Évaluation de la qualité des contributions et des contributeurs sur plateformes de crowdsourcing

Constance Thierry

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Constance THIERRY

Évaluation de la qualité des contributions et des contributeurs sur plateformes de *crowdsourcing*

Thèse présentée et soutenue à Lannion, le 14 décembre 2021
Unité de recherche : IRISA

Rapporteurs avant soutenance :

Sihem AMER-YAHIA	Directrice de recherche, LIG, Grenoble
Allel HADJALI	Professeur des universités, LIAS, ENSMA, Poitiers

Composition du Jury :

Président :	Éric LEFÈVRE	Professeur des universités, LGI2A, Université d'Artois
Examinateurs :	Sihem AMER-YAHIA	Directrice de recherche, LIG, Grenoble
	Allel HADJALI	Professeur des universités, LIAS, ENSMA, Poitiers
	Yolande LE GALL	Maître de conférences, IRISA, Lannion
	Jean-Christophe DUBOIS	Maître de conférences, IRISA, Lannion
Dir. de thèse :	Arnaud MARTIN	Professeur des universités, Université de Rennes 1

Remerciements

J'ai rapidement su au cours de mes études que je voulais faire une thèse de doctorat et grâce à un grand nombre de personnes, j'ai pu réaliser ce rêve. C'est pourquoi je souhaiterais aujourd'hui les remercier.

En premier lieu, je souhaite remercier mon directeur de thèse Arnaud Martin et mes encadrants Jean-Christophe Dubois et Yolande Le Gall pour m'avoir permis de faire mon projet de fin d'étude d'ingénieur sous leur supervision et de m'avoir fait confiance pour continuer mon travail en doctorat. J'ai beaucoup appris auprès d'eux, et le temps et les conseils avisés qu'ils m'ont offert m'ont été précieux. Ils ont su être présents pour la rédaction de ce manuscrit et cette thèse aurait été toute autre sans eux.

Je souhaite également exprimer ma gratitude à Sihem Amer-Yahia et Allel Hadjali pour la qualité des rapports qu'ils ont rendu à la suite de la lecture de cette thèse. Je leur suis reconnaissante pour leur participation en tant que membre du jury de soutenance, mais aussi membre de mon comité de suivi individuel de thèse. Je souhaite aussi remercier Éric Lefevre pour sa participation à mon jury de thèse en tant que président et pour les retours qu'il m'a fait lors de la soutenance.

Je n'oublie pas que c'est grâce aux financements du Conseil départemental des Côtes d'Armor et le projet ANR *Headwork* que j'ai pu réaliser ces trois années de thèse au sein de l'équipe DRUID de l'IRISA.

J'ai apprécié réaliser ma thèse dans l'équipe DRUID, et bien que celle-ci soit scindée sur deux sites, je me suis sentie intégrée aussi bien à Lannion qu'à Rennes. Pour cela, je tiens à remercier Zoltan Miklos et David Gross-Amblard pour leur bienveillance, car c'est notamment grâce à eux que je me sentais inclus à Rennes. J'ai trouvé particulièrement intéressants nos échanges sur le *crowdsourcing* avec David Gross-Amblard.

J'ai eu la chance de prendre part au projet IRDICS qui s'est fait en partenariat entre les équipes DRUID et LOKI de l'INRIA Lille. Je suis contente d'avoir travaillé avec les membres de l'équipe LOKI sur ce projet, ce fut très instructif pour moi d'échanger avec une équipe de recherche d'un autre domaine. C'est pourquoi je remercie l'équipe LOKI et

plus particulièrement Géry Casiez, Thomas Pietrzak et Sylvain Malacria qui m'ont permis de développer de nouvelles connaissances en IHM.

J'ai eu le plaisir d'effectuer des missions d'enseignements à l'ENSSAT et à l'IUT de Lannion. Je tiens à témoigner ma reconnaissance aux équipes pédagogiques de ces deux établissements qui m'ont accueillie chaleureusement et m'ont accompagné dans la découverte du monde de l'enseignement. Je remercie plus particulièrement : Hélène Jaudoins, Virginie Thion, Isabelle Henrio, Ludovic Lietard, Bertrand De Villeneuve, Gildas Quiniou et Jean-Christophe Vialat.

Je remercie également Jonathan Chevelu et David Le Roy membres de l'équipe organisatrice du challenge Ada Lovelace, car grâce à ce projet, j'ai pu évoluer et gagner en confiance en moi ce qui a eu des répercussions bénéfiques dans ma thèse.

Je suis heureuse d'avoir pu compter sur le soutien de ma famille qui bien que n'étant pas du domaine, s'est efforcée de comprendre la magie de l'informatique. Je remercie ma mère pour m'avoir encouragée et mon père pour son aide lorsque je collectais des photos d'oiseaux et qu'il m'a transmis ses propres clichés annotés des noms scientifiques.

Je souhaite également remercier mes amis pour avoir été présent dans les meilleurs moments comme dans les plus difficiles. Un grand merci à Marine Bouloy mon amie de toujours et Marine Bécu, je suis heureuse que bien que nous ayons pris des chemins éloignés, nous soyons restées si proches, votre amitié m'a vraiment été essentielle au cours de ces trois ans. Merci également à tous mes amis que j'ai eus le bonheur de rencontrer sur Lannion : Clément Caresmel, Erwan Colin, Corentin Lavaud, Romain Mercier, Stella-Maria Profizi, Jules Courjault et Maxime Duris. Merci également à Marie-Anne Lacroix, qui est quelqu'un de très attentionnée et qui a su m'écouter lorsque j'avais des doutes et des inquiétudes, je lui en suis très reconnaissante. J'ai hâte de pouvoir assister à sa soutenance de thèse, et je lui souhaite à elle, à Alexis Vershelde et à Benoît Fournier bon courage pour la fin de leur thèse. À mon ami Emmanuel Doumard qui a choisi de commencer une thèse, je lui souhaite de vivre pleinement cette belle expérience.

Finalement, je remercie ma moitié qui a su croire en moi quand moi, je n'y arrivais pas et qui est toujours resté à mes côtés. Merci également à mes trois boules de poils : Hadès, Calypso et Tensoon, vos ronronnements et vos câlins sont toujours agréables et encourageants.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

Chapitre 1

Introduction

Résumé : Le *Big data* traduit l'accumulation d'un grand nombre de données en temps réel, de nombreux secteurs d'activité y ont recours comme le médical, la finance ou encore les transports. Il existe différentes approches pour collecter les données, parmi elles le *crowdsourcing* qui consiste en l'externalisation de tâches en ligne sur des plateformes où des individus vont venir apporter leur contribution. Une des problématiques du *Big data* est de s'assurer de la crédibilité des sources d'information générant les données, mais aussi de vérifier la qualité de leur contenu pour une exploitation optimale. Après avoir introduit le sujet de la thèse dans ce chapitre, nous exposons les principales contributions réalisées et le plan du manuscrit.

Sommaire

1.1 Introduction

Cette thèse porte sur la modélisation de données provenant de plateformes de *crowdsourcing*. Les données collectées de la sorte sont massives et s'inscrivent dans le contexte du *Big data* qui est introduit dans ce chapitre. Le domaine du *crowdsourcing* inclut plusieurs problématiques notamment celles du contrôle du *workflow* des données et de l'estimation de leur qualité pour une agrégation optimale. Le projet ANR *Headwork* qui est également présenté dans ce chapitre répond à certaines de ces problématiques.

1.1.1 Le *Big data*

Le terme *Big data* caractérise un ensemble massif de données collectées afin d'être analysées et fouillées, en vue d'en extraire des informations exploitables, ou d'être utilisées dans le cadre de projet d'apprentissage automatique. Laney 2001 définit le *Big data* d'après un modèle tridimensionnel : le volume, la variété et la vélocité. Dans le cas du *Big Data*, le volume de données recueillies est très grand et inclut de la variété, car les données peuvent être brutes, non-structurées ou encore semi-structurées. La vélocité traduit le fait que les données sont produites et récoltées en temps réel.

Le *Big data* a pris son essor grâce au développement du web qui a facilité l'acquisition rapide d'un large volume de données diversifiées. Les données peuvent être collectées auprès de différentes sources comme des sociétés. Il peut s'agir d'entreprises qui commercialisent des informations sur leurs clients ou d'autres spécialisées dans le recueil de données comme les plateformes de *crowdsourcing*. Les données peuvent également provenir des réseaux sociaux, où les utilisateurs sont quotidiennement actifs. L'estimation des éléments d'influence grâce aux données de réseaux sociaux permet d'optimiser les campagnes marketing d'entreprises. Les objets connectés sont aussi une importante source d'information. Les montres connectées par exemple, permettent d'acquérir des informations sur la géolocalisation de leur porteur. Certaines grandes chaînes de restaurations déterminent ainsi l'emplacement optimal pour leur enseigne local en analysant ces données de géolocalisation associées à d'autres informations comme la démographie ou les préférences.

L'analyse et l'exploitation de ces données massives ont des retombées dans de très nombreux domaines. Les entreprises privées les traitent pour le profilage d'utilisateurs et le développement de nouveaux produits. Dans la recherche scientifique, les données sont par exemple employées en informatique pour constituer des corpus d'entraînement pour

de l'apprentissage automatique. Nous pouvons aussi mentionner l'utilisation des données collectées grâce aux sciences participatives pour l'étude de la biodiversité, comme le recensement des oiseaux de jardin. En politique, le *Big data* permet de faire des prédictions sur les élections. Il est également utile en médecine. On peut ainsi citer l'application *COVID Near You*[1] développée par des chercheurs de la Harvard Medical School afin de demander à la foule touchée par la Covid-19 de spécifier leurs symptômes en temps réel. Les données recueillies par cette plateforme permettent aux experts d'estimer les variations de la propagation géolocalisée du virus au cours du temps. Nous avons mentionné quelques exemples de domaines d'exploitation du *Big Data*, mais il en existe d'autres encore : la communication, l'écologie, la finance...

Les secteurs d'applications où le Big data est pertinent sont nombreux, et l'exploitation des données revêt une importance capitale voire parfois vitale. C'est pourquoi il est essentiel de vérifier la crédibilité des sources d'information, leur provenance et la qualité de leur contenu avant leur utilisation. Dans le cadre de cette thèse nous nous intéressons à ces problématiques pour des données issues de plateformes de *crowdsourcing*. Ces plateformes génèrent de grandes quantités de données et sont par conséquent une source de *Big data*. Les données recueillies sont utilisées par des particuliers, pour la recherche ou encore par des entreprises dans des domaines très variés allant des sciences sociales aux sciences exactes. Nous introduisons le *crowdsourcing* dans la section suivante.

1.1.2 Le *crowdsourcing*

Le *crowdsourcing* consiste à employer un large nombre de personnes grâce à internet afin de contribuer à des tâches permettant l'acquisition de données. Il s'agit bien d'une source du *Big data*, car un volume important de données, de grande variété, est collecté de la sorte. Les contributions recueillies en temps réel peuvent prendre des formes multiples en fonction des instructions associées à la tâche à réaliser. L'application *COVID Near You* présentée dans la section précédente comme une application médicale du *Big data* est également une forme de *crowdsourcing*. Nous pouvons aussi mentionner l'encyclopédie en ligne Wikipédia[2] comme une plateforme de *crowdsourcing* et une représentation importante du *Big data*. Dans son ensemble, le *crowdsourcing* est à la portée de tout un chacun et impacte de nombreux domaines d'applications scientifiques, sociaux et économiques.

1. https ://hms.harvard.edu/news/crowdsourcing-covid-19 (27/10/2021)
2. https ://fr.wikipedia.org/ (27/10/2021)

Malheureusement, cette ouverture des plateformes à la foule bien que positive, car elle génère de la diversité dans les contributions, apporte également son lot de problématiques. En effet, la variété des profils des contributeurs qui composent la foule induit des contributions de qualité inégale. Afin de résoudre cette problématique, nous proposons dans cette thèse un modèle, nommé MONITOR pour la modélisation des contributions et contributeurs dans les plateformes de *crowdsourcing*. L'élaboration de MONITOR s'inscrit dans le contexte du projet ANR *Headwork* présenté dans la section suivante.

1.1.3 Le projet Headwork

Le projet ANR *Headwork*[3] propose des *workflows* complexes d'analyse de données à partir desquels les contributeurs peuvent interagir d'une question à une autre. Il s'agit par exemple de poser en priorité une question aux contributeurs qui possèdent les meilleures connaissances du domaine, ou encore de tenir compte de la réponse du contributeur à la question q pour poser la question $q + 1$. Le projet tient également compte des imperfections des données collectées grâce à la modélisation des contributeurs et la gestion de l'incertitude et de l'imprécision de leur réponse. Dans cet objectif, *Headwork* s'appuie sur l'expertise des plateformes de *crowdsourcing* et des équipes de recherche spécialisées dans la gestion des données et la modélisation des *workflows*.

Cette thèse participe au projet *Headwork* par la modélisation des contributeurs et de leurs réponses qui peuvent être imprécises et incertaines. La section suivante introduit les principales contributions réalisées au cours de ces trois années.

1.2 Contributions

Le projet *Headwork* nécessite de tenir compte de la spécificité des contributions humaines dans les plateformes de *crowdsourcing*. Or les données collectées dans les plateformes de *crowdsourcing* ne permettent pas d'identifier les imperfections relatives aux contributions humaines ce qui a un impact dommageable sur l'agrégation des réponses. C'est pourquoi nous souhaitons enrichir les informations recueillies sur les plateformes de *crowdsourcing* afin de faciliter la prise de décision finale par l'employeur. Pour ce faire nous prenons en compte l'incertitude et l'imprécision des réponses. Cependant, il n'existe pas d'interface pour collecter des données imprécises et incertaines dans les plateformes ce

3. http ://headwork.gforge.inria.fr/index (11/09/2021)

qui nous a amenés à définir une interface adéquate. Puisque cette interface de recueil de données imparfaite est nouvelle pour le contributeur, nous avons étudié son utilisation de l'imprécision et de l'incertitude en fonction de la difficulté de la tâche qui lui est demandée. Nous proposons également MONITOR, un modèle défini pour estimer le profil du contributeur et faciliter l'agrégation des contributions imprécises et incertaines. Grâce à l'interface définie, nous avons réalisé des campagnes de *crowdsourcing* afin de tester MONITOR sur des données réelles. L'ensemble de ces contributions sont développées dans cette section.

Proposition d'une nouvelle interface pour le *crowdsourcing*

La façon dont la tâche est communiquée au contributeur a un impacte sur la qualité de ses réponses. Pourtant il n'existe pas d'interface banalisée pour le *crowdsourcing* et celles existantes ne permettent pas de capturer les hésitations éventuelles du contributeur.

Au cours de cette thèse, nous avons défini dans le cadre d'un projet PEPS en collaboration avec l'équipe LOKI d'INRIA Lille [4] une nouvelle interface de *crowdsourcing* qui offre davantage de capacité d'expression au contributeur. Grâce à l'interface proposée, le contributeur peut sélectionner plusieurs réponses en cas d'indécision et donner sa confiance en sa contribution. L'interface permet ainsi d'enrichir les données recueillies dans l'objectif d'améliorer les résultats. Or, d'après nos expériences, l'interface que nous proposons permet au commanditaire de la campagne un gain financier et qualitatif comparé à l'utilisation d'une interface traditionnelle de *crowdsourcing*. La publication de Thierry et al. 2020a introduit cette interface.

Mise en évidence de la relation entre : difficulté, certitude et imprécision dans les plateformes de *crowdsourcing*

L'étude des contributions recueillies grâce à l'interface proposée nous permet de constater un lien entre l'imprécision et l'incertitude d'une réponse et la difficulté de la question. Plus une question est difficile plus elle amène le contributeur à douter de la réponse à renseigner. La difficulté éprouvée dans la réalisation de la campagne peut varier d'un contributeur à l'autre d'après ses connaissances du domaine de la tâche, c'est pourquoi la difficulté éprouvée par le contributeur est révélatrice de son profil. Par exemple, un ornithologue a davantage de facilités pour l'annotation de photos d'oiseaux qu'un néophyte,

4. https ://loki.lille.inria.fr/ (13/09/2021)

il sera donc plus aisé pour le premier de réaliser une tâche de ce type que pour le second.

Les expériences menées au cours de la thèse ont montré que la difficulté d'une question impacte la certitude du contributeur en sa réponse. De plus, nous avons établi que plus la tâche est difficile, plus le contributeur utilise la possibilité qui s'offre à lui d'être imprécis.

Validation de l'hypothèse de Philippe Smets

Dans un contexte de *crowdsourcing*, l'imprécision du contributeur est révélatrice de son hésitation entre plusieurs choix de réponse. La certitude du contributeur traduit sa confiance en l'exactitude de sa réponse. La moyenne des valeurs de certitude des réponses correcte indépendamment des mauvaises réponses permet de constater que les réponses correctes ont des valeurs de certitudes plus élevées que les mauvaises réponses. La certitude renseignée par le contributeur est par conséquent une information pertinente à intégrer dans le processus d'agrégation des réponses. Or Smets 1997 émet l'hypothèse suivante : plus un individu est imprécis plus il est certain et réciproquement, plus il est précis moins il est certain. Dans le cadre d'une activité de *crowdsourcing*, l'hypothèse de Smets 1997 se traduit par le fait qu'en sélectionnant plusieurs réponses le contributeur est plus confiant sur la qualité de sa participation. D'après la réciproque lorsqu'il sélectionne moins de réponses, il est moins confiant. L'étude de données de *crowdsourcing* imparfaites collectées grâce à l'interface proposé montre que pour des questions de même difficulté les contributeurs qui ont choisi plusieurs réponses sont plus certain de leur contribution que ceux qui ont sélectionné une unique réponse. Grâce à l'analyse de données provenant de différentes compagnes de *crowdsourcing* l'hypothèse de Smets 1997 est validée dans ce contexte par Thierry et al. 2021. Il est donc intéressant de permettre au contributeur d'être imprécis et de le modéliser afin de l'inclure dans le processus d'agrégation des réponses, ce que fait MONITOR.

Définition de MONITOR

Les méthodes d'estimation de profil existantes considèrent indépendamment la qualification du contributeur et son comportement dans la réalisation de la tâche. Ces approches amènent fréquemment à une estimation binaire du profil indiquant s'il faut accepter ou refuser la contribution. Or une estimation plus fine du profil permettrait de mettre en évidence des contributeurs volontaires, mais éprouvant des difficultés et requérant une aide. De plus, la plupart des méthodes qui s'intéressent à un panel de profils requièrent des données d'or qui servent de valeurs de référence pour la classification du contributeur.

Cependant, il n'est pas toujours possible d'avoir des données d'or dans les campagnes ce qui restreint l'application de ces méthodes.

MONITOR est un modèle permettant l'estimation du profil d'un contributeur grâce à sa qualification pour la tâche, mais également de son comportement lors de la réalisation de la campagne. En effet, les travaux existants permettant de déterminer le profil considèrent soit la qualification du contributeur soit son comportement. Il n'existe pas à l'heure actuelle de modèle considérant ces deux éléments sans l'utilisation de données d'or. MONITOR se fonde sur la théorie des fonctions de croyance pour la modélisation de l'ensemble des éléments le constituant. L'évolution de ces travaux au cours de la thèse nous a permis de réaliser plusieurs publications : Thierry et al. 2018, Thierry et al. 2019, et Thierry et al. 2020b.

Création de bases de données incertaines et imprécises pour la validation des travaux

Au cours de la thèse, plusieurs campagnes de *crowdsourcing* ont été définies, implantées et réalisées sur une plateforme dédiée à cette activité afin de collecter l'ensemble des données nécessaires à nos expériences. En effet, peu de travaux portent sur l'utilisation des fonctions de croyance dans un contexte de *crowdsourcing*. Parmi les études existantes, la moitié d'entre elles utilise des données générées pour leurs tests, alors que l'autre moitié utilise des données provenant de campagnes de *crowdsourcing* mais uniquement issues de réponses précises. Nous avons tenu à réaliser l'intégralité de nos expériences sur des données réelles, provenant de campagnes de *crowdsourcing*, et dont les contributions peuvent être imprécises. Nous avons ainsi effectué des campagnes de *crowdsourcing* et constitué nos propres bases de données incertaines et imprécises. L'ensemble de ces données est à la disposition de la communauté scientifique [5].

L'ensemble des contributions introduites ici sont détaillées dans les prochains chapitres de cette thèse dont le plan est présenté dans la section suivante.

1.3 Plan de la thèse

Cette thèse porte sur l'évaluation de la qualité des contributions et des contributeurs sur plateformes de *crowdsourcing* grâce à la théorie des fonctions de croyance. La définition

5. https ://gitlab.inria.fr/cthierry/imprecise_uncertain_dataset (29/10/2021)

du *crowdsourcing* est introduite dans le chapitre 2, qui explique notamment les différents types de *crowdsourcing* existants et les plateformes associées. Sur ces plateformes deux acteurs interagissent toujours sans se rencontrer et généralement sans communiquer. Il s'agit des employeurs qui sont les commanditaires des campagnes de *crowdsourcing*, et les contributeurs qui répondent au besoin de l'employeur. Le chapitre 2 présente les types d'employeurs qui utilisent les plateformes, une étude démographique des contributeurs, et les raisons qui poussent ces deux parties à recourir au *crowdsourcing*. Le *crowdsourcing* rencontre plusieurs problématiques : juridiques, sociales ou techniques exposées dans ce chapitre ainsi que les travaux existants proposant de les résoudre.

Nous présentons dans le chapitre 3 les définitions mathématiques nécessaires à la compréhension de la théorie des fonctions de croyance. Ce chapitre inclut un descriptif des opérateurs de combinaison et de décision utilisés. De même, les travaux existants, portant sur l'utilisation des fonctions de croyance dans un contexte de *crowdsourcing* sont également exposés.

Le chapitre 4 introduit l'interface définie au cours de la thèse afin d'offrir davantage de capacité d'expression au contributeur. Le protocole expérimental défini pour la validation de l'interface est présenté dans un premier temps, puis les résultats obtenus sont exposés. L'étude des résultats montre notamment l'impact de la difficulté de la tâche sur les contributions. L'intérêt pour l'employeur d'utiliser l'interface proposée plutôt qu'une interface de *crowdsourcing* plus traditionnelle est également développée. Ce chapitre inclut une étude des retours faits par les contributeurs après avoir utilisé l'interface proposée.

Au cours de cette thèse, nous avons défini un modèle permettant de modéliser le profil des contributeurs grâce aux données collectées à l'aide d'une interface semblable à celle définie dans le chapitre 4. MONITOR, le modèle proposé pour estimer le profil d'un contributeur, est introduit dans le chapitre 5. Le chapitre 6 présente les campagnes de *crowdsourcing* réalisées pour collecter les données requises pour nos expériences. Des comparaisons sont faites entre les résultats obtenus par MONITOR et d'autres études présentées dans l'état de l'art pour l'estimation du profil. De même, nous comparons l'agrégation des contributions modélisées par des fonctions de croyance à des méthodes d'agrégations plus couramment employées dans les plateformes de *crowdsourcing* et introduites dans le chapitre 2.

Finalement le chapitre 7 conclut ce manuscrit et expose les perspectives de recherche qui peuvent être envisagées dans la continuité de cette thèse.

Chapitre 2

Le crowdsourcing

Résumé : Le *crowdsourcing* consiste en l'externalisation de tâches à une foule de contributeurs sur une plateforme dédiée. Les différents types de plateformes et les acteurs engagés dans une activité de *crowdsourcing* sont présentés dans ce chapitre. Les problématiques du domaine et l'état de l'art associé sont également exposés.

Sommaire

2.1 Introduction

La proposition par une organisation d'une activité de réflexion à une foule d'individus est un phénomène plus ancien qu'il n'y paraît et qui existait bien avant internet. Par exemple, en 1714, le gouvernement britannique publie le *Longitude Act*, une loi récompensant de vingt mille livres la personne qui établirait une méthode pour la détermination de la longitude d'un navire en pleine mer. Ce prix a été remporté des années plus tard par l'horloger John Harrison. Cette forme de production participative a également pris des aspects moins scientifiques et plus artistiques, comme des concours pour la création d'affiches ou de logos. La foule participant à ces productions reste généralement dans un cercle géographique proche en raison des limites des moyens de communication de l'époque.

Par la suite, la mondialisation et l'expansion du web 2.0 ont permis de développer de manière exponentielle un éventail de possibilités. De nombreux services ont été adaptés à internet, avec notamment la possibilité pour une foule d'individus de participer à une activité d'intérêt commun. L'externalisation de tâches par un organisme prend alors une ampleur mondiale, ce qui intensifie la taille de la foule qu'il est possible de regrouper et la diversité des individus qui la compose. De plus, les tâches réalisables par le biais du web sont très variées, il peut s'agir de tests produits, de sondages, de résolution de problèmes... L'univers des possibles est vaste sur le web. C'est dans ce contexte qu'a été introduite pour la première fois la notion de *crowdsourcing* par Howe 2006.

Le plan de ce chapitre est le suivant, la section 2.2 définit le *crowdsourcing*. La section 2.3 présente les différents types de plateformes de *crowdsourcing* existants et la section 2.4 les acteurs qui y interagissent. La section 2.5 détaille les problématiques du domaine et la section 2.6 expose l'état de l'art associé. La section 2.7 conclut ce chapitre.

2.2 Définition du *crowdsourcing*

Le terme *crowdsourcing* est un néologisme introduit par Howe 2006 combinant les termes *"crowd"* et *"outsourcing"*. Il consiste à l'externalisation de tâches non réalisables par l'ordinateur, car non automatisables ou nécessitant une expertise humaine, à une foule d'individus sur une plateforme dédiée. Les utilisateurs de ces plateformes sont appelés contributeurs, ce sont eux qui résolvent les tâches disponibles.

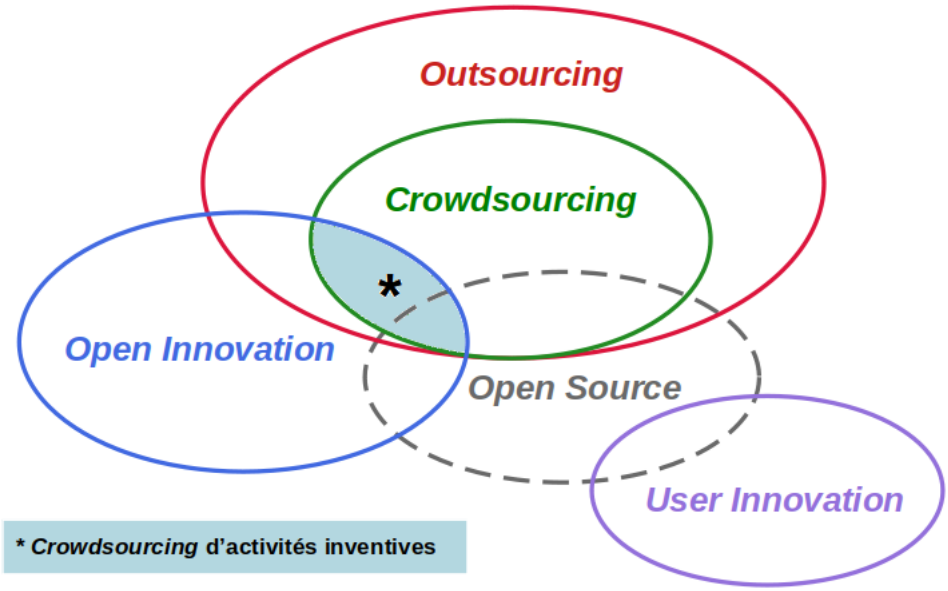


FIGURE 2.1 – Schéma présentant le positionnement du *crowdsourcing*

Il faut veiller à distinguer le *crowdsourcing* des autres activités décentralisées ou communautaires. GUITTARD et al. 2010 et SCHENK et al. 2012 différencient dans leurs travaux le *crowdsourcing* de l'*Open Innovation*, du *User Innovation* et de l'*Open Source*. Ces trois termes et leur positionnement par rapport au *crowdsourcing* sont expliqués dans les paragraphes suivants et illustrés par la figure 2.1.

Open innovation

Dans le cadre de l'*Open innovation*, l'entreprise collabore avec des partenaires externes sur un projet innovant. Cette collaboration permet à l'entreprise d'accroître sa R&D ou de promouvoir des inventions tout en se protégeant légalement grâce à des brevets. Dans le cas où l'entreprise valorise ses inventions, la collaboration ne se positionne plus dans un contexte d'*Outsourcing*. Comme pour le *crowdsourcing*, le processus d'innovation est externalisé par l'entreprise qui partage ses connaissances avec ses partenaires. Cependant, il y a un réel partenariat dans le cadre de l'*Open Innovation* qui implique un échange entre les acteurs, ce qui n'est pas le cas du *crowdsourcing*. En effet, lorsque l'employeur dépose une tâche sur une plateforme de *crowdsourcing*, il n'interagit pas par la suite avec la foule. De plus, l'*Open Innovation*, est employée uniquement dans le contexte d'innovation. Le *crowdsourcing*, lui, n'est pas exclusivement utilisé pour l'innovation, bien que cela soit une possibilité pour le *crowdsourcing* d'activités inventives défini en section 2.3.2.

User Innovation

Pour l'*Open Innovation*, c'est l'entreprise qui identifie un besoin utilisateur. Dans le cadre du *User Innovation*, c'est l'utilisateur, appelé *lead user* par Schenk et al. 2012, qui détermine ce besoin et est au centre du processus d'innovation. Le *User Innovation* n'est donc pas une démarche d'*Outsourcing* puisqu'il n'y a pas d'externalisation de la part de l'entreprise. C'est en effet l'utilisateur qui propose une solution novatrice à son problème. Cette approche est diamétralement opposée au *crowdsourcing* en raison de la nature de la relation entre la foule et l'entreprise. Pour le *User Innovation*, l'entreprise ne collabore qu'avec des utilisateurs qui font usage de leur produit ou service. En revanche, dans les plateformes de *crowdsourcing* la foule est diversifiée voire anonyme de sorte que l'employeur ne s'adresse pas nécessairement à des connaisseurs du domaine. De plus, le *crowdsourcing* ne se limite pas à des activités de recherche et offre un large panel de possibilités contrairement au *User Innovation* qui est exclusivement utilisé dans un contexte de R&D.

Open Source

Pour l'*Open Source*, l'entreprise et l'utilisateur collaborent au développement ouvert de logiciels. L'*Open Source* n'est pas toujours de l'*Outsourcing* car la collaboration entre l'utilisateur et l'entreprise ne résulte pas nécessairement de l'externalisation d'une activité par l'entreprise. Dans le cadre de l'*Open Source* l'entreprise renonce à ses droits de propriété sur le code logiciel de sorte que les utilisateurs puissent y accéder et proposer des évolutions. Pour le *crowdsourcing*, l'*Open Innovation* et le *User Innovation* l'entreprise reste propriétaire des contributions collectées.

Le *crowdsourcing* se définit donc comme l'externalisation, par un employeur, de tâches à des contributeurs aux profils variés sur des plateformes spécifiques au domaine. Les tâches externalisables peuvent porter sur une activité de recherche ou d'innovation, mais le *crowdsourcing* ne se limite pas à cela et offre bien d'autres activités possibles exposées dans la section suivante.

2.3 Les plates-formes

De nombreuses typologies de plateformes et de tâches existent. Par exemple, pour ce qui est de la structure de la plateforme, Chittilappilly et al. 2016 différencient dans

leurs travaux les plateformes autonomes des métaplateformes. Les premières ne requièrent pas d'autres plateformes et existent par elles-mêmes, comme *Amazon mechanical Turk*[1] (AMT) ou iStockPhoto[2]. Les secondes, quant à elles, nécessitent l'utilisation de plateformes autonomes pour leur fonctionnement, ce qui est notamment le cas de Quadrant of Euphoria. Cette plateforme, développée par Chen et al. 2010 dans le cadre d'un projet de recherche académique, fait appel en partie à la foule de AMT pour des retours utilisateurs.

Guittard et al. 2010 définissent la typologie d'une plateforme d'après l'attente de l'employeur. Les auteurs différencient ainsi le *crowdsourcing* intégratif, qui consiste en la mise en commun d'éléments complémentaires apportés par la foule, du *crowdsourcing* sélectif où l'employeur choisit une contribution parmi l'ensemble de celles renseignées par la foule. Cependant, hormis la différence dans la structuration de la plateforme, la diversité des tâches réalisables induit une diversité de plateformes existantes, si bien que différentes nomenclatures apparaissent dans la littérature. Guittard et al. 2010 et Schenk et al. 2012 définissent trois catégories : routinière, complexe et créative. Burger-Helmchen et al. 2011 distinguent les plateformes de *crowdsourcing* d'activités routinières, d'apport de contenu et d'activité inventives qui peuvent être assimilées aux tâches complexes de Guittard et al. 2010. Les sections suivantes présentent ces différentes activités.

2.3.1 Apport de contenu

Comme son nom l'indique, le *crowdsourcing* de contenu consiste à apporter de la matière sur un sujet. Il peut s'agir par exemple, de regrouper un ensemble d'informations comme sur la plateforme Wikipédia[3] mais également de constituer un corpus d'images sur iStockPhoto. Le *crowdsourcing* d'apport de contenu peut également être utilisé à des fins scientifiques comme c'est le cas pour Galaxyzoo[4] où les contributeurs sont invités à classifier bénévolement des galaxies. La foule sur ces plateformes est importante et hétérogène afin de contribuer à une diversité du contenu et de corréler les informations apportées. Les contributeurs participent le plus souvent à titre gracieux (Wikipédia) mais ils peuvent sur certaines plateformes recevoir un micropaiement (iStockPhoto).

1. https ://www.mturk.com (27/10/2021)
2. https ://www.istockphoto.com (27/10/2021)
3. https ://fr.wikipedia.org (27/10/2021)
4. https ://www.zooniverse.org/projects/zookeeper/galaxy-zoo/ (27/10/2021)

2.3.2 Activité inventive

Dans les plateformes de *crowdsourcing* d'activités inventives, l'objectif pour le contributeur est d'apporter une solution, une idée ou des connaissances sur des sujets de complexité variable. Le plus souvent, notamment pour les tâches nécessitant un haut niveau d'expertise dans le domaine, un prix élevé est accordé pour récompenser le travail effectué. Ces plateformes instaurent une stratégie de compétitivité, de sorte que seul le contributeur qui propose la réponse la plus pertinente à la problématique posée est rémunéré. La foule est de taille restreinte sur ces plateformes, car ce n'est pas la quantité de réponses qui intéresse l'employeur mais leur qualité. Une autre raison à cette foule peu nombreuse est que les personnes qui participent à l'activité inventive sont principalement expertes du domaine. Une plateforme représentative de ce secteur est Innocentive [5] où des employeurs, souvent de grands groupes industriels ou des entités gouvernementales comme la NASA, proposent des activités de recherche et développement. Par exemple, au cours de l'année 2021, une tâche a été proposée pour la formulation de nouveaux vaccins à ARN messager contre la covid-19 ne nécessitant qu'une seule injection. La personne qui résoudra ce problème recevra une récompense de 40 000$.

2.3.3 Activité routinière

Les plateformes d'activités routinières proposent des microtâches simples, qui ne requièrent pas de qualification particulière, et réalisables dans des délais assez courts. La foule sur ces plateformes est grande, avec des profils diversifiés. Contrairement aux plateformes d'activités inventives, sur les plateformes d'activités routinières les contributeurs sont tous rémunérés, mais assez faiblement puisque le travail proposé est très simple. Par exemple, pour une tâche d'activité routinière, une image est présentée au contributeur avec un intitulé et il est demandé au contributeur si l'intitulé est en adéquation ou non avec l'image. AMT est une plateforme d'activités routinières utilisée partout à travers le monde. À l'échelle française, l'entreprise Wirk [6] (anciennement Foule Factory) propose également ce type de service.

5. https ://www.innocentive.com/ (27/10/2021)
6. https ://www.wirk.io/ (27/10/2021)

2.3.4 Activité créative

D'après Guittard et al. 2010 et Schenk et al. 2012, le *crowdsourcing* d'activités créatives est ancien, car ses origines remontent avant même l'existence d'Internet lors d'événements comme des concours d'affiche artistiques. Dans ces plateformes, l'employeur est à la recherche d'une nouvelle création répondant à un cahier des charges qu'il a défini. Les contributeurs renseignent un ensemble de propositions répondant au besoin du financeur qui choisira celle qu'il estime être la plus pertinente. Sur les plateformes d'activités créatives, ce n'est pas la quantité de contributions qui compte, mais la créativité de celles-ci. En effet, l'objectif de l'employeur est d'obtenir une idée ou une réalisation originale. Par exemple, la plateforme *Crowdspring*[7] est dédiée à la création de logos, de noms d'entreprise ou de designs de nouveaux produits. La compagnie LG a ainsi fait appel aux contributeurs de *Crowdspring* pour définir le design de leurs futurs téléphones portables. La rémunération pour ce type d'activité est d'un niveau intermédiaire entre celle d'une activité et celle d'une activité inventive.

Après avoir présenté les principaux types de plateformes de *crowdsourcing* existants, la section suivante décrit les acteurs qui y sont associés, à savoir les employeurs et les contributeurs.

2.4 Les acteurs

L'employeur propose une activité sur une plateforme de *crowdsourcing* où une foule de contributeurs est présente et va réaliser le travail demandé. Cette section présente ces deux acteurs et leurs intérêts pour le *crowdsourcing*.

2.4.1 Les employeurs

L'employeur peut être un particulier qui fait appel à du *crowdsourcing* de contenu, mais de manière plus générale, les entités qui recourent au *crowdsourcing* sont des laboratoires de recherche ou des entreprises.

7. https ://www.crowdspring.com (27/10/2021)

La recherche

L'annotation ou la création de corpus de référence consiste le plus souvent en un travail simple mais fastidieux s'il est réalisé par une seule personne. Dans le cadre d'une campagne de *crowdsourcing*, il peut facilement être décomposé en sous-tâches et confié à une foule de contributeurs. C'est pourquoi, des équipes de recherche font appel à des plateformes d'activité routinières pour de l'annotation de données. Il est notamment démontré par Snow et al. 2008 que l'annotation de langage naturel par des contributeurs d'AMT est pertinente. Fort 2017 présente *Zombilingo*[8], un jeu ayant pour objectif de faire annoter la syntaxe de phrases par des joueurs. Ce jeu reproduit ainsi le principe du *crowdsourcing* où l'annotation syntaxique constitue une tâche complexe qui est décomposée en sous-tâches plus simples présentées à la foule par le biais d'une interface ludique. Les joueurs contribuent bénévolement à la tâche en s'amusant, ce qui facilite leur participation et accroît leur intérêt et leur volonté de bien faire. Avant de pouvoir jouer, les contributeurs doivent réaliser une phase d'entraînement sous la forme d'un tutoriel. En effet, l'objectif poursuivi ici par les auteurs est de former les contributeurs afin d'améliorer leur expertise pour l'annotation de syntaxe. Grâce à *Zombilingo*, les créateurs du jeu ont pu collecter un grand nombre d'annotations en un court laps de temps.

Dans un autre domaine, Chu et al. 2015 ont créé KATARA un système qui utilise une base de connaissances et le *crowdsourcing* pour l'annotation et la "réparation" de données. Mais l'utilisation du *crowdsourcing* pour la science ne se restreint pas à l'informatique. Ainsi, grâce aux données de *Foldit*, un jeu expérimental sur le repliement de protéines, Khatib et al. 2011 ont découvert la structure tridimensionnelle d'une protéase rétrovirale.

L'industrie

Comme le souligne Felstiner 2011, les entreprises utilisent le *crowdsourcing* car elles trouvent facilement et rapidement de la main d'œuvre sur les plateformes d'apport de contenu et d'activités routinières. De plus, faire appel à une foule de contributeurs plutôt qu'à une société du domaine est moins onéreux. Ainsi, Howe 2006 donne l'exemple d'une entreprise qui, au lieu d'embaucher un photographe professionnel, a fait le choix de recourir à iStockphoto pour des contraintes de coût et de temps.

Une entreprise peut aussi utiliser le *crowdsourcing* d'activités inventives pour répondre à une problématique de R&D, ou du *crowdsourcing* d'activités créatives pour le dévelop-

8. https ://zombilingo.org/ (26/10/2021)

pement de nouveaux produits. Il existe également des plateformes, qu'Innocent et al. 2017 qualifient d'"*ideagora*", où l'objectif n'est pas de chercher à résoudre un problème, mais d'aller au-devant des besoins des utilisateurs. La plateforme LEGO Ideas [9] a notamment été développée par la société LEGO dans ce but. Les contributeurs y proposent des créations inédites et si l'œuvre obtient le nombre minimal de votes requis et l'approbation de LEGO, elle est alors commercialisée.

Le *crowdsourcing* sous ses différents traits offre donc de multiples avantages pour l'entreprise qui peut faire appel à un grand nombre de contributeurs aux profils variés sans que cela ne nécessite d'efforts logistiques importants. En effet, il n'est pas nécessaire pour l'employeur de fournir d'installations à la foule ni d'engager des managers pour en assurer l'encadrement. L'employeur n'a pas besoin non plus d'effectuer de recrutement puisque les contributeurs vont d'eux-mêmes chercher les tâches.

Après avoir introduit les différents employeurs présents sur les plateformes de *crowdsourcing*, la section suivante expose la diversité des contributeurs.

2.4.2 Les contributeurs

D'après Guittard et al. 2010, la foule comporte un grand nombre d'individus hétérogènes anonymes. Nous nous appuyons ici sur l'étude démographique des contributeurs d'AMT réalisée par Ross et al. 2009. Pour effectuer leur analyse les auteurs ont utilisé la plateforme AMT sur laquelle 537 contributeurs ont répondu à leur questionnaire. La majorité des répondants (57%) sont des citoyens des États-Unis d'Amérique, 32% sont Indiens et le reste de la foule arbore des nationalités variées. D'après l'étude de Ross et al. 2009, le *crowdsourcing* n'a pas de genre puisque 55% des contributeurs ayant participé à la campagne sont des femmes. La population de la plateforme est relativement jeune avec 40% de la foule entre 18 et 24 ans, et si l'on ajoute à cet intervalle la tranche d'âge des 25-30 ans cela représente 62% des individus. En matière d'éducation, là encore les contributeurs ont des niveaux de diplôme très diversifiés allant du baccalauréat à un diplôme d'études supérieures. D'ailleurs, 33% sont des étudiants à temps plein ou à temps partiel, et le reste de la foule se répartit équitablement entre des travailleurs à temps plein, partiel ou au chômage lors de l'étude. D'après l'analyse de Ross et al. 2009, certains répondants étaient nouveaux sur la plateforme tandis que 3% y sont présents

9. https ://ideas.lego.com/ (26/10/2021)

depuis plus de deux ans. Le temps passé par les contributeurs sur les plateformes est très variable, pouvant aller de moins d'une heure par semaine à plus de 30 heures. Par exemple, la majorité de la foule (46%) passe entre 1 et 5 heures sur AMT. En terme de revenu, les contributeurs dont le revenu annuel est inférieur à 10 k$ constituent 27% de la foule. Pour 18% des répondants, la rémunération reçue pour leur contribution est souvent voire toujours une nécessité pour leur fin de mois. La rémunération constitue la principale raison de participation des contributeurs dans les plateformes de *crowdsourcing*, mais ce n'est toutefois pas la seule. Le contributeur peut aussi tout simplement aimer réaliser une activité qui l'intéresse sans qu'il n'y ait de rémunération comme pour *Zombilingo* et *Galaxizoo*. Les personnes qui constituent la foule apprécient notamment la flexibilité que leur accorde le *crowdsourcing*. Felstiner 2011 souligne la grande liberté offerte au contributeur dans la planification de son activité puisqu'il peut choisir : son lieu de travail, ses horaires et la tâche réalisé, ce que ne permet pas un travail ordinaire en entreprise.

La section suivante introduit les problématiques du *crowdsourcing* que rencontrent employeurs et contributeurs.

2.5 Problématiques du crowdsourcing

Les problématiques sont de trois sortes dans les plateformes de *crowdsourcing* : juridiques, sociales et techniques. Chacun de ces thèmes est abordé dans les sections suivantes.

2.5.1 Problématiques juridiques

La problématique juridique principale dans le *crowdsourcing* est l'absence de cadre légal pour le contributeur. En effet, il n'y a pas de réglementation qui définit les droits du contributeur, à commencer par la rémunération minimale pour son travail. D'après Felstiner 2011, la foule reçoit la plupart du temps de faibles salaires sans avantages en contrepartie et le contributeur n'a aucune sécurité de l'emploi. Ces constatations rejoignent celles de Silberman et al. 2010 qui vont plus loin et soulignent le risque pour le contributeur de ne pas être rémunéré pour son travail si ce dernier est discrédité, à tort, par l'employeur. Dans cette éventualité, non seulement le contributeur n'est pas payé, mais sa réputation peut être ternie injustement ce qui peut lui être préjudiciable pour de futures campagnes de *crowdsourcing*. Or, pour certains contributeurs, cette activité de

crowdsourcing constitue leur principale source de revenus.

Silberman et al. 2010 mettent également en garde contre certaines tâches qui constituent des escroqueries à l'encontre du contributeur dans le but d'implanter un virus informatique sur sa machine ou de voler des données personnelles. Par exemple, un employeur malhonnête demande de poursuivre une étude en cliquant sur un lien qui installera un logiciel malveillant sur l'ordinateur du contributeur.

Sur certaines plateformes, il est possible pour le contributeur de faire un retour sur l'employeur ou sur la tâche, mais ce n'est pas une généralité ni une obligation. Pour faire face à ces problèmes, des communautés de contributeurs se créent : Silberman et al. 2010 donnent l'exemple de Turkopticon [10] où les contributeurs peuvent faire un retour sur leurs expériences. Ils peuvent ainsi échanger sur les employeurs et mettre en garde contre d'éventuels abus.

2.5.2 Problématiques sociales

La motivation du contributeur dans sa réalisation de la tâche ainsi que son profil ont un impact sur la qualité des données collectées ce qui peut engendrer des problèmes pour l'employeur.

Motivation du contributeur

La motivation du contributeur sur les plateformes de *crowdsourcing* fait partie des challenges que Chittilappilly et al. 2016 considèrent comme les plus importants. Kaufmann et al. 2011 différencient les motivations intrinsèques et extrinsèques dans le cadre du *crowdsourcing*.

Les motivations intrinsèques définissent l'intérêt et le plaisir que le contributeur trouve dans la réalisation de la tâche sans l'attente d'une récompense externe. La formulation de la tâche, comme l'utilisation d'une interface ludique par exemple, peut influencer positivement la motivation intrinsèque du contributeur.

Les motivations extrinsèques caractérisent chez le contributeur une motivation externe. Il est démontré par Harris 2011 que la modulation de la motivation du contributeur par des variations de la gratification accordée pour le travail a un impact sur la qualité des contributions. La rémunération a donc un poids important sur la motivation extrinsèque du contributeur et doit être bien réfléchie par l'employeur. Cependant, Kazai et al. 2013

10. https ://turkopticon.net/ (28/10/2021)

concluent dans leur article qu'une rémunération élevée, bien qu'elle suscitera une motivation pour les contributeurs qualifiés attire également les contributeurs malveillants. D'après Gadiraju et al. 2015, ces contributeurs sont motivés uniquement par la rémunération, et non pas par l'envie d'accomplir consciencieusement la tâche. Ils renseignent des contributions de mauvaise qualité en répondant rapidement et aléatoirement, mais cela est une problématique relative au profil du contributeur.

Profil du contributeur

Comme le montre l'étude démographique de Ross et al. 2009 présentée dans la section 2.4.2, la foule est très diversifiée ce qui induit une diversité dans les profils des contributeurs qui la compose. Il n'existe pas de définition arrêtée du profil du contributeur dans les plateformes de *crowdsourcing*. Ce terme peut renvoyer aussi bien à des critères démographiques, qu'à l'expérience, la personnalité ou la motivation. Kazai et al. 2011 ont réalisé une première étude de l'impact de la personnalité du contributeur et du temps de réponse sur la qualité des contributions. Dans une seconde étude, Kazai et al. 2012 considèrent cette fois les éléments démographiques conjointement à la personnalité, toujours en comparaison de la justesse des réponses. Afin d'étudier la personnalité d'un contributeur, les auteurs utilisent les cinq traits de personnalité définis par Goldberg 1990 et Goldberg 1993 dans le modèle OCEAN, aussi appelé *Big Five* :

— l'Ouverture à l'expérience : ouverture d'esprit, curiosité, volonté de participer à des activités nouvelles.

— la Conscienciosité : conscience morale, respect des consignes, réflexion plutôt que précipitation.

— l'Extraversion : caractère optimiste, énergie positive, sociabilité.

— l'Agréabilité : compassion et coopération.

— le Névrosisme : sentiments négatifs (peur, colère ...), instabilité émotionnelle.

Dans le cadre du *crowdsourcing* l'ouverture à l'expérience caractérise la volonté du contributeur à participer à la tâche ; la conscienciosité est relative à son comportement lors de la réalisation de la tâche ; et l'extraversion peut avoir un impact sur sa contribution s'il est amené à travailler en équipe. D'après les études de Kazai et al. 2011 et Kazai et al. 2012, la localisation géographique du contributeur, son ouverture à l'expérience et sa conscienciosité ont un impact sur la justesse des réponses. Le profil du contributeur, aussi bien en terme démographique qu'en terme de personnalité, est donc bien révélateur de la

qualité des données qu'il fournit. C'est pourquoi il est important pour l'employeur d'avoir une connaissance du profil du contributeur pour traiter au mieux ses réponses, alors même que l'anonymisation des individus qui composent la foule constitue un obstacle. De plus, Penna et al. 2012 soulignent la difficulté de faire la distinction entre les contributeurs fiables et ceux qui ne le sont pas en l'absence de données d'or. Les données d'or correspondent à des questions pour lesquelles les réponses sont déjà connues par l'employeur. Elles permettent ainsi d'avoir une estimation de la qualification du contributeur et de son sérieux. Malheureusement, toutes les campagnes de *crowdsourcing* ne permettent pas de disposer de données d'or.

2.5.3 Problématiques techniques

Les problèmes techniques rencontrés sur les plateformes sont liés à la définition de la tâche, à son assignation et à l'agrégation des réponses.

Définition de la tâche

La définition de la tâche comprend sa segmentation (si elle est complexe), sa formulation et l'ergonomie de l'interface utilisée. Cette première étape est essentielle. Ainsi, comme l'indique Lease 2011, bien que la tâche paraisse évidente pour son créateur, si de mauvais résultats sont obtenus cela signifie potentiellement que la conception n'est pas idéale. Si les consignes sont mal exprimées (trop complexes, ambiguës ...), il y a un risque élevé qu'elles ne soient pas correctement interprétées par le contributeur qui fournira alors des réponses erronées. Ceci rejoint les propos de Kittur et al. 2013 qui affirment que la qualité des contributions est également relative au design de la tâche et pas uniquement à l'expertise du contributeur, bien que cet élément ait aussi un impact. Des instructions peu claires peuvent par conséquent être à l'origine de réponses médiocres. Dans leur article Kittur et al. 2013 citent un contributeur qui rapporte que trop souvent les tâches sont mal conçues et il existe un certain degré d'incompréhension entre le travailleur et le concepteur de la tâche. Ces éléments sont en accord avec les études de Rahmanian et al. 2014 et H. Zhang et al. 2012 qui montrent que le design de la tâche a un impact sur la qualité des données collectées. Il est donc essentiel pour l'employeur de considérer avec intérêt la définition de la tâche, mais il n'existe pas à l'heure actuelle de protocole officiel pour cela.

Assignation

Ce sont les contributeurs sur les plateformes de *crowdsourcing* qui choisissent les tâches qu'ils souhaitent effectuer. Or certaines tâches requièrent parfois des compétences spécifiques, et ne peuvent par conséquent pas être réalisées par tout un chacun. C'est pourquoi sur certaines plateformes, l'employeur peut exiger que le contributeur réussisse un test de qualification avant de pouvoir participer à la tâche proposée. Cependant, il reste possible qu'un contributeur malveillant réussisse le test de qualification et renseigne par la suite des contributions erronées. D'après Mavridis et al. 2016, la plateforme devrait idéalement assigner ou proposer la tâche au contributeur le plus qualifié du domaine en priorité. Chittilappilly et al. 2016 considèrent le recrutement du contributeur comme une problématique importante du domaine, au même titre que le contrôle de la qualité des réponses et leur agrégation.

Agrégation des réponses

Plusieurs éléments viennent influer sur la qualité des données provenant de campagnes de *crowdsourcing*. C'est pour cette raison qu'il est nécessaire pour l'employeur d'appliquer une méthode robuste de combinaison des réponses face aux éventuelles imperfections des données. La méthode traditionnellement utilisée est le vote majoritaire, *Majority Voting* (MV) en anglais, qui consiste à sélectionner la réponse renseignée par le plus grand nombre de contributeurs. Cependant, cette méthode ne permet pas de modéliser la potentielle incertitude du contributeur sur sa réponse. Or, imaginons qu'un contributeur donne une réponse A dont il est absolument certain, et que la majorité du reste de la foule renseigne la réponse B sans certitude. Le MV s'accorde sur la réponse B alors même que l'employeur aurait raison d'en douter. La méthode de vote ne tient pas non plus compte du profil du contributeur. Or, si un expert est opposé à un groupe important de néophytes il serait plus raisonnable de donner davantage de crédit à l'expert.

Cette section résume les problématiques juridiques, sociales et techniques rencontrées sur les plateformes de *crowdsourcing*. L'objectif de ce travail n'étant pas de traiter des problématiques associées aux droits des contributeurs, elles ne seront pas développées par la suite. La section suivante présente les approches existantes qui ont été réfléchies pour répondre aux problématiques sociales et techniques.

2.6 L'existant

Cette section présente l'état de l'art des approches répondant aux problématiques sociales et techniques présentées dans les sections 2.5.2 et 2.5.3. D'après Chittilappilly et al. 2016, une campagne de *crowdsourcing* se décompose en trois phases : l'initialisation, l'implantation et la finalisation. L'initialisation de la campagne inclut toutes les étapes requises avant sa mise en ligne sur une plateforme. L'implantation est le déroulement de la campagne et la finalisation consiste au traitement des données collectées. Les approches existantes pour chacune de ces phases sont présentées dans les sections suivantes.

2.6.1 Initialisation de la campagne

Chittilappilly et al. 2016 incluent dans la phase d'initialisation l'ensemble des étapes requises avant que la campagne de *crowdsourcing* ne soit mise à la disposition de la foule. En effet, lorsque l'employeur décide de réaliser une campagne afin de subvenir à un besoin, il doit prendre un temps de réflexion pour sa création. Cette étape a un impact sur la qualité des données collectées comme il est expliqué dans la section 2.5.3. Ceci rejoint les propos d'Allahbakhsh et al. 2013 qui affirment que le contrôle de la qualité dans les plateformes de *crowdsourcing* dépend du profil du contributeur (sa réputation et son expertise), mais aussi du design de la tâche. D'après les auteurs, le design de la tâche se caractérise par la définition du travail que donne l'employeur aux contributeurs, l'interface employée, la granularité de la tâche (sa décomposition en sous-tâches) et la gratification.

Les paragraphes suivants exposent les bonnes pratiques que devraient suivre les employeurs lors de la création de campagnes. Les éléments de la littérature relatifs à la conception de campagnes de *crowdsourcing*, la définition d'interfaces et les méthodes employées pour motiver le contributeur dans sa contribution sont aussi développés.

Bonnes pratiques

Marcus et al. 2015 renseignent un ensemble d'éléments, complémentaires aux informations données par Chittilappilly et al. 2016, que l'employeur doit garder à l'esprit lors de la phase de conception. Les auteurs expliquent ainsi que l'employeur doit prêter une attention particulière à la décomposition de tâches en sous-tâches et préférer pour l'interface les questions fermées, car il est plus facile pour le contributeur d'y répondre. L'employeur prendra également soin de détailler les instructions et d'ajouter des exemples

pertinents afin de faciliter la compréhension de la foule. D'après cet article, la plupart des contributeurs apprécient d'avoir des instructions détaillées de façon progressive pour réaliser la tâche correctement. Une fois que la tâche est définie et l'interface conçue, il est important de faire tester la campagne par une personne extérieure qui n'est pas familière du domaine, de façon à avoir ses retours. Le coût de la campagne est également un élément à prendre en compte. L'employeur a le devoir moral de payer convenablement les contributeurs. Il est notamment recommandé par Marcus et al. 2015 d'offrir une bonification aux contributeurs qui ont le mieux réussi la tâche.

Afin d'aider l'employeur dans la création de la tâche, certaines plateformes comme AMT [11] et Appen [12] fournissent un ensemble de conseils et d'outils. Les plateformes insistent elles aussi sur la nécessité de transmettre aux contributeurs des instructions claires. AMT permet à l'employeur de donner autant de détails que nécessaire dans la définition de la tâche pour une meilleure compréhension du contributeur. Appen rappelle que la plupart des erreurs des contributeurs peuvent être évitées en suivant quelques règles de bonnes pratiques. Il s'agit pour l'employeur de ne pas demander l'impossible au contributeur. Il ne faut, par exemple, pas demander des informations sur un site qui requiert une inscription préalable. Il est également de bon usage de faire un retour aux contributeurs sur la validité de leur travail en temps réel, grâce à l'utilisation entre autres de données d'or. Le design de l'interface est aussi un élément crucial : plus la tâche est complexe, plus l'interface devra permettre de simplifier le travail du contributeur.

Définition de la tâche

La définition de la tâche est composée des informations données à l'employeur avant la campagne. Il s'agit d'une courte description du travail à réaliser, de la durée qu'il demande, de la paie octroyée et éventuellement de la qualification requise nécessitant alors une sélection de la foule. D'après Gadiraju et al. 2014, les contributeurs choisissent leur tâche en premier lieu d'après la rémunération, puis le temps requis et finalement le sujet. C'est pourquoi ces éléments doivent être mûrement réfléchis. De plus, Difallah et al. 2012 indiquent qu'une formulation sophistiquée des tâches constitue une bonne barrière contre les contributeurs malveillants. Cependant, une formulation sophistiquée nécessite une charge de travail plus importante de l'employeur dans la définition de la tâche qui doit être plus recherchée.

11. *Create an Amazon Mechanical Turk project* (09/09/2021)
12. *How to Build a Successful Task for the Crowd* (09/09/2021)

Conception de l'interface

L'interface utilisateur doit être facile d'utilisation afin d'attirer les contributeurs et leur donner envie de participer. C'est un élément clef d'une campagne de *crowdsourcing* réussie. D'après Finnerty et al. 2013, la qualité des données obtenues est relative au design de la tâche et non uniquement aux contributeurs. Une interface simple offre de bien meilleurs résultats qu'une interface complexe. De plus, selon Yang et al. 2016, un meilleur design de la tâche et des composants plus interactifs permettent de diminuer la complexité perçue par le contributeur. Bien que différents exemples d'interfaces existent, il n'y a pas encore de modèle généralisé pour le *crowdsourcing*. En effet, la diversité des tâches compromet la possibilité d'un unique modèle d'interface. De plus, de légers changements d'une interface peuvent avoir des répercutions importantes sur les données collectées. Roitero et al. 2018 considèrent une tâche où une assertion est présentée au contributeur pour laquelle il doit donner le degré de confiance qu'il accorde à cette assertion. Les auteurs s'interrogent dans l'article sur l'échelle de confiance à employer. Bien qu'ils ne soient pas parvenus à définir la meilleure échelle à utiliser, ils ont constaté que l'utilisation de différentes échelles de confiance a un impact sur les données collectées notamment sur l'accord entre les contributeurs répondant à une même question.

Jurgens 2013 confrontent trois interfaces pour l'annotation de corpus. L'objectif de la tâche reste le même : déterminer le sens d'un mot. Les trois interfaces étudiées sont les suivantes : *Likert Rating*, *Select and rate* et *MaxDiff*. Pour la *Likert rating*, tous les sens du mot sont proposés au contributeur qui doit les noter. Une échelle de Likert est un ensemble de propositions permettant d'établir un degré d'accord de la part du répondant. Ces échelles sont souvent composées de 5 ou 7 choix de réponse avec une forme de symétrie, de sorte que si le premier élément est "Absolument d'accord" le dernier sera "Absolument pas d'accord". Pour un nombre de réponses impaires, l'élément central est associé à l'absence d'avis comme "Ne sait pas". Dans le cas de l'interface *Select and rate*, la tâche est décomposée en deux sous-tâches. Pour commencer, tous les sens d'un mot sont proposés au contributeur qui doit choisir de manière binaire si ce sens est approprié. Ensuite, la tâche consiste à noter les sens qui ont été les plus fréquemment choisis à la manière de l'interface *Likert rating*. Finalement pour la troisième interface, *MaxDiff*, les contributeurs doivent choisir le sens qui est le plus approprié pour le mot proposé et celui qui l'est le moins. Au vu des expérimentations menées par Jurgens 2013, le mot pour lequel le sens doit être défini joue un rôle important, car il est représentatif de la complexité de la question. De même, le choix de la méthodologie d'annotation relative à l'interface

employée est important pour la qualité des résultats. Une comparaison sur le temps de réponse pour les taches *Likert Rating* et *Select and rate* montre que ces deux interfaces ont approximativement les mêmes besoins cognitifs. Mais c'est l'interface *MaxDiff* qui offre le meilleur accord inter annotateur.

Rahmanian et al. 2014 comparent trois interfaces pour la classification d'images. La première interface utilise le classement d'images, la deuxième est un tri direct qui ordonne les images par l'action de "glisser-déposer" et la troisième demande au contributeur de noter les images. D'après les auteurs, la notation est l'approche la plus performante, car le classement nécessite de comparer chaque image les unes par rapport aux autres et demande du temps, de même que le tri direct qui est distrayant. En effet, des fonctions distrayantes inutiles et un nombre élevé de clics augmentent la charge cognitive de la tâche et ont tendance à donner de mauvais résultats de la part des utilisateurs. Par exemple, il est plus complexe pour un contributeur de noter une image en la comparant à d'autres que de noter cette image de façon isolée.

Comme le montre les exemples précédents, des questionnaires fermés sont le plus souvent employés dans les plateformes de *crowdsourcing*. Il s'agit généralement de QCMs. Il n'existe pas à notre connaissance d'étude ciblée sur l'utilisation des QCMs pour les interfaces de *crowdsourcing*. Cependant, les travaux de Diaz et al. 2008 portent sur l'intérêt des QCMs dans le milieu scolaire. Les auteurs soulignent que les QCMs à choix unique restreignent l'étudiant à une réponse précise. Cependant, si ce dernier n'est pas convaincu par son choix, en cas de doute, l'étudiant doit faire un choix aléatoire entre plusieurs réponses qui ne sont pas nécessairement totalement fausses. Il manque d'après Diaz et al. 2008, dans les QCMs à choix unique, la possibilité pour l'étudiant d'exprimer son ignorance, son imprécision et son incertitude. Les auteurs proposent Ev-MCQ, un QCM permettant aux étudiants d'exprimer une réponse plus flexible. Grâce à Ev-MCQ, les réponses imparfaites facilitent l'identification des questions problématiques pour les étudiants. Dans ce papier, les QCMs ne peuvent avoir qu'une bonne réponse, car il s'agit d'une évaluation universitaire. Or, dans les tâches de *crowdsourcing*, il n'y a pas nécessairement une unique bonne réponse attendue. Le contributeur dans les plateformes de *crowdsourcing* peut éprouver cette même hésitation dans sa réponse que l'étudiant et se sentir restreint dans ses choix face à un QCM à choix unique. Il serait intéressant d'offrir au contributeur davantage de possibilités d'expression par le biais d'une interface permettant de renseigner ses hésitations comme l'a fait Diaz et al. 2008 en milieu scolaire.

Motivation

Comme il est expliqué dans la section 2.5.2, la source principale de motivation du contributeur dans les plateformes de *crowdsourcing* est la rémunération offerte pour son travail qui constitue une motivation extrinsèque. Cependant, comme le souligne Mason et al. 2009 augmenter la rémunération augmente le nombre de travailleurs, mais pas nécessairement la qualité des données collectées. Les auteurs constatent également que l'accroissement de la difficulté des tâches s'accompagne de la diminution du nombre de tâches réalisées par les contributeurs. Rogstadius et al. 2011, en accord avec Mason et al. 2009, observent que rehausser la gratification offerte pour la tâche n'améliore pas la qualité des contributions. D'après les auteurs, la présence d'une motivation intrinsèque améliore la justesse des résultats. Pour ce faire, la motivation intrinsèque doit être plus importante que la motivation extrinsèque. Il est également constaté que l'augmentation de la difficulté de la tâche s'accompagne de l'augmentation du temps nécessaire à sa réalisation, jusqu'à un certain point. Passé ce point de rupture, accroître la difficulté provoque une diminution du temps.

Le problème de la motivation du contributeur est plus fort encore dans les plateformes de *crowdsourcing* de contenu où la gratification est faible voire absente. Dans ce contexte, Kamar et al. 2016 cherchent à maintenir la motivation intrinsèque de l'utilisateur afin que celui-ci continue d'apporter sa contribution sur la plateforme. Pour ce faire, des messages sont envoyés au contributeur en temps réel avant qu'il ne quitte la plateforme, afin de l'inciter à poursuivre sa participation. Les auteurs montrent que ces messages encourageants accroissent significativement le nombre de contributions faites par un utilisateur sans que la qualité de son travail n'en soit diminuée. Toujours dans l'optique de motiver le contributeur en interagissant avec lui, Gadiraju et al. 2014 utilisent des questions distrayantes au cours de la campagne afin de préserver l'attention du contributeur, et plus tard d'identifier les mauvais contributeurs.

Finalement, Fort 2017 exposent un type particulier d'interface de *crowdsourcing* : "le jeu ayant un but". Ce type d'interface est intéressant, car elle permet d'influer sur la motivation intrinsèque du contributeur grâce à l'aspect ludique de la tâche. Des tâches laborieuses, voire complexes, peuvent ainsi être décomposées et proposées sous forme de jeu au contributeur, c'est le cas de *Zombilingo* présenté dans la section 2.4.1.

Après avoir énoncé les approches existantes pour l'initialisation de la campagne de *crowdsourcing* nous abordons dans la section suivante son déroulement sur la plateforme.

2.6.2 Déroulement de la campagne

Les campagnes se déroulent de différentes façons en fonction de leur conception par l'employeur. Pour certaines tâches, la foule doit réaliser une phase d'entraînement afin que le contributeur se forme pour le travail à effectuer avant d'accomplir la tâche. Pour d'autres campagnes, il n'y a pas de phase d'entraînement, mais un niveau d'expertise est tout de même attendu de la part du contributeur. Dans ce cas, deux approchent existent. La première consiste à faire passer un test de qualification à la foule de sorte que seuls les contributeurs qui ont réussi le test peuvent y participer. Dans le cas de la seconde, la tâche est assignée au contributeur, ce n'est plus lui qui choisit le travail qu'il souhaite réaliser. L'utilisation d'une phase d'entraînement, d'un test d'expertise ou de la sélection du contributeur sont trois moyens utilisés afin d'améliorer la qualité des données collectées et sont explicités dans les paragraphes suivants. Une dernière possibilité peut être envisagée pour accroître la qualité des contributions au cours de la campagne de *crowdsourcing*, il s'agit d'interagir avec le contributeur lorsqu'il réalise la tâche. Cette possibilité est également présentée dans cette section.

Phase d'entraînement

La phase d'entraînement consiste à proposer aux contributeurs le même travail que celui qu'il devra effectuer au cours de la campagne. Lors de l'entraînement les bonnes réponses sont connues de l'employeur ce qui permet de valider ou corriger le contributeur. Le et al. 2010 montrent par l'expérience que, grâce à la réalisation d'une phase d'entraînement, le contributeur est plus à même de réaliser le travail demandé par l'employeur que s'il n'est pas entraîné. C'est dans l'objectif d'observer les effets de l'apprentissage que Matsubara et al. 2018 présentent au contributeur une réponse de référence durant une phase de correction, après qu'il ait accompli la tâche. Différentes réponses de référence sont testées dans cet article :

— la réponse correcte

— une réponse aléatoire

— une réponse obtenue par apprentissage automatique sur les réponses correctes

— une réponse obtenue par apprentissage automatique sur les réponses d'un humain

D'après les résultats de l'article, les réponses proposées par l'apprentissage automatique humain sont plus propices pour le contributeur à l'acquisition des connaissances requises pour la tâche.

Test de qualification

Pour certaines campagnes, seuls les contributeurs démontrant les compétences appropriées pour la tâche sont autorisés à participer. Afin d'accorder le droit aux contributeurs de réaliser la campagne, plusieurs approches existent. Dans certains cas, la foule doit effectuer un test de qualification spécifique à la campagne. Dans d'autres cas, c'est la plateforme de *crowdsourcing* qui propose aux contributeurs de passer un test pour monter en compétence. Cette certification permet au contributeur d'avoir accès à des tâches plus complexes ou qui demandent davantage de sérieux. Dans les travaux de Y. Zhang et al. 2012, un contributeur peut participer à une tâche s'il a les compétences requises et s'il n'a pas mauvaise réputation. Pour définir la réputation d'un contributeur, le modèle des auteurs nécessite un administrateur qui vérifie la résolution des tâches effectuées par le contributeur. La réputation de ce dernier croît s'il a exécuté convenablement la tâche, dans le cas contraire elle décroît. Dans cette même optique, Folorunso et al. 2015 accordent à chaque contributeur un degré de qualification et un degré de confiance qui sont utilisés pour calculer un niveau de priorité. Si le niveau de priorité du contributeur est supérieur à la valeur seuil de la tâche, il est autorisé à réaliser celle-ci.

L'étude de Gadiraju et al. 2017 est également intéressante, car plutôt qu'exploiter la réputation du contributeur, les auteurs considèrent sa capacité d'auto-évaluation. Cet intérêt pour l'auto-évaluation est dû à la constatation par les auteurs de l'impact de l'effet Dunning-Kruger sur les données de *crowdsourcing*. L'effet Dunning-Kruger est un biais cognitif selon lequel les personnes les moins qualifiées dans un domaine surestiment leurs compétences. À l'inverse, les personnes les plus qualifiées auraient tendance à sous-estimer leur niveau de compétence. Gadiraju et al. 2017 montrent que les contributeurs les moins compétents surestiment leurs aptitudes en accord avec l'effet Dunning-Kruger. En revanche, au vu de l'étude des auteurs, les contributeurs compétents ne se sous-estiment pas. Il est ainsi préconisé dans l'article de réaliser une présélection des contributeurs en fonction de leur compétence et leur capacité d'auto-évaluation. Une comparaison est faite entre deux campagnes de *crowdsourcing* avec, pour l'une, une présélection des contributeurs par leur compétence et, pour l'autre, une présélection d'après la compétence et la capacité d'auto-évaluation. Les contributeurs choisis au vu de leur compétence et auto-évaluation ont de meilleures performances en terme de précision des résultats comparés à ceux retenus pour leur seule compétence.

Assignation de la tâche

Des études récentes portent sur l'assignation de la tâche au contributeur. Contrairement au *crowdsourcing* traditionnel où c'est le contributeur qui postule pour une tâche, cette fois c'est l'employeur qui cherche à recruter le contributeur qui répondra au mieux à ses besoins. Tong et al. 2014 ont élaboré le système *Crowdcleaner* qui assigne une tâche aux contributeurs les plus pertinents grâce à une approche probabiliste. Boim et al. 2012 emploient également un système utilisant les probabilités, il s'agit de *AskIt !*. Ce dernier sélectionne la question à poser et le contributeur à qui elle est adressée afin de réduire l'incertitude sur les réponses. Roy et al. 2013 ont, quant à eux, défini le système *SmartCrowd* qui fait un apprentissage des compétences du contributeur lorsque celui-ci réalise des tâches. Grâce à cet apprentissage, *SmartCrowd* affecte ensuite de nouvelles tâches en fonction des compétences des contributeurs. Afin d'assigner des tâches aux contributeurs, Mavridis et al. 2016 utilisent une taxonomie des compétences des utilisateurs et des tâches. Les auteurs commencent par sélectionner le contributeur ayant la compétence exacte pour la tâche, et s'ils ne le trouvent pas, élargissent leur possibilité de sélection à un contributeur avec une compétence plus large. Dans le cas où aucun contributeur n'est trouvé, c'est la compétence requise pour la tâche qui est élargie.

Interaction avec le contributeur

L'interaction avec le contributeur peut l'amener à reconsidérer son travail, comme dans l'étude de Dow et al. 2012. Les auteurs s'intéressent à la possibilité pour le contributeur de reprendre son travail après une auto-correction ou un retour extérieur. Ils comparent trois scénarios de *crowdsourcing* : pas de correction de la réponse, une auto-correction et une correction externe par le retour d'un expert. Notons que les auteurs mentionnent la possibilité d'un retour par les pairs pour la correction externe. Cette approche induit d'identifier les contributeurs les plus performants afin de leur proposer ce genre de travail, ce qui n'est pas étudié dans l'article, car il s'agit d'une autre problématique du domaine. Pour les campagnes de *crowdsourcing*, des questions ouvertes sont posées aux contributeurs afin que le retour des experts puisse améliorer la qualité de la réponse. Les résultats obtenus montrent que l'auto-correction et la correction externe permettent toutes deux de collecter des réponses de meilleure qualité.

Drapeau et al. 2016 demandent également une auto-correction au contributeur après une interaction avec lui. Les auteurs ont créé le système *MicroTalk* qui peut être décom-

posé en trois étapes : *Assess, Justify, Reconsider*. *MicroTalk* entraîne les contributeurs sur des données d'or avant la réalisation de tâches. Lors de la réalisation de la tâche, le contributeur répond à une question (*Assess*). Ensuite il renseigne son raisonnement en complément de sa réponse (*Justify*). Puis un argument opposé à la réponse est présenté au contributeur, il lui est alors demandé de confirmer sa réponse ou d'en choisir une nouvelle (*Reconsider*). *MicroTalk* est comparé à un système traditionnel de *crowdsourcing* qui n'exploite que *Assess*. *MicroTalk* est plus performant que ce dernier, car il offre de meilleurs résultats en terme d'exactitude des réponses.

Une fois la campagne terminée, il n'est plus possible d'améliorer la qualité des données, cependant un traitement approprié de l'information permet d'en tirer le meilleur parti. La section suivante expose ces traitements post-campagne.

2.6.3 Finalisation : contrôle de qualité

Marcus et al. 2015 incluent dans leur ensemble de bonnes pratiques de faire réaliser la tâche par plusieurs contributeurs et d'agréger les données. Cependant, la foule est très variée sur les plateformes de *crowdsourcing* ce qui génère des contributions de qualités inégales. La finalisation de la campagne a pour objectif de déterminer la pertinence des réponses pour optimiser leur agrégation. Pour le contrôle de la qualité, deux éléments entrent en compte : le profil du contributeur et la méthode d'agrégation employée. Certaines méthodes d'agrégation tiennent compte du profil du contributeur et le profil est parfois estimé grâce à l'agrégation des réponses c'est pourquoi nous différencions trois approches. La première permet uniquement l'agrégation des réponses, la seconde exclusivement l'estimation du profil et la troisième combine l'estimation du profil à l'agrégation des réponses.

Agrégation des réponses

La méthode la plus couramment employée dans les plateformes de *crowdsourcing* pour l'agrégation des réponses est le MV. Elle est par exemple utilisée par Nguyen 2015. Le MV repose sur l'hypothèse que la majorité de la foule a raison et consiste à sélectionner la réponse qui a été choisie par le plus grand nombre de contributeurs. Cette méthode présente l'avantage d'être simple à implanter, car la réponse X d'un contributeur c à la

question q est modélisée par une fonction indicatrice :

$$\begin{cases} r_{cq}(X) = 1, X \subset \Omega \text{ si le contributeur choisit la réponse } X \\ r_{cq}(Y) = 0, Y \in \Omega \backslash X, \text{ sinon} \end{cases} \tag{2.1}$$

Dans l'équation (2.1), Ω est l'ensemble des réponses proposées au contributeur pour la question q. Si deux réponses ou plus comptabilisent un nombre maximal de votes égaux, l'employeur est contraint de faire appel à des experts pour lever l'indécision. Une autre possibilité est de choisir aléatoirement une réponse parmi celles ayant reçu le plus de voix.

La limite de cette approche réside dans le fait que toutes les contributions ont un poids identique lors de l'agrégation. Le MV n'est par conséquent pas robuste face aux contributeurs au comportement malveillant ou manquant de qualifications. En effet, une réponse donnée aléatoirement a une importance égale à une réponse sérieuse, ce qui peut être dommageable pour la qualité des données après leur agrégation. De plus, en admettant que la foule soit constituée uniquement de contributeurs consciencieux et qu'aucun contributeur n'a un comportement négatif, cette méthode reste imparfaite, car tous les contributeurs n'ont pas le même niveau de compétences. L'employeur devrait accorder davantage de crédit à la contribution d'un expert du domaine plutôt qu'à un novice ce qui n'est pas le cas avec le MV.

Lee et al. 2010 résolvent le problème des contributeurs malveillants et améliorent les résultats obtenus par MV en réalisant une présélection des réponses. Ainsi, la méthode *Honeypot*, définie par les auteurs, se fonde sur le même principe d'agrégation que le MV, mais les réponses des contributeurs qui ne sont pas considérés dignes de confiance ne sont pas prises en compte. Ces contributeurs sont identifiés grâce à un test de qualification utilisant des données d'or.

Khattak et al. 2011 vont encore plus loin et proposent ELICE [13], un système qui considère la précision des contributeurs et la difficulté de la tâche pour une meilleure agrégation des réponses. Des données d'or sont utilisées pour une partie des réponses à classifier, les classes des données restantes sont approchées. Hung et al. 2013 comparent le MV avec les méthodes *Honeypot* et ELICE : les résultats expérimentaux montrent qu'une meilleure précision des résultats est obtenue avec ELICE. Ceci est en accord avec les travaux de Whitehill et al. 2009 qui ont fait des tests avec ELICE et le MV et observent de meilleurs résultats avec ELICE qu'avec le MV. Faire un pré-traitement des réponses avant l'agrégation par MV est donc concluant. Malheureusement, l'estimation

13. Expert Label Injected Crowd Estimation

de la fiabilité du contributeur requiert l'utilisation de données d'or pour *Honeypot* et ELICE, et ces données ne sont pas toujours accessibles à l'employeur. Des données d'or sont souvent utilisées pour déterminer le profil ou la fiabilité d'un contributeur, car il est extrêmement difficile de le faire en leur absence. C'est ce que montrent Penna et al. 2012 : les auteurs concluent qu'une faible quantité de données d'or est suffisante, mais nécessaire pour traiter convenablement les contributions.

Une autre approche possible pour l'agrégation des réponses en l'absence de connaissance du profil du contributeur consiste à utiliser la théorie des ensembles flous. La théorie des ensembles flous a été introduite par Zadeh 1996, elle permet de modéliser l'imprécision d'assertions. Usuellement, les mesures floues sont estimées grâce à des vérités terrain comme des données d'or ou de l'apprentissage automatique. Wagner et al. 2012 emploient cette théorie pour calculer la précision des contributeurs et l'accord au sein de la foule. D'après les auteurs, plus le nombre de contributeurs est important, moins ceux-ci sont d'accord entre eux, mais les points sur lesquels ils s'accordent sont essentiels.

Estimation du profil

L'estimation du profil se fait souvent par la classification des contributeurs en différents groupes d'après leurs aptitudes pour la tâche. Encore une fois, différentes approches existent pour atteindre cet objectif. Par exemple, plutôt qu'utiliser les ensembles flous pour l'agrégation des réponses, Folorunso et al. 2015 emploient cette théorie pour estimer le profil du contributeur : très qualifié, qualifié, semi-qualifié ou non-qualifié. Cette qualification permet ensuite au contributeur d'accéder à certaines tâches.

Blanco 2012 recourt à des machines à vecteurs supports (SVM) pour différencier les contributeurs malveillants du reste de la foule. Les SVMs sont des classifieurs qui reposent sur la recherche d'un hyperplan de marge optimale. Cet hyperplan va séparer les données tout en maximisant la distance entre lui et les observations. Grâce à l'utilisation de SVMs, Blanco 2012 identifient les *spammers* au sein de la foule. Un *spammer* est un contributeur malveillant qui répond aléatoirement afin d'effectuer la tâche plus rapidement dans le but de participer à un nombre maximal de campagnes de *crowdsourcing* pour augmenter ses gains en peu de temps. Ces contributeurs ne sont absolument pas consciencieux dans la réalisation de la tâche, c'est pourquoi il est essentiel de les identifier pour traiter leurs réponses en conséquence et ne pas les rémunérer voire les signaler à la plateforme. Afin d'identifier les *spammers*, Blanco 2012 estime le profil du contributeur d'après le nombre de tâches qu'il a accomplies, son temps moyen de réponse et la moyenne des réponses cor-

rectes attendue aux données d'or. L'auteur compare les résultats de classification obtenus par les SVMs avec ceux obtenus grâce à des arbres de décision. Le pourcentage de précision dans l'identification des *spammers* est plus élevé pour les SVMs que pour les arbres de décision. L'intérêt de cette approche réside dans son utilisation conjointe de la qualification du contributeur (la moyenne des bonnes réponses) et son comportement (temps moyen de réponse) pour estimer s'il s'agit ou non d'un *spammer*. Cependant, des données d'or sont encore nécessaires et la caractérisation du profil est binaire : soit le contributeur est un *spammer*, soit il ne l'est pas. Or il serait intéressant pour les contributeurs qui ne sont pas malveillants de différencier leur capacité. Cela permettrait d'accorder plus de poids aux réponses des contributeurs experts du domaine voire éventuellement leur demander dans un système plus complexe d'aider les contributeurs non experts.

Estimation du profil et agrégation des réponses

L'algorithme *Expectation Maximisation* (EM) est la méthode de l'état de l'art la plus couramment employée pour l'estimation conjointe du profil du contributeur et l'agrégation des réponses. Cet algorithme proposé par Dempster et al. 1977 permet l'estimation de données manquantes. Il est itératif et composé de deux phases : *"Expectation"* qui estime les données inconnues grâce aux paramètres courants, et *"Maximisation"* qui calcule les nouveaux paramètres d'après les données courantes. Toutes deux sont répétées jusqu'à la convergence de l'algorithme. Dawid et al. 1979 appliquent EM dans un cadre similaire à celui du *crowdsourcing*, l'algorithme 1 est donné en annexe 1 page 159. Dans cet algorithme, K individus répondent à un total de I questions pour lesquelles J réponses sont proposées. Pour chaque individu $k \in K$ un taux d'erreur individuel $\pi_{jl}^{(k)}$ est estimé d'après l'équation (2.3). Le taux d'erreur individuel $\pi_{jl}^{(k)}$ représente la probabilité que l'individu k renseigne la réponse $l \in J$ alors que la bonne réponse est $j \in J$. Par conséquent, $\pi_{jj}^{(k)}$ est la probabilité que l'individu k choisisse la bonne réponse. Les taux d'erreur sont des probabilité conditionnelles de sorte que pour chaque bonne réponse j et individus k :

$$\sum_{l=1}^{J} \pi_{jl}^{(k)} = 1 \tag{2.2}$$

Le nombre de fois où un individu k donne la réponse l à la question i est noté $n_{il}^{(k)}$. Et l'indicatrice de la véracité de la réponse j à la question i est notée T_{ij}. Grâce aux valeurs

de $n_{il}^{(k)}$ et T_{ij} il est possible d'estimer les valeurs de $\Pi_{jl}^{(k)}$ d'après l'équation :

$$\pi_{jl}^{(k)} = \frac{\sum_{i=1}^{I} T_{ij} n_{il}^{(k)}}{\sum_{l=1}^{J} \sum_{i=1}^{I} T_{ij} n_{il}^{(k)}} \tag{2.3}$$

Les valeurs de T_{ij} permettent également d'estimer la probabilité p_j que la réponse j choisit au hasard soit correcte :

$$p_j = \frac{\sum_{i=1}^{I} T_{ij}}{I} \tag{2.4}$$

Dawid et al. 1979 indiquent que lorsque les taux d'erreur $\pi_{jl}^{(k)}$ et les probabilités p_j sont connues mais que la vraie réponse à la question i est inconnue il est possible de réaliser une estimation des valeurs de T_{ij} grâce au théorème de Bayes et aux données :

$$p(T_{ij} = 1 | données) = \frac{\prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{jl}^{(k)})^{n_{il}^{(k)}} p_j}{\sum_{q=1}^{J} \prod_{k=1}^{K} \prod_{l=1}^{J} (\pi_{ql}^{(k)})^{n_{il}^{(k)}} p_q} \tag{2.5}$$

L'algorithme 1 qui correspond à EM de Dawid et al. 1979 utilise les équations (2.3), (2.4) et (2.5) pour estimer les valeurs de $\pi_{jl}^{(k)}$, p_j et T_{ij}.

L'algorithme EM de Dawid et al. 1979 permet l'estimation de la bonne réponse et des taux d'erreurs individuels des personnes, c'est pourquoi cette méthode a été adaptée par la suite au *crowdsourcing* dans différents travaux. Whitehill et al. 2009 ont ainsi développé GLAD dans un contexte de classification d'images. L'algorithme estime la précision du contributeur, la difficulté d'une image à classer et sa classe. L'algorithme de Whitehill et al. 2009 est généralisé par Welinder et al. 2010 qui introduisent un concept dimensionnel de difficulté de l'image et le combinent avec une définition plus large de la compétence de l'annotateur. Ipeirotis et al. 2010 et Wang et al. 2011 attribuent un score de qualité aux contributeurs en utilisant EM et Raykar et al. 2012 calculent la sensibilité et la spécificité du contributeur grâce à cet algorithme. À la différence des auteurs cités précédemment Yan et al. 2010 estiment que le "profil" de l'utilisateur n'est pas constant tout au long de l'expérience et emploient EM pour modéliser le profil du contributeur.

Les expériences de Khattak et al. 2011 ; Raykar et al. 2012 ; Raykar et al. 2010 ; Whitehill et al. 2009 montrent que EM offre de meilleurs résultats que le MV pour l'agrégation des réponses. Hung et al. 2013 réalisent une étude comparative plus complète en confrontant le MV, *Honeypot* et ELICE à EM et 3 algorithmes qui en sont issus : SLEM [14], GLAD et ITER [15]. D'après les tests des auteurs, les méthodes qui utilisent EM ont un temps d'exécution plus important, il est donc préférable de ne pas les appliquer en ligne pour cette raison. Mais, en contrepartie, ces algorithmes itératifs offrent une meilleure précision des résultats pour l'agrégation des réponses, tout particulièrement EM et SLEM. De plus, EM et SLEM sont plus robustes face aux *spammers* que ITER et GLAD. Cependant, parmi toutes ces méthodes, seuls le MV, Honeypot et EM sont adaptables aux questions à choix multiples. En effet, les autres approches ne considèrent que des réponses à des questions binaires. L'utilisation de l'algorithme EM est donc très intéressante dans le cadre du *crowdsourcing* puisqu'il permet d'agréger les réponses d'après la qualification estimée du contributeur. Néanmoins, la qualification ne suffit pas à elle seule pour définir le profil du contributeur et EM n'estime pas son comportement qui peut tout de même impacter la qualité des données.

Cet état de l'art nous a permis de constater qu'il n'existe pas encore de méthodes pour l'agrégation des réponses qui tiennent compte du profil du contributeur estimé sans données d'or par sa qualification et son comportement.

2.7 Conclusion

Le *crowdsourcing* se caractérise par l'externalisation par un employeur de tâches non réalisables par ordinateur à une foule de contributeurs sur des plateformes dédiées. Les tâches qu'il est possible d'effectuer grâce au *crowdsourcing* sont très variées, c'est pourquoi il existe plusieurs types de plateformes. Nous en différencions quatre : apport de contenu, activité inventive, activité routinière, activité créative.

Les problématiques du domaine sont d'ordre juridique, technique et social. Il manque ainsi un cadre légal au contributeur afin de le protéger d'éventuels abus de l'employeur. Pour l'employeur, la problématique principale est l'obtention de données de qualité à moindre coût. Plusieurs éléments impactent la qualité des contributions collectées. Il y a la motivation et le profil du contributeur sur lesquelles l'employeur peut essayer d'avoir

14. *Supervised Learning from Multiple Experts*
15. *Iterative learning*

une maîtrise grâce à l'assignation de la tâche. Mais avant cela, la définition de la campagne de *crowdsourcing* joue un rôle crucial. Si l'employeur ne donne pas d'indications claires sur le travail à réaliser et une interface ergonomique cela impactera le travail du contributeur même s'il est de bonne volonté. Finalement, une fois les réponses collectées, le choix de la méthode d'agrégation est également délicat.

Pour résoudre ces problématiques techniques et sociales, il est possible de s'intéresser aux différentes phases de la campagne de *crowdsourcing*. En amont de la campagne, lors de son initialisation, la définition de la tâche et la conception de l'interface ont un rôle crucial. Cependant, il n'existe pas de protocole établi pour une définition optimale de la tâche ni d'interfaces de référence. Il est aussi possible de travailler sur la problématique de la motivation du contributeur au cours de cette première phase d'initialisation de la campagne en commençant par établir la rémunération appropriée. Afin de susciter une motivation intrinsèque chez le contributeur, l'employeur peut lui proposer une tâche ludique ou entrer en interaction avec lui, mais cela est plus complexe à mettre en place.

Lors de la phase de mise en ligne de la campagne, demander à la foule de réaliser un entraînement pour la tâche avant de commencer la campagne permet d'améliorer la qualité des contributions. Une autre solution est de faire passer un test de qualification au contributeur de sorte que seuls ceux ayant réussi le test puissent participer. Cependant, ces deux approches nécessitent que l'employeur soit en possession de données d'or qui ne sont pas toujours accessibles à tous les types de tâches. Une autre solution est d'assigner la tâche aux contributeurs d'après leur qualification. Dans ce cas, ce n'est plus le contributeur qui choisit la tâche et cela peut diminuer son intérêt pour le *crowdsourcing* puisqu'un des avantages principaux pour lui est la possibilité de choisir son travail.

Finalement, pour le contrôle de la qualité des données après la campagne, il s'agit d'analyser le profil du contributeur et de trouver une solution optimale pour l'agrégation des réponses. Traditionnellement, la méthode d'agrégation utilisée est le vote majoritaire, mais cette méthode ne tient pas compte du profil du contributeur ni de l'incertitude sur ses réponses. Nous citons dans ce chapitre une méthode plus élaborée qui utilise les SVNs pour estimer le profil du contributeur, mais cette dernière n'offre pas de méthode d'agrégation des réponses et nécessite des données d'or. Des approches utilisant la théorie des ensembles flous permettent quant à elles de modéliser le profil du contributeur ou d'agréger les réponses. Cependant bien que cette théorie soit pertinente pour la modélisation de l'imprécision elle ne se prête pas à la modélisation de la certitude du contributeur. Finalement, de nombreux auteurs utilisent l'algorithme EM car il permet de calculer l'in-

certitude sur la réponse et certains travaux tiennent compte d'un score de qualité du contributeur. Cependant, ce score de qualité n'est pas totalement révélateur du profil du contributeur puisqu'il ne considère que sa qualification sans prendre en considération son comportement.

En conclusion, il n'existe pas de méthodes qui ne requièrent pas de données d'or pour l'estimation du profil du contributeur qui tiennent compte à la fois de la qualification de du contributeur pour la tâche et de son comportement dans la réalisation de celle-ci. Il n'existe pas non plus de méthodes d'agrégation qui considèrent l'imprécision et l'incertitude du contributeur sur sa réponse tout en tenant compte de son profil.

Cette thèse se focalise uniquement sur des données collectées dans les plateformes d'activités routinières, ainsi, dans la suite de ce document, toute allusion au *crowdsourcing* fait référence à ce type de plateforme. De plus, nous nous intéressons à des données de *crowdsourcing* qui peuvent être imprécises et incertaines inhérentes à toute contribution humaine. Afin de traiter ces imperfections, nous utilisons la théorie des fonctions de croyance qui est introduite dans le chapitre suivant.

Chapitre 3

La théorie des fonctions de croyance

Résumé : La théorie des fonctions de croyance permet la modélisation et l'agrégation de données de sources d'information imparfaites. Nous présentons dans ce chapitre les bases de la théorie et son utilisation pour la classification et le *crowdsourcing*.

Sommaire

3.1 Introduction

La théorie des fonctions de croyance est aussi nommée théorie de Dempster-Shafer, car elle fut introduite par Dempster 1967 et formalisée par Shafer 1976. Elle est une généralisation des approches floues et probabilistes et permet de modéliser l'imprécision et l'incertitude de sources imparfaites d'information.

L'imprécision caractérise l'apport d'information d'une assertion. Par exemple, prenons la question suivante : "Quelle est la date de parution de la première édition de Bilbo le Hobbit de J. R. R. Tolkien ?". La réponse "Dans les années 30" est imprécise, car elle couvre un intervalle de 10 ans, alors que la réponse "Le 21 septembre 1937." est précise.

Nous différencions la certitude associée au résultat d'un événement aléatoire comme un lancer de dé, de la certitude épistémique qui traduit la connaissance de la source sur le sujet considéré. Pour la même question que précédemment, une personne qui a étudié la littérature anglaise sera plus sûre de sa réponse qu'une personne qui ignore à quelle époque a vécu J. R. R. Tolkien. La certitude aléatoire est le plus souvent modélisée par une probabilité. Nous traitons la certitude épistémique dans ce chapitre et dans l'ensemble de ce manuscrit.

L'imprécision et l'incertitude relatives à l'information sont traduites dans la théorie de Dempster-Shafer par des fonctions de croyance, ces dernières sont introduites dans la section 3.2. La section 3.3 explique les procédés applicables aux fonctions de croyance pour faciliter la combinaison qui est abordée dans la section 3.4. Après la combinaison, différentes méthodes de décision peuvent être employées, celles-ci sont présentées section 3.5. La section 3.6 expose plusieurs utilisations de la théorie des fonctions de croyance, notamment dans un contexte de *crowdsourcing*. La section 3.7 vient conclure ce chapitre.

3.2 Fonctions de croyance

On appelle cadre de discernement $\Omega = \{r_0, ..., r_n\}$ l'ensemble des classes ou hypothèses r_i qui sont exclusives et exhaustives. Les fonctions de croyance sont définies sur l'espace puissance :

$$2^\Omega = \{\emptyset, \{r_0\}, \{r_1\}, \{r_0 \cup r_1\}, ..., \Omega\} \tag{3.1}$$

L'élément Ω représente l'ignorance, $\emptyset$ symbolise l'ouverture au monde hors du cadre de discernement. Dans le cadre du *crowdsourcing*, un contributeur c est une source imparfaite, et Ω est constitué de l'ensemble des choix de réponse possibles à une question q. Il existe trois principaux types de fonctions de croyance pour représenter l'incertitude et l'imprécision : les fonctions de masse, de crédibilité et de plausibilité.

3.2.1 Fonctions de masse

Les fonctions de masse m^Ω sont définies sur 2^Ω et à valeurs dans $[0, 1]$. Elles respectent la condition de normalisation :

$$\sum_{X \in 2^\Omega} m^\Omega(X) = 1 \tag{3.2}$$

La masse $m^\Omega(X)$ caractérise le degré de croyance accordé par une source d'information S à l'élément $X \in 2^\Omega$, plus la masse est élevée, plus la croyance est forte. Lorsque $m^\Omega(\emptyset) = 0$ cela signifie qu'une ouverture au monde hors du cadre de discernement n'est pas possible, on dit alors être en monde fermé ou monde clos. De plus, pour $m^\Omega(\Omega) > 0$, la fonction est dite non-dogmatique.

Un élément $X \in 2^\Omega$ tel que $m^\Omega(X) > 0$ est appelé élément focal, et la réunion des éléments focaux constitue le noyau. Si seuls les singletons de Ω sont des éléments focaux alors m^Ω est une probabilité, la fonction est alors appelée fonction de masse bayésienne.

Il existe des fonctions de masse spécifiques, celles-ci sont présentées dans les paragraphes ci-dessous. Dans la suite de cette thèse, la fonction de masse associée au contributeur c pour la question q et le cadre de discernement Ω est notée m^Ω_{cq}. Le contributeur c est considéré comme une source d'information S_c.

Les fonctions de masse catégoriques Lorsque la source S_c est absolument certaine de sa réponse, toute la croyance est accordée à $X \in 2^\Omega$. La réponse X peut être imprécise dans le cas où elle est une réunion de classes appartenant au cadre de discernement.

$$m^\Omega_{cq}(X) = 1, X \in 2^\Omega \tag{3.3}$$

Si X est un singleton $r_i \in \Omega$, alors non seulement la réponse est totalement certaine, mais elle est aussi précise. Dans le cas spécifique où $m^\Omega_{cq}(\Omega) = 1$, la fonction de masse traduit une ignorance complète de la part de c.

Les fonctions de masse à support simple Cette fonction de masse traduit une réponse incertaine et imprécise de la part de la source d'information.

$$\begin{cases} m^{\Omega}_{cq}(X) = \omega_{cq} \text{ avec } X \in 2^{\Omega} \setminus \Omega, \omega_{cq} \in [0,1] \\ \\ m^{\Omega}_{cq}(\Omega) = 1 - \omega_{cq} \\ \\ m^{\Omega}_{cq}(Y) = 0, \ \forall Y \in 2^{\Omega} \setminus \{X, \Omega\} \end{cases} \tag{3.4}$$

Le contributeur c a une connaissance incertaine car il croit partiellement en sa réponse X mais pas totalement puisqu'une masse non nulle est présente sur Ω. Cette fonction est également notée $X^{\omega_{cq}}$.

Fonction de masse consonante Cette fonction de masse a pour particularité d'avoir des éléments focaux emboîtés. Le contributeur estime un ensemble de réponses r_i comme approprié pour la question posée, et parmi cet ensemble, sa certitude varie pour des sous-ensembles de réponses inclus les uns dans les autres. Une fonction de masse avec pour éléments focaux $X_1, X_2, ..., X_n \in 2^{\Omega}$ tels que $X_1 \subset X_2 \subset ... \subset X_n \subseteq \Omega$ est une fonction de masse consonante.

Les fonctions de masse modélisent le degré de croyance élémentaire de la source. La section suivante introduit les fonctions de crédibilité et de plausibilité qui représentent respectivement la croyance minimale et la croyance maximale de la source.

3.2.2 Fonctions de crédibilité et plausibilité

Les fonctions de crédibilité, notées bel_{cq} et de plausibilité, pl_{cq} d'un contributeur c répondant à la question q sont définies dans les paragraphes suivants.

Fonction de crédibilité Ces fonctions de croyance modélisent la croyance minimale d'une source, soit la certitude que l'ensemble des informations fournies par le contributeur c soutiennent la réponse X. Pour tout élément $X \in 2^{\Omega}$ considéré, la fonction de crédibilité regroupe les masses des éléments focaux inclus dans X :

$$bel_{cq}(X) = \sum_{Y \subseteq X} m^{\Omega}_{cq}(Y) \tag{3.5}$$

$bel_{cq}(X)$ représente la part totale de croyance soutenant X. Il est possible qu'un degré de croyance élémentaire ne puisse être affecté précisément à un élément X ce qui se traduit par $m^{\Omega}_{cq}(X) = 0$. Cependant, cela ne signifie pas que la réalisation de X est impossible si la croyance minimale en X est non nulle, soit $bel_{cq}(X) > 0$. La crédibilité de l'ignorance est définie par : $bel_{cq}(\Omega) = 1 - m^{\Omega}_{cq}(\emptyset)$.

Fonction de plausibilité Ces fonctions modélisent la croyance maximale accordée à l'élément X par la source S_c. Pour tout élément $X \in 2^{\Omega}$ considéré, la fonction de plausibilité somme les masses des éléments focaux intersectés par X :

$$pl_{cq}(X) = \sum_{Y \cap X \neq \emptyset} m^{\Omega}_{cq}(Y) \tag{3.6}$$

La fonction de plausibilité, mesure à quel point S_c considère sa réponse comme plausible. Pour $X \in 2^{\Omega}$ nous avons : $pl_{cq}(X) \geq bel_{cq}(X)$.

Une fois l'information des sources modélisée par des fonctions de croyance, il est possible de réaliser l'agrégation de l'information grâce aux opérateurs de combinaison existant. Cependant, s'il existe un désaccord entre les sources S_c, cela a des répercussions lors de l'agrégation et il en résulte une masse non nulle sur l'ensemble vide qui traduit le conflit de la combinaison. Les méthodes exposées dans la section suivante facilitent la combinaison de l'information en diminuant le conflit éventuel.

3.3 Dynamique des fonctions de croyance

Dans cette section, les méthodes permettant de faciliter la combinaison de fonctions de croyance sont abordées.

3.3.1 Affaiblissement

Le coefficient d'affaiblissement $\alpha_c \in [0, 1]$ modélise la confiance accordée à la source c. L'affaiblissement d'une fonction de masse est défini de la manière suivante :

$$\begin{cases} m^{\Omega,\alpha}_{cq}(X) = \alpha_c m^{\Omega}_{cq}(X), \ \forall X \in 2^{\Omega} \setminus \Omega \\ \\ m^{\Omega,\alpha}_{cq}(\Omega) = 1 - \alpha_c(1 - m^{\Omega}_{cq}(\Omega)) \end{cases} \tag{3.7}$$

Plus ce coefficient est grand, plus la source est estimée fiable. Si $\alpha_c = 0$ la source n'est absolument pas fiable et la totalité de la masse est reportée sur l'ignorance Ω. L'affaiblissement augmente les intervalles $[bel_{cq}, pl_{cq}]$ et diminue ainsi le conflit résultant de la combinaison des fonctions de croyance.

3.3.2 Raffinement et grossissement du cadre de discernement

Si deux sources s'expriment sur deux cadres de discernement différents, mais compatibles, il est possible de combiner l'information grâce au raffinement ou au grossissement d'un des cadres de discernement. Par exemple, si un contributeur c_1 répond à un QCM pour l'identification d'oiseaux avec les propositions suivantes : $\Omega_1 = \{$ Rouge-gorge, Pigeon, Mésange ... $\}$. Un second contributeur c_2 participe également à ce questionnaire, mais avec des réponses plus explicites de l'espèce d'oiseau : $\Omega_2 = \{$ Rouge-gorge familier, Pigeon ramier, Pigeon biset, Mésange charbonnière, Mésange nonnette, Mésange à tête noire ... $\}$. Grâce à une fonction de raffinement R, $X \in 2^{\Omega_1}$ peut être exprimée sur 2_2^{Ω} :

$$m_{c_2q}^{\Omega_2}(R(X)) = m_{c_1q}^{\Omega_1}(X), \forall X \in 2^{\Omega_1} \tag{3.8}$$

Pour un grossissement de la fonction de masse, il faut utiliser la réciproque de R. Cependant ces fonctions ne s'appliquent que pour des cadres de discernement compatibles, dans le cas contraire il est nécessaire de se tourner vers les méthodes d'extension vide et de marginalisation décrites dans la section suivante.

3.3.3 Extension vide et marginalisation

L'extension vide est la projection de Ω sur un produit cartésien de cadres de discernement incluant Ω. Cette opération est décrite par l'équation :

$$m^{\Omega\uparrow\Omega\times\Theta}(B) = \begin{cases} m^{\Omega}(A) \text{ si } B = A \times \Theta \\ \\ 0 \text{ sinon} \end{cases} \tag{3.9}$$

Delmotte et al. 2004 présentent la *balloning extension* pour l'extension vide d'un sous-ensemble du cadre de discernement sur ce dernier. Soit un cadre de discernement

$\Omega' \subset \Omega$, et une fonction $m^{\Omega'}$ définie sur $2^{\Omega'}$, la *balloning extension* de Ω' sur Ω est :

$$m^{\Omega'\uparrow\Omega}(B) = \begin{cases} m^{\Omega'}(A) \text{ si } A \subseteq \Omega', B = A \cup \overline{\Omega'} \\ \\ 0 \text{ sinon} \end{cases} \tag{3.10}$$

La marginalisation est la projection du produit de cadres de discernement sur l'un d'eux : $\Omega \times \Theta \downarrow \Theta$. Soit le produit cartésien $\Omega \times \Theta$, la marginalisation permettant de se projeter sur Ω est donnée par l'équation :

$$m^{\Omega\times\Theta\downarrow\Omega}(B) = \sum_{A\subseteq\Omega\times\Theta, A\downarrow\Omega=B} m^{\Omega\times\Theta}(A) \tag{3.11}$$

Grâce aux méthodes d'extension vide et de marginalisation, il est possible d'effectuer des projections de Ω sur $\Omega \times \Theta$ et réciproquement. Ces opérations sont fort utiles pour la combinaison de fonctions de croyance possédant des cadres de discernement différents.

3.3.4 Opérations complémentaires

Cette section présente la distance de Jousselme et la décomposition canonique de fonctions de masse qui sont des opérations complémentaires qui peuvent être employées dans certains cas pour la combinaison de l'information.

Distance de Jousselme Cette distance est définie par Jousselme et al. 2001 afin d'estimer la proximité entre deux fonctions de masse. Soit deux fonctions de masse m_1 et m_2 ayant le même cadre de discernement Ω, la distance est donnée par :

$$d_J(m_1, m_2) = \sqrt{\frac{1}{2}(m_1 - m_2)^T \underline{\underline{D}}(m_1 - m_2)} \tag{3.12}$$

$$\underline{\underline{D}}(X, Y) = \begin{cases} 1 \text{ si } X = Y = \emptyset \\ \\ \frac{|X\cap Y|}{|X\cup Y|} \forall X, Y \in 2^\Omega \end{cases} \tag{3.13}$$

Dans l'équation (3.12), $\underline{\underline{D}}$ est une matrice de taille $2^\Omega \times 2^\Omega$ fondée sur la dissimilarité de Jaccard. Plus la distance d_J est élevée, plus les fonctions de masse sont différentes, et réciproquement, plus elle est faible plus elles sont semblables.

Décomposition canonique de fonction de masse Une fonction de masse non-dogmatique dont les éléments focaux $X_i \subset \Omega$ sont distinct peut se décomposer de manière unique :

$$m = \bigcap_{X \subset \Omega} X^{\omega(X)} \tag{3.14}$$

Dans l'équation 3.14, $X^{\omega(X)}$ est une fonction de masse à support simple dont l'élément focal X à pour masse $\omega(X) \in]0, 1]$. L'opérateur utilisé pour la combinaison des $X^{\omega(X)}$ pour retrouver m est l'opérateur de combinaison conjonctive décrit dans la section suivante.

3.4 Combinaison des sources d'informations

Pour la fusion de l'information, les sources s'expriment toutes sur le même cadre de discernement Ω et le même sujet. Si une combinaison sur des cadres de discernement différents Ω et Θ est souhaitée, il faut effectuer une des opérations vues dans la section précédente pour se ramener à un même cadre de discernement avant la combinaison. Dans l'hypothèse où Ω et Θ sont compatibles, il est possible d'utiliser les méthodes de raffinement ou grossissement du cadre de discernement. Dans le cas contraire, l'extension vide ou la marginalisation doivent être employées.

Les notations suivantes sont utilisées : K contributeurs c répondent à la question q d'après l'ensemble de propositions Ω. Pour toutes les équations définies dans cette section $X \in 2^\Omega$ il est possible de réaliser la moyenne de fonctions de masse :

$$m_{Moy}(X) = \frac{1}{K} \sum_{c=1}^{K} m_{cq}^{\Omega}(X) \tag{3.15}$$

Cette combinaison simpliste des fonctions de masse permet notamment de rester en monde clos. De nombreuses autres règles de combinaison existent dans la théorie des fonctions de croyance, cette section en recense quelques unes, Martin 2019 en présente davantage. Les méthodes conjonctives sont différenciées des disjonctives et des approches hybrides.

3.4.1 Opérateurs de combinaison conjonctive

Cette section présente plusieurs opérateurs utilisant uniquement une combinaison conjonctive de l'information. La combinaison conjonctive nécessite que toutes les sources soient fiables, distinctes et indépendantes.

Combinaison Conjonctive Cet opérateur permet de réduire l'imprécision sur les éléments focaux et d'accroître la croyance sur les éléments concordants.

$$m^{\Omega}_{Conj}(X) = \left(\bigcap_{c=1}^{K} m^{\Omega}_{cq}\right)(X) = \sum_{Y_1 \cap \ldots \cap Y_N = X} \prod_{c=1}^{N} m^{\Omega}_{cq}(Y_c) \tag{3.16}$$

La masse $m^{\Omega}_{Conj}(\emptyset)$ représente le conflit global de la combinaison.D'autres opérateurs permettent de rester en monde clos après combinaison, par exemple l'opérateur de combinaison conjonctive normalisée de Dempster ou celui de Yager 1987.

Combinaison conjonctive normalisée de Dempster Cet opérateur permet une répartition équitable du conflit $k = m^{\Omega}_{Conj}(\emptyset)$ sur les éléments focaux.

$$m^{\Omega}_{D}(X) = \frac{1}{1-k} m^{\Omega}_{Conj}(X) \tag{3.17}$$

Opérateur de combinaison conjonctive de Yager 1987 Les auteurs ont fait le choix de transférer toute la masse du conflit sur l'ignorance d'après l'équation :

$$\begin{cases} m^{\Omega}_{Y}(X) = m^{\Omega}_{Conj}(X) \\ m^{\Omega}_{Y}(\Omega) = m^{\Omega}_{Conj}(\Omega) + m^{\Omega}_{Conj}(\emptyset) \\ m^{\Omega}_{Y}(\emptyset) = 0 \end{cases} \tag{3.18}$$

Règle PCR La règle PCR, nommée ainsi pour *Proportional Conflict Redistribution* est introduite par Smarandache et al. 2005. Cette règle permet la redistribution du conflit partiel des masses sur les éléments de 2^{Ω} qui sont à l'origine de ce conflit. Cette opération se décompose en trois étapes. Pour commencer la combinaison conjonctive des fonctions de masse est effectuée. Ensuite, il faut procéder au calcul du conflit des masses avant de finalement redistribuer ce conflit. Smarandache et al. 2005 proposent cinq règles

PCR pour la redistribution du conflit, numéroté PCR1 à PCR5, Martin et al. 2006 complète cet ensemble avec la règle PCR6. Soit deux sources d'information et m_1 et m_2 leurs fonctions de masse associées, la règle PCR s'écrit :

$$m_{PCR6}(X) = m_{Conj}(X) + \sum_{\substack{Y \in 2^\Omega \\ X \cap Y \neq \emptyset}} \left(\frac{m_1(X)^2 m_2(Y)}{m_1(X) + m_2(Y)} + \frac{m_2(X)^2 m_1(Y)}{m_2(X) + m_1(Y)} \right) \tag{3.19}$$

Règle LNS Parfois, l'opérateur de combinaison conjonctive ne permet pas d'obtenir des résultats décidables. C'est notamment le cas lorsque le nombre de sources à combiner est élevé ou ne sont pas toutes fiables, comme pour des sources humaines. L'opérateur de combinaison LNS donné par l'équation (3.20) et proposé par Zhou et al. 2017 présente l'intérêt de diminuer la contrainte de fiabilité des sources. En effet, la règle LNS requiert pour son application que les sources soient indépendantes cognitivement et majoritairement fiables. Pour cette règle, plus une source est cohérente avec d'autres, plus elle est fiable. Une décomposition canonique des fonctions de masse m^Ω_{cq} est réalisée pour chaque contributeur c pour la même question q afin d'obtenir l'ensemble de fonctions de masse à support simple $\{X_k^{\omega_{cq}}, X_k \subset \Omega\}$. Les fonctions de masses $X_k^{\omega_{cq}}$ sont ensuite regroupées en K *clusters*, K étant le nombre distinct de X_k. Chaque *cluster* est constitué d'un nombre s_k de fonctions de masse à support simple.

$$m^\Omega_{LNS} = \bigcap_{k=1,\ldots,K} (X_k)^{1 - \alpha_k \left(1 - \prod_{c=1}^{s_k} \omega_{cq}\right)} \tag{3.20}$$

$$\alpha_k = \frac{s_k}{\sum_{i=1}^{K} s_i} \tag{3.21}$$

Dans l'équation (3.20), α_k correspond au nombre moyen de fonctions de masse à support simple présentes dans le *cluster* k par rapport au nombre total de fonctions de masse générées par la décomposition.

Règle conjonctive prudente La règle conjonctive prudente, *Cautious rule* en anglais, est définie par Denœux 2006. Cette règle peut être appliquée à des sources d'information dépendantes. Soit deux fonctions de masses m_1 et m_2, l'application de la règle conjonctive

prudente à ces fonctions est donnée par l'équation suivante :

$$m_{1\textcircled{\cap}2} = m_1 \textcircled{\cap} m_2 = \textcircled{\cap}_{A\subseteq\Omega} A^{(1-\omega_1(A))\wedge(1-\omega_2(A))} \tag{3.22}$$

Dans l'équation (3.22), le symbole $\wedge$ est l'opérateur de minimum et ω_i la masse obtenue après la décomposition canonique de m_i par la combinaison conjonctive.

3.4.2 Opérateurs de combinaison disjonctive et hybride

Cette section introduit tout d'abord l'opérateur de combinaison disjonctive, puis des opérateurs hybrides qui utilisent conjointement des opérateurs différents.

Combinaison disjonctive La combinaison disjonctive nécessite qu'au moins une des sources soit fiable et que toutes soient cognitivement indépendantes.

$$m^{\Omega}_{Dis}(X) = \sum_{Y_1 \bigcup \ldots \bigcup Y_K = X} \prod_{c=1}^{K} m^{\Omega}_{cq}(Y_c) \tag{3.23}$$

Cette combinaison offre une perte totale du conflit et permet donc de rester en monde clos. En revanche, elle élargit les éléments focaux, ce qui entraîne une perte de spécificité.

***Combinaison With Adapted Conflict* (CWAC)** Proposée par Lefèvre et al. 2013, cette règle de combinaison utilise les opérateurs conjonctifs des équations (3.16) et (3.17).

$$m_{CWAC}(X) = \gamma_1 m_{Conj}(X) + \gamma_2 m_D(X) \tag{3.24}$$

Dans l'équation (3.24) γ_1 et γ_2 sont calculés d'après la distance de Jousselme qui permet une estimation du conflit entre les fonctions de masse combinées.

***Robust Combination Rules* (RCR)** Les règles de combinaison conjonctive et disjonctive sont les opérateurs majeurs dans la théorie des fonctions de croyance. Plusieurs opérateurs les utilisant sont proposés par : Liang-zhou et al. 2005 ; Martin et al. 2007 ; Osswald et al. 2006. Nous revenons maintenant sur l'opérateur RCR défini par Florea et al. 2009 qui repose sur les règles conjonctive et disjonctive.

$$m_{RCR}(X) = \gamma_1(k) m_{Dis}(X) + \gamma_2(k) m_{Conj}(X) \tag{3.25}$$

Dans cette équation k est le conflit global, $\gamma_1(k)$ et $\gamma_2(k)$ sont définis en fonction de k.

Après avoir présenté les opérations de combinaison pour la théorie des fonctions de croyance, nous allons maintenant aborder dans la section suivante la prise de décision.

3.5 Décision

La combinaison des sources d'information permet de calculer la masse m_{Comb} qui est le résultat de l'agrégation des fonctions de masse. Il existe différentes stratégies pour la prise de décision sur m_{Comb} dans la théorie des fonctions de croyance. Il s'agit de considérer le maximum de plausibilité, de crédibilité ou de probabilité pignistique. Pour ces trois méthodes, la décision se fait toujours sur Ω. Une autre solution, proposée par Essaid et al. 2014, consiste à prendre une décision grâce à la distance de Jousselme.

Maximum de plausibilité La fonction de plausibilité pl est donnée dans la section 3.2.2 par l'équation (3.6). Choisir le maximum de plausibilité revient à choisir $r_d \in \Omega$ tel que :

$$pl(r_d) = \max_{r_i \in \Omega} pl(r_i) \tag{3.26}$$

La fonction pl indique la croyance maximale accordée à un élément de 2^Ω. Par conséquent, choisir son maximum revient à une décision optimiste qui manque parfois de pouvoir discriminant. Cette méthode de décision est optimale pour les fonctions de masse dérivées de probabilités.

Maximum de crédibilité Soit la fonction de crédibilité bel définie par l'équation (3.5) de la section 3.2.2 dédiée. Le maximum de crédibilité est $r_d \in \Omega$ tel que :

$$bel(r_d) = \max_{r_i \in \Omega} bel(r_i) \tag{3.27}$$

Lorsque le résultat de la combinaison des fonctions de croyance ne porte que sur des singletons, l'élément r_d choisi d'après le maximum de crédibilité est le même que celui obtenu par le maximum de plausibilité. La décision d'après le maximum de crédibilité est pessimiste et plus sélective ce qui peut entraîner des erreurs dans le choix effectué.

Maximum de probabilité pignistique Le niveau crédal qui consiste en la modélisation et la manipulation de l'information se différencie du niveau pignistique qui permet

la prise de décision. La probabilité pignistique $betP$ est introduite par Smets 1990 et est un compromis entre les fonctions de plausibilité et crédibilité :

$$bel(X) \leq betP(X) \leq pl(X) \ \forall X \in 2^{\Omega} \tag{3.28}$$

La probabilité pignistique est calculée comme suit :

$$betP(X) = \sum_{Y \in 2^{\Omega}, Y \neq \emptyset} \frac{|X \cap Y|}{|Y|} \frac{m^{\Omega}(Y)}{1 - m^{\Omega}(\emptyset)} \tag{3.29}$$

Le maximum de probabilité est obtenu pour r_d tel que :

$$betP(r_d) = \max_{r_i \in \Omega} betP(r_i) \tag{3.30}$$

L'utilisation de $betP$ offre un bon compromis entre les fonctions pl et bel, notamment en cas d'ambiguïté sur la réponse.

Décision sur la distance Essaid et al. 2014 proposent de prendre une décision en utilisant la distance de Jousselme. Pour ce faire, d_J est calculée entre la masse résultant de la combinaison de l'information m_{Comb} et une fonction de masse catégorique m_X ; $X \in 2^{\Omega}$ est l'unique élément focal de cette fonction :

$$D = \arg\min_{X \in 2^{\Omega}} d_J(m_{Comb}, m_X) \tag{3.31}$$

L'élément $X \in 2^{\Omega}$ choisi est celui qui permet de minimiser la distance entre la combinaison des fonctions de croyance et la fonction de masse catégorique. Cette solution de décision est intéressante, car X n'est pas nécessairement un singleton.

La théorie des fonctions de croyance étant introduite, nous abordons dans la section suivante ses applications dans le cadre du *crowdsourcing*.

3.6 Fonctions de croyance et *crowdsourcing*

Les fonctions de croyance sont principalement utilisées dans le cadre de la classification supervisée ou semi-supervisée. D'après Denoeux 2018, il existe trois manières d'exploiter la théorie de Dempster-Shafer pour la classification. La première est la fusion

de classifieurs qui consiste à modéliser leurs sorties par des fonctions de croyance afin de les combiner. La deuxième méthode désigne ce que l'auteur nomme calibration crédibiliste, les décisions de classifieurs statistiques sont converties en fonctions de croyance. La troisième approche est un classifieur principalement conçu par des fonctions de croyance et appelé classifieur crédibiliste. Denoeux et al. 2006 font la distinction entre les classifieurs *distance-based* et *model-based*. Les *distance-based* sont fondés sur des distances comme les K-plus proches voisins crédibilistes de Denoeux 1995. Les *model-based* reposent sur l'utilisation du théorème de Bayes généralisé par Smets 1993 en remplaçant les probabilités conditionnelles par des fonctions de croyance. Cette approche a par la suite été reprise dans les travaux de Fixsen et al. 1997, et Appriou 1998 proposent une approche similaire sous un autre angle de vue. Delmotte et al. 2004 comparent les fonctions de croyance et notamment l'utilisation du théorème de Bayes généralisé aux probabilités, et en concluent que les deux théories ont des performances similaires. Dans certains cas, une divergence de résultats sur la décision peut être observée entre les deux approches due à la différence de modélisation de l'ignorance.

Nous pouvons également mentionner l'exemple de la classification hiérarchique de Maalel et al. 2014 qui recourt à la théorie des fonctions de croyance. D'autres approches encore, consistent à employer un algorithme EM crédibiliste, ce que font Jraidi et al. 2007, Vannoorenberghe 2007 et Côme et al. 2009. Les expériences réalisées par Côme et al. 2009 confirment l'intérêt d'utiliser des étiquettes imparfaites et imprécises comme solution au bruit. De plus, les résultats de Vannoorenberghe 2007 montrent que la version crédibiliste d'EM offre de meilleures performances comparées à l'algorithme originel.

Dans les plateformes de *crowdsourcing* les fonctions de croyance sont employées pour l'estimation du profil du contributeur ou l'agrégation des réponses. L'utilisation de ces fonctions est intéressante, car tout comme les probabilités elles permettent de considérer l'incertitude d'une contribution. La théorie de Dempster-Shafer permet également de modéliser l'imprécision des réponses ce qui peut être un véritable avantage pour le *crowdsourcing*. En effet, Smets 1997 émet l'hypothèse que plus un individu est imprécis, plus il est certain et à l'inverse, plus il est précis moins il est certain. C'est pourquoi autoriser le contributeur à être imprécis pourrait être un avantage pour l'employeur puisque les réponses de la foule seraient alors plus certaines. Peu d'éléments de la littérature portent, à notre connaissance, sur la théorie des fonctions de croyance appliquée au *crowdsourcing*.

Mais parmi ceux existants nous différencions les approches pour l'agrégation des réponses, l'estimation du profil du contributeur et la réalisation conjointe de ces deux opérations.

3.6.1 Agrégation des réponses

Koulougli et al. 2016 ont défini la méthode CASCAD pour la modélisation et l'agrégation de contributions issues de plateformes de *crowdsourcing*. Afin d'appliquer CASCAD, le contributeur doit réaliser un test de qualification en amont de la campagne pour déterminer son expertise d'après l'échelle suivante : ignorant, peu compétent, moyennement compétent, compétent ou expert. Durant la campagne, le participant sélectionne une à plusieurs réponses et assigne à ses contributions son degré d'incertitude ce qui permet de construire des fonctions de masse. Ces fonctions sont affaiblies par un coefficient α relatif à l'expertise du contributeur précédemment établie, après quoi les réponses sont agrégées en cascade par l'opérateur de combinaison de Dempster (équation (3.17)). CASCAD est comparée au vote majoritaire et à l'algorithme EM de Dawid et al. 1979, des données générées sont employées pour les tests. CASCAD obtient un taux de bonne réponse plus élevé que EM et le MV à partir du moment où trois éléments focaux sont utilisés. En revanche, cette méthode est plus coûteuse en temps d'exécution et en mémoire.

3.6.2 Estimation du profil

Deux approches existent pour l'estimation du profil : la méthode de Ouni et al. 2017 qui nécessite des données d'or et celle de Ben Rjab et al. 2016 qui s'en affranchit.

Ouni et al. 2017 estiment l'expertise des contributeurs en vue de distinguer les "experts" E des "non-experts" NE d'après le cadre de discernement $\Omega = \{E, NE\}$. Les auteurs construisent un graphe orienté sur les réponses attendues grâce aux données d'or, un autre graphe est défini d'après les réponses données par le contributeur. Afin de comparer le graphe de référence au graphe du contributeur, des fonctions de masse sont calculées pour chaque nœud :

— m_1^Ω compare la position d'un nœud entre les deux graphes.

— m_2^Ω mesure la proportion de nœuds de même distance au point de départ du graphe que le nœud considéré.

— m_3^Ω et m_4^Ω mesurent les erreurs d'inversion entre les nœuds précédents et suivants d'un nœud donné.

La fonction de masse pour l'ensemble du graphe est calculée d'après la moyenne des fonctions de masse m_1^Ω, m_2^Ω, m_3^Ω et m_4^Ω ce qui permet d'estimer l'expertise du contributeur. Afin de tester leur modèle, les auteurs ont utilisé des données réelles obtenues grâce à des campagnes de *crowdsourcing* d'évaluation de la qualité d'enregistrements sonores. Une des campagnes a été accomplie par une foule de contributeurs vivant en Asie et une autre par des contributeurs résidant aux États-Unis d'Amérique. Les degrés d'expertise calculés respectivement pour ces deux campagnes sont comparés et il apparaît que l'expertise des contributeurs américains est plus élevée que celle des asiatiques. Ce phénomène est expliqué dans l'article comme résultant des différences culturelles des deux continents, car les extraits musicaux proposés à l'écoute étaient de sonorité occidentale. Cependant, il n'est pas toujours possible pour l'employeur d'avoir des données d'or, c'est pourquoi l'approche de Ben Rjab et al. 2016 qui n'est pas soumise à cette contrainte est intéressante.

Afin d'identifier les contributeurs experts de leur domaine sans avoir recours à des données d'or, Ben Rjab et al. 2016 calculent un degré de précision DP_c, équation (3.33), et un degré d'exactitude DE_c, équation (3.32), sur la réponse. Soit E_C l'ensemble des contributeurs, E_{Q_c} l'ensemble des questions auxquelles un contributeur c a répondu et Ω_q le cadre de discernement associé à la question q.

$$\begin{cases} DE_c = 1 - \dfrac{1}{|E_{Q_c}|} \displaystyle\sum_{q \in E_{Q_c}} d_J(m_c^{\Omega_q}, m_{E_C|c}^{\Omega_q}) \\ \\ m_{E_C|c}^{\Omega_q}(X) = \dfrac{1}{|E_C| - 1} \displaystyle\sum_{j \in E_C|c} m_j^{\Omega_q}(X) \end{cases} \tag{3.32}$$

Le degré d'exactitude DE_c traduit l'exactitude globale des réponses du contributeur c comparé aux réponses agrégées du reste de la foule. Ce degré n'est cependant pertinent que si la majorité de la foule a raison. L'équation repose sur la mesure de conflit de Martin et al. 2008 où d_J est la distance de Jousselme et al. 2001.

$$\begin{cases} DP_c = \dfrac{1}{|E_{Q_c}|} \displaystyle\sum_{q \in E_{Q_c}} \delta_c^{\Omega_q} \\ \\ \delta_c^{\Omega_q} = 1 - \displaystyle\sum_{X \in 2^{\Omega_q}} m_c^{\Omega_q}(X) \frac{log_2(|X|)}{log_2(|\Omega_q|)} \end{cases} \tag{3.33}$$

Le degré d'imprécision DP_c mesure la dispersion des réponses pondérée par leur masse.

Le degré global d'expertise du contributeur DG_c est calculé par les auteurs en pondérant les degrés DE_c et DP_c par un coefficient $\beta_c \in [0, 1]$:

$$DG_c = \beta_c DE_c + (1 - \beta_c) DP_c \tag{3.34}$$

L'étude de Ben Rjab et al. 2016 propose une comparaison avec une approche probabiliste mesurant l'expertise d'un contributeur. Des données générées sont utilisées pour les expériences. Les résultats obtenus par les auteurs montrent que le calcul de DG_c est plus pertinent pour l'évaluation d'experts que l'approche probabiliste.

3.6.3 Estimation du profil et agrégation des réponses

La méthode d'Abassi et al. 2018 nommée CGS-BLA [1] emploie également des données d'or. CGS-BLA permet l'estimation du profil du contributeur et l'agrégation de ses réponses. Pour ce faire, les auteurs définissent trois profils, l'"Expert", le "Bon contributeur" et le "Mauvais contributeur". Pour identifier le profil du contributeur c, trois mesures de la précision de la réponse sont calculées :

— le taux de réponse de c en accord avec le corpus de référence composé des données d'or

— le taux de réponse de c en accord avec les réponses agrégées par vote majoritaire

— la proportion de réponses du reste de la foule qui sont similaires à la réponse de c

Une fois ces mesures obtenues, l'algorithme de classification *k-mean* est appliqué avec k=3 afin de classifier les contributeurs. Les réponses de c sont modélisées par des fonctions de croyance et affaiblies par une valeur α_c d'après le profil du contributeur. Si c est expert $\alpha_c = 1$, la réponse du contributeur est inchangée, à l'inverse, pour un "Mauvais contributeur" $\alpha_c = 0$, aucun crédit n'est accordé à la réponse du contributeur et toute la croyance est assimilée à l'ignorance. Un "Bon contributeur" se voit quant à lui attribuer un affaiblissement égal à son taux de bonnes réponses sur les données d'or. Les fonctions de masse sont ensuite combinées par l'opérateur CWAC (équation (3.24)) pour chaque question puis les probabilités pignistiques de chaque réponse sont calculées pour la prise de décision. CGS-BLA offre de meilleurs résultats en termes de précision comparés au vote majoritaire et ELICE.

1. *Clustering approach of the Gold Standards based Belief Label aggregation*

3.7 Conclusion

La théorie des fonctions de croyance permet de modéliser l'imprécision et l'incertitude des informations provenant de sources imparfaites. Les fonctions de croyance peuvent être des fonctions de masse, de crédibilité ou de plausibilité. Les premières représentent la croyance élémentaire de la source, les deuxièmes la croyance minimale et les troisièmes la croyance maximale. Différents opérateurs de combinaison existent pour l'agrégation de l'information. Une fois les fonctions de croyance combinées, il est possible de se ramener à un cadre probabiliste pour la prise de décision ou de rester à un niveau crédal en utilisant la distance de Jousselme.

Les fonctions de croyance peuvent être employées pour la classification qu'elle soit supervisée, semi-supervisée ou non supervisée. Dans le cadre du *crowdsourcing* un contributeur est une source imparfaite qui peut renseigner une réponse imprécise et incertaine. Une contribution peut donc aisément être modélisée par une fonction de croyance. Pourtant cette théorie est peu employée dans le *crowdsourcing*. Kougougli et al. 2016 utilisent la théorie des fonctions de croyance pour l'agrégation des réponses. Ouni et al. 2017 et Ben Rjab et al. 2016 l'emploient pour l'estimation du profil du contributeur. Finalement seuls Abassi et al. 2018 proposent une méthode permettant à la fois l'estimation du profil et l'agrégation des réponses. Parmi ces quatre approches, seuls les travaux de Ben Rjab et al. 2016 n'utilisent pas de données d'or. Comme il n'est pas toujours possible pour l'employeur d'avoir de telles données dans les campagnes de *crowdsourcing* , les trois autres méthodes sont par conséquent fort contraignantes. De plus, la méthode d'Abassi et al. 2018 est testée sur des réponses précises alors qu'un des principaux atouts de la théorie des fonctions de croyance est la modélisation de l'imprécision des sources. Kougougli et al. 2016 et Ben Rjab et al. 2016 autorisent le contributeur à être imprécis, mais leurs expérimentations sont réalisées sur des données générées.

Au cours de cette thèse, afin de réaliser les expériences sur des données réelles, nous avons élaboré une interface pour le recueil de données imprécises et incertaines dans le cadre du *crowdsourcing*. Le chapitre suivant introduit l'interface définie et sa validation.

Chapitre 4

Définition d'une interface de *crowdsourcing*

Résumé : L'indécision du contributeur sur la réponse peut impacter négativement la qualité des données collectées. Pour pallier ce problème, nous proposons d'enrichir les informations collectées, mêmes imparfaites, dans l'objectif d'améliorer la prise de décision. Pour ce faire, une interface de *crowdsourcing* est définie afin d'offrir au contributeur répondant la possibilité d'être imprécis et de renseigner sa confiance dans les réponses données. Ce chapitre présente l'interface proposée et le protocole expérimental suivi pour sa validation via des campagnes de *crowdsourcing*.

Sommaire

4.1 Introduction

Il est expliqué dans la section 2.6.1 que le plus souvent les employeurs ont recours à des questionnaires à réponses fermées comme les QCMs dans les plateformes de *crowdsourcing*. Il s'agit de questions pour lesquelles un ensemble de réponses est proposé au contributeur. En fonction de l'interface utilisée le répondant peut sélectionner une unique réponse, à l'aide de boutons radio par exemple, ou une à plusieurs réponses dans le cas de cases à cocher. L'interprétation des contributions est laborieuse, car certaines questions peuvent induire une situation d'indécision chez le contributeur dont la certitude peut alors différer du reste de la foule.

Actuellement très peu d'interfaces de *crowdsourcing* permettent à l'utilisateur de moduler sa réponse afin d'y intégrer ses éventuelles hésitations. Nous pouvons néanmoins citer l'étude de Kazai et al. 2013 qui demande au contributeur son degré de connaissance sur le sujet et d'autres informations. Par ailleurs, la plupart des études existantes de Diaz et al. 2008, Farrell 2006 et Khan et al. 2001 s'appliquent au milieu éducatif. Dans ce milieu, les questionnaires ont toujours une bonne réponse établie et l'objectif est l'évaluation des connaissances d'un étudiant. Or, dans les plateformes de *crowdsourcing*, il n'est pas fréquent d'avoir des données d'or. De plus, l'intérêt pour l'employeur n'est pas d'estimer le niveau du contributeur, mais bien d'obtenir des données pertinentes en vue de leur agrégation.

Nous introduisons dans ce chapitre une interface de *crowdsourcing* qui permet de prendre en considération l'hésitation éventuelle du contributeur à travers son imprécision et sa certitude. L'imprécision traduit l'indécision du contributeur pour plusieurs réponses, et la certitude sa confiance en la véracité de sa sélection. L'interface est réfléchie en considérant l'hypothèse de Smets 1997 d'après laquelle, "plus un individu est imprécis plus il est certain, et réciproquement plus il est précis moins il est certain". Le contributeur peut choisir une à plusieurs réponses s'il en éprouve le besoin tout en indiquant sa certitude dans sa contribution. Nous employons la théorie des fonctions de croyance pour modéliser l'imperfection des données collectées et procéder à leur agrégation.

Un autre objectif de cette interface est de diminuer le nombre de répondants nécessaire sans impacter négativement la qualité des données agrégées. Afin de valider l'interface et l'hypothèse de Smets 1997 (section 3.6) des campagnes de *crowdsourcing* ont été réalisées sous la forme de QCMs dont chaque question admet une unique bonne réponse,

l'hypothèse est aussi vérifiée par Thierry et al. 2021 pour une autre base de donnée provenant de *crowdsourcing*. La section 4.2 détaille le protocole expérimental suivi et la collecte des données sur la plateforme de *crowdsourcing* Crowdpanel[1]. La section 4.3 présente l'analyse des données réalisée pour vérifier les corrélations entre : la difficulté de la tâche, la certitude du contributeur, son imprécision et son taux de bonne réponse. Nous comparons ensuite différentes méthodes d'agrégation des réponses dans la section 4.4. Les résultats des sections 4.3 et 4.4 sont également explicités par Thierry et al. 2020a. Nous souhaitons que cette nouvelle interface soit bénéfique, non seulement pour l'employeur, mais aussi pour les contributeurs, c'est pourquoi les retours des utilisateurs vis-à-vis de l'interface sont considérés dans la section 4.5. La section 4.6 conclut ce chapitre.

4.2 Le protocole expérimental

Nous avons réalisé quatre campagnes de *crowdsourcing* avec quatre interfaces différentes faisant appel distinctement puis conjointement à de l'imprécision et de l'incertitude de la part du contributeur,et ce afin d'observer l'impact de ces deux éléments sur sa façon de répondre. Parmi ces quatre campagnes il y a donc une campagne témoin sans incertitude ni imprécision, une campagne pour laquelle la certitude est demandée, une autre où le contributeur peut être imprécis et une dernière où il peut être imprécis tout en donnant sa certitude. La section 4.2.1 expose le protocole suivi et les interfaces définies. La section 4.2.2 revient sur la collecte des contributions au sein de la plateforme Crowdpanel.

4.2.1 Les campagnes

Afin d'identifier les corrélations entre la difficulté de la tâche, l'imprécision de la réponse du contributeur et sa certitude il est essentiel de pouvoir quantifier la difficulté de la question posée. De plus, la question doit s'affranchir d'éventuels biais de connaissance qui vont faire varier la difficulté de la tâche suivant le contributeur. En effet, Kazai et al. 2012 montrent que différents éléments comme la géolocalisation de l'individu, son âge, son niveau d'étude ou encore sa personnalité ont un fort impact sur la justesse de ses réponses. Par exemple, une photo d'oiseau est présentée au contributeur et le nom de l'espèce lui est demandé. Un contributeur sans connaissance ne reconnaîtra pas l'oiseau, un amateur pourra renseigner le nom commun et un expert le nom scientifique.

1. https ://crowdpanel.io/ (28/10/2021)

Pour pallier ce problème, nous utilisons une tâche reposant sur la perception visuelle qui ne nécessite aucun savoir de la part du contributeur. À chaque question, un ensemble de segments parallèles est présenté au répondant qui doit sélectionner le plus grand. La difficulté de la tâche est inversement proportionnelle à la différence de taille entre le bon segment et les autres. Ceci permet de contrôler finement la difficulté tout en ayant toujours une unique bonne réponse.

Puisque nous souhaitons étudier l'impact de l'imprécision et de l'incertitude sur la réponse du contributeur, quatre expériences faisant appel de manière distincte à ces deux éléments sont définies. Ces expériences donnent chacune lieu à une campagne de *crowdsourcing*. Nous avons suivi les conseils énoncés dans la section 2.6.1 pour la conception de l'interface et avons fait tester chaque expérience par des volontaires au sein du laboratoire avant d'effectuer les campagnes. Pour chacune d'entre elles, cinq segments sont présentés au contributeur et il lui est demandé de sélectionner celui qu'il estime être le plus grand. Parmi les cinq segments proposés, l'un d'entre eux est le segment à sélectionner et les quatre autres sont des segments de contrôle, de taille identique : 40 mm. Le segment à sélectionner fait : $40+\delta$ mm, où δ est la taille augmentée du segment. Les segments sont tous espacés de 20 mm. D'une question à une autre, la position du plus grand segment change aléatoirement, de sorte que tous les segments sont augmentés une fois d'une valeur δ ce qui permet d'éviter un apprentissage de la part du contributeur au cours de l'expérience. Dans le cas où $\delta = 0$ mm, le segment que le contributeur doit trouver a la même taille que les quatre segments de contrôle, soit 40 mm. Lorsque les cinq segments sont de taille identique, le bon segment est indiscernable ce qui a pour objectif de placer le contributeur dans une situation d'indécision absolue. Chaque expérience est composée de plusieurs blocs de questions et chaque bloc contient une question d'attention afin de garder le contributeur attentif au cours de la tâche et de s'assurer de son sérieux.

Pour chaque campagne, une page d'instruction est présentée au contributeur avant qu'il ne commence la tâche afin de lui expliquer le travail à réaliser. Pour les expériences permettant des réponses imparfaites, il est indiqué dans cette page d'instruction que le contributeur ne sera pas pénalisé si sa contribution est incertaine ou imprécise. Les caractéristiques des quatre expériences sont décrites dans les paragraphes suivants et résumées par le tableau 4.1 page 67.

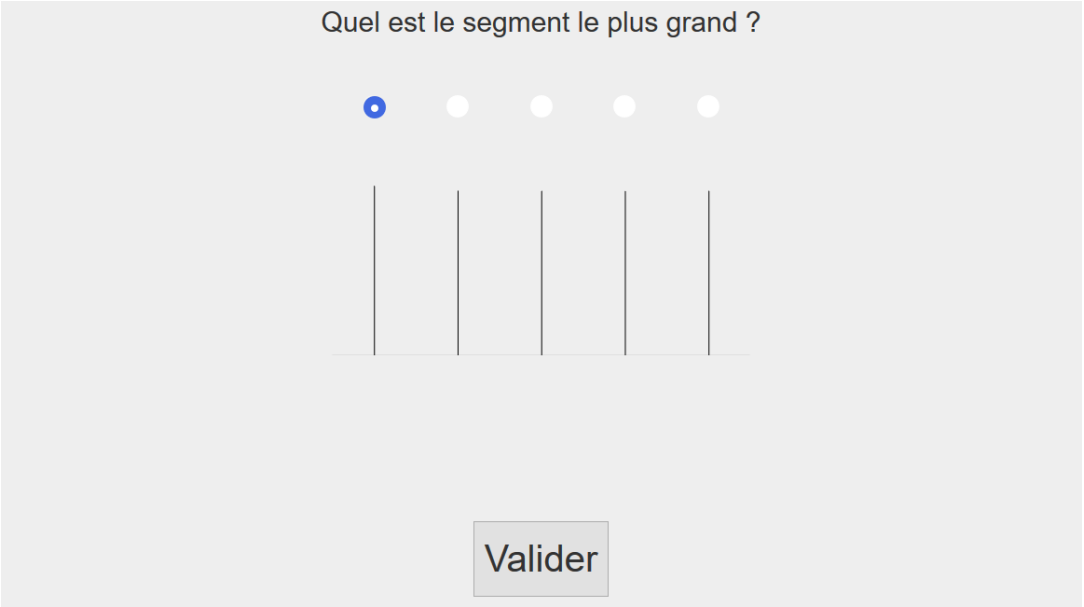


FIGURE 4.1 – Interface de l'**expérience 0**. Le contributeur choisit un unique segment.

Expérience 0 Cette expérience témoin est la plus représentative des campagnes habituelles de *crowdsourcing*. Elle est réalisée afin d'avoir une comparaison des réponses et du temps mis par le contributeur pour réaliser la tâche d'une part classiquement, de façon précise et sans mentionner sa certitude, et d'autre part dans l'approche proposée, avec possibilité d'être imprécis et en indiquant sa certitude. Le temps de réalisation d'une campagne est un élément important pour l'employeur car il est un des facteurs du coût avec la taille de la foule.

Le contributeur doit sélectionner un unique segment, ainsi des boutons radio sont utilisés sur l'interface présentée par la figure 4.1. Pour cette expérience, la certitude du contributeur n'est pas demandée et il lui est impossible de la donner. L'ensemble de difficulté utilisé est :

$$\Delta_0 = \{0\ mm, 0.3\ mm, 0.6\ mm, 0.9\ mm, 1.2\ mm, 1.8\ mm, 2.4\ mm\}$$

Lorsque les cinq segments sont de taille identique, le contributeur est dans l'incapacité d'identifier le bon segment et le taux de bonne réponse attendu est égal à l'équiprobabilité : 20%, soit une chance sur cinq de choisir aléatoirement le bon segment.

FIGURE 4.2 – Interface de l'**expérience 1**. Le contributeur choisit un unique segment et la certitude associée à sa réponse.

Expérience 1 L'objectif de cette expérience est d'observer l'évolution de la certitude du contributeur face à différents degrés de difficulté. L'interface utilisée est présentée par la figure 4.2, sur laquelle cinq segments verticaux sont alignés vers le bas. Il est demandé au contributeur de sélectionner le segment qu'il estime être le plus grand. Le contributeur est contraint de donner une réponse précise, ainsi il ne peut sélectionner qu'un seul segment grâce aux boutons radio. Il doit également préciser sa certitude en sa réponse. Pour ce faire, il choisit une assertion sur l'échelle de Likert à sept points qui lui sont proposés. Les éléments composant cette échelle sont des degrés d'accord : "Totalement pas d'accord", "Pas d'accord", "Plutôt pas d'accord", "Neutre", "Plutôt d'accord", 'D'accord", "Totalement d'accord".

L'expérience 1 est réalisée avec deux ensembles de valeurs δ afin d'observer si la certitude du contributeur est relative à la difficulté de la question ou à la difficulté globale de la campagne.

$$\Delta_1 = \{0\ mm, 0.3\ mm, 0.6\ mm, 0.9\ mm, 1.2\ mm, 1.5\ mm\}$$

$$\Delta_2 = \{0\ mm, 0.3\ mm, 0.6\ mm, 1.2\ mm, 1.8\ mm, 2.4\ mm\}$$

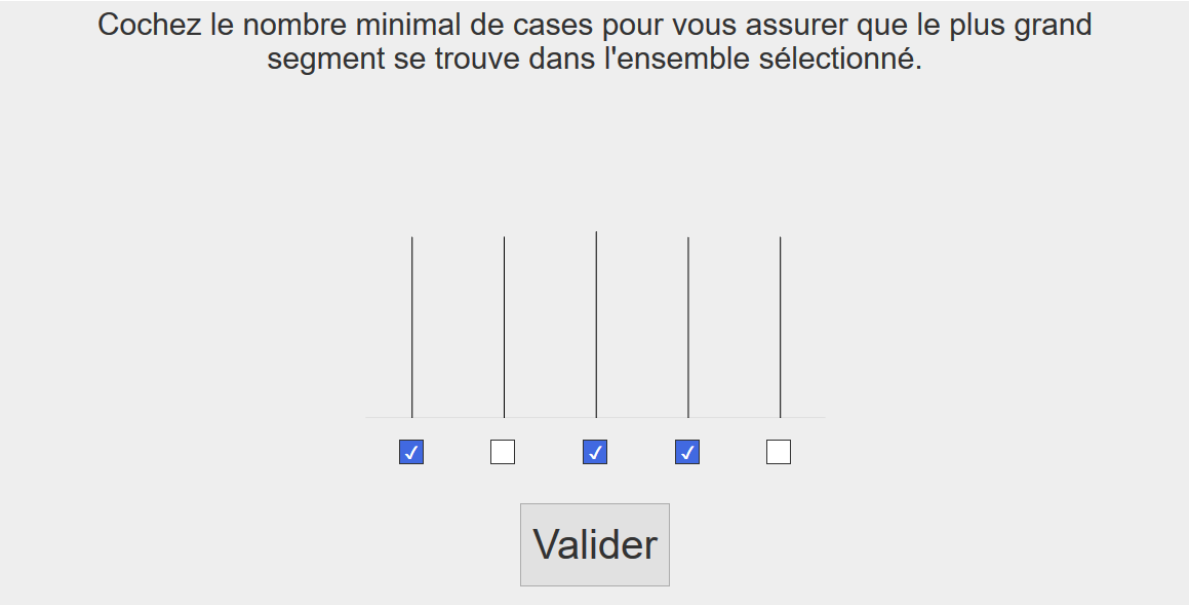


FIGURE 4.3 – Interface de l'**expérience 2**. Le contributeur choisit un à cinq segments.

La foule qui réalise cette expérience est scindée en deux groupes de même taille, de sorte qu'un groupe répond aux questions avec l'ensemble de difficulté Δ_1 et l'autre Δ_2.

Comme pour l'expérience 0, pour $\delta = 0$ le taux de bonne réponse attendu est 20% puisque les 5 segments sont de tailles identiques. Pour cette valeur de δ les contributeurs devraient donc être incertains de leur réponse puisqu'il leur est impossible de déterminer visuellement quel est le bon segment à sélectionner et qu'ils ne peuvent sélectionner qu'un seul segment.

Expérience 2 Le but de cette expérience est d'observer le comportement de l'utilisateur vis-à-vis de l'imprécision lorsqu'il est confronté à des questions de difficulté variable.

Dans cette expérience, dont l'interface est présentée figure 4.3, le contributeur a la possibilité de sélectionner plusieurs segments en cas d'hésitation. Des cases à cocher lui permettent de choisir de 1 à 5 segments. L'objectif pour le contributeur au cours de cette expérience est de sélectionner le plus petit ensemble de segments tout en étant certain que le plus grand d'entre eux y est bien inclus. Ainsi, pour les valeurs élevées de δ pour lesquelles le plus grand segment est facilement identifiable, il est attendu que le contributeur n'en sélectionne qu'un seul. La certitude du contributeur n'est pas demandée, et il ne peut la renseigner en aucune façon. Les tests préliminaires à la campagne réalisés en laboratoire ont montré que deux ensembles de difficulté ne sont pas nécessaires à

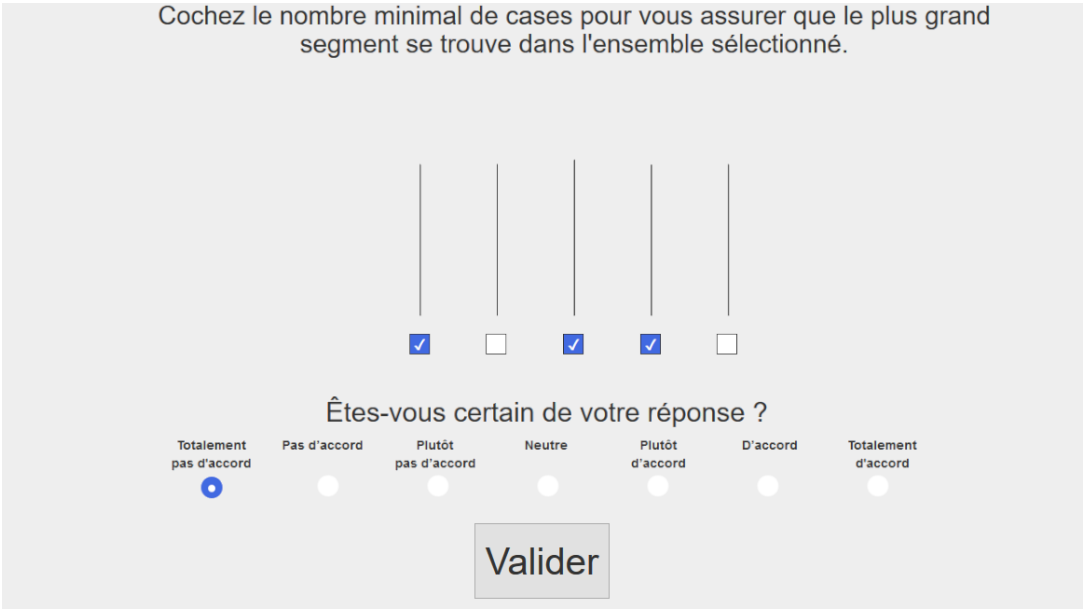


FIGURE 4.4 – Interface de l'**expérience 3**. Le contributeur choisit 1 à 5 segments et la certitude associée à sa réponse.

l'expérience. C'est pourquoi pour cette expérience et les suivantes nous avons fait le choix d'utiliser l'ensemble de difficulté Δ_0 qui est composé de valeurs δ communes à Δ_1 et Δ_2. L'ensemble Δ_0 n'inclut pas toutes les valeurs de l'union de Δ_1 et Δ_2 afin que la campagne de *crowdsourcing* ne dure pas trop longtemps pour le contributeur. Puisque le contributeur peut être imprécis, nous attendons de lui qu'il sélectionne les cinq segments lorsque ceux-ci sont de taille identique pour $\delta = 0\ mm$.

Expérience 3 Cette expérience vise à mettre en évidence les fluctuations de la certitude du contributeur en fonction de l'imprécision de ses réponses.

L'interface utilisée pour l'expérience (figure 4.4), permet au contributeur d'être imprécis en cas d'hésitation, grâce aux cases à cocher, et il lui est demandé de renseigner sa certitude globale dans ses choix. L'échelle de Likert utilisée pour la certitude est la même que pour l'expérience 1. Comme pour l'expérience 2, l'interface permet de sélectionner 1 à 5 segments et l'ensemble de difficulté des questions est Δ_0.

Questions d'attention Pour les quatre campagnes de *crowdsourcing* les questions d'attention se fondent sur le même principe : la question qui a été posée précédemment est de nouveau adressée au contributeur qui doit renseigner une contribution identique à celle

FIGURE 4.5 – Question d'attention de l'expérience 3 avec imprécision et certitude.

qu'il a déjà donné. Pour ce faire les mêmes segments qu'à la question précédente sont affichés et il est demandé au contributeur de redonner les mêmes réponses. L'interface employé est très ressemblante à celle des autres questions de la campagne. Dans l'exemple donné par la figure 4.5 le contributeur doit cocher les mêmes cases et donner la même certitude car il s'agit de l'expérience 3 pour laquelle il peut être imprécis et donner sa certitude.

Dans la section suivante nous détaillons la mise en place des campagnes de *crowdsourcing* pour les quatre expériences décrites ici.

4.2.2 La collecte de données réelles

Au total quatre campagnes de *crowdsourcing* sont menées, une par expérience. Pour réaliser une expérience le contributeur doit utiliser un système d'exploitation Windows ou Apple, car les interfaces utilisent l'application Pointingserver de la bibliothèque libpointing développée par CASIEZ et al. 2011. Celle-ci n'est disponible que pour ces deux systèmes d'exploitation. L'application Pointingserver est exploitée pour connaître les dimensions de l'écran du contributeur en *mm* afin de s'assurer que les tailles des segments qui lui sont présentés sont identiques indépendamment de la taille de l'écran utilisé.

Lorsqu'un contributeur participe, à une expérience il ne peut pas réaliser une autre

expérience que nous proposons, de sorte que la foule est intégralement renouvelée pour chaque campagne et il n'y a pas d'apprentissage possible de la part d'un contributeur d'une expérience à une autre. Les cinq segments sont augmentés chacun une fois d'une valeur δ pour chaque ensemble de difficulté ce qui compose un bloc de questions. Ainsi, pour l'expérience 1, Δ_1 comme Δ_2 sont composés de 6 valeurs de δ, ce qui fait un bloc de 30 questions. Pour les trois autres expériences, Δ_0 inclut 7 valeurs de δ, un bloc est donc constitué de 35 questions. À la fin de chaque expérience, un questionnaire est adressé aux contributeurs afin d'avoir un retour sur leur ressenti au cours des expériences.

Pour la collecte de données réelles nous utilisons Crowdpanel [2], qui est une plateforme de *crowdsourcing* française. Les contributeurs sont rémunérés 10 € de l'heure sur Crowdpanel quelque soit la campagne réalisée. Pour chaque campagne nous indiquons à la plateforme la taille de la foule souhaitée et la durée de la tâche. Chaque expérience a été implantée sur un serveur Node.js et la plateforme redirige la foule sur l'interface web dédiée à la campagne.

Pour l'expérience 0 , l'ensemble Δ_0 est utilisé avec 2 blocs de questions. À la différence des autres expériences, c'est principalement le temps de réponse qui nous intéresse, c'est pourquoi la foule qui réalise cette expérience est de taille plus restreinte et composée de 25 contributeurs, pour un total de :

7 Valeurs de δ × 5 Positions × 2 Blocs × 25 Contributeurs = 1750 réponses

Pourl'expérience 1, 50 contributeurs ont réalisé la campagne avec Δ_1 et 50 autres avec Δ_2. Tous ont répondu à 3 blocs de questions, soit :

2 Ensembles Δ × 6 Valeurs de δ × 5 Positions × 3 Blocs × 50 Contributeurs = 9000 réponses

L'objectif de cette répartition des contributeurs est de déterminer si leur certitude est absolue, ou relative à la difficulté globale des questions. Après cette expérience, nous avons conclu qu'il n'est plus nécessaire d'avoir deux ensembles de difficulté, c'est pourquoi Δ_0 est utilisé. De même, 100 contributeurs sont répartis uniformément sur Δ_1 et Δ_2. Afin de conserver la même taille de foule pour les expériences suivantes, 100 contributeurs sont interrogés avec l'ensemble de difficulté Δ_0.

2. https ://crowdpanel.io/ (29/10/2021)

Campagne	Difficulté	Imprécision	Certitude	Taille de la foule	Nombre de réponses	Durée moyenne
Expérience 0	Δ_0	Non	Non	25	1750	12 min
Expérience 1	Δ_1 / Δ_2	Non	Oui	100	9000	18 min
Expérience 2	Δ_0	Oui	Non	100	7000	13 min
Expérience 3	Δ_0	Oui	Oui	100	7000	16 min

TABLE 4.1 – Récapitulatif des campagnes de *crowdsourcing* réalisées.

Pour les campagnes des expériences 2 et 3, seul l'ensemble de difficulté Δ_0 est utilisé, il est répété dans 2 blocs de questions, soit :

7 VALEURS DE δ × 5 POSITIONS × 2 BLOCS × 100 CONTRIBUTEURS = 7000 RÉPONSES

Le tableau 4.1 résume pour chaque campagne l'ensemble de difficulté utilisé, la possibilité pour le contributeur d'être imprécis et la nécessité de renseigner sa certitude. Il inclut également la taille de la foule, le nombre de réponses collectées et la durée moyenne nécessaire aux contributeurs pour réaliser la campagne.

Dans la section suivante, les données collectées sont analysées afin d'étudier les corrélations entre la difficulté et la réponse du contributeur.

4.3 Influence de la difficulté de la tâche sur la réponse du contributeur

La difficulté de la tâche dans les expériences réalisées est relative à la taille augmentée de δ du plus grand segment. Plus δ est grand, plus la question est simple pour le contributeur car le bon segment se distingue de façon évidente des quatre segments témoins. Les corrélations entre la difficulté de la tâche et les éléments qui constituent la réponse du contributeur sont analysées dans cette section. La certitude associée à la réponse, son imprécision éventuelle et sa véracité sont ainsi considérées.

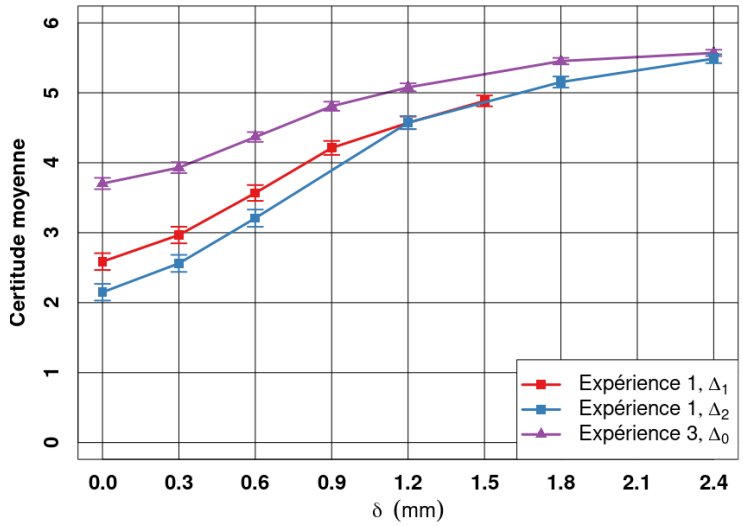


FIGURE 4.6 – Certitude moyenne des contributeurs pour les expériences 1 et 3 en fonction de δ. Intervalles de confiance : 95%.

4.3.1 Difficulté et certitude

À chaque élément de l'échelle de Likert proposée au contributeur pour renseigner sa certitude, est associé un entier $n_c \in [0, 6]$. Plus le contributeur indique être certain de sa réponse, plus n_c est élevé, de sorte que pour $n_c = 0$, il est totalement incertain et à l'opposé, pour $n_c = 6$, il est totalement certain.

La moyenne des n_c est réalisée pour chaque valeur de difficulté δ dans les expériences 1 et 3 pour lesquelles le contributeur doit renseigner sa certitude. Dans la figure 4.6, la moyenne des certitudes est représentée par la courbe rouge pour les contributeurs qui ont réalisé l'expérience 1 avec Δ_1 et par la courbe bleue pour ceux qui ont fait l'expérience avec Δ_2. La courbe violette, quant à elle, présente les résultats des contributeurs de l'expérience 3. D'après la figure 4.6, pour les trois ensembles Δ_0, Δ_1 et Δ_2, la certitude moyenne de la foule augmente lorsque δ augmente et que la difficulté diminue.

Lorsque la question est la plus complexe, pour $\delta = 0$ mm, il est impossible de distinguer le bon segment des témoins car ils sont tous de taille identique. Cette question a pour objectif de placer le contributeur dans une situation d'indécision extrême. Pour l'expérience 1, la foule ne pouvait sélectionner qu'un seul segment, c'est pourquoi nous attendions de ces contributeurs qu'ils soient totalement incertains de leur réponse. La certitude moyenne devrait donc avoisiner 0. Cependant, sur la figure 4.6, les contributeurs

de l'expérience 1 ont une certitude moyenne comprise entre 2 et 3 ce qui correspond aux degrés "Plutôt incertain" et "Ni certain ni incertain". Il apparaît que les contributeurs admettent difficilement qu'ils sont incertains de leur réponse alors même qu'il leur a été spécifié qu'ils ne seraient pas pénalisés pour cela.

La figure 4.6 présente également les intervalles de confiance à 95% de chaque courbe. Si un point n'est pas inclus dans cet intervalle, cela montre qu'il est significativement éloigné du reste de la courbe. Or pour $\delta \in [0, 0.6]$, les courbes de l'expérience 1 sont espacées et leurs intervalles de confiance disjoints. Cette différence de certitude moyenne entre les deux foules pour des valeurs de δ identiques est liée aux difficultés globales de chaque campagne. En effet, pour Δ_2 avec $\delta \in [1.8, 2.4]$ le segment le plus grand est très facilement perceptible. Ainsi, lorsque le contributeur voit ensuite apparaître un segment avec une faible valeur de δ, la question lui apparaît beaucoup plus complexe. Pour le dernier point $\delta = 1.2\ mm$ commun aux deux ensembles Δ_1 et Δ_2, les courbes de l'expérience 1 se superposent ce qui montre que les contributeurs ont la même certitude en la véracité de leur réponse. Il se trouve que cette valeur de certitude est comprise entre 4 et 5, soit entre les éléments "Plutôt certain" et "Certain" de l'échelle proposée au contributeur. Cette certitude plus importante, commune à l'ensemble des contributeurs de l'expérience 1, s'explique par le fait que pour cette valeur de δ le taux de bonne reconnaissance des contributeurs calculé d'après le nombre de segments correctement identifiés par la foule, est proche de 100% (figure 4.9). Le plus grand segment pour cette valeur de δ est donc facilement identifiable par la majorité des contributeurs. En effet, pour $\delta = 2.4\ mm$, le plus long segment est strictement distinct des témoins. Il s'agit de la question la plus facile de l'ensemble, et les intervalles des valeurs de certitude des expériences 1 et 3 se superposent alors que les ensembles de difficulté utilisés sont différents et que le contributeur peut être imprécis dans l'expérience 3. La certitude donnée par le contributeur à une question est donc relative à la difficulté globale de la campagne de *crowdsourcing*.

Finalement, pour l'expérience 3, pour laquelle la foule a eu l'opportunité d'être imprécise, la certitude moyenne est plus élevée comparée à l'expérience 1 où le contributeur doit donner une unique réponse. De plus, lorsque les segments sont de taille identique ($\delta = 0\ mm$), il y a un écart de 18.7% et 25.8% pour les moyennes des certitudes entre l'expérience 3 et l'expérience 1 avec respectivement Δ_1 et Δ_2. En offrant la possibilité aux contributeurs d'être imprécis, ils sont plus certains de leurs réponses.

Nous sommes ainsi amenés à étudier la capacité du contributeur à donner des réponses imprécises en fonction de la difficulté de la question dans la section suivante.

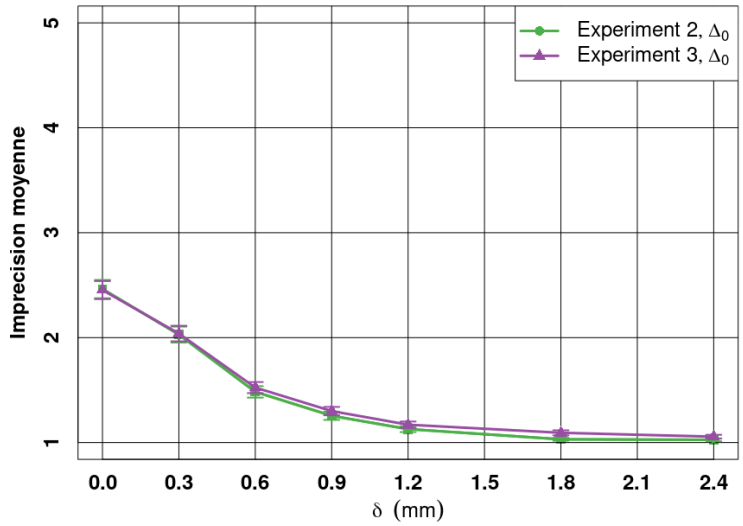


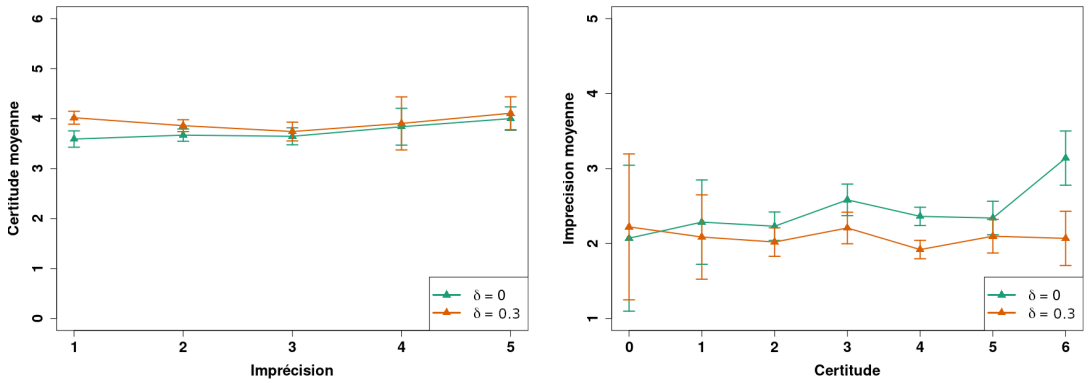
FIGURE 4.7 – Imprécision moyenne des contributeurs pour les expériences 2 et 3 en fonction de δ. Intervalles de confiance : 95%.

4.3.2 Difficulté et imprécision

L'imprécision du contributeur à une question est égale au nombre de segments qu'il a sélectionnés. Il est possible de sélectionner 1 à 5 segments pour les expériences 2 et 3 pour lesquelles le contributeur peut être imprécis, c'est pourquoi l'imprécision moyenne de la foule est étudiée sur l'intervalle $[1, 5]$ sur la figure 4.7. Pour cette figure, l'imprécision moyenne des contributeurs est calculée pour chaque valeur de $\delta \in \Delta_0$.

D'après cette figure, l'imprécision moyenne décroît avec la difficulté de la tâche. Lorsque la tâche est la plus simple, $\delta \in [1.8, 2.4]$, le contributeur est précis et sélectionne une unique réponse. À l'opposé, la foule sélectionne bien davantage de segments lorsque la question est plus complexe. Cependant, entre 2 et 3 segments sont sélectionnés en moyenne lorsque les 5 segments sont identiques pour $\delta = 0\ mm$. Pour cette question qui est la plus difficile et où le plus grand segment est indiscernable des 4 témoins, nous attendions du contributeur qu'il sélectionne les 5 segments. De manière similaire aux contributeurs de l'expérience 1 qui n'admettent pas d'être incertains, la majorité de la foule pour les expériences 2 et 3 n'ose pas être totalement imprécise alors qu'il est spécifié que cela n'est pas pénalisant.

Les intervalles de confiance à 95% des deux courbes sont joints, ce qui montre que les contributeurs ont la même utilisation de l'imprécision pour ces deux expériences. Pour-



(a) Certitude moyenne en fonction de l'imprécision

(b) Imprécision moyenne en fonction des degrés de certitude

FIGURE 4.8 – Certitude et imprécision moyenne pour les valeurs de $\delta = \{0, 0.3\}$ de l'expérience 3.

tant, la certitude du contributeur en sa réponse est demandée dans l'expérience 3. Cela montre que demander la certitude du contributeur n'a pas d'impact sur son utilisation de l'imprécision. Ce point est intéressant car nous avons montré dans la section précédente que l'imprécision du contributeur a un réel impact sur sa certitude puisque dans l'expérience 3 la foule est plus sûre de ses réponses.

La figure 4.8a présente les valeurs moyennes de certitude pour chaque degré d'imprécision pour les valeurs de $\delta = \{0, 0.3\}$. Ces deux valeurs de δ ont été choisies car, d'après la figure 4.7, il s'agit des niveaux de difficulté pour lesquels l'imprécision moyenne est la plus élevée. La figure 4.8 de manière symétrique montre les moyennes de l'imprécision des réponses d'après la valeur de certitude donnée. Nous constatons grâce à la figure 4.8a que la certitude moyenne associée aux réponses croît à partir d'une contribution incluant 3 segments pour $\delta = 0.3$ mm. De manière plus générale, la certitude moyenne croît avec l'imprécision pour $\delta = 0$ mm. La certitude moyenne est tout de même plus élevée lorsque le contributeur ne choisit qu'un seul segment pour $\delta = 0.3$ mm, mais pour cette difficulté, il est possible d'observer un segment de taille supérieure aux autres et il est par conséquent normal que les contributeurs aptes à identifier le bon segment soient certains de leur réponse. De plus, d'après la figure 4.8, pour $\delta = 0$ mm, l'imprécision moyenne du contributeur augmente avec les valeurs de certitude, donc plus le contributeur est certain, plus l'imprécision de sa réponse est grande. Cependant, ce n'est pas le cas pour $\delta = 0.3$ mm car le contributeur n'est pas positionné dans une situation d'indécision aussi

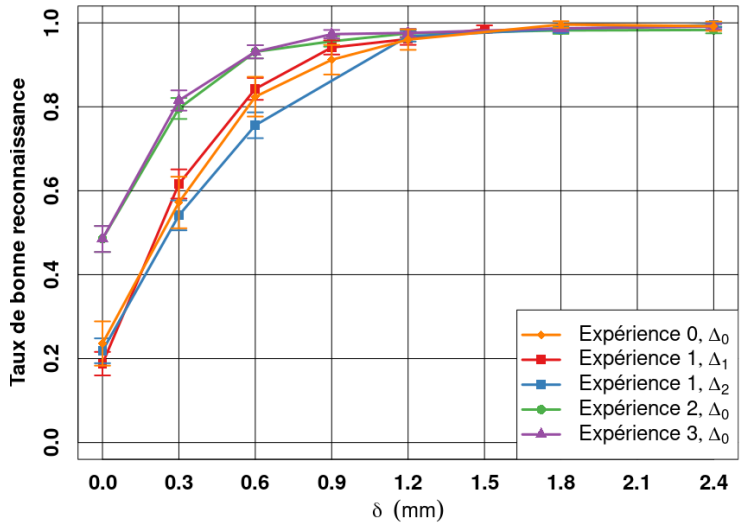


FIGURE 4.9 – Taux de bonne reconnaissance des contributeurs en fonction de δ. Intervalles de confiance : 95%.

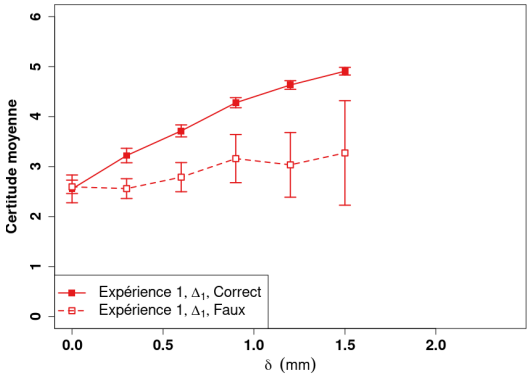
forte que pour $\delta = 0$ *mm*. Par conséquent, lorsque le contributeur est dans une situation d'indécision totale, plus il est imprécis plus il est certain, et réciproquement, plus il est précis moins il est certain ce qui valide l'hypothèse de SMETS 1997 présentée en section 3.6. Ces résultats sont également observés pour d'autres données imprécises et incertaines provenant de campagnes de *crowdsourcing* par THIERRY et al. 2021.

Après avoir analysé ici les relations entre la difficulté de la tâche, la certitude et l'imprécision du contributeur, la section suivante compare cette difficulté au taux de bonne reconnaissance du contributeur.

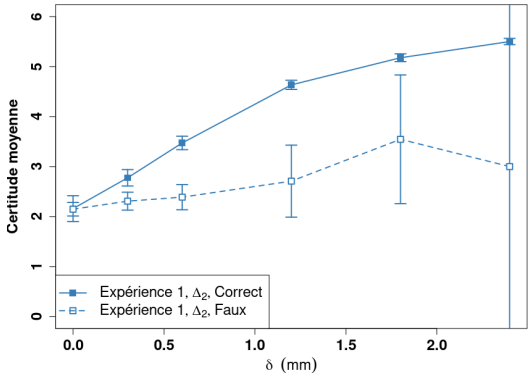
4.3.3 Difficulté et taux de bonne reconnaissance

Le taux de bonne reconnaissance (TBRec) représente le nombre de fois où les réponses renseignées par les contributeurs sont justes. Pour les expériences 2 et 3, dans le cas où le contributeur est imprécis, si le plus grand segment est inclus dans l'ensemble sélectionné alors la réponse est considérée comme correcte.

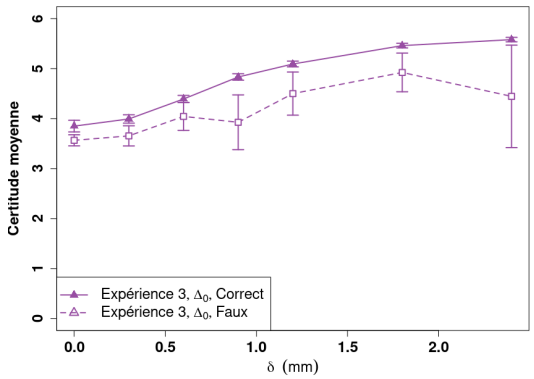
La figure 4.9 présente une comparaison des TBRec pour les quatre expériences et les trois ensembles de difficulté. Pour l'ensemble des expériences, le TBRec augmente lorsque δ augmente et la difficulté de la question diminue. Les taux des expériences 0 et 1, pour $\delta = 0$ avoisinent 0.2. Cette valeur correspond à l'équiprobabilité de choisir aléatoirement



(a) Expérience 1, Δ_1



(b) Expérience 1, Δ_2



(c) Expérience 3, Δ_0

FIGURE 4.10 – Certitude moyennes des réponses correctes et fausses en fonction de δ pour les expériences 1 et 3

le bon segment parmi les cinq affichés par l'interface. Pour les expériences 2 et 3 les TBRec sont plus élevés car les contributeurs ont eu la possibilité d'être imprécis. Pour ces expériences avec imprécision le TBRec est d'environ 0.5 car l'imprécision moyenne des contributeurs pour $\delta = 0$ est de 2.5 d'après la figure 4.7. Si les contributeurs avaient sélectionné les 5 segments pour $\delta = 0$ alors le TBRec serait de 100%.

L'accroissement de la certitude des contributeurs sur la figure 4.6 est en adéquation avec l'accroissement du TBRec du plus grand segment sur la figure 4.9. Pour la figure 4.10, pour les expériences 1 et 3, pour chaque valeur de δ, les moyennes des valeurs de certitude des réponses correctes (courbe pleine) et des réponses incorrectes (courbe en pointillé) sont calculées. Les intervalles de confiance à 95% sont également présents. Pour les trois graphiques de la figure 4.10, les valeurs de certitude moyenne des réponses correctes

sont supérieures à celles des réponses fausses, que le contributeur soit précis ou non. La certitude que le contributeur a en sa réponse est donc un bon indicateur de sa véracité.

Nous avons constaté dans cette section qu'offrir l'opportunité au contributeur d'être imprécis est bénéfique puisqu'il est ainsi plus certain de sa contribution. Dans la section suivante nous montrons que notre interface présente également un intérêt pour l'employeur car elle permet de diminuer la taille de la foule requise pour la campagne de *crowdsourcing*.

4.4 Agrégation des réponses et coût des campagnes

L'intérêt pour l'employeur dans les plateformes de *crowdsourcing* d'activités routinières est d'obtenir des données pour un faible coût. Nous montrons dans cette section qu'en fonction de l'interface et de la méthode d'agrégation utilisées, l'employeur peut diminuer le coût de la campagne sans impacter négativement la qualité des données collectées. Nous réalisons ainsi une comparaison des TBR de la foule pour trois méthodes de modélisation et d'agrégation : le MV (section 2.6.3), EM (section 2.6.3) et la théorie des fonctions de croyance (chapitre 3). Les méthodes d'agrégation utilisées sont tout d'abord introduites, puis la comparaison des résultats est effectuée.

4.4.1 Méthodes d'agrégation utilisées

Différentes études portent sur l'agrégation des réponses dans les plateformes de *crowdsourcing* (section 2.6.3). Les paragraphes ci-dessous reviennent sur le MV, EM et les fonctions de croyance en détaillant les paramètres utilisés pour nos expériences.

Vote majoritaire Cette méthode d'agrégation couramment employée dans les plateformes de *crowdsourcing* est détaillée section 2.6.3. Il s'agit pour l'employeur de choisir la réponse ayant recueilli le plus de votes. Habituellement cette méthode est utilisée pour agréger des réponses précises, elle a été adaptée dans cette thèse aux réponses imprécises afin de procéder à une comparaison avec EM et les fonctions de croyance. L'ensemble des éléments de réponses possibles est appelé Ω et il est composé des cinq segments présentés au contributeur c. Pour chaque question q à laquelle répond c, une fonction indicatrice $\mathbb{1}_{cq}$ est associée aux éléments de Ω, donnée par :

$$\begin{cases} \mathbb{1}_{cq}(X) = 1, X \in \Omega \text{ si le contributeur choisit le segment } X \\ \mathbb{1}_{cq}(Y) = 0, Y \in \Omega \backslash X, \text{ sinon} \end{cases} \tag{4.1}$$

(0) Totalement incertain	(1) Incertain	(2) Plutôt incertain	(3) Ni certain ni incertain	(4) Plutôt certain	(5) Certain	(6) Totalement certain
0.2	0.3	0.4	0.5	0.6	0.7	0.8

TABLE 4.2 – Valeurs numériques ω associées à l'échelle de certitude proposée.

Les fonctions indicatrices sont ensuite sommées sur l'ensemble des contributeurs pour chaque question. Le segment qui est sélectionné par le plus grand nombre de contributeurs est validé par le MV.

Expectation Maximization (EM) L'algorithme EM et ses variantes utilisées pour le *crowdsourcing* sont présentées en section 2.6.3. Dans le cadre des expériences réalisées au cours de cette thèse, l'algorithme de DAWID et al. 1979 est implanté, le pseudo code 1 utilisé est donné en annexe 1 159. La matrice de valeurs T_{ij} indiquant la probabilité de la véracité de la réponse j à la question i est initialisée par MV, la méthode employée est la même que celle décrite dans le paragraphe ci-dessus. La réponse r choisie à la question i est telle que : $T_{ir} = \max_j T_{ij}$.

Fonctions de croyance La théorie des fonctions de croyance permet de modéliser l'incertitude et l'imprécision de sources imparfaites, elle est introduite dans le chapitre 3. Dans le contexte du *crowdsourcing* les contributeurs sont considérés comme des sources d'information. Le cadre de discernement Ω est composé pour nos expériences des cinq segments que le contributeur peut sélectionner. Une contribution à une question q est modélisée par une fonction de masse à support simple m^{Ω}_{cq}, équation (5.1), avec X la réponse du contributeur c qui peut être composée de 1 à 5 segments. Lorsque la certitude du contributeur en sa réponse est demandée, expériences 1 et 3, une valeur numérique correspondante est utilisée comme masse ω. Le tableau 4.2 répertorie les valeurs numériques ω associées aux différents éléments de l'échelle de certitude proposée au contributeur. Pour l'expérience 2, pour laquelle la certitude du contributeur n'est pas connue, nous avons choisi d'utiliser pour toutes les contributions $\omega = 0.5$, ce qui correspond à une certitude neutre pour le contributeur : "Ni certain, ni incertain". Les valeurs de ω appartiennent à l'intervalle $[0.2, 0.8]$ afin de ne pas avoir de fonction de masse catégorique sur la réponse du contributeur. Les fonctions de masses sont ensuite agrégées pour chaque question afin d'obtenir les fonctions de masse m^{Ω}_{q}. Les fonctions m^{Ω}_{q} sont ensuite transformées en pro-

Δ_1 (mm)	0	0.3	0.6	0.9	1.2	1.5
MV	**0.13**	1.00	1.00	1.00	1.00	1.00
EM	**0.13**	1.00	1.00	1.00	1.00	1.00
BF (Moyenne)	**0.13**	1.00	1.00	1.00	1.00	1.00
BF (Conjonctive)	**0.13**	1.00	1.00	1.00	1.00	1.00
Δ_2 (mm)	0	0.3	0.6	1.2	1.8	2.4
MV	**0.27**	**0.93**	1.00	1.00	1.00	1.00
EM	**0.33**	1.00	1.00	1.00	1.00	1.00
BF (Moyenne)	**0.13**	**0.93**	1.00	1.00	1.00	1.00
BF (Conjonctive)	**0.13**	**0.93**	1.00	1.00	1.00	1.00

TABLE 4.3 – Taux de bonne réponse pour l'agrégation des contributions pour chaque valeur de δ pour l'**expérience 1**.

Δ_0 (mm)	0	0.3	0.6	0.9	1.2	1.8	2.4
MV	**0.15**	1.00	1.00	1.00	1.00	1.00	1.00
EM	**0.10**	1.00	1.00	1.00	1.00	1.00	1.00
BF (Moyenne)	**0.20**	1.00	1.00	1.00	1.00	1.00	1.00
BF (Conjonctive)	**0.10**	1	1.00	1.00	1.00	1.00	1.00

TABLE 4.4 – Taux de bonne réponse pour l'agrégation des contributions pour chaque valeur de δ pour l'**expérience 2**.

babilités pignistiques, le segment avec la probabilité la plus élevée est considéré comme le plus grand. La section suivante présente les résultats obtenus pour la comparaison des trois méthodes d'agrégation pour l'ensemble des expériences menées.

4.4.2 Comparaison des méthodes d'agrégation

Les résultats obtenus après agrégation pour chaque valeur de δ sont ensuite comparés aux réponses attendues afin de calculer pour chaque niveau de difficulté le taux de bonne réponse (TBR). Pour les fonctions de croyance, les résultats de l'agrégation des contribution par l'opérateur conjonctif, équation (3.16), sont comparées aux résultats obtenus par la moyenne, équation (3.15). Les TBR sont donnés pour les expériences 1 à 3 dans les tableaux 4.3, 4.4 et 4.5.

Ce TBR est toujours égal à 1 pour $\delta > 0.3$ pour les trois expériences, indépendamment de la méthode d'agrégation employée. Pour $\delta = 0.3$, le taux est égal à 1 pour toutes les méthodes d'agrégation pour les contributeurs ayant réalisé l'expérience avec Δ_0 et Δ_2.

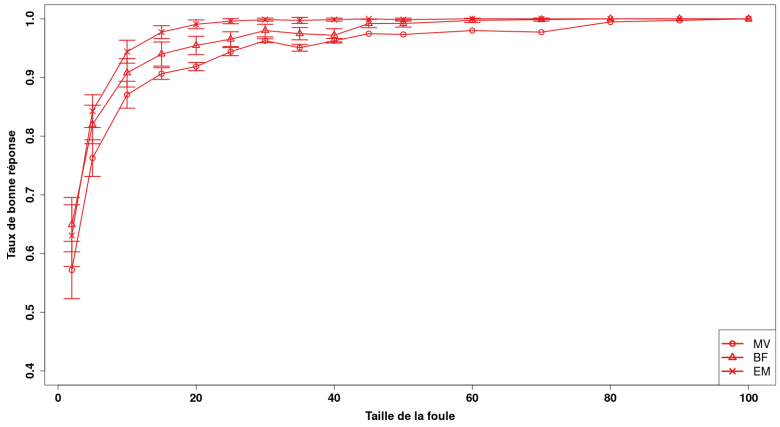
Δ_0 (mm)	0	0.3	0.6	0.9	1.2	1.8	2.4
MV	**0.2**	1.00	1.00	1.00	1.00	1.00	1.00
EM	**0.1**	1.00	1.00	1.00	1.00	1.00	1.00
BF (Moyenne)	**0.2**	1.00	1.00	1.00	1.00	1.00	1.00
BF (Conjonctive)	**0.2**	1.00	1.00	1.00	1.00	1.00	1.00

TABLE 4.5 – Taux de bonne réponse pour l'agrégation des contributions pour chaque valeur de δ pour l'**expérience 3**.

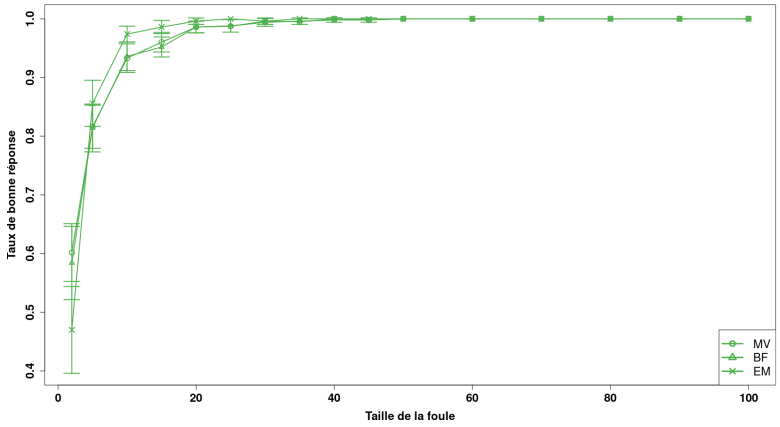
Le TBR est égal à 0.93 pour cette valeur de δ pour les agrégations par MV et fonction de croyance, il est de 1 pour l'algorithme EM. Lorsque les cinq segments sont de tailles égales, $\delta = 0$, le TBR est beaucoup plus faible que pour les autres valeurs de difficulté puisque le bon segment est indiscernable des témoins.

Afin d'observer l'impact du nombre de contributeurs sur la qualité des contributions, le TBR est calculé pour différentes tailles de foule pour $\delta = 0.3$ car il s'agit des questions les plus difficiles après $\delta = 0$. En effet, pour $\delta = 0$ les TBR seront toujours proches de la probabilité de choisir aléatoirement le bon segment, soit 0.2, alors que pour $\delta = 0.3$ le TBR peut atteindre une valeur de 1 pour l'intégralité de la foule d'après les tableaux 4.3, 4.4 et 4.5. De plus, cette valeur de δ est commune aux trois ensembles de difficulté Δ_0, Δ_1 et Δ_2. Les TBR obtenus pour les fonctions de croyance avec une agrégation par la moyenne et par l'opérateur conjonctif sont identiques pour toutes les valeurs de difficulté, excepté pour $\delta = 0$ de l'expérience 2 (tableau 4.4) où l'opérateur conjonctif procure un résultat légèrement plus faible que la moyenne. Pour l'agrégation des réponses pour différentes tailles de foule, nous avons choisi d'utiliser la moyenne des fonctions de masse car des tests complémentaires ont montré qu'il n'est pas toujours possible d'utiliser l'opérateur conjonctif lorsque les contributeurs sont trop en conflit.

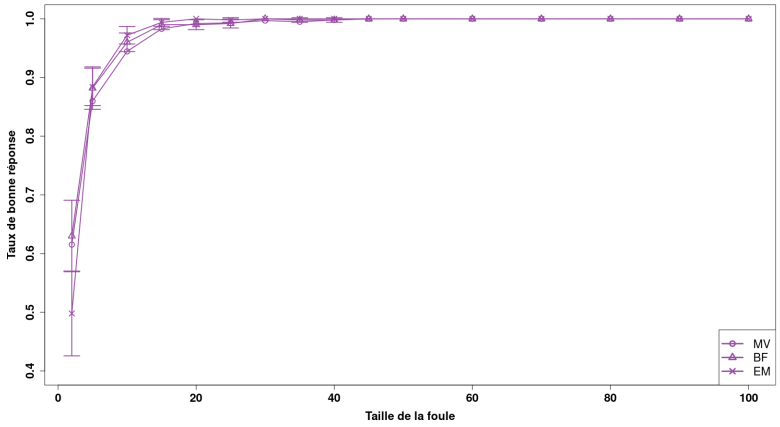
Les graphiques de la figure 4.11 présentent pour les expériences 1 à 3 l'évolution des TBR pour les trois méthodes d'agrégation pour une taille de foule T_f allant de 2 jusqu'à 100 contributeurs. Seules les questions pour lesquelles $\delta = 0.3$ sont utilisées. Les réponses des contributeurs de l'expérience 1 qui ont réalisé la campagne de *crowdsourcing* pour Δ_1 ne sont pas différenciées de celles de Δ_2. Pour réaliser les graphiques de la figure 4.11, un groupe de $T_f \in \{2, 4, 5, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55,\ 60, 70, 80, 90, 100\}$ contributeurs, choisis aléatoirement parmi les 100 contributeurs à l'expérience, est constitué. Les données sont agrégées pour cet ensemble de contributeurs et le TBR est calculé d'après les résultats obtenus après agrégation. Ce procédé de sélection des contributeurs, agrégation des données et calcul du TBR est répété 50 fois pour chaque valeur de T_f. Les



(a) Expérience 1



(b) Expérience 2



(c) Expérience 3

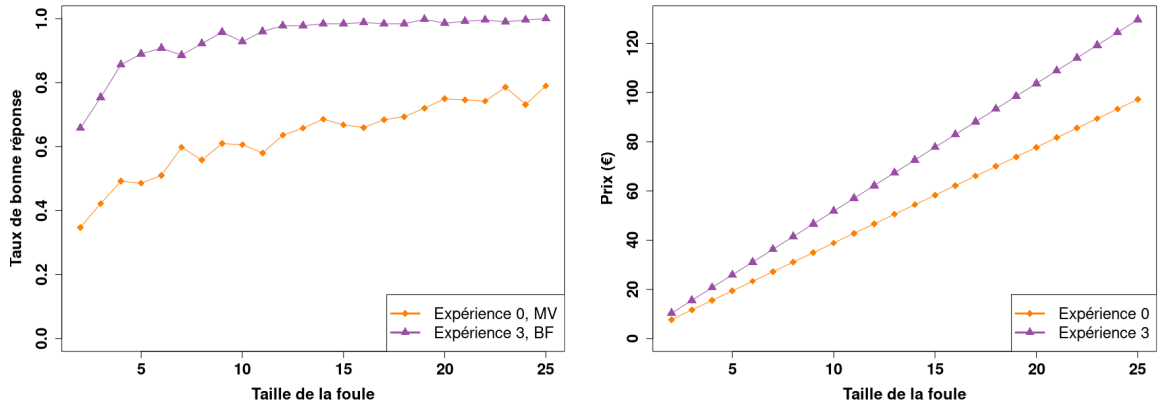
FIGURE 4.11 – Comparaison de l'évolution du taux de bonne réponse pour $\delta = 0.3mm$ pour une taille de foule croissante d'après 3 méthodes d'agrégation : MV, BF, EM.

graphiques 4.11a, 4.11b et 4.11c présentent pour chaque expérience et taille de foule T_f la moyenne des 50 TBR obtenus et les intervalles de confiance à 95%. Pour les trois graphiques et méthodes d'agrégation, les TBR moyens augmentent avec la taille de la foule jusqu'à la valeur maximale de 1. Le TBR est, de plus, toujours supérieur à 0.2, aussi l'estimation du plus grand segment par la foule ne repose pas sur la chance.

Une foule de 100 contributeurs n'est pas nécessaire pour avoir un TBR de 1. D'après les graphiques, l'interface et la méthode d'agrégation des données utilisées ont un impact sur le TBR et par conséquent sur la taille de la foule nécessaire. En effet, les TBR des expériences 2 et 3 avec imprécision sont supérieurs à ceux de l'expérience 1 où le contributeur ne peut sélectionner qu'un seul segment. Grâce aux interfaces permettant l'imprécision des réponses, la taille de la foule nécessaire afin d'avoir un TBR de 1 est inférieure à celle requise pour les interfaces traditionnelles qui contraignent le contributeur à être précis. En effet, pour l'expérience 3 (figure 4.11c), 45 contributeurs suffisent pour obtenir un TBR de 1 quelque soit la méthode d'agrégation utilisée alors qu'il en faut 90 avec l'expérience 1 (figure 4.11a).

Pour les trois expériences, l'algorithme EM est la méthode qui offre les meilleurs TBR, mais l'écart entre les courbes diminue entre chaque expérience. La modélisation et l'agrégation des réponses par la théorie des fonctions de croyance offre de meilleurs résultats que le MV, et pour l'expérience 3 avec imprécision et certitude où le modèle est des plus pertinents, les TBR obtenus sont très proches de ceux de EM.

L'interface de l'expérience 3 permettant au contributeur d'être imprécis tout en donnant sa certitude sur sa réponse est intéressante pour l'employeur car la taille de la foule requise pour réaliser la campagne est moindre que dans les interfaces traditionnelles qui ne considèrent pas l'imprécision. Néanmoins, le temps nécessaire au contributeur pour être imprécis et donner sa certitude (Expérience 3) est plus long que pour des réponses précises sans certitude demandée (Expérience 0) comme l'indique le tableau 4.1 page 67. Or le temps est une autre variable significative du coût de la campagne pour l'employeur, c'est pourquoi nous allons maintenant comparer les coûts des campagnes pour les expériences 0 et 3. La figure 4.12a compare les taux moyens de bonne réponse pour les contributions de l'expérience 3 modélisées par la théorie des fonctions de croyance et celles de l'expérience 0 pour laquelle le MV est utilisé. La méthode suivie pour obtenir cette figure est la même que pour la figure 4.11. Un ensemble de T_f contributeurs est choisi 50 fois et la moyenne des TBR est réalisée. Pour l'expérience 0, la foule est plus restreinte et composée



(a) Taux de bonne réponse pour une agrégation des réponses par fonction de croyance (expérience 3) et vote majoritaire (expérience 0).

(b) Comparaison des prix des campagnes de *crowdsourcing* des expériences 0 et 3 pour différentes tailles de foule.

FIGURE 4.12 – Comparaison des prix des expériences 0 et 3 avec les taux de bonne réponse obtenus pour $\delta = 0.3mm$.

de 25 contributeurs c'est la campagne témoin. La taille de la foule varie donc entre 2 et 25 contributeurs pour les graphiques de la figure 4.12. La figure 4.12b présente les coûts des campagnes de *crowdsourcing* associées aux expériences 0 et 3 en fonction du nombre de contributeurs T_f qui composent la foule. L'expérience 3 dure 16 minutes et coûte 5,18€ par contributeur à l'employeur contre 3,88€ pour 12 minutes pour l'expérience 0. L'expérience 0 est bien plus rapide à réaliser pour le contributeur et moins coûteuse pour l'employeur pour un même nombre de contributeurs comparée à l'expérience 3. Cependant, l'agrégation des données de l'expérience 0 avec le MV permet d'obtenir un TBR maximal de 0.79 pour l'intégralité de la foule, pour ce même nombre de contributeurs, le TBR est de 1 pour l'expérience 3. Pour les 25 contributeurs qui ont participé à l'expérience 0 la campagne de *crowdsourcing* a coûté 97.20€. En comparaison, 5 contributeurs de l'expérience 3 suffisent à obtenir un TBR de 0.856, ce qui coûte 20.74€ à l'employeur. Le contributeur met donc davantage de temps à répondre aux questions avec la possibilité d'être imprécis tout en donnant sa certitude, contrairement aux interfaces traditionnelles sans imprécision et certitude. En revanche, grâce à la modélisation des réponses par la théorie des fonctions de croyance, l'employeur peut faire appel à une foule plus petite pour un TBR identique voire meilleur.

Après avoir analysé dans cette section l'agrégation des contributions collectées, nous étudions dans la section suivante les retours des utilisateurs sur l'interface.

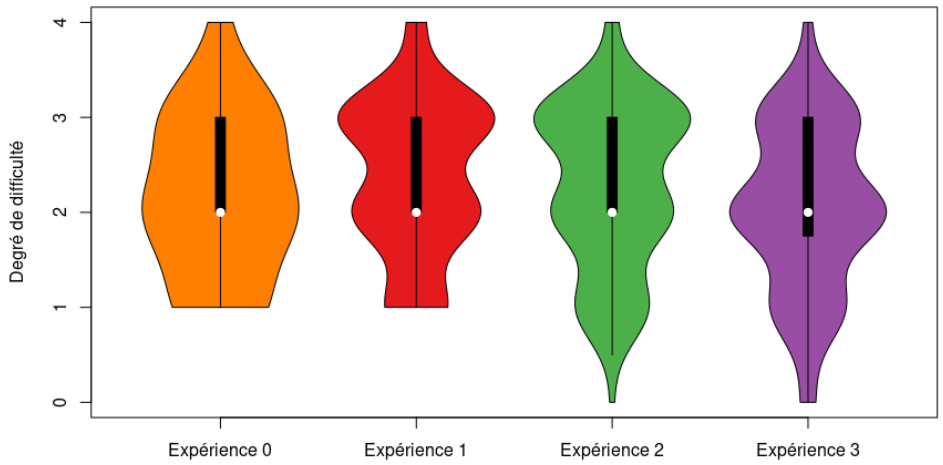


FIGURE 4.13 – Difficulté ressentie par le contributeur pour sélectionner le plus grand segment pour les quatre expériences.

4.5 Retour utilisateur

À la fin de chaque campagne de *crowdsourcing*, des questions sont posées au contributeur afin d'avoir un retour sur son expérience utilisateur vis-à-vis de l'interface employée. Le questionnaire utilisé pour l'expérience 3 avec incertitude et imprécision est donné en annexe 3 page 160. Dans cette section les réponses collectées aux questionnaires de fin sont analysées.

Pour les quatre expériences, il est demandé au contributeur la difficulté qu'il éprouve pour sélectionner le plus grand segment. Une échelle de Likert à cinq degrés lui est proposée : "Très facile" (0), "Facile" (1), "Ni facile ni difficile" (2), "Difficile" (3), "Très Difficile" (4). La figure 4.13 présente les diagrammes en violon associés aux réponses. Pour l'ensemble des campagnes, la majorité des contributeurs éprouve des difficultés à différencier le bon segment des quatre témoins. Pour les expériences 0 et 1 pour lesquelles il n'est pas possible pour le contributeur d'être imprécis, aucun contributeur n'a spécifié qu'il lui était "Très facile" (0) de choisir le plus grand segment, à la différence des expériences 2 et 3. Pour l'expérience 3, la proportion de contributeurs ayant indiqué que la tâche est "Difficile" (3) est plus faible que pour les trois autres expériences. De plus, il s'agit de la seule expérience pour laquelle le premier quartile est inférieur à la médiane, qui a une valeur de 2. Grâce à l'interface lui permettant d'être imprécis tout en indiquant sa certi-

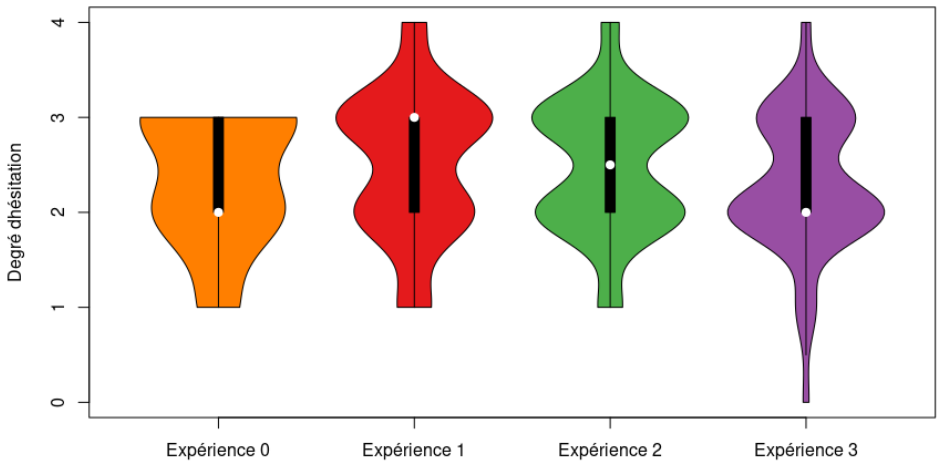
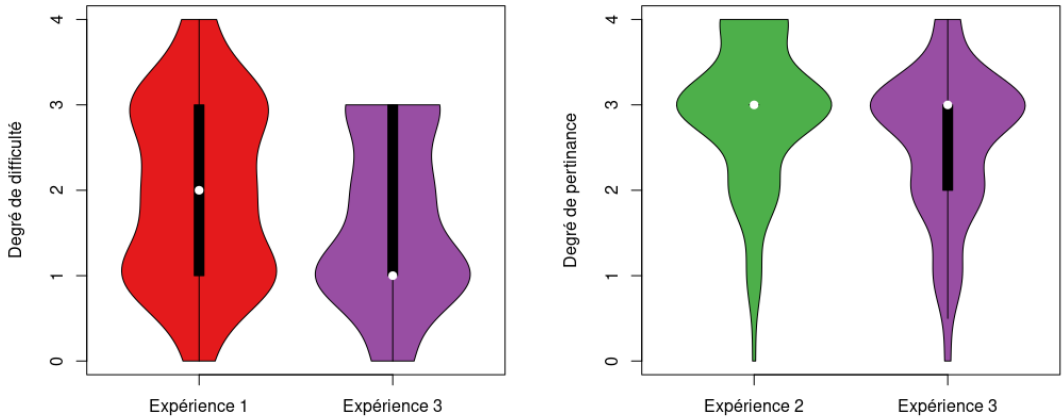


FIGURE 4.14 – Fréquence d'hésitation du contributeur pour les quatre expériences.

tude, le contributeur éprouve moins de difficultés à réaliser la tâche. Cette diminution de la difficulté de la tâche est un élément positif de l'amélioration des conditions de travail du contributeur et peut être un facteur de motivation pour lui.

Les expériences menées ont pour objectif d'amener le contributeur dans des situations d'indécision de difficultés variables afin d'étudier la capacité du contributeur à être imprécis et incertain. Dans le questionnaire final il est demandé au contributeur d'estimer sa capacité à répondre, à savoir s'il a hésité : "Jamais" (0), "Rarement" (1), "Occasionnellement" (2), "Souvent" (3), "Toujours" (4). Les diagrammes en violon obtenus d'après les réponses collectées sont donnés sur la figure 4.14. Pour les expériences 0, 1 et 2, les contributeurs affirment avoir au minimum hésité "Occasionnellement", ce qui est très positif puisqu'il s'agit d'un facteur important au cours de ce travail. Pour l'expérience 1, 7% de la foule hésite toujours sur le bon segment contre respectivement 5% et 3% pour les expériences 2 et 3. Aucun contributeur affirme être toujours hésitant dans ses réponses pour l'expérience 0 alors qu'il n'était pas possible d'être imprécis ou de donner sa certitude. Cependant, au vu des pourcentages obtenus, il devrait y avoir une moyenne de contributeurs entre 1.75 et 0.75 hésitants parmi la foule de l'expérience 0 de taille plus restreinte, ce qui explique l'absence de contributeurs toujours hésitants. Les contributeurs sont moins hésitants pour les expériences 2 et 3. L'imprécision permet donc de diminuer l'hésitation puisqu'en cas de doute ils peuvent sélectionner plusieurs segments. Ceci est



(a) Difficulté d'exprimer une certitude pour les foules des expériences 1 et 3.

(b) Pertinence de sélectionner plusieurs segments pour les expériences 2 et 3.

FIGURE 4.15 – Pertinence de l'imprécision d'après les contributeurs.

un élément positif pour l'employeur qui peut avoir davantage confiance dans les réponses de la foule.

Pour les expériences 1 et 3 avec certitude, le contributeur est questionné sur la difficulté ressentie à donner sa certitude, la même échelle de Likert que pour la difficulté de la tâche est utilisée. D'après les diagrammes en violon de la figure 4.15a, il est moins difficile pour la foule de donner sa certitude pour l'expérience 3 avec imprécision que pour l'expérience 1. En permettant au contributeur d'être imprécis, il est plus simple pour ce dernier d'auto-évaluer la certitude qu'il a en sa réponse, notamment parce qu'il est plus certain de cette dernière.

Finalement, pour les expériences 2 et 3 avec imprécision, il est demandé au contributeur s'il estime pertinent de pouvoir être imprécis, ainsi l'ensemble de réponses suivant est proposé : "Absolument pas pertinent" (0), "Non pertinent" (1), "Plus ou moins pertinent" (2), "Pertinent" (3), "Très pertinent" (4). La figure 4.15b présente les diagrammes en violon des réponses. Les deux diagrammes ont une répartition des contributeurs similaire et montrent que les contributeurs des deux expériences considèrent majoritairement qu'avoir la possibilité d'être imprécis est pertinent.

Les retours des contributeurs sont très positifs puisqu'ils montrent que l'interface leur permettant d'être imprécis tout en leur demandant leur certitude, diminue leur ressenti

de la difficulté de la tâche et leurs hésitations. Lorsqu'ils peuvent être imprécis, ils ont moins de difficulté à donner leur certitude et ils sont nombreux à trouver pertinent de pouvoir donner un ensemble de plusieurs réponses.

4.6 Conclusion

Dans ce travail, une interface de *crowdsourcing* offrant davantage de possibilités d'expression au contributeur a été mise au point. L'objectif de cette interface est de permettre la modélisation des imperfections des réponses inhérentes aux contributions humaines afin de les intégrer au processus final de fusion de l'information grâce à la théorie des fonctions de croyance. L'interface ainsi proposée offre au contributeur la possibilité d'être imprécis en cas de doute sur la réponse, tout en donnant sa certitude, ceci est dévellopé par Thierry et al. 2020a.

Afin de valider cette interface quatre expériences ont été réalisées, chacune donnant lieu à une campagne de *crowdsourcing*. Pour les quatre expériences une tâche reposant sur la perception visuelle est utilisée afin de ne pas introduire de biais de connaissance, d'avoir une maîtrise de la difficulté des questions et de connaître la bonne réponse. Un contributeur ayant participé à une campagne n'est pas autorisé à en réaliser une autre afin d'éviter tout biais d'apprentissage.

Les expériences réalisées ont montré que la difficulté de la question, et plus généralement de la campagne, a un impact sur la réponse du contributeur. En effet, la certitude de la foule augmente lorsque la difficulté de la question diminue et la certitude du contributeur est relative à la difficulté globale de la campagne. De plus, plus la question est difficile, plus le contributeur est imprécis dans sa réponse. Nos expériences nous ont notamment permis de valider l'hypothèse de Smets 1997 d'après laquelle plus un individu est imprécis plus il est certain de sa réponse et réciproquement, plus il est imprécis moins il est certain. Grâce à l'interface proposée, lorsque le contributeur hésite entre plusieurs réponses, il peut être imprécis et ainsi plus certain de sa réponse que s'il avait dû effectuer un choix unique. Cette hypothèse est également validée par Thierry et al. 2021 qui effectuent leur expérimentations sur une autre base de donnée provenant de *crowdsourcing*.

Au sein des plateformes de *crowdsourcing*, le vote majoritaire est la méthode d'agrégation des données la plus couramment employée. Elle consiste à accorder une voix à la réponse de chaque contributeur. Il est également possible d'utiliser l'algorithme EM pour l'agrégation des contributions. Les données collectées par la campagne de *crowdsourcing* de

l'expérience 3 intègrent de l'imprécision et de l'incertitude. La Théorie des fonctions de croyance est particulièrement adaptée pour modéliser ce type d'information et offrir des méthodes d'agrégation performantes. Nos travaux présentent une comparaison entre les TBR obtenus pour chacune des trois méthodes, MV, EM et fonctions de croyance, réalisée pour différentes tailles de foule pour les expériences 1 à 3. Les résultats obtenus montrent que TBR du MV est inférieur à ceux des fonctions de croyance et de EM qui demeure la méthode la plus efficace. Par ailleurs, plus la taille de la foule augmente, plus l'écart entre les taux de bonne réponse de ces trois méthodes diminue. De plus, il est important de souligner l'apport de l'incertitude et de l'imprécision dans les informations recueillies. En effet, quelque soit la méthode utilisée, les taux de bonne réponse sont plus élevés dans l'expérience 3 avec des écarts réduits. Bien que plus longues à réaliser, les campagnes avec imprécision et certitude restent plus intéressantes financièrement pour l'employeur. Elles nécessitent en effet une foule bien plus restreinte que les campagnes habituelles pour un taux de bonne réponse égal.

Finalement, les contributeurs considèrent que la campagne est moins difficile lorsqu'il leur est possible d'être imprécis et de donner leur certitude. Ils hésitent moins dans leurs réponses et trouvent pertinent d'avoir la possibilité d'être imprécis au besoin.

Chapitre 5

MONITOR

Résumé : La foule sur les plateformes de *crowdsourcing* est très diversifiée et inclut différents profils de contributeurs, certains apportant leurs contributions avec un sérieux variable. La diversité des profils génère ainsi des données de qualité inégale. Or, le vote majoritaire, qui est la méthode d'agrégation des réponses communément employée dans les plateformes, accorde une importance égale à chaque contribution. Pour palier ce problème, nous proposons une méthode, MONITOR, permettant d'estimer le profil du contributeur et d'agréger les données collectées en prenant en considération leurs imperfections éventuelles.

Sommaire

5.1 Introduction

Les contributeurs sur les plateformes de *crowdsourcing* sont nombreux et présentent des profils très diversifiés. La plupart font preuve de sérieux dans la réalisation de la tâche et possèdent les compétences requises pour le travail demandé. Il existe également au sein de la foule des personnes disposant d'une connaissance approfondie sur le domaine proposé. À l'opposé, quelques contributeurs peu scrupuleux ont un comportement malveillant, répondant aléatoirement et rapidement dans le but de réaliser le plus grand nombre de tâches sur une courte période afin de maximiser la gratification obtenue. La diversité des profils de contributeurs composant la foule induit des réponses de qualité inégale. Cette diversité est problématique, car la méthode d'agrégation la plus utilisée pour le traitement des données est le vote majoritaire qui accorde un poids égal à chaque contribution. C'est dans ce contexte qu'est défini MONITOR. Ce modèle permet l'estimation du profil du contributeur et une agrégation des réponses en conséquence. MONITOR est présenté dans les travaux de Thierry et al. 2018 ; Thierry et al. 2019 ; Thierry et al. 2020b.

MONITOR utilise la théorie des fonctions de croyance pour modéliser l'imprécision et l'incertitude inhérente aux contributions. Ceci est expliqué dans la section 6.5.1. Afin d'estimer le profil d'un contributeur, le modèle considère à la fois sa qualification, présentée section 5.3, et son comportement, défini section 5.4. Les différents profils déterminés par MONITOR ainsi que la méthode employée pour l'agrégation des réponses sont exposés dans la section 5.5.

5.2 Les motivations du modèle

MONITOR est l'acronyme de *MOdelling uNcertainty and Inaccuracy on daTa from crOwdsourcing platfoRms*, car ce modèle considère des réponses pour lesquelles le contributeur peut être imprécis tout en renseignant sa certitude. La théorie des fonctions de croyance est utilisée afin de modéliser l'imprécision et l'incertitude des contributions. Cette section revient dans un premier temps sur le type de données utilisées par MONITOR, puis dans un second temps, la modélisation des contributions est présentée. Enfin l'estimation du profil dans sa globalité est introduite.

Les données

Les données utilisées proviennent de campagnes de *crowdsourcing* constituées exclusivement de questions fermées. Ce type de campagne est principalement réalisé sur les plateformes d'activités routinières où les tâches consistent fréquemment à la réalisation de QCMs. Lorsqu'il réalise la tâche, le contributeur a la possibilité d'être imprécis en sélectionnant plusieurs réponses, il doit également renseigner sa certitude en sa contribution. Les réponses collectées de la sorte sont donc potentiellement imprécises avec un niveau de certitude variable. Le temps de réponse du contributeur à une question est sauvegardé, car il fait partie des données utilisées par MONITOR. Les campagnes de *crowdsourcing* incluent aussi une à plusieurs questions d'attention similaires à celles employées dans le chapitre 4 afin de s'assurer du sérieux du contributeur. Ces informations sont intégrées dans MONITOR.

La modélisation des réponses

Soit une question q, l'ensemble des réponses associées à q composent le cadre de discernement $\Omega_q = \{r_1, ..., r_K\}$. La question étant fermée, cette modélisation s'effectue en monde clos. Le contributeur c répond à la question q par la contribution $X \in 2^{\Omega_q}$, qui peut être imprécise, et à laquelle il associe une certitude de valeur ω_{cq}. Cette contribution est modélisée par une fonction de masse à support simple $(X^{\omega_{cq}})$:

$$\begin{cases} m_{cq}^{\Omega_q}(X) = \omega_{cq} \text{ avec } X \in 2^{\Omega_q} \setminus \Omega \\ m_{cq}^{\Omega_q}(\Omega_q) = 1 - \omega_{cq} \end{cases} \tag{5.1}$$

Cette modélisation caractérise le fait que c croit partiellement en sa réponse. Avant leur agrégation, les fonctions de masse $m_{cq}^{\Omega_q}$ sont affaiblies par un coefficient d'affaiblissement (équation (3.7)) en accord avec le profil du contributeur.

L'estimation du profil

La plupart des éléments de l'état de l'art se contentent d'évaluer la qualification du contributeur pour la tâche, c'est-à-dire d'estimer s'il a les compétences requises ou s'il est expert du domaine. Afin d'estimer la qualification du contributeur, MONITOR considère également l'imprécision et la certitude dans sa réponse. Plus les réponses du contributeur sont précises et certaines, plus il est qualifié pour la tâche.

Cependant, si la qualification d'un individu influence fortement la qualité de son travail, elle ne constitue pas la seule composante à prendre en considération. En effet, Mehmood et al. 2016 affirment dans leur étude que la personnalité d'un employé a un impact significatif sur ses performances professionnelles. La personnalité d'un individu est définie dans cette étude d'après le modèle des *Big Five* (section 2.5.2) et sa performance d'après cinq critères : la qualité du travail, la discipline et l'attention, la coopération, la responsabilité des résultats et la responsabilité des risques. D'après Mehmood et al. 2016 l'Ouverture à l'expérience, la Conscienciosité, l'Agréabilité et l'Extraversion ont un fort impact positif sur les performances professionelles de l'individu. Ceci est en accord avec les travaux de Kazai et al. 2012 où les auteurs affirment que l'Ouverture à l'expérience et la Conscienciosité du contributeur sont fortement liées à la justesse de ses réponses. Afin d'analyser la Consciensiosité du contributeur dans sa réalisation de la tâche, MONITOR considère son comportement à travers la réflexion et l'attention. En effet, plus un contributeur est attentif est réfléchis dans ses réponses, plus il est consciencieux dans la réalisation de son travail.

L'estimation conjointe de la qualification (section 5.3) et du comportement (5.4) du contributeur par MONITOR permet d'estimer son profil (section 5.5), ce qu'illustre le schéma de la figure 5.1. Pour un contributeur c, les fonctions de masse modélisant l'imprécision, la certitude, la réflexion et l'attention sont calculées pour chaque question q de la campagne. La combinaison individuelle de chacun de ces éléments est ensuite réalisée avant qu'une conversion des cadres de discernement n'aboutisse à l'obtention d'une unique fonction de masse qui permet ensuite d'estimer le profil du contributeur. L'ensemble de ces opérations sont détaillées dans les sections suivantes.

5.3 Qualification du contributeur

Dans les plateformes de *crowdsourcing*, l'employeur n'a généralement pas connaissance du niveau réel de qualification du contributeur vis-à-vis de la tâche. Il peut éventuellement évaluer ce niveau en faisant passer un test de qualification au contributeur avant que celui-ci ne participe à la campagne, mais cela nécessite des données d'or. Or, comme mentionné dans le chapitre 2, il n'est pas toujours possible d'avoir des données de référence suivant le type de tâche considéré. Différentes méthodes sont exposées dans l'état de l'art, section 2.6.3 pour l'estimation de l'expertise. Cependant, il existe peu de méthodes qui utilisent les fonctions de croyance pour estimer l'expertise du contributeur. Ouni et al. 2017 et Abassi

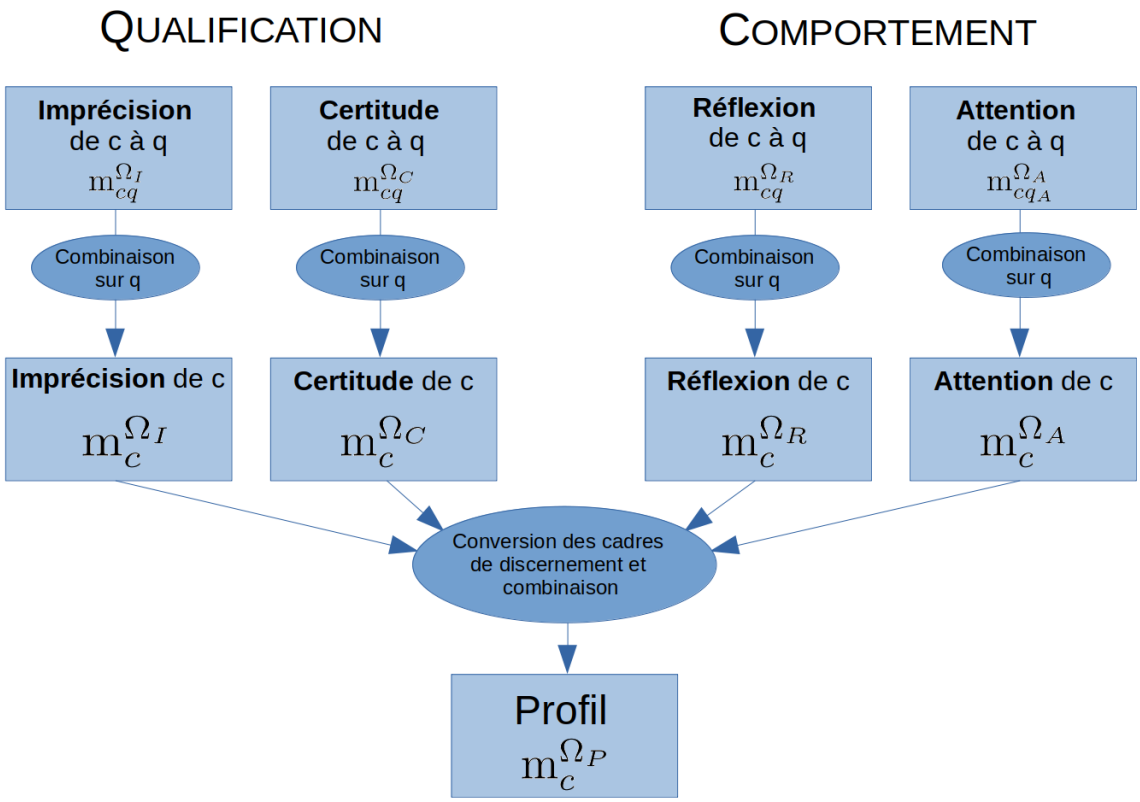


FIGURE 5.1 – Schéma de la méthode employée pour l'estimation du profil du contributeur.

et al. 2018 proposent deux approches différentes expliquées dans la section 3.6, mais les auteurs utilisent des données d'or et traitent des réponses précises. Or, dans le chapitre 4, nous avons montré qu'il est pertinent d'offrir au contributeur la possibilité d'être imprécis dans ses contributions. De plus, il n'est pas toujours possible pour l'employeur d'avoir des données d'or. C'est pour cette raison que nous avons proposé un modèle, MONITOR, qui n'y a pas recours. Ce même choix a été effectué dans les travaux de BEN RJAB et al. 2016 portant sur une mesure de l'expertise du contributeur en l'absence de données d'or et en considérant des réponses potentiellement imprécises.

Dans MONITOR la qualification définit une appréciation de la valeur professionnelle d'une personne suivant la certitude et l'imprécision de ses contributions. Nous abordons dans la section suivante l'estimation par MONITOR de l'imprécision globale du contributeur dans sa réalisation de la tâche.

5.3.1 Estimation de l'imprécision des réponses du contributeur

Les données utilisées proviennent de campagnes de *crowdsourcing* pour lesquelles le contributeur a eu la possibilité d'être imprécis s'il le souhaitait. L'imprécision dans MONITOR qualifie la capacité du contributeur à être précis P ou imprécis IP dans ses réponses. Le cadre de discernement associé est donc : $\Omega_I = \{P, IP\}$. Ainsi, plus un contributeur est qualifié pour la tâche plus ses réponses sont précises.

La fonction de masse définie sur 2^{Ω_I} pour un contributeur c renseignant la réponse $X \in 2^{\Omega_q}$ à la question q, est donnée par les équations suivantes :

$$\omega_{cq}^{I} = \frac{\log_2 |X|}{\log_2(imp_{MAX})} \tag{5.2}$$

$$\begin{cases} m_{cq}^{\Omega_I}(P) = \alpha^I * (1 - \omega_{cq}^{I}) \\ m_{cq}^{\Omega_I}(IP) = \alpha^I * \omega_{cq}^{I} \\ m_{cq}^{\Omega_I}(\Omega_I) = 1 - \alpha^I \end{cases} \tag{5.3}$$

$\alpha^I \in [0, 1]$ dans l'équation (5.3) est un coefficient d'affaiblissement et ω_{cq}^{I} la masse associée aux éléments de Ω_I. Le calcul de la masse ω_{cq}^{I} s'inspire du degré de précision DP_c de Ben Rjab et al. 2016 qui est rappelé section 3.6 par l'équation (3.33). Contrairement aux auteurs qui utilisent $\log_2 |\Omega_q|$, nous avons $\log_2(imp_{MAX})$ avec imp_{MAX} l'imprécision maximale que l'employeur autorise au contributeur ($imp_{MAX} \leq |\Omega_q|$). Plus le contributeur est imprécis plus ω_{cq}^{I} est élevée. Pour une réponse précise, X est un singleton et la masse $\omega_{cq}^{I} = 0$ grâce à la fonction logarithme et $m_{cq}^{\Omega_I}$ devient une fonction de masse à support simple avec pour élément focal P. Respectivement, si $X = \Omega_q$, $m_{cq}^{\Omega_I}$ est une fonction de masse à support simple dont l'élément focal est IP.

À la différence du degré DP_c de Ben Rjab et al. 2016, le quotient des logarithmes de l'équation (5.2) n'est pas multipliée par $m_{cq}^{\Omega_q}(X)$. En effet, cela peut éventuellement fausser le calcul de la masse sur l'imprécision, par exemple si le contributeur c est très imprécis et sélectionne toutes les réponses qui lui sont proposées, alors $X = \Omega_q$. Admettons que ce contributeur reste très incertain de sa réponse malgré tout, de sorte que $m_{cq}^{\Omega_q}(X) = 0$ alors $m_{cq}^{\Omega_q}(X)\frac{\log_2 |X|}{\log_2 |\Omega_q|} = 0$ ce qui signifierait que c est totalement précis alors que ça n'est pas le cas. Bien qu'il soit intéressant de tenir compte de $m_{cq}^{\Omega_q}(X)$ cela peut donc avoir un

impact très négatif sur l'estimation de la précision du contributeur.

Pour calculer l'expertise, Ben Rjab et al. 2016 utilisent à la fois le degré de précision DP_c et le degré d'exactitude sur la réponse du contributeur DE_c. Cependant, les expériences menées et exposées dans la section 6.4.2 montrent que le degré d'exactitude du contributeur n'est pas garant de sa bonne réponse. C'est pourquoi MONITOR ne considère pas l'exactitude des réponses pour estimer la qualification du contributeur, mais la certitude du contributeur.

5.3.2 Estimation de la certitude globale du contributeur

Un contributeur qualifié est sûr de sa réponse. Aussi, on peut considérer que plus il est qualifié, plus il est certain et plus l'employeur peut avoir confiance en sa contribution. Le cadre de discernement utilisé est le suivant : $\Omega_{\mathcal{C}} = \{C, IC\}$, C signifie que le contributeur est certain et IC à l'inverse, qu'il est incertain. Lorsque le contributeur c répond à la question q, il renseigne sa certitude ω_{cq} sur la justesse de sa contribution. La valeur de ω_{cq} est incluse dans l'intervalle $[\omega_{MIN}, \omega_{MAX}]$, avec $\omega_{MIN} < \omega_{MAX}$ pour obtenir une valeur de $\omega_{cq}^{\mathcal{C}}$ entre 0 et 1. La valeur ω_{MIN} signifie que c n'est pas certain de sa réponse. À l'opposé, pour $\omega_{cq} = \omega_{MAX}$ le contributeur est absolument sûr que sa réponse est correcte. La fonction de masse associée à la certitude sur 2^{Ω_C} est :

$$\omega_{cq}^{\mathcal{C}} = \frac{\omega_{cq} - \omega_{MIN}}{\omega_{MAX} - \omega_{MIN}} \tag{5.4}$$

$$\begin{cases} m_{cq}^{\Omega_C}(\mathcal{C}) = \alpha^{\mathcal{C}} * \omega_{cq}^{\mathcal{C}} \\ m_{cq}^{\Omega_C}(\mathcal{IC}) = \alpha^{\mathcal{C}} * (1 - \omega_{cq}^{\mathcal{C}}) \\ m_{cq}^{\Omega_C}(\Omega_{\mathcal{C}}) = 1 - \alpha^{\mathcal{C}} \end{cases} \tag{5.5}$$

Dans l'équation (5.5), $\alpha^{\mathcal{C}} \in [0, 1]$ est le coefficient d'affaiblissement de la fonction $m_{cq}^{\Omega_C}$. D'après l'équation (5.4), $\omega_{cq}^{\mathcal{C}}$ croît avec ω_{cq} renforçant la croyance que le contributeur c est certain C. Pour $\omega_{cq} = \omega_{MIN}$, $\omega_{cq}^{\mathcal{C}} = 0$ et $m_{cq}^{\Omega_C}(C) = 0$, ce qui fait de $m_{cq}^{\Omega_C}$ une fonction de masse à support simple d'élément focal IC.

Parallèlement à la qualification du contributeur, son comportement est estimé afin de déterminer son profil. La section suivante introduit cette nouvelle modélisation.

5.4 Comportement du contributeur

D'après Kazai et al. 2012 et Mehmood et al. 2016, l'Ouverture à l'expérience, la Conscienciosité et l'Agréabilité sont les trois traits de personnalité corrélés à des contributions de bonne qualité (section 2.5.2).

MONITOR estime la Consciensiosité du contributeur d'après son comportement en modélisant sa réflexion et son attention. En effet, une personne qui prend le temps de la réflexion est consciencieuse dans la réalisation de la tâche et renseignera des réponses en lesquelles l'employeur peut avoir confiance. À l'inverse, quelqu'un qui répond rapidement avec, par conséquent, un temps de réflexion très court, peut s'avérer être un mauvais contributeur comme un expert. Un mauvais contributeur répond précipitamment et sans réfléchir, car ses réponses sont aléatoires. À l'opposé, l'expert répond prestement, car il a besoin d'un temps de réflexion moindre comparé au reste de la foule en raison de sa meilleure connaissance sur le domaine. L'estimation de l'attention est alors nécessaire pour différencier ces deux profils, car le mauvais contributeur n'est pas attentif à la tâche contrairement à l'expert.

Nous présentons dans les sections suivantes l'estimation de la réflexion et de l'attention du contributeur par MONITOR.

5.4.1 Temps de réflexion pris par le contributeur

Difallah et al. 2012 spécifient que le temps de réponse est un bon indicateur de contributions aléatoires. De plus, Gadiraju et al. 2015 mettent en évidence dans leur annexe la corrélation entre le taux de bonne réponse des contributeurs et leur temps de réponse. C'est pour ces raisons que la modélisation de la réflexion par MONITOR repose sur l'utilisation du temps de réponse T_{cq} du contributeur c à la question q.

Nous faisons l'hypothèse que la question q nécessite un temps de réponse minimum T_{0q}. La valeur de T_{0q} est estimée de façon différente celon la campagne de *crowdsourcing*. Si $T_{cq} < T_{0q}$, cela indique que le contributeur n'a pas pris le temps minimum requis pour réfléchir à la question. Il est alors considéré comme non réfléchi dans sa contribution (NR). A contrario, pour $T_{cq} \geq T_{0q}$ le contributeur est réfléchi (R). Le cadre de discernement utilisé pour la réflexion est donc $\Omega_R = \{R, NR\}$, la fonction de masse modélisant la

réflexion sur 2^{Ω_R} est définie par :

$$\begin{cases} x = \dfrac{T_{cq} - T_{0q}}{T_{0q}} \\ \\ \omega_{cq}^R = \dfrac{\arctan(x)}{\pi} + \dfrac{1}{2} \end{cases} \tag{5.6}$$

$$\begin{cases} m_{cq}^{\Omega_R}(R) = \alpha^R \omega_{cq}^R \\ \\ m_{cq}^{\Omega_R}(NR) = \alpha^R(1 - \omega_{cq}^R) \\ \\ m_{cq}^{\Omega_R}(\Omega_R) = 1 - \alpha^R \end{cases} \tag{5.7}$$

Dans l'équation (5.7), $\alpha_R \in [0, 1]$ est un coefficient d'affaiblissement. La valeur $x \in [-1, +\infty[$ de l'équation (5.6) n'est pas à valeurs dans $[0, 1]$ car elle est négative si $T_{cq} < T_{0q}$ et supérieure à 1 pour $T_{cq} > 2 * T_{0q}$. La fonction arctangeante est utilisée dans le calcul de ω_{cq}^R afin de se ramener à des valeurs de l'intervalle $[0, 1]$. Cette fonction est choisie car le temps de réponse d'un contributeur à une question peut beaucoup varier au sein de la foule pour une même question. Cependant, lorsque ce temps de réponse est strictement supérieur à $2 * T_{0q}$ la masse sur l'élément R doit être proche de 1 ce qui est possible grâce aux asymptotes d'arctangeante.

5.4.2 Attention du contributeur lors de la campagne

Généralement, des questions d'attention sont posées au contributeur au cours des campagnes de *crowdsourcing*afin de s'assurer de son sérieux. MONITOR utilise ces questions pour estimer l'attention du contributeur. Cependant, les questions d'attention q_A doivent être d'un type spécifique. Il s'agit de reposer au contributeur la question q qui précède q_A et de lui demander de renseigner exactement les mêmes réponses. Si le contributeur est attentif (A), il se souvient de ses anciennes réponses. Dans le cas contraire (NA), celles-ci sont différentes. Il est ainsi possible d'estimer l'attention du contributeur en calculant la proximité de la réponse d'origine à la réponse de la question d'attention par une distance. Dans le cas où les réponses sont modélisées par des fonctions de masse, la distance de Jousselme et al. 2001 d_J (équation (3.12)) peut être utilisée car elle tient compte de la cardinalité de la réponse (imprécision) et des valeurs des masses (certitude).

La réponse $X \in 2^{\Omega_q}$ d'un contributeur c à la question q est modélisée de manière analogue par une fonction de masse à support simple $X^{\omega_{cq}}$ donnée par l'équation (5.1). La réponse $Y \in 2^{\Omega_q}$ à la question d'attention q_A, qui reprend la question q, est modélisée de manière analogue par $Y^{\theta_{cq}}$, avec $\theta_{cq} \in [0, 1]$ la certitude donnée à q_A. La fonction de masse associée à l'attention est donnée par :

$$\omega_{cq}^{A} = d_J(X^{\omega_{cq}}, Y^{\theta_{cq}}) \tag{5.8}$$

$$\begin{cases} m_{cq_A}^{\Omega_A}(A) = \alpha^A * (1 - \omega_{cq}^A) \\ m_{cq_A}^{\Omega_A}(NA) = \alpha^A * \omega_{cq}^A \\ m_{cq_A}^{\Omega_A}(\Omega_A) = 1 - \alpha^A \end{cases} \tag{5.9}$$

Dans l'équation (5.8), d_J est la distance de Jousselme et al. 2001 entre $X^{\omega_{cq}}$ et Y_{cq}^{θ}. Si $X = Y$ et $\omega = \theta$, alors les contributions sont identiques et la distance est nulle. $\omega_{cq}^A = 1$ et $m_{cq_A}^{\Omega_A}$ devient une fonction de masse à support simple d'élément focal A. Ainsi, en utilisant d_J, plus la réponse et la certitude données à la question d'attention sont proches de celles données à la question originelle, plus la masse accordée à l'élément A est importante.

Les fonctions de masse $m_{cq}^{\Omega_I}$, $m_{cq}^{\Omega_C}$, $m_{cq}^{\Omega_R}$ et $m_{cq_A}^{\Omega_A}$ sont calculées pour chaque question q auxquelles a répondu le contributeur c. Ces fonctions sont ensuite combinées sur leur cadre de discernement respectif et sur l'ensemble de la campagne afin d'obtenir pour c, son imprécision $m_c^{\Omega_I}$, sa certitude $m_c^{\Omega_C}$, sa réflexion $m_c^{\Omega_R}$ et son attention $m_c^{\Omega_A}$. Une fois ces quatre éléments obtenus, il est possible d'estimer le profil du contributeur comme le montre le schéma 5.1 page 91.

5.5 Profil et agrégation

La foule se compose de nombreux contributeurs au profil varié, si bien que plusieurs auteurs, présentés dans les sections 2.6.3 et 3.6, ont proposé différents ensembles de profils. Dans le cadre de cette thèse, quatre profils de contributeur sont définis et estimés par MONITOR. Ils constituent le cadre de discernement $\Omega_P = \{Expert, Bon, Moyen, Mauvais\}$. Nous avons choisi ces quatre profils, car nous estimons qu'un contributeur peut être ex-

cellent, correct ou médiocre dans sa réalisation de la tâche. Un contributeur qui excelle dans la réalisation de la tâche est expert, un contributeur dont les réponses sont correctes est bon. Et finalement ceux dont les contributions médiocres sont des contributeurs moyen ou mauvais, cela se différencie de leur sérieux dans leur travail. En effet, il est à notre sens dommage de pénaliser de la même façon un contributeur qui donne de mauvaises réponses, mais en étant consciencieux et un contributeur qui donne de mauvaises réponses, car il n'est pas attentif. L'on pourrait au contraire envisager d'apporter une aide au contributeur moyen pour qu'il s'améliore dans la réalisation de la tâche. C'est pourquoi la distinction est faite entre le contributeur moyen et le mauvais.

L'expert

Ce contributeur possède d'excellentes connaissances sur le domaine de la tâche. Il est donc plus qualifié que le reste de la foule. Cette qualification supérieure se caractérise par des réponses précises et certaines. Il réalise plus rapidement la tâche que la majorité des contributeurs car ses réponses sont instinctives et par conséquent non-réfléchies, mais il n'en reste pas moins attentif à son travail.

Le bon contributeur

Il dispose de connaissances moindres que l'expert, ce qui peut induire chez lui un doute sur la réponse à renseigner. En cas d'hésitation, le bon contributeur est imprécis dans sa contribution afin d'être sûr de celle-ci. Il prend le temps qui lui est nécessaire pour réfléchir à sa réponse et est attentif à son travail.

Le contributeur moyen

Il possède lui aussi des connaissances plus limitées que l'expert sur le domaine de la tâche. Ceci l'amène parfois, comme le bon contributeur, à hésiter sur la réponse à sélectionner et à être ainsi imprécis. Il est réfléchi et attentif dans son travail. Cependant, à la différence du bon contributeur, cette imprécision renforce peu sa certitude en sa sélection.

Le mauvais contributeur

Cet individu ne manque pas nécessairement de qualification pour la tâche, mais il s'en désintéresse et répond au plus vite afin de finir au plus tôt la campagne. Il fait partie

Caractéristique	**Cadre de discernement**	**Élément**	**Conversion**
Imprécision	Ω_I	P	$\{Expert, Mauvais\}$
		IP	$\{Bon, Moyen\}$
Certitude	Ω_C	C	$\{Expert, Bon, Mauvais\}$
		IC	$\{Moyen\}$
Réflexion	Ω_R	R	$\{Bon, Moyen\}$
		NR	$\{Expert, Mauvais\}$
Attention	Ω_A	A	$\{Expert, Bon, Moyen\}$
		NA	$\{Mauvais\}$

TABLE 5.1 – Conversion des cadres de discernements Ω_I, Ω_C, Ω_R, Ω_A

des rares contributeurs au profil malveillant attiré uniquement par l'appât du gain et n'étant pas consciencieux dans la réalisation de son travail. Ses réponses sont toujours précises puisque cela lui permet de ne pas perdre de temps. Elles sont certaines, car ils souhaitent compenser leur manque de sérieux auprès de l'employeur par leur assurance ou encore parce qu'ils n'ont pas une bonne perception de leurs réelles aptitudes pour la tâche. Néanmoins, il est possible de le différencier de l'expert grâce à son comportement. En effet, bien que le mauvais contributeur réponde vite et de façon non réfléchie comme l'expert, il est non attentif car ses contributions rapides sont souvent aléatoires. Il est ainsi plus difficile pour lui de se souvenir de ses réponses lors des questions d'attention.

Agrégation

Une fois les fonctions de masse $m_c^{\Omega_I}$, $m_c^{\Omega_C}$, $m_c^{\Omega_R}$, $m_c^{\Omega_A}$ calculées, une conversion des cadres de discernement Ω_I, Ω_C, Ω_R, Ω_A est réalisée pour se ramener au cadre de discernement sur le profil Ω_P. Le tableau 5.1 résume la façon dont sont convertis les éléments des différents cadres de discernement.

Ainsi, une personne qui est estimée précise P par MONITOR peut-être un $Expert$ ou un $Mauvais$ contributeur. A l'inverse si elle est imprécise IP il s'agit d'un contributeur Bon ou $Moyen$. Grâce à la conversion des cadres de discernement, il est ensuite possible de combiner les fonctions de masse de la qualification et du comportement sur Ω_P. La combinaison permet d'obtenir une unique fonction de masse $m_c^{\Omega_P}$ pour c :

$$m_c^{\Omega_P} = \frac{\alpha_I m_c^{\Omega_I \to \Omega_P} + \alpha_C m_c^{\Omega_C \to \Omega_P} + \alpha_R m_c^{\Omega_R \to \Omega_P} + \alpha_A m_c^{\Omega_A \to \Omega_P}}{\alpha_I + \alpha_C + \alpha_R + \alpha_A} \tag{5.10}$$

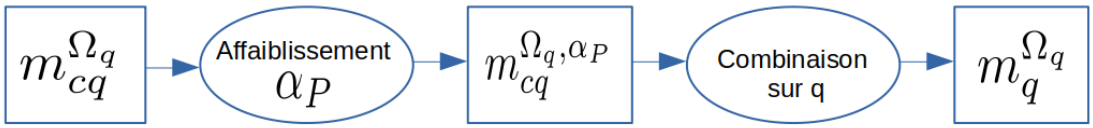


FIGURE 5.2 – Schéma de l'agrégation des contributions.

Dans l'équation (5.10), $\Omega_x \rightarrow \Omega_P$ symbolise la conversion du cadre de discernement Ω_x sur Ω_P fournie par le tableau 5.1. Les coefficients α_x sont utilisés pour moduler le poids accordé à chaque élément qui définit le profil du contributeur. Cette fonction est ensuite transformée en probabilité pignistique afin de prendre une décision sur le profil du contributeur.

En conséquence, grâce à l'estimation du profil du contributeur, ses réponses sont traitées en vue de leur agrégation, schématisée sur la figure 5.2. La réponse du contributeur c à la question q est modélisée par la fonction de masse à support simple $m_{cq}^{\Omega_q}$ donnée par l'équation (5.1) page 89. Cette fonction est par la suite affaiblie d'une valeur $\alpha_P \in [0, 1]$ suivant le profil du contributeur. L'objectif de cette démarche est d'accorder plus de poids aux réponses des contributeurs qualifiés et consciencieux. Ainsi, pour un expert α_P sera proche de 1. À l'opposé, un mauvais contributeur recevra une valeur α_P proche de 0, de sorte que ses contributions n'impactent pas négativement la qualité des données après l'agrégation. Les fonctions de masse affaiblies $m_{cq}^{\Omega_q,\alpha_P}$ sont combinées pour chaque question q pour l'ensemble de la foule : $m_q^{\Omega_q}$. Les fonctions $m_q^{\Omega_q}$ sont finalement transformées en probabilités pignistiqueS afin de prendre une décision sur la réponse.

5.6 Conclusion

La diversité des profils des contributeurs qui composent la foule a une incidence sur la qualité des données collectées. Afin de pallier ce problème, MONITOR est conçu pour proposer une estimation du profil du contributeur, à partir de sa qualification et de son comportement. La qualification représente l'imprécision du contributeur et sa certitude. Plus un contributeur est précis et certain, plus il est qualifié pour la tâche. Pour le comportement, le temps de réflexion pris pour réaliser la tâche et les réponses aux questions d'attention sont analysés. Un bon contributeur prend le temps de réfléchir à sa réponse et est attentif. L'imprécision, la certitude, la réflexion et l'attention sont modélisées par des fonctions de masse avec les cadres de discernement respectifs : Ω_I, Ω_C, Ω_R, Ω_A. Une

conversion des cadres de discernement est réalisée pour se ramener à Ω_P qui inclut les quatre types de profils définis pour MONITOR, à savoir : l'expert, le bon contributeur, le contributeur moyen et le mauvais contributeur. Cette conversion des cadres de discernement permet une agrégation des fonctions de masse associées à la qualification et au comportement du contributeur pour obtenir son profil. Les réponses d'un contributeur sont finalement affaiblies par un coefficient en accord avec son profil afin de donner plus de poids aux réponses d'un expert comparées à celles d'un mauvais contributeur. Les réponses affaiblies sont finalement agrégées pour chaque question.

MONITOR est un modèle complexe établi pour l'estimation du profil du contributeur et l'agrégation des réponses, il est présenté par Thierry et al. 2018 ; Thierry et al. 2019 ; Thierry et al. 2020b. Plusieurs campagnes de *crowdsourcing* ont été réalisées afin de collecter les données nécessaires pour tester le modèle. Le chapitre suivant présente ces campagnes et les expérimentations réalisées pour la validation de MONITOR.

Chapitre 6

Expériences et résultats

Résumé : Des campagnes de *crowdsourcing* ont été réalisées afin de collecter les données nécessaires pour tester MONITOR. Ce chapitre présente les campagnes menées et les expériences effectuées sur les différents éléments qui permettent l'estimation du profil du contributeur. L'agrégation des réponses par la théorie des fonctions de croyance est également comparée ici au vote majoritaire qui est traditionnellement utilisé et à l'algorithme EM.

Sommaire

6.1 Introduction

MONITOR modélise l'imprécision et la certitude de réponses de contributeurs provenant de plateformes de *crowdsourcing* grâce à la théorie des fonctions de croyance. Des contributions imprécises pour lesquelles le contributeur a spécifié sa certitude sont nécessaires afin de tester le modèle. Cependant, dans les plateformes de *crowdsourcing* les réponses sont généralement précises et le contributeur ne peut pas renseigner sa confiance en sa réponse. C'est pourquoi une interface offrant davantage de capacité d'expression, décrite dans le chapitre 4, a été définie au cours de cette thèse. Cette interface a par la suite été utilisée afin de collecter les données nécessaires pour tester MONITOR. Le modèle peut être décomposé en 3 étapes distinctes, le calcul des éléments qui composent la qualification et le comportement du contributeur, l'estimation du profil et l'agrégation des réponses affaiblies par le profil. Les expériences menées pour chacune de ces étapes sont explicitées dans ce chapitre. De plus, une comparaison de MONITOR avec d'autres approches de l'état de l'art est également effectuée.

La section 6.2 présente les données utilisées pour les expériences. La section 6.3 réalise une comparaison de l'estimation de la précision du contributeur et sa réflexion lors de la tâche avec des approches existantes. Les sections 6.4 et 6.5 exposent respectivement les expériences réalisées pour l'estimation du profil du contributeur par MONITOR et l'agrégation des réponses. Finalement la section 6.6 conclut ce chapitre.

6.2 Acquisition de données réelles

Les données présentées dans la section 4.2.2 ont permis de valider l'interface de *crowdsourcing* proposée pour la collecte de réponses imparfaites. Néanmoins la tâche utilisée est artificielle et éloignée du type de tâche rencontré d'ordinaire dans les plateformes de *crowdsourcing*. En effet, cette tâche reposait sur la perception visuelle afin qu'il n'y ait pas de biais de connaissance de la part du contributeur. Cependant, puisque MONITOR permet de déceler des niveaux de qualification différents chez les contributeurs, afin de tester convenablement le modèle les réponses doivent provenir d'une tâche pour laquelle les contributeurs n'ont pas tous une qualification égale. Ainsi un nouveau type de tâche consistant en l'annotation de photos d'oiseaux a été défini et a donné lieu à cinq campagnes de *crowdsourcing* différentes, résumées dans le tableau 6.1 et explicitées ici.

Campagne	Réponses	Taille de la foule	Nombre de réponses
multi_oiseaux_précis	Précises	100	5000
multi_oiseaux_imprécis	Imprécises	100	5000
10_oiseaux_précis	Précises	50	2500
10_oiseaux_imprécis	Imprécises	50	2500
10_oiseaux_dynamique	Imprécises	51	2990

TABLE 6.1 – Récapitulatif des campagnes de *crowdsourcing* réalisées

Campagnes initiales

Deux premières campagnes de *crowdsourcing* ont été réalisées dans un premier temps et consistaient pour le contributeur à trouver le nom de l'espèce d'oiseau correspondant à la photo qui lui était présentée parmi 5 espèces proposées. Nous nous sommes efforcés dans l'élaboration des questions d'introduire des niveaux de difficulté différents. Par exemple, pour une question difficile une photo d'aigle est présentée au contributeur et les cinq éléments de réponses sont des espèces d'aigles différentes. A l'inverse, pour une question plus simple une photo de goéland est présentée au contributeur et les quatre autres réponses sont des noms d'espèces de canard. Pour une même photo, les réponses étaient présentées dans un ordre aléatoire afin d'éviter les biais de sélection. De plus, les questions étaient également posées dans un ordre aléatoire aux différents contributeurs, de sorte que lorsqu'un contributeur c répond à une question q un second contributeur c' répond lui à la question q'.

La plateforme Wirk (Crowdpanel) est utilisée pour réaliser les campagnes de *crowdsourcing*. Les utilisateurs de la plateforme résidant en France, les oiseaux utilisés pour la campagne sont tous d'espèces visibles en France métropolitaine. Ces campagnes de *crowdsourcing* incluent 3 questions d'attention pour lesquelles il est demandé au contributeur de redonner la même réponse que celle renseignée à la question précédente. Dans les deux campagnes, le contributeur doit donner sa réponse, la valider, puis spécifier sa certitude d'après l'échelle de Likert suivante : "Totalement incertain","Incertain", "Plutôt incertain", "Neutre", "Plutôt certain", "Certain", "Totalement certain". Finalement après avoir donné sa réponse et sa certitude il peut valider sa contribution afin de passer à la question suivante. Pour la première campagne (multi_oiseaux_précis) le contributeur doit donner une réponse précise en sélectionnant un unique nom d'oiseau. Pour la seconde campagne (multi_oiseaux_imprécis) le contributeur peut être imprécis et sélectionner jusqu'à l'intégralité des noms d'oiseaux qui lui sont proposés. Les interfaces utilisées pour

ces campagnes sont présentées par les figures 7.1 et 7.2 en Annexe 2. Les foules qui ont contribué à ces deux campagnes sont composées de 100 contributeurs, et un contributeur autorisé à faire la première campagne ne peut pas participer à la seconde. Chaque contributeur doit annoter 50 photos d'oiseaux, ce qui fait un total de 5000 réponses pour chacune de ces campagnes.

Ces données sont très utiles pour tester MONITOR, cependant, les 50 espèces d'oiseaux à identifier sont toutes distinctes. De plus, parmi les quatre réponses additionnelles proposées à une question, certains noms d'oiseau ne font pas partie des photos présentées dans les autres questions. Il n'est ainsi pas possible de réaliser une matrice de confusion sur les réponses et par conséquent d'appliquer l'algorithme EM. C'est pourquoi des campagnes de *crowdsourcing* complémentaires ont été réalisées afin de collecter les données nécessaires pour une comparaison de MONITOR avec l'algorithme EM.

Campagnes avec 10 oiseaux

Pour ces nouvelles campagnes, dix espèces d'oiseaux sont choisies et sont proposées comme éléments de réponse aux contributeurs tout au long de la campagne. Afin d'observer la capacité du contributeur à être imprécis en cas d'hésitation, les dix oiseaux présentés sont composés de sous-groupes provenant de même famille d'oiseaux. Les familles et espèces d'oiseaux choisies pour ces campagnes sont données dans le tableau 6.2. Ces dix noms sont présentés à chaque contributeur dans un ordre différent afin d'éviter un éventuel biais de sélection. Cet ordonnancement des noms est tout de même fixe pour un contributeur tout au long de la campagne. Tout comme pour les campagnes multi_oiseaux_précis et multi_oiseaux_imprécis les questions sont posées dans un ordre aléatoire. La même échelle est utilisée pour la certitude et 3 questions d'attention sont également posées. Il n'est cependant cette fois plus nécessaire de valider la réponse pour le contributeur avant de pouvoir donner sa certitude.

Les campagnes de *crowdsourcing* ont été réalisées sur la plateforme Crowdpanel et chaque campagne a requis la participation de 50 contributeurs. La taille de la foule est diminuée de moitié car, comme il est montré dans les tests effectués dans la suite de ce chapitre, une foule de 50 contributeurs est suffisante pour obtenir un taux de bonne réponse élevé après agrégation des données. Comme pour les autres campagnes, un contributeur ayant participé à une expérience ne peut pas participer à une autre. Pour chacune des dix espèces d'oiseaux qui composent l'ensemble de réponses proposées, 5 photos d'un oiseau

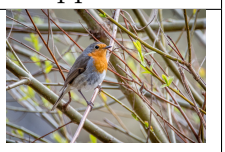
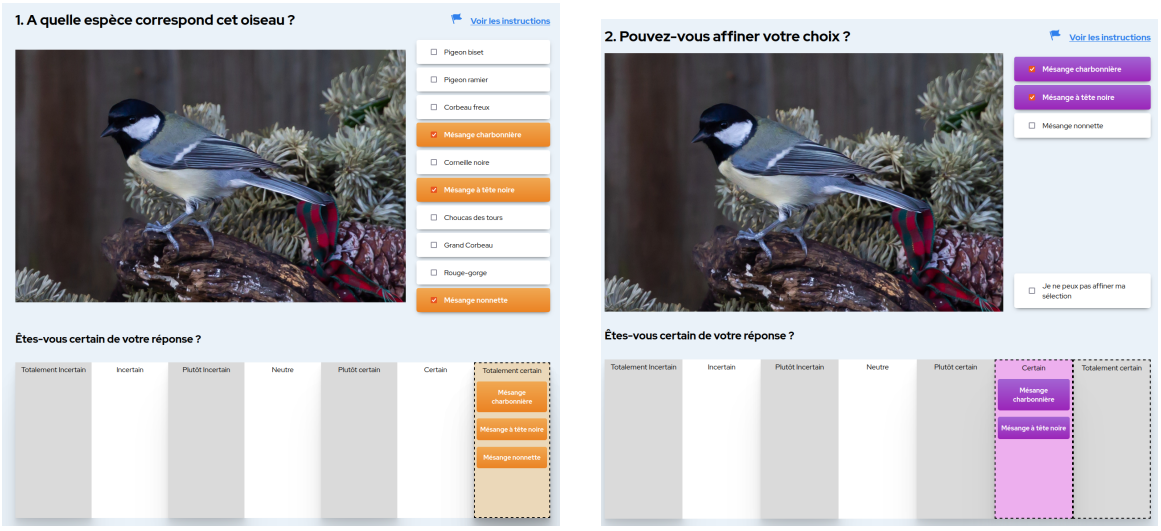
Famille	Nom commun	Apparence
Muscicapidae	Rouge-gorge familier	
Columbidae	Pigeon ramier	
	Pigeon biset	
Paridae	Mésange charbonnière	
	Mésange nonette	
	Mésange à tête noire	
Corvidae	Choucas des tours	
	Corneille noire	
	Grand Corbeau	
	Corbeau freux	

TABLE 6.2 – Les dix espèces d'oiseaux utilisées dans les trois dernières campagnes de *crowdsourcing*



(a) Étape 1 : première sélection (b) Étape 2 : affinement de la sélection

FIGURE 6.1 – Interface utilisée pour la campagne 10_oiseaux_dynamique dans le cas d'une réponse imprécise

sont présentées au contributeur, de sorte que ce dernier répond à 50 questions. Ainsi 2500 données sont collectées pour les expériences 10_oiseaux_précis, qui requière des réponses précises, et 10_oiseaux_imprécis, pour laquelle le contributeur peut choisir jusqu'à cinq noms d'oiseau. Le nombre de données est différent pour 10_oiseaux_dynamique car pour cette expérience, le contributeur peut être imprécis et sélectionner au maximum cinq noms d'oiseau. Dans le cas où le contributeur est imprécis (figure 6.1a), il lui est demandé dans un second temps s'il est capable de restreindre son choix de réponse tout en redonnant sa nouvelle certitude (figure 6.1). Lorsqu'il lui est offert de restreindre sa sélection, seules les éléments de réponses précédemment choisi lui sont de nouveau proposé. A l'inverse, s'il est précis mais pas "totalement certain" de sa réponse, il lui est proposé d'élargir sa sélection s'il en éprouve le besoin. Dans ce cas, la première réponse choisit est conservée lors de l'étape 2 et il peut la compléter en sélectionnant de nouveaux noms. Ces interactions avec le contributeur ont accru le nombre de réponses collectées, et donc le temps de sollicitation du contributeur.

Après avoir introduit les données utilisées pour tester MONITOR nous présentons les expériences menées pour valider le modèle dans les sections suivantes.

6.3 Comparaison des éléments de MONITOR avec l'existant

MONITOR analyse quatre éléments pour estimer le profil du contributeur : son imprécision, sa certitude, son attention et sa réflexion. Afin de valider notre modèle nous souhaitions faire une comparaison de ces quatre points avec des estimations équivalentes existantes. Cependant, à notre connaissance, aucune méthode ne permet d'estimer la certitude du contributeur pour sa réponse ni son attention lors de la campagne. Nous n'avons pu comparer le modèle proposé sur ces deux éléments. Cette section présente donc les comparaisons du calcul de la précision et de la réflexion du contributeur avec des méthodes existantes.

6.3.1 Comparaison de la précision avec le degré de Ben Rjab et al. 2016

Afin de calculer la fonction de masse associée à l'imprécision du contributeur $m_c^{\Omega_I}$, décrite section 5.3.1, nous nous inspirons du degré de précision DP_c de Ben Rjab et al. 2016, équation (3.33). C'est pourquoi cette section fait une comparaison de la masse calculée par MONITOR pour caractériser l'imprécision du contributeur $m_c^{\Omega_I}(P)$ avec le degré de précision de Ben Rjab et al. 2016. Puisque les réponses $X \in 2^{\Omega_q}$ sont modélisées par des fonctions de masse à support simple le degré de précision de Ben Rjab et al. 2016 peut être écrit pour X différent de Ω_q de la manière suivante :

$$DP_c = \frac{1}{|E_{Q_c}|} \sum_{q \in E_{Q_c}} m_{cq}^{\Omega_q}(X) \left(1 - \frac{log_2(X)}{log_2(\Omega_q)}\right) \tag{6.1}$$

À la différence de $m_c^{\Omega_I}(P)$, DP_c inclut la masse $m_{cq}^{\Omega_q}(X)$ dans son calcul. De plus, pour $m_c^{\Omega_I}(P)$ le logarithme de la cardinalité d'une réponse est divisé par le logarithme de l'imprécision maximale autorisée au contributeur imp_{MAX}, alors que pour DP_c la division se fait par le logarithme de Ω_q.

Puisque $m_c^{\Omega_I}(P)$ et DP_c calculent l'imprécision du contributeur, la comparaison est effectuée sur les données collectées lors des campagnes où le contributeur peut être imprécis : multi_oiseaux_imprécis, 10_oiseaux_imprécis, 10_oiseaux_dynamique. Pour la campagne multi_oiseaux_imprécis, $|\Omega_q| = 5$ puisque cinq réponses sont proposées au contributeur à chaque question, alors que pour 10_oiseaux_imprécis et 10_oiseaux_dynamique,

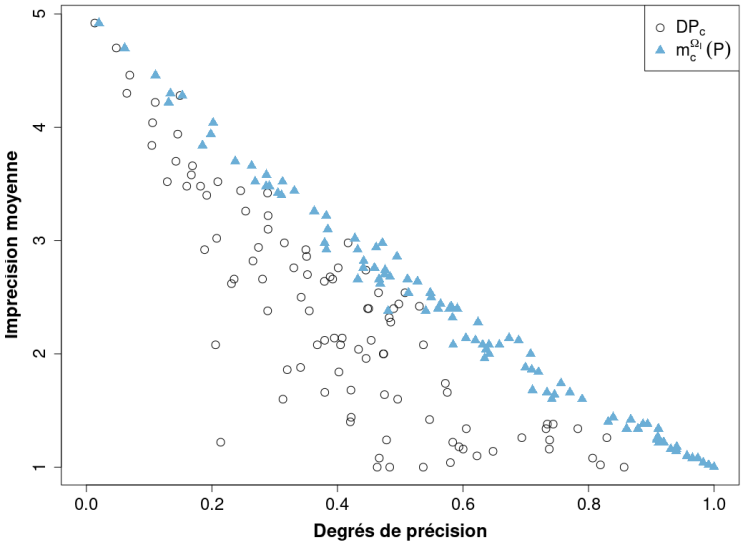


FIGURE 6.2 – Comparaison des degrés pour la campagne **multi_oiseaux_imprécis**.

$|\Omega_q| = 10$ car 10 réponses sont offertes. Pour ces trois campagnes le contributeur peut choisir de sélectionner jusqu'à 5 noms d'oiseau, ainsi $imp_{MAX} = 5$ dans chaque cas.

Les figures 6.2, 6.3 et 6.4 présentent les comparaisons de $m_c^{\Omega_I}(P)$ et DP_c pour les trois campagnes. Pour ces trois figures, un contributeur c est représenté par deux points de même ordonnée qui correspond à son imprécision moyenne. Ces points ont comme abscisse la valeur de DP_c pour l'un et celle de $m_c^{\Omega_I}(P)$ pour l'autre. Par exemple, sur la figure 6.2 les deux points qui ont pour ordonnée une imprécision moyenne de 5 et des valeurs DP_c et $m_c^{\Omega_I}(P)$ proches correspondent tous deux au même contributeur. Le contributeur peut sélectionner au maximum 5 noms d'oiseau, l'imprécision varie donc entre 1 et 5, et les valeurs de DP_c et $m_c^{\Omega_I}(P)$ sont inclues dans l'intervalle $[0, 1]$.

Pour la figure 6.2, $DP_c \leq m_c^{\Omega_I}(P)$ car $imp_{MAX} = |\Omega_q|$, comme le montre la démonstration en annexe 4 page 164. Nous constatons sur cette figure que plus un contributeur est précis en moyenne, plus son imprécision est proche de 1 sur le graphique, plus l'écart entre les valeurs de DP_c et $m_{cq}^{\Omega_q}$ se creuse. Par exemple, un contributeur a un degré DP_c proche de 0.2 sur ce graphique alors que son imprécision moyenne est très proche de 1 de même que $m_c^{\Omega_I}(P)$. Comme pour ce graphique $imp_{MAX} = |\Omega_q|$, l'écart entre les valeurs de DP_c et $m_c^{\Omega_I}(P)$ est uniquement dû à la masse $m(X)$ dans le calcul de DP_c qui amoindrit la qualité de l'estimation de l'imprécision du contributeur.

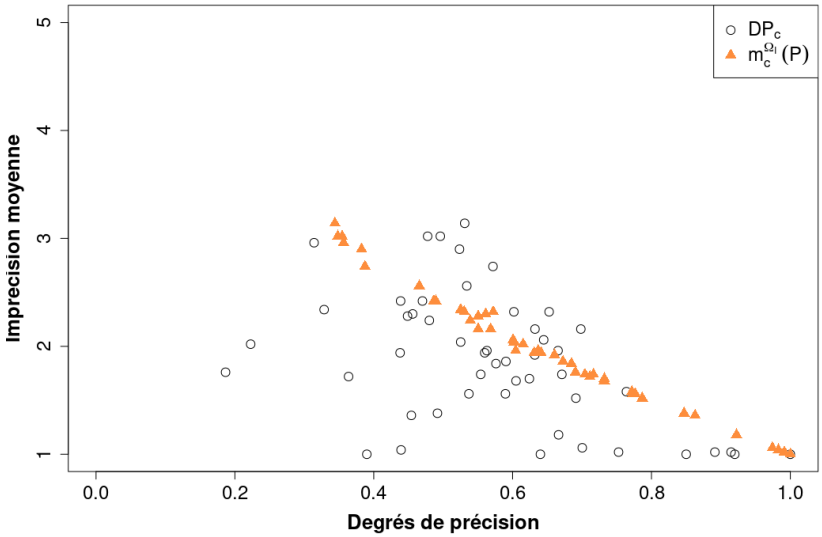


FIGURE 6.3 – Comparaison des degrés pour la campagne **10_oiseaux_imprécis**.

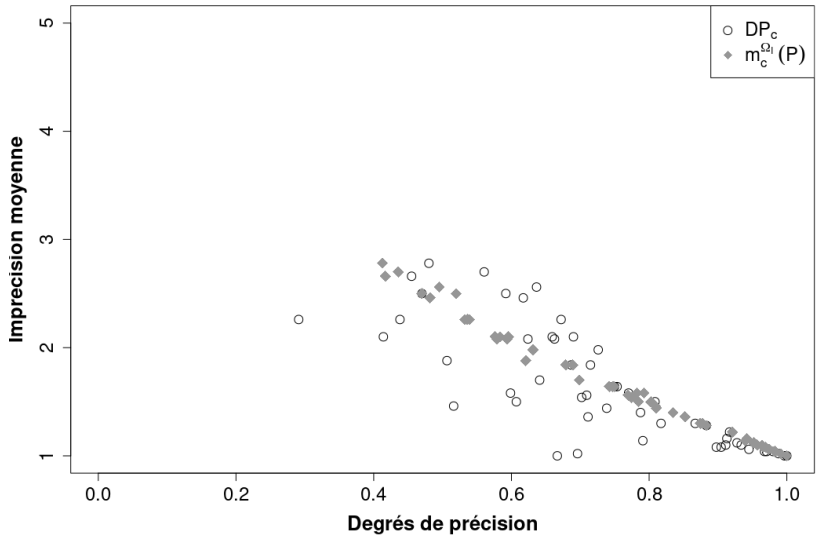


FIGURE 6.4 – Comparaison des degrés pour la campagne **10_oiseaux_dynamique**.

$\|X\|$	1	2	3	4	5
$\gamma(\|X\|)$	1.00	0.82	0.61	0.35	0.00

TABLE 6.3 – Valeurs de $\gamma(|X|)$ en fonction de l'imprécision de la réponse.

Pour les figures 6.3 et 6.4 des expériences 10_oiseaux_imprécis et 10_oiseaux_dynamique, $imp_{MAX} < |\Omega_q|$ ce pourquoi $m_c^{\Omega_I(P)} \leq DP_c$ dans le cas où $\gamma(|X|) \leq m_{cq}^{\Omega_q}(X)$ (la démonstration est donnée en annexe 4 page 164) avec :

$$\gamma(|X|) = \frac{log_2|\Omega_q|}{log_2(imp_{MAX})} \frac{log_2(imp_{MAX}) - log_2|X|}{log_2|\Omega_q| - log_2|X|} \tag{6.2}$$

Le tableau 6.3 répertorie les valeurs numériques obtenues pour γ pour $imp_{MAX} = 5$ et $|\Omega_q| = 10$. La condition $\gamma(|X|) \leq m_{cq}^{\Omega_q}(X)$ est par exemple vérifiée lorsque $|X| = 4$ et que le contributeur est certain de sa réponse, ce qui correspond à $m_{cq}^{\Omega_q}(X) = 0.83$.

Nous observons sur les figures 6.3 et 6.4 que l'imprécision moyenne ne dépasse pas 3 alors qu'elle atteint une valeur proche de 5 pour certains contributeurs de la figure 6.2. Ceci s'explique par le fait que bien qu'il est possible pour le contributeur de sélectionner jusqu'à 5 oiseaux pour les campagnes 10_oiseaux_imprécis et 10_oiseaux_dynamique cela n'est pas nécessaire. En effet, les 10 espèces proposées peuvent être regroupées d'après leur famille, soit 1 Muscicapidae, 2 Columbidae, 3 Paridae et 4 Corvidae (voir le tableau 6.2). Il est possible d'hésiter entre des oiseaux d'une même famille mais plus difficilement avec des oiseaux de différentes familles, le pigeon et le corbeau sont par exemple très éloignés. C'est pourquoi l'imprécision moyenne ne doit pas excéder 4 en toute logique, ce qui corrobore avec les résultats observés.

Sur les figures 6.3 et 6.4, les valeurs de DP_c se rapprochent de $m_c^{\Omega_I}(P)$ en comparaison avec la figure 6.2, surtout pour la figure 6.4 dont les données sont issues de la campagne de *crowdsourcing* dynamique. Les valeurs des deux estimations de l'imprécision restent cependant distinctes à cause de l'utilisation de $m(X)$ dans le calcul de DP_c, mais aussi parce que $imp_{MAX} < |\Omega_q|$. Dans l'ensemble, d'après la lecture des graphiques, $m_c^{\Omega_I}(P)$ est plus représentatif de l'imprécision moyenne du contributeur que DP_c.

Nous avons réalisé dans cette section une comparaison d'un élément de l'estimation de la qualification du contributeur, à savoir l'estimation de son imprécision, avec l'existant. Dans la section suivante nous faisons de même pour le comportement avec la réflexion du contributeur pour la tâche.

6.3.2 Comparaison de la réflexion avec Komarov et al. 2013

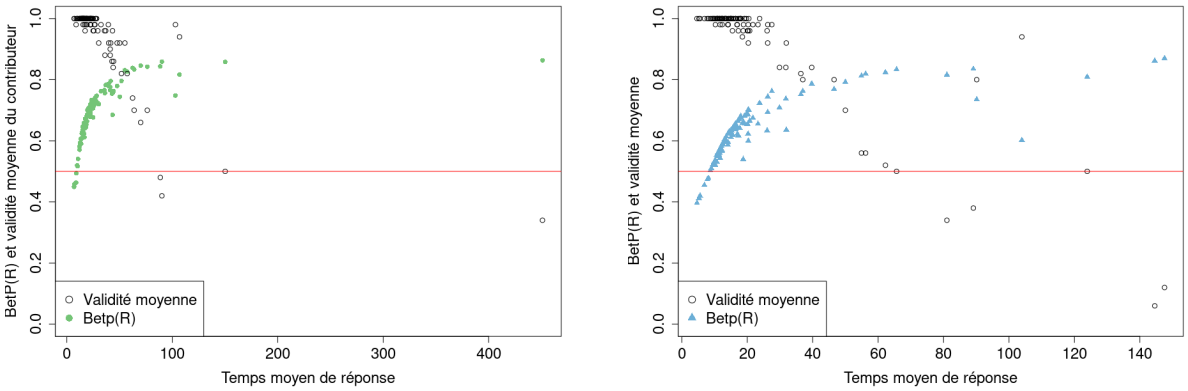
MONITOR utilise le temps de réponse du contributeur à une question T_{cq} et le compare à un temps minimal attendu T_{0q} afin de calculer la fonction de masse associée à la réflexion du contributeur.

Nous comparons le calcul de la réflexion de MONITOR avec la méthode statistique d'exclusion de contributeurs marginaux de Komarov et al. 2013 car ces derniers utilisent également le temps de réponse. D'après les auteurs un contributeur est considéré comme marginal si son temps de réponse est trop éloigné des temps de réponse de l'ensemble de la foule. Un contributeur qui répond trop rapidement est suspecté de répondre aléatoirement et un contributeur qui est trop long est jugé comme non pertinent par les auteurs. Afin de déterminer les contributeurs marginaux Komarov et al. 2013 calculent l'écart interquartile (IQR) des temps de réponse, soit la différence entre le troisième ($Q3$) et le premier quartile ($Q1$). Un contributeur est estimé marginal et donc exclu de l'étude des auteurs si le calcul de son temps de réponse n'est pas inclus dans l'intervalle $[Q1 - 3 * IQR, Q3 + 3 * IQR]$. Deux mesures de temps de réponse sont utilisées : la log-transformée du temps moyen de réponse par participant et la log-transformée du temps maximal de réponse par participant.

Un taux de marginalité du contributeur est calculé en s'inspirant de la méthode d'exclusion de Komarov et al. 2013 afin d'obtenir un élément de comparaison avec la réflexion. Pour calculer ce taux de validité, à chaque question, une fonction indicatrice est associée à la validité de la réponse du contributeur. Si $T_{cq} \in [Q1 - 3 * IQR, Q3 + 3 * IQR]$ la contribution est validée et la fonction indicatrice vaut 1. La moyenne des fonctions indicatrices est effectuée pour chaque contributeur afin d'obtenir leur taux de validité. L'algorithme 2 est donné en annexe 1 page 159.

Dans les expériences menées avec MONITOR, T_{0q} est égal au premier quartile de l'ensemble des temps de réponse de la foule à la question q, et $\alpha_R = 0.8$ de façon arbitraire. Pour chaque contributeur la moyenne des fonctions de masse $m_{cq}^{\Omega_R}$ est calculée afin d'obtenir $m_c^{\Omega_R}$ et la probabilité pignistique $Betp(R)$ est calculée. La probabilité $BetP(R)$ que le contributeur soit réfléchi est ensuite comparée au taux de validité du contributeur.

Pour chaque contributeur, le temps moyen de réponse aux questions T_c est calculé pour la totalité de la campagne. Il s'agit du temps moyen présenté par l'axe des abscisses pour les figures 6.5 et 6.6. L'axe des ordonnées correspond pour un contributeur respectivement à la probabilité pignistique qu'il soit réfléchi au cours de l'ensemble de la campagne et au taux de validité inspiré de la méthode de sélection de Komarov et al. 2013. Sur chaque



(a) multi_oiseaux_précis (b) multi_oiseaux_imprécis

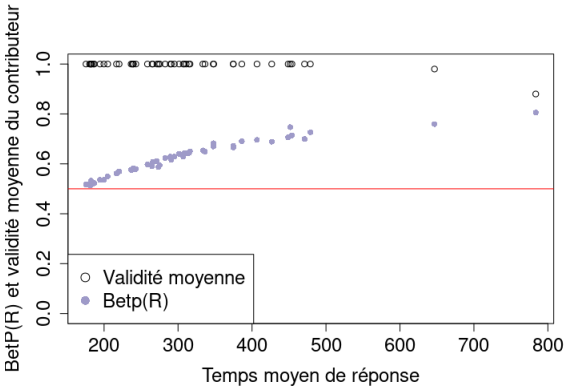
FIGURE 6.5 – Probabilité pignistique que le contributeur soit réfléchi et validité moyenne du contributeur pour les expériences multi_oiseaux.

Expérience	Taux de validité < 0.5		Taux de validité ≥ 0.5		% Bonne classification
	TBRec < 0.5	TBRec ≥ 0.5	TBRec< 0.5	TBRec≥ 0.5	
multi_oiseaux_précis	0	4	66	30	30
multi_oiseaux_imprécis	1	2	71	27	28
10_oiseaux_précis	0	0	38	12	24
10_oiseaux_imprécis	0	0	6	44	88
10_oiseaux_dynamique	0	0	4	47	92

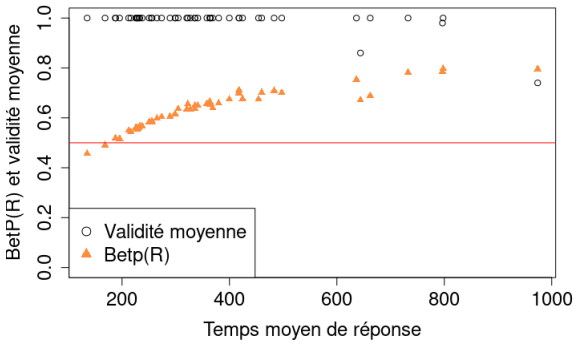
TABLE 6.4 – Comparatif des taux de validité des contributeurs et de leur taux de bonne reconnaissance d'oiseaux.

graphique de ces figures un contributeur est représenté par deux points, un point pour la réflexion estimée et un autre pour sa validité moyenne. D'après les figures, seules les expériences avec de multiples oiseaux présentent des contributeurs dont la validité moyenne est inférieure à 0.5 et par conséquent dont les réponses ne doivent pas être considérées. Pour ces deux expériences présentées figure 6.5, la majorité des contributeurs ont des temps moyens de réponse cours ce qui facilite l'apparition de contributeurs marginaux avec des temps moyens de réponse plus grands. Cela est contraire aux expériences avec 10 oiseaux de la figure 6.6 où les temps moyens de réponse sont plus dispersés.

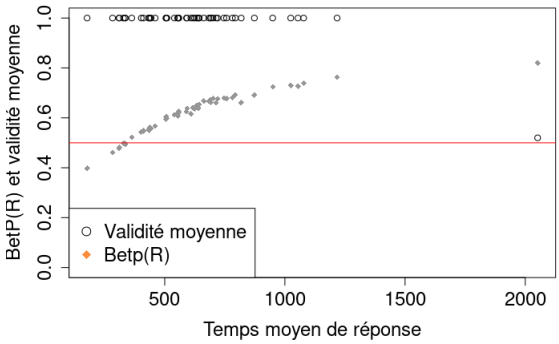
D'après les figures 6.5 et 6.6 certains contributeurs sont estimés non-réfléchis au cours de la campagne par MONITOR, $BetP(R) < 0.5$. Ceci signifie que soit ils n'ont pas pris le temps de réfléchir car ils répondent aléatoirement, soit ils ont une expertise supérieure à la moyenne pour la tâche et ont besoin d'un temps de réflexion moindre.



(a) 10_oiseaux_précis



(b) 10_oiseaux_imprécis



(c) 10_oiseaux_imprécis

FIGURE 6.6 – Probabilité pignistique que le contributeur soit réfléchi et validité moyenne du contributeur pour les expériences 10_oiseaux.

Dans les tableaux 6.4 et 6.5, nous faisons un comparatif, du taux de validité pour le premier et de la probabilité pignistique sur la réflexion pour le second, avec le taux de bonne reconnaissance du contributeur (TBRec). Pour le taux de validité, KOMAROV et al. 2013 excluent de leur analyse les contributeurs pour lesquels le temps de réponse n'est pas inclus dans l'intervalle temporel voulu. Ainsi, nous calculons le pourcentage de bonne classification entre les taux de validité et les TBRec des contributeurs. Pour la réflexion, il n'est pas possible de calculer ce pourcentage de bonne classification immédiatement car parmi les contributeurs non-réfléchis il y a à la fois les contributeurs malveillants et les experts. Comme nous l'avons observé sur les figures 6.5 et 6.6, tous les contributeurs des expériences avec 10 oiseaux ont un taux de validité supérieur ou égal à 0.5. D'après le tableau 6.4, les meilleurs pourcentages de bonne classification sont obtenus pour les

Expérience	betP(R)< 0.5		betP(R) ≥ 0.5	
	TBRec < 0.5	TBRec ≥ 0.5	TBRec < 0.5	TBRec ≥ 0.5
multi_oiseaux_précis	15	0	51	34
multi_oiseaux_imprécis	13	0	59	29
10_oiseaux_précis	0	0	38	12
10_oiseaux_imprécis	0	2	6	42
10_oiseaux_dynamique	2	5	2	42

TABLE 6.5 – Comparatif des probabilités pignistiques que les contributeurs soient réfléchis avec leur taux de bonne reconnaissance d'oiseaux.

campagnes 10_oiseaux_imprécis et 10_oiseaux_dynamique pour lesquelles les 10 mêmes noms d'oiseau sont toujours proposés au contributeur et il peut être imprécis. Pour la campagne multi_oiseau_imprécis le contributeur peut également être imprécis mais les noms d'oiseau proposés changent à chaque question, or pour cette campagne 72 contributeurs ont un $TBRec < 0.5$. En plus de permettre au contributeur d'être imprécis, utiliser les mêmes propositions de réponses pour toutes les questions semble améliorer le TBRec. Cette constatation est la même pour les TBRec du tableau 6.5.

Pour les campagnes multi_oiseaux, précis et imprécis, l'ensemble des contributeurs estimés non réfléchis par MONITOR ont un taux de bonne réponse inférieur à 0.5 et sont donc a priori des contributeurs marginaux. Or ces contributeurs marginaux ont un taux de validité supérieur à 0.5 d'après le tableau 6.4 alors qu'ils devraient également être estimés marginaux. Le problème de la méthode statistique de KOMAROV et al. 2013 est qu'elle peut valider des contributeurs comme marginaux si leurs temps de réponse sont assez éloignés de l'intervalle $[Q1 - 3 * ICQ, Q3 + 3 * ICQ]$. Ceci est observable sur la figure 6.5, où les temps de réponses moyens sont très courts et les taux de validité sont regroupés. Les taux inférieurs à 0.5 le sont pour les contributeurs qui ont un temps moyen de réponse très supérieur aux autres.

La méthode de KOMAROV et al. 2013 parvient bien à identifier des contributeurs marginaux ici, mais principalement ceux qui ont un temps moyen de réponse élevé et un TBRec supérieur ou égal à 0.5. Ce sont donc des contributeurs dont les réponses sont pertinentes, ce qui n'est pas bénéfique à l'agrégation des contributions. A contrario, certains contributeurs sont validés par cette approche statistique mais ont un TBRec inférieur à 0.5. MONITOR identifie plus de contributeurs non-réfléchis, et cette catégorie inclut à la fois des contributeurs malveillants et des experts, ce pourquoi certains ont $TBRec < 0.5$ et d'autres $TBRec \geq 0.5$. L'analyse des données de cette section nous a

également permis d'observer qu'en plus de permettre à la foule d'être imprécise, utiliser le même ensemble de propositions de réponses pour chaque question accroît les TBRec.

Après avoir fait dans cette section une comparaison de l'imprécision et la réflexion du contributeur avec l'existant, nous présentons dans la suivante les expérimentations réalisées sur le profil dans son intégralité.

6.4 Profil

Cette section revient sur les tests réalisés pour l'estimation du profil du contributeur par MONITOR et les comparaisons faites avec Ben Rjab et al. 2016 et EM. Pour ces expériences seules les données incluant des réponses imprécises sont utilisées puisqu'elles sont plus pertinentes pour l'analyse du profil. En effet, il n'est pas intéressant de calculer la masse sur l'imprécision du contributeur $m_c^{\Omega_I}$ sur des données précises. Les données manipulées ici proviennent par conséquent des campagnes : multi_oiseau_imprécis, 10_oiseaux_imprécis et 10_oiseaux_dynamique.

6.4.1 Apprentissage semi-supervisé pour déterminer α_P

Afin de calculer la masse sur le profil du contributeur $m_c^{\Omega_P}$, une conversion des cadres de discernement associés à la qualification et au comportement du contributeur est réalisée. Les fonctions de masse converties sont pondérées par des coefficients α_x comme l'indique l'équation (5.10) page 98. Il n'est pas anodin de définir les poids à accorder à chaque élément du profil, c'est pourquoi nous avons scindé les résultats de nos campagnes en deux jeux de données afin de réaliser un apprentissage semi-supervisé. Les probabilités pignistiques sur le profil des contributeurs sont calculées pour des valeurs α_I, α_C, α_R, $\alpha_A \in [0, 10]$. Pour chaque réponse X renseignée à une question q, le booléen $estValide$ qui indique la validité de la réponse est divisé par le nombre d'éléments sélectionnés par le contributeur $|X|$. La moyenne de ces quotients sur l'ensemble des questions nous permet d'obtenir le taux de bonne réponse du contributeur :

$$TBR_c = \sum_{q \in E_{Q_c}} \frac{estValide}{|X|} \tag{6.3}$$

Ce taux TBR_c permet de calculer les taux de bonne classification de MONITOR et des autres approches présentées dans la section suivante.

	Mauvais	Moyen	Bon	Expert
TBR_c	≤ 0.2	$]0.2, 0.5[$	$[0.5, 0.85[$	≥ 0.85

TABLE 6.6 – Tableau récapitulatif des profils d'après les valeurs des taux de bonne réponse des contributeurs.

	Min	Q1	Moyenne	Q3	Max
multi_oiseaux_imprécis	0.16	0.30	0.44	0.52	0.96
10_oiseaux_imprécis	0.29	0.39	0.44	0.46	0.92
10_oiseaux_dynamique	0.36	0.45	0.52	0.55	0.92

TABLE 6.7 – Résumé des taux de bonne réponse des contributeurs pour les **données d'apprentissage**.

Avant de déployer les campagnes de *crowdsourcing* en ligne, trois ornithologues ont réalisé la campagne multi_oiseaux_imprécis afin de les valider. Le taux de bonne réponse moyen des Experts au cours de la campagne est de 0.89, nous considérons donc qu'un contributeur dont $TBR_c \geq 0.85$ peut être reconnu comme un expert en classification d'oiseau pour cette tâche. Il n'est pas attendu des Experts qu'ils aient un taux de bonne réponse de 1.00 car certaines espèces ou photos peuvent rendre plus difficile l'identification de l'oiseau. De manière arbitraire nous estimons qu'une personne capable d'identifier un oiseau sur deux ou plus ($TBR_c \geq 0.5$) est un bon contributeur. En revanche, un contributeur qui n'identifie qu'un oiseau sur les cinq proposés ou moins ($TBR_c \leq 0.2$) est un mauvais contributeur car ses réponses sont aléatoires. Finalement, un contributeur Moyen a un taux de bonne réponse inclus entre celui du Bon et du Mauvais contributeur : ses réponses sont moins pertinentes que le Bon contributeur mais pas aléatoires contrairement au Mauvais. L'ensemble de ces informations sont résumées dans le tableau 6.6.

Pour les données d'apprentissage de la campagne multi_oiseaux_imprécis, le minimum des valeurs TBR_c est de 0.16 d'après le tableau 6.7 et le maximum de 0.96 avec un troisième quartile de 0.52, ce qui signifie que cet ensemble de données d'apprentissage inclut bien les quatre types de profil à identifier. Ce n'est pas le cas pour les données d'apprentissage des campagnes 10_oiseaux, il semblerait qu'aucun contributeur Mauvais n'y ait participé. Ceci peut s'expliquer par la taille de la foule plus restreinte pour ces deux campagnes.

Les tableaux 6.8, 6.9 et 6.10 présentent les valeurs de coefficients α retenues après l'apprentissage, les taux de bonne réponse moyen des contributeurs TBR_c d'après leur profil et les taux de bonne classification (TBC). Pour la campagne multi_oiseaux_imprécis trois

Coefficients α				$TBR_c(apprentissage)$				TBC	
α_I	α_R	α_C	α_A	Expert	Bon	Moyen	Mauvais	Apprentissage	Test
1	2	6	1	0.59	0.51	0.42	/	0.56	0.63
1	7	1	1	/	0.52	0.42	/	0.58	0.57
2	4	3	1	0.59	0.51	0.42	/	0.56	0.55

TABLE 6.8 – Apprentissage sur les données multi_oiseaux_imprécis.

Coefficients α				$TBR_c(apprentissage)$				TBC	
α_I	α_R	α_C	α_A	Expert	Bon	Moyen	Mauvais	Apprentissage	Test
1	2	6	1	0.67	0.45	0.41	0.40	0.72	0.4
1	2	7	0	0.58	0.45	0.41	/	0.72	0.4
2	4	3	1	0.67	0.45	0.41	0.40	0.72	0.36
2	4	4	0	0.58	0.45	0.41	/	0.72	0.36

TABLE 6.9 – Apprentissage sur les données 10_oiseaux_imprécis.

ensembles de coefficients sont retenus après l'apprentissage semi-supervisé et sont donnés dans le tableau 6.8. Pour cette campagne aucun Mauvais contributeur n'est identifié par l'apprentissage alors que nous avons la certitude qu'il y en a au moins un parmi la foule. Le taux moyen de bonne réponse des contributeurs estimés Expert par MONITOR est de 0.59 ce qui est bien supérieur au troisième quartile de TBR_c d'après le tableau 6.7 mais inférieur à la valeur de 0.85 attendue des Experts. Pour les contributeurs estimés comme Bon lors de l'apprentissage TBR_c est d'environ 0.5 ce qui est bien inclus dans l'intervalle théorique souhaité pour les bons contributeurs. De même, TBR_c pour les contributeurs Moyen est de 0.42 ce qui est également inclus dans l'intervalle souhaité pour ce profil. Le meilleur taux TBR_c est obtenu pour $\alpha_R = 7$ et les trois autres coefficients sont égaux à 1, aussi c'est principalement la réflexion qui joue un rôle dans l'estimation du profil, pour ces valeurs le TBC des données de test est de 0.57.

D'après le tableau 6.9 deux ensembles de valeurs α sont communs à ceux du tableau 6.8. Pour ces deux ensembles les taux de bonne réponse des Experts sont maximaux à une valeur de 0.67 ce qui est encore hors de l'intervalle attendu des Experts mais toujours supérieur au troisième quartile de l'ensemble de données. De même, TBR_c est inférieur à 0.45 pour les bons contributeurs ce qui est inférieur à la valeur minimale de 0.5 attendue d'un bon contributeur. En revanche, pour les contributeurs moyens TBR_c est bien inclus dans $]0.2, 0.5[$. Certains contributeurs sont de plus estimés Mauvais alors que d'après le tableau 6.7 il n'y a pas de Mauvais contributeurs pour cette campagne. Les taux de bonne

Coefficients α				$TBR_c(apprentissage)$				TBC	
α_I	α_R	α_C	α_A	Expert	Bon	Moyen	Mauvais	Apprentissage	Test
1	4	5	0	0.63	0.48	0.51	/	0.28	0.36
2	7	1	0	0.63	0.48	0.51	/	0.28	0.36

TABLE 6.10 – Apprentissage sur les données 10_oiseaux_dynamique.

classification pour l'apprentissage sont meilleurs que ceux des données de test.

Les taux de bonne classification pour la campagne 10_oiseaux_dynamique sont les plus mauvais, dans les meilleurs des cas nous obtenons un taux de 0.28 pour les données d'apprentissage. Pour cette campagne il est parfois demandé au contributeur s'il lui est possible de préciser ou élargir sa sélection de noms d'oiseaux, l'éventuelle réponse modifiée du contributeur n'est pas considérée ici. Pour les deux ensembles de valeurs retenus $\alpha_A = 0$ signifie que l'attention du contributeur ne doit pas être considérée pour ce jeu de données. De plus les TBR_c moyens des contributeurs d'après leur profil ne sont pas en accord avec ceux attendus. MONITOR ne semble pas s'adapter aussi bien à cette campagne qu'aux autres, ceci est probablement dû à l'aspect itératif du questionnaire qui n'est actuellement pas considéré par le modèle.

D'après les trois tableaux 6.8, 6.9 et 6.10, nous constatons que le coefficient α_I est faible puisqu'il est égal à 1 ou 2 pour des taux de bonne classification optimaux lors de l'apprentissage. Ce sont principalement les coefficients α_R et α_C qui ont de l'importance puisque le coefficient α_A prend une valeur de 0 ou 1. Ceci signifie que les fonctions de masse sur la réflexion et la certitude ont davantage d'importance comparées à l'imprécision et l'attention pour l'estimation du profil.

La section suivante fait une comparaison entre l'expertise estimée par MONITOR et celle de BEN RJAB et al. 2016 et de EM.

6.4.2 Comparaison avec d'autres estimations de l'expertise

BEN RJAB et al. 2016 utilisent également la théorie des fonctions des croyance pour déterminer le profil du contributeur c'est pourquoi nous comparons ici l'estimation faite par MONITOR avec celle des auteurs. Une autre approche pour l'estimation du profil est de considérer la matrice de confusion sur les réponses du contributeur établie par EM, que nous étudions dans cette section. MONITOR considère en plus de la qualification du contributeur et son comportement pour estimer son profil ce qui n'est pas le cas des deux autres méthodes BEN RJAB et al. 2016 et de EM.

Profil estimé	DE	DP	DG	TBR
Expert	0.751	0.257	0.504	0.384
Bon	0.748	0.104	0.426	0.266
Moyen	0.711	0.389	0.550	0.550
Mauvais	0.618	0.616	0.617	0.810

TABLE 6.11 – Moyenne des valeurs de DE_c, DP_c, DG_c et TBR_c d'après les groupes de profils (*clustering* sur DE_c et DP_c) pour la campagne **multi_oiseau _imprécis**.

Profil estimé	DE	DP	DG	TBR
Expert	0.653	0.504	0.579	0.400
Bon	0.629	0.301	0.465	0.358
Moyen	0.576	0.656	0.616	0.444
Mauvais	0.424	0.915	0.669	0.570

TABLE 6.12 – Moyenne des valeurs de DE_c, DP_c, DG_c et TBR_c d'après les groupes de profils (*clustering* sur DE_c et DP_c) pour la campagne **10_oiseau _imprécis**.

Comparaison avec Ben Rjab et al. 2016

Afin d'estimer le profil du contributeur c, BEN RJAB et al. 2016 calculent son degré d'exactitude des réponses DE_c et son degré de précision DP_c donnés par les équations (3.32) et (3.33) page 54. Les auteurs proposent ensuite de réaliser un *clustering* grâce à *k-mean* avec $k = 2$ pour différencier les contributeurs experts des non-experts. L'article indique que parmi les deux ensembles obtenus, celui ayant la valeur moyenne de DE_c la plus élevée est constitué d'experts.

Nous utilisons pour l'estimation des profils, d'après l'approche de BEN RJAB et al. 2016, l'algorithme des *k-mean* avec $k = 4$, car MONITOR définit quatre types de profil de contributeurs. Puisque pour les auteurs le groupe d'experts a la plus grande valeur moyenne DE_c, nous considérons également une expertise croissante du profil du contributeur d'après DE_c.

Pour les tableaux 6.11, 6.12 et 6.13 les profils sont estimés d'après la méthode de BEN RJAB et al. 2016 en réalisant un *clustering* sur DE_c et DP_c. Pour chaque groupe de contributeurs obtenu, les moyennes de DE_c, DP_c, DG_c et TBR_c sont calculées. Le degré DG_c de BEN RJAB et al. 2016 est la somme des degrés DE_c et DP_c pondérés par un coefficient β. L'équation (3.34) de DG_c est donnée page 55. Dans les expériences réalisées dans cette section, $\beta = 0.5$ car c'est la valeur pour laquelle les auteurs obtiennent les

Profil estimé	DE	DP	DG	TBR
Expert	0.628	0.446	0.537	0.418
Bon	0.600	0.645	0.622	0.473
Moyen	0.584	0.756	0.670	0.592
Mauvais	0.502	0.935	0.718	0.597

TABLE 6.13 – Moyenne des valeurs de DE_c, DP_c, DG_c et TBR_c d'après les groupes de profils (*clustering* sur DE_c et DP_c) pour la campagne **10_oiseau _dynamique**.

meilleurs résultats dans leur papier. Le taux de bonne réponse est la moyenne des valeurs TBR_c calculées d'après l'équation (6.3).

D'après le tableau 6.11 qui présente les résultats de la campagne multi_oiseau_imprécis, la valeur moyenne DE la plus élevée est de 0.751, ce groupe de contributeurs est par conséquent désigné comme Experts. Ensuite $DE = 0.748$ est la seconde valeur plus élevée et est associée aux Bons contributeurs, $DE = 0.711$ aux contributeurs Moyens et $DE = 0.618$ aux Mauvais. Le même raisonnement est suivi dans les tableaux 6.12 et 6.13 pour l'assignation des profils aux groupes de contributeurs.

Les trois tableaux montrent que l'estimation du profil d'après les valeurs moyennes de DE_c est mauvaise car elle n'est pas représentative du taux de bonne réponse du contributeur. Par exemple, pour le tableau 6.11 les experts ont un taux de bonne réponse moyen TBR inférieur à 0.5 et pour les Mauvais contributeurs $TBR = 0.810$, les deux catégories de profils sont donc mal estimées. C'est également le cas des tableaux 6.12 et 6.13 où les experts ont les plus faibles taux de bonne réponse. Ceci est confirmé par le calcul des taux de bonne classification des profils donnés par la colonne "Clustering DE,DP" du tableau 6.14 qui est de 0.17 pour la campagne multi_oiseau_imprécis, 0.24 pour 10_oiseaux_imprécis et 0.14 pour 10_oiseaux_dynamique. Il n'est ainsi pas possible de se référer au degré d'exactitude des réponses DE_c pour estimer l'expertise du contributeur dans nos expériences.

En comparaison, le tableau 6.14 inclut les taux de bonne classification des profils par MONITOR. Pour les campagnes multi_oiseau_imprécis et 10_oiseaux_imprécis les coefficients d'affaiblissement utilisés sont les suivants : $\alpha_2 = 1$, $\alpha_3 = 2$, $\alpha_4 = 6$, $\alpha_5 = 1$. Pour la campagne 10_oiseaux_dynamique : $\alpha_2 = 2$, $\alpha_3 = 7$, $\alpha_4 = 1$, $\alpha_5 = 0$, et le changement potentiel de la réponse du contributeur n'est pas inclus dans les calculs. Ces valeurs de α sont choisies d'après les résultats de la section 6.4.1. L'estimation du profil par MONITOR offre de meilleurs taux de bonne classification en comparaison de l'estimation

de Ben Rjab et al. 2016 avec un *clustering* sur DE et DP et une assignation des profils d'après DE.

Cependant, d'après les observations des tableaux 6.11, 6.12 et 6.13, le degré global d'expertise DG croît avec le taux de bonne réponse. Nous avons donc réalisé une nouvelle phase de *clustering*, toujours avec *k-mean* et $k = 4$, mais cette fois en utilisant DG plutôt que DE et DP. Les taux de bonne classification (voir tableau 6.14) obtenus de la sorte sont de 0.4 pour la campagne multi_oiseau_imprécis, 0.42 pour 10_oiseau_imprécis et 0.51 pour 10_oiseaux_dynamique. Il y a donc une grande amélioration en utilisant DG pour le *clustering* plutôt que DE et DP. Les taux de bonne classification de multi_oiseau_imprécis et 10_oiseau_imprécis restent inférieurs à ceux obtenus par MONITOR. Cependant, pour la campagne 10_oiseaux_dynamique le taux de bonne classification est plus élevé avec l'utilisation de DG plutôt que MONITOR. Il est déjà constaté dans la section 6.4.1 que le taux de bonne classification est le plus faible pour cette campagne. Cela est probablement dû à un changement de comportement du contributeur lorsqu'il lui est demandé de modifier sa réponse.

Il est plus intéressant d'utiliser MONITOR que le degré de Ben Rjab et al. 2016 pour des campagnes de *crowdsourcing* ne redemandant pas au contributeur de modifier sa réponse car le modèle que nous proposons offre de meilleurs pourcentages de bonne classification. De plus, au cours de nos expériences, nous avons constaté que le calcul du degré DE peut rapidement devenir coûteux en temps de calcul en fonction de la taille du cadre de discernement Ω_q sur la réponse. Dans la section suivante nous comparons l'estimation de l'expertise de MONITOR avec celle de EM.

Comparaison avec EM

Il est possible d'estimer l'expertise des contributeurs en utilisant les matrices de confusion calculées par l'algorithme EM. Dans les expériences, réalisées la précision du contributeur est obtenue grâce à sa matrice de confusion. Une fois les valeurs de précision obtenues pour l'intégralité de la foule, un *clustering* est réalisé avec $k = 4$. Les résultats du *clustering* sont ensuite comparés aux valeurs attendues du tableau 6.6 et le taux de bonne classification est calculé et inclus dans le tableau 6.14.

Il n'est pas possible d'appliquer EM aux données de la campagne multi_oiseau_imprécis car les réponses proposées à la question q ne sont pas les mêmes que celles proposées à la question $q + 1$ ce qui empêche d'établir les matrices de confusion. Il n'y a ainsi pas de valeur dans la ligne associée à cette campagne pour EM. Nous constatons pour les cam-

Campagne	MONITOR	Clustering DE,DP	Clustering DG	EM
multi_oiseau_imprécis	0.47	0.17	0.4	/
10_oiseaux_imprécis	0.56	0.24	0.42	0.62
10_oiseaux_dynamique	0.31	0.14	0.51	0.61

TABLE 6.14 – Récapitulatif des pourcentages de bonne classification des profils pour : MONITOR, BEN RJAB et al. 2016, EM.

pagnes 10_oiseaux_imprécis et 10_oiseaux_dynamique que EM est l'approche qui offre les meilleurs résultats pour l'estimation du profil comparé à MONITOR et BEN RJAB et al. 2016. L'écart entre les taux de bonne classification est de 0.06 pour la campagne 10_oiseaux_imprécis mais de 0.3 pour 10_oiseaux_dynamique car l'estimation du profil par MONITOR est la plus mauvaise pour cette campagne.

L'utilisation de EM pour estimer le profil du contributeur est plus pertinente que MONITOR, notamment dans le cas de campagnes de *crowdsourcing* où une question peut être reposée au contributeur pour qu'il fasse évoluer sa réponse. Dans le cas d'une campagne plus simple où la question n'est posée qu'une fois au contributeur, l'écart des taux de bonne classification entre EM et MONITOR reste faible. De plus, comme il est expliqué dans cette section, EM ne peut malheureusement pas être utilisé si l'ensemble de réponses proposées au contributeur change d'une question à une autre contrairement à MONITOR qui n'est pas impacté par cela. L'estimation du profil du contributeur a pour finalité d'affaiblir ses réponses d'après son profil, ainsi la section suivante présente les expérimentations réalisées pour l'agrégation des contributions.

6.5 Agrégation des contributions

Cette section présente les expérimentations réalisées pour l'agrégation des réponses avec le modèle proposé. Nous faisons dans la section 6.5.1 un comparatif de différents opérateurs de combinaison existant dans la théorie des fonctions de croyance afin de déterminer le plus intéressant à appliquer à des données de *crowdsourcing*. La section 6.5.2 fait la comparaison de l'agrégation des réponses par le MV, EM et MONITOR. Enfin, la section 6.5.3 met en parallèle la modélisation des réponses utilisées par MONITOR d'une approche similaire à celle de CASCADE de KOULOUGLI et al. 2016.

Campagne	Conjonctif (3.17)	Yager (3.18)	RCR (3.25)	*Cautious* (3.22)	LNS (3.20)	Moyenne (3.15)
multi_oiseaux_précis	**0.92**	**0.92**	0.88	0.13	0.68	0.86
multi_oiseaux_imprécis	0.9	0.9	0.9	0.19	0.9	**0.96**
10_oiseaux_précis	0.58	0.58	0.58	0.1	**0.66**	0.55
10_oiseaux_imprécis	0.72	0.72	0.72	0.1	0.7	**0.74**
10_oiseaux_dynamique	0.78	0.79	0.80	0.1	0.8	**0.84**

TABLE 6.15 – Comparaison des taux de bonne réponse pour différents opérateurs de combinaison pour les données de chaque campagne.

6.5.1 Comparaison de différents opérateurs de combinaison

Cette section présente la comparaison de différents opérateurs de combinaison appliqués aux fonctions de masse modélisant les réponses des contributeurs $m_{cq}^{\Omega_q}$. Dans un premier temps des tests sont effectués afin de déterminer le coefficient α utilisé pour affaiblir $m_{cq}^{\Omega_q}$ comme il est indiqué dans la section . Les réponses sont donc combinées par l'opérateur conjonctif normalisé, équation (3.17), pour les cinq jeux de données pour des valeurs de $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$. L'opérateur conjonctif est choisi car l'affaiblissement des réponses a plus d'impact, contrairement à d'autres opérateurs comme la moyenne ou la règle LNS. Les tests montrent que les meilleurs résultats sont obtenus pour $\alpha = 0.8$ et $\alpha = 0.9$. Nous avons choisi d'utiliser $\alpha = 0.9$ pour l'affaiblissement des fonctions de masse avant leur combinaison dans les expériences de cette section.

Nous utilisons la librairie *ibelief* pour réaliser nos expérimentations, notamment pour réaliser la combinaison des fonctions de masse. Dans cette librairie, les fonctions $\gamma_1(k)$ et $\gamma_2(k)$ requises dans l'équation (3.25) de l'opérateur RCR (page 49) sont :

$$\gamma_1(k) = \frac{log(\frac{1+x}{k+x})}{log(\frac{1+x}{x})} \tag{6.4}$$

$$\gamma_2(k) = \frac{log(\frac{(1+x)^k * (k+x)^{(1-k)}}{x})}{log(\frac{1+x}{x})} \tag{6.5}$$

Dans ces équations $k = m_{Conj}(\emptyset)$ et $x = 0.9$ dans la librairie *ibelief*.

Le tableau 6.15 répertorie les TBR obtenus pour chaque opérateur de combinaison présenté pour les cinq jeux de données. D'après ce tableau la règle *Cautious* est à proscrire pour des données de *crowdsourcing* puisqu'elle offre le TBR le plus faible, inférieur à 0.2 pour les cinq jeux de données. L'opérateur conjonctif normalisé et l'opérateur de Yager ont

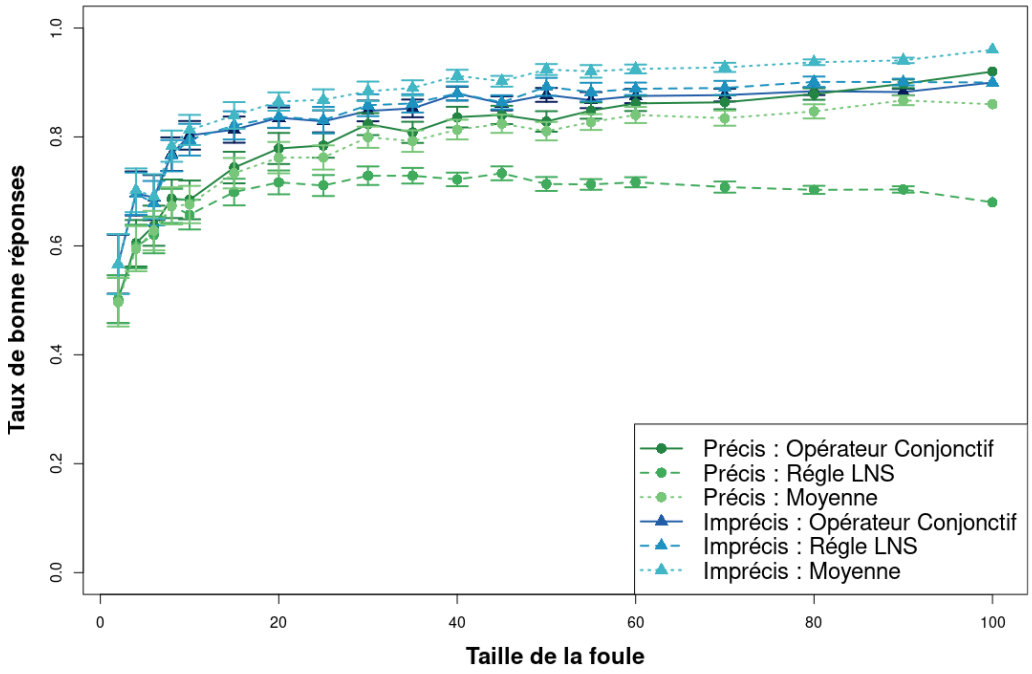


FIGURE 6.7 – Comparaison des opérateurs : Conjonctif, LNS et moyenne pour les campagnes **multi_oiseaux_précis et multi_oiseaux_imprécis.**

le meilleur TBR pour la campagne multi_oiseau_précis mais c'est la seule. De même la règle LNS offre le meilleur résultat pour la campagne 10_oiseaux_précis mais uniquement pour cette campagne. Pour les trois campagnes restantes c'est l'opérateur de moyenne qui est le plus performant. Comparons maintenant les opérateurs conjonctif, LNS et moyenne en faisant varier la taille de la foule. Nous faisons le choix de préférer l'opérateur conjonctif à celui de Yager car ils offrent globalement des résultats identiques. De plus, l'opérateur conjonctif normalisé répartit la masse associée à $\emptyset$ sur les éléments focaux restant alors que l'opérateur de Yager l'ajoute à l'ignorance Ω.

Pour la figure 6.7 l'agrégation des réponses est réalisée pour chaque opérateur et pour une taille n de la foule croissante, telle que :

$$n \in \{2, 4, 6, 8, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 70, 80, 90, 100\}$$

Pour chaque valeur de n les contributeurs dont les réponses sont agrégées sont sélectionnées aléatoirement, puis le TBR est calculé sur les 50 photos. Ce procédé de sélection de

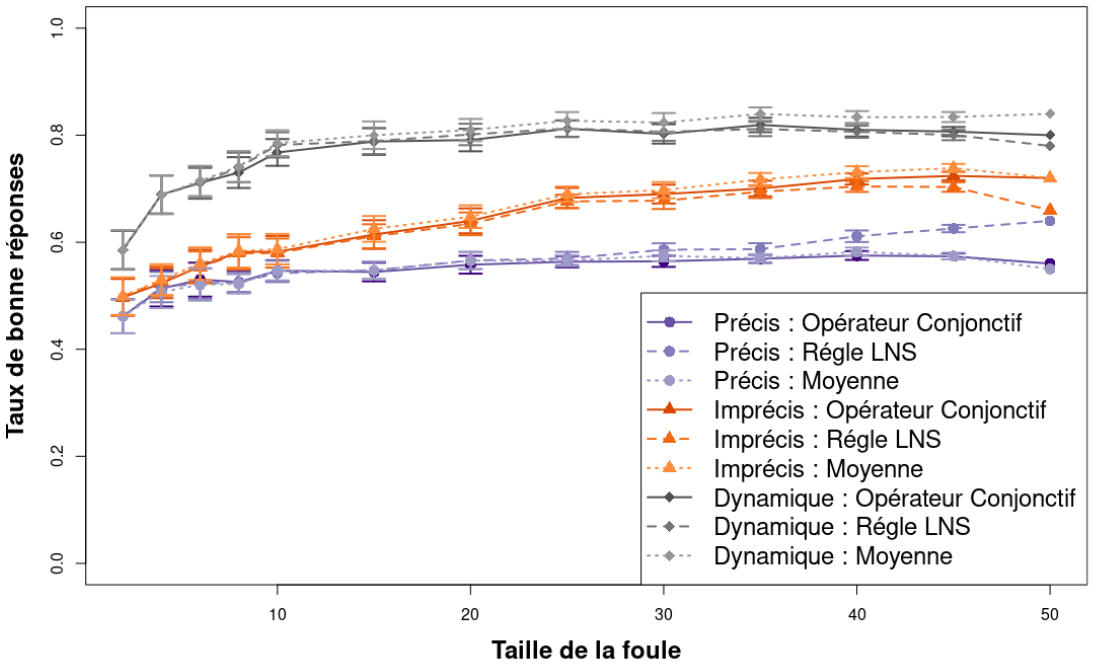


FIGURE 6.8 – Comparaison des opérateurs : Conjonctif, LNS et moyenne pour les **10_oiseaux_précis, 10_oiseaux_imprécis et 10_oiseaux_dynamique**.

la foule, agrégation des contributions et calcul des TBR est réalisé 50 fois pour chaque valeur de n afin d'obtenir un TBR moyen pour n et un intervalle de confiance à 95%. La figure 6.8 est réalisée de manière analogue à la figure 6.7, seulement les tailles n de la foule sont inférieures ou égales à 50 car il s'agit de la taille maximale de la foule pour ces campagnes.

La figure 6.7 montre des résultats proches pour la campagne imprécise, alors que l'écart entre la règle LNS et les deux autres opérateurs croît avec la taille de la foule pour la campagne précise. La figure 6.8 présente également des résultats proches entre les différents opérateurs appliqués à un même jeu de données pour une taille de foule inférieure à 30 contributeurs. Dans l'ensemble, d'après ces deux figures, les campagnes imprécises offrent de meilleurs résultats que celles précises ce qui est en accord avec les résultats du chapitre 4. De plus, l'opérateur de moyenne semble le plus pertinent des trois car pour la figure 6.7, la moyenne offre les meilleurs résultats pour la campagne imprécise. Pour la campagne précise, les intervalles de confiance des TBR de la moyenne et de l'opérateur conjonctif se croisent jusqu'à $n = 70$ ce qui montre une proximité des

Campagne	α_I	α_R	α_C	α_A
multi_oiseaux_précis	0	2	6	2
multi_oiseaux_imprécis	1	2	6	1
10_oiseaux_précis	0	2	6	2
10_oiseaux_imprécis	1	2	6	1
10_oiseaux_dynamiques	2	7	1	0

TABLE 6.16 – Coefficients utilisés pour l'estimation des profils par MONITOR pour les différentes campagnes.

résultats. De plus, sur la figure 6.8 les TBR de la moyenne sont constamment les plus élevés pour les campagnes avec des données imprécises. Pour la campagne 10_oiseaux_précis, c'est la règle LNS qui présente les meilleurs résultats, mais les intervalles de confiance de la règle LNS et de la moyenne se coupent jusqu'à une taille de foule de $n = 35$ contributeurs.

Après avoir établi un comparatif de plusieurs opérateurs de combinaison nous avons observé que la moyenne reste la méthode offrant le plus couramment le TBR le plus élevé. C'est donc l'opérateur de moyenne que nous employons pour l'agrégation des réponses dans la section suivante.

6.5.2 Comparaison de MV, EM, et MONITOR

Dans cette section nous comparons l'agrégation des réponses par le vote majoritaire, l'algorithme EM et MONITOR. L'objectif de MONITOR est d'affaiblir par un degré approprié les fonctions de masse afin d'améliorer les résultats obtenus par l'agrégation des données comme c'est le cas avec le taux de bonne reconnaissance. Pour ce faire, l'affaiblissement α_P dépend du profil du contributeur. Les coefficients utilisés pour l'estimation des profils pour les différentes campagnes sont donnés dans le tableau 6.16.

Apprentissage des valeurs optimales α_P

Afin d'identifier les meilleures valeurs de α_P, les jeux de données sont de nouveaux divisés en deux pour avoir un ensemble de données d'apprentissage et un ensemble de test. Nous considérons que l'Expert a de meilleures connaissances du domaine de la tâche et par conséquent davantage de crédit doit être accordé à ses réponses. Ainsi pour cette catégorie nous testons les valeurs $\alpha_{Expert} \in [0.5, 1.0]$. Le Bon contributeur a des connaissances moindres que l'Expert c'est pourquoi les valeurs de α_{bon} sont testées sur un intervalle plus restreint $[0.5, 0.85]$. Le contributeur moyen, bien que volontaire dans sa réalisation de la

tâche manque de capacité, ainsi ses réponses peuvent encore être davantage affaiblies et l'intervalle utilisé pour les tests est $[0.2, 0.7]$. Finalement le Mauvais contributeur a des réponses aléatoires, c'est pourquoi l'employeur ne peut pas lui faire confiance et les valeurs de $\alpha_{Mauvais}$ testées sont inclues dans l'intervalle $[0.0, 0.2]$.

Lorsque les données d'apprentissage sont modélisées par des fonctions de croyance, toutes affaiblies indifféremment par un coefficient $\alpha_P = 0.9$ et agrégées par la moyenne, les TBR obtenus sont les suivants :

— multi_oiseaux_précis : $TBR = 0.67$

— multi_oiseaux_précis : $TBR = 0.94$

— 10_oiseaux_précis : $TBR = 0.64$

— 10_oiseaux_imprécis : $TBR = 0.74$

— 10_oiseaux_dynamiques : $TBR = 0.80$

Pour les données des expériences multi_oiseaux_précis et multi_oiseaux_imprécis plusieurs ensembles de valeurs α_P permettent d'avoir des TBR respectivement de 0.68 et 0.96. Ces TBR obtenus par apprentissage de α_P sont meilleurs que ceux obtenus avec un affaiblissement de 0.9 indépendamment du profil du contributeur.

Pour l'expérience 10_oiseaux_imprécis, les TBR calculés pour différentes valeurs α_P sur les données d'apprentissage sont, au mieux, égaux à 0.64, ce qui est le TBR obtenu pour un affaiblissement de 0.9 commun à l'ensemble des contributeurs. En effet, sur l'ensemble des contributeurs pour cette campagne, 44 sont estimés comme "moyen" par MONITOR et seulement 6 sont bons. Le manque de diversité dans les profils estimés fait que l'apprentissage sur l'affaiblissement d'après le profil a peu d'impact. Puisque les valeurs de α_P utilisées pour l'apprentissage n'ont pas d'impact sur le calcul du TBR nous avons choisi d'expérimenter les valeurs de α_P du tableau 6.17 sur les données de tests. Pour les valeurs d'affaiblissement de ce tableau nous obtenons dans les trois cas un TBR de 0.6 sur les données de tests, alors que pour un affaiblissement commun à tous les contributeurs de 0.9 le TBR est de 0.54.

Pour les données d'apprentissage de la campagne 10_oiseaux_imprécis, dans les meilleurs cas les TBR sont égaux à 0.86 ce qui est bien supérieur à la valeur de 0.74 pour un affaiblissement commun $\alpha = 0.9$. Les valeurs de α_P requises pour $TBR = 0.86$ sont répertoriées dans le tableau 6.17. Ce tableau comptabilise trois ensembles de valeurs α_P, pour ces trois ensembles nous avons :

$$\alpha_{Expert} > \alpha_{Bon} > \alpha_{Moyen} \geq \alpha_{Mauvais}$$

α_{Expert}	α_{Bon}	α_{Moyen}	$\alpha_{Mauvais}$	TBR test
0.95	0.50	0.20	0.20	0.6
1.00	0.50	0.20	0.20	0.6
1.00	0.50	0.25	0.20	0.6

TABLE 6.17 – Affaiblissement α_P pour $TBR = 0.86$ pour les données d'apprentissage de la campagne 10_oiseaux_imprécis.

α_{Expert}	α_{Bon}	α_{Moyen}	$\alpha_{Mauvais}$	TBR test
0.50	0.85	0.20	[0.00,0.20]	0.84
0.85	0.50	0.20	[0.00,0.20]	0.78
0.90	0.55	0.20	[0.00,0.20]	0.80
0.95	0.55	0.30	[0.00,0.20]	0.76
1.00	0.60	0.25	[0.00,0.20]	0.78

TABLE 6.18 – Affaiblissement α_P pour $TBR = 0.84$ pour les données d'apprentissage de la campagne 10_oiseaux_dynamique.

Les TBR calculés sur les données de tests d'après les valeurs α_P du tableau 6.17 sont égaux à 0.6 qui est également le TBR obtenu pour ces mêmes données avec un affaiblissement commun de 0.9.

Le tableau 6.17 présente les valeurs de α_P qui permettent d'obtenir un TBR de 0.84 pour les données d'apprentissage contre 0.80 pour ces mêmes données avec un affaiblissement de 0.9. Excepté pour la première ligne du tableau pour laquelle $\alpha_{Bon} > \alpha_{Expert}$ nous avons bien pour les autres lignes des valeurs de α_P décroissantes de façon logique en accord avec le profil associé. Cependant pour les données de test, pour un affaiblissement de 0.9 commun à tous les contributeurs, le TBR est de 0.82. Seule la première ligne de valeurs du tableau 6.17 avec $\alpha_{Bon} > \alpha_{Expert}$ permet d'améliorer ce TBR. Cela s'explique par le taux de bonne classification des profils beaucoup plus faible pour cette campagne que pour les autres (voir tableau 6.10 page 118).

Comparaison du MV et des fonctions de croyance pour les données des campagnes multi_oiseaux

Pour les campagnes multi_oiseaux_précis et multi_oiseaux_imprécis, l'ensemble de réponses change d'une question à une autre et chaque espèce d'oiseau présentée en photo est unique dans la base de données. Puisqu'il n'y a pas de répétition dans les noms d'oiseaux attendus, il n'est pas possible de construire la matrice de confusion requise par EM. Ainsi pour ces deux ensembles de données la comparaison est réalisée exclusivement

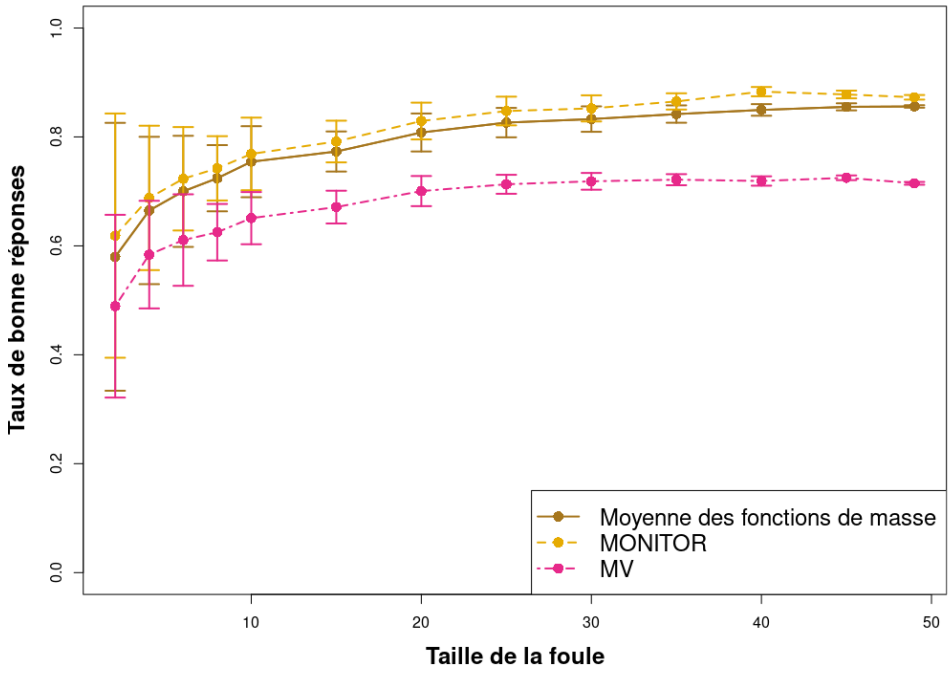


FIGURE 6.9 – Comparaison des fonctions de croyance au MV pour la campagne **multi_oiseaux_précis**.

entre le vote majoritaire et les fonctions de croyance. Les figures 6.9 et 6.10 présentent la comparaison entre le MV, la moyenne des fonctions de masse faite avec un coefficient d'affaiblissement $\alpha_P = 0.9$ pour tous les profils de contributeurs et l'agrégation réalisée par MONITOR. La combinaison faite par MONITOR utilise les valeurs d'affaiblissement suivantes d'après les profils des contributeurs :

$$\alpha_{Expert} = 1.00, \alpha_{Bon} = 0.85, \alpha_{Moyen} = 0.65, \alpha_{Mauvais} = 0.20$$

Pour les figures 6.9 et 6.10, les TBR sont calculés de manière analogue à la section 6.5.1 pour des tailles n de foule différentes. Ainsi, n contributeurs sont choisis aléatoirement parmi l'ensemble de contributeurs dédiés aux données de test et leurs réponses sont agrégées. Cette méthode de sélection et d'agrégation est réalisée 50 fois pour chaque taille n de foule et la moyenne des TBR est réalisée de sorte à obtenir les courbes des figures et leur intervalles de confiance à 95%.

Pour ces deux figures, les TBR sont croissants avec la taille n de la foule. Le TBR

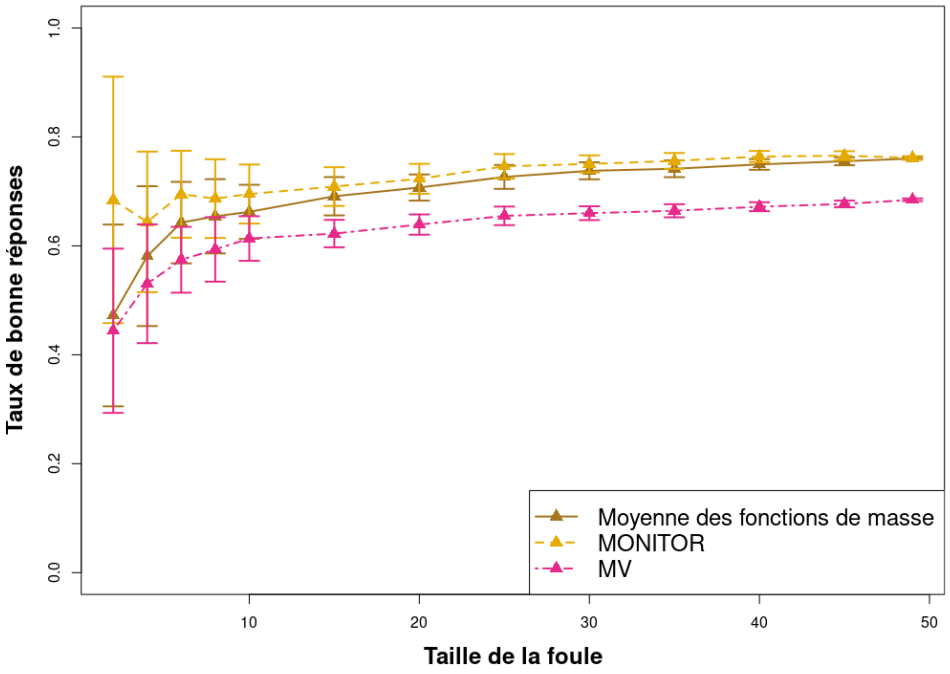


FIGURE 6.10 – Comparaison des fonctions de croyance au MV pour la campagne **multi_oiseaux_imprécis**.

de MONITOR est plus grand que celui des fonctions de croyance affaiblies de manière identique par $\alpha = 0.9$. La moyenne des fonctions de croyance donne de meilleurs résultats que le MV.

Comparaison du MV, de EM et des fonctions de croyance pour les données 10_oiseaux

L'utilisation des 10 mêmes espèces d'oiseau tout au long des campagnes 10_oiseaux_précis, 10_oiseaux_imprécis et 10_oiseaux_dynamique permet cette fois une comparaison du MV et des fonctions de croyance avec EM. Pour les données 10_oiseaux_dynamique, l'itération éventuelle sur la réponse n'est pas utilisée, seule la première contribution renseignée est exploitée. Pour obtenir les figures 6.11, 6.12 et 6.13 les contributions des jeux de test sont agrégées 25 fois pour une taille n de la foule.

Pour les figures 6.11 et 6.12 les valeurs α_P utilisées par MONITOR sont :

$$\alpha_{Expert} = 1.00, \alpha_{Bon} = 0.50, \alpha_{Moyen} = 0.25, \alpha_{Mauvais} = 0.20$$

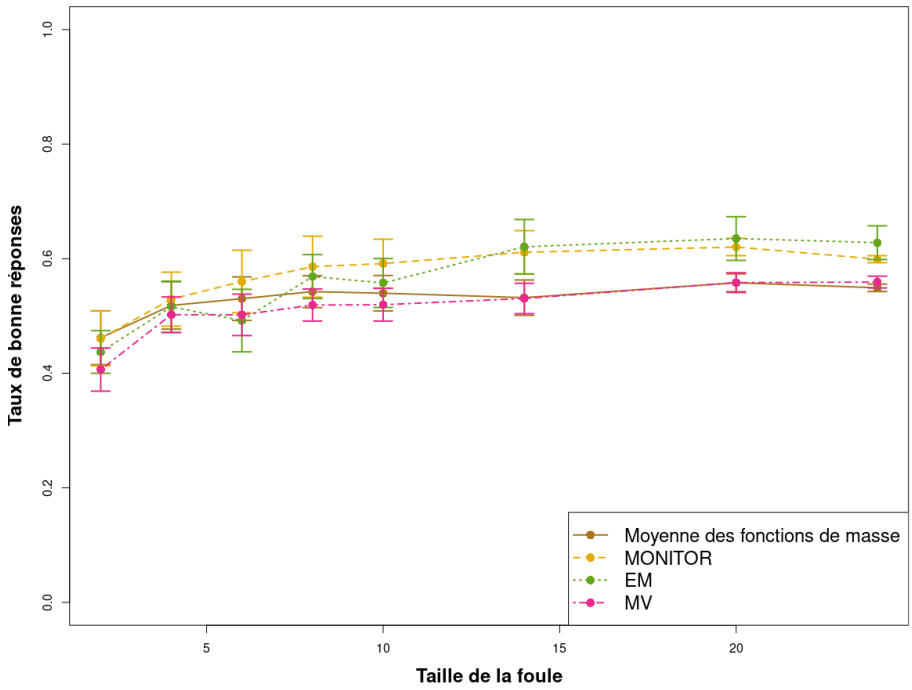


FIGURE 6.11 – Comparaison des méthodes d'agrégation pour la campagne **10_oiseaux_précis**.

La figure 6.11 montre que MONITOR donne de meilleurs résultats que le MV et la moyenne des fonctions de masse. Pour une foule de plus de 14 contributeurs, EM a de meilleurs résultats, mais la différence avec MONITOR n'est pas significative.

Pour la figure 6.12, les fonctions de croyance agrégées par la moyenne offrent globalement de meilleurs résultats que les autres méthodes d'agrégation, EM et MONITOR ont des TBR inférieurs au vote majoritaire.

Pour la figure 6.13, les coefficients d'affaiblissement dépendant du profil utilisés par MONITOR sont :

$$\alpha_{Expert} = 0.50, \alpha_{Bon} = 0.85, \alpha_{Moyen} = 0.20, \alpha_{Mauvais} = 0.20$$

Nous avons choisi ces valeurs, car ce sont celles donnant les meilleurs résultats sur les données de tests d'après le tableau 6.18. D'après la figure 6.13 MONITOR offre les TBR les plus élevés pour une taille de foule croissante, allant jusqu'à dépasser EM. Le vote majoritaire donne bien les plus mauvais résultats.

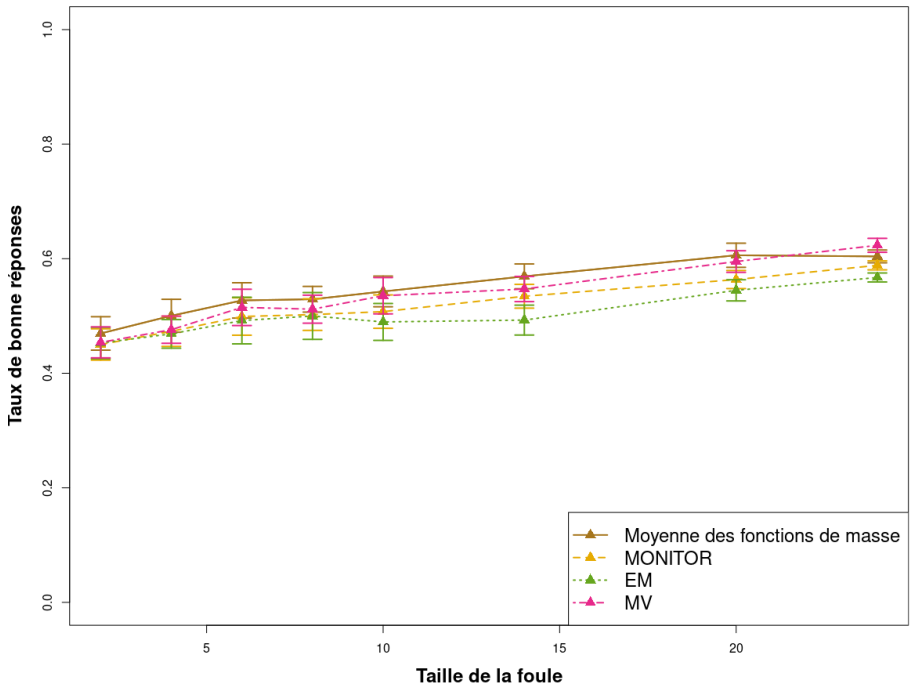


FIGURE 6.12 – Comparaison des méthodes d'agrégation pour la campagne **10_oiseaux _imprécis**.

Conclusion sur la comparaison des agrégations

Le tableau 6.19 résume les TBR obtenus pour chaque méthode d'agrégation pour la totalité des données de test. Pour les campagnes avec des espèces d'oiseaux différentes à identifier à chaque question (multi_oiseaux), MONITOR permet une amélioration des TBR comparé aux fonctions de croyance affaiblies par un même coefficient indépendamment du profil. Cette moyenne des fonctions de croyance ayant elle-même un meilleur TBR que le MV. Il n'est pas possible d'appliquer EM à ces campagnes. Pour les campagnes avec 10 oiseaux à identifier (10_oiseaux), la méthode d'agrégation donnant les meilleurs TBR change d'une campagne à une autre. Nous constatons par conséquent que la définition de la campagne a un fort impact sur les données collectées et la méthode d'agrégation à employer par la suite.

Nous nous sommes concentrées dans cette section sur l'utilisation de fonctions de masse à support simple. Cependant, nous pouvons imaginer l'utilisation d'autres types de fonctions de croyance dans le cadre du *crowdsourcing* comme des fonctions de masse consonantes, ce qui est fait dans la section suivante.

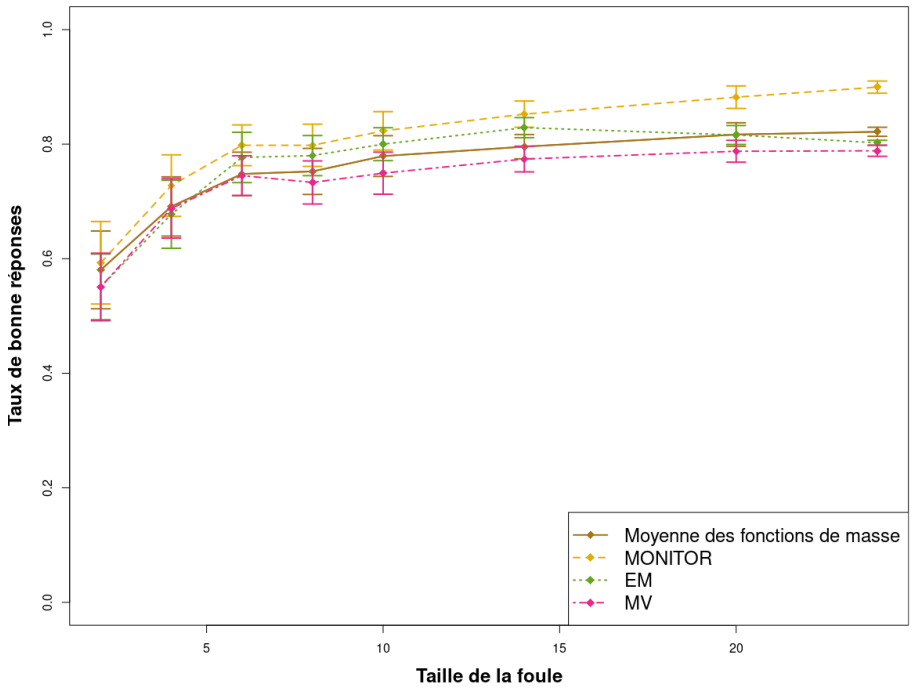


FIGURE 6.13 – Comparaison des méthodes d'agrégation pour la campagne **10_oiseaux_dynamique**.

6.5.3 Comparaison de fonctions de masse consonantes et à support simple

Nous souhaitions initialement réaliser une comparaison de MONITOR avec la méthode CASCADE de KOULOUGLI et al. 2016, qui propose une autre modélisation des réponses par les fonctions de croyance. Cependant deux problèmes se sont posés à nous, tout d'abord nous n'avons pas connaissance du test réalisé par les auteurs pour définir la qualification du contributeur, ensuite nous ne sommes pas en possession de données permettant d'utiliser le même type de fonctions de masse. Nous souhaitons tout de même confronter notre approche à une autre modélisation proche de CASCADE, nous avons ainsi réalisé la campagne de *crowdsourcing* 10_oiseaux_dynamique dont les données permettent de générer des fonctions de masse consonantes introduites dans la section 3.2.1.

Lors de la campagne 10_oiseaux_dynamique, une même question q peut être posée deux fois au contributeur c qui peut alors élargir ou préciser sa première réponse X_1 par une seconde réponse X_2 s'il le souhaite. Soit Ω_q l'ensemble des réponses proposé

Campagne	MV	EM	BF (Moyenne)	MONITOR
multi_oiseaux_précis	0.71	/	0.86	**0.86**
multi_oiseaux_imprécis	0.70	/	0.78	**0.8**
10_oiseaux_précis	0.56	**0.62**	0.54	0.54
10_oiseaux_imprécis	**0.64**	0.56	0.6	0.6
10_oiseaux_dynamique	0.77	0.80	0.82	**0.84**

TABLE 6.19 – Tableau comparatif des taux de bonne réponse pour les différentes méthodes d'agrégation sur les **données de test**.

et $X_1, X_2 \in 2^{\Omega}_q$. Si la première réponse du contributeur X_1 est précise et qu'il élargit sa seconde réponse X_2 alors $X_1 \subset X_2$, et réciproquement si X_1 est plus imprécise que X_2 alors $X_2 \subset X_1$. Lors de sa première sélection X_1, le contributeur renseigne un degré de certitude de valeur numérique $w_{cq1} \in [0, 1]$. S'il fait le choix de renseigner une seconde réponse X_2 il doit indiquer sa nouvelle certitude dont la valeur numérique est notée $w_{cq2} \in [0, 1]$.

S'il n'est pas demandé au contributeur de modifier sa sélection ou s'il ne souhaite pas le faire, la contribution est modélisée par une fonction de masse à support simple comme il est proposé dans la section 6.5.1 par l'équation (5.1). Dans le cas où le contributeur modifie sa réponse X_1 au profit de la réponse X_2, avec $X_1 \subset X_2$, alors la contribution est modélisée par une fonction de masse consonante comme suit :

$$\begin{cases} m^{\Omega_q}_{cq}(X_1) = \delta_1 * w_{cq1} \\ \\ m^{\Omega_q}_{cq}(X_2) = \delta_2 * w_{cq2} \\ \\ m^{\Omega_q}_{cq}(\Omega) = 1 - \delta_1 * w_{cq1} - \delta_2 * w_{cq2} \end{cases} \tag{6.6}$$

Dans l'équation (5.3), les coefficients δ_1 et δ_2 assurent que la fonction de masse appartient bien à l'intervalle $[0, 1]$, ainsi :

$$\delta_1 + \delta_2 = 1 \tag{6.7}$$

Le tableau 6.20 fait l'état des données provenant de la campagne oiseaux_10_xp3, cette base inclut un total de 2990 contributions. Au total il y a 1417 entrées pour lesquelles le contributeur a tout d'abord choisi une unique réponse $|X_1| = 1$, puis, il lui a été proposé d'élargir sa sélection de sorte que $|X_2| > 1$. Sur ces 1417 fois où il est proposé au contributeur d'être imprécis, il n'y a que 88 fois où une seconde réponse X_2 est donnée. De même, il y a au total 1133 fois où le contributeur renseigne une réponse imprécise,

Données	Nombre de réponses	Sous ensemble de données	Nombre de réponses du sous ensemble
oiseaux_10_xp3	2990	$\|X_1\| = 1$ et $\|X_2\| > 1$	1417
		$\|X_1\| > 1$ et $\|X_1\| > \|X_2\|$	1133
oiseaux_10_xp3 (consonante)	440	$\|X_1\| = 1$ et $\|X_2\| > 1$	88
		$\|X_1\| > 1$ et $\|X_1\| > \|X_2\|$	352

TABLE 6.20 – Nombres de contributions précises puis imprécises ($|X_1| = 1$ et $|X_2| > 1$) et imprécises puis moins imprécises ($|X_1| > |X_2|$).

$|X_1| > 1$ et il lui est proposé de restreindre sa sélection de sorte que $|X_1| > |X_2|$, au total 352 contributions font état d'un changement de réponse. Sur les 2990 données de la base oiseaux_10_xp3, 440 permettent de calculer des fonctions de masse consonantes, les données restantes sont modélisées par des fonctions de masse à support simple.

Dans la suite de cette section, nous notons m_1 la fonction de masse à support simple associée à la réponse X_1 du contributeur en absence de toutes modifications, et $m_{1,2}$ la fonction de masse consonante associée aux réponses X_1 et X_2.

Sur la figure 6.14, l'axe des ordonnées présente les TBR des données agrégées calculés sur les 50 photos pour les 2990 données en combinant fonctions de masse à support simple et consonantes. Les TBR des fonctions consonantes seules sur les 440 données permettant cette modélisation apparaissent également. Les opérateurs de combinaison utilisés sont l'opérateur conjonctif normalisé, équation (3.17) et la moyenne, équation (3.15). Nous refaisons ici une comparaison de ces deux opérateurs car les fonctions de croyance employées sont différentes de celles utilisées par MONITOR. L'axe des abscisses porte les différentes valeurs de δ_1 utilisées dans l'équation (6.7) pour le calcul des fonctions de masse consonantes. Plus δ_1 est grand, plus la valeur accordée à la première réponse X_1 donnée par le contributeur est importante, et par conséquent, comme $\delta_2 = 1-\delta_1$, la valeur accordée à la seconde réponse est faible. Ainsi pour $\delta_1 = 1$ seule la première réponse est considérée pour la fonction de masse consonante de sorte que $m_{1,2}$ est équivalente à m_1. D'après la figure, une agrégation conjointe de m_1 et $m_{1,2}$ offre un meilleur TBR plutôt que de ne considérer que les fonctions $m_{1,2}$. Il n'y a cependant que 440 réponses sur les 2990 collectées qui permettent de générer des fonctions de masse consonantes comme l'indique le tableau 6.20. Ainsi beaucoup moins de réponses sont prises en compte pour les calculs des TBR pour $m_{1,2}$ seule. Dans les deux cas, l'agrégation par la moyenne donne de meilleurs résultats que l'agrégation par l'opérateur conjonctif.

Pour l'agrégation des fonctions m_1 et $m_{1,2}$, le TBR augmente pour $\delta_1 \geq 0.4$ et diminue

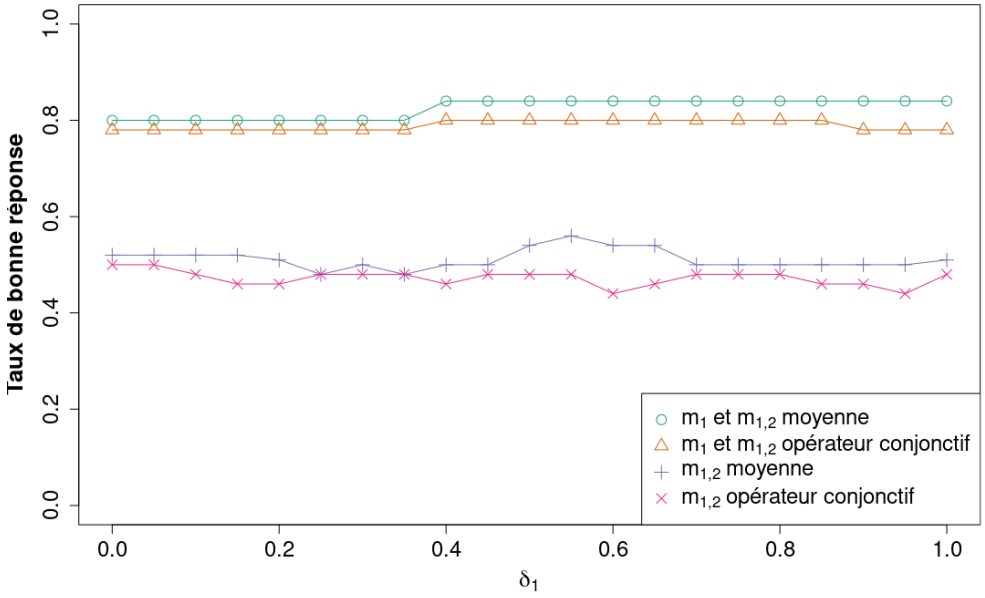


FIGURE 6.14 – Comparaison des taux de bonne réponse en fonction de δ_1 pour les données de la campagne oiseaux_10_xp3.

	$TBR_{MAX}(m_1, m_{1,2})$	$TBR(m_1)$
Moyenne	0.84	0.84
Opérateur Conjonctif	0.80	0.78

TABLE 6.21 – Comparaison des meilleurs taux de bonne réponse de $m_1 + m_{1,2}$ avec m_1 seule.

par la suite pour $\delta_1 \geq 0.9$ pour l'opérateur conjonctif. En revanche, les résultats de l'agrégation des fonctions $m_{1,2}$ seules ne semblent pas indiquer de corrélation évidente entre δ_1 et le TBR.

Le TBR maximal obtenu pour l'agrégation de m_1 et $m_{1,2}$ par la moyenne est de 0.84 contre 0.80 pour l'opérateur conjonctif avec les mêmes fonctions. Dans le cas où toutes les réponses ne sont modélisées que par m_1 uniquement, nous obtenons un TBR de 0.84 avec la moyenne et 0.78 avec l'opérateur conjonctif, ces résultats sont répertoriés dans le tableau 6.21. Pour la moyenne, les TBR sont dans le meilleur des cas égaux. Pour l'opérateur conjonctif l'utilisation de fonctions de masse consonantes permet d'améliorer le TBR lorsque $\delta_1 < \delta_2$, soit lorsque plus de poids est accordé à la seconde réponse.

Pour les 440 données modélisées uniquement par les fonctions de masse consonantes

	$TBR_{MAX}(m_{1,2})$	$TBR(m_1)$
Moyenne	0.56	0.51
Opérateur Conjonctif	0.50	0.48

TABLE 6.22 – Comparaison des meilleurs taux de bonne réponse de $m_{1,2}$ avec m_1 seule.

$m_{1,2}$, le TBR maximal est de 0.56 pour la moyenne et 0.50 pour l'opérateur conjonctif (voir tableau 6.22). Dans le cas où seules les premières réponses X_1 de ces données sont modélisées par des fonctions de masse à support simple alors nous obtenons un TBR de 0.51 pour la moyenne et 0.48 pour l'opérateur conjonctif. Les fonctions de masse consonantes améliorent donc les TBR.

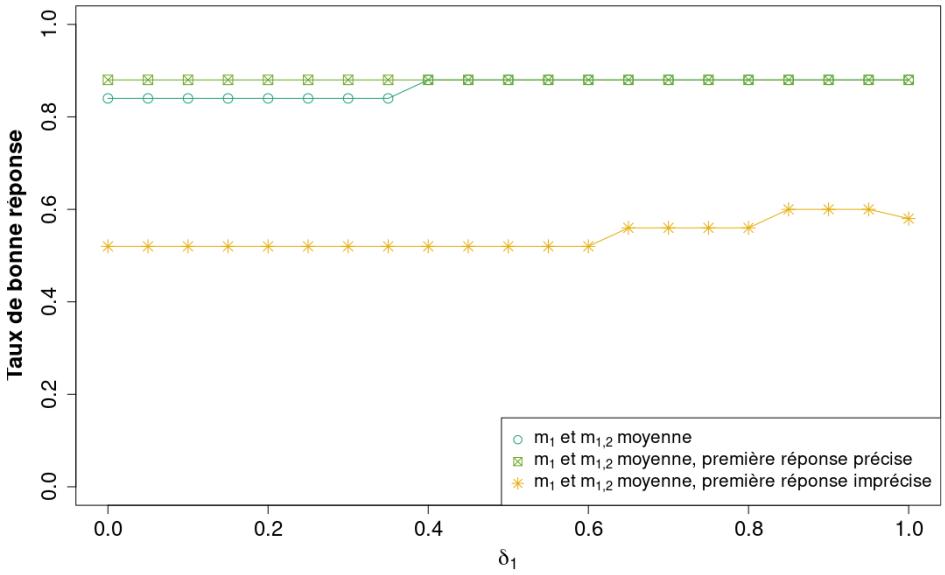
Intéressons nous maintenant à la dissociation de l'agrégation des réponses d'abord précises puis imprécises des autres. Pour ce faire, nous scindons pour chacune des courbes de la figure 6.14 notre ensemble de données en deux :

— les données où le contributeur sélectionne un unique nom d'oiseau pour sa première réponse ($|X_1| = 1$) puis éventuellement il élargit sa sélection ($|X_2| > |X_1|$)

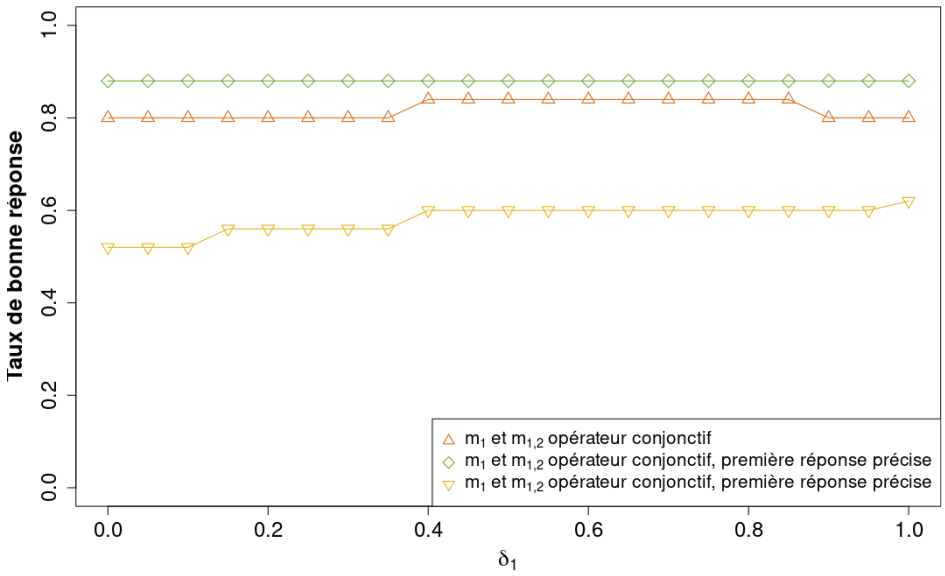
— les données où le contributeur est tout d'abord imprécis ($|X_1| > 1$) puis éventuellement il restreint sa sélection ($|X_1| > |X_2|$)

Pour ces deux ensembles, les réponses sont modélisées en utilisant les fonctions m_1 et $m_{1,2}$ pour la figure 6.15 et uniquement $m_{1,2}$ pour la figure 6.16 en faisant varier la valeur de δ_1. Les fonctions de croyance sont agrégées par la moyenne pour les figures 6.15a et 6.16a, et par l'opérateur conjonctif pour les figures 6.15b et 6.16b. Dans les quatre cas les TBR calculés sont comparés aux courbes de la figure 6.14. Contrairement à la figure 6.14 où le TBR est calculé sur les 50 photos à annoter, les taux des figures 6.15 et 6.16 sont calculés sur 25 photos. En effet, ces 25 photos sont toutes annotées et permettent de calculer $m_{1,2}$ seule. Pour les 25 photos restantes les contributeurs sont toujours précis et certains, ou alors ils ne changent pas de réponse de sorte que X_2 est inexistante.

Pour la figure 6.15, il y a 1417 réponses d'abord précises puis imprécises et 1133 réponses imprécises puis plus précises voir tableau 6.20. Les TBR des contributions d'abord imprécises puis précises sont beaucoup plus faibles (courbe jaune), comparés aux deux autres courbes. Les taux des courbes jaunes croissent avec δ_1 ce qui signifie que donner plus de poids à la réponse imprécise est pertinent ici. La courbe des réponses précises puis imprécises est continue, peu importe la valeur de δ_1. Cela s'explique par le fait que seules 88 réponses sur les 1417 permettent de faire des fonctions de masse consonantes donc leur poids dans l'agrégation est faible.

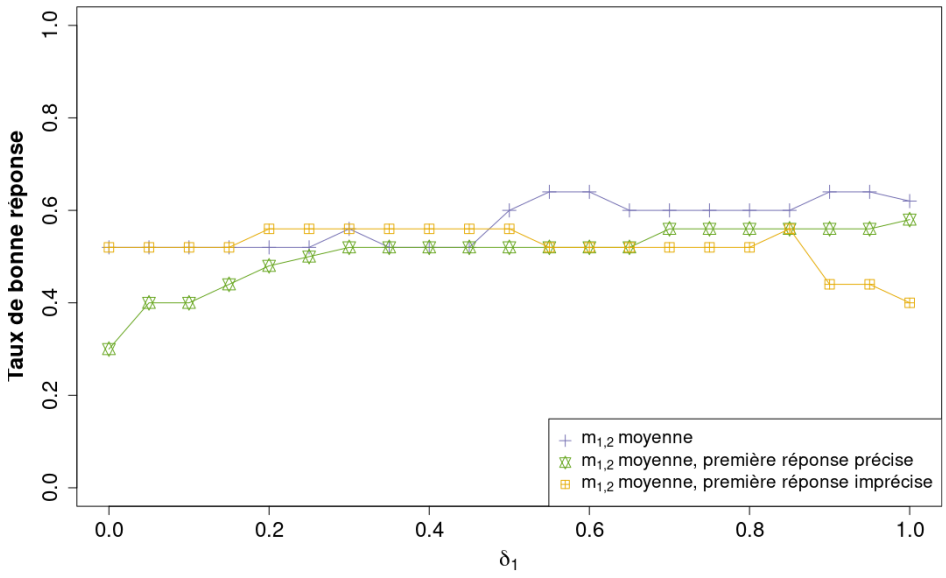


(a) Taux de bonne réponse pour une agrégation par la moyenne.

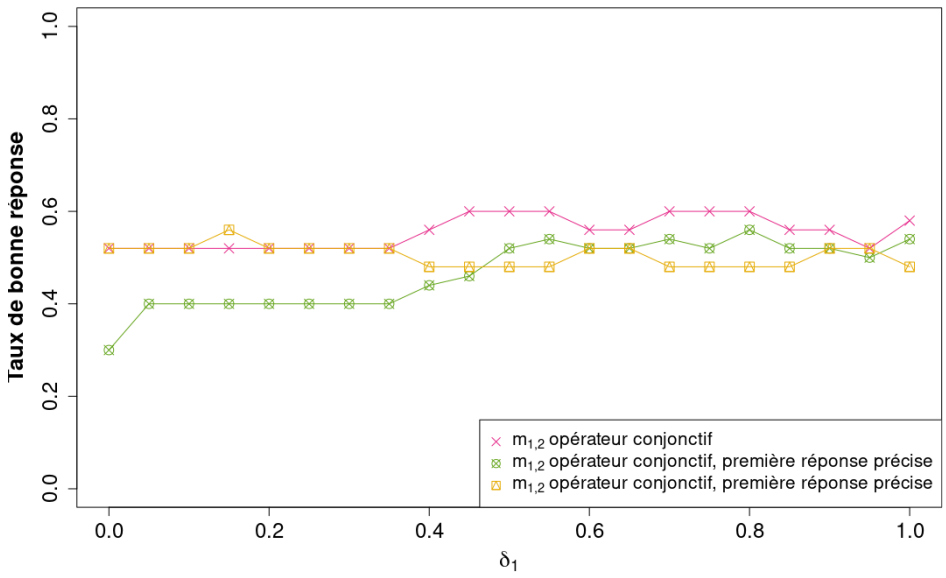


(b) Taux de bonne réponse pour une agrégation par l'opérateur conjonctif.

FIGURE 6.15 – Taux de bonne réponse pour des fonctions de masse à support simple (m_1) et des fonctions de masse consonantes ($m_{1,2}$).



(a) Taux de bonne réponse pour une agrégation par la moyenne.



(b) Taux de bonne réponse pour une agrégation par l'opérateur conjonctif.

FIGURE 6.16 – Taux de bonne réponse pour des fonctions de masse consonantes (m6).

Pour la figure 6.16, il y a 88 réponses d'abord précises puis imprécises et 352 réponses imprécises puis plus précises (voir tableau 6.20). D'après les figures 6.16a et 6.16b, lorsque δ_1 augmente, le TBR pour les contributions d'abord précises puis imprécises (courbe verte) augmente aussi, et à l'opposé, le TBR pour les contributions d'abord imprécises puis précises diminue. Il semble plus intéressant lorsque les réponses sont modélisées uniquement par des fonctions de masse consonantes de donner davantage de poids à la réponse la plus précise des deux.

Les fonctions de croyance consonantes permettent une légère amélioration des TBR des données collectées. Cette modélisation bien qu'intéressante apporte un faible gain sur la qualité des données car les contributeurs utilisent peu la possibilité de changer de réponse. Ainsi l'impact des fonctions de masse consonantes est limité lors de l'agrégation. Il faudrait réaliser un travail approfondi sur l'interface pour exploiter le potentiel de cette approche, ceci est détaillé dans les perspectives de la thèse.

6.6 Conclusion

Dans ce chapitre nous présentons les résultats expérimentaux de la validation de MONITOR. Afin de réaliser nos tests sur des données réelles nous avons réalisé cinq campagnes de *crowdsourcing* qui consistent en l'annotation de photos d'oiseaux. Pour toutes les campagnes, il est demandé au contributeur de renseigner sa certitude en sa réponse. Pour les deux premières campagnes, les réponses proposées changeaient à chaque question, pour l'une il est exigé du contributeur une réponse précise (multi_oiseaux_précis), pour la seconde le contributeur peut être imprécis et sélectionner plusieurs réponses (multi_oiseaux_imprécis). Pour les trois autres campagnes, seules dix espèces d'oiseaux sont à identifier, et les dix noms sont proposés à chaque question. Pour la campagne 10_oiseaux_précis le contributeur doit sélectionner un unique nom, alors que pour les campagnes 10_oiseaux_imprécis et 10_oiseaux_dynamique il peut être imprécis et choisir jusqu'à 5 noms d'oiseaux. Ces données nous ont permis de réaliser des expériences sur les éléments qui composent le profil, sur le profil lui même et l'agrégation des réponses.

Pour estimer le profil du contributeur, MONITOR considère sa qualification pour la tâche à travers son imprécision et sa certitude, mais aussi son comportement, afin de savoir si les réponses du contributeur sont réfléchies et s'il est attentif. Nous avons comparé l'estimation de l'imprécision par MONITOR au degré de Ben Rjab et al. 2016 dont elle s'inspire. Notre modélisation de l'imprécision du contributeur est plus en accord

avec l'imprécision moyenne réelle de ce dernier que le degré proposé par les auteurs.

Pour la vérification de la réflexion du contributeur, nous comparons notre méthode de calcul utilisant les fonctions de croyance à une approche statistique de sélection de contributeurs qu'utilisent Komarov et al. 2013. L'approche des auteurs est limitée par la répartition statistique des temps de réponse des contributeurs qui peut être source d'erreurs. MONITOR offre une estimation pertinente de la réflexion en fonction du temps de réponse du contributeur. Cependant, la réflexion ne peut être considérée seule pour exclure un contributeur car un contributeur non-réfléchi peut être un *spammer* comme un expert. Il est ainsi essentiel de considérer la réflexion conjointement à l'attention afin de déterminer le comportement du contributeur.

Les expériences réalisées pour l'estimation du profil par MONITOR en comparaison de l'estimation de Ben Rjab et al. 2016 et de EM montrent que MONITOR offre un meilleur taux de bonne classification du profil que Ben Rjab et al. 2016 tout en étant moins coûteux en temps de calcul. En revanche, une estimation du profil du contributeur grâce à EM reste plus performante que les estimations obtenues par MONITOR. Cependant, il n'est pas toujours possible d'appliquer l'algorithme EM aux données collectées, notamment dans le cas où l'ensemble de réponses proposé au contributeur change d'une question à une autre. MONITOR en revanche n'est pas impacté par cette problématique et est applicable à tout type de données imprécises et incertaines.

Pour les données multi_oiseau, il n'est pas possible d'effectuer une comparaison avec EM. Les résultats obtenus sur ces données montrent que l'affaiblissement des contributions d'après les profils estimés par MONITOR offrent de meilleurs taux de bonne réponse qu'un affaiblissement commun à l'ensemble des données. Pour les deux campagnes, les fonctions de croyance offrent de meilleurs résultats que le MV traditionnellement employé dans les plateformes de *crowdsourcing*. Pour les campagnes avec 10 espèces d'oiseaux récurrentes à identifier, la méthode donnant les meilleurs taux de bonne réponse change d'une campagne à une autre entre le MV, EM et MONITOR. Ceci montre que la définition de la campagne a un fort impact sur les contributions collectées et la méthode d'agrégation à employer par la suite. Pour des campagnes similaires à celles avec une multitude d'oiseaux, où un apprentissage par le contributeur est plus complexe, la méthode la plus appropriée à employer est MONITOR. Dans le cas de questions redondantes, il est possible d'utiliser EM comme MONITOR. Le modèle peut encore être optimisé en vue d'une utilisation dynamique durant la campagne de *crowdsourcing*, cela fait l'objet de nos perspectives de recherche développées dans la section suivante.

Chapitre 7

Conclusion

Résumé Dans ce chapitre nous concluons cette thèse et présentons les perspectives de recherche qu'elle amène. Ce travail est réalisé dans le cadre du projet ANR Headwork qui a pour objectif de définir une nouvelle plateforme de *crowdsourcing* qui permet une meilleure gestion des *workflows* des contributeurs et des contributions. Notre but dans ce projet est la modélisation des contributions et des contributeurs au sein de la plateforme. Pour réaliser au mieux nos expériences nous avons également été amenées à définir une nouvelle interface de *crowdsourcing* qui offre au contributeur la possibilité d'être imprécis et incertain. Les résultats que nous avons établis dans les chapitres 4 et 6 sont concluants mais réalisés dans un cadre statique, ce qui nous amène à réfléchir à la réalisation de nos travaux dans un contexte dynamique à l'avenir.

Sommaire

7.1 Conclusion

Le *crowdsourcing* est une nouvelle forme de travail massivement participatif et dématérialisé qui s'inscrit dans l'air du *Big Data*. Ce mode de travail permet à un employeur d'externaliser une tâche sur une plateforme en ligne où une foule de contributeurs la réalise. Il existe plusieurs types de plateformes de *crowdsourcing*. Dans le cadre de cette thèse nous nous focalisons sur le *crowdsourcing* d'activités routinières qui consiste en la réalisation de micro-tâches par le contributeur contre une faible gratification. Les problématiques qui touchent le domaine sont d'ordres juridique, social et technique mais en l'absence de connaissances en droit nous nous consacrons uniquement aux problématiques techniques et sociales.

Parmi les problématiques sociales, la diversité des profils des contributeurs est un enjeu important. En effet, les tâches de *crowdsourcing* d'activités routinières sont simples et accessibles à tous c'est pourquoi des profils variés composent la foule. Certains contributeurs ont un comportement malveillant et il faut traiter leurs réponses en conséquence. D'autres à l'inverse font preuve de sérieux dans leur travail et peuvent même avoir une expertise particulière pour la tâche. L'idéal est de considérer le profil lors de l'agrégation des réponses, mais, à l'heure actuelle, il n'existe pas de méthode clairement définie pour cela. En effet, le vote majoritaire couramment employé pour l'agrégation des réponses ne tient compte ni de l'incertitude sur la réponse du contributeur, ni du profil de ce dernier. Une autre approche consiste à utiliser l'algorithme EM qui utilise des probabilités et une matrice de confusion sur la réponse du contributeur. Cependant, si cette matrice de confusion est révélatrice de la qualification du contributeur, elle ne rend pas compte de son comportement qui est pourtant un élément important. A dire vrai, il n'existe pas de méthode qui permette d'estimer le profil du contributeur d'après sa qualification pour la tâche et son comportement conjointement en l'absence de données d'or et qui permette l'agrégation des contributions.

Avant l'agrégation des réponses et la considération du profil du contributeur l'employeur doit en premier lieu s'assurer que la tâche qu'il met en ligne est clairement définie. Plusieurs retours utilisateurs et travaux de l'existant montrent que la qualité des données collectées peut être impactée négativement par une tâche mal explicitée ou une interface inappropriée. L'existant montre que l'interface utilisateur doit être ergonomique et simple pour ne pas être trop distrayante. Cependant, il n'existe pas d'interface généralisée pour le *crowdsourcing*, et il n'est généralement pas possible pour le contributeur de s'exprimer

sur sa contribution.

Au cours de cette thèse, nous avons créé une interface pour le contributeur dans les plateformes de *crowdsourcing* introduite dans l'article Thierry et al. 2020a. Là où les interfaces traditionnelles requièrent du contributeur une réponse précise, l'interface que nous proposons lui permet de sélectionner une à plusieurs réponses en cas de doute tout en donnant sa certitude. L'interface a été testée par des contributeurs sur la plateforme de *crowdsourcing Crowdpanel*. Les données collectées ont permis de mettre en évidence la relation entre la difficulté de la tâche, la certitude du contributeur et son usage de la possibilité d'être imprécis. La certitude du contributeur diminue lorsque la difficulté de la tâche augmente, et il est plus imprécis lorsque la tâche est plus complexe. En comparant les données d'une campagne avec imprécision avec celle d'une campagne où le contributeur est contraint d'être précis nous constatons qu'en étant imprécis le contributeur est plus certain. L'analyse des données nous a ainsi permis de valider, dans un contexte de *crowdsourcing*, l'hypothèse de Smets 1997 d'après laquelle plus un contributeur est imprécis plus il est certain, et réciproquement plus il est précis moins il est certain. Ce travail a fait l'objet d'une publication Thierry et al. 2021. D'après les retours des utilisateurs, en utilisant l'interface que nous proposons avec imprécision ils éprouvent moins de difficulté à réaliser la campagne de *crowdsourcing* que dans le cas où ils sont contraints de choisir une unique réponse. L'interface leur permet notamment de moins hésiter avant de répondre. Cette interface est également intéressante pour l'employeur car bien qu'elle soit plus coûteuse en temps pour le contributeur qu'une interface traditionnelle, elle permet de diminuer la taille de la foule requise et de réaliser un gain financier tout en ayant un meilleur taux de bonne réponse.

Afin de traiter l'imprécision et l'incertitude des réponses collectées grâce à cette interface nous utilisons la théorie des fonctions de croyance qui permet de modéliser ces imperfections. Cette théorie est utilisée par MONITOR qui est le modèle défini au cours de cette thèse pour représenter les contributions et estimer les profils des contributeurs. Pour réaliser l'estimation du profil du contributeur, MONITOR considère à la fois sa qualification pour la tâche et son comportement. La qualification du contributeur caractérise l'imprécision de ses réponses et sa certitude associée. Le comportement est représentatif de la réflexion du contributeur et de son attention dans la réalisation de la tâche. Le poids accordé à la réponse du contributeur est ensuite pondéré par le profil estimé par MONITOR lors de l'agrégation des contributions.

Afin de tester MONITOR sur des données réelles des campagnes de *crowdsourcing* ont

été réalisées avec l'interface proposée. Il s'agit pour le contributeur d'annoter des images d'oiseaux. Une partie des résultats des expériences réalisées est publiée dans Thierry et al. 2018 ; Thierry et al. 2019 ; Thierry et al. 2020b. Les expériences réalisées montrent que l'utilisation de l'algorithme EM reste plus pertinente que MONITOR. Cependant la définition des campagnes ne permet pas toujours d'utiliser cet algorithme, contrairement à MONITOR qui est moins contraignant dans son contexte d'utilisation. Les résultats pour l'estimation du profil par MONITOR restent encourageants au vu de la difficulté d'estimer le profil en l'absence de données d'or. En affaiblissant les fonctions de masse qui modélisent les contributions par les profils des contributeurs correspondants nous obtenons une amélioration du taux de bonne réponse en comparaison à l'agrégation des fonctions sans pondération. De plus la modélisation et l'agrégation des réponses par des fonctions de croyance offrent de meilleurs résultats en comparaison au vote majoritaire usuellement utilisé dans les plateformes de *crowdsourcing* et parfois même de meilleurs résultats que EM.

Suite aux travaux de cette thèse plusieurs perspectives de recherche, que nous présentons dans la section suivante, apparaissent.

7.2 Perspectives

MONITOR est défini et testé dans un contexte statique, nous envisageons de faire évoluer le modèle de sorte qu'il puisse être utilisé dans un contexte dynamique et réaliser une estimation du profil des contributeurs en ligne sur la plateforme Headwork et non après la campagne. De plus, après les campagnes de *crowdsourcing* réalisées grâce à l'interface proposée dans cette thèse, nous imaginons d'améliorer l'interface de sorte qu'elle s'adapte davantage au besoin du contributeur. Les sections suivantes présentent ces perspectives envisagées à court, moyen et long terme.

7.2.1 Perspectives à court terme

Afin d'estimer la réflexion du contributeur dans sa réponse MONITOR fait une comparaison du temps de réponse du contributeur T_{cq} à un temps de réponse théorique attendu T_{0q}. A l'heure actuelle T_{0q} est calculé en fin de campagne d'après les temps de réponse de tous les contributeurs à la question q. La réflexion du contributeur est un élément de MONITOR plus vague à estimer, comparé à l'imprécision, la certitude et l'attention.

C'est pourquoi nous pensons intéressant dans les mois à venir d'essayer de calculer la réflexion du contributeur par une approche floue. Cette idée nécessite un travail sur la mise en relation de la théorie des ensembles flous et la théorie des fonctions de croyance.

Puisque nous envisageons une utilisation de MONITOR en ligne sur Headwork, il faut l'adapter à une utilisation dynamique en commençant par simuler la collecte des contributions au cours du temps. Il s'agit ainsi de tester l'estimation de la qualification et celle du comportement en traitant les données non pas dans leur totalité mais en les ajoutant progressivement comme les réponses arrivent au cours du temps lors d'une campagne de *crowdsourcing*. Cette simulation peut se faire sur les données déjà collectées durant la thèse en traitant les données d'après l'ordre dans lequel elles ont été recueillies par exemple. L'objectif de ces tests est notamment de déterminer les coûts en temps de calcul et en mémoire pour une estimation du profil en temps réel afin de travailler par la suite sur ces problématiques éventuelles.

7.2.2 Perspectives à moyen terme

Nous avons défini une interface de *crowdsourcing* qui offre au contributeur la possibilité d'être imprécis en sélectionnant plusieurs réponses en cas d'hésitation tout en donnant sa certitude sur sa contribution. Pour une de nos expériences, nous avons fait évoluer cette interface afin d'introduire une nouvelle interactivité avec le contributeur. Ainsi, une image d'oiseau est présentée lors de la tâche avec une proposition de dix noms dans le but que le contributeur identifie l'oiseau. Dans le cas où le contributeur est imprécis et sélectionne plusieurs noms, la question avec la même image est reposée afin qu'il diminue sa sélection si possible tout en indiquant sa nouvelle certitude. A l'inverse, si le contributeur est précis en choisissant une unique réponse mais n'est pas totalement sûr de cette dernière, il lui est proposé d'élargir sa sélection en choisissant d'autres noms afin d'être plus certain. Cette interface ne repose la question qu'une fois au contributeur. Il serait intéressant d'améliorer l'interface de sorte que la question ne soit posée que lorsque le contributeur souhaite réellement modifier sa réponse où lorsqu'il y a un réel intérêt sur l'indécision de la réponse. Il serait éventuellement possible de reposer la question plus d'une itération afin de lui permettre d'améliorer au mieux sa contribution. Cette version améliorée de l'interface nécessiterait *a priori* de recourir à de l'apprentissage automatique.

Il serait souhaitable sur le moyen terme que l'estimation du comportement et de la qualification soit réalisable en temps réel sans aller jusqu'à l'estimation du profil dans un premier temps car cette estimation se fait avec de l'apprentissage. L'objectif à ce

terme est d'optimiser la méthode d'agrégation de l'imprécision, la certitude, la réflexion et l'attention pour l'estimation du profil afin qu'elle soit moins laborieuse et ne requière pas nécessairement un apprentissage qui peut être pénalisant en temps réel.

7.2.3 Perspectives à long terme

Sur le long terme nous souhaitons que l'évaluation du profil soit réalisable en temps réel au cours de la campagne de *crowdsourcing*, avec une intégration de MONITOR à Headwork. Grâce à cette évaluation en ligne du contributeur il serait alors possible d'interagir avec lui pour l'aider à réaliser au mieux la tâche mais aussi dissuader les contributeurs malveillants. On peut ainsi imaginer que si un contributeur manque de sérieux au cours de la tâche, lorsque cela est établi par MONITOR, la plateforme lui envoie un avertissement pour qu'il se reconcentre à moins de se voir exclure de la campagne. Dans un autre cas, où deux contributeurs sont identifiés comme un expert pour l'un et un contributeur bienveillant mais en difficulté pour l'autre, il serait opportun de proposer à l'expert de venir en aide à un contributeur moins qualifié et, réciproquement, proposer au contributeur en difficulté une aide par un tiers. De même, lorsque l'interface proposée sera suffisamment évoluée on peut imaginer que les questions s'adaptent au profil estimé du contributeur au cours du temps.

L'ensemble de ces perspectives sont proposées dans le contexte du *crowdsourcing*, mais on peut à long terme imaginer une application de MONITOR dans d'autres contextes comme des cours en ligne de type MOOC. En effet, avec la situation de la crise sanitaire causée par la covid-19, les cours en ligne ont pris un essor important et se sont beaucoup diversifiés. Grâce à MONITOR il serait possible de déterminer si la personne est en difficulté lors du cours et de lui proposer une aide par un autre élève, qui lui, aurait été estimé comme "expert" par le modèle. Cela permettrait de créer un système d'entre-aide et créer une forme de cohésion sociale dans le cours tout en permettant aux personnes en difficulté de s'améliorer.

RÉFÉRENCES

Publications

Thierry, C., Martin, A., Dubois, J.-C. et Le Gall, Y. (2021), « Validation of Smets' hypothesis in the crowdsourcing environment », in*6th International Conference on Belief Functions* (cf. p. 6, 59, 72, 84, 145).

Thierry, C., Casiez, G., Dubois, J.-C., Le Gall, Y., Malacria, S., Martin, A., Pietrzak, T. et Uro, P. (2020a), « Interface de Recueil de Données Imparfaites pour le CrowdSourcing », in*EGC 2020-Humains et IA, travailler en intelligence Atelier de la conférence* (cf. p. 5, 59, 84, 145).

Thierry, C., Dubois, J.-C., Le Gall, Y. et Martin, A. (2018), « Modélisation du profil des contributeurs dans les plateformes de crowdsourcing », in*27èmes rencontres francophones sur la logique floue et ses applications* (cf. p. 7, 88, 100, 146).

— (2019), « Modeling uncertainty and inaccuracy on data from crowdsourcing plateforms : MONITOR », in*Proceedings of the 31st International Conference on Tools with Artificial Intelligence* (cf. p. 7, 88, 100, 146).

— (2020b), « Modelisation de l'incertitude et de l'imprecision de donnees de crowdsourcing : MONITOR », in*Extraction et Gestion des Connaissances (EGC)* (cf. p. 7, 88, 100, 146).

Bibliographie

Abassi, L. et Boukhris, I. (2018), « A worker clustering-based approach of label aggregation under the belief function theory », in*Applied Intelligence*, p. 1-10 (cf. p. 55, 56, 90).

Allahbakhsh, M., Benatallah, B., Ignjatovic, A., Motahari-Nezhad, H. R., Bertino, E. et Dustdar, S. (2013), « Quality control in crowdsourcing systems : Issues and directions », in*IEEE Internet Computing* 17.*2*, p. 76-81 (cf. p. 23).

Ben Rjab, A., Kharoune, M., Miklos, Z. et Martin, A. (2016), « Characterization of experts in crowdsourcing platforms », in*International Conference on Belief Functions*, Springer, p. 97-104 (cf. p. 53-56, 91-93, 107, 115, 118, 119, 121, 122, 140, 141).

Blanco, H. H. R. (2012), « Machine-Learning for Spammer Detection in Crowd-Sourcing », in*Human Computation AAAI Technical Report* (cf. p. 33).

Boim, R., Greenshpan, O., Milo, T., Novgorodov, S., Polyzotis, N. et Tan, W.-C. (2012), « Asking the right questions in crowd data sourcing », in*2012 IEEE 28th International Conference on Data Engineering*, IEEE, p. 1261-1264 (cf. p. 30).

Burger-Helmchen, T. et Pénin, J. (2011), « Crowdsourcing : définition, enjeux, typologie », in*Management & Avenir* 41, p. 254-269 (cf. p. 13).

Chen, K.-T., Chang, C.-J., Wu, C.-C., Chang, Y.-C. et Lei, C.-L. (2010), « Quadrant of euphoria : a crowdsourcing platform for QoE assessment », in*IEEE Network* 24.*2*, p. 28-35 (cf. p. 13).

Chittilappilly, A. I., Chen, L. et Amer-Yahia, S. (2016), « A survey of general-purpose crowdsourcing techniques », in*IEEE Transactions on Knowledge and Data Engineering* 28.*9*, p. 2246-2266 (cf. p. 12, 19, 22, 23).

Chu, X., Morcos, J., Ilyas, I. F., Ouzzani, M., Papotti, P., Tang, N. et Ye, Y. (2015), « Katara : A data cleaning system powered by knowledge bases and crowdsourcing », in*Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, ACM, p. 1247-1261 (cf. p. 16).

Côme, E., Oukhellou, L., Denoeux, T. et Aknin, P. (2009), « Learning from partially supervised data using mixture models and belief functions », in*Pattern recognition* 42.*3*, p. 334-348 (cf. p. 52).

Create an Amazon Mechanical Turk project ((09/09/2021)), url : `https://docs.aws.amazon.com/AWSMechTurk/latest/RequesterUI/CreatingYourBatchofHITs.html` (cf. p. 24).

Dawid, A. P. et Skene, A. M. (1979), « Maximum likelihood estimation of observer error-rates using the EM algorithm », in*Applied statistics*, p. 20-28 (cf. p. 34, 35, 53, 75).

Delmotte, F. et Smets, P. (2004), « Target identification based on the transferable belief model interpretation of Dempster-Shafer model », in*IEEE Transactions on Systems, Man, and Cybernetics-Part A : Systems and Humans* 34.*4*, p. 457-471 (cf. p. 44, 52).

Dempster, A. P. (1967), « Upper and Lower Probabilities Induced by a Multivalued Mapping », in*The Annals of Mathematical Statistics* 38, p. 325-339 (cf. p. 40).

Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977), « Maximum likelihood from incomplete data via the EM algorithm », in*Journal of the royal statistical society. Series B (methodological)*, p. 1-38 (cf. p. 34).

Denœux, T. (2006), « The cautious rule of combination for belief functions and some extensions », in*2006 9th International Conference on Information Fusion*, IEEE, p. 1-8 (cf. p. 48).

Denoeux, T. (1995), « A k-nearest neighbor classification rule based on Dempster-Shafer theory », in*IEEE transactions on systems, man, and cybernetics* 25.*5*, p. 804-813 (cf. p. 52).

— (2018), « Logistic regression, neural networks and dempster-shafer theory : a new perspective », in*arXiv preprint arXiv :1807.01846* (cf. p. 51).

Denoeux, T. et Smets, P. (2006), « Classification using belief functions : relationship between case-based and model-based approaches », in*IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 36.*6*, p. 1395-1406 (cf. p. 52).

Diaz, J., Rifqi, M., Bouchon-Meunier, B., Jhean-Larose, S. et Denhiére, G. (2008), « Imperfect answers in multiple choice questionnaires », in*European Conference on Technology Enhanced Learning*, Springer, p. 144-154 (cf. p. 26, 58).

Difallah, D. E., Demartini, G. et Cudré-Mauroux, P. (2012), « Mechanical Cheat : Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms », in *CrowdSearch*, p. 26-30 (cf. p. 24, 94).

Dow, S., Kulkarni, A., Klemmer, S. et Hartmann, B. (2012), « Shepherding the crowd yields better work », in*Proceedings of the ACM 2012 conference on computer supported cooperative work*, p. 1013-1022 (cf. p. 30).

Drapeau, R., Chilton, L. B., Bragg, J. et Weld, D. S. (2016), « Microtalk : Using argumentation to improve crowdsourcing accuracy », in*Fourth AAAI Conference on Human Computation and Crowdsourcing* (cf. p. 30).

Essaid, A., Martin, A., Smits, G. et Yaghlane, B. B. (2014), « Uncertainty in ontology matching : a decision rule-based approach », in*International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Springer, p. 46-55 (cf. p. 50, 51).

Farrell, G. (jan. 2006), *A comparison of an innovative web-based assessment tool utilizing confidence measurement to the traditional multiple choice, short answer and problem solving questions* (cf. p. 58).

Felstiner, A. (2011), « Working The Crowd : Employment And Labor Law In The Crowdsourcing Industry », in*Berkeley Journal of Employment & Labor Law* 32, p. 142-204 (cf. p. 16, 18).

Finnerty, A., Kucherbaev, P., Tranquillini, S. et Convertino, G. (2013), « Keep it simple : Reward and task design in crowdsourcing », in*Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, ACM, p. 14 (cf. p. 25).

Fixsen, D. et Mahler, R. P. (1997), « The modified Dempster-Shafer approach to classification », in*IEEE Transactions on Systems, Man, and Cybernetics-Part A : Systems and Humans* 27.*1*, p. 96-104 (cf. p. 52).

Florea, M. C., Jousselme, A.-L., Bossé, É. et Grenier, D. (2009), « Robust combination rules for evidence theory », in*Information Fusion* 10.*2*, p. 183-197 (cf. p. 49).

Folorunso, O. et Mustapha, O. A. (2015), « A fuzzy expert system to Trust-Based Access Control in crowdsourcing environments », in*Applied Computing and Informatics* 11.*2*, p. 116-129 (cf. p. 29, 33).

Fort, K. (2017), « Experts ou (foule de) non-experts ? La question de l'expertise des annotateurs vue de la myriadisation (crowdsourcing) », in*Corela. Cognition, représentation, langage* HS-21 (cf. p. 16, 27).

Gadiraju, U., Fetahu, B., Kawase, R., Siehndel, P. et Dietze, S. (2017), « Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks », in*ACM Transactions on Computer-Human Interaction (TOCHI)* 24.*4*, p. 30 (cf. p. 29).

Gadiraju, U., Kawase, R. et Dietze, S. (2014), « A taxonomy of microtasks on the web », in*Proceedings of the 25th ACM conference on Hypertext and social media*, ACM, p. 218-223 (cf. p. 24, 27).

Gadiraju, U., Kawase, R., Dietze, S. et Demartini, G. (2015), « Understanding malicious behavior in crowdsourcing platforms : The case of online surveys », in*Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, p. 1631-1640 (cf. p. 20, 94).

Goldberg, L. R. (1990), « An alternative" description of personality" : the big-five factor structure. », in*Journal of personality and social psychology* 59.*6*, p. 1216 (cf. p. 20).

— (1993), « The structure of phenotypic personality traits. », in*American psychologist* 48.*1*, p. 26 (cf. p. 20).

Guittard, C. et Schenk, E. (2010), « Le Crowdsourcing : Une typologie des pratiques d'externalisation vers la foule », in*XIXème conférence de l'AIMS* (cf. p. 11, 13, 15, 17).

Harris, C. (2011), « You're hired ! an examination of crowdsourcing incentive models in human resource tasks », in*Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, Hong Kong, China, p. 15-18 (cf. p. 19).

How to Build a Successful Task for the Crowd ((09/09/2021)), url : `https://appen.com/blog/recreate-how-to-build-a-successful-task-for-the-crowd/` (cf. p. 24).

Howe, J. (2006), « The Rise of Crowdsourcing », in*Wired Magazine* (cf. p. 10, 16).

Hung, N. Q. V., Tam, N. T., Tran, L. N. et Aberer, K. (2013), « An evaluation of aggregation techniques in crowdsourcing », in*International Conference on Web Information Systems Engineering*, Springer, p. 1-15 (cf. p. 32, 36).

Innocent, M., Gabriel, P. et Divard, R. (2017), « Comprendre l'expérience de participation des meilleurs contributeurs dans un contexte de crowdsourcing d'activités inventives », in*Recherche et Applications en Marketing (French Edition)* 32.*1*, p. 3-21 (cf. p. 17).

Ipeirotis, P. G., Provost, F. et Wang, J. (2010), « Quality Management on Amazon Mechanical Turk », in*KDD-HCOMP'10* (cf. p. 35).

Jousselme, A.-L., Grenier, D. et Bossé, É. (2001), « A new distance between two bodies of evidence », in*Information fusion* 2.*2*, p. 91-101 (cf. p. 45, 54, 95, 96).

Jraidi, I. et Elouedi, Z. (2007), « Belief classification approach based on generalized credal EM », in*European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty*, Springer, p. 524-535 (cf. p. 52).

Jurgens, D. (2013), « Embracing ambiguity : A comparison of annotation methodologies for crowdsourcing word sense labels », in*Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 556-562 (cf. p. 25).

Kamar, E., Horvitz, E., Bowyer, A. et Miller, G. (2016), « Intervention strategies for increasing engagement in crowdsourcing : Platform, predictions, and experiments », in(cf. p. 27).

Kaufmann, N., Schulze, T. et Veit, D. (2011), « More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk. », in*AMCIS*, t. 11, p. 1-11 (cf. p. 19).

Kazai, G., Kamps, J. et Milic-Frayling, N. (2011), « Worker Types and Personality Traits in Crowdsourcing Relevance Labels », in*20th ACM Conference on Information and Knowledge Management, CIKM* (cf. p. 20).

— (2012), « The Face of Quality in Crowdsourcing Relevance Labels : Demographics, Personality and Labeling Accuracy », in*Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, Maui, Hawaii, USA : ACM, p. 2583-2586 (cf. p. 20, 59, 90, 94).

— (2013), « An analysis of human factors and label accuracy in crowdsourcing relevance judgments », in*Information retrieval* 16.*2*, p. 138-178 (cf. p. 19, 58).

Khan, K., Davies, D. et Gupta, J. (2001), « Formative self-assessment using multiple true-false questions on the Internet : feedback according to confidence about correct knowledge », in*Med Teach.* 23.*2*, p. 158-163 (cf. p. 58).

Khattak, F. K. et Salleb-Aouissi, A. (2011), « Quality control of crowd labeling through expert evaluation », in*Proceedings of the NIPS 2nd Workshop on Computational Social Science and the Wisdom of Crowds*, t. 2, p. 5 (cf. p. 32, 36).

Kittur, A., Nickerson, J. V., Bernstein, M. S., Gerber, E. M., Shaw, A., Zimmerman, J., Lease, M. et Horton, J. J. (2013), « The Future of Crowd Work », in*16th ACM Conference on Computer Supported Coooperative Work (CSCW 2013), Forthcoming* (cf. p. 21).

Komarov, S., Reinecke, K. et Gajos, K. Z. (2013), « Crowdsourcing performance evaluations of user interfaces », in*Proceedings of the SIGCHI conference on human factors in computing systems*, ACM, p. 207-216 (cf. p. 111, 113, 114, 141).

Koulougli, D., Hadjali, A. et Rassoul, I. (2016), « Handling query answering in crowdsourcing systems : A belief function-based approach », in*Fuzzy Information Processing Society (NAFIPS), 2016 Annual Conference of the North American*, IEEE, p. 1-6 (cf. p. 53, 56, 122, 133).

Laney, D. (2001), « 3D Data Management : Controlling Data Volume, Velocity and Variety », in(cf. p. 2).

Le, J., Edmonds, A., Hester, V. et Biewald, L. (2010), « Ensuring quality in crowdsourced search relevance evaluation : The effects of training question distribution », in*CSE 2010* (cf. p. 28).

Lease, M. (2011), « On quality control and machine learning in crowdsourcing. », in*Human Computation* 11.*11* (cf. p. 21).

Lee, K., Caverlee, J. et Webb, S. (2010), « The social honeypot project : protecting online communities from spammers », in*Proceedings of the 19th international conference on World wide web*, ACM, p. 1139-1140 (cf. p. 32).

Lefèvre, E. et Elouedi, Z. (2013), « How to preserve the conflict as an alarm in the combination of belief functions ? », in*Decision Support Systems* 56, p. 326-333 (cf. p. 49).

Liang-zhou, C., Wen-kang, S., Yong, D. et Zhen-fu, Z. (2005), « A new fusion approach based on distance of evidences », in*Journal of Zhejiang University-Science A* 6.*5*, p. 476-482 (cf. p. 49).

Maalel, W., Zhou, K., Martin, A. et Elouedi, Z. (2014), « Belief hierarchical clustering », in*International Conference on Belief Functions*, Springer, p. 68-76 (cf. p. 52).

Marcus, A. et Parameswaran, A. (2015), « Crowdsourced Data Management : Industry and Academic Perspectives », in*Foundations and Trends® in Databases* 6.*1-2*, p. 1-161 (cf. p. 23, 24, 31).

Martin, A. (2019), *Conflict Management in Information Fusion with Belief Functions.* (Cf. p. 46).

Martin, A., Jousselme, A.-L. et Osswald, C. (2008), « Conflict measure for the discounting operation on belief functions », in*Information Fusion, 2008 11th International Conference on*, IEEE, p. 1-8 (cf. p. 54).

Martin, A. et Osswald, C. (2007), « Toward a combination rule to deal with partial conflict and specificity in belief functions theory », in*Information Fusion, 2007 10th International Conference on Information Fusion*, IEEE, p. 1-8 (cf. p. 49).

Mason, W. et Watts, D. J. (2009), « Financial incentives and the performance of crowds », in*Proceedings of the ACM SIGKDD workshop on human computation*, ACM, p. 77-85 (cf. p. 27).

Mavridis, P., Gross-Amblard, D. et Miklós, Z. (2016), « Using hierarchical skills for optimized task assignment in knowledge-intensive crowdsourcing », in*Proceedings of the 25th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, p. 843-853 (cf. p. 22, 30).

Mehmood, M. S., Mehmood, A. et Siddique, M. (2016), « Personality Traits Nexus Employee's Performance : An Application of Big Five Personality Dimensions Model », in*Abasyn Journal of Social Sciences–Special Issue : AIC*, p. 101-119 (cf. p. 90, 94).

Nguyen, A. T. (2015), « Combining Crowd and Expert Labels using Decision Theoretic Active Learning », in*Association for the Advancement of Artificial Intelligence* (cf. p. 31).

Osswald, C. et Martin, A. (2006), « Understanding the large family of Dempster-Shafer theory's fusion operators-a decision-based measure », in*2006 9th International Conference on Information Fusion*, IEEE, p. 1-7 (cf. p. 49).

Ouni, H., Martin, A., Gros, L., Kharoune, M. et Miklos, Z. (2017), « Une mesure d'expertise pour le crowdsourcing », in*Extraction et Gestion des connaissances (EGC)* (cf. p. 53, 56, 90).

Penna, N. D. et Reid, M. D. (2012), « Crowd & Prejudice : An Impossibility Theorem for Crowd Labelling without a Gold Standard », in*Collective Intelligence* (cf. p. 21, 33).

Rahmanian, B. et Davis, J. G. (2014), « User interface design for crowdsourcing systems », in*Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, ACM, p. 405-408 (cf. p. 21, 26).

Raykar, V. C. et Yu, S. (2012), « Annotation models for crowdsourced ordinal data », in*Journal of Machine Learning Research* (cf. p. 35, 36).

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L. et Moy, L. (2010), « Learning From Crowds », in*Journal of Machine Learning Research* (cf. p. 36).

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J. et Vukovic, M. (2011), « An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets », in*Fifth International AAAI Conference on Weblogs and Social Media* (cf. p. 27).

Roitero, K., Demartini, G., Mizzaro, S. et Spina, D. (2018), « How Many Truth Levels ? Six ? One Hundred ? Even More ? Validating Truthfulness of Statements via Crowdsourcing », in*Proceedings of the 2nd International Workshop on Rumours and Deception in Social Media* (cf. p. 25).

Ross, J., Zaldivar, A., Irani, L. et Tomlinson, B. (2009), « Who are the turkers ? worker demographics in amazon mechanical turk », in*Department of Informatics, University of California, Irvine, USA, Tech. Rep* (cf. p. 17, 20).

Roy, S. B., Lykourentzou, I., Thirumuruganathan, S., Amer-Yahia, S. et Das, G. (2013), « Crowds, not drones : modeling human factors in interactive crowdsour-

cing », in*DBCrowd 2013-VLDB Workshop on Databases and Crowdsourcing*, CEUR-WS, p. 39-42 (cf. p. 30).

Schenk, E. et Guittard, C. (2012), « Une typologie des pratiques de Crowdsourcing : l'externalisation vers la foule, au-delà du processus d'innovation1 », in*Management international/International Management/Gestiòn Internacional* 16, p. 89-100 (cf. p. 11-13, 15).

Shafer, G. (1976), *A mathematical theory of evidence*, t. 42, Princeton university press (cf. p. 40).

Silberman, M., Irani, L. et Ross, J. (2010), « Ethics and tactics of professional crowdwork », in*XRDS : Crossroads, The ACM Magazine for Students* 17.*2*, p. 39-43 (cf. p. 18, 19).

Smarandache, F. et Dezert, J. (2005), « Information fusion based on new proportional conflict redistribution rules », in*2005 7th International Conference on Information Fusion*, t. 2, IEEE, 8-pp (cf. p. 47).

Smets, P. (1990), « Constructing the pignistic probability function in a context of uncertainty », in*Uncertainty in Artificial Intelligence* (cf. p. 51).

— (1993), « Belief functions : the disjunctive rule of combination and the generalized Bayesian theorem », in*International Journal of approximate reasoning* 9.*1*, p. 1-35 (cf. p. 52).

— (1997), « Imperfect Information : Imprecision and Uncertainty », in*Uncertainty Management in Information Systems : From Needs to Solutions*, sous la dir. de Motro, A. et Smets, P., Boston, MA : Springer US, p. 225-254 (cf. p. 6, 52, 58, 72, 84, 145).

Snow, R., O'Connor, B., Jurafsky, D. et Ng, A. Y. (2008), « Cheap and fast—but is it good ? : evaluating non-expert annotations for natural language tasks », in*Proceedings of the conference on empirical methods in natural language processing*, Association for Computational Linguistics, p. 254-263 (cf. p. 16).

Tong, Y., Cao, C. C., Zhang, C. J., Li, Y. et Lei, C. (2014), « CrowdCleaner : Data cleaning for multi-version data on the web via crowdsourcing », in*Data Engineering (ICDE), 2014 IEEE 30th International Conference* (cf. p. 30).

Vannoorenberghe, P. (2007), « Estimation de modèles de mélanges finis par un algorithme EM crédibiliste », in*Traitement du signal* 24 (cf. p. 52).

Wagner, C. et Anderson, D. T. (2012), « Extracting meta-measures from data for fuzzy aggregation of crowd sourced information », in*2012 IEEE International Conference on Fuzzy Systems*, IEEE, p. 1-8 (cf. p. 33).

Wang, J., Ipeirotis, P. G. et Provost, F. (2011), « Managing crowdsourcing workers », in*The 2011 nter conference on business intelligence*, p. 10-12 (cf. p. 35).

Welinder, P., Branson, S., Perona, P. et Belongie, S. J. (2010), « The multidimensional wisdom of crowds », in*Advances in neural information processing systems*, p. 2424-2432 (cf. p. 35).

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J. R. et Ruvolo, P. L. (2009), « Whose vote should count more : Optimal integration of labels from labelers of unknown expertise », in*Advances in neural information processing systems*, p. 2035-2043 (cf. p. 32, 35, 36).

Yager, R. R. (1987), « On the Dempster-Shafer framework and new combination rules », in*Information sciences* 41.*2*, p. 93-137 (cf. p. 47).

Yan, Y., Rosales, R., Fung, G., Schmidt, M., Hermosillo, G., Bogoni, L., Moy, L. et Dy, J. (2010), « Modeling annotator expertise : Learning when everybody knows a bit of something », in*Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, p. 932-939 (cf. p. 35).

Yang, J., Redi, J., Demartini, G. et Bozzon, A. (2016), « Modeling task complexity in crowdsourcing », in*Fourth AAAI Conference on Human Computation and Crowdsourcing* (cf. p. 25).

Zhang, H., Law, E., Miller, R., Gajos, K., Parkes, D. et Horvitz, E. (2012), « Human computation tasks with global constraints », in*Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, p. 217-226 (cf. p. 21).

Zhang, Y. et Van Der Schaa, M. (2012), « Reputation-based incentive protocols in crowdsourcing applications », in*INFOCOM, 2012 Proceedings IEEE*, IEEE, p. 2140-2148 (cf. p. 29).

Annexe 1

Algorithm 1 Function EM(table : data, table : users, real : I, J, K)

1: Estimation des valeurs initiales des T_{ij} par vote majoritaire
2: Estimation des valeurs $\pi_{jl}^{(k)}$ et p_j
3: Estimation des valeurs T_{ij}
4: **while** Not Convegence **do**
5: Estimation des valeurs $\pi_{jl}^{(k)}$ et p_j
6: Estimation des valeurs T_{ij}
7: **end while**
8: **return** T_{ij}

Algorithm 2 Function TauxValidite(table : time, real : Q)

```
1: degreValidite <- 0
2: for q in c(1 :Q) do
3:    if  (time[q] < Q1 - 3*ICQ) || (time[q] > Q3 + 3*ICQ)  then
4:       degreValidite <- degreValidite + 1
5:    end if
6: end for
7: return  degreValidite/Q
```

Annexe 2

Questions relatives à l'expérience 3 avec imprécision et incertitude

Pour vous, sélectionner le(s) segment(s) le(s) plus long(s) était ?

— Très facile

— Facile

— Ni facile ni difficile

— Difficile

— Très difficile

Pour vous, sélectionner plusieurs segments pour donner la bonne réponse était ?

— Très pertinent

— Pertinent

— Moyennement pertinent

— Peu pertinent

— Pas du tout pertinent

Pour vous, exprimer votre certitude sur la réponse était ?

— Très facile

— Facile

— Ni facile ni difficile

— Difficile

— Très difficile

En sélectionnant plusieurs segments, vous étiez plus certain de votre réponse :

— Je suis tout à fait d'accord

— Je suis d'accord

— Je suis moyennement d'accord

— Je ne suis pas vraiment d'accord

— Je ne suis absolument pas d'accord

Comment évalueriez-vous votre capacité à répondre ?

— J'ai pratiquement toujours hésité avant de répondre

— J'ai souvent hésité avant de répondre

— J'ai occasionnellement hésité avant de répondre

— J'ai rarement hésité avant de répondre

— Je n'ai jamais hésité avant de répondre

Si vous avez hésité avant de répondre, c'est parce que :

— La tâche est devenue fastidieuse

— Vous hésitiez entre plusieurs segments

— Autre raison (à spécifier) :

Combien de campagnes de crowdsourcing avez-vous réalisées avant celle-ci ?

Avez-vous d'autres commentaires sur cette expérience ?

— Non

— Oui :

Vous êtes ?

— Une femme

— Une homme

— Autre

Quel âge avez-vous ?

En général, vous aimez résoudre des problèmes :

— Je suis tout à fait d'accord
— Je suis d'accord
— Je suis moyennement d'accord
— Je ne suis pas vraiment d'accord
— Je ne suis absolument pas d'accord

En général, vous êtes attentif aux détails :

— Je suis tout à fait d'accord
— Je suis d'accord
— Je suis moyennement d'accord
— Je ne suis pas vraiment d'accord
— Je ne suis absolument pas d'accord

En général, vous êtes méticuleux :

— Je suis tout à fait d'accord
— Je suis d'accord
— Je suis moyennement d'accord
— Je ne suis pas vraiment d'accord
— Je ne suis absolument pas d'accord

ANNEXE 3

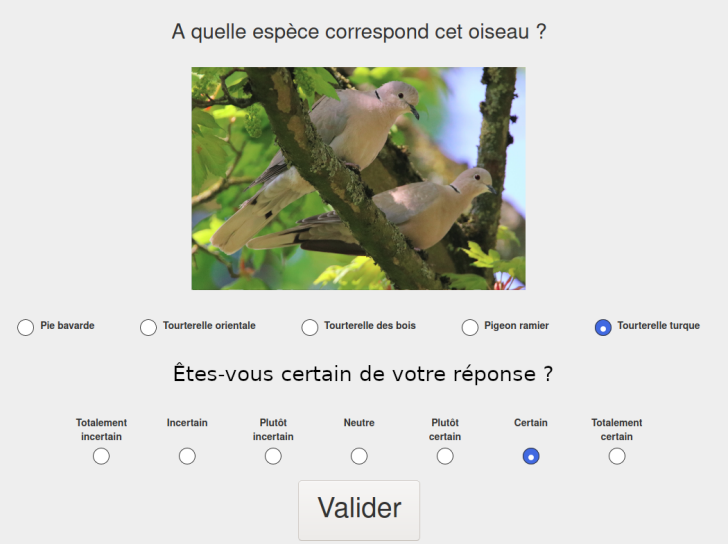


FIGURE 7.1 – Interface utilisée pour la campagne de *crowdsourcing* oiseaux_xp1.

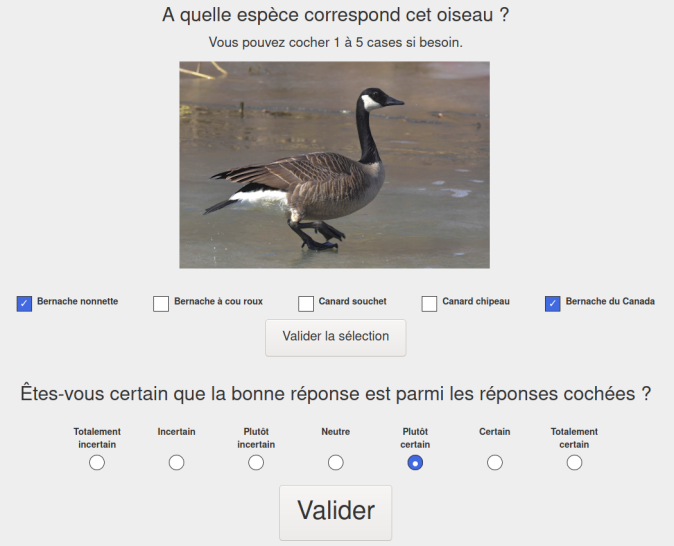


FIGURE 7.2 – Interface utilisée pour la campagne de *crowdsourcing* oiseaux_xp2.

ANNEXE 4

Démonstration.

$$0 \leq m_{cq}^{\Omega_q}(X) \leq 1$$

$$0 \leq m_{cq}^{\Omega_q}(X)\left(1 - \frac{log_2|X|}{log_2|\Omega_q|}\right) \leq \left(1 - \frac{log_2|X|}{log_2|\Omega_q|}\right)$$

$$0 \leq DP_c \leq m_c^{\Omega_I}(P)$$

□

Démonstration.

$$1 - \frac{log_2|X|}{log_2(imp_{MAX})} \leq m_{cq}^{\Omega_q}(X)\left(1 - \frac{log2|X|}{log_2|\Omega_q|}\right)$$

$$\frac{1}{log_2(imp_{MAX})}(log_2(imp_{MAX}) - log_2|X|) \leq \frac{m_{cq}^{\Omega_q}(X)}{log_2|\Omega_q|}(log_2|\Omega_q| - log_2|X|)$$

$$\frac{log_2|\Omega_q|}{log_2(imp_{MAX})}\frac{(log_2(imp_{MAX}) - log_2|X|)}{(log_2|\Omega_q| - log_2|X|)} \leq m_{cq}^{\Omega_q}(x)$$

$$\gamma(|X|) \leq m(X)$$

□

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

Titre : Évaluation de la qualité des contributions et des contributeurs sur plateformes de *crowdsourcing*

Mot clés : Imprécision, incertitude, fonction de croyance, *crowdsourcing*

Résumé : Le *crowdsourcing* est l'externalisation de tâches à une foule de contributeurs sur des plateformes dédiées. Les tâches sont simples et accessibles à tous, c'est pourquoi la foule est constituée de profils très diversifiés, ce qui induit des contributions de qualité inégales. La méthode d'agrégation la plus employée dans les plateformes ne prend pas en considération les imperfections des données relatives aux contributions humaines ce qui impacte les résultats obtenus. L'ensemble des travaux de cette thèse tend à solutionner la problématique de la qualité des données de *crowdsourcing*. Nous proposons ainsi une nouvelle interface pour le *crowdsourcing* offrant davantage de capacité d'expression au contributeur. Les expériences menées nous ont permis de mettre en évidence une corrélation entre la difficulté de la tâche, la certitude du contributeur et l'imprécision de sa réponse. Nous avons également validé l'hypothèse de Ph. Smets d'après laquelle plus une personne est imprécise plus elle est certaine et réciproquement plus elle est précise moins elle est certaine. Une fois cette hypothèse validée, nous avons élaboré le modèle MONITOR pour l'estimation du profil du contributeur et l'agrégation des réponses grâce à la théorie des fonctions de croyance qui permet de modéliser les imperfections. L'intégralité de nos expérimentations est réalisée sur des données réelles provenant de campagnes de *crowdsourcing*.

Title: Evaluation of the quality of contributions and contributors on crowdsourcing platforms

Keywords: Imprecision, uncertainty, belief function, crowdsourcing

Abstract: Crowdsourcing is the outsourcing of tasks to a crowd of contributors on dedicated platforms. The tasks are simple and accessible to all, that's why the crowd is made of very diverse profiles, but this induces contributions of unequal quality. The aggregation method most used in platforms does not take into account the imperfections of the data related to human contributions, which impacts the results obtained. The work of this thesis aims at solving the problem of data quality in crowdsourcing platforms. Thus, we propose a new interface for crowdsourcing offering more expression capacity to the contributor. The experiments carried out allowed us to highlight a correlation between the difficulty of the task, the certainty of the contributor and the imprecision of his answer. We also validated the hypothesis of Ph. Smets according to which the more imprecise a person is, the more certain he is, and conversely the more precise he is, the less certain he is. Based on this hypothesis, we develop the MONITOR model for the estimation of the contributor's profile and the aggregation of the answers thanks to the theory of belief functions which allows to model imperfections. All our experiments are performed on real data coming from crowdsourcing campaigns.