Île de France". This indicator represents a "level of habitats " that reflects the potential spaces or habitats present on the territory, biodiversity home run. The methodology used to calculate this indicator is described in [Liénart, 2009]. The BIOMOS indicator gives a value to areas according to their ordinary and remarkable biodiversity. The area hosting a remarkable biodiversity has a score between 1 and 4 ($\{1, 2$ and $4\}$) while the areas hosting only an ordinary biodiversity are scored between 0.1 and 0.8 ($\{0.1, 0.3, 0.6$ and $0.8\}$). We presented some examples of types of areas with their associated scores according to the BIOMOS indicator and asked the experts if this score seems representative (at least from an ordinal point of view). It seemed that the experts globally agreed on these scores. Thus, we decided to find thresholds to separate the value levels so that these two scores can be transformed into one single indicator (from 0.1 to 0.8 if there is no remarkable biodiversity and from 1 to 4 otherwise) that can be transformed into our five value levels scales used to represent the *Value of the environment*. It appeared that a target that has no remarkable biodiversity cannot have a *value of the environment* strictly higher than $C3$ (but can have a *value of the environment* of $C3$). Then asking them to assign some types of area into value levels we came to the conclusion that the limits of the value levels are 0.5 (between $C1$ and $C2$), 0.7 (between $C2$ and $C3$), 0.9 (between $C3$ and $C4$) and 3 (between $C4$ and $C5$). The types of area that were shown to the experts

are presented along with their BIOMOS indicators and their associated value levels in the Appendix B.2.

While looking for documentation to evaluate the *vulnerability* and the *value of the environment* the user may also have a look at the "cahiers d'habitats Natura 2000" [Bensettiti et al., 2001] [Bensettiti et al., 2004a] [Bensettiti et al., 2002] [Bensettiti et al., 2005] [Bensettiti et al., 2004b] (in French), written by the *Muséum National d'Histoire Naturelle* (MNHN), that list the different biological habitats that can be observed in France and propose a synthesis for each of them in the form of a sheets. At the end of each of these sheet, a paragraph untitled *biological and ecological value* deals with biological evaluation. Although it is not presented as a single indicator this short paragraph may be useful to the user when performing a risk study.

3.2.6 Adaptation of this methodology to the impact on ground targets

The hierarchy of criteria that we just described is not adapted to evaluate the severity on ground targets. Indeed the receiving volume and the mobility of water are data that are not meaningful in this context. As a simple proposition in order to be able to deal with lack of receiving volume, the toxicologist proposed to make an equivalence of volume of water for each meter square. The interest of the mobility of the receiving water while evaluating the severity on a surface water target is that a mobile water can evacuate the product out of the target. On a ground target this phenomenon does not exist. Thus, a ground target will always be considered as not mobile. Therefore, these solutions could allow the user to evaluate the biodiversity severity on a ground target.

3.2.7 Conclusion on the hierarchy of criteria

We built a hierarchy of criteria that aims at providing a global indicator representing the expected strength of the impact of a scenario of accidental pollution on the surrounding biodiversity. For each studied target of surface water the user will have to provide an evaluation of the scenario on six criteria; the expected *concentration* of the product in the target, the *maximum acceptable concentration* of the product, the *product persistence*, the *mobility of water*, the *value of the environment* on the considered target and the *vulnerability* of the biodiversity. Among these criteria the user may have some difficulties to provided precise information on the expected concentration of product in the target. She may use some software for mechanics of fluids to deal

with this issue or create a set of rule to provide an approximation. We did not neither provide a precise methodology to evaluate the criteria *value of the environment* and *vulnerability* although we indicated a direction by giving reference points that were based on interactions with an expert in ecology and on a document created by the "Direction régionale de l'équipement d'Île-de-France".

We were helped by experts while building this hierarchy and it was finally validated by them. For sure, other hierarchies of criteria could have been built to deal with the same problem. However, as the reader is about to see, we built this hierarchy using a methodology that we think is adapted to the problem together with experts from disciplines that are related to this topic.

3.3 Construction process for the hierarchy of criteria

In the previous section, we presented the hierarchy of criteria that will allow the user to obtain a synthetic value of the environmental impact of a toxic leak. We are now about to describe the methodical reasoning and process that allowed us to build to it. We will see that this issue had already been studied by the INERIS and a method was proposed. We will explain why it is useful to continue this work. We used a constructive approach and this process included repeated interactions with various experts as explained in Subsection 3.3.4 that led us to successive modifications. The reasoning that we followed was made with the help of various stakeholders and was mainly inspired by the Value Focused Thinking [Keeney, 1992] approach. Nevertheless, some differences with Value Focused Thinking must be mentioned.

3.3.1 Adapting the Value Focused Thinking approach to our problem

The methodology that we used in this document is widely inspired by the Value Focused Thinking approach [Keeney, 1992]. Indeed, it seems important to the author of this thesis to have real understanding of the values of the decision maker, to understand and formalise the role of each criterion. As well, using a hierarchy of criteria seems quite adapted to our problem for mainly two reasons. At first, our problem is made of several sub-problems that require the expertise of different experts (in particular toxicologists and ecologists). Then for an expert, considering too many criteria sumltaneously may

be a too important congnitive load for the experts. The hierarchy of criteria here allows us to split our problem in several smaller (with less criteria) sub-problems that can be treated separately by the relevant experts. Value Focused Thinking [Keeney, 1992] provides several methodological tools that aim at dealing with such problems. However, for several reasons saying that we properly followed this methods may be considered as abusive. Indeed, we are dealing with an assessment problem while the Value Focused Thinking approach seems mainly suitable to decision problems. In Value Focused Thinking the objects that are evaluated are actions that the user may apply to improve the state of the world according to her perspective. As well, an important place is given to the concept of objectives i.e. changes that would be welcome to the decision maker. The Value Focused Thinking also aims at finding new alternatives which in our case is not an issue. If we would have used the Value Focused Thinking approach properly to deal with our problem, this all process would have aimed at measuring the attractiveness of possible alternatives available to an industry, the government or any other actor to improve the overall situation according to this actor. It is likely that this network would have been different according to the actor that is considered. These possible models would have helped the user to find possible actions in order to reach a certain number of objectives that must be measured and expressed through attributes. This more general framework should have taken in consideration objectives relative to the cost of the possible measures or their socio-economical impact. It is not sure that these possible frameworks would have led to the creation of the Biodiversity Severity Index as an objective in the fundamental or mean-ends networks. However, the creation of this indicator is needed for reasons (legal, working habits) that are external to the frame of this research. Furthermore, in the Value Focused Thinking approach, the user may create two networks, the *mean-ends objective network* and the *fundamental objective network* that are then combined in a hybrid objective hierarchy. A large portion of the mean-ends objective network is composed by actions ("maintain vehicles properly", "educate about safety"...). This does not fit to our problem. However, we could not neither simply abandon the mean-ends objective network. Although what we are evaluating is not an action, evaluation of accident has this common particularity with action evaluations, that their consequences may only been observed in the future. Now they can only be approximately foreseen through evaluation of causal factors. Then, in our case we directly built the global network without building the mean-ends objective network and the fundamental objective network. Finally we would like to use a more flexible version of this tool that would allow us to use not fully compensatory approaches (like ELECTRE TRI for example), while as stated in the discussion in Value Focused Thinking the approach concerning the aggregation is mainly utilitarian. Indeed in a context of environmental assessment issues it seems, as mentioned in Subsection 1.4.5, that a totally compensatory approach may not be

appropriate for the all hierarchy of criteria.

Here is a list of some points on which our methodology will differ.

- We do not use the concept of objective which we think is not adapted to our problem. Instead we will use the term criterion to represent the information on which we base our judgement. In Value Focused Thinking, Keeney uses the word *attribute* to talk this concept but the word *criterion* more specifically refers to an ordered information.
- We directly build a final hierarchy without creating two hierarchies (the fundamental objective network and the mean-ends objective network) and mix them together.
- We keep the possibility to use any MCAP at the different sub-problems rather than fix our choice to use a utilitarian method.

However, we were widely inspired by the Value Focused Thinking methodology, for instance:

- This work is based on the idea that a careful definition of the criteria and a clear understanding of why we are interested in these criteria is a major issue in multi-criteria decision and sorting problems.
- We will use a hierarchy of criteria as a tool to represent, understand, elicit and aggregate the criteria of the problem.
- The creation of the hierarchy of criteria and the interactions that helped us to build it are mainly based on the philosophy and the advises provided in Value Focused Thinking for instance the questions asked to the experts to find new criteria or to find the sub-criteria of a criterion.

3.3.2 First proposition from the INERIS

Méthode d'évaluation de la gravité des conséquences environnementales d'un accident industriel [Roux, 2013] was a first daft made by Pierre Roux and then developed by Christophe Duval, both engineers at the INERIS, to represent the expected severity of a scenario of accident from an environmental point of view (both biodiversity and uses). After having presented statistics about industrial accidents in France, the author

makes the statement that today, risks studies only take into account the possible short term consequences on human lives, for mainly two reasons. At first, even taking into account only these consequences, risks studies are long and costly. Then, the current level of knowledge on the mechanisms of pollution and impact assessment models is insufficient. In order to evaluate the severity of effects of a given scenario of accident, two questions are asked:

- What is the risk for a potential pollution caused by a scenario of accident to impact the studied target?
- How vulnerable is the studied target?

In order to make this evaluation the author proposed a framework in which each scenario is evaluated on each target (visual description in Figure 3.4). Then, the global severity of the scenario is determined by the target on which the effect is the worst. So as to determine the severity of a scenario on a target, three scores are created: the source module, the transfer module and the target module. The source module and the target module are expressed on a scale of 0 to 100. The transfer is made to decrease the global score it is expressed on a scale from 0 to 1. The product of these scores will be global score that is considered as the assessment of severity of the impact on the target. Here is how these modules are defined:

- **Source module**: The source module aims at representing the destructive potential of the leak once it is out of the industrial plant. It is obtained by a cross tabulation on the volume of liquid that would escape from its original location and the type of product of the leak (see the Table of the source score on biodiversity severity on Table 3.2). Two tables are provided according to the type of concern that is threatened: one table valuing the biological target according to the expected importance of the biodiversity that is located on it and one for the use importance (mainly direct use) of resources on the considered target. It is important to notice that in case of natural resources the released volume is not taken into account, following the principle that if the considered liquid is released even in a small quantity, then the good will not be used by the society following a precaution principle.
- **Transfer module**: The first goal of the transfer module is to enumerate the possibly impacted targets. It was created to represent the possible presence of obstacles that would avoid the liquid of the leak to get to the considered target or that would slow the movement of the liquid down, allowing a human intervention

to limit the impact. This module may also express a diminution of the score obtained in the source module due to the existence of some obstacle that might not avoid all the liquid to reach the target but at least should reduce the volume reaching it. The final result of this module can be considered as a score where the conclusion "this target cannot be impacted" is represented by the score 0 and a *division of the source module source score by 10* as a score of 10^{-1} (see Table 3.3).

- **Target module**: The target module aims at evaluating the environmental importance of the considered targets in the affected perimeter delimited in the previous module. This evaluation is made on a scale from 1 to 100. This assessment is specific to each target and is based on occupation territories and the presence of natural resources used by humans as seen in Table 3.4.

Finally a table is proposed to determine the final classification of the severity according to the global score as seen in Figure 3.5.

Type of product	Volume rejected (in m^3)	Score
Substances labelled H400, pesticides	≥ 10	100
	≤ 10	80
Substances labelled H290, strong acids, strong alkalis	≥ 100	100
	≤ 100	40
fertilizer	≥ 100	100
	≤ 100	40
Food	≥ 100	100
	≤ 100	40

Table 3.2: Table for the source score on biodiversity severity. Source: [Roux, 2013]

Remarks about this approach

This document is to my knowledge the first made to evaluate the expected severity of a scenario of accident. All the steps to get a practical result are explained from the listing of the possible targets to the final aggregation and allows the user to practice the severity evaluation completely. This work is currently being continued by Christophe Duval so as to improve it. This obtained index can be seen as made through a hierarchy of criteria where the source module score is a sub-criterion.

Type of transfer	Criteria for reducing the danger of score	Score
Soil runoff	Beyond 250 m distance	0.5
	high permeability soil surface	0.1
Infiltration into the soil	Low permeability soil in unsaturated zone (silt, slightly fissured rock) and water table depth $\geq$ 10 m	0.1

Table 3.3: Table for the transfer score. Source: [Roux, 2013]

Type of protected area	Score
National park - kernel area, National Nature Reserve, integral biological reserve	100
Regional Nature Reserve, national hunting and wildlife reserve, arrested biotope protection	50
Marine park, sensitive natural area	25
No protection	1

Table 3.4: Table for the target score. Source: [Roux, 2013]

Nevertheless, both the scores for the three modules and the final aggregation seem to be obtained approximatively (mainly powers of 10). No mention is made about any formal method to find them. The creator of this method did not use the methodological tools that we presented earlier in this document. The author of this method did not properly list the criteria that must be considered and the aggregation procedure which is used is a product of scores that looks rather arbitrary. Furthermore, this methodology was created by only one engineer without interviewing experts in toxicology nor in ecology to use their expertise on these crucial topics. Thus, we thought it was possible to improve this evaluation to make it more representative of human's valuation of severity of pollution and more adapted to the issue of risks of accidental pollutions. This process requires using methodological tools from multi-criteria decision theory and interacting with appropriate experts. The main aim of this thesis is to fill this gap

Severity value level	Score
Disastrous (C5)	[5000;10 000]
Catastrophic (C4)	[1000;5000]
Important (C3)	[100;1000]
Serious (C2)	[10;100]
Moderate (C1)	[0;10]

Table 3.5: Table for the global classification from the global score. Source: [Roux, 2013]

and we are about to described how we did so.

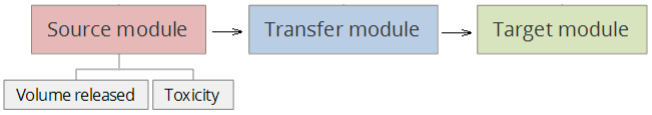

Figure 3.4: Graphic interpretation of the model described in *Méthode d'évaluation de la gravité des conséquences environnementales d'un accident industriel*

3.3.3 First modifications of the hierarchy of criteria

In order to find the appropriate hierarchy of criteria we applied some Value Focused Thinking advises and we asked ourselves the question "why is the transfer module important in this process?". The obvious response to this question was that it impacts the volume that will reach the considered target. Thus, the volume that reaches the considered target is a criterion that must be taken into account and that admits the target module as a sub-criterion with a causal factor. The question comes next to know what could be the other criteria that influence it. And the other obvious criterion that was found was the volume of liquid that escapes from the source. Then, as explained in Subsection 2.1.6.1 we asked ourselves why we are interested in the volume that gets to the target and we got to the conclusion that while studying a water target (lakes, rivers, etc), we are interested in this information because it will determine the *exposure of the local biodiversity to the product*. The *exposure of the local biodiversity to the product* also depends on the *receiving volume* of water i.e. the volume of water initially contained in the considered target and the *mobility of water* in the target. To the question "why are we interested in the *exposure of the local biodiversity to the product* of liquid in the target?" we arrived to the conclusion that it impacts, as well

as the toxicity of the liquid, the *destructive potential of the leak* which, as well as the environmental value of the target, impacts the severity of the given scenario on the considered target".

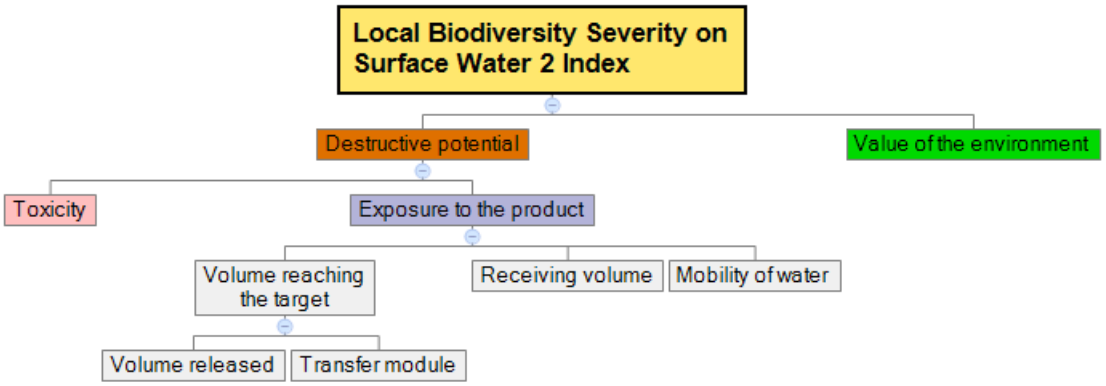

Figure 3.5: First modification of the hierarchy of criteria

Following the reasoning used in Subsection 3.3.2 the "index of biological severity" is divided into indices describing the impacts of the scenario on different types of target, specifically surface water, ground water or soil. The thus obtained framework is graphically described in Figure 3.5. The next modifications only concern the local biodiversity severity and its sub-criteria. Hence, thereafter we will not mention the upper part of the hierarchy.

3.3.4 Interacting with experts

At this point we understood that this index would be more meaningful if created with the help of scientific experts. The two main fields that are involved in this process are biodiversity and toxicology. We decided to work with the *Museum National d'Histoire Naturelle de Paris* (MNHN). The Museum National d'Histoire Naturelle de Paris founded in 1793 is a French research institution and dissemination of naturalistic scientific culture. In particular, they published the "cahiers d'habitats Natura 2000" [Bensettiti et al., 2002] that inventory and describes the ecological habitats on the French soil. Thus, their expertise seemed totally appropriate to our to the biodiversity evaluation that is needed in our problem. We met the MNHN four times and they

helped us to validate the model and aggregate the *destructive potential* the *vulnerability of the target* and the *importance of the environment* into the *local Biodiversity Severity Index*.

On toxicology we were supported by Eric Thybaud, a toxicologist working at the INERIS, that accorded us 5 interviews. He helped us to validate the model and to aggregate the *toxicity*, the *concentration* and the *residence time* into the *destructive potential*.

Posing the bases of these interactions

During these interactions with experts our objectives were to:

- Listen to their remarks to step back and see our problem from a different perspective
- Use their expertise to create an appropriate hierarchy of criteria
- Use their expertise to aggregate the criteria at each node of the hierarchy

First of all it seemed important to create the bases of our interaction with the experts by:

1) Presenting our problem to the experts: the context, our objectives, our constraints and the multi-criteria approach that we chose to use.
2) Concerting on the definition of the crucial concepts so as to make sure that we will speak in a common language and avoid potential misunderstandings.
3) Understanding at which part of the process and how their expertise may be useful.
4) Listening to their comments to look at our problem from another perspective and making sure that we are not missing an important issue.
5) Checking the tools that already exist in their scientific fields that may help us ("does such an evaluation already exist?","does a scale on the value of the environment exist?").

We had five meetings with the expert in toxicology of the INERIS and four with the experts in ecology of the MNHN. These meeting generally lasted 2 to 3 hours. The

meeting with the experts in ecology were generally organized together with Alexandre Robert researcher at the MNHN but the other stakeholder did not attend to all of the meetings. The two first meeting were maid in small groups: 3 to 4 experts from the MNHN, one engineer from the INERIS (Christophe Duval) and myself. The third one was made with a larger group: Meltem Ozturk (LAMSADE), about 20 experts from the MNHN. Finally the last meeting was an interview with Meltem Ozturk, Alexandre Robert (MNHN) and myself.

Experts's skepticism regarding our approach

During the interviews with the MNHN many of the researchers were skeptical for mainly one reason. Indeed, to many of them it seemed that we were creating an algorithm that could replace their expertise. They argued that what we were doing was an extreme simplification of all the mechanisms involved in the process of impact on biodiversity and in the evaluation of the importance of this impact. Indeed, the form itself of the index, one indicator on a discrete scale with few value levels, cannot picture the complexity of the description of the possible consequences of an industrial accident on the biodiversity. Then, they argued that we will not be able to get all the necessary information to evaluate the severity of the impact. Finally, an algorithm or a mathematical rule will never be able to replace the human judgement experience and intuition.

It is likely that, while creating such synthetic indices, analysts have to face similar remarks from scientists working on the ground. To these legitimate concerns, we answered that we are not trying to replace their expertise by an algorithm, we are building a tool. We accept that this tool will be an approximation of the judgement that a team of experts might have obtained with a more detailed survey. Nevertheless, we need such an index in our risk management process and this index must be created systematically. Indeed, we want this index to be neutral in the sense that every scenarios and every industrial plant is treated the same way. We are also concerned about minimize the necessary resources for this risk evaluation. Specifically, we want the user of our method to resort as little as possible to any kind of expert during risk studies.

3.3.5 Second modification based on the *Muséum National d'Histoire Naturelle*'s expertise

At first, we put the bases of our interactions as described in Subsection 3.3.4. The experts of the MNHN felt competent to help us on the part that is above the *destructive potential* criterion and bellow the *local biodiversity severity index*.

To the questions mentioned in Subsection 3.3.4 we added two extra questions: "What could be a scale to represent the severity of a scenario of accidental pollution?" and "Could the biodiversity equilibrium be modified by an industrial accident?". We also wanted to validate our current hierarchy of criteria or make it evolve in the light of their expertise.

It came out that the ecological resilience of the considered target is an important criterion to take into account. This criterion is highly influenced by the lifetime of the living organism present of the considered target. The ecological resilience is generally considered as the ability of a target to recover quickly to its initial state after being disturbed by an external stress but also sometimes as the quantity of disturbance needed so that the target will never turn back to its original state by its own [Connell and Sousa, 1983][Gunderson, 2000]. Given that the vast majority of targets in France are constantly or regularly subject to stress directly or indirectly linked to human activities it is very unlikely that a single event would change the ecological equilibrium of a target. Thus, the second definition of resilience previously given ("the quantity of disturbance needed so that the target will never turn back to its original state by its own") is not appropriate in this situation and we should better consider the expected time to return to the original ecological equilibrium.

They also mentioned the fact that the degradation time of the product in the target must be taken into account. This criterion will be called *persistence* of the product.

About the ecosystem that exist in the deep ground water, the experts argued that it is not known well enough to be efficiently taken into account in our process. As well, the season during which the accident happen influences the severity of its consequences on biodiversity. However, the season at which the accident occurs is a criterion that influences the severity of a pollution. However, given the fact that in most of the cases we will not be able to foresee the period at which a scenario of accidental pollution could happen, we decided not to consider the season as a criterion in the hierarchy of criteria.

Finally, to their knowledge, excepted monetarization, there does not exist a standard scale or unit to measure formally the severity of a pollution nor the value of the

environment.

We then, oriented the discussion to the research of an appropriate hierarchy of criteria implementing, as explained in Subsection 2.1.6.1, the procedure that consists in asking them for each criterion why they are interested in it. This step was made before the experts saw the hierarchy of criteria obtained in Subsection 3.3.3 in order to avoid influencing them. Taking into account the previous remarks, we obtained a hierarchy of criteria that was actually very close to the previous one. However, this new hierarchy of criteria included the *impact on the target on biodiversity* which can be defined as the extent to which the biodiversity would be modified by the considered scenario. This criterion should not take into account the importance given by society to the biodiversity on the target. This information is taken into account in an other criterion. The *impact on the target on biodiversity* is influenced through a causal factor by the *exposure to the product* and the *toxicity of the product*. As explained earlier, we also integrated the resilience and the persistence of the product as new criteria in the hierarchy. Hence, the hierarchy that we obtained at the end of this meeting is illustrated in Figure 3.6.

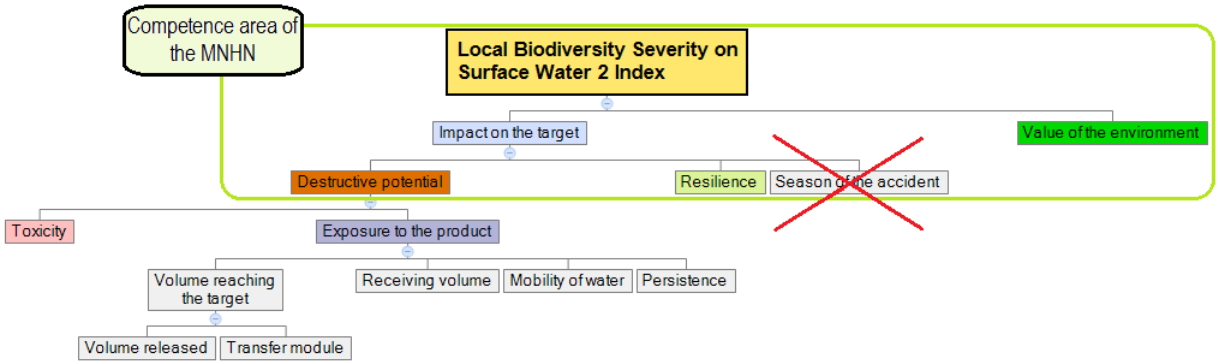

Figure 3.6: Second modification of the hierarchy of criteria

3.3.6 Validation of the previously obtained model with a toxicologist

As was done with the MNHN, at the beginning of our interviews with the toxicologist, we put the bases of our interactions as defined in Subsection 3.3.4. Then, we looked for sub-problems in which a toxicologic expertise could be useful. The expert declared

that his expertise would be adapted to the process of aggregating data to create the destructive potential. But he declared that he was not competent enough in biology to evaluate the resilience. We came to the conclusion that his expertise would mainly help us to find the destructive potential. He might also help us aggregate the destructive potential and the resilience to get the impact on biodiversity, but only if an explicit syntax is created for the resilience.

Then, we got on with enumerating the criteria that should be taken into account in this evaluation process. As we did with the MNHN, we made this work before the expert could see the hierarchy of criteria that was already obtained. However, the hierarchy of criteria that was obtained was identical to the one obtained interacting with the MNHN. It has reinforced our belief that this hierarchy of criteria was appropriate to this problem.

As a novelty, the expert thought that the toxicity could be decomposed into chemical toxicity and physical toxicity. In our context, we should consider as physical toxicity the toxicity due to the pH of a product and the toxicity due to the creation of a layer of liquid that would avoid water to exchange oxygen with the air.

3.3.7 Validating the destructive potential as a criterion

Another topic was to check with the expert that the destructive potential is indeed an intermediate step while evaluating the possible consequences of a scenario. The goal of this question was to decide if this criterion should be included in the hierarchy or if we should better aggregate the *toxicity* of the liquid, the *exposition* to the product and the *resilience* of the target to get the impact on biodiversity (illustration of the two possibilities in Figure 3.7). To the toxicologist, it seemed quite intuitive to work with the *destructive potential*. To validate this criterion we also must check that its sub-criteria and itself are independent to the other criteria in the hierarchy as defined in Subsection 2.2.6. We proceeded as described in Subsection 3.4.1 to check it and validate this criterion. We could consider that there exists a dependence between the destructive potential and other criteria that depend on the target (such as the value of the environment) in the sense that some target could be more or less resistant to some specific products than others but several arguments could advocate to consider the destructive potential independently from the criterion *value of the environment*, the *resistance* or the *resilience* of the target. First, the information that will be taken into account in the criterion *destructive potential* that could interact with other criteria is the *toxicity*. It will be expressed with an ordered scale and will not contain any information about specificities of the product. The information outside the *destructive*

potential criterion that could interact with the *toxicity* are the *value of the environment* and the *resistance* and the *resilience* of the target. But if some specific product could have higher/lower impact on some specific target due to some specificity of both the target and the product, knowing only the level of toxicity of a product, the level of importance and the level of *resistance* or *resilience* of a target this "matching" could not be predicted. Then, it was considered that products that are toxic are generally toxic to most of the target and that thus including the dependence would not be really meaningful compared to the cost of integrating them.

Thus, the conclusion was that the destructive potential should be included in the hierarchy of criteria.

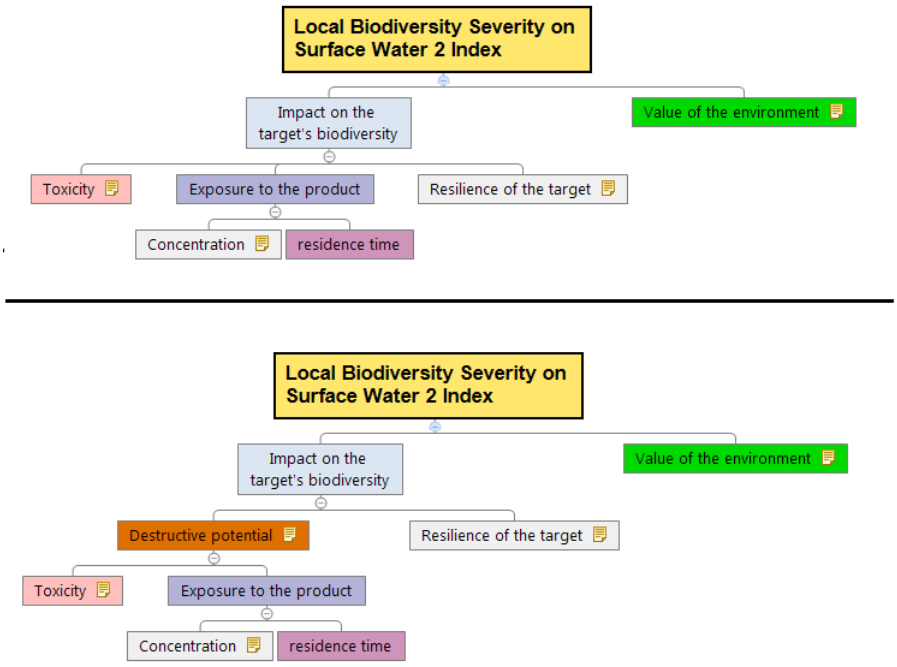

Figure 3.7: Illustration of the hierarchy with (above) and without (below) the destructive potential criterion.

3.3.8 Including the *residence time* and the concentration as new criteria

While wondering "why are we interested in the water mobility?" it came out that the answer would be "because it impacts the *residence time* of the product on the

target". And the other criterion that would influence the *residence time* seemed to be the *persistence* criterion. While trying to understand how to represent the *mobility of water* it came out that it could simply represented as a boolean variable (moving water/standing water). Indeed, a more precise information could be difficult and expensive to obtain and would probably not be useful to our process. As for made with the water mobility, it seemed concerning the residence time that two main situations could appear; a long residence time (the product being still present at the target weeks after the scenario happened) and short time residence (the product being degraded or elapsed quickly). The *persistence* criterion can also be represented as boolean variable without to much loss of valuable information. Indeed, toxic product may whether degrade them-self while impacting the environment or not. Thus, we decided to include the *residence time* criterion as a boolean variable that is equal to "long" if the product is not persistent and if the target is a standing water and equal to "short" otherwise.

While the persistence and the mobility of water were grouped together into the *residence time* it seemed quite natural to group the *volume reaching the target* and the *receiving volume* inside the *concentration* criterion. Indeed the impact on living organisms directly depends on the concentration rather than the volume released.

The current model at this time is described in Figure 3.8.

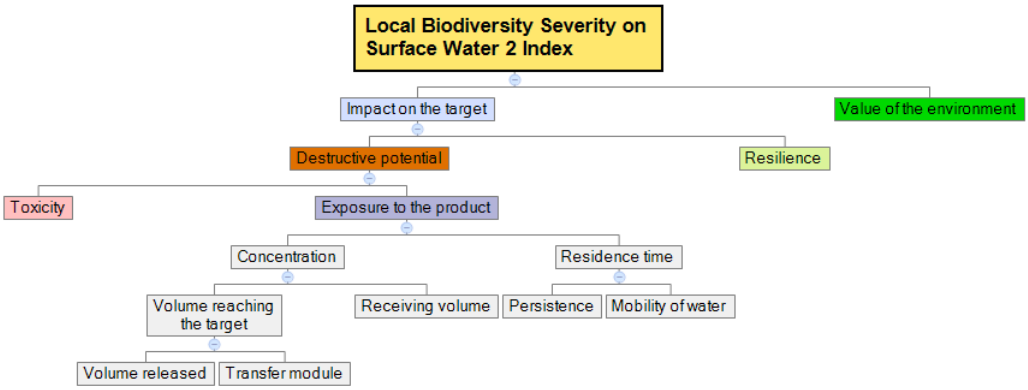

Figure 3.8: Third modification of the hierarchy of criteria

3.3.9 Adding the *resistance* criterion by interacting with the *Muséum National d'Histoire Naturelle*

After some meetings with the toxicologist expert we met the MNHN expert again. Our main motivations for this second meeting was to understand better the *resilience criterion* and to have a first idea on how this criterion together with the *destructive potential* influences the *impact on the target*. Some interesting remarks were made by the experts among which the following are particularly worth being mentioned. At first, they told us that the more the target is connected to other targets the more resilient it will be. We say that a target is connected if it can be recolonized from other surrounding target after being biologically impacted. Then, there is not a total independence between the impacts on several target. A target will recover quicker if the surrounding targets are not impacted (the recolonization will be easier). However, we concluded that taking into account the interactions between several target would highly complexify the method. We decided not to take them into account although we know that they exist.

Targets that are already impacted by humans will be more resilient. The resistance of the target is also important. The ecological resistance [Grimm and Wissel, 1997] of a target is considered as the ability of a target to remain unchanged or sparsely changed by an external stress. The resistance and the resilience are two different notions and a resistant target is not necessarily resilient (and vice versa). We concluded that the resistance criterion was to be added to the model. It was supposed to be given as an input of the model and would influence the short term impact on the target as well as the destructive potential. The short term impact would influence the long term impact as well as the resilience.

The thus obtained model is described in Figure 3.9.

3.3.10 From an exhaustive model to a "user friendly" one

At this point we thought that the hierarchy of criteria was probably quite accurate with regard to connections that the criteria have on each other. However, as the reader may observe, the current model at this time was quite complex and we wanted to simplify it for mainly two reasons. At first, we might have had a lack of resources to make all these elicitation, in particular concerning the expert availability. Then, each elicitation induces an inevitable error (or noise). Stacking too many aggregation

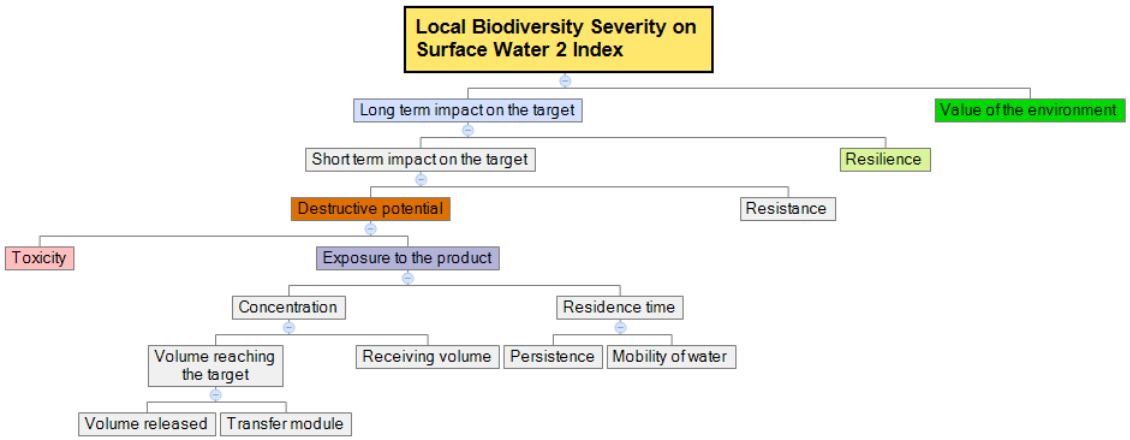

Figure 3.9: Fourth modification of the hierarchy of criteria

might accentuate this noise. Thus, we decided to simplify our model by removing some intermediate criteria.

First of all, it seemed quite intuitive and adapted to the experts of the *Muséum National d'Histoire Naturelle* to add a criterion to replace the *resilience* and the *resistance* by the *vulnerability* as explained in Subsection 3.1.1. These three information being subjectively understood and intuitively described, the evaluation of the vulnerability is probably not less accurate nor more difficult to the experts than the evaluation of the resistance and the resilience could be. This data should be given in an input of our methodology. Then, we decided to remove the impact criterion to directly aggregate into the *local biodiversity severity index* the *destructive potential*, the *vulnerability* and the *value of the environment*. Finally we removed the exposure criterion to directly aggregate into the destructive potential the *concentration* of the product, the *residence time* and the *toxicity of the product*. The concentration should be obtained with the use of a physical model. There still remains two aggregations: the destructive potential, rather relative to toxicology and the local biodiversity severity index rather relevant to ecology. The comparison between the exhaustive model and the user friendly one is illustrated in Figure 3.10

3.3.11 Why we stopped the evolution and accepted the hierarchy as it is.

After these last modifications were done and after they were validated during the next meeting with both the MNHN experts and the toxicologist it was considered

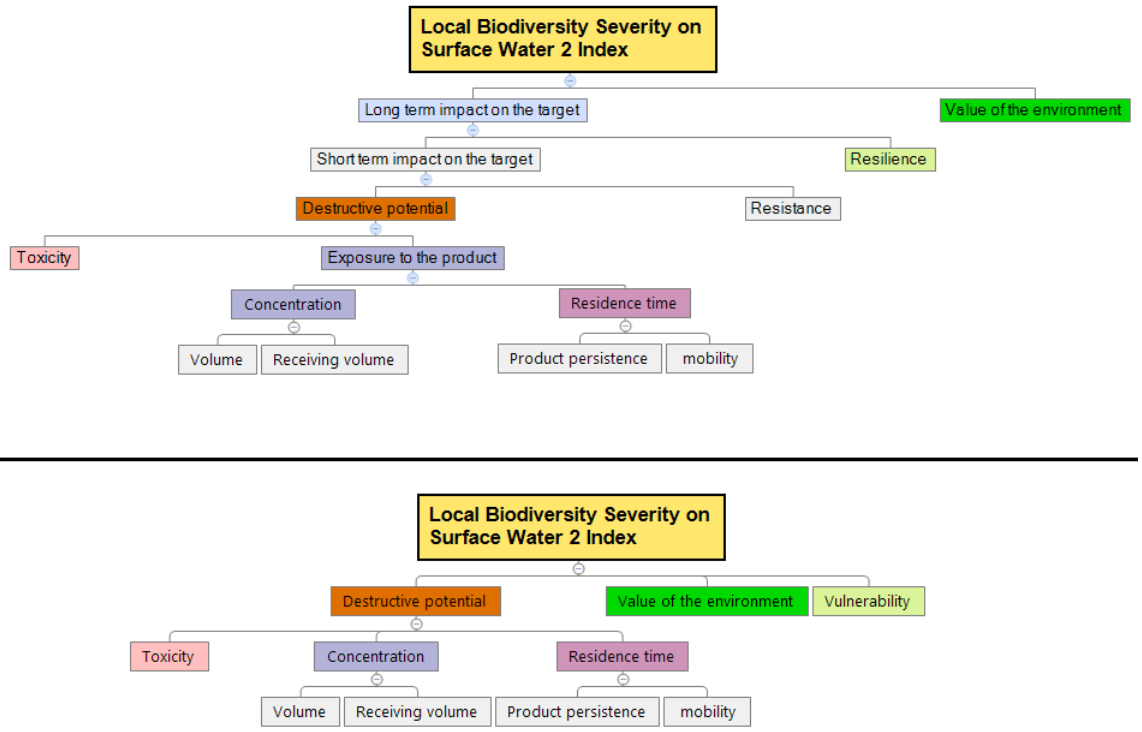

Figure 3.10: Comparison between the *exhaustive model* (above) and the *user friendly model* (bellow)

that this hierarchy of criteria was convenient in this context. We are aware that this hierarchy is probably not the only one that could be appropriate to this problem and that other researchers could have used an other approach and come up to an other one. Nevertheless, here are some positive comments that we can make about this hierarchy of criteria with respect the list of good properties made in Subsection 2.1.6:

- We managed to use criteria understood in the same way by every stakeholder involved in this process.
- Every maximal sub-family of criteria of this hierarchy (as defined in Subsection 2.1.6) is a coherent family of criteria.
- Each sub-problem can be treated independently given that there is no dependence between a criterion and an other criterion located in an other sub-problem. By the way we will later see that there is probably no dependency between criteria.
- Every sub-problem contains a limited number of criteria which makes it easier to manage for the experts.

- Every relation between a criterion and its sub-criteria seems quite natural to all the actors involved in the process.
- There was a limited number of aggregations to be done.

However, the criteria located at the bottom of the hierarchy are not all easy to obtain. For instance the prediction of the expected concentration of liquid in the target requires the use of physic models that are not included in our problem. The value of the environment and the vulnerability of a target are not directly available data even if they can be evaluated by experts in particular in France, the MNHN with the help of some tools such as those presented in Subsection 3.2.5

3.4 Aggregation to get the destructive potential

At this point we had a hierarchy of criteria that was obtained through a progressive process made of interactions with experts and we were interested in the aggregation that was to be done at its sub-problems. This raises various technical questions about what is the role of each criterion in each sub-problem and how the criteria interact with each other in the experts or decision maker's minds as described in Subsection 2.2.6. The following section describes the way that we practically dealt with these issues in the sub-problem of the creation of the *destructive potential* (illustrated in Figure 3.11).

Aggregating the *toxicity*, the *concentration* and the *residence time* into the *destructive potential* is a sorting problem with 3 criteria and 5 categories. As mentioned in Subsection 3.3.4 this part of our work was made with the help of an expert in toxicology with which we interacted during five interviews (2 to 3 hours each).

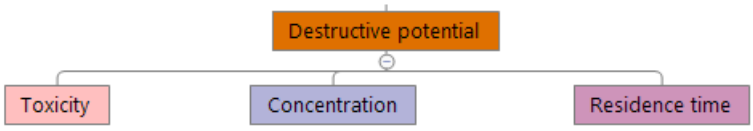

Figure 3.11: Illustration of the sub-problem of the construction of the *destructive potential*

3.4.1 Finding possible dependences between sub-criteria of the destructive potential

In order to choose which aggregation method should be used to obtain the destructive potential, we first need to understand if there could be dependencies between its sub-criteria as explained in Subsection 2.2.6. Indeed most of the MCAPs (ELECTRE methods, UTADIS, MACBETH...) cannot represent preferences that induce dependence between criteria while others, in particular Choquet integrals can. The first step of this process consisted in explaining to the decision maker what dependence between criteria is, either positive and negative synergy. In order to do so, after having defined the concept, we presented an example of decision context where dependencies may exist to the expert. This example was based on the choice of a housing where five criteria would be considered; the area of the apartment, the attractiveness of the surrounding, the price, connexion with public transports and the presence of a parking space. The expert had no difficulty to understand that there could be a redundancy between the connexion with public transports and the presence of a parking space in the sense that the absence of a parking place could be more problematic if the connexion with public transport is bad. Then we showed several sets of objects (here theoretical houses) (a, b, c, d) such as $a = (v, w, x, y, z)$, $b = (v', w', x', y', z)$, $c = (v, w, x, y, z')$, $d = (v', w', x', y', z')$, where v is the area, w is the attractiveness of the surrounding, x the price, y the connexion with public transports and z the presence of a parking space (boolean value) and the expert understood that is legitimate to simultaneously express the following preferences: $a > b$ and $c < d$ (illustration Figure 3.12).

Then, we talked with the expert about the possibility to find this situation in our problem. The global feeling was that it was not appropriate to our context. In order to validate this intuition we showed several sets of objects (here toxic leaks) (a, b, c, d) such as $a = (x, y, z)$, $b = (x', y', z)$, $c = (x, y, z')$, $d = (x', y', z')$, where x is the toxicity, y is the concentration of product in the target after the scenario happened and z is the residence time. The expert did not expressed simultaneously preferences that supposed dependences. We also made this same test with (a, b, c, d) such as $a = (x, y, z)$, $b = (x', y, z')$, $c = (x, y', z)$, $d = (x', y', z')$ and with (a, b, c, d) such as $a = (x, y, z)$, $b = (x, y', z')$, $c = (x', y, z)$, $d = (x', y', z')$ to look for other possible dependencies. We deduced from this experience that there does not exist any significant dependence between the sub-criteria of the destructive potential.

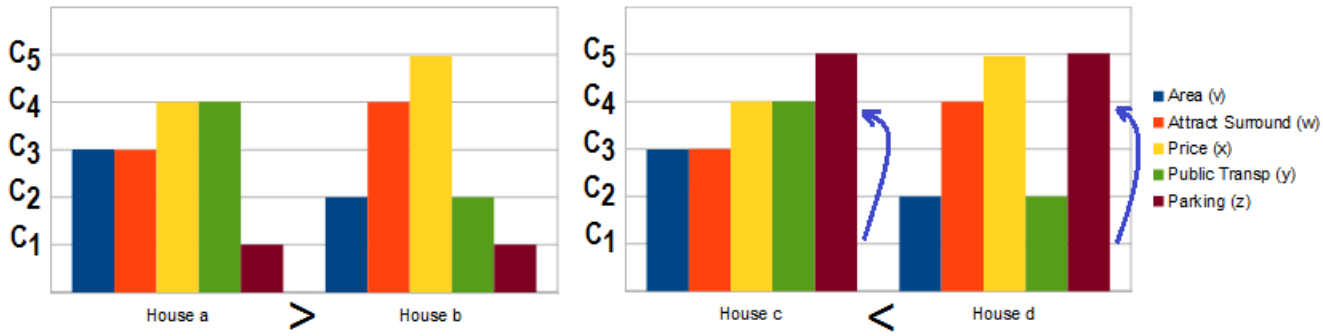

Figure 3.12: Graphic illustration of dependency between criteria. The four objects a, b, c, d are expressed on five criteria (area, attractiveness of the surrounding, price, the connexion to public transportation, presence of a parking place). These criteria are expressed on a scale of c_1 ("very bad") to c_5 ("very good"), the presence of a parking place has only two levels c_1 ("no parking place"), c_5 ("parking place provided". We can see here that adding a parking place to both a and b changes the order of preference between these two objects.

3.4.2 About additive methods for the destructive potential

At first, we presented to the expert in toxicology several MCAPs for multi-criteria decision aiding; ELECTRE TRI [Mousseau et al., 2000], Choquet integrals [Labreuche and Grabisch, 2003], rule based methods [Stefanowski and Vanderpooten, 2001] so as to deal with the aggregation issue as described in Section 3.4. Many of the most popular aggregation methods in multi-criteria decision aiding are based on additive methods as described in Subsection 2.2.3. While looking for an appropriate method to aggregate the sub-criteria of the destructive potential we had to wonder if additive methods are appropriate. One problematic point in using additive methods in our context was that for obvious reasons, if the concentration of liquid of the leak is equal to 0 then we would like the destructive potential to be null and likewise if the toxicity is null. However, using an additive method, if the concentration is null still the global score of the destructive potential will be equal to the score of the toxicity plus the score of the residence time thus can be hight (likewise with toxicity). Then one could argue that with concentration equal 0 or non toxic product the scenario should not even be studied but then we should treat the case where the concentration tends to 0.

Intuitively it seemed more appropriate to use a multiplicative method which means that objects have a utility on every criterion depending upon its value on each criterion and the global score of the object is the product of its score every criteria. Indeed, it seems legitimate to think that for a fixed residence time the score should be proportional

to scores on the concentration and the toxicity. However, multiplicative utilities can always be represented through additive utilities. Indeed given that both $exp(x)$ and $ln(x)$ are strictly increasing function and that $s = \prod_{i \in N} f(g_i(a)) = exp(\sum_{i \in N} ln(f(g_i(a))))$ we can deduce that if there exists a set of utility functions $f(x)$ such that (a is preferred to b)$\Leftrightarrow \prod_{i \in N} f(g_i(a)) > \prod_{i \in N} f(g_i(b))$ then there exists a utility function $f_+(x) = ln(f(x))$ such that (a is preferred to b)$\Leftrightarrow \sum_{i \in N} f_+(g_i(a)) > \sum_{i \in N} f_+(g_i(b))$.

In the case of a transformation from a multiplicative method to an additive one, the problem mentioned earlier for using additive methods in our situation would be visible by the fact that the logarithm function is not defined at 0. However toxic leaks with a zero concentration or toxicity will not be taken into account. So we concluded that we may use a multiplicative method.

3.4.3 Defining the aggregation method for the destructive potential

Given that any multiplicative method can be represented as an additive method as explained in Subsection 3.4.2 we chose to evaluate the destructive potential through the use of UTADIS method. Nevertheless, we knew it was also possible that the main base on which the expert would judge these scenarios would be "how many times do we have the maximal acceptable concentration?". In order to make this elicitation, we prepared a list of scenarios of leaks expressed by the type of product, the maximal acceptable toxicant concentration of the product, the residence time and the expected concentration in the target after the scenario happens. In this case we showed to the expert some "types of product" that correspond to the maximal acceptable toxicant concentration. The idea was that we did not want to induce the expert to choose to base his judgement on the ratio $\frac{Concentration}{MATC}$ ($MATC$ being the maximum acceptable toxicant concentration). We tried to distribute the values of these scenarios as well as possible to cover every possible new scenario. The idea was to compare the obtained result to the ratio $\frac{Concentration}{MATC}$ to see if our first intuition was justified. We also asked for each scenario, after being assigned, how much should the concentration raise so that the scenario would be assigned to a higher category and conversely how much should it decrease so that the scenario change its assignment to a lower category. By doing so, some new scenarios are added to the learning set. Furthermore, these new scenarios correspond to objects that are close to the frontier between two categories (or value levels) and thus might help us to approximate it well.

After having presented the program, we begun the elicitation, showing scenarios and asking the expert to assign them as explained before. Quickly, he told us that his reasoning was based on the ratio $\frac{Concentration}{MATC}$. Thus, it was not necessary to elicit his preferences. Indeed, while using a disaggregation approach, the elicitation process aims at proposing an aggregation method that would approximate the decisions or assignment that would be made by a decision maker. The interest of this approach comes from the fact that the decision maker is not himself able to formally describe the rules that support his decision or judgement. In this situation, the expert knows these rules and thus their approximation through an elicitation process is not necessary. Therefore, we moved from a disaggregation approach, where the decision maker or expert gives some examples of judgement, to an aggregation approach where the decision maker or expert can describe the way that synthetic evaluation works in his mind.

However, the value of $\frac{Concentration}{MATC}$ is not appropriate to represent the destructive potential given that it has no intuitive meaning for the ecology experts that will use it as an input. For that purpose, we created a scale of five value levels described in Subsection 3.2.4 whose definition was accepted by both the toxicologist and the ecology experts. Furthermore, the value of $\frac{Concentration}{MATC}$ does not take into account the *residence time* that was defined as a relevant criterion by the expert. Thus, we created a set of rules to assign any scenario to one of the five value levels of the scales on the *destructive potential*. In order to do so, we first asked the expert to give an assignment that he is certain of. Here are his answers:

- If the concentration of the product is below the maximal admissible toxicant concentration then there should not be any impact on the environment regardless of the residence time. Indeed such a concentration is even considerate as admissible in normal circumstances. Thus the destructive potential will be assigned as $C1$.
- If the concentration is higher than 1000 times the maximal admissible toxicant concentration then the biodiversity will generally be totally destructed regardless of the residence time ($C5$).
- If the concentration is between 100 and 1000 times the maximal admissible toxicant concentration then the biodiversity will be strongly impacted is the residence time is short ($C4$) and totally destructed ($C5$) if it is long.
- If the concentration is between 1 and 10 times the maximal admissible toxicant concentration it will generally not have any impact if the product does not stay too long but it could have a low impact if it does remain on the target. Then

the destructive potential will be assigned to category $C2$ if the residence time is long.

Then, the "undetermined area" in our model was restricted. We could not assign to any category the leaks whose concentrations are between 1 and 100 times the maximal admissible toxicant concentration nor the leak whose concentrations are between 1 and 10 times the maximal admissible toxicant concentration with a long residence time as illustrated in table 3.6. On the other side, one category was not defined, the category $C3$. Together with the expert, we filled this gap by placing a limit at 50 times the maximal admissible toxicant concentration. It was determined as fair to state that a long time residence would increase the category of the destructive potential by 1.

Thus the assignment rule is described in the table 3.7.

$\frac{Cons}{MATC}$	Short residence time	Long residence time
≤ 1	$C1$	$C1$
$]1, 10]$	$C1$	$C2$
$]10, 50]$	?	?
$]50, 100]$	?	?
$]100, 1000]$	$C4$	$C5$
> 1000	$C5$	$C5$

Table 3.6: Table of the destructive potential (incomplete).

$\frac{Cons}{MATC}$	Short residence time	Long residence time
≤ 1	$C1$	$C1$
$]1, 10]$	$C1$	$C2$
$]10, 50]$	$C2$	$C3$
$]50, 100]$	$C3$	$C4$
$]100, 1000]$	$C4$	$C5$
> 1000	$C5$	$C5$

Table 3.7: Table of the destructive potential (completed)

Regarding the risk that the liquid creates a layer that avoid water to exchange oxygen with the air (if the liquid is oil for instance), the expert told us that the impact on the environment essentially depends on the proportion of the target that is covered by the layer and not the concentration, given that this case happens with products that

are not miscible with water. According to his experience on real accidents and tests he came to the conclusion that the destructive potential could in this case be calculated as follow:

- If the layer covers less than 10% of the target then its impact on biodiversity will be negligible. Then the destructive potential can be assigned to category 1.
- If the layer covers between 10% and 50% of the target then the destructive potential will be assigned to category 2.
- If the layer covers between 50% and 70% of the target then the destructive potential will be assigned to category 3.
- If the layer covers between 70% and 90% of the target then the destructive potential will be assigned to category 4.
- If the layer covers more than 90% of the target then almost all the life of the target will disappear, we will assign the destructive potential to category 5.

This aggregation may look basic to the reader. One could wonder in which extent the interaction between the analyst and the expert was useful to the process and if this part of the work could have been done with the expert alone. However, the existence of the *destructive potential* itself was found through a decision aiding methodology. As well, the choice of the criteria that should be taken into account while measuring the *destructive potential* and the scales on which these criteria are expressed was made during these interactions. Finally, we believe that the discussion that we had about the aggregation (possible dependencies or vetos...) helped us and the expert to have a clearer understanding of what really matters while evaluating the destructive potential and how to measure it. Furthermore we found a scale to represent the possible levels of *destructive potential* that will allow us to make a link with the sub-problem located above in the hierarchy of criteria: the creation of the *local biodiversity severity index*.

3.5 Aggregation to get the Local Biodiversity Severity Indices

As we have seen earlier, the destructive potential that was obtained is not itself an end but aims at being aggregated as a criterion, together with the *value of the environment* and the *vulnerability*, to obtain the *local biodiversity severity index* (illustrated in Figure

3.13). As we did to obtain the destructive potential, the task of constructing the aggregation process requires interacting with experts in order to make it adapted to the human values and knowledge. Here the interactions were made during four interviews with the experts of the Museum National d'Histoire Naturelle (MNHN) which expertise in ecological matters seems convenient for this sub-problem while the three criteria that are considered are the *destructive potential*, the *vulnerability* and the *value of the environment*. This sub-problem being a sorting problem, we will consider here the value level representing the *local biodiversity severity index* as the category in a sorting problem. In this section, we will test and compare several MCAPs and elicitation methods, direct and indirect to converge toward a MR-Sort method that suits to the experts of the MNHN.

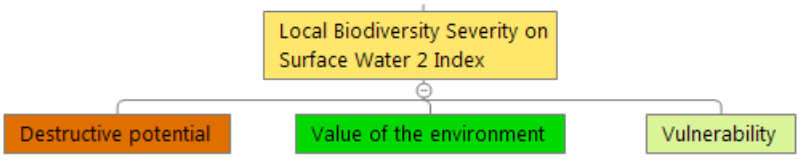

Figure 3.13: Illustration of the sub-problem of the construction of the *local biodiversity severity index*

3.5.1 Frame of this sub-problem

Scales used for the vulnerability, the value of the environment and the local biodiversity severity index

At first as for every sub-problem, we needed to state which scale will be used to represent the criteria given in input and found as the output of this sorting. We saw earlier that the destructive potential is expressed on a semantically defined discrete scale of 5 value levels. The *vulnerability*, the *value of the environment* and the *local biodiversity severity index* are concepts for which no formal scales exist from their scientific fields. Furthermore, we do not expect them to be found through a formal calculation but rather to be intuitively judged by an expert. Thus, a semantically defined discrete scale of 5 value levels as described in Subsection 3.2.4 also seems appropriate to express these three values.

Several possible MCAPs

In order to find the MCAP that is appropriate to the construction of the *local biodiversity severity index*, we needed to get some information about the compensation in this sub-problem. Interacting with the experts as described in Subsection 3.4.1 we got to the conclusion that, although there could be some specific product that interact more or less with some specific environment, no dependency was to be taken into account between the three criteria involved in this sub-problem. Then, contrarily to what happened while studying the construction of the destructive potential we did not have an intuition about what could be a good MCAP for this aggregation.

Testing several preference learning algorithms

We decided to elicit the preferences of the experts of the MNHN and to test the obtained data with several preference learning algorithms for the sorting problem: UTADIS, DRSA, the evolutionary algorithm for MR-Sort presented in Subsection 2.3.6 and the Dominance Based Monte Carlo (DBMC) algorithm presented later in this document. The choice of the three first algorithms was based on several properties. At first, these algorithms are framed by three MCAP's (utility, set of rules, outranking) which are the three main families of MCAP. Thus, these methods may be seen as rather representative of the sorting methods. Furthermore, these methods are recognized and count a high number of publications. Then, they run in a reasonable computable time. Finally, they are based on assignment examples which in our case is the information that we asked to the expert. Indeed, we had to use a disaggregation approach to elicit the preference parameters of the experts. Indeed, using any direct elicitation method requires asking the experts questions about the preference parameters of the MCAP such as "what do you think may be the weight of the criterion *destructive potential*". These preference parameters being different for every MCAP we would have had to do an elicitation process for every method. The time of interaction with the experts being limited we chose to use a disaggregation process which as we are about to see requires from the experts for all the methods one same type preference information: a learning i.e. a set of assignment of scenarios to categories.

Given that no particular MCAP seems intuitively adapted to this problem, we were interested in using methods that are not based on MCAPs. However, as explained in Subsection 4.1.4, we were not fully satisfied by the properties of these methods. Thus, we created a method, described in chapter 4 that aims at finding an "average" of all the complete assignments that respect the learning set which is being given to us.

3.5.2 The questionnaire

We gave a questionnaire to the experts of the MNHN so as to use the answer as a learning set. The questionnaire (presented in Table 3.9) is composed of 20 scenarios of toxic leak to be assigned to a category representing the severity of the pollution as defined earlier in this document. The questionnaire was filled by 6 groups of 2-3 experts that were created informally. The groups had 30 minutes to fill the questionnaire.

Using argument strength assessment

We decided to ask the experts to give us argument strength assessment as described in Subsection 2.3.3. With this type of fuzzy assessmennt proposed in [Cailloux, 2012], the experts give for each object $a \in A'$ and each category c an integer score representing level of confidence for the assertion "I think that the object a should be sorted in the category c". This information can then be used to obtain intervals of categories or to aggregate the preferences of a group into an exact assignment (similar to the preference of one decision maker).

We decided to use this type of information for mainly two reasons. First of all, we wanted the experts to express their judgements in a way that may be less categorical than an exact assignment. Then, this format of information allows to aggregate the preferences of several experts in one synthetic learning set as we will later see.

In our case, the learning set of each of the groups of expert consists in distributing for each scenario 7 points on the categories proportionally to the correctness of the assertion "I think that this scenario should be sorted in this category". We forced the experts to give at least 4 points to one category so that there is a category that is the most preferred one (assignment). For each scenario a and each category c, $\nu_a(c)$ is the number of groups that give 4 points or more (assignments) for the scenario a to the category c. For each scenario a and each category c, $\tau_a(c)$ is the sum of all the points that were given for the scenario a to the category c for all the experts.

Choosing the learning set

In order to find the appropriate set to propose, we generated randomly a large number of scenarios. Then we selected a set of scenarios that seems to cover the criteria space. Indeed, we could not just pick 20 scenarios randomly given that even if they would have been drawn uniformly the sample being very small the result would probably not have been balanced.

In order to find a fairly balanced sample we picked the scenarios according to the sum of their values on the criteria. Let us call $\Lambda(\alpha)$ the set of all the sets of 3 integer (x, y, z) between 1 and 5 such that $x+y+z=\alpha$ i.e. $\Lambda(\alpha) = \{(x, y, z) \in [\![1, 5]\!]^3 | x+y+z=\alpha\}$. We decided to pick among the randomly found scenarios, for each α between 3 and 15 a predefined number of scenarios from $\Lambda(\alpha)$ i.e. for each α a predefined number of scenarios such that $g_{PotDes}(a) + g_{Vuln}(a) + g_{ValEnv}(a) = \alpha$. Table 3.8 presents the sizes of $\Lambda(\alpha)$ and the number of scenarios that we chose from each $\Lambda(\alpha)$. The idea here was to pick a distribution of scenarios regarding the $\Lambda(\alpha)$s that was symmetrical and almost single picked with a pick in 9 just as the size of the $\Lambda(\alpha)$s. In practice, this method is also useful to make sure that you do not have the same scenario twice.

Given that we had the feeling that a scenario with a null *destructive potential* would be classified as null severity, we added the scenario with a null destructive potential and a very hight value of the environment and vulnerability so as to allow the preference learning algorithm to take this specificity into account. The results obtained from the 6 groups can be found in the appendix C.

α	3	4	5	6	7	8	9	10	11	12	13	14	15
$\lvert\Lambda(\alpha)\rvert$	1	3	6	10	15	18	19	18	15	10	6	3	1
Scenarios in $\Lambda(\alpha)$	0	0	1	0	3	4	4	4	3	0	1	0	0

Table 3.8: Sizes of $\Lambda(\alpha)$ and Number of scenarios that were chosen from each $\Lambda(\alpha)$.

Scenarios	Dest pot	Val Env	Vuln
S1	C4	C2	C3
S2	C5	C3	C5
S3	C4	C1	C2
S4	C4	C2	C1
S5	C1	C5	C5
S6	C2	C3	C3
S7	C2	C4	C4
S8	C3	C4	C1
S9	C5	C3	C2
S10	C5	C1	C4
S11	C3	C4	C2
S12	C3	C1	C5
S13	C2	C5	C1
S14	C3	C1	C4
S15	C5	C1	C1
S16	C4	C5	C1
S17	C2	C1	C2
S18	C4	C3	C4
S19	C3	C4	C4
S20	C4	C5	C4

Table 3.9: Table defining the scenarios proposed as a learning set to elicit the *Local Biodiversity Severity Indices*.

3.5.3 Processing the data

Coherence among the different learning sets received

While looking at the learning sets that we received, the question arised to know whether these learning sets are kind of similar or if very different opinions were expressed. In order to have an idea on this we looked at four data.

At first, for every pair of groups we looked at the percentage of scenario that are not assigned to the same category (with 4 point or more) of the severity by the two groups presented in Table 3.10. We saw that the proportion of different perception between groups is rather high. However, while looking for each pair of groups at the percentage of scenarios that are not assigned (with 4 points or more) to the same category nor

to an adjacent category we saw that this percentage is generally very low as shown in Table 3.11.

We also looked for every pair of groups (and shown in Table 3.12) the percentage of pairs of scenarios a and b such that the two groups disagree on which of a and b is the more severe (according to the category that receives 4 points or more). Here, we can see that the experts mainly agree order on the order of severity of scenarios. We deduced that the experts mainly have the same perception of the value levels regarding the *destructive potential*, the *value of the environment* and the *vulnerability* but still have different perceptions on the *severity* of a leak. We can also notice that the groups of experts felt more comfortable attributing argument strength assessment to the categories in order to express their view in a less categorical way. Thus, it is not very surprising that pairs of groups may assign a scenario to adjacent categories given that themselves do not give a complete approbation to one particular category for each scenario.

As well, we compared for each scenario proposed the standard deviation of the values on the criteria to the standard deviation of the assignments made by the groups (result described on Figure 3.14). We observed that there seem to be a connection between these two values in the sense that the expert mainly disagree on their judgements where the values on the criteria are very unbalanced. Conversely, the experts agree on their judgements when the value on the criteria are balanced in particular when they are both good or bad.

	Group 2	Group 3	Group 4	Group 5	Group 6
Group 1	60%	60%	55%	35%	65%
Group 2		60%	45%	50%	60%
Group 3			75%	70%	80%
Group 4				40%	25%
Group 5					60%

Table 3.10: Percentage of different assignments of scenarios by pairs of group.

	Group 2	Group 3	Group 4	Group 5	Group 6
Group 1	5%	0%	0%	5%	0%
Group 2		10%	5%	5%	15%
Group 3			5%	5%	10%
Group 4				0%	0%
Group 5					0%

Table 3.11: Percentage of different assignments of scenarios by pairs of group with a difference of two categories or more.

	Group 2	Group 3	Group 4	Group 5	Group 6
Group 1	6%	4%	9%	6%	13%
Group 2		5%	9%	4%	16%
Group 3			7%	4%	13%
Group 4				3%	3%
Group 5					9%

Table 3.12: Percentage by pairs of groups of pairs of scenarios that are not ranked in the same order by the two groups.

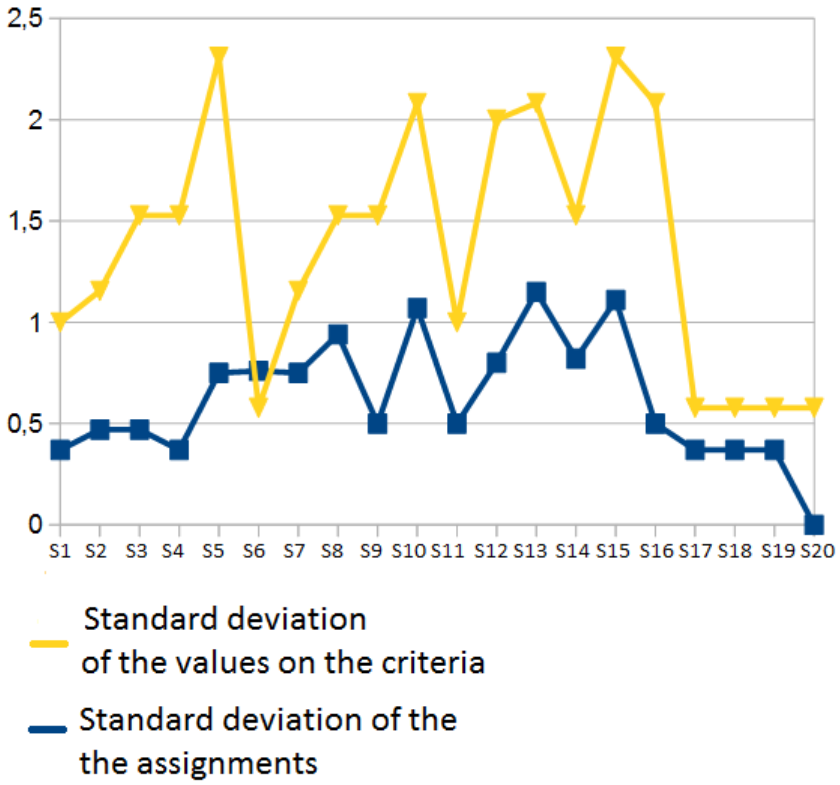

Figure 3.14: Illustration of the interrelationship between the balance between criteria and the convergence on the assignment by the different groups. The abscissa represents the different scenarios while the ordinate represents the standard deviation of the values on the criteria and the standard deviation of the assignments of scenarios to categories. Here while talking of assignments we refer to the category, for each scenario and each group of experts, to which the 4 points or more were given.

Synthetic learning set

As we just saw, the assignments made by the groups of experts may be similar on some scenarios but are different on others. However, the preference learning algorithms that we will use require the use of one single learning set (scenarios and assignments). Hence, we decided to aggregate these learning sets into a synthetic one. The synthetic preference information is constructed on all the experts by summing the points of all the experts for every scenario and every category. Then, for each scenario the median category is chosen i.e $min\{c' \in C | \sum_{\substack{c' \in C \\ c' \leq c}} \tau(c') \geq \frac{1}{2} \sum_{c' \in C} \tau(c')\}$, where $\tau(c)$ is the sum for a given scenario of all the points that were given to the category c for all the experts. The synthetic assignment is shown on Table 3.13.

Scenarios	Dest Pot	Val Env	Vuln	S-LS
S1	C4	C2	C3	C3
S2	C5	C3	C5	C5
S3	C4	C1	C2	C2
S4	C4	C2	C1	C2
S5	C1	C5	C5	C1
S6	C2	C3	C3	C3
S7	C2	C4	C4	C4
S8	C3	C4	C1	C3
S9	C5	C3	C2	C5
S10	C5	C1	C4	C5
S11	C3	C4	C2	C3
S12	C3	C1	C5	C3
S13	C2	C5	C1	C2
S14	C3	C1	C4	C3
S15	C5	C1	C1	C3
S16	C4	C5	C1	C4
S17	C2	C1	C2	C1
S18	C4	C3	C4	C4
S19	C3	C4	C4	C4
S20	C4	C5	C4	C5

Table 3.13: Table defining the scenarios that were proposed to the groups of experts and the synthetic assignment that was found (S-LS).

Violations of monotonicity

Over the questionnaire that were filled by the experts of the MNHN, two of them contained violation of the monotonicity (a dominates b but is assigned in a worse category) despite the small number of scenarios. These violation of monotonicity can be observed in Appendix C (group 1 and 6). The synthetic learning set does not contain any violation of monotonicity.

Comparison of the preference learning algorithms

In order to choose the most adapted algorithm for the elicitation of this sub-problem, we decided to apply a k-fold validation on the synthetic learning set with each of the

4 selected elicitation algorithms for the sorting problem (namely MR-Sort, UTADIS, DRSA and Dominance Based Monte Carlo).

The k-fold validation, later presented in more detail in Subsection 4.3.2, is a method that aims at measuring how efficient is a preference learning algorithm. Basically, it consists in learning the preference of a decision maker with only one part of the learning set ($\frac{k-1}{k}$ times the size of the learning set), then predicting the rest of the assignments (the part that we did not look at) and finally calculates which proportion of these predictions were misclassified. The idea behind that choice is that if two algorithms after having learnt from this learning set give two different assignments we cannot know which result would be the closest to the expert's judgement but we can assume that an algorithm that would be able to predict accurately an assignment that was formulated without looking at it may be accurate concerning an assignment that was not formulated. The percentage of misclassification during the k-fold validation is presented on Table 3.14[2][3][4].

	DRSA	MRSORT	UTADIS	DBMC
k=2	65%	49.5%	67%	39.5%
k=10	60%	42%	62%	37,8%
k=20	40.9%	45%	45%	39,5%

Table 3.14: Result of the k-fold validation applied with 4 algorithms on the synthetic data set obtained from the MNHN experts on the evaluation of the scenarios. Here the percentage of misclassification is written in each cell.

As we can see on Table 3.14, the results of the k-fold validation are rather bad with every algorithms. However, we can observe that the results with the Dominance Based Monte Carlo algorithm are relatively better than with the other two.

3.5.4 Aggregating the criteria with a MR-Sort model

As we saw on this learning set, the MR-Sort preference learning algorithm and the Dominance Based Monte Carlo algorithm seems to reflect better the preference of the

[2]The .isf files to run the tests can be found at https://drive.google.com/file/d/0B5VxOh1ccEY5QXpOaTNWcDZCYTA/view?usp=sharing

[3]The files used to run the tests as well as the test's code for UTADIS and MRSORT can be found at https://drive.google.com/file/d/0B5VxOh1ccEY5aFgyWHZGY2VJa28/view?usp=sharing

[4]The files used to run the tests as well as the test's code for the DBMC algorithm can be found at https://drive.google.com/file/d/0B5VxOh1ccEY5Sl90bDhZZWVRc2M/view?usp=sharing

expert that the other two algorithms. Then, between these two algorithms we think that MR-Sort affords the advantage of being based on an understandable MCAP while, as the reader will see in chapter 4, the DBMC algorithm may be seen by a user as a black box. Indeed, in public decision making and big stakes problem in general it is important for the actors involved in the process to understand how and why a decision is being taken. If our methodology is used by a public organisation, it may affect the fact that a project of an industrial plant may be accepted as it is or not. To either the public authorities, the industrial manager of potential opponent to this project there would be a need for justification.

We chose to use the heuristic algorithm to find the MR-Sort parameters (presented in Subsection 2.3.6) which could be criticised arguing that the algorithm based on a mixed integer programming (presented in Subsection 2.3.5) would have returned a better MR-Sort model. By a better MR-Sort model we mean a model that returns a result closer to the expressed learning set. Indeed, both the mixed integer programming and the heuristic algorithm for induction of a MR-Sort model try to minimize the number of object that are assigned to different categories in the learning set and in the result of the algorithm. Yet, we will show in Table 3.17 that the scenarios that are misclassified cannot be assigned in the good category with any MR-Sort model. Thus, we believe that a mixed integer programming algorithm would have given a similar result. We preferred to use the heuristic rather than the algorithm based on a MIP because we wanted to compare this algorithm to others and thus we performed a k-fold validation. In a k-fold validation the elicitation algorithm is repeated many times and thus a low complexity algorithm such as the heuristic was preferred to the one based on a MIP which has a higher computational complexity.

Description of the rule

Applying the preference learning algorithm described in Subsection 2.3.6 we obtained the following MR-Sort parameter set:

- The weights $w_{PotDest} = w_{Vuln} = w_{ValEnv} = 0.333$
- The concordance level is $s = 0.5$
- The profiles described in Table 3.15

In order to use the MR-Sort method we assume here that the values on the criteria are expressed as integer values ("$C3$" $\rightarrow 3$).

	Dest Pot	Val Env	Vuln
Profile b_5	4.67	4.53	3.80
Profile b_4	3.30	3.76	3.66
Profile b_3	2.05	2.80	2.69
Profile b_2	2.03	1.46	1.71

Table 3.15: Profiles delimiting the categories expressed on a continuous scale as provided by the MR-Sort heuristic algorithm.

We deduce from the voting powers and the concordance level that in order to outrank a profile, an object should be higher than this profile on at least two criteria (regardless to which two criteria). Given that the criteria take integer values, an object a will be sorted in category c if it is at least as good as the profile associated on two or three criteria and if it at least as good as the category above c on zero or one criterion only, the profiles being defined on ordinal scales as described on Table 3.16:

	Dest Pot	Val Env	Vuln
Profile b_5	$C5$	$C5$	$C4$
Profile b_4	$C4$	$C4$	$C4$
Profile b_3	$C3$	$C3$	$C3$
Profile b_2	$C3$	$C2$	$C2$

Table 3.16: Profiles delimiting the categories expressed on ordinal scales (lower bounds).

We remind the reader that the definition of the value levels on the vulnerability, the value of the environment, the destructive potential and the local biodiversity can be found in Subsection 3.2.4.

Comparison of the synthetic learning set with the results given by the MR-Sort model

In order to check if the MR-Sort that obtained is appropriate to this aggregation we tested whether or not it returns the synthetic learning set that was made with the experts questionnaires. We saw on table 3.17 that the MR-Sort model does not return the same assignment as the synthetic learning set in 4 scenarios. These scenarios are very unbalanced scenarios for which the perception of the groups did not converge. This difference between the learning set and the result of the MR-Sort method probably comes from the fact that a scenario that is very severe (resp. not severe at all) on two

criteria will be considered as very severe (not severe at all) regardless to the value of scenario of the third criterion. Yet, this abstraction of one criterion is probably not made by the experts. Then, these 4 unbalanced scenarios cannot be represented by a MR-Sort model unless the concordance threshold is equal to one (or superior to 0.666 which is equivalent) and a concordance threshold equal to one creates a very inflexible model that would probably not return many other preferences expressed in the learning set. Thus, we decided to deal with this issue by adding vetoes on the criteria which leads to the creation of an other MR-Sort model.

Scenarios	Dest Pot	Val Env	Vuln	S-LS	MR-Sort
S1	C4	C2	C3	C3	C3
S2	C5	C3	C5	C5	C5
S3	C4	C1	C2	C2	C2
S4	C4	C2	C1	C2	C2
S5	C1	C5	C5	C1	C5
S6	C2	C3	C3	C3	C3
S7	C2	C4	C4	C4	C4
S8	C3	C4	C1	C3	C3
S9	C5	C3	C2	C5	C3
S10	C5	C1	C4	C5	C5
S11	C3	C4	C2	C3	C3
S12	C3	C1	C5	C3	C3
S13	C2	C5	C1	C2	C1
S14	C3	C1	C4	C3	C3
S15	C5	C1	C1	C3	C1
S16	C4	C5	C1	C4	C4
S17	C2	C1	C2	C1	C1
S18	C4	C3	C4	C4	C4
S19	C3	C4	C4	C4	C4
S20	C4	C5	C4	C5	C5

Table 3.17: Table defining the scenarios that were proposed to the groups of experts, the synthetic assignment that was found (S-LS) and the result obtained using the MR-Sort model.

3.5.5 A MR-Sort model with vetoes

Although the use of a veto threshold is generally not considered in MR-Sort, some variants of this method allow it [Leroy et al., 2011]. Following the previously described

reasoning, we decided to adopt a MR-Sort model with vetoes. Let us remind the reader that by this, we mean that for some categories there may exist a veto profile v such that an object a cannot be assigned to them (or to a higher category) if a does not dominates v.

We also opted for an elicitation in which only one expert was involved. Indeed, during the first three meetings, most of the participant just came once. Thus, we though that their understanding of the problem, the model and the criteria was not sufficient to base our elicitation on them. Hence, we chose to have one other meeting with only one expert Alexandre Robert, that participated to the three first meetings.

At the beginning of this meeting, we started by recalling the definition of each criterion, each value level on the criteria and comparing these value levels to the various reference points mentioned in Subsection 3.2.5.

Then, we asked the expert to assign 13 scenarios to categories of severity. The scenarios proposed to the expert as well as the assignments that he made are presented in Table 3.18.

Scenarios	Dest Pot	Val Env	Vuln	LS
S1	C2	C5	C1	C2
S2	C3	C1	C2	C1
S3	C4	C3	C5	C3
S4	C3	C3	C3	C2
S5	C2	C4	C5	C3
S6	C4	C2	C3	C2
S7	C3	C1	C4	C2
S8	C4	C4	C2	C3
S9	C2	C2	C5	C1
S10	C5	C3	C1	C3
S11	C3	C4	C3	C4
S12	C3	C5	C2	C4
S13	C4	C5	C5	C5

Table 3.18: Table of the learning set obtained from the expert at the fourth meeting. The column LS represents the answers of the experts on the scenarios that were proposed to him.

We wanted to introduce some veto profiles as described in Subsection 2.2.2 and we identify two possible veto phenomenons for the sorting problem. One that we will here

call positive veto phenomenon avoids a scenario with a low value on a given criterion ("C1" for instance) to be assigned to a high category regardless to its values on the other two criteria. For instance, we could say that such a veto phenomenon exists if a scenario with a value of "C2" ("rather low value") or lower on the criterion *value of the environment* cannot be assigned to the category "C3" ("medium pollution") or higher. The other one, that we will here call negative veto phenomenon, avoids a scenario with a high value on a given criterion ("C5" for instance) to be assigned to a low category regardless to its values on the other two criteria. For instance, we could say that such a veto phenomenon exists if a scenario with a value of "C5" ("very high value") on the criterion *destructive potential* cannot be assigned to the category "C2" or lower ("rather low pollution").

In order to test the presence of vetoes we added several new scenarios presented in Table 3.19.

In our context, if there was no veto positive phenomenon then, regardless to the value of the profiles b_i, the scenario $a=$ ("Total annihilation of the local biodiversity" C5, "Very low value" C1, "Very high vulnerability" C5) should be assigned to category $C5$ given that it would be more severe or as severe than the profile delimiting the categories $C4$ and $C5$ on two criteria (whatever this is it cannot be more severe than $C5$ in these two criteria).

Yet, the expert told us that this scenario should according to him be assigned to the category $C2$. Indeed, according to his perception even the total disappearing of the biodiversity located on a very poor target cannot be considered as more severe than $C2$. The reader should pay attention to the fact that we are only measuring the impact on this precise target and not on other surrounding targets that may be impacted as well. This information allows us to think that there exists a positive veto phenomenon on the criterion "value of the environment".

Similar questions were asked and similar conclusions were made with the other two criteria. Therefore, we conclude that in this sub-problem a MR-Sort model should be applied with a positive veto phenomenon for the three criteria. On the criteria "destructive potential" and "vulnerability" a value C1 avoids the scenario to be assigned to category C4.

At the opposite, if there would be a negative veto threshold on the criterion *Value of the environment* scenario $b=$ ("No impact on biodiversity" C1, "Very high value" C5, "Very low vulnerability" C1) should not be assigned to the category $C1$. As the reader may see on the Table 3.19 the last three scenarios having a $C5$ evaluation and two $C1$ evaluation are all assigned to $C1$ "No impact or negligible pollution".

In order to represent this positive veto phenomenon we decided to model this sorting problem as a MR-Sort model with vetoes as presented in Subsection 2.2.2. We model MR-Sort in such way that the value level $C1$ is "the worst" and $C5$ is "the best" for the three criteria and the categories. Indeed, in MR-Sort, veto thresholds are used so that it may avoid an object a to outrank a given profile b_h. Thus, if we consider that the meaning of a outranks a profile b_h is that "the scenario a is at most as severe as the scenario represented by the profile b_h" then the effect of adding veto profiles would be to create a negative veto phenomenon. Conversely, if the outranking relation means "is at least as severe" then it creates a positive veto phenomenon. In our situation, a positive veto phenomenon being desired, we will then consider that the outranking relation aSb means "a is at least as severe b"

Scenarios	Dest Pot	Val Env	Vuln	LS
S14	C1	C5	C5	C3
S15	C5	C1	C5	C2
S16	C5	C5	C1	C3
S17	C5	C1	C1	C1
S18	C1	C5	C1	C1
S19	C1	C1	C5	C1

Table 3.19: Table of the learning set proposed to the expert to test the presence of veto phenomenons. The column LS represents the answers of the experts on the scenarios that were proposed to him. We can see with the scenarios 14, 15 and 16 that there is a positive veto phenomenon while the last three show the absence of a negative veto phenomenon.

We decided to use an algorithm for MR-Sort based on a mixed integer programming close to the one described in Subsection 2.3.5 on the learning set made of the 13 first assignments plus the 6 assignments proposed to test the veto phenomenons. The method that we used consists in first applying the algorithm "MRSortIdentifyIncompatibleAssignments" that find the largest subsets of the Θ (the objects in the learning set) such that there exists a MR-Sort model (with or without veto depending on what the user prefers) which is compatible with the assignments given in the learning set. The user can then chose which assignment of the learning set she choose to abandon and use the mixed integer programming algorithm to find a MR-Sort model compatible with the chosen subset of Θ. The choice of a mixed integer programming in our case was justified by the fact that it offers a guaranty to respect as well as possible the learning set which is not true for the heuristic method.

Two subsets of size 17 were proposed by "MRSortIdentifyIncompatibleAssignments"

to run the elicitation algorithm: The one excluding the scenarios $\{S1, S4\}$ and $\{S4, S9\}$. By the way, we can see here the advantage of including vetoes. Indeed, applying "MRSortIdentifyIncompatibleAssignments" without including vetoes, we see that the biggest subsets of the learning set that can be respected without including vetoes are of cardinality 12. In any case, if we wanted to exclude only two preferences of the learning set, the scenario $S4$ is to be deleted. The question arise then to decide whether the scenario $S1$ or $S9$ should also be deleted to find a compatible model. We observed that a destructive potential equal to "weak impact on biodiversity" (expected on an averagely vulnerable target) combined with a "very low vulnerability" may not have any impact on the biodiversity which make coherent the assignment to the category $C1$ despite the "very high value of the environment". Therefore, abandoning the scenario $S1$ seems more reasonable than abandoning the scenario $S9$.

We applied the elicitation algorithm for MR-Sort based on the mixed integer programming on the learning set minus the scenarios $S1$ and $S4$ and we obtained the following MR-Sort parameter set:

- The weights $w_{PotDest} = w_{Vuln} = w_{EnvVal} = 0.333$
- The concordance level is $s = 0.5$
- The profiles described in Table 3.20
- The veto profiles described in Table 3.21

In order to use the MR-Sort method we assume here that the values on the criteria are expressed as integer values, ("$C3$" $\rightarrow 3$).

	Dest Pot	Env Val	Vuln
Profile b_5	4.501	4.501	3.499
Profile b_4	3.000	4.001	3.000
Profile b_3	2.501	2.501	2.500
Profile b_2	2.001	2.001	2.001

Table 3.20: Profiles delimiting the categories expressed on a continuous scale as provided by the MR-Sort heuristic algorithm.

	Dest Pot	Val Env	Vuln
Veto profile v_5	1.001	3.001	1.001
Veto profile v_4	1.001	3.001	1.001
Veto profile v_3	-0.001	2.001	-0.001
Veto profile v_2	-0.001	-0.001	-0.001

Table 3.21: Veto profiles expressed on a continuous scale as provided by the MR-Sort heuristic algorithm.

Once again given that the criteria take integer values, an object a will be sorted in category c_j if it outranks the profile b_j but not b_{j+1} and that the outranking relation aSb_j means that a is higher than b_j on two or three criteria and higher than the veto profile v_j on the three criteria the profiles and the veto profiles can be defined on ordinal scales as shown on Tables 3.22 and 3.23:

	Dest Pot	Val Env	Vuln
Profile b_5	$C5$	$C5$	$C4$
Profile b_4	$C3$	$C5$	$C3$
Profile b_3	$C3$	$C3$	$C3$
Profile b_2	$C3$	$C3$	$C3$

Table 3.22: Profiles delimiting the categories expressed on ordinal scales (lower bounds).

	Dest Pot	Val Env	Vuln
Veto profile v_5	$C2$	$C4$	$C2$
Veto profile v_4	$C2$	$C4$	$C2$
Veto profile v_3	$C1$	$C3$	$C1$

Table 3.23: Veto profiles of the categories expressed on ordinal scales (lower bounds).

The resulting assignment of the object in the learning set by this MR-Sort model are shown in Table 3.24.

Scenarios	Dest Pot	Val Env	Vuln	LS	MR-Sort
S1	C2	C5	C1	C2	C1
S2	C3	C1	C2	C1	C1
S3	C4	C3	C5	C3	C3
S4	C3	C3	C3	C2	C3
S5	C2	C4	C5	C3	C3
S6	C4	C2	C3	C2	C2
S7	C3	C1	C4	C2	C2
S8	C4	C4	C2	C3	C3
S9	C2	C2	C5	C1	C1
S10	C5	C3	C1	C3	C3
S11	C3	C4	C3	C4	C4
S12	C3	C5	C2	C4	C4
S13	C4	C5	C5	C5	C5
S14	C1	C5	C5	C3	C3
S15	C5	C1	C5	C2	C2
S16	C5	C5	C1	C3	C3
S17	C5	C1	C1	C1	C1
S18	C1	C5	C1	C1	C1
S19	C1	C1	C5	C1	C1

Table 3.24: On this table, the column LS represents the answers of the experts on the scenarios that were proposed to him. MR-Sort represents the assignments obtained by the MR-Sort model with vetoes.

While looking at the profiles and the category profiles exposed on tables 3.22 and 3.23 the reader can observe that the profiles b_2 and b_3 and the veto profiles v_4 and v_5 are identical. This means that the profile b_3 and the veto profile v_5 has no impact on the MR-Sort assignment. The frontier between the category $C2$ and $C3$ is determined only by the veto profile v_3, while the frontier between the category $C4$ and $C5$ is determined only by the profile b_5. However, this feature does avoid these two categories to contain several scenarios in the MR-Sort assignment of the learning set.

The reader may observe on the Table 3.24 that the thus obtained MR-Sort model does indeed not return the scenarios $S1$ and $S4$ which as the method MRSortIdentifyIncompatibleAssignments suggested would not be returned. However, they are both assigned to adjacent categories with the MR-Sort model and in the learning set.

This second model is different from the first one on mainly two points. At first,

we abandoned the idea of including the preferences of a large number of experts that would have been aggregated in one single learning set. Instead, we worked with one single expert that was the most familiar with the topic. This allowed us to have better interactions and to obtain a monotonic learning set that is quite compatible with a MR-Sort model. In similar situations, I would recommend to analysts to interact either with one expert or with one group of expert that does not change over time.

The second difference is the use of a veto threshold that allows more flexibility for the model which can then get closer to the expert's judgment. In our situation, the learning set being small, adding veto threshold did not bring additional difficulties nor problematic computing time.

As a result, while the first MR-Sort model had 4 misclassifications over a learning set of 20 scenarios (20%), this second one has 2 misclassification over a learning set of 17 scenarios ($\approx 11.7\%$). Furthermore, in this second model, the scenarios that are being misclassified are always assigned to an adjacent category. For instance, the scenario (C1, C5, C5) was part of both learning sets, in the first learning set it was assigned to category 1 while in the second learning set it was assigned to category 3 (according to the idea that a scenario that would have "No impact expected on an averagely vulnerable target" may have an impact on a very vulnerable target). The first model assigned this scenario to category 5 while the second assigned it to category 3 due to the use of the veto profiles.

3.6 Aggregation of the Biodiversity Severity Index

We now have a procedure that, considering a given scenario of toxic leak and a given target that may be impacted, if a user can provide the input criteria *concentration of liquid in the target*, *toxicity of the liquid*, *mobility of the water*, *persistence of the liquid*, *environmental value of the target* and *vulnerability of the target*, would allow us to obtain an evaluation of the *local biodiversity severity index*. Our final goal is to obtain the Biodiversity Severity Index which would represent the global expected severity of a toxic leak on the whole surrounding biodiversity. Indeed, as stated earlier we think that the impact on the use must be treated separately.

3.6.1 Description of the aggregation procedure

The *local biodiversity severity indices*, the *biodiversity severity index on surface waters* and the *biodiversity severity index* measure values of the same nature. The only difference between them is their frame. Hence, it seems appropriate to represent the *biodiversity severity index on surface waters* and the *biodiversity severity index* with the same scale as the one used to represent the *local biodiversity severity indices* i.e. discrete scale of five value levels: "no impact or negligible pollution", "low pollution", "medium pollution", "serious pollution", "Ecological disaster". In order to aggregate together the various *local biodiversity severity indices* we cannot use the multi-criteria tools described in this document. Indeed, we do not know how many *local biodiversity severity indices* may exist and thus, how many criteria we would have. This aggregation was not properly studied and what follows must be considered as a trail for this aggregation. We believe that the *biodiversity severity index on surface waters* and the *biodiversity severity index* should be at least as severe as the worst *local biodiversity severity index*. Furthermore, it looks natural to think that a scenario s_1 that would impact one small target with a local biodiversity severity of c would be less severe that a scenario s_2 that would impact a larger target with a local biodiversity severity of c. Here, the massive aspect of a pollution is considered. As a possible threshold to decide whether a pollution is massive or not we decided to tell that if a pollution affects with a similar severity an area of more than 10 square kilometres it can be considered as massive (regardless to it intensity on the target that are impacted). Thus, one possibility would be to fix the following rule to evaluate the *biodiversity severity index on surface waters* and the *biodiversity severity index*:

- If the most impacted target is impacted with local biodiversity severity of c and the total area of the targets that are impacted with local biodiversity severity of c is lower than 10 square kilometre, then the *biodiversity severity index on surface waters* is equal to the category c.
- If the most impacted target is impacted with local biodiversity severity of c which is lower than $C5$ and the total area of the targets that are impacted with local biodiversity severity of c is higher than 10 square kilometre, then the *biodiversity severity index on surface waters* is equal to the category above c.
- If the most impacted target is impacted with local biodiversity severity of $C5$ then the *biodiversity severity index on surface waters* is equal to the category $C5$ regardless to the global area of the targets for which the impact is equal to $C5$.

Then it looks natural to say that the *biodiversity severity index* is equal to the max of the *biodiversity severity index on surface waters* and the *biodiversity severity index on ground targets*.

3.6.2 Practical example of the evaluation of a scenario of accidental pollution

We will now illustrate how the *biodiversity severity index* could by build while performing a risk evaluation (illustrated in Figure 3.15). The example that we are about to study is totally fictive and is not based on the study of a real industrial plant. Let us assume that a factory of alimentary preservative, in Créteil, Val de marne (94000), is subject to a risk study that includes risks of accidental pollution. After the description of the industrial plant was made a list of the possible scenarios is defined. Among them the scenario a represents a leak of biphenyl. If the leak a would happen it would reach the lake of Créteil. Let us study the *local biodiversity severity index* of the scenario a on this target. At first we will evaluate the destructive potential of the leak. The

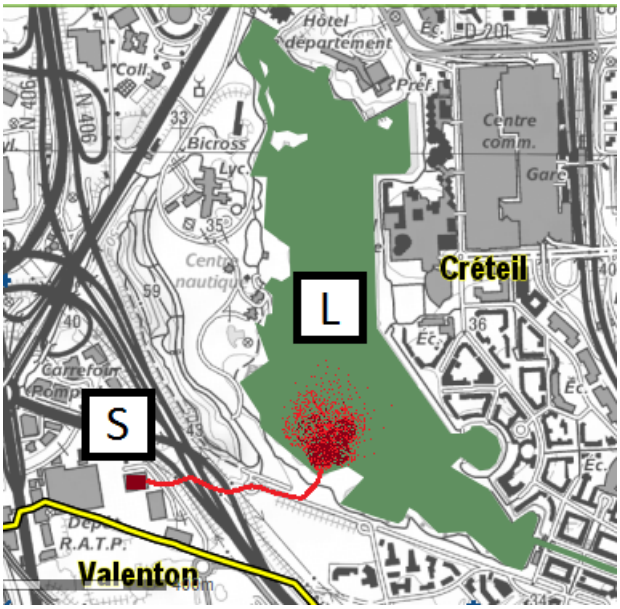

Figure 3.15: Illustration of the scenario of toxic leak in the lake of Créteil. **S** represents the source of the leak and **L** represents the lake of Créteil.

expected concentration of biphenyl in the lake after the scenario happen was estimated within a range from 33 ± 5 $\mu g\,l^{-1}$, with the use of physic models. The water of the lake is static and biphenyl is a persistent product, hence the *residence time* is defined

as "long". The maximal acceptable concentration of the biphenyl is equal to 4 $\mu g\,l^{-1}$ [5]. Then, according to the rules defined in subsection 3.4.3, the destructive potential of the scenario a on the lake of Créteil is equal to $C3$, "relatively large impact on the biodiversity". Although the lake of Créteil is not a protected area, according to the @d maps of ordinary and remarkable biodiversity in Île de France, it is considered as having a high level of ordinary biodiversity (0.8 on a scale of 0 to 1) and a level of remarkable biodiversity higher than 0 (1 on a scale of 0 to 4) which is a rare feature in cities located at less that 10 kilometres from Paris. Thus, let us assume that the expert or the group of experts that would have participated to the risk study estimated the value of the environment to "medium value" $C3$. Furthermore, given that the lake of Créteil is no connected to other surface waters the experts estimated the vulnerability of the lake as "rather high vulnerability", $C4$. According to these values and to the MR-Sort model defined in Subsection 3.5.4 the value of the *local biodiversity severity index* is equal to "medium pollution", $C3$ (illustrated in Figure 3.16). The severity of

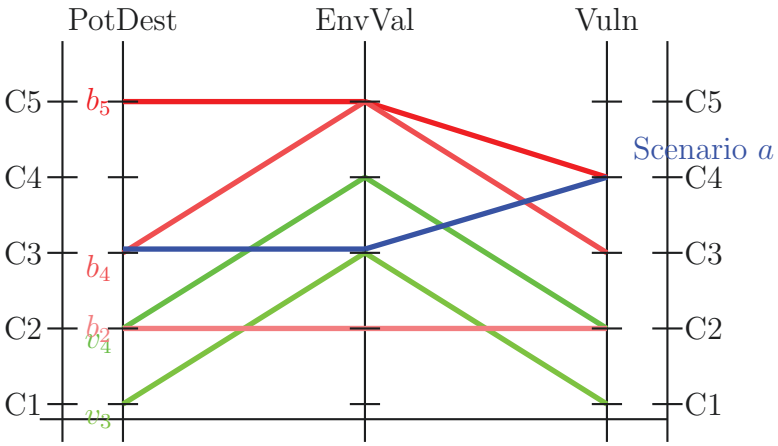

Figure 3.16: Graphic representation of the evaluation of the local biodiversity severity index using the MR-Sort model found in Subsection 3.5.4. Here the profiles are represented by the red lines while the veto profiles are represented by the green lines. The scenario described in this section is represented by the blue line. The profile b_3 and the veto profile v_5 were removed from this figure because they have no impact of the sorting due to their redundancy with the lower profile and veto profile. We can see here that this scenario should be assigned to category C3.

the scenario a on other targets is lower that $C3$ as no other possibly impacted target

[5]Data obtained on the INERIS website https://www.google.fr/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjz9YOl6OHOAhVFIsAKHXHYAfMQFggeMAA&url=http%3A%2F%2Fwww.ineris.fr%2Fsubstances%2Ffr%2Fsubstance%2Fpdf%2F30&usg=AFQjCNF66HOTrZSWMxogEL0_cPRn-K7fqA&sig2=l5GYL-6aj6AXyyOTiIE0LA

has an important biodiversity value and the area of the lake of Créteil is lower than 1 square kilometre. Thus the *biodiversity severity index* of the scenario a is estimated to $C3$ medium pollution. This example is illustrated with the values on the different criteria on Figure 3.6.2. Then according the evaluation of its likelihood, the scenario a may be considered as acceptable for biodiversity or not. If the risk is acceptable for the biodiversity and the resources then we may say that the environmental risk is acceptable.

Conclusion

In this chapter, we proposed a method that allows a possible user to evaluate the expected severity of a scenario of accidental pollution. We described the steps that led us to it:

1) Finding the appropriate family of criteria. This task was made by interacting with several experts trying to identify all the elements that may influence what we are trying to measure. For that purpose, the *value focused thinking* concept [Keeney, 1992] was very useful to us, in particular the questions that Keeney advises to ask to the decision maker, in order to create a constructive common reflection on the values.

2) We organized the criteria into a hierarchy of criteria. We did so trying to understand what makes these criteria important for us and the relation that exists between them either mean-ends relation of specifications as defined in Subsection 2.1.6. Once again this task was mainly influenced by the value focused thinking concept.

3) We proposed for each criterion i a scale to represent the different values that an object a may have on these i and that allow. This scale is also a tool that allows that different actors of this process to communicate about the criteria. In our situation, these scales are mainly ordinal, some are continuous and standardized (the *concentration* and the *maximal acceptable concentration*) others are discrete and have an intuitive meaning (for instance the vulnerability is expressed with the scale {very low vulnerability, rather low vulnerability, medium vulnerability, rather high vulnerability, very high vulnerability, }). In the second case it was important to carefully choose the appropriate number of value levels in these scales and to create reference points so that the subjectivity inherent in this type of scale can be decreased.

4) We aggregated chose at each sub-problem of the hierarchy of criteria an aggregation method and a set of preference parameters which we believe is adapted to the experts's preferences (or perception).

Three aggregations are applied to find the *destructive potential*, the *local biodiversity severity indices* and the *biodiversity severity index*. The aggregation that leads to the *destructive potential* is a rule based method that is found through a direct elicitation (aggregation approach) with an expert in toxicology. The possible aggregation that leads to the *biodiversity severity index* is a slightly modified max that would take into account the total area that is impacted. The aggregation that leads to the creation of the *local biodiversity severity indices* is an MR-Sort model that was found through a direct elicitation (aggregation approach) with experts from the MNHN. However this model is an evolution of an MR-Sort model that was found through an indirect elicitation (disaggregation approach) the experts of the MNHN. For this elicitation, we compared several methods for preference learning and one of them, the Dominance Based Monte Carlo algorithm, was created by me during this thesis. We finally opted for the use of an MR-Sort model which provides an understandable aggregation procedure which in a public decision context is an important advantage. Nevertheless, the reader can observe on the Figure 3.14 that the results of the k-fold validation applied to the learning set obtained from the MNHN experts is sensibly better with the Dominance Based Monte Carlo algorithm than with the other tested algorithms. We believe that this preference learning algorithm is interesting and may in other contexts be used for real world applications. Hence, the next chapter aims at introducing the Dominance Based Monte Carlo algorithm, studying its theoretical properties and its practical performances.

The main perspectives concern the adaptation of the destructive potential on ground target and the improvement of the aggregation of the several Local Biodiversity Severity Indices into a single Biodiversity Severity Index. The second one being more subjective than the other matters that we dealt with here, it may be interesting to call for decision makers (either authorities, citizens, industrial managers etc) instead of experts.

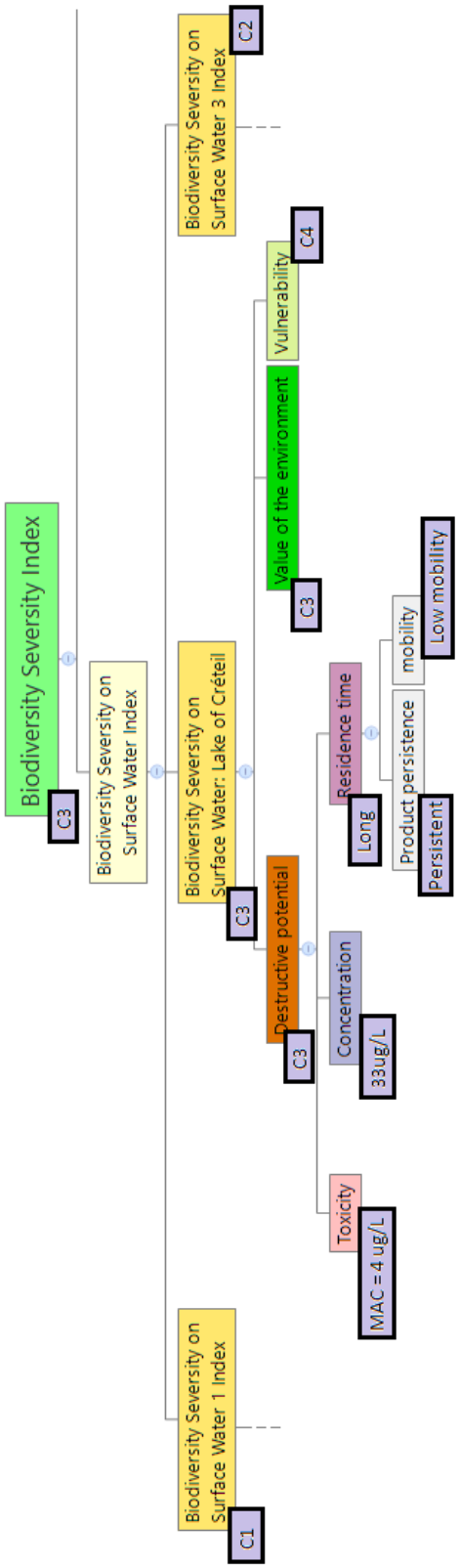

Figure 3.17: Illustration of the values of the scenario of toxic leak on the criteria.

Chapter 4

Dominance Based Monte Carlo algorithm

Contents

4.1 Basic ideas about the Dominance Based Monte Carlo algorithm . . . 166

4.1.1 Using the monotonicity and the learning set as a frame . . . 166

4.1.2 A model free approach . . . 167

4.1.3 A stochastic functioning . . . 167

4.1.4 From the existing methods in the literature to the Dominance Based Monte Carlo algorithm . . . 168

4.2 Description of the algorithm . . . 169

4.2.1 Theoretical basis and notation . . . 170

4.2.2 Process of a trial . . . 172

4.2.3 Collecting and using the trial's information . . . 178

4.2.4 Aggregating the DBMC Vector . . . 179

4.2.5 Expected properties . . . 181

4.2.6 Summary of the theoretical properties . . . 185

4.2.7 Theoretical properties of the AverageDBMC and the ModeDBMC . . . 185

4.3 Experimental validations and comparison to other sorting algorithms . . . 187

4.3.1 Stability . . . 187

4.3.2 Presentation of the k-fold cross validation 190
4.3.3 Comparison of the DBMC algorithm with other elicitation algorithms through a k-fold validation 192
4.4 What are the problems to which this algorithm can be adapted? 201
4.5 Perspectives 202
4.5.1 Modified idiosyncrasy: An other experimental test for the efficiency of the algorithm 202
4.5.2 Possible applications of the Dominance based Monte Carlo . 203

"On voit, par cet Essai, que la théorie des probabilités n'est, au fond, que le bon sens réduit au calcul; elle fait apprécier avec exactitude ce que les esprits justes sentent par une sorte d'instinct, sans qu'ils puissent souvent s'en rendre compte.",
Pierre-Simon de Laplace, Théorie analytique des probabilités , 1812.

During the conception of this thesis, we proposed an algorithm for multi-criteria elicitation of the sorting problem (as defined in Subsection 2.1.3). We named it the Dominance Based Monte Carlo algorithm (DBMC algorithm). As explained in Section 4.4, we saw that this elicitation algorithm was efficient to restore a learning set provided in the sub-problem of the local biodiversity severity indices although an other algorithm was finally preferred to it. In this chapter we will first mention the underlying ideas of this algorithm before describing its functioning. We will analysis its mathematical properties and show some results of the tests that we made to evaluate its performances and compare it to other elicitation algorithms.

4.1 Basic ideas about the Dominance Based Monte Carlo algorithm

The Dominance Based Monte Carlo algorithm (DBMC) is an elicitation algorithm for the sorting problem. It requires the use of discrete finite scales on the criteria and it is based on examples of assignments given by the decision maker (such as "I think that the object a should be assigned to category 2, the object b should be assigned to category 1..."). This could be considered as a disaggregation approach (as defined in Subsection 2.3.2). We can find in literature several methods of elicitation by a disaggregation approach for multi-criteria sorting problem as presented in Section 2.2. Their attractiveness comes from the fact that they only require a judgement that may fit well with the decision maker's expectations. The DBMC method is based on monotonicity and the respect of the learning set which are very common in multi-criteria decision aiding but presents two specificities: a probabilistic approach and the absence of a proper MCAP, which can be found in existing methods but are usually not coupled.

4.1.1 Using the monotonicity and the learning set as a frame

In multi-criteria decision aiding, the monotonicity principle as defined in Subsection 2.2.1 is well accepted as a desirable property and non respect is considered as pathological. The respect of the learning set is generally considered as a good property in multi-criteria decision aiding. We say that an assignment respects the learning set if every object assigned in the learning set is assigned in the same category. Most of the disaggregation elicitation methods for multi-criteria sorting problem respect the learning set as long as this learning set is compatible with their associated MCAP.

For this reason, these two principles will be the base of this algorithm as detailed in Section 4.2. Indeed, while in other multi-criteria methods these two concepts are only seen as good properties, in this algorithm, as we will later see, they are imposed and thus compose the only real frame of this method.

4.1.2 A model free approach

Ian Stewart [Stewart and Cohen, 1995] says that "If our brains were simple enough for us to understand them, we'd be so simple that we couldn't". This joke illustrates well the incapacity of man to understand his own reasoning. In particular, the human judgement is probably too complex to be described by simple rules and may not be totally deterministic according to many authors, mainly from psychology and behaviour analysis [Regenwetter et al., 2011][Luce, 1995][Carbone and Hey, 2000]. In some contexts a MCAP may be a good frame that help the decision maker to build a coherent reasoning. But in others it could be that none of the existing MCAP can be rigorously defended. However, as described in Section 2.3 most of the elicitation algorithms for the sorting problem are structured around a MCAP and thus a set of explicit rules although some methods such as ORCLASS are not. One idea of the *Dominance Based Monte Carlo algorithm* was to use a stochastic functioning based on a Monte Carlo principle to extend the learning set given by the decision maker to every possible object without being framed by a MCAP.

4.1.3 A stochastic functioning

The idea of this algorithm is to propose an assignment that may be seen as an "average" (not in the arithmetic mean sense) of all the assignments that respect monotonicity and the learning set by generating several random assignments that respect these two properties and aggregating them into a synthetic assignment. To do so, the principle of this algorithm consists in generating randomized assignments that respect the learning set and monotonicity to aggregate them in a synthetic assignment. The justification for this method comes from the idea (that can be disputed) that only the assignments that respect the monotonicity and the learning set are valid and that hence, an assignment that would have a status of centrality among these valid assignments could be considered as a reasonable one. However, this non-deterministic property of the output may look disturbing. Indeed human generally have difficulties to accept that decisions are taken randomly (more so if the decision has important consequences). Nevertheless,

we will prove theoretically that the result of our method converges almost surely and we will observe that in practice the convergence is rather quick and effective.

4.1.4 From the existing methods in the literature to the Dominance Based Monte Carlo algorithm

The method that we are about to describe was mainly influenced by three existing methods namely DRSA described in Subsection 2.3.9, ORCLASS described in Subsection 2.3.8 that can both be seen as model free and SMAA described in Subsection 2.2.5 that has stochastic functioning.

Indeed the functioning of SMAA methods is based on a Monte Carlo process just as the DBMC but two main differences are to be highlighted. At first the SMAA methods mainly deal with the ranking problem with the notable exceptions of SMAA classification [Salminen et al., 2007] that deals the nominal classification problem (non ordered categories) and SMAA TRI [Figueira et al., 2004] which consists in a robustness analysis for the sorting problem. Then, SMAA methods are based on a MCAP and at each trial the preference parameters of this MCAP is set randomly. As stated earlier the DBMC algorithm differentiates from the SMAA methods by not being based on a MCAP so as to be more flexible in term of capacity to accept preference information and to return results that do not fit with a specified frame.

ORCLASS method and DRSA both are based on monotonicity and are model free methods in the sense that they are not framed by a MCAP (although DRSA finds a set of rule but it can represent any monotonic preferences as demonstrated earlier). Two properties made DRSA difficultly compatible with our problem. At first, DRSA may return a set of rules that assign some objects to an interval of categories which is not an acceptable output in many contexts (for instance in ours) where one category is expected. Then, the objective in DOMLEM (an algorithm mainly used in DRSA method to find set of rules) to obtain a set of rules "as simple as possible" may lead to an oversimplified set of rules that would not represent the decision maker's reasoning.

The ORCLASS method has an important influence to the DBMC methodology. Indeed there is the idea of adding successively some new objects to the learning set and observing the impact on the remaining possible monotonic assignments which, as explained later, is the main idea of the DBMC algorithm. However, as we demonstrated in Subsection 2.3.8 this method is mainly adapted to problems with 2 categories and interactions with the decision maker may be too numerous if the number of possible

objects (combinations of criteria) is not particularly low. Thus, the idea of this algorithm was to finish arbitrarily the ORCLASS process a large number of times so as to find a kind of average of the assignments found.

We may mention that some tournament method for social choice [Moulin et al., 2016] uses probabilistic processes through markov chains. However, both the goal that is intended in these tournaments (social choice) and the method that is being used are very different from what I am going to propose in this chapter.

Finally, the method which according to me is the most similar to what I am going to present is the logistic and choquistic regression approach. Although no random experiment is made in these methods, the obtained assignment arise from a probability distribution. As the reader will later see, the Dominance Based Monte Carlo algorithm approximates the probability at each random step (here named a trial), for a given object a to be sorted in a given category c, repeating several times this random step. In logistic and choquistic regression, this probability is supposed to be calculated through a regression process. Thus there is no need to approximate it. However, this probability is supposed to follow a predefined pattern (in logistic regression $\mathbb{P}(f_k(a) = 1) = (1 + exp(-k_0 - k_1 \cdot g_1(a) - ... - k_n \cdot g_n(a))^{-1}$ for instance).

Our will was to create a method that is not subject to any pattern in order to make it adaptable to any type of preference that a decision maker may express as long as monotonicity is respected. We did so by combining the Monte Carlo principle existing in SMAA methods and the model free property of DRSA and ORCLASS.

4.2 Description of the algorithm

Monte Carlo algorithms [Metropolis and Ulam, 1949] [Bortz et al., 1975] [Cazenave and Helmstetter, 2005] [Sloan and Woniakowski, 1998] are a family of algorithms that are used in many applications such as integral calculation, artificial intelligence or to simulate the time evolution of some processes occurring in nature. These algorithms are based on two main steps: Making a large number of randomized and independent experiments, that we will later call trials, and draw conclusions or take a decision in the light of some observations and statistics on the results of these trials.

The Dominance Based Monte Carlo algorithm works as follows. First we choose randomly an object a among A and we assign it randomly to a category among the category in which it can be assigned without violating monotonicity according the other objects that are already assigned. Then we choose an other object randomly and assign

it randomly to a category and so on and so forth until every each object is fixed in a category. This random assignment will be considered as a *trial*. Obviously, randomness has an important impact in this assignment, which is considered as a bad characteristic in multi-criteria decision aiding. In order to reduce this impact and to converge we make a large number of trials, probabilistically independent from each other, we collect the information about the consecutive results of the trials into a vector that we will later call a DBMC Vector and we aggregate it to get a single synthetic assignment. I would like to emphasise the fact that this algorithm requires the use of discrete finite scales on all the criteria.

4.2.1 Theoretical basis and notation

Notation for the sorting problem

We define a sorting context as a $5-tuple$ $S =< N, V, A, C, L >$ containing:

- $N = \{1, ..., n\}$ is a finite set of criteria.
- Each criterion i is expressed on a discrete and finite scale $v_i = \{v_{i,1}, ..., v_{i,\beta_i}\} \subset \mathbb{N}$. We will thereafter assume that all the criteria are to be maximized. Let $V = \prod_{i \in N} v_i$, β_i being the number of grades in the scale of the criterion i.
- A is the set of objects to be sorted. Here it is considered that any combination of values on the criteria must be sorted i.e., $A = \prod_{i \in N} v_i$ and $|A| = m = \prod_{i \in N} \beta_i$. We denote by $g_i(a), a \in A, i \in N$ the value of a on the criterion i.
- $C = \{1, 2, ..., r\} \subset \mathbb{N}$ is a set of r ordered categories in which the objects are to be sorted. We will thereafter assume that the higher a category is, the better it is.
- L is a set of learning set given by the decision maker with $L= < \Theta, f_l >$, $\Theta \subseteq A$ being the Learning Examples Set i.e., the set of objects that are supposed to be assigned by the decision maker and $f_l : \Theta \to C$ being the assignment of these examples. Basically we will consider here that the objects are assigned to one category in the learning and not to an interval of categories.
- The expected output of this problem is an assignment. An assignment is a function $f : A \to C$ that assign every possible object a to a category $f(a)$. Thereafter, we will denote by $DBMC(S, T)$, the result of the Dominance Based Monte Carlo algorithm on the sorting context S with T.

Monotonicity

Monotonicity being a central principle of this algorithm we here recall some formal definitions and notation about monotonicity although some are already defined in Subsection 2.2.1:

Given two objects $a, b \in A$, we say that a weakly dominates b if a is at least as good as b on every criteria. This relation is also denoted by aDb. We will thereafter simply use the term "dominates" to mention "weak domination".

For any object $a \in A$ we call dominating cone of a, also denoted by $D^+(a) = \{a' \in A : a'Da\}$, the set of all objects that dominate a. For any object $a \in A$ we call dominated cone of a, also denoted by $D^-(a) = \{a' \in A : aDa'\}$, the set of all objects that are dominated by a. We say that an assignment f respects monotonicity if, for any $a, b \in A$ such as aDb, we have $f(a) \geq f(b)$.

The learning set

Given that we impose the respect of monotonicity and the learning set, the possible remaining assignments are then constraint to a smaller space. This space will obviously be empty if the learning set does not respect monotonicity. Therefore, from now on, we assume that the learning set respects monotonicity unless we explicitly say otherwise. At the end of this chapter we will see how non-monotonic learning set can be modified to become monotonic. For instance, if $a'Da$ and aDa'' with a' already assigned to category 4 and a'' already assigned to category 2, then we can deduce that a can only be assigned to category 2, 3 or 4 (we can say that a can be assigned to the interval of categories $[2, 4]$). In order to introduce this notion of "interval assignment" not violating the monotonicity and respecting the learning set we define the notion of a "necessary interval assignment", also denoted by $\gamma : A \to \Delta$, (Δ being the set of category intervals). In other words this necessary assignment represents the fact that any object a will be assigned to a category at least as good as the best sorted object that is dominated by a and at most as good as the worst sorted object that weakly dominates a.

An object a is said fixed with an interval assignment γ if $\gamma_{min}(a) = \gamma_{max}(a)$. We say that an assignment f is compatible with an interval assignment γ, also denoted by $f \sqsubset \gamma$, if $\forall a \in A$, $\gamma_{min}(a) \leq f(a) \leq \gamma_{max}(a)$.

Figure 4.2 shows the interval assignments that we may have in the beginning of DBMC algorithm. In this figure, we have two criteria (10 levels for each) and three categories.

As we see 5 objects belong to the learning sets (one of them in category 1, two of them in category 2 and two of them in category 3). These five assignments constrain the interval assignments of the remained objects : red ones in category 3, orange-colored ones in interval $[2, 3]$, yellow ones in category 2, light green ones in interval $[1, 2]$, ...

	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	[1,2]	[1,2]	[1,3]	[1,3]	[1,3]	[1,3]
2	1	1	1	1	[1,2]	[1,2]	[1,3]	[1,3]	[1,3]	[1,3]
3	1	1	1	1	[1,2]	2	[2,3]	[2,3]	[2,3]	[2,3]
4	[1,2]	[1,2]	[1,2]	[1,2]	[1,3]	[2,3]	[2,3]	[2,3]	[2,3]	[2,3]
5	[1,2]	[1,2]	[1,2]	[1,2]	[1,3]	[2,3]	[2,3]	[2,3]	[2,3]	[2,3]
6	[1,2]	[1,2]	[1,2]	2	[2,3]	[2,3]	[2,3]	[2,3]	[2,3]	[2,3]
7	[1,3]	[1,3]	[1,3]	[2,3]	[2,3]	[2,3]	[2,3]	[2,3]	3	3
8	[1,3]	[1,3]	[1,3]	[2,3]	[2,3]	3	3	3	3	3
9	[1,3]	[1,3]	[1,3]	[2,3]	[2,3]	3	3	3	3	3
10	[1,3]	[1,3]	[1,3]	[2,3]	[2,3]	3	3	3	3	3

Figure 4.1: Illustration of the necessary interval assignment with two criteria with ten value levels on each, three categories and 5 objects in the learning set (in the squares)

4.2.2 Process of a trial

In the DBMC algorithm each trial will be a random completion of the learning set that respects monotonicity. To do so, we iteratively choose a random object and assign it to a random category among the categories in which it could be sorted. Formally it

works as explained in algorithm 6.

Algorithm 6: Random completion - trial

Data: Sorting context $S =< N, V, A, C, L >$

Result: Sorting $f_t : A \to C$ monotonic and compatible with L

for *All* $a \in \Theta$ **do**
  $f_t(a) \leftarrow f_l(a)$
while $\exists a \in A$ *such that* $\gamma_{min} \neq \gamma_{max}$ **do**
  Choose randomly an object χ with a uniform distribution over A;
  Choose randomly a category Δ with a uniform discrete distribution between $\gamma_{min}(\chi)$ and $\gamma_{max}(\chi)$;
  Add the information $< \chi, \Delta >$ to the learning set;
  $\gamma_{max}(\chi) \leftarrow \Delta$;
  $\gamma_{min}(\chi) \leftarrow \Delta$;
  for *All* $a^- \in D^-(\chi)$ **do**
    $\gamma_{max}(a^-) \leftarrow min\{\Delta, \gamma_{max}(a^-)\}$
  for *All* $a^+ \in D^+(\chi)$ **do**
    $\gamma_{min}(a^+) \leftarrow max\{\Delta, \gamma_{min}(a^+)\}$
for *All* $a \in A$ **do**
  $f_t(a) \leftarrow \gamma_{max}(a)$ (or $\gamma_{min}(a)$ they are equal at this point)

We say that we add an information $< a, c >$, $a \in A, c \in C$ to the learning set when we impose the fact that the object a should be sorted in category c, i.e., $\Theta \leftarrow \Theta \cup \{a\}$ and $f_l(a) \leftarrow c$.

Uniform distribution

We say that a randomly chosen object χ follows a uniform distribution over $A' \subseteq A$, also denoted by $\chi \sim \mathbb{U}(A')$, if $\forall a', a'' \in A'$, $\mathbb{P}(\chi = a') = \mathbb{P}(\chi = a'')$. While implementing the DBMC algorithm, we chose randomly the next object to be randomly sorted χ by choosing randomly and independently each of its values on the criteria. The value of χ on the criterion i follows a uniform discrete law on v_i, i.e., $g_i(\chi) \sim \mathbb{U}(v_i)$. We are about to prove that this is equivalent to say that χ follows a uniform law on A.

Proposition 1 *A random object χ follows a uniform distribution over A if and only if, on each criterion $i \in N$, $g_i(\chi)$ follows a uniform distribution over v_i i.e., $\chi \sim \mathbb{U}(A) \Leftrightarrow \forall i \in N, g_i(\chi) \sim \mathbb{U}(v_i)$ and $\forall i, j \in N, g_i(\chi)$ is independent to $g_j(\chi)$.*

Proof : $\Leftarrow$ Let $\chi \in A$ and let χ be a random object such as $\forall i \in N, g_i(\chi) \sim \mathbb{U}(v_i)$ and $\forall i, j \in N, g_i(\chi)$ is independent to $g_j(\chi)$. Then $\forall i \in N, \forall a \in A$, $\mathbb{P}(g_i(\chi) = g_i(a)) = \frac{1}{\beta_i}$ and $\mathbb{P}(\chi = a) = \prod_{i \in N} \frac{1}{\beta_i} = \frac{1}{m}$ (β_i being the number of value levels for the criterion i).

$\Rightarrow$ Let us consider that $\chi \sim \mathbb{U}(A)$ then $\forall i \in N, \forall \alpha \in v_i$, the number of objects χ such as $g_i(\chi) = \alpha$ is equal to $\prod_{\substack{i \in N \\ j \neq i}} \beta_j = \frac{m}{\beta_i}$ then the probability $\mathbb{P}(g_i(\chi) = \alpha) = \frac{1}{\beta_i}$. ■

Properties of the random completion

Regarding each trial, there are several properties that we would like to test. We want the learning set and monotonicity to be respected and we want the algorithm to end properly in the sense that there will not be infinite loops and we will not get to a point where, for a reason, it is not possible to finish the random completion process. It seems quite obvious that the learning set is respected given that, by definition, these assignments are imposed from the beginning. Concerning the risk of an infinite loop, the only loop that exists here will stop when all the alternatives are fixed. At each iteration, the object that is chosen is fixed. Thus there cannot be any infinite loop. Furthermore the only case in which the process could not keep running would be if for a given object a the lower bound would be higher than the higher bound (bound inversion). We are about to show that this case cannot happen and that the obtained assignment respects monotonicity condition.

Proposition 2 *If we assume that the learning set respects the monotonicity, then it will not be stopped by a bound inversion and the assignment f_t obtained by Algorithm 6 respects also the monotonicity.*

Proof : We will show that at each step, the interval assignment γ verifies $\forall a, b \in A$ if aDb then $\gamma_{max}(a) \geq \gamma_{max}(b)$, $\gamma_{min}(a) \geq \gamma_{min}(b)$ and $\forall a \in A$, $\gamma_{max}(a) \geq \gamma_{min}(a)$.

Let us assume that at a given step St the interval assignment respects these properties. Let us now assume that we add the object $a \in A$ in category c such that $\gamma_{min}(a) \leq c \leq \gamma_{max}(a)$ (bounds of the object a before it was added to the learning set).
Now let a^+ and a^- be two objects of A such as $a^+ D a^-$, we have different possibilities:

i. $a^+, a^- \in A \backslash (D^-(a) \cup D^+(a))$: nothing changed in γ for a^+ and a^- and there couldn't be new violation of monotonicity about these objects nor could $\gamma_{min}(a^+)$ become higher than $\gamma_{max}(a^+)$.

ii. $a^+ \in A\backslash(D^-(a) \cup D^+(a))$ and $a^- \in D^+(a)$: not possible given that $a^+ D a^-$.

iii. $a^- \in A\backslash(D^-(a) \cup D^+(a))$ and $a^+ \in D^-(a)$: not possible given that $a^+ D a^-$.

iv. $a^- \in A\backslash(D^-(a) \cup D^+(a))$ and $a^+ \in D^+(a)$: nothing changes for $\gamma_{max}(a^+)$, $\gamma_{min}(a^-)$ and $\gamma_{max}(a^-)$. $\gamma_{min}(a^+)$ might increase so no new violation of the monotonicity with respect to $\gamma_{min}(a^-)$. If $\gamma_{min}(a^+)$ increases then, $\gamma_{min}(a^+) = c \leq \gamma_{max}(a) \leq \gamma_{max}(a^+)$ then we still have $\gamma_{max}(a^+) \geq \gamma_{min}(a^+)$.

v. $a^+ \in A\backslash(D^-(a) \cup D^+(a))$ and $a^- \in D^-(a)$: similar to the item iv.

vi. $a^+, a^- \in D^+(a)$ then $\gamma_{min}(a^+) \leftarrow max\{\gamma_{min}(a^+), c\}$ while $\gamma_{min}(a^-) \leftarrow max\{\gamma_{min}(a^-), c\}$. Initially $\gamma_{min}(a^+) \geq \gamma_{min}(a^-)$ then $max\{\gamma_{min}(a^+), c\} \geq max\{\gamma_{min}(a^-), c\}$. Moreover $c \leq \gamma_{max}(a) \leq \gamma_{max}(a^+)$ then $\gamma_{max}(a^+) \geq \gamma_{min}(a^+)$ (same with a^-).

vii. $a^+, a^- \in D^-(a)$: similar to the item vi.

Thus these two properties are still respected at the step $St + 1$. If the learning set is monotonic then the interval assignment at the beginning of the process is the same as if all the elements of the learning set were added successively to an empty learning set. The initial interval assignment as well as the following ones and thus the output assignment of the random completion will then by recurrence also respect these two properties. ■

Distribution of a trial

As we just saw, the result of a trial is subject to randomness. Thus, the question arises to know whether or not we can find its distribution. It appeared to us that the distribution of a trial could theoretically be found modelling this problem with the use of a Markov chain where the state space would be all the possible learning sets that are compatible with the original learning set and that are monotonic. Indeed, at each step of the algorithm (that can be represented as a sorting context), the probability distribution of the next step only depends on the current step. However we decided for now to leave beside the calculation of this distribution through the use of a Markov chain. Indeed this task seems to be extremely difficult due to mainly two obstacles:

- The transition matrix seems very complex to model. Indeed, the state space is made of possible learning sets i.e., each state would be a pair $< \Theta, f_l >$, where $\Theta \subseteq A$ would be a subset of A and where $f_l : \Theta \to C$ would be a function of Θ

in C. This form makes these states difficult to use as the indices of a transition matrix.

- Furthermore, the number of states in the state space of this Markov chain would be difficult to calculate precisely but in any case, it is higher than the number of subsets of A which is equal to 2^m.
- Finally, this Markov chain is not positive recurrent given that the size of the learning set increases at each step. Thus there cannot be no stationary probability distribution on this Markov chain but instead, to calculate the distribution of the result of a trial, we would have to calculate the limit of the transition matrix power T when T tends to infinity which given the complexity of the transition matrix seems to be an extremely difficult task.

Nevertheless, even not knowing exactly this distribution, some problems can be solved. We wanted to know if every assignment that respects monotonicity and the learning set could be returned with a probability higher than 0. We are about to prove that the answer to this question is "yes". We also wondered if every assignment that respects monotonicity and the learning set would have the same probability to be returned. And we found that the answer to this question is "no".

Proposition 3 *While doing a random completion, any assignment that is compatible with the current necessary interval assignment and respects monotonicity can be chosen with a probability higher than* $\frac{1}{r^m}$ *(r being the number of categories and m being the number of objects).*

Proof : Let us denote by $\{a_1, a_2, ..., a_m\}$ the permutation of objects of A representing the order in which these objects are selected by Algorithm 6 and let us consider an assignment f_0. Then, given that during a random completion, at each one of the m steps, an object a is assigned randomly to a category with a uniform distribution among $C' \subseteq C$ and $|C| = r$, the probability that the assignment f_t obtained by the random completion is equal to f_0 must be higher than $\frac{1}{r^m}$. Formally $\mathbb{P}(f_t(a_1) = f_0(a_1) \cap f_t(a_2) = f_0(a_2) \cap ... \cap f_t(a_m) = f_0(a_m))$
$=\mathbb{P}(f_t(a_1) = f_0(a_1)) \times \mathbb{P}(f_t(a_2) = f_0(a_2) | f_t(a_1) = f_0(a_1)) \times ... \times \mathbb{P}(f_t(a_m) = f_0(a_m) | f_t(a_1) = f_0(a_1) \cap ... \cap f_t(a_{m-1}) = f_0(a_{m-1})) \geq \frac{1}{r^m}$. ■

We also wondered if every assignment that respects monotonicity and the learning set would have the same probability to be returned. For this, we analysed first of all

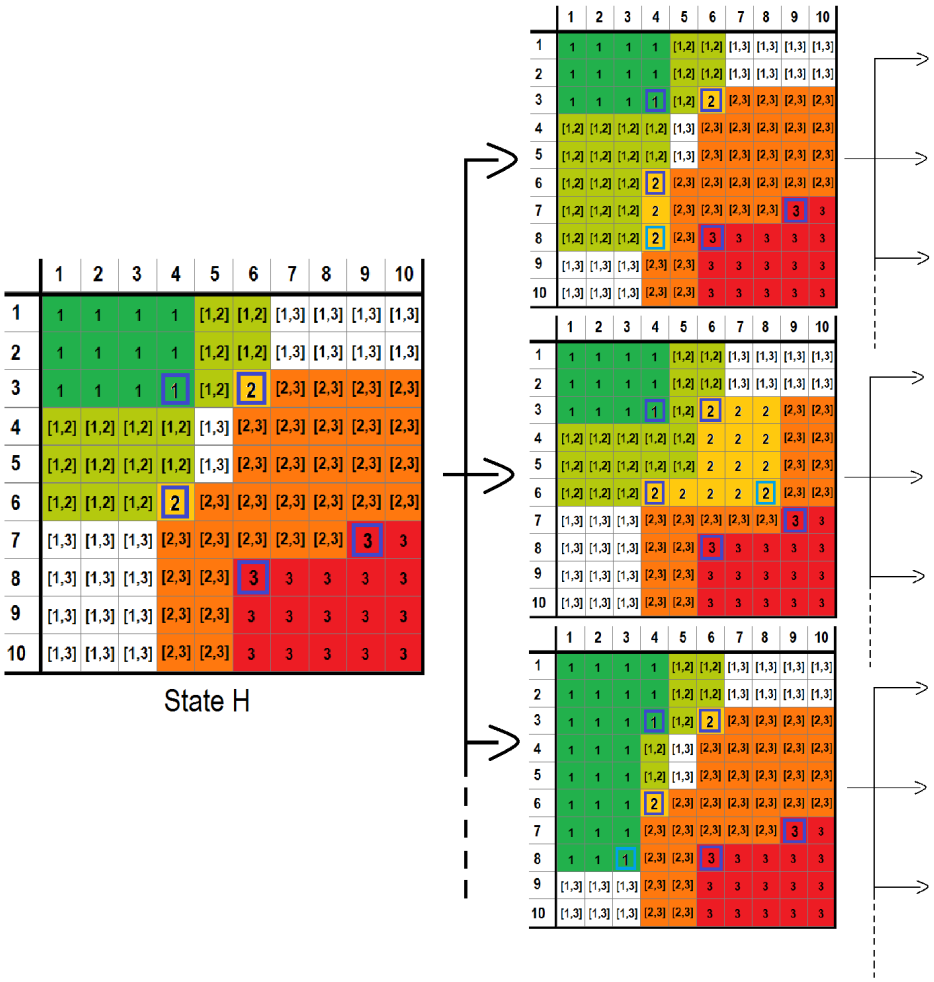

Figure 4.2: Illustration of modelization of the trial distribution through a Markov chain. We can see a state H and some of its following states.

a very special case which has an empty learning set and where the first chosen object of Algorithm 6 is an object having the maximum value in all the criteria (dominating object).

Lemma 1 *If the learning set is empty then, the probability for all the objects to be sorted in category 1 is higher than* $(m \times r)^{-1}$

Proof : If at the first step the dominating object is chosen and the category 1 is

attributed then all the other objects will be sorted in category 1. ■

This lemma shows us that all the assignments do not have the same probability to be found with Algorithm 6.

Proposition 4 *While doing a trial in DBMC from an empty learning set, all the possible assignments do not have the same probability to be obtained at the end of the trial.*

Proof : The number of possible assignments with no training set is strictly higher than $(m \times r)$, thus all the possible assignments cannot have a probability to be returned equal or higher than $(m \times r)^{-1}$. ■

It might have been preferable to use a uniform distribution on the possible assignments for the trials. However, modelling a uniform distribution over the possible assignments would probably require to list all of them first which given their number would make it not applicable from either time or space computational point of view.

Computational complexity of a trial

Theoretically in the worst case the previously described algorithm could run infinitely if for example the same object would be chosen at each step eternally. While programming the algorithm we decided, when the chosen object was already fixed to stop the step and launch it again. This has no impact on the result of the algorithm as adding object where it is already fixed does not change γ. Thereby, if we neglect the computing time of a step that has been aborted, we find that the number of steps in a trial cannot be higher than m. Furthermore, at each step we modify the necessary interval assignment γ (γ_{min} and γ_{max}) at most $|D^+(a) \cup D^-(a)| \leq m$ times. Thus, the computational complexity of a trial is in $O(m^2)$.

4.2.3 Collecting and using the trial's information

To apply the DBMC algorithm, we first complete randomly T times the original sorting context. Then, for every object, we note in a DBMC vector in what category it has been assigned at every trial. Finally, we aggregate these vectors to assign each object to the category in which it has "globally" been assigned during the T trials.

Definition 1 *We call a DBMC Vector of S,* $\varphi : A \to \mathbb{N}^T$ *the vector in which we store the results of* T *random completions of* S *such as* $\forall T \in \mathbb{N}, a \in A, k \in \{1, ..., T\}$, $\varphi_k(a)$ *the category in which* a *has been assigned at the* k^{th} *trial.*

Formally we get a DBMC Vector of S as follows:

Algorithm 7: Building a DBMC Vector

Data: Sorting context $S =< N, V, A, C, L >$
Result: DBMC Vector, $\varphi : A \to \mathbb{N}^T$

```
for j from 1 to T do
    S' ← S;
    Complete randomly S';
    (as defined in Algorithm 6) for a ∈ A do
        φ_j(a) ← f_t(a);
```

For instance, assuming that during the 8 trials of the DBMC algorithm an object a was assigned successively to the categories 2, 3, 5, 3, 4, 3, 4 and 2, we would obtain the DBMC Vector (2, 3, 5, 3, 4, 3, 4, 2).

4.2.4 Aggregating the DBMC Vector

The DBMC Vector cannot be used directly for two reasons. Firstly, the format of this information, $\mathbb{N}^T$, is very uncomfortable and we cannot draw direct conclusion from it. Then, it is subject to randomness. Thereby, once this information is collected we need to aggregate it so that we can assign each object to the category in which it has "globally" been assigned during the T trials. In statistic these methods are generally refereed to as measures of central tendency. We are about to present several methods with which we can perform this task. While computing the algorithm, we stocked this information into a smaller vector that we called *frequency vector*, $FV : A \to \mathbb{N}^r$ such as $\forall c \in C, a \in A$. $FV_c(a)$ represents the number of times that a has been sorted in category c among the T trials. In practice, this change enables us to require less memory space than what we would have required using the DBMC Vector and allows us to apply the algorithm in the same way from every other points of view. Indeed, the measures of central tendency that we use are not influenced by the order of the assignments during the process. Nevertheless, from a theoretical perspective, the use of the OWA operators (presented thereafter) has no relevant signification while applied

on the frequency vector. Therefore here, we will use the DBMC Vector rather than the Frequency vector as a way of stocking the information.

Definition 2 *Let us thereby call a measure of central tendency [Weisberg, 1992] [McCluskey and Lalkhen, 2007]* $\Gamma : \mathbb{N}^T \to \mathbb{N}$ *an operator whose function is to synthesize the vector* z *given as a parameter, that verify the boundary condition i.e.,* $min(z) \leq \Gamma(z) \leq max(z)$ *and the symmetry condition i.e.,* $\Gamma(z_1, z_2, \ldots, z_T) = \Gamma(i_{\pi(1)}, i_{\pi(2)}, \ldots, i_{\pi(T)})$ *if* π *is a permutation map. Generally three measures of central tendency are considered, the mode, the arithmetic mean and the median. We can observe that both the median and the arithmetic mean are particular cases of a more general measure called OWA operators [Yager, 1988]. Let us now define these concepts:*

- *The mode, here denoted by* $Mode(z)$ *returns the number that was found the more often in the vector* z*.*
- *An OWA operator of dimension* T *is a mapping* $OWA : \mathbb{R}^T \to \mathbb{R}$ *that has an associated collection of weights* $W = [w_1, \ldots, w_T]$ *lying in the unit interval and summing to one and with:* $OWA(z_1, \ldots, z_T) = \sum_{j=1}^{T} w_j b_j$ *where* b_j *is the* j^{th} *largest of the* z_i*. As mentioned earlier, the arithmetic mean and the median are two specific examples of OWA operators. Indeed, when* $W = [\frac{1}{T}, \frac{1}{T}, ..., \frac{1}{T}]$ *the OWA operator is equivalent to the arithmetic mean and if the* $W_{\frac{T}{2}} = 1$ *and* $W_j = 0, \forall j \neq \frac{T}{2}$ *the OWA operator is equivalent to the median. We will thereafter denote by* $Average(z)$ *the arithmetic mean and by* $Median(z)$ *the median of the vector* z*.*

- Let us denote by $\Lambda(a)$ the random variable that represents the assignment of the object a by the DBMC algorithm with T trials, using the mode as the measure of central tendency (ModeDBMC algorithm). Formally $\Lambda(a) = Mode(\varphi(a))$
- Let us denote by $\eta(a)$ the random variable that represents the assignment of the object a by the DBMC algorithm with T trials, using the median as the measure of central tendency (MedianDBMC algorithm). Formally $\eta(a) = Median(\varphi(a))$
- Let us denote by $\delta(a)$ the random variable that represents the assignment of the object a by the DBMC algorithm with T trials, using the average as the measure of central tendency (AverageDBMC algorithm). Formally $\delta(a) = Average(\varphi(a))$

For obvious reasons these operators both verify the boundary and the symmetry conditions. The reader should pay attention to the fact that the AverageDBMC algorithm

induces a sum between the values of the categories in C. If the method is used in a context in which the categories have a cardinal meaning it could eventually make sense (if in the context, we consider that the difference between two adjacent categories is invariable). However, in sorting problems the categories generally have an ordinal meaning and then adding categories that could even be defined without the use of numerical values makes no sense. By the way this remark could be done about any OWA operator that would involve a vector W with more than one strictly positive value, given that it would induce a sort of weighted sum. Thus, the only OWA operators that may be accepted in a sorting problem with ordinal categories is an OWA operator with only a 1 and T -1 0's which is equivalent to different percentiles. Given that we do not want to be particularly optimistic nor pessimistic, it seems to us more legitimate to choose the median value. By the way in a two categories sorting context these three variant would be equivalent. Given that, as demonstrated in Subsection 4.2.5 we have no guarantee that the result of the ModeDBMC will be monotonic, the MedianDMBC algorithm of simply DBMC algorithm is the one that is the most appropriate to a multi-criteria sorting context.

Applying the median as the central tendency measure on the example cited above where we would have obtained the DBMC Vector (2, 3, 5, 3, 4, 3, 4, 2), the result of the DBMC algorithm would be $Median((2,3,5,3,4,3,4,2)) = Median((2,2,3,3,3,4,4,5) = 3$.

4.2.5 Expected properties

There are some good properties that we would like our method to fulfil. We would like to know that, when the number of trials grows, the result of the method becomes less aleatory. Thus we will later prove that the results of the methods converge almost surely when T tends to infinity. Other properties that we expect to be satisfied, are to respect the learning set and monotonicity. We will thereby see that it fulfils these conditions.

Convergence

We say that the sequence X_T converges almost surely towards X when $T \to \infty$ if $\mathbb{P}(\lim_{T\to\infty} X_T = X) = 1$. Almost sure convergence implies convergence in probability (by Fatou's lemma) and thus is considered as a strong convergence (although it does not necessarily implies convergence in mean). By definition of the algorithm, it is considered that the $\varphi_k(a)$ (the assignments of the object a over the trials) are i.i.d. For

any $a \in A, c \in C$ et $\rho_{a,c} = \mathbb{P}(\varphi_k(a) = c)$ be the probability on one trial that the object a is sorted in category c. We thereafter demonstrate that $\eta(a)$ almost surely converges to the lowest category $c \in C$ such as $\mathbb{P}(\varphi_k(a) \leq c) \geq \frac{1}{2}$ as $T \to \infty$.

Theorem 1 *The median in probability, here denoted by ϵ_a is the lowest category such as $\mathbb{P}(\varphi_k(a) \leq \epsilon_a) \geq \frac{1}{2}$. $\eta(a)$ converges to ϵ_a a.s when $T \to \infty$.*

Proof : By definition $\eta(a)$ is equal to the lowest category $c' \in C$ such as $\sum_{k=1}^{T} 1_{\{\varphi_k(a) \leq c'\}} \geq \frac{1}{2} \times T$. Thus, for any $c \in C$, $\mathbb{P}(\eta(a) > c) = \mathbb{P}(\sum_{k=1}^{T} 1_{\{\varphi_k(a) \leq c\}} < \frac{1}{2} \times T)$.
Let us call $\alpha_c = \mathbb{P}(\varphi_k(a) \leq c)$.
Then, $\sum_{k=1}^{T} 1_{\{\varphi_k(a) \leq c\}} \sim \beta(\alpha_c, T)$ (binomial distribution with parameters α_c, T)
$\Rightarrow \mathbb{P}(\sum_{k=1}^{T} 1_{\{\varphi_k(a) \leq c\}} < \frac{1}{2} \times T) \to \Phi((\frac{1}{2} \times T - T \times \alpha_c)/(\sqrt{T \times \alpha_c \times (1 - \alpha_c)}))$
$= \Phi((\sqrt{T} \times (\frac{1}{2} - \alpha_c))/(\sqrt{\alpha_c \times (1 - \alpha_c)}))$ when $T \to \infty$ (de Moivre-Laplace theorem, Φ being the standard notation for the cumulative distribution function of the standard normal distribution).
Thus $\frac{1}{2} - \alpha_c > 0 \Rightarrow \eta(a) > c \quad a.s$ and $\frac{1}{2} - \alpha_c < 0 \Rightarrow \eta(a) \leq c \quad a.s$
$\Rightarrow \eta(a) > \epsilon_a - 1$ and $\eta(a) \leq \epsilon_a$ $a.s$
$\Rightarrow \eta(a) = \epsilon_a$. ■

Monotonicity

We prove in this section that monotonicity is respected while using the OWA operators.

Definition 3 *Let us state that the DBMC assignment $f : A \to C$ is such that $f(a) = \Gamma(\varphi(a))$, φ being a DBMC Vector and Γ being a measure of central tendency.*

Proposition 5 *Any DBMC sorting that is based on an OWA operator respects monotonicity.*

Proof : Let us consider $a, b \in A$ such as aDb. Since every random completion is monotonic, we have $\varphi(a) \geq \varphi(b)$ and since OWA operators are monotonic (*i.e.* if $\lambda_0, \lambda_1 \in \mathbb{R}^k$

and $\lambda_0 \geq \lambda_1$ then, for any OWA operator op, we have $op(\lambda_0) \geq op(\lambda_1)$)[Fullér, 1996], we have $f(a) = \Gamma(\varphi(a)) \geq \Gamma(\varphi(b)) = f(b)$. ■

However, as already mentioned, monotonicity property is not necessarily respected with the ModeDBMC algorithm is broadly broken in practice when the learning set is small which is the reason why we prefer using the median DBMC algorithm. Let us give a theoretical example in which it could be broken. Let us assume that in a sorting context there exist two objects a_0, a_1 that are not included in the learning set, such that a_0 is dominated by a_1. Category 1 is the worst category and category 4 is the best. Let us assume that after running the algorithm with 11 trials , a_0 and a_1 are sorted as illustrated in Figures 4.3 and 4.4. This feature is possible, for instance if at each trial a_0 and a_1 are the two objects that are chosen first. As we can see at the second and at the fifth trials the object a_1 is one category higher than a_0 due to the fact that it dominates it. However the object a_0 was sorted with the highest frequency in category 3 (5 times) while the object a_1 was sorted with the highest frequency in category 1 (4 times). Thus, a_0 would be sorted better than the a_1 with the ModeDBMC procedure. With the median value they would both have been sorted in category 2.

Trial	1	2	3	4	5	6	7	8	9	10	11	Mode
a_0	C3	**C3**	C2	C1	**C3**	C1	C1	C3	C1	C3	C2	*C3*
a_1	C3	**C4**	C2	C1	**C4**	C1	C1	C3	C1	C3	C2	*C1*

Figure 4.3: Results of the 11 trials for a_0 and a_1

	C1	C2	C3	C4
a_0	4/11	2/11	**5/11**	0/11
a_1	**4/11**	2/11	3/11	2/11

Figure 4.4: Frequency of assignment for each object and each category

In order to give a practical illustration of the violations monotonicity while using the ModeDBMC we applied it on a model with 2 criteria expressed on scales of 40 value levels, 7 categories, 100 trials and an empty learning set (represented on Figure 4.5). In each cell we wrote the category in which each object was sorted. Indeed, on the graphic representation of the result, we can observe that monotonicity is widely violated. However, we also observe that, when increasing the number of trials, these violation of monotonicity generally disappear as illustrated in Figure 4.6, applying the previously described experience with 100 000 trials.

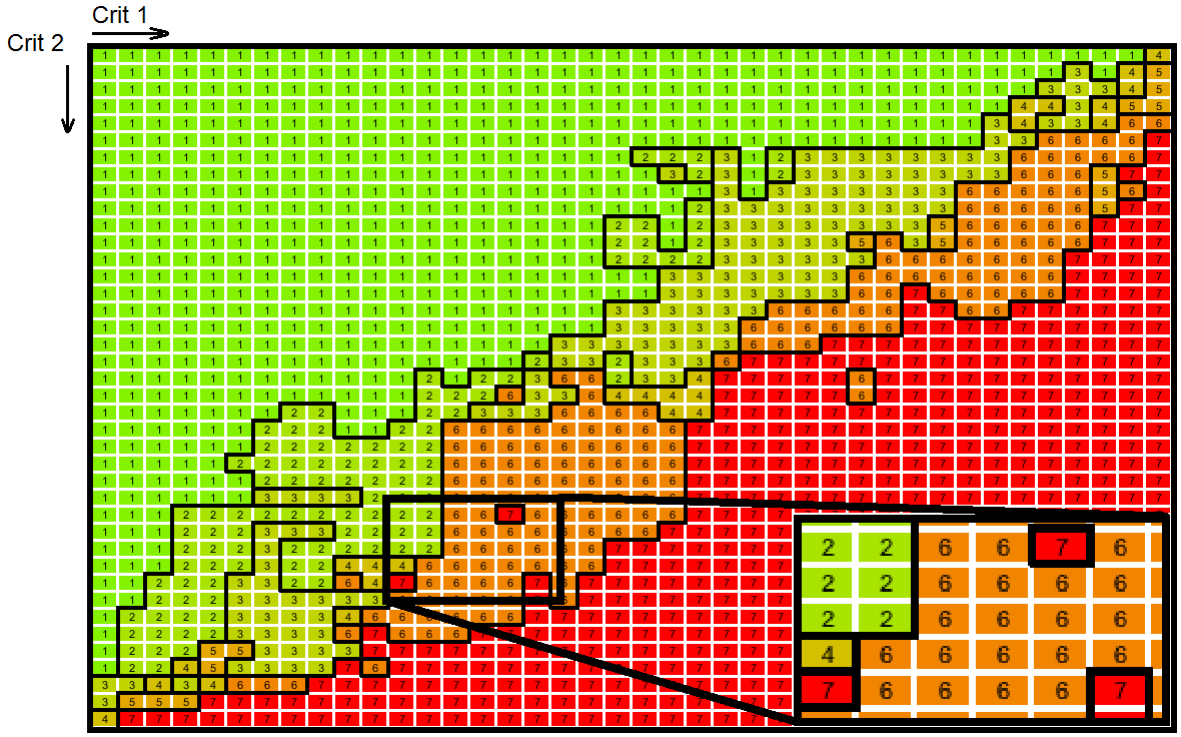

Figure 4.5: Illustration of violations of monotonicity with the modeDBMC in a practical test with 2 criteria (40×40), 7 categories and 100 trials.

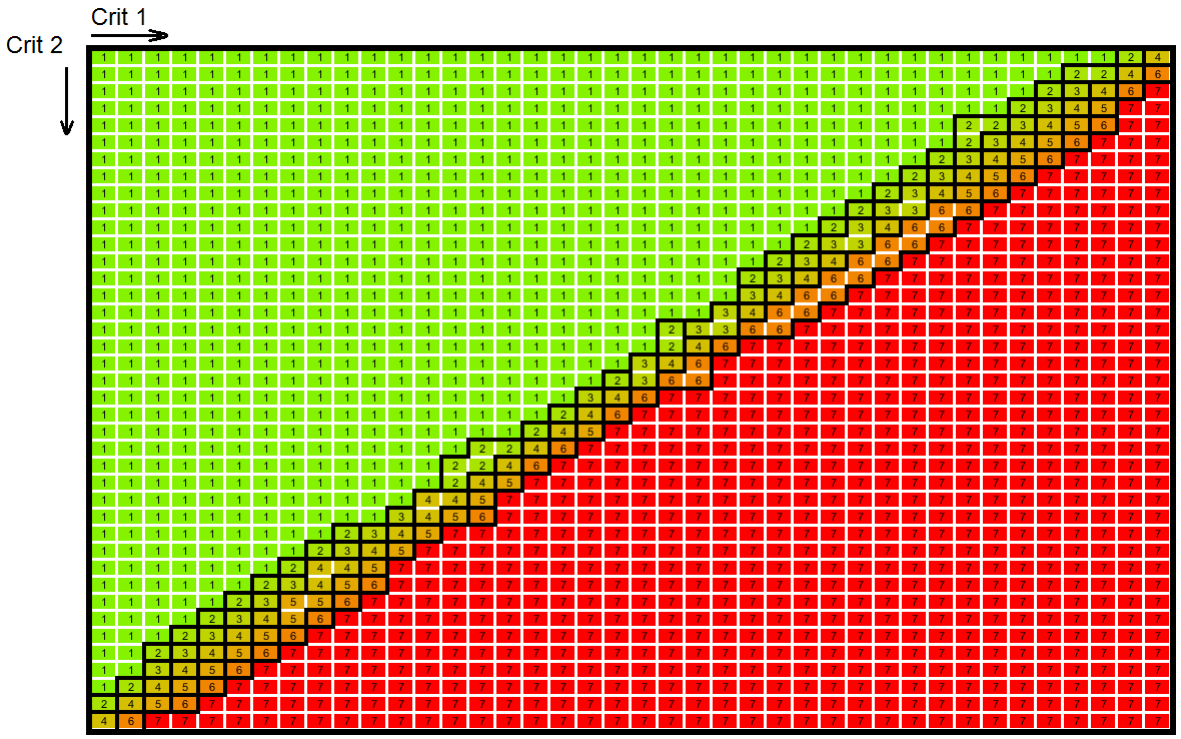

Figure 4.6: Illustration of the disappearance of violations of monotonicity with the modeDBMC while increasing the number of trial in a practical test. Here there are 2 criteria (40×40), 7 categories and 100 000 trials.

Respect of the learning set

Proposition 6 *While using the DBMC algorithm the learning set is respected.*

Proof : As we saw, at each trial the learning set is respected. Thus, if an object a is included in the learning set the DBMC Vector will contain T times the category in which it has been sorted by the decision maker *i.e.* $\varphi(a) = \{f_l(a), ..., f_l(a)\}$. Since we defined the aggregation operators such as $min(z) \leq \Gamma(z) \leq max(z)$ then $\Gamma(\varphi(a)) = f_l(a)$. ■

4.2.6 Summary of the theoretical properties

To summarise the previous section we can notice several theoretically demonstrated properties regarding the MedianDBMC algorithm.

- Monotonicity is respected.
- The learning set is respected.
- Despite its non-deterministic nature, this algorithm converges almost surely to the median in probability.
- Regarding the distribution of a trial, at each step every possible assignment may be found but not uniformly.
- The worst case complexity of this algorithm is in $O(m^2 \times T)$.

4.2.7 Theoretical properties of the AverageDBMC and the ModeDBMC

Although we explained earlier the theoretical properties that make the MedianDBMC algorithm our favourite one, the AverageDBMC algorithm and the ModeDBMC algorithm may be interesting in some contexts. Thus, here are the theoretical proofs practised on the DBMC algorithm (Median DBMC algorithm) that may be interesting to study for its two variants. We saw that the monotonicity is respected with the AverageDBMC given that the average aggregation is an OWA operator while the ModeDBMC is not necessarily monotonic. As well we are about to prove that the convergence almost sure demonstrated for the DBMC algorithm can also be demonstrated with its two variants.

Convergence with the ModeDBMC

We will thereafter prove that $\Lambda(a)$ almost surely converges to $c \in C$ such as $\rho_{a,c}$ is the highest among C as $T \to \infty$ ($\rho_{a,c} = \mathbb{P}(\varphi_k(a) = c)$). To do so we will prove that when $T \to \infty$ the more probable it is that an object get sorted in a category the more often it will be sorted in that category.

Proposition 7 *The number of time that a was sorted in category c among T trials follows a binomial distribution i.e.* $FV_c(a) \sim \beta(\rho_{a,c}, T)$

Proof : All trials are i.i.d. and $1_{\{\varphi_k(a)=c\}}$ follows a bernouilli distribution thus $FV_c(a) = \sum_{k=1}^{T} 1_{\{\varphi_k(a)=c\}} \sim \beta(\rho_{a,c}, T)$. ■

Proposition 8 $\forall c_i, c_j \in C$ *such as* $\rho_{a,c_i} > \rho_{a,c_j}$ *then* $\underset{T\to\infty}{Lim}\ FV_{c_i}(a) > \underset{T\to\infty}{Lim}\ FV_{c_j}(a)$ *a.s.*

Proof : Let us consider δ such as $\rho_{a,c_i} > \delta > \rho_{a,c_j}$.

$\mathbb{P}(FV_{c_i}(a) < T \times \delta) =$
$\mathbb{P}((FV_{c_i}(a) - T \times \rho_{a,c_i})/(\sqrt{T \times \rho_{a,c_i} \times (1 - \rho_{a,c_i})}))$
$< (T \times \delta - T \times \rho_{a,c_i})/(\sqrt{T \times \rho_{a,c_i} \times (1 - \rho_{a,c_i})}))$ (central limit theorem)
$= \Phi((T \times \delta - T \times \rho_{a,c_i})/(\sqrt{T \times \rho_{a,c_i} \times (1 - \rho_{a,c_i})}))$
$= \Phi((\sqrt{T} \times (\delta - \rho_{a,c_i}))/(\sqrt{\rho_{a,c_i} \times (1 - \rho_{a,c_i})})) \underset{T\to\infty}{\to} 0.$
$\Rightarrow FV_{c_i}(a) \geq T \times \delta$ *a.s.*

Likewise $\mathbb{P}(Nb_{a,c_j,n} < T \times \delta) \underset{T\to\infty}{\to} 1. \Rightarrow FV_{c_j}(a) < T \times \delta$ *a.s.*

Thus $\underset{T\to\infty}{Lim}\ FV_{c_i}(a) > \underset{T\to\infty}{Lim}\ FV_{c_j}(a)$ *a.s.* ■

Theorem 2 $\Lambda(a)$ *converges a.s to* $\max_{c\in C} \rho_{a,c}$.

Proof : By definition of the ModeDBMC, $\Lambda(a) = \max_{c\in C}(FV_c(a))$ then given proposition 2 $\underset{T\to\infty}{Lim}\ \Lambda(a) \underset{a.s.}{=} \max_{c\in C} \rho_{a,c}$. ■

Convergence with the AverageDBMC

We are about to prove that $\delta(a)$ almost surely converges to $\mathbb{E}(\varphi_k(a))$ as $T \to \infty$. This proof directly derives from the law of large numbers.

Theorem 3 *$\delta(a)$ converges a.s to $\mathbb{E}(\varphi_k(a))$ (k is not important here given that the $\varphi_k(a)$ are i.i.d).*

Proof : By definition of the , $\delta(a) = \frac{1}{T}\sum_{k=1}^{T} \varphi_k(a)$. Thus given that $\varphi_k(a)$ are i.i.d we can apply the law of large numbers and say that $\delta(a)$ converges a.s. to $\mathbb{E}(\varphi_k(a))$. ■

4.3 Experimental validations and comparison to other sorting algorithms

It is always useful to make practical tests on a preference elicitation algorithm to illustrate how it reacts in practice and evaluate its performances in addition to theoretical proven properties. This is especially true while speaking of an algorithm that may be seen by the user as a black box, which increases the need for a justification. Here, the tests that we present aim at answering two questions. At first the Dominance Based Monte Carlo algorithm being a stochastic algorithm, we would like to know to which extent randomness impacts its result. Then, we made tests to evaluate the ability of this algorithm to restore a part of a learning set while looking at the rest of it.

4.3.1 Stability

We proved in 4.2.5 that the result of the Dominance Based Monte Carlo algorithm converges almost surely when the number of trials grows to infinity. But saying that does not necessarily mean that this convergence is observed in practice (the result could converge almost surely but start converging when the number of trials is higher than 10^{100}). Thus, we made tests to assess the practical stability of the algorithm. To test the stability of the DBMC algorithm on a sorting context $S =< N, V, A, C, L >$ with T trials, Ω stability rounds and a learning set of size τ, we proceed as follows. We iteratively do Ω times the following stability round in which two complete assignments provided by the DBMC algorithm in similar contexts are compared. To do so, a random

complete assignment f is chosen that respects monotonicity. We simply perform a random completion of the sorting context S (as described in Algorithm 6) with an empty leaning set. Then, we randomly (uniformly) choose a set of objects $A' \subset A$ such that $|A'| = \tau$ which is considered as the learning set. With this learning set, the Dominance Based Monte Carlo algorithm is ran twice and we obtain two complete assignments on A, f_1 and f_2. We count the number of objects in A that are not assigned to the same category with f_1 and f_2. Then, we perform the next stability round i.e., the same experience with a new random complete assignment f. At the end of the algorithm, we look at the average percentage of objects sorted differently in f_1 and f_2 across the stability rounds. This number will be called the stability score (a low stability score means that the algorithm is stable). The formal description of this test is given in Algorithm 8.

Algorithm 8: Stability test

Data: Sorting context $S =< N, V, A, C, L = \emptyset >$, number of trial T, number of stability rounds Ω

Result: Average number of difference between two complete assignments provided by the DBMC algorithm

```
counter ← 0
for i from 1 to Ω do
    Create a random complete assignment f of the sorting context S (algorithm 6).
    Select randomly A' ⊂ A (with a uniform distribution) such that |A'| = τ.
    L ←< A', f >
    f1 ← DBMC(S, T)
    f2 ← DBMC(S, T)
    for a ∈ A do
        if f1(a) ≠ f2(a) then
            counter ← counter + 1
return counter / Ω
```

In order to illustrate the evolution of the stability with the number of trials, we applied the stability test on a model of 3 criteria, both of them expressed on a scale of 10 value levels with 50 stability rounds, with several fixed values for the number of categories (2, 5 and 7) and the size of the learning set (0 and 50). This test was made several times with different numbers of trials so that we can plot it with the stability and observe the correlation. The results of these tests are shown in figures 4.7 and 4.8.

The reader may observe that the convergence is effective in practice when the number of trials increases. Indeed, the convergence starts being rather good with 100 trials.

Therefore, in the following we will use 100 trials while applying the DBMC algorithm. Furthermore, we can see that the more categories there are the less stable the DBMC algorithm is. Moreover, we can also see that the larger the training set is the more stable the DBMC algorithm is. To conclude, we think that the disturbing property of this algorithm to be non deterministic has a low impact in practice if we use a large enough number of trials.

nb Trials	10	100	200	500	1000
2 categories	12.9%	4.5%	3.3%	1.3%	0.5%
5 categories	37.5%	14%	9.2%	5.1%	4.2%
7 categories	49.9%	20.3%	13.5%	7.9%	5.5%

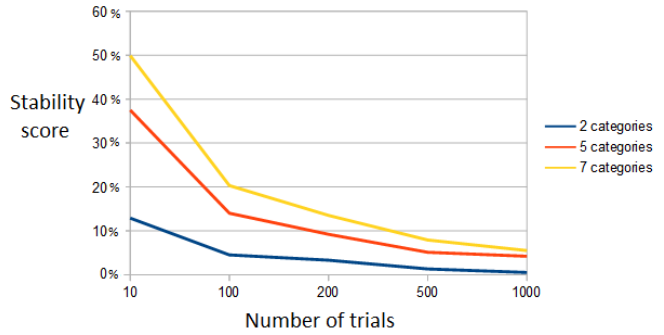

Figure 4.7: Result of the stability test (stability score) with 50 stability rounds in a context with 3 criteria (10x10x10) and 0 elements in the learning set. The number of trials varies from 10 to 1000.

nb Trials	10	100	200	500	1000
2 categories	5%	1.6%	1.3%	0.7%	0.5%
5 categories	16%	5.2%	3.7%	2.3%	1.5%
7 categories	21.3%	7.2%	5.1%	3.3%	2.2%

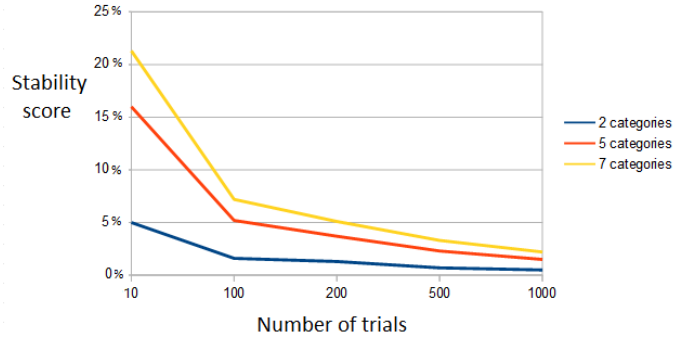

Figure 4.8: Result of the stability test (stability score) with 50 stability rounds in a context with 3 criteria (10x10x10) and 50 elements in the learning set. The number of trials varies from 10 to 1000.

size of the learning set	0	5	10	20	50	100
2 categories	4.5%	3.2%	2.5%	2.3%	1.4%	1.2%
5 categories	14%	9.9%	8.6%	7.1%	5%	3.9%
7 categories	19.4%	14.9%	12.4%	10.2%	7.4%	5.4%

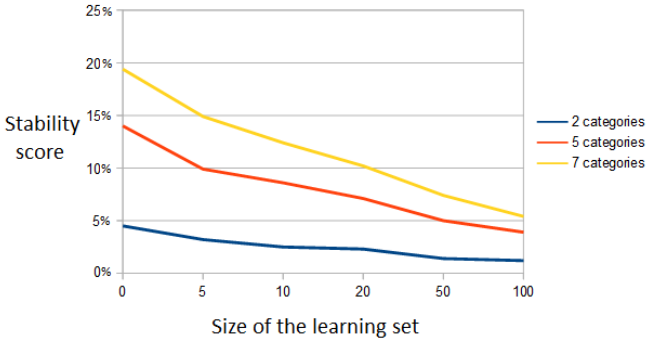

Figure 4.9: Result of the stability test (stability score) with 50 stability rounds in a context with 3 criteria (10x10x10) and 100 trial. The size of the learning set varies from 0 to 100.

4.3.2 Presentation of the k-fold cross validation

We wanted to assess how pertinent are the assignments made by the DBMC algorithm according to the decision makers preferences. In order to make this evaluation we practised a k-fold validation.

The k-fold cross validation [Ron, 1995] is a model validation technique for classification and sorting methods mainly used in machine learning. It aims at assessing how the results of a statistical analysis will generalize to an independent data set. It is generally used to evaluate the accuracy of a prediction technique or of a preference elicitation method. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset). One round of k-fold cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). In k-fold cross-validation, the original sample is randomly uniformly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k-1$ subsamples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k subsamples used exactly once as the validation data. The k results from the folds are then generally averaged to produce a single estimation. The advantage

of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used, but in general k remains an unfixed parameter. To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds. The formal description of this test is given in Algorithm 9. A more complete information consists in considering not only the proportion of objects that are misclassified but also the proportion of objects that are sorted with a distance of 1 category between the training set and the prediction, a distance of 2 categories etc. Although the Dominance Based Monte Carlo algorithm is not a statistical analysis method, as a preference elicitation algorithm aims at predicting the judgement made by a decision maker the k-fold validation may be appropriate as a test to evaluate the accuracy of this prediction. As well it will highlight the problems in which it performs well and those in which it does not. In our test we will mainly apply the k-fold validation with $k = 10$ and $k = 2$ which are the most commonly use [McLachlan et al., 2005].

Algorithm 9: k-fold validation

Data: Sorting context $S =< N, V, A, C, L >$, number of trial T, number of rounds Ω

Result: Average prediction error

```
counter ← 0
for i from 1 to Ω do
    Partition L =< Θ, f_l > in k parts {L_1, ..., L_k}.
    for j from 1 to k do
        L^{-j} ← L/L_j
        f' ← DBMC(< N, V, A, C, L^{-j} >, T)
        for a ∈ Θ_j do
            if f'(a) ≠ f_l(a) then
                counter ← counter + 1
return counter / (Ω × |A|)
```

The k-fold validation as it is defined earlier just counts the number of misclassification and does not take into account the how far from the original assignment the prediction that was made by the tested preference elicitation algorithm is. In order to take this distance between the learning set assignment and the prevision we can use the average $L1$ loss measure. Using the $L1$, during the k-fold validation, at each round, instead of counting the number of misclassification, the test will sum up, for each element a of the test dataset, the distance between the assignment of a in the learning set and the

prediction that was made by the preference elicitation algorithm. Formally the k-fold validation with the $L1$ measure is described in Algorithm 10.

Algorithm 10: k-fold validation with L1 measure

Data: Sorting context $S =< N, V, A, C, L >$, number of trial T, number of rounds Ω

Result: Average prediction error

```
counter ← 0
for i from 1 to Ω do
    Partition L =< Θ, f_l > in k parts {L_1, ..., L_k}, with L_γ =< Θ_γ >.
    for j from 1 to k do
        L^{-j} ← L/L_j
        f' ← DBMC(< N, V, A, C, L^{-j} >, T)
        for a ∈ Θ_j do
            counter ← counter + |f_l(a) − f'(a)|
return counter / (Ω × |A|)
```

4.3.3 Comparison of the DBMC algorithm with other elicitation algorithms through a k-fold validation

The other elicitation algorithms

The result of a k-fold validation presented apart may be difficult to interpret. Indeed, we do not know what is a good performance for this test given that it may depend on various factors (the number of criteria, the size of the learning set, the number of categories, etc). Thus, it may be useful to compare the performances of the DBMC algorithm with those of other preference elicitation algorithm for multi-criteria sorting problem on the same learning sets. In order to do so we used the k-fold validation with several other algorithms for ordinal preference elicitation: UTADIS, logistic regression, choquistic regression, DRSA and the evolutionary algorithm for MR-Sort and a heuristic for 2-additive NCS. Here are some reasons why we chose to compare the DBMC with these three algorithms:

- These algorithms are framed by several MCAP's (utility, set of rules, outranking...) which cover the main families of MCAP. Thus, these methods may be seen as rather representative of the sorting methods.
- They run in a reasonable computable time.

- We chose to use the evolutionary algorithm for eliciting MR-Sort parameters because the algorithm is the most adapted to data set with a relatively large number of assignment examples. Indeed, a version with a mixed integer programming (MIP) is also available [Leroy et al., 2011] that contains $m \times (n+1)$ binary variables. In [Mousseau et al., 2001] it is mentioned that similar problems with 400 binary variables can be solved within 90 minutes. While performing a 10 validation with 50 rounds the algorithm is to be ran 500 times. Although we perform the k-fold validation on relatively small datasets (172 assignments, 5 criteria for the "paper evaluation" dataset), we cannot use the MIP variant due to its computational complexity. By contrast, as mentioned in [Sobrie et al., 2013], the evolutionary algorithm can perform on a model composed of 10 criteria with 1000 assignment within two minutes.
- Some of these methods were already tested with the use of a 2-fold validation in the articles [Sobrie et al., 2015] and [Fallah Tehrani and Huellermeier, 2013].

The learning sets

The k-fold validation aims at illustrating the ability of a learning algorithm to learn from a training set and reproduce and predict a hidden part. In the context of a sorting problem the training set will consist in a set of assignments possibly provided by a decision maker (the learning set). In order to test the performance of the Dominance Based Monte Carlo algorithm we applied the k-fold validation on several sets of preferences or assignments.

1) **The learning set provided by the MNHN**: This learning set is obtained by aggregating several expert's preferences on how severe they think that a scenario of accidental pollution would be according to its destructive potential, the vulnerability of the impacted target and the importance of the biodiversity in this target. The three criteria are evaluated on a discrete scale of five value level ($C1$ to $C5$). The scenario are to be sorted in a set of five categories. This learning set was described in more detail in subsection 3.5.2.

2) We tested the Dominance Based Monte Carlo algorithm through a k-fold validation made on three data sets provided by Ali Fallah Tehrani, Weiwei Cheng, Eyke Hüllermeier. These data sets are particularly adapted to our context due the fact that they cover assignments from sets of criteria expressed on finite discrete scales which is a necessary condition to apply the DBMC algorithm. We chose to use the data set named "breast cancer", "car evaluation" and "paper evaluation". We presented in Table 4.1 several properties of these learning sets.

The "Lecture evaluation" data set (LEV) [1] comes from the Weka data base. It contains examples of anonymous lecturer evaluations, collected at the end of MBA courses. The students were asked to score their lecturers based on four criteria such as oral skills and contribution to their professional/general knowledge. The output is a global evaluation of each lecturer's performance, expressed on an ordinal scale from 0 to 4 (an assignment).

The "Car Evaluation" (CEV)[2] data set comes from the UCI database. It represents the evaluation of a car based on 6 attributes describing a car, namely, buying price, price of the maintenance, number of doors, capacity in terms of persons to carry, the size of luggage boot, estimated safety of the car. The output is the global assignment of the car in 4 categories unacceptable, acceptable, good, very good.

The "Breast cancer" (BCC)[3] data set comes from the UCI database provided by the Oncology Institute of Ljubljana. The instances are described by 7 attributes and are classified in two categories.

In order to compare the results obtained with the DBMC algorithm to results obtained with other methods we created similar learning sets with only two categories as it is made in [Sobrie et al., 2015] i.e., for car evaluation binarized this evaluation into unacceptable versus not unacceptable (acceptable, good or very good) and for lecturers evaluation we binarized the output value by distinguishing between good (score 3 to 4) and bad evaluation (score 0 to 2). The choice of these learning sets was due to the relatively low number of combinations of criteria (less than 5 000) which make the DBMC algorithm run in an acceptable time (some few seconds).

Dataset	Size LS	Nb Crit	Nb Cat	Nb Comb	Monot Viol
Car evaluation (CEV)	1728	6	4	1728	$< 0.1\%$
Breast cancer (BCC)	278	7	2	4536	7.5%
lectures evaluation (LEV)	1000	4	5	625	5.2%

Table 4.1: Several properties of the datasets used for the evaluation and the comparison of the elicitation algorithms. "Size LS" represents the number of assignments provided in the learning set. "Nb Crit" is the number of criteria, "Nb Cat" is the number of categories, "Nb comb" is the number of combinations on the criteria. The column "Monot Viol" gives an idea of how violated is monotonicity in the learning set. It represents the percentage of the pairs of objects a, b with aDb that violate the monotonicity.

[1] Available at https://github.com/oso/pymcda/tree/master/datasets/lev.csv
[2] Available at https://github.com/oso/pymcda/tree/master/datasets/cev.csv
[3] Available at https://github.com/oso/pymcda/tree/master/datasets/bcc.csv

The sources of the results

The k-fold cross validation is provided by jMaf (version of November 2016) software for DRSA [4]. We programmed it for UTADIS [5] based on the code of the UTADIS program that was provided by Patrick Meyer, Sébastien Bigaret, Richard Hodgett and Alexandru-Liviu Olteanu on their MCDA package for the GNU R statistical software on github [6]. The results of the k-fold validation with MR-SORT on the data set of the MNHN was also programmed with the code of MR-Sort from the MCDA package. The results of the k-fold validation with MR-Sort and 2-additive NCS on the "breast cancer", "car evaluation" and "paper evaluation" data sets were provided in [Sobrie et al., 2015]. We decided to use these results rather than running them by our own way for computational complexity reasons. Similarly, the results of the k-fold validation with the logistic regression method and the choquistic regression method were provided in [Fallah Tehrani and Huellermeier, 2013]. For the Dominance Based Monte Carlo algorithm, the tests were programmed in Java [7].

Approximations to deal with violations of monotonicity

The DBMC algorithm requires a monotonic learning set while violations of monotonicity are quite common in ordinal learning sets and there are violations of monotonicity in the learning set that we used as shown in Table 4.1. This property may be seen as a weak point of the DBMC algorithm compared to other methods that allow the use of non-monotonic learning sets (such as DRSA or the MIP for MR-Sort for instance). However, given that the assignments obtained by all these methods is expected to be monotonic, it is impossible for a preference elicitation algorithm to obtain an assignment which is fully compatible with the learning set, as the DBMC algorithm does, and simultaneously to accept non-monotonic learning sets. In order to deal with learning set involving some violations of monotonicity, two possibilities are studied that we will call the *delete approximation* of the sorting context and the *relaxed approximation* of the sorting context. The *delete approximation* of a sorting context consists in creating a similar sorting context in which every pair of objects a and b such that a dominates b and b is sorted in a better category than a is deleted from the learning set. It is important to notice that while creating the delete approximation the inconsistent pairs of

[4]The .isf files to run the tests can be found at https://drive.google.com/file/d/0B5VxOh1ccEY5QXpOaTNWcDZCYTA/view?usp=sharing

[5]The files used to run the tests as well as the test's code for UTADIS and MRSORT can be found at https://drive.google.com/file/d/0B5VxOh1ccEY5aFgyWHZGY2VJa28/view?usp=sharing

[6]Find this package at: https://github.com/paterijk/MCDA

[7]The files used to run the tests as well as the test's code for the DBMC algorithm can be found at https://drive.google.com/file/d/0B5VxOh1ccEY5Sl90bDhZZWVRc2M/view?usp=sharing

objects are marked first and only after every pair is marked the object are erased from the learning let. Therefore, the result of the delete approximation does not depend on the order in which the pairs of objects are studied. The *relaxed approximation* of a sorting context consists in creating a sorting context in which every pair of objects a and b such that a dominates b and b is sorted in a better category than a is relaxed. Here, we mean by relaxed that the assignment of the objects a and b are intervals (as defined in Subsection 4.2.1) and that $\gamma_{max}(a) \leftarrow \gamma_{max}(b)$ and $\gamma_{min}(b) \leftarrow \gamma_{min}(a)$. For instance, if an object a that was assigned to category 2 dominates an object b that was assigned to category 4 they will both be assigned to the interval $[2, 4]$ i.e., $\gamma_{min}(a) = \gamma_{min}(b) = 2$ and $\gamma_{max}(a) = \gamma_{max}(b) = 4$. The reader may notice that the interval assignment that is obtained with this relaxed approximation verifies the interval monotonicity defined in Subsection 4.2.2 i.e., $aDb \Rightarrow \gamma_{min}(a) \geq \gamma_{min}(b)$ and $\gamma_{max}(a) \geq \gamma_{max}(b)$. Hence, the algorithm can then be ran with no additional difficulty. We are aware that other methods exist to re-assign objects from the learning set (in particular isotonic regression [Kotlowski and Slowinski, 2013]), however, we chose to use the *relaxed approximation* and the *delete approximation* because it avoids properly assigning objects to categories without the consent of the decision maker. In case where an object would be found twice (or more times) in the learning set with different assignments, given that every object weakly dominates itself, then the *relaxed approximation* would assign it between the lowest and the highest category in which it is assigned in the learning set. The *delete approximation* would delete both these assignments.

During the k-fold validation that we made with the Dominance Based Monte Carlo algorithm, the training data set was modified so that the algorithm could be ran on it. However, the testing data set remained unchanged and thus may contain violation of monotonicity. Hence, while comparing the Dominance Based Monte Carlo algorithm to other preference elicitation algorithms, the learning set that was used was similar for all the algorithms.

Results of the k-fold validation tests

The results of the 2-fold validation, practised with several preference elicitation algorithms on the previously described data sets is show on Table 4.2. By 2-fold validation we mean that 50% of the learning set is used as a training dataset while the other 50% is used as a test dataset. In each cell the number at the left represent the average percentage of misclassification across the rounds while the number at the right of the cell represents the standard deviation of the percentage of misclassification. As mentioned earlier, here the category set was binarized. The tests were made with 50 rounds while speaking of the Dominance Based Monte Carlo algorithm we talk here about the

median DBMC and we run it with 100 trials. Indeed, as stated in Subsection 4.3.1 the result of the DBMC is rather stable with 100 trials. D DBMC (resp. R DBMC) here represent the DBMC where the approximation that was used to make the learning data set monotonic is the Delete approximation (resp. Relaxed approximation). The results provided for the MR-Sort heuristic and the Non compensatory sorting heuristic come from the article [Sobrie et al., 2015], the Jmaf program for DRSA algorithm contains a k-fold validation and the other results were obtained by tests that were programmed by the author of this document.

	DRSA	NCS	MR-Sort	UTADIS	R DBMC	D DBMC
CEV	$4.91 \pm 0.41\%$	$12.6 \pm 2.63\%$	$13.9 \pm 1.19\%$	$6.9 \pm 0.71\%$	$3.72 \pm 0.28\%$	$3.59 \pm 0.28\%$
LEV	$18.76 \pm 0.35\%$	$14.92 \pm 1.88\%$	$15.92 \pm 1.22\%$	$15.01 \pm 1.31\%$	$18.67 \pm 1.12\%$	$18.66 \pm 1.10\%$
BCC	$25.95 \pm 1.33\%$	$26.72 \pm 3.45\%$	$27.5 \pm 3.79\%$	$28.70 \pm 1.11\%$	$25.92 \pm 0.63\%$	$25.96 \pm 0.62\%$

Table 4.2: Results of the 2-fold validation tests. Here the categories were binarized as described in this document. Percentage of misclassification with its standard deviation. The k-fold validation test with DRSA is the one proposed in the jMAF software, the k-fold validation test with UTADIS was coded in R from the CRAN R repository MCDA (`https://github.com/cran/MCDA`), the k-fold validation for the DBMC algorithm was programmed in java ("R DBMC", meaning Relaxed approximation and "D DBMC", meaning Delete approximation). The k-fold validation test with the 2-additive heuristic for NCS algorithm and with the heuristic for MR-Sort were provided in [Sobrie et al., 2015] (Table 3)

Looking at the results on Table 4.2, we can observe that all the preference elicitation algorithms perform better on the car evaluation dataset (CEV) than on the lecture evaluation data set (LEV) and have their worst performance on the breast cancer dataset (BCC). Several explanations can be proposed. At first, we can observe on Table 4.1 that this is consistent with the proportion of violations of monotonicity. Indeed, all these methods providing a monotonic output, they are unable to restore any non monotonic assignment in the learning set. We can also observe that the breast cancer has the smallest learning set with the highest number of combinations of criteria which makes the elicitation algorithms learn with less examples on a model where more possible complete assignments are possible.

We can observe that the performances of the DBMC algorithm is relatively good on the car evaluation database (CEV). This algorithm being based on monotonicity it may appear relevant to think that it performs well with model with little violations of monotonicity.

The results on the breast cancer database (BCC) are really similar with all the algo-

rithms, DRSA and the DBMC being a bit better than the other three algorithms. The results of the DRSA algorithm and of the DBMC algorithm on the lecture evaluation database (LEV) are less good than the results with the three others.

We also wanted to assess the performance of the DBMC algorithm of data set with more than 2 categories. Thus, we ran the k-fold validation test with the car evaluation dataset and the lecture evaluation dataset without binarizing the category set. The result presented in Table 4.3 does not include the breast cancer dataset given that it is already binarized in its initial state.

	DRSA	UTADIS	R DBMC	D DBMC
CEV (5 cat)	$22.11 \pm 0.54\%$	$9.88 \pm 0.43\%$	$6.61 \pm 0.41\%$	$6.64 \pm 0.37\%$
LEV (7 cat)	$54.21 \pm 0.78\%$	$41.23 \pm 1.97\%$	$60.07 \pm 2.37\%$	$59.79 \pm 2.27\%$

Table 4.3: Results of the 2-fold validation tests (percentage of misclassification with its standard deviation). Here the categories were not binarized. The 2-fold validation of DRSA is the one proposed in the jMAF software, the 2-fold validation of UTADIS was coded in R from the CRAN R repository MCDA (https://github.com/cran/MCDA), the 2-fold validation for the DBMC algorithm was programmed in java ("R DBMC", meaning Relaxed approximation and "D DBMC", meaning Delete approximation).

We can see in Table 4.3 that the binarization of the categories that was made in Table 4.2 had a real impact on the performances of the elicitation algorithms. Indeed, all the results of k-fold validation are clearly higher than those observed on Table 4.2. In particular, the rate of misclassification with DRSA on the car evaluation increases dramatically. Given that the categories are being binarized with the category 1 on the one hand and the categories 2, 3 and 4 on the other hand an explanation may be that many misclassification happen between these three last categories and thus were not counted in Table 4.2.

We were also interested in knowing, when objects are assigned to the wrong category, at "How far are we from the good category?". In order to answer this question we presented on Table 4.4 the result of the 2-fold validation with the $L1$ measure for the Dominance Based Monte Carlo algorithm and compared with the results of the ordinal logistic regression (OLR) and ordinal choquistic regression (OCR), UTADIS and DRSA. The $L1$ measure in the k-fold validation is a measure of misclassification in which each misclassification is weighted by the distance between the assignment in the learning set and the assignment found by the preference elicitation algorithm. The reader may observe that when there are only two categories, the $L1$ 2-fold validation

returns the same result as the normal 2-fold validation (also called 0-1 measure for the 2-fold validation).

	OLR	OCR	DRSA	UTADIS	R DBMC	D DBMC
CEV (5 cat)	23.10 ± 0.75	10.97 ± 3.6	26.32 ± 0.92	10.32	7.2 ± 1.19	7.2 ± 1.19
LEV (7 cat)	42.64 ± 1.48	41.84 ± 1.87	74.69 ± 3.36	45.2	72.27 ± 4.01	73.84 ± 3.52

Table 4.4: Results of the 2-fold validation tests with $L1$ measure. Here the categories were not binarized. Average L1-loss measure with its standard deviation. The 2-fold validation with the ordinal logistic regression and ordinal choquistic regression were provided in [Fallah Tehrani and Huellermeier, 2013] (Table 2), the 2-fold validation for the DBMC algorithm was programmed in java ("R DBMC", meaning Relaxed approximation and "D DBMC", meaning Delete approximation).

While looking at Table 4.4 we see that, once again, the performance of the DBMC algorithm and the DRSA algorithm on the lecture evaluation database (LEV) are very close. By the way we also observe that on the lecture evaluation database their L1 measure are dramatically higher that the 0-1 measure observed in Table 4.3 which means that the objects that are assigned by these elicitation algorithms in the wrong category are often assigned to a category which is not adjacent to the category in which they were assigned in the learning set. By comparison we can see that le L1 measure of the UTADIS method is only 4 point higher than the 0-1 measure (percentage of misclassifications) which means that the objects that assigned to the wrong category with UTADIS are mainly assigned in an adjacent category. On this database, the DBMC algorithm and the DRSA algorithm show bad performances compared to the other three elicitation algorithm. Concerning the car evaluation database (CEV) we can remark that the L1 measures of UTADIS and the DBMC algorithm have performance that are relatively close to the one of choquistic regression, the DBMC algorithm being a little better. On this database with the DBMC, UTADIS and DRSA the L1 measure is relatively close to the 0-1 measure (percentage of misclassifications) given in Table 4.3 which means that the object are mainly assigned to the good category or to an adjacent one.

We showed in Table 4.5 a more precise description of the results of the 2-fold validation for the Dominance Based Monte Carlo algorithm in which we do not only see what percentage of object are not returned in the good category by what percentage is returned in an adjacent category (diff=1) what percentage is returned in a category distant by two levels (diff=2) etc.

	diff=0	diff=1	diff=2	diff=3	diff=4
CEV	94.40%	5.98%	0.55%	$< 0.01\%$	
LEV	39.92%	47.71%	11.02%	1.25%	0.09%

Table 4.5: Distance between the learning set assignment and the prediction of the Dominance Based Monte Carlo algorithm with relaxed approximation.

We recall in Table 4.6 the results of the k-fold validation obtained while comparing the elicitation algorithms for the local biodiversity severity indices based on the preferences of the experts of the MNHN. This sub-problem was a sorting problem with 3 criteria, both expressed on scales of 5 value levels, hence 125 combinations and 5 categories. What we can see here is that the DBMC performs quite well on this learning set. One possible conclusion is that it is quite adapted to little learning sets.

	DRSA	MRSORT	UTADIS	DBMC
k=2	65%	49.5%	67%	39.5%
k=10	60%	42%	62%	$37,8\%$
k=20	40.9%	45%	45%	$39,5\%$

Table 4.6: Result of the k-fold validation (percentage of misclassification) applied with 4 algorithms on the synthetic data set obtained from the MNHN experts on the evaluation of the scenarios in Subsection 3.5.3. Here the percentage of misclassification are written in each cell.

Looking at the results of the k-fold validation, several conclusion can be made. At first we can see that the results obtained while using the delete approximation and the relaxed approximation are very similar. Then, the DBMC algorithm is quite efficient on the car evaluation (CEV) and breast cancer (BCC) data sets while it is less efficient on the lecture evaluation (LEV) data set. The reader may observe that the lecture evaluation contains more assignments in the learning set than the number of combinations on the criteria. While applying a 2-fold validation with the DBMC algorithm, the learning set is cut in two and the part used as a training data set (with which the DBMC will learn) contains almost the same number of assignments than the number of combinations. Therefore, the DBMC algorithm has a very little limited room to manoeuvre. One possibility to explain the relatively bad performance of the DBMC algorithm on the Lecture evaluation data set (LEV) may be that this algorithm does not perform well when the number of assignment in the learning set is too high compared to the number of combinations possibly due to its incapacity to reassign object to different categories when the learning set is violated. Finally, the performances of

the DBMC algorithm often seem to be relatively close to those of the DRSA algorithm. This similarity can be due to the fact that these methods are both model free and based on monotonicity.

4.4 What are the problems to which this algorithm can be adapted?

As described in Subsection 4.2.6 and in Section 4.3, this algorithm is characterized by several properties, theoretical and practical that make it well adapted to some multi-criteria sorting problems and less or not adapted to others.

At first, due to it functioning this algorithm requires as an input the use of discrete finite scales on the criteria. In many decision problems, some criteria that are considered may be expressed on continuous scales such as a number of square meters, or a price. A possibility may be to discretize the continuous values so as to make them adapted to the DBMC algorithm. But this discretization might create a loss of information and could be subject to controversy if not operated carefully. Furthermore, the computational complexity of this algorithm being proportional to the square of the number of theoretically possible objects as defined in Subsection 2.1.5, if the number of criteria or the number of value levels in scales is too high this complexity might make the algorithm not computable (which should be taken into account if some continuous scales are to be discretized).

Then, this algorithm requires the learning set given as an input to respect the monotonicity. In practice, while interviewing decision makers, violations of monotonicity are frequent as observed on the data sets presented in Subsection 4.3.3. Nevertheless, it is possible to approximate the learning set given as an input to make it in accordance with monotonicity as explained in Subsection 4.3.3. However, if the number of violations of the monotonicity is too high, the approximation might be very different from the learning set given as an input and not represent the decision maker's preferences.

Thus, this algorithm seems particularly adapted to multi-criteria classification problems with criteria expressed on discrete scales with a limited number of theoretically possible alternatives (combinations of values on the criteria). As an example, we ran our algorithm with 100 000 theoretically possible objects and 100 trials (see in Section 4.2) within approximately 2 hours and a half. Hence, from this dimensional point of view it may be suitable to a relatively large panel of multi-criteria sorting problems such as for example problems with 7 criteria expressed on scales of 5 value levels. Then,

due to its model free particularity, it may be adapted to problems where the human reasoning is not based on an explicitly formulated MCAP. Finally, this algorithm is proved to return the learning set given by the decision maker as long as this learning set respects monotonicity. The methods that are based on a MCAP could not return a result that do not suit to this MCAP. For instance most of them (UTADIS, MR-Sort...) assume that there is independence between the criteria (as defined in Subsection 2.2.6). Our algorithm does not need such assumption.
Finally, we should mention the legitimacy of the method. The fact is that the DBMC algorithm works as a black box. A learning set is given as an input and a complete assignment is found as an output without an intuitive process that can be followed step by step by the decision maker. From this point of view, this method may not be adapted to some context in which more justification is needed. In our case, the legitimacy may come from practical tests such as the k-fold validation (presented in Subsection 4.3.2) that may demonstrate that it is a good representation of the decision maker's mind.

4.5 Perspectives

4.5.1 Modified idiosyncrasy: An other experimental test for the efficiency of the algorithm

In its original version proposed in [Sobrie et al., 2013] this test consists in using a complete assignment obtained with a MCAP that is not the one used in the tested elicitation method, looking at the assignment of a limited proportion of the dataset called the learning dataset of the objects and looking at how accurately the algorithm returns these same assignments (those that it learnt with). If there is no violation of the monotonicity principle (which should not occur with any MCAP) the DBMC algorithm should return 100% of the learning set. However, we modified this tests to make it relevant by learning on one part of the data set (the learning dataset) and measuring the accuracy of the complete assignment obtained by the DBMC on the

rest of the dataset. The formal description of this test is given in Algorithm 11.

Algorithm 11: Idiosyncrasy

Data: Sorting context $S =< N, V, A, C, L = \emptyset >$, number of trial T, number of rounds m,a parameter set σ of an MCAP Ω, proportion α of the objects to be used as the learning set

Result: Average prediction error

$counter \leftarrow 0$
for $a \in A$ **do**
  $f_l(A) \leftarrow$ classification of a by Ω with the parameters σ
$L \leftarrow < A, f_l >$ **for** *i from 1 to m* **do**
  Select $A' \subset A$ randomly such that $|A'| = \alpha|A|$
  $L' \leftarrow < A', f_l >$
  $f' \leftarrow DBMC(< N, V, A, C, L' >, T)$
  for $a \in A/A'$ **do**
    if $f'(A) \neq f_l(A)$ **then**
      $counter \leftarrow counter + 1$
return $\frac{count \times (1-\alpha)}{\Omega \times |A|}$

In practice we already have some results with this experimental tests made on the DBMC algorithm and on two elicitation algorithms UTADIS and the heuristic algorithm for MR-Sort described in subsection 2.3.6. However, the results that we have and the conclusions that we have made yet are not sufficient to be added in this thesis.

4.5.2 Possible applications of the Dominance based Monte Carlo

As mentioned earlier, the Dominance based Monte Carlo algorithm is a multi-criteria preference learning methods that suits quite well to contexts in which both the criteria and the category set are expressed through an ordinal scale with a limited number of value levels. One idea of possible application is related to the Oxfam program named "Behind the brands". Oxfam is an NGO focused on the alleviation of global poverty. They created a program named "Behind the brands" that consists in rating the biggest 10 food companies according to 7 criteria, three related to the respect of the workers, three of them related to the respect of the environment and one related to the financial transparency. However, these rates are not easily interpretable and it may not be easy to compare these companies. A possible application of the Dominance based Monte

Carlo algorithm to real life problem could be to create a smart-phone application that would ask the user to assign profiles of companies expressed on these 7 criteria to 5 categories representing how globally responsible the company is according to the user. Then, the algorithm would be ran and while shopping the user may make her choice considering the price and the responsibility category. When the scores of the companies are updated by Oxfam then the global score is updated too.

Conclusion

This thesis is built around two main axis: the construction of the *Biodiversity Severity Index* and the creation of the Dominance Based Monte Carlo algorithm.

The *Biodiversity Severity Index*

In order to deal with the first issue properly, we first had to introduce the scientific fields and methodological tools that we used in this thesis, namely risk management and multi-criteria decision aiding. In chapter 1 we introduced the problem as it was presented to us together with the scientific fields and tools that are generally associated to risk management. The legal context of risk management in France was presented, in particular the "Arrêté du 29 septembre 2005" and the "Circulaire du 10/05/2010", as well as its main public actors in France, namely the DREALs and the INERIS. We saw that all the methods that exist to evaluate the acceptability of accidental risks require an evaluation of the probability of the different scenarios and an evaluation their severity.

We showed how the current risk management methodologies that is provided by the different institutions around the world mainly consider the scenarios based on their expected impact on the human life. However, there is a lack of consideration for the environmental dimension in this field and the impact on biodiversity in particular. According to every sources relevant to environmental evaluation, while evaluating the risk of accidental pollution, taking the risk of losses on potential uses that human could make of the environment is important. Yet, we decided not to explore this dimension of the problem and leave it to economists of the environment that would be more appropriate to investigate it. We came to the conclusion that in order to take the damages on biodiversity into account, a *Biodiversity Severity Index* is needed to evaluate the expected severity of a scenario of accidental pollution. Then, we defined

formally the first topic of this thesis as the creation a methodology that would assess the expected severity of a scenario of an accidental pollution on the biodiversity.

All the risk evaluation methods and all legislations on risks aggregate several criteria to evaluate the expected severity of the scenario and thus use (voluntarily or not) some multi-criteria aggregation procedures. For instance, the method presented while describing the risk evaluation procedure currently in use, in subsection 1.2.2 aims at aggregating the number of potential victims in an area that could be touched by an accident with the expected strength of the effect on these persons (see table 1.6) to obtain the severity of the scenario. It could be seen as a rule based method with two criteria. However, the use of multi-criteria decision aiding as a scientific field to support this task is not commonly used by the different institution in charge of the risk management. Therefore, the choice of the Multi-Criteria Aggregation Procedure and the preference parameters that are associated to it generally done intuitively. The idea here was to make these choices based on a formal reasoning and on multi-criteria methodology. Multi-criteria decision aiding is a discipline at the center of which the interactions between the different actors is placed. In our topic we dealt with experts of different fields and with different background. Therefore, it was important for us to organize formally our interactions with them so as to insure a better communication.

We came to the conclusion that the construction of an indicator would benefit from the use of a hierarchy of criteria. Indeed, we think that the involvement of several scientific fields in this problem makes it adapted to a modeling through a hierarchy of criteria. In this way, our problem could be decomposed in several distinct sub-problems and for each of them we could choose the most adapted method. Multi-criteria decision making provides several formal tools to find a good family of criteria and to organize it as a hierarchy of criteria.

In chapter 2 we introduced the scientific field of multi-criteria decision aiding. We insisted on two topics: the construction of a hierarchy of criteria, the concept of the preference elicitation. On the construction of the hierarchy of criteria, we particularly focused on *value focused thinking* [Keeney, 1992] which explores in depth this topic. Then, we stated that every sub-problem of the hierarchy must contain a Multi-Criteria Aggregation Procedure and we presented several of them.

Finally, we presented the concept of preference and preference elicitation. We introduced several elicitation techniques for different MCAPs and based on different operating modes (mixed integer programming, rough set, evolutionary algorithm) so that the reader can be familiar with these tools used during the construction of the BSI.

Recall of the main contributions on the *Biodiversity Severity Index*

The first task of the creation of the Biodiversity Severity Index consisted in finding the appropriate hierarchy of criteria. We constructed the hierarchy described in section 3.1 based on a formal reasoning partly inspired by *Value Focused Thinking* [Keeney, 1992], interacting with experts so that the choice of this hierarchy instead of other possible ones can be scientifically defended. In particular:

- We managed to use criteria understood in the same way by every stakeholder involved in this process.
- Every maximal sub-family of criteria of this hierarchy (as defined in subsection 2.1.6) is a coherent family of criteria.
- Each sub-problem can be treated independently given that there is no dependence between a criterion and an other criterion located in an other sub-problem.
- Every sub-problem contains a limited number of criteria which makes it easier to manage for the elicited actors.
- Every relation between a criterion and its sub-criteria seems quite natural to all the actors involved in the process.
- There was a limited number of aggregations to be done.

The scales on which all the criteria are expressed were also chosen so as to be accepted and easy to use by both the experts and the potential users. We used standardized scales when such scales exist and we created and semantically defined scales when no such scale exist. When using such scales we proposed reference points so that their meaning can be understood quite uniformly by different users. Three aggregation were to be done in order to obtain the Biodiversity Severity Index:

- The aggregation of the *toxicity* with the *residence time* and the expected *concentration* of product in a given target to obtain the destructive potential of a scenario in a leak. The methodology for this aggregation was built with repeated interactions with an expert in toxicology. We initially opted for disaggregation approach for elicitation but these interactions helped the expert to define a set of rules that he felt is appropriate in this context. In other words, we moved to a aggregation approach for elicitation.

- The aggregation of the *destructive potential* of a leak on a target with the *environmental value* of the target and the *vulnerability* of the target to obtain the *Local Biodiversity Severity Index*. The methodology used for this aggregation was built with several interaction with groups of experts from the Muséum National d'Histoire Naturelle (MNHN). The criteria used in this sub-problem being defined on semantically defined scales the definition of the reference points for each of these criteria was a crucial step that really helped all the stakeholders to have a better understanding of these criteria and to obtain a coherent result.

  We tested several algorithm and compared them with a k-fold validation. Finally we chose a MR-Sort model with veto. We saw that the introduction of a veto dramatically improved the coherence between the preference information given by the expert and the result given by the MR-Sort model.

- The aggregation of several *Local Biodiversity Severity Index* to obtain the *Biodiversity Severity Index*. This part of the problem was treated more superficially but our idea is to propose a simple modified max.

Pros and cons of the Biodiversity Severity Indicator

The main weakness that I identify in this work is the lack of a real methodology for the aggregation of several *Local Biodiversity Severity Index* to obtain the *Biodiversity Severity Index*. A trail is given in section 3.6.1 where we proposed a modified version of the max for which if the most impacted target is impacted with a severity c and the total area of the targets impacted with a severity of c is higher than 10 square kilometres the global severity of the studied scenario is increase to the category above c. However, in order to make it scientifically acceptable, more work should be made about it, to compare it to other possible aggregation procedures, study the consequences of its use in this sub-problem, present and discuss it with several experts. This work would probably lead to modify the threshold of 10 square kilometres and we may even change this aggregation to something totally different.

The main advantage that I think this work offers is to be created in cooperation with several experts with different expertises. The creation of the hierarchy of criteria, the choice of the aggregation methods for each sub-problem and the choice of the preference parameters for these aggregation methods were all obtained according to their expertises which increases the method's legitimacy.

Perspectives on the Biodiversity Severity Index

As a perspective, we think about improving this process by involving more decision makers. Indeed, until now we focused on involving experts instead of decision makers because the lower part of the hierarchy is mainly made of mean-end relations that require expertise. However, for the upper part of the hierarchy we could consider call out public decision makers, industrial managers or citizens to aggregate together the biodiversity severity of several target, which is relative to preferences rather than to expertise.

The Dominance Based Monte Carlo

The second axis of this thesis deals with the creation of a preference elicitation method for the multi-criteria sorting problem. This method combines two properties that are not frequently met by the other methods for preference elicitation: a model free approach and a stochastic approach. It may be seen as a mix between DRSA, ORCLASS and SMAA methods. Indeed among other things, its model free particularity connects it to the two firsts while its stochastic functioning connects it to SMAA methods.

We described its functioning and some theoretical properties, including some related to its stochastic nature. The Dominance Based Monte Carlo algorithm offers a guaranty to return correctly the learning set if it is monotonic, unlike methods based on an MCAP that cannot return a learning set which is not compatible with their MCAP (for instance a learning set which does not respect independence between the criteria cannot be represented by many methods such as UTADIS or MRSORT). Then, we showed that despite its stochastic property (that may be seen as not very reassuring for a multi-criteria sorting method), the convergence that was demonstrated theoretically was observed in practice with a relatively low number of trials.

Finally, we compared the Dominance Based Monte Carlo performances to other reference learning algorithms with a k-fold validation on several data sets. The conclusion that we can draw is that it is that its performance are rather similar to those of DRSA, while compared to UTADIS, MR-Sort or NCS its performances were better on one dataset, rather similar on one dataset and worst on one dataset. The computational complexity of this method being proportional to the number of theoretically possible objects (combinations of values on the criteria) the use of this method is recommended for small and medium size problems. Moreover, the reader may notice that the computational complexity of the Dominance Based Monte Carlo algorithm is not related to the size of the learning set (the number of assignment given by the decision maker) unlike most of the elicitation methods, in particular those based on a Mixed Integer

Programming. Thus, it is adapted to problems with a relatively large learning set. Given that the Dominance Based Monte Carlo algorithm works as a black box, it may not be suitable to problems in which justification is important. However, it may be used in contexts with less strategic issues, where the human judgment or any ordinal assignment is to be reproduced and no MCAP seems adapted.

Perspectives of improvements of the algorithm

As a perspective, we think of improving the Dominance Based Monte Carlo algorithm by adding the possibility to impose independence between all the criteria or a set of criteria. As well, we could imagine imposing that if an object a is better than an object b on a criteria i with a difference higher than a veto v_i then a could not be assigned to a lower category than b, regardless to their values on the other criteria.

Perspective of new tests to do on the Dominance Based Monte Carlo

We propose to create a test named model returning validation in which all the theoretically possible objects of a model (combinations of values on the criteria) are assigned to a category with the use of a given MCAP M and a set of preference parameters. Then a predefined proportion of the objects are randomly chosen and used as a training dataset while the rest is used as a testing dataset as it is done with the k-fold validation. The idea here would be to compare the percentage of error obtained with the Dominance Based Monte Carlo to the percentage of error obtained using other preference elicitation algorithms in particular algorithms based on the M. This test was already programmed and several results are already obtained. However, the analysis that we made of these results was not sufficient to be included in this thesis. This is an avenue for some future research.

Appendices

A Description of the DOMLEM algorithm

This appendix describes the DOMLEM algorithm which is used in DRSA (presented in subsection 2.3.9) and aims at finding a set of rules that matches the with the rough set obtained earlier in this method. For the sake of simplicity, in the following we present the general scheme of the DOMLEM algorithm only for a case of type certain rules from upward unions of decision categories. There after we will call an complex E a conjunction of elementary conditions e. We will denote by $[E]$ the set of all the objects in A matching the complex E.

We say that an object supports a decision rule if it matches both condition and decision parts of the rule. On the other hand, an object is covered by a decision rule if it matches the condition part of the rule. We say that a rule is minimal if there is no other more general rule assigning objects to the same union or sub-union of categories (if this rule could not be simplified without changing the result). A set of rule is said complete if it covers all objects given in the learning set in such so that consistent objects are re-assigned to their original categories and inconsistent objects are assigned to clusters of categories referring to this inconsistency.

A set of rule is said minimal if all its rules are minimal and if the suppression of any rule would make it not complete. In order to find a minimal set of rules returning the information given as an input several algorithms are available in rough set theory.

The DOMLEM algorithm iteratively create a set of rules for each upward unions of decision category as described here:

Algorithm 12: DOMLEM algorithm

Data: Input: L_{upp} - a family of lower approximations of upward unions of decision categories: $\{\underline{P}(c^{\geq}_{nbCategory}), \underline{P}(c^{\geq}_{nbCategory-1}), ..., \underline{P}(c^{\geq}_{2})\}$; Output: $R_{\geq}$ a set of $D_{\geq}$Decision rules

```
1 R_≥ = ∅
2 for Each B ∈ L_upp do
3     E ← findRules(B)
4     for Each rule E ∈ E do
5         if E is a minimal rule then
6             R_≥ ← R_≥ ∪ E
```

Algorithm 13: findRules procedure

```
1  G ← B
2  E ← ∅
3  while G ≠ ∅ do
4      E ← ∅
5      {Creating a new rule E through a greedy procedure}
6      S ← G
7      while (E = ∅) or [E] ⊆ B do
8          best ← ∅
9          for Each criterion i ∈ N do
10             Cond ← {g_i(a) ≥ r_i : ∃a ∈ S g_i(a) = r_i}
11             for Each elem ∈ Cond do
12                 if evaluate({elem} ∪ E) isBetterThan evaluate({best} ∪ E) then
13                     best ← elem
14                 { Finding the "best" condition possible (evaluation explained
                   later)}
15         E ← E ∪ {best}
16         {Adding the obtained "best" condition to the rule} S ← S ∩ [best]
17     for Each elementary condition e ∈ E do
18         if [E − {e}] ⊆ B then
19             E ← E − {e}
20             {Removing useless conditions}
21         E ← E ∪ E
22         G ← B − ∪_{E∈E}[E]
```

Here the choice of the next condition it made through the use of the function evaluate(E). A candidate E for a condition part of a rule is in this version of DOMLEM the complex E with the highest ratio $\frac{|[E] \cap G|}{|[E]|}$ i.e. a rule that covers mainly elements of G.

B Reference points for the criteria of the Biodiversity Severity Index

We presented in subsection 3.2.5 the scales on which the criteria will be evaluated. We mentioned several sets of reference points that are proposed to help the user of this method to give a value of a scenario on the different criteria. We are presenting here

these reference points.

B.1 Reference points of the *Local Biodiversity Severity Indices*

Value level	Value level name	Description of the scenario used as a reference
C2	Low pollution	Due to a fire accident in an industrial plant in Ferme de la Garenne, the water used by the firemen brought to the Seine river a large volume of ammoniac. In the following days, the concentration of ammoniac in the Seine river nearby the industrial plant was 15 times superior to the maximal acceptable concentration. Some dead fish are found and fishing is prohibited on the Seine river on a distance of 50 kilometres for two months
C3	Medium pollution	An accident in an industrial plant causes a leak of biphenyl in the lake of Créteil. Following this accident, the concentration of biphenyl in the lake of Créteil is 25 times superior to the maximal acceptable concentration. The lake of Créteil is a wet area of 40 acres. We observe many dead fishes and some dead birds on and around the lake. Fishing is prohibited on the lake until further notice.

Figure 10: Table defining the reference points of the *Local Biodiversity Severity Indices*

Value level	Value level name	Description of the scenario used as a reference
C4	Serious pollution	An road accident involving a road tank creates a leak of approximately 10 cubic meters of sulfuric acid in the "Natural park of the Cotentin marshes". The leak create in the surrounding surface water a concentration of sulfuric acid 56 times superior to the maximal acceptable concentration. A high mortality of all type of wild life is observed around the leak although some wild life remains alive.
C5	Ecological disaster	This reference is a real event happening in Brazil in 2015. In November 21, 2015, 60 millions of cubic meters of red muds (a very basic liquid) escape from a Aluminum factory. The leak reaches the Rio Doce killing all the wild life in this river located in a protected tropical forest before flowing to the sea where it largely impacts the coral reef.

Figure 11: Table defining the reference points of the *Local Biodiversity Severity Indices*

B.2 Reference points of the *Value of the environment*

Area type used as a reference point for the value of the environment	Type of biodiversity	BIOMOS score	Value level
Component of the interregional arc of the remarkable biodiversity, (in French "Composante de l'arc inter-régional de biodiversité remarquable")	Remarkable	4	$C5$
Forest massif of more that 2000 ha, (in French "Massif forestier de plus de 2000 ha")	Remarkable	2	$C4$
ZNIEFF, natural area of ecological interest for flora and fauna Type 1, (in French "Znieff, Zone naturelle d'intérêt écologique, faunistique et floristique, Type 1")	Remarkable	2	$C4$
ZICO, Important area for the protection of birds, (in French "ZICO, zone importante pour la protection d'oiseaux")	Remarkable	2	$C4$
ZNIEFF, natural area of ecological interest for flora and fauna Type 2, (in French "Znieff, Zone naturelle d'intérêt écologique, faunistique et floristique, Type 2")	Remarkable	1	$C4$
Forest, wetland of an area higher than 1 ha, (in French "Forêt ou zone humide d'une superficie supérieure à 1 ha")	Remarkable	1	$C4$

Figure 12: Table defining the reference points of the *Local Biodiversity Severity Indices*

Area type used as a reference point for the value of the environment	Type of biodiversity	BIOMOS score	Value level
Forest, wood of an area lower than 1 ha, (in French "Forêt ou bois d'une superficie inférieure à 1 ha")	Ordinary	0.8	$C3$
Wetland of an area lower than 1 ha, (in French "Zone humide d'une superficie inférieure à 1 ha")	Ordinary	0.8	$C3$
Rural vacant area, (in French "Espace rural vacant")	Ordinary	0.8	$C3$
Forest clearings, (in French "Clairière en forêt")	Ordinary	0.8	$C3$
Watercourse, (in French "Cours d'eau")	Ordinary	0.8	$C3$
Urban vacant area, (in French "Terrain vacant en mileu urbain")	Ordinary	0.6	$C2$
Urban park or large garden, (in French "Parc ou grand jardin")	Ordinary	0.6	$C2$
Orchard, (in French "Verger")	Ordinary	0.6	$C2$
Rail transport allowances, (in French "Emprise de transports férés")	Ordinary	0.3	$C1$
Individual home garden, (in French "Jardin d'habitat individuel")	Ordinary	0.3	$C1$
Poplar plantations, (in French "Peupleraies")	Ordinary	0.1	$C1$
Cemetery, (in French "Cimetière")	Ordinary	0.1	$C1$
Golf field, (in French "Terrain de golf")	Ordinary	0.1	$C1$

Figure 13: Table defining the reference points of the *Local Biodiversity Severity Indices*

C Elicitation of the preferences of the experts of the MNHN for Local Biodiversity Severity indices

	C1	C2	C3	C4	C5
S1			4	3	
S2				4	3
S3	2	5			
S4		3	4		
S5	7				
S6		3	4		
S7			4	3	
S8				3	4
S9				1	6
S10			2	5	
S11			1	6	
S12			4	3	
S13			2	5	
S14			2	5	
S15		1	4	2	
S16				3	4
S17	5	2			
S18			1	6	
S19				5	2
S20					7

Table 7: Argument strength assignment of the sentence "I think that the scenario S? should be assigned to the category C?". Group 1. Here we can see several violations of monotonicity. Scenario S2 [5, 3, 5] in category 4 is incompatible with scenario S9 [5, 3, 2] in category 5. Scenario S11 [3, 4, 2] in category 4 is incompatible with scenario S8 [3, 4, 1] in category 5. Scenario S19 [3, 4, 4] in category 4 is incompatible with scenario S8 [3, 4, 1] in category 5. Scenario S12 [3, 1, 5] in category 3 is incompatible with scenario S14 [3, 1, 4] in category 4.

	Cat1	Cat2	Cat3	Cat4	Cat5
S1			5	2	
S2					7
S3		4	3		
S4	3	4			
S5	7				
S6	5	2			
S7	1	6			
S8		2	4	1	
S9					7
S10					7
S11		3	4		
S12					
S13		4	3		
S14			5	2	
S15				6	1
S16				6	1
S17	6	1			
S18			1	5	1
S19			1	5	1
S20					7

Table 8: Argument strength assignment of the sentence "I think that the scenario S? should be assigned to the category C?". Group 2.

	Cat1	Cat2	Cat3	Cat4	Cat5
S1			6	1	
S2					7
S3		4	3		
S4		4	3		
S5	4	3			
S6			5	2	
S7				7	
S8				7	
S9				1	6
S10				1	6
S11				4	3
S12				3	4
S13				3	4
S14			1	6	
S15			1	6	
S16					7
S17		5	2		
S18				2	5
S19				1	6
S20					7

Table 9: Argument strength assignment of the sentence "I think that the scenario S? should be assigned to the category C?". Group 3.

	Cat1	Cat2	Cat3	Cat4	Cat5
S1		1	5	1	
S2				5	2
S3		3	4		
S4		4	3		
S5	7				
S6		2	5		
S7			2	5	
S8			5	2	
S9				7	
S10		1	4	2	
S11			4	3	
S12		2	4	1	
S13	3	4			
S14		4	3		
S15	3	4			
S16			1	5	1
S17	6	1			
S18			1	6	
S19				6	1
S20					7

Table 10: Argument strength assignment of the sentence "I think that the scenario S? should be assigned to the category C?". Group 4.

	Cat1	Cat2	Cat3	Cat4	Cat5
S1		1	5	1	
S2				4	3
S3		5	2		
S4	3	4			
S5	7				
S6		5	2		
S7		1	5	1	
S8		4	3		
S9				4	3
S10				4	3
S11			3	4	
S12		3	4		
S13	3	4			
S14		3	4		
S15	1	5	1		
S16				3	4
S17	6	1			
S18				5	2
S19			2	5	
S20					7

Table 11: Argument strength assignment of the sentence "I think that the scenario S? should be assigned to the category C?". Group 5.

	Cat1	Cat2	Cat3	Cat4	Cat5
S1			2	5	
S2				7	
S3			7		
S4		7			
S5			7		
S6			7		
S7				7	
S8			7		
S9				7	
S10		7			
S11			7		
S12			7		
S13			7		
S14		7			
S15	7				
S16				7	
S17	7				
S18				7	
S19				7	
S20					7

Table 12: Argument strength assignment of the sentence "I think that the scenario S? should be assigned to the category C?". Group 6. Here we can observe a violation of monotonicity. Indeed, scenario S10 [5, 1, 4] in category 2 is incompatible with scenario S3 [4, 1, 2] in category 3

	Cat1	Cat2	Cat3	Cat4	Cat5
S1	0	2	27	13	0
S2	0	0	0	20	22
S3	2	21	19	0	0
S4	6	26	10	0	0
S5	32	3	7	0	0
S6	5	12	23	2	0
S7	1	7	11	23	0
S8	0	6	19	13	4
S9	0	0	0	20	22
S10	0	8	6	12	16
S11	0	3	19	17	3
S12	0	5	19	7	4
S13	6	12	12	8	4
S14	0	14	15	13	0
S15	11	10	6	14	1
S16	0	0	1	24	17
S17	30	10	2	0	0
S18	0	0	3	31	8
S19	0	0	3	29	10
S20	0	0	0	0	42

Table 13: Sum of the argument strength assignment for all the groups

	Cat1	Cat2	Cat3	Cat4	Cat5
S1	0	0	5	1	0
S2	0	0	0	4	2
S3	0	4	2	0	0
S4	0	5	1	0	0
S5	5	0	1	0	0
S6	1	1	4	0	0
S7	0	1	2	3	0
S8	0	1	3	1	1
S9	0	0	0	3	3
S10	0	1	1	2	2
S11	0	0	3	3	0
S12	0	0	4	0	1
S13	0	3	1	1	1
S14	0	2	2	2	0
S15	1	2	1	2	0
S16	0	0	0	3	3
S17	5	1	0	0	0
S18	0	0	0	5	1
S19	0	0	0	5	1
S20	0	0	0	0	6

Table 14: Number of groups that gave 4 points of more to the scenario S? and the category C?

Bibliography

[Haz, 2012] (2012). *All Hazards Risk Assessment Methodology Guidelines*. Public Safety Canada.

[UKC, 2015] (2015). *National Risk Register of Civil Emergencies*. CabinetOffice.

[Ackermann and Eden, 2011] Ackermann, F. and Eden, C. (2011). Strategic management of stakeholders: Theory and practice. *Long Range Planning*, 44(3):179–196.

[Aristotle, 50BC] Aristotle (350BC). *The Nicomachean Ethics*.

[Aspe, 1989] Aspe, C. (1989). Histoire de l'écologie. *Revue française de sociologie*, 30(2):345–346.

[Aven, 2012] Aven, T. (2012). The risk concept: historical and recent development trends. *Reliability Engineering & System Safety*, 99:33–44.

[Balakrishnan, 2013] Balakrishnan, N. (2013). *Handbook of the logistic distribution*. CRC Press.

[Bana e Costa et al., 2003] Bana e Costa, C. A., de Corte, J.-M., and Vansnick, J.-C. (2003). Macbeth. Working Paper LSEOR 03.56, London School of Economics and Political Science, London, UK. © 2003 London School of Economics and Political Science.

[Bana e Costa et al., 2016] Bana e Costa, C. A., De Corte, J.-M., and Vansnick, J.-C. (2016). *Multiple Criteria Decision Analysis: State of the Art Surveys*, chapter On the Mathematical Foundations of MACBETH, pages 421–463. Springer New York, New York, NY.

[Bateman et al., 2002] Bateman, I. J., Carson, R. T., Day, B., Hanemann, M., Hanley, N., Hett, T., Jones-Lee, M., Loomes, G., Mourato, S., Özdemiroglu, E., et al. (2002). Economic valuation with stated preference techniques: a manual.

[Baybutt, 2012] Baybutt, P. (2012). Understanding risk tolerance criteria. Technical report, Pimatech inc. Colombus, Ohio, USA.

[Bell et al., 2012] Bell, H., P. Hughey, E., Prizzia, R., and Clark, A. (2012). Risks, hazards, and crisis in public policy. *Wiley Online Library*.

[Belton and Stewart, 2002] Belton, V. and Stewart, T. (2002). *Multiple Criteria Decision Analysis: An Integrated Approach*, page 2. Springer US.

[Bensettiti et al., 2004a] Bensettiti, F., Bioret, F., Roland, J., and Lacoste, J.-P. (2004a). *Cahiers d'habitats Natura 2000. Connaissance et gestion des habitats et des espèces d'intérêt communautaire. Tome 2-Habitats côtiers.* Museum national d'Histoire Naturelle.

[Bensettiti et al., 2005] Bensettiti, F., Boullet, V., Chavaudret-Laborie, C., and Deniaud, J. (2005). *Cahiers d'habitats Natura 2000. Connaissance et gestion des habitats et des espèces d'intérêt communautaire. Tome 4 (vol.1)-Habitats agropastoraux.* Museum national d'Histoire Naturelle.

[Bensettiti et al., 2002] Bensettiti, F., Gaudillat, V., and Haury, J. (2002). *Cahiers d'habitats Natura 2000. Connaissance et gestion des habitats et des espèces d'intérêt communautaire. Tome 3-Habitats humides.* Museum national d'Histoire Naturelle.

[Bensettiti et al., 2004b] Bensettiti, F., Herard-Logereau, K., J., V. E., and C., B. (2004b). *Cahiers d'habitats Natura 2000. Connaissance et gestion des habitats et des espèces d'intérêt communautaire. Tome 5-Habitats rocheux.* Museum national d'Histoire Naturelle.

[Bensettiti et al., 2001] Bensettiti, F., Rameau, J.-C., and Chevallier, H. (2001). *Cahiers d'habitats Natura 2000. Connaissance et gestion des habitats et des espèces d'intérêt communautaire. Tome 1-Habitats forestiers.* Museum national d'Histoire Naturelle.

[Bentham, 1780] Bentham, J. (1780). *An introduction to the principles of morals and legislation*. Courier Corporation.

[Błaszczyński et al., 2009] Błaszczyński, J., Greco, S., Słowiński, R., and Szelg, M. (2009). Monotonic variable consistency rough set approaches. *International journal of approximate reasoning*, 50(7):979–999.

[Bortz et al., 1975] Bortz, A. B., Kalos, M. H., and Lebowitz, J. L. (1975). A new algorithm for monte carlo simulation of ising spin systems. *Journal of Computational Physics*, 17(1):10–18.

[Bouyssou, 1986] Bouyssou, D. (1986). Some remarks on the notion of compensation in mcdm. *European Journal of Operational Research*, 26(1):150–160.

[Bouyssou et al., 2012] Bouyssou, D., Jacquet-Lagrèze, E., Perny, P., Slowiński, R., Vanderpooten, D., and Vincke, P. (2012). *Aiding Decisions with Multiple Criteria: Essays in Honor of Bernard Roy*. International Series in Operations Research & Management Science. Springer US.

[Bouyssou and Marchant, 2007a] Bouyssou, D. and Marchant, T. (2007a). An axiomatic approach to noncompensatory sorting methods in mcdm, i: The case of two categories. *European Journal of Operational Research*, 178(1):217–245.

[Bouyssou and Marchant, 2007b] Bouyssou, D. and Marchant, T. (2007b). An axiomatic approach to noncompensatory sorting methods in mcdm, ii: More than two categories. *European Journal of Operational Research*, 178(1):246–276.

[Brans and Vincke, 1985] Brans, J. P. and Vincke, P. (1985). A preference ranking organisation method (the promethee method for multiple criteria decision-making). *Management Science*, 31(6):647–656.

[Brown et al., 1973] Brown, G., Copeland, T., and Millward, M. (1973). Monadic testing of new products : an old problem and some partial solutions.

[Brownlow and Watson, 1987] Brownlow, S. and Watson, S. (1987). Structuring multi attribute value hierarchies. *Journal of the Operational Research Society*.

[Cailloux, 2012] Cailloux, O. (2012). *Indirect elicitation of multicriteria sorting models*. Theses, Ecole Centrale Paris.

[Carbone and Hey, 2000] Carbone, E. and Hey, J. D. (2000). Which error story is best? *Journal of Risk and Uncertainty*, 20(2):161–176.

[Cazenave and Helmstetter, 2005] Cazenave, T. and Helmstetter, B. (2005). Combining tactical search and monte-carlo in the game of go. *CIG*, 5:171–175.

[Checkland and Scholes, 1990] Checkland, P. and Scholes, J. (1990). *Soft Systems Methodology in Action*. Wiley, John and Sons, Inc., New York, NY, USA.

[Connell and Sousa, 1983] Connell, J. H. and Sousa, W. P. (1983). On the evidence needed to judge ecological stability or persistence. *American Naturalist*, pages 789–824.

[Cox, 2008] Cox, A. (2008). What's wrong with risk matrices? *Risk analysis*, 28(2):497–512.

[De Dianous and Fiévez, 2006] De Dianous, V. and Fiévez, C. (2006). Aramis project: A more explicit demonstration of risk control through the use of bowtie diagrams and the evaluation of safety barrier performance. *Journal of Hazardous Materials*, 130(3):220–233.

[Devaud et al., 1980] Devaud, J., Groussaud, G., and Jacquet-Lagreze, E. (1980). Utadis: Une méthode de construction de fonctions d'utilité additives rendant compte de jugements globaux. *European Working Group on Multicriteria Decision Aid, Bochum*.

[Dias and Tsoukiàs, 2004] Dias, L. and Tsoukiàs, A. (2004). On the constructive and other approaches in decision aiding. In Hengeller Antunes, C., Figueira, J., and Clímaco, J., editors, *Proceedings of the 56th meeting of the EURO MCDA working group*, pages 13–28. CCDRC, Coimbra.

[Dodgson et al., 2009] Dodgson, J., Spackman, M., Pearman, A., and Phillips, L. (2009). Multi-criteria analysis: a manual. Economic history working papers, London School of Economics and Political Science, Department of Economic History.

[Doumpos et al., 2009] Doumpos, M., Marinakis, Y., Marinaki, M., and Zopounidis, C. (2009). An evolutionary approach to construction of outranking models for multicriteria classification: The case of the electre tri method. *European Journal of Operational Research*, 199(2):496–505.

[Doumpos and Zopounidis, 2002] Doumpos, M. and Zopounidis, C. (2002). *Multicriteria decision aid classification methods*, volume 73. Springer Science & Business Media.

[Drake, 1992] Drake, L. (1992). The non-market value of the swedish agricultural landscape. *European review of agricultural economics*, 19(3):351–364.

[Eden and Ackermann, 1998] Eden, C. and Ackermann, F. (1998). *Making Strategy: The Journey of Strategic Management*. SAGE Publications.

[Ellul, 1967] Ellul, J. (1967). *The technological society*. Vintage.

[Enserink et al., 2010] Enserink, B., Hermans, L., Kwakkel, J., Thissen, W., Koppenjan, J., and Bots, P. (2010). *Policy analysis of multi-actor systems*. Lemma The Hague.

[Fallah Tehrani and Huellermeier, 2013] Fallah Tehrani, A. and Huellermeier, E. (2013). Ordinal choquistic regression. In *EUSFLAT Conf.*

[Figueira et al., 2004] Figueira, J., Tervonen, T., Almeida-Dias, J., Lahdelma, R., and Salmiken, P. (2004). Smaa-tri: a parameter stability analysis method for electre tri. In *NATO advanced research workshop*, pages 20–24.

[Fitzgerald, 2015] Fitzgerald, R. D. (2015). *Social Impact Of The Industrial Revolution*. The University of Houston.

[Freedman, 2009] Freedman, D. A. (2009). *Statistical models: theory and practice*, page 128. Cambridge University press.

[Freeman, 2010] Freeman, R. E. (2010). *Strategic management: A stakeholder approach*. Cambridge University Press.

[Fullér, 1996] Fullér, R. (1996). Owa operators in decision making. *Exploring the limits of support systems, TUCS General Publications*, 3:85–104.

[Gallai et al., 2011] Gallai, N., Vaissiere, B., Potts, S., and Salles, J. (2011). Assessing the monetary value of global crop pollination services. *Oxford University Press*.

[Grabisch and Roubens, 2000] Grabisch, M. and Roubens, M. (2000). Application of the choquet integral in multicriteria decision making. *Fuzzy Measures and Integrals-Theory and Applications*, pages 348–374.

[Graham and Rhomberg, 1996] Graham, J. D. and Rhomberg, L. (1996). How risks are identified and assessed. *The annals of the American Academy of Political and Social Science*.

[Greco et al., 2001a] Greco, S., Matarazzo, B., and Slowinski, R. (2001a). Rough sets theory for multicriteria decision analysis. *European journal of operational research*, 129(1):1–47.

[Greco et al., 2001b] Greco, S., Matarazzo, B., Slowinski, R., and Stefanowski, J. (2001b). *An Algorithm for Induction of Decision Rules Consistent with the Dominance Principle*, pages 304–313. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Grimm and Wissel, 1997] Grimm, V. and Wissel, C. (1997). Babel, or the ecological stability discussions: an inventory and analysis of terminology and a guide for avoiding confusion. *Oecologia*.

[Gunderson, 2000] Gunderson, L. H. (2000). Ecological resilience-in theory and application. *Annual review of ecology and systematics*, pages 425–439.

[Hershey and Taiwo, 1999] Hershey, H. F. and Taiwo, A. (1999). Rating the rating scales. *Journal of Marketing Management*.

[Hillson and Murray-Webster, 2007] Hillson, D. and Murray-Webster, R. (2007). *Understanding and Managing Risk Attitude*. Gower.

[Hufschmidt et al., 1983] Hufschmidt, M. M., James, D. E., Meister, A. D., Bower, B. T., and Dixon, J. A. (1983). *Environment, natural systems, and development: an economic valuation guide*. Johns Hopkins University Press, Baltimore, MD.

[Jacquet-Lagrèze, 1979] Jacquet-Lagrèze, E. (1979). De la logique d'agrégation de critères vers une logique d'agrégation-désagrégation de préférences et de jugements. *Cahiers de l'ISMEA, Série sciences de gestion*, 13:839–859.

[Jacquet-Lagreze and Siskos, 1982] Jacquet-Lagreze, E. and Siskos, J. (1982). Assessing a set of additive utility functions for multicriteria decision-making, the uta method. *European Journal of Operational Research*, 10(2):151–164.

[Jacquet-Lagrèze and Siskos, 2001] Jacquet-Lagrèze, E. and Siskos, Y. (2001). Preference disaggregation: 20 years of mcda experience. *European Journal of Operational Research*, 130(2):233–245.

[Jan Duijm et al., 2008] Jan Duijm, N., Fiévez, C., Gerbec, M., Hauptmanns, U., and Konstandinidou, M. (2008). Management of health, safety and environment in process industry. *Safety Science*, 46(6):908–920. Occupational Safety and Risk at ESREL 2006.

[Jochem, 2006] Jochem, J. (2006). The economic value ofnatural andenvironmentalresources. *Institute for Applied Environmental Economics*.

[Kaler, 2002] Kaler, J. (2002). Morality and strategy in stakeholder identification. *Journal of Business Ethics*, 39(1-2):91–100.

[Katossky and Marical, 2011] Katossky, A. and Marical, F. (2011). *Evaluation économique des services rendus par les zones humides-Complémentarité des méthodes de monétarisation*. Commissariat général au développement durable.

[Keeney, 1992] Keeney, R. (1992). *Value-Focused Thinking. A Path to Creative Decision Making*. Harvard University Press, Cambridge.

[Keeney and Raiffa, 1994] Keeney, R. and Raiffa, H. (1994). Decisions with multiple objectives preferences and value tradeoffs, cambridge university press, cambridge, new york, 1993, 569 pages, isbn 0-521-44185-4 (hardback), 0-521-43883-7 (paperback). *Behavioral Science*, 39(2):169–170.

[Khakzad et al., 2012] Khakzad, N., Khan, F., and Amyotte, P. (2012). Dynamic risk analysis using bow-tie approach. *Reliability Engineering & System Safety*, 104:36–44.

[Kirch, 2008] Kirch, W., editor (2008). *Encyclopedia of Public Health*, chapter Level of Measurement, pages 851–852. Springer Netherlands, Dordrecht.

[Kotlowski and Slowinski, 2013] Kotlowski, W. and Slowinski, R. (2013). On nonparametric ordinal classification with monotonicity constraints. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2576–2589.

[Labreuche and Grabisch, 2003] Labreuche, C. and Grabisch, M. (2003). The choquet integral for the aggregation of interval scales in multicriteria decision making. *Fuzzy Sets and Systems*, 137(1):11–26.

[Labreuche et al., 2014] Labreuche, C., Tehrani, A. F., and Hüllermeier, E. (2014). Choquistic utilitaristic regression. In *DA2PL 2014*.

[Lahdelma et al., 1998] Lahdelma, R., Hokkanen, J., and Salminen, P. (1998). Smaa-stochastic multiobjective acceptability analysis. *European Journal of Operational Research*, 106(1):137–143.

[Lahdelma and Salminen, 2001] Lahdelma, R. and Salminen, P. (2001). Smaa-2: Stochastic multicriteria acceptability analysis for group decision making. *Operations Research*, 49(3):444–454.

[Lahdelma and Salminen, 2010] Lahdelma, R. and Salminen, P. (2010). *Stochastic Multicriteria Acceptability Analysis (SMAA)*, pages 285–315. Springer US, Boston, MA.

[Landry et al., 1983] Landry, M., Malouin, J.-L., and Oral, M. (1983). Model validation in operations research. *European Journal of Operational Research*, 14(3):207–220.

[Lave, 1987] Lave, L. (1987). Health and safety risk analyses: Information for better decisions. *Science*.

[Le Roux, 2011] Le Roux, T. (2011). Accidents industriels et régulation des risques: l'explosion de la poudrerie de grenelle en 1794. *Revue d'histoire moderne et contemporaine (1954-)*, 58(3):34–62.

[Lehmann and Hulbert, 1972] Lehmann, D. R. and Hulbert, J. (1972). Are three-point scales always good enough? *Journal of Marketing Research*, 9(4):444–446.

[Leroy et al., 2011] Leroy, A., Mousseau, V., and Pirlot, M. (2011). Learning the parameters of a multiple criteria sorting method. In *International Conference on Algorithmic DecisionTheory*, pages 219–233. Springer.

[Liénart, 2009] Liénart, S. (2009). *Intégrer la biodiversité dans les projets d'aménagement de la ville: qualifier les espaces pour contribuer aux choix d'aménagement*. Direction régionale et interdépartementale de l'équipement et de l'aménagement" of the French region Île de France.

[List et al., 2006] List, J. A., Sinha, P., and Taylor, M. H. (2006). Using choice experiments to value non-market goods and services: evidence from field experiments. *Advances in economic analysis & policy*, 5(2).

[Lorrillière et al., 2012] Lorrillière, R., Denis, C., and Alexandre, R. (2012). The effects of direct and indirect constraints on biological communities. *Ecological Modelling*.

[Luce, 1995] Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology*, 46(1):1–27.

[Mammeri, 2013] Mammeri, L. (2013). *A multiple criteria decision aiding tool for evaluating the overall comfort on board trains*. Theses, Université Paris Dauphine-Paris IX.

[Mateo and Ramón San, 2012] Mateo, J. and Ramón San, C. (2012). *Multi Criteria Analysis in the Renewable Energy Industry*, chapter Multi-Attribute Utility Theory, pages 63–72. Springer London, London.

[McCluskey and Lalkhen, 2007] McCluskey, A. and Lalkhen, A. G. (2007). Statistics ii: Central tendency and spread of data. *Continuing Education in Anaesthesia, Critical Care & Pain*, 7(4):127–130.

[McLachlan et al., 2005] McLachlan, G., Do, K.-A., and Ambroise, C. (2005). *Analyzing microarray gene expression data*, volume 422. John Wiley & Sons.

[McPhee, 2005] McPhee, I. (2005). Risk and risk management in the public sector. Technical report, Australian Institute of Company Directors, in conjunction with the Institute of Internal Auditors Australia Public Sector Governance and Risk Forum.

[Metropolis and Ulam, 1949] Metropolis, N. and Ulam, S. (1949). The monte carlo method. *Journal of the American Statistical Association*, 44(247):335–341. PMID: 18139350.

[Miller-George, 1956] Miller-George, A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63(2):81–97.

[Mitchell et al., 1997] Mitchell, R. K., Agle, B. R., and Wood, D. J. (1997). Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. *The Academy of Management Review*, 22(4):853–886.

[Moulin et al., 2016] Moulin, H., Brandt, F., Conitzer, V., Endriss, U., Lang, J., and Procaccia, A. D. (2016). *Handbook of Computational Social Choice*. Cambridge University Press.

[Mousseau, 2003] Mousseau, V. (2003). *Elicitation des préférences pour l'aide multicritère à la décision*. PhD thesis, Mémoire présenté en vue de l'obtention de l'habilitation à diriger des recherches, Université Paris-Dauphine.

[Mousseau et al., 2001] Mousseau, V., Figueira, J., and Naux, J.-P. (2001). Using assignment examples to infer weights for electre tri method: Some experimental results. *European Journal of Operational Research*, 130(2):263–275.

[Mousseau and Slowinski, 1998] Mousseau, V. and Slowinski, R. (1998). Inferring an electre tri model from assignment examples. *Journal of Global Optimization*, 12(2):157–174.

[Mousseau et al., 2000] Mousseau, V., Slowinski, R., and Zielniewicz, P. (2000). A user-oriented implementation of the electre tri method integrating preference elicitation support. *Computers and Operations Research*, 27.

[Nunes and van den Bergh, 2001] Nunes, P. A. and van den Bergh, J. C. (2001). Economic valuation of biodiversity: sense or nonsense? *Ecological economics*, 39(2):203–222.

[O.A.B., 2004] O.A.B., H. (2004). Application of value-focused thinking on the environmental selection of wall structures. *Journal of Environmental Management*, 70(2):181–187.

[Ortwin, 2005] Ortwin, R. (2005). White paper on risk governance; towards an integrative approach. Technical report, IRGC.

[Ostanello and Tsoukiàs, 1993] Ostanello, A. and Tsoukiàs, A. (1993). An explicative model of public interorganizational interactions. *European Journal of Operational Research*, 70(1):67–82.

[Ott et al., 2008] Ott, W., Baur, M., Kaufmann Rolf Frischknecht, Y., and Steiner, R. (2008). Assessment of biodiversity losses. Technical report, New Energy Externalities Developments for Sustainability.

[Öztürk and Tsoukiàs, 2008] Öztürk, M. and Tsoukiàs, A. (2008). Bipolar preference modeling and aggregation in decision support. *International Journal of Intelligent Systems*, 23(9):970–984.

[Pascual et al., 2010] Pascual, U., Muradian, R., Brander, L., Gómez-Baggethun, E., Martín-López, B., Verma, M., Armsworth, P., Christie, M., Cornelissen, H., Eppink, F., et al. (2010). The economics of valuing ecosystem services and biodiversity. *TEEB-Ecological and Economic Foundation*.

[Pawlak and Slowinski, 1994] Pawlak, Z. and Slowinski, R. (1994). Rough set approach to knowledge-based decision support. *International Transactions in Operational Research*, 99(1):48–57.

[Pinheiro et al., 2014] Pinheiro, P. R., Machado, T. C. S., and Tamanini, I. (2014). *Handing the Classification Methodology ORCLASS by Tool OrclassWeb*, pages 61–72. Springer International Publishing, Cham.

[Regenwetter et al., 2011] Regenwetter, M., Dana, J., and Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, 118(1):42.

[Rolland, 2008] Rolland, A. C. (2008). *Ordinal preferences aggregation rules with reference points for decision aiding*. Theses, Université Pierre et Marie Curie-Paris VI.

[Ron, 1995] Ron, K. (1995). A study of cross validation and bootstrap for accuracy estimation and model selection. In *IJCAI*.

[Rousval, 2005] Rousval, B. (2005). *Aide multicritère à l'évaluation de l'impact des transports sur l'environnement*. Theses, Université Paris Dauphine-Paris IX.

[Roux, 2013] Roux, P. (2013). *Méthode dévaluation de la gravité des conséquences environnementales d'un accident industriel*. INERIS, DRA.

[Roy, 1978] Roy, B. (1978). Electre iii: Un algorithme de classement fondé sur une représentation floue des préférences en présence de critères multiples. *Cahiers du CERO*, 20(1).

[Roy and Bouyssou, 1993] Roy, B. and Bouyssou, D. (1993). *Aide Multicritère à la Décision : Méthodes et Cas*. Economica, Paris.

[Roy and Damart, 2002] Roy, B. and Damart, S. (2002). L'analyse coûts-avantages, outil de concertation et de légitimation? *Metropolis*, 108-109:7–16.

[Saaty, 1990] Saaty, T. L. (1990). Desicion making by the analytic hierarchy process: Theory and applications how to make a decision: The analytic hierarchy process. *European Journal of Operational Research*, 48(1):9–26.

[Salminen et al., 2007] Salminen, P., Yevseyeva, I., Miettinen, K., and Risto, L. (2007). Smaa-classification-a new method for nominal classification.

[Sarin, 2001] Sarin, R. K. (2001). Multi-attribute utility theory. In *Encyclopedia of Operations Research and Management Science*. Springer US.

[Sheng et al., 2005] Sheng, H., Fui-Hoon Nah, F., and Siau, K. (2005). Strategic implications of mobile technology: A case study using value-focused thinking. *The Journal of Strategic Information Systems*, 14(3):269–290. The Future is UNWIRED: Organizational and Strategic Perspectives.

[Siskos et al., 2005] Siskos, Y., Grigoroudis, E., and Matsatsinis, N. F. (2005). *Multiple Criteria Decision Analysis: State of the Art Surveys*, chapter UTA Methods, pages 297–334. Springer New York, New York, NY.

[Sloan and Woniakowski, 1998] Sloan, I. H. and Woniakowski, H. (1998). When are quasi-monte carlo algorithms efficient for high dimensional integrals? *Journal of Complexity*, 14(1):1–33.

[Slowinski et al., 2002] Slowinski, R., Greco, S., and Matarazzo, B. (2002). Axiomatization of utility, outranking and decision-rule preference models for multiple-criteria classification problems under partial inconsistency with the dominance principle. *Control and Cybernetics*, 31(4):1005–1035.

[Sobrie et al., 2013] Sobrie, O., Mousseau, V., and Pirlot, M. (2013). *Learning a Majority Rule Model from Large Sets of Assignment Examples*, pages 336–350. Springer Berlin Heidelberg.

[Sobrie et al., 2015] Sobrie, O., Mousseau, V., and Pirlot, M. (2015). *Algorithmic Decision Theory: 4th International Conference, ADT 2015, Lexington, KY, USA, September 27-30, 2015, Proceedings*, chapter Learning the Parameters of a Non Compensatory Sorting Model, pages 153–170. Springer International Publishing, Cham.

[Stefanowski and Vanderpooten, 2001] Stefanowski, J. and Vanderpooten, D. (2001). Induction of decision rules in classification and discovery-oriented perspectives. *International Journal of Intelligent Systems*, 16(1):13–27.

[Stevens, 1946] Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684):677–680.

[Stevenson, 2010] Stevenson, A., editor (2010). *Oxford dictionary of English*. Oxford University Press.

[Stewart and Cohen, 1995] Stewart, I. and Cohen, J. S. (1995). *The Collapse of Chaos: Discovering Simplicity in a Complex World*. PP Science.

[Tamiz et al., 1998] Tamiz, M., Jones, D., and Romero, C. (1998). Goal programming for decision making: An overview of the current state-of-the-art. *European Journal of Operational Research*, 111(3):569–581.

[Tehrani et al., 2011] Tehrani, A. F., Cheng, W., and Hüllermeier, E. (2011). Choquistic regression: Generalizing logistic regression using the choquet integral. In *EUSFLAT Conf.*, pages 868–875.

[Thomson and Monje, 2015] Thomson, K. and Monje, C. (2015). Secretarial officers modal administrators. Technical report, Office of the Secretary Of Transportation.

[Tsoukiàs, 2008] Tsoukiàs, A. (2008). From decision theory to decision aiding methodology. *European Journal of Operational Research*, 187(1):138–161.

[Verma et al., 1981] Verma, S., Tonk, I., and Dalela, R. (1981). *Determination of the Maximum Acceptable Toxicant Concentration (MATC) and the safe Concentration for Certain Aquatic Pollutants*. Acta Hydrochimica et Hydrobiologica.

[Walker and Duncan, 1967] Walker, S. H. and Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179.

[Warren, 1973] Warren, S. M. (1973). The effects of scaling on the correlation coefficient: A test of validity. *Journal of Marketing Research*, 10(3):316–318.

[Weisberg, 1992] Weisberg, H. (1992). Central tendency and variability sage university paper series on quantitative application in social sciences, series no. 07-038. a. virding, ed.

[Wiley, 2009] Wiley, J. (2009). *Appendix A: Understanding and Using F-N Diagrams*, pages 109–117. Center for Chemical Process Safety.

[Worcester and Burns, 1975] Worcester, R. M. and Burns, T. R. (1975). Statistical examination of relative precision of verbal scales. *Journal of the Market Research Society*, 17(3):181–197.

[Yager, 1988] Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decisionmaking. *IEEE Trans. Systems, Man, and Cybernetics*, 18(1):183–190.

[Yevseyeva, 2007] Yevseyeva, I. (2007). *Solving Classification problems with Multi criteria decision aiding Approaches*. PhD thesis, University of Jyvaskyla.

[Zopounidis and Doumpos, 1997] Zopounidis, C. and Doumpos, M. (1997). A multicriteria decision aid methodology for the assessment of country risk. *European Research on Management and Business Economics*, 3(3):13–33.

[Zopounidis and Doumpos, 2001] Zopounidis, C. and Doumpos, M. (2001). A preference disaggregation decision support system for financial classification problems. *European Journal of Operational Research*, 130(2):402–413.

Résumé

Cette thèse s'appuie sur deux axes. L'un appliqué traite de la création d'un indicateur dont le but est d'évaluer la gravité attendue des conséquences d'un scénario de pollution accidentelle. J'ai choisi d'utiliser des outils méthodologiques appartenant au domaine de l'aide multi-critères à la décision pour traiter ce premier sujet. Ce problème impliquant plusieurs disciplines scientifiques, j'ai choisi de le diviser en plusieurs sous-problèmes à travers une arborescence de critères. J'ai également impliqué plusieurs experts, notamment en toxicologie et en écologie afin de mieux prendre en compte les aspects liés à ces deux disciplines dans la création de cet indicateur.
L'étude des méthodes de tri multicritère effectuée lors des recherches sur le premier axe m'a amené à en proposer une nouvelle que j'ai nommé algorithme du Dominance Based Monte Carlo (DBMC). Cet algorithme a comme particularités de n'être pas fondé sur un modèle et de fonctionner de manière stochastique. Nous avons étudié ses propriétés théoriques, en particulier nous avons démontré qu'en dépit de sa nature stochastique, le résultat de l'algorithme Dominance Based Monte Carlo converge presque surement. Nous avons également étudié son comportement et ses performances pratiques à travers un test nommé k-fold cross validation et les avons comparés aux performances d'autres algorithmes d'élicitation des préférences pour le tri multi-critères.

Abstract

This thesis is based on two main axes. The first one deals with the creation of an indicator that aims at evaluating the expected severity of the consequences of a scenario of accidental pollution. In order to create this methodology of evaluation, I chose to use methodological tools from multi-criteria decision aiding. So as to deal with the complexity of this problem, i decided to split it into several sub-problems using a hierarchy of criteria, being mainly inspired by the "value focused thinking approach". In this work, I interacted with several experts in toxicology and in ecology in order to better deal with every aspect of this problem.
While studying several elicitation methods for the multi-criteria sorting problem, I proposed a new one that I named Dominance Based Monte Carlo algorithm (DBMC), which brings me to the secons axis of this thesis. This elicitation algorithm has two main specificities: being model free and a stochastic functionning. In this thesis, we study its theoretical properties. In particular, we prove that despite its stochastic nature, the result of the Dominance Based Monte Carlo algorithm converges almost surely. We also study its practical performances through a test named k-fold validation and we compared these performances to those of other elicitation algorithms for the sorting problem.

Mots Clés

Aide multi-critères à la décision
Élicitation des préférences
Études de risques

Keywords

Multi-criteria decision aiding
Preference elicitation
Risk studies