



HAL
open science

Interprétation floue des expressions numériques approximatives

Sébastien Lefort, Marie-Jeanne Lesot, Elisabetta Zibetti, Charles Tijus,
Marcin Detyniecki

► **To cite this version:**

Sébastien Lefort, Marie-Jeanne Lesot, Elisabetta Zibetti, Charles Tijus, Marcin Detyniecki. Interprétation floue des expressions numériques approximatives. 25èmes Rencontres Francophones sur la Logique Floue et ses Applications, Nov 2016, La Rochelle, France. hal-01368177

HAL Id: hal-01368177

<https://hal.sorbonne-universite.fr/hal-01368177>

Submitted on 19 Sep 2016

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

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

Interprétation floue des expressions numériques approximatives

Approximate numerical expressions : a fuzzy computational model

S. Lefort¹

M.-J. Lesot¹

E. Zibetti²

C. Tijus²

M. Detyniecki^{1,3}

¹ Sorbonne Universités, UPMC Univ Paris 06, CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris

² Laboratoire CHArt-LUTIN, EA 4004, Université Paris 8, 2 rue de la liberté, 93526, Saint-Denis - Cedex 02, France

³ Polish Academy of Sciences, IBS PAN, Warsaw, Poland

{sebastien.lefort,marie-jeanne.lesot,marcin.detyniecki}@lip6.fr

{ezibetti,tijus}@univ-paris8.fr

Résumé :

Les Expressions Numériques Approximatives, ou ENA, sont des expressions linguistiques de la forme “*environ x*”, où x est un nombre. Cet article propose un modèle d’interprétation des ENA basé sur un compromis entre saillance cognitive des nombres et plage de valeurs dénotées. Contrairement aux modèles de la littérature qui représentent les ENA par des intervalles, nous proposons de les représenter par des nombres flous et d’en caractériser le support, le noyau et la 0,5-coupe. Une étude expérimentale, basée sur des données réelles collectées met en évidence les performances du modèle.

Mots-clés :

Expressions Numériques Approximatives, nombres flous, fronts de Pareto, saillance cognitive.

Abstract:

Approximate Numerical Expressions ANEs are linguistic expressions of the form “*about x*”, where x is a number. This paper proposes a model based on a compromise between cognitive salience of numbers and range of denoted values. Unlike the models from the literature, which represent ANEs by intervals, we propose to represent them by fuzzy numbers and to characterise their support, kernel and 0.5-cut. An experimental study, based on collected real data highlights the performances of the model.

Keywords:

Approximate Numerical Expression, Fuzzy numbers, Pareto frontiers, Cognitive salience

1 Introduction

Les Expressions Numériques Approximatives (ENA) sont des expressions linguistiques vagues dont la forme générale est “*environ x*”, où x est un nombre. Elles sont utilisées quotidiennement pour dénoter des plages imprécises de valeurs, par ex., “Berlin est situé à *environ 900km* de Paris”; “Le patient a de la fièvre depuis *environ une semaine*”. Dans le cadre des Interfaces Homme-Machine, l’inter-

prétation des ENA pose un double problème : d’un côté l’estimation de la plage de valeurs qu’elles dénotent, de l’autre leur représentation dans les systèmes d’information, comme intervalles de valeurs ou comme sous-ensembles flous.

D’un point de vue linguistique, Lasersohn [8] propose de formaliser le vague en général, au delà du cas des expressions numériques, par des halos pragmatiques. Ceux-ci sont définis comme l’union de l’entité à laquelle se réfère explicitement l’expression vague et d’autres entités, de même type sémantique, implicitement dénotées. Par exemple, dans la proposition “*il y a environ 100 participants à la conférence*”, le halo pragmatique de l’expression “*environ 100*” correspond à 100 exactement et à une plage de valeurs autour de 100 (par ex., [90; 110]). Interpréter une ENA revient donc à estimer la plage de valeurs incluses dans son halo pragmatique, formalisée comme un intervalle.

Une autre approche naturelle consiste à représenter les ENA par des nombres flous [5]. A notre connaissance, toutefois, aucune étude n’a été réalisée pour caractériser empiriquement les fonctions d’appartenance correspondant à des ENA dans le cadre de la langue naturelle.

Cet article propose un modèle caractérisant le support, le noyau et la 0,5-coupe de nombres flous correspondant à des ENA de la forme “*environ x*”, pour $x \in \mathbb{N}^*$. Le modèle est basé sur des points critiques issus de fronts de Pareto, représentant un compromis entre la saillance cog-

nitive des nombres et leur distance à la valeur de référence x . Une étude empirique est conduite pour collecter des données réelles et réaliser une validation expérimentale mettant en évidence les performances du modèle.

La section 2 décrit les travaux précédents et les modèles existants. Le modèle proposé est présenté dans la section 3. La procédure de collecte des données est décrite dans la section 4. La section 5 est consacrée à l'étude expérimentale et ses résultats. Enfin, les conclusions et travaux futurs sont discutés dans la section 6.

2 Etat de l'art

Cette section introduit d'abord les notations et définitions des dimensions et propriétés des ENA. Deux modèles, issus de la littérature, estimant l'intervalle de valeurs dénotées sont ensuite présentés.

2.1 Définitions et notations

Les ENA considérées sont de la forme “*environ x*”, pour $x \in \mathbb{N}^*$. Dans le système décimal, x s'écrit $x = \sum_{i=0}^q a_i \cdot 10^i$, avec $a_i \in \llbracket 0, 9 \rrbracket$. Deux types de dimensions de x , détaillées dans le tableau 1, peuvent être distingués : arithmétiques et cognitive.

Dimensions arithmétiques. La granularité $Gran(x)$ est la puissance de 10 à laquelle x appartient, la magnitude relative $R_m(x)$ est la valeur de son dernier chiffre significatif et la précision $Prec(x)$ est le produit de la granularité par la magnitude relative. Nous nous attendons à ce que ces dimensions influencent l'interprétation des ENA. Par exemple, la précision doit refléter le fait que la taille de l'intervalle de “*environ 30050*” soit comparable à celle de “*environ 150*”, 50 étant leur partie commune.

Deux types d'entiers peuvent être distingués à partir de ces dimensions. Les nombres *ronds* sont classiquement définis comme des multiples de 10 avec un seul chiffre significatif (par ex. 50 ou 8000). Nous proposons de définir les nombres *pseudo-ronds* comme des multiples de 10 avec au moins deux chiffres significatifs (par ex. 320 ou 8150).

Dimension cognitive. Nous proposons de considérer une autre caractéristique tenant compte de l'aspect cognitif des nombres. En effet, il a été observé que certains nombres apparaissent plus fréquemment que d'autres dans des corpus [6, 2]. La complexité $Cpx(x)$ a pour but de capturer cette saillance. En premier lieu, il apparaît que plus un nombre possède de chiffres significatifs, plus sa fréquence est faible [6]. En second lieu, les nombres dont le dernier chiffre significatif est 5 et, dans une moindre mesure, 2, apparaissent plus fréquemment. Pour des raisons de symétrie autour des multiples de 10, nous proposons de traiter les nombres avec $R_m(x) = 8$ (par ex. $18 = 20 - 2$) comme ceux avec $R_m(x) = 2$ (par ex. $22 = 20 + 2$). Nous proposons de formaliser Cpx comme le nombre de chiffres significatifs auquel est soustrait un bonus pour capturer ces cas particuliers, si le nombre de chiffres significatifs est supérieur à 1.

La fonction de bonus distingue alors trois catégories, selon le dernier chiffre significatif $R_m(x)$ et respectant l'ordre de leur fréquence d'apparition : $B(x_1) > B(x_2) > B(x_3)$, pour $x_1, x_2, x_3 \in \mathbb{N}$ tels que $R_m(x_1) = 5$, $R_m(x_2) \in \{2, 8\}$ et $R_m(x_3) \notin \{2, 5, 8\}$. Nous proposons de fixer arbitrairement ces valeurs à 0,5, 0,25 et 0. La figure 1 illustre la complexité $Cpx(x)$ pour tous les entiers x compris entre 400 et 500.

2.2 Modèle des échelles (SBM)

La première approche d'interprétation des ENA a été proposée dans une perspective linguistique. Le modèle des échelles (SBM) [7, 9, 10] est basé sur des systèmes d'échelles $S = \{s_1, \dots, s_n\}$, où les s_i sont des niveaux de granularité tels que $s_i < s_{i+1}$. On peut mentionner comme exemples le système temporel, $S = \{1 \text{ min}, 5 \text{ min}, 15 \text{ min}, \dots\}$, ou le système décimal, $S = \{1, 10, 100, \dots\}$.

L'interprétation d'une expression numérique peut être réalisée à tout niveau de granularité. Par exemple, dans le système décimal, “100” peut être interprétée au niveau 1, 10 ou 100. Plus la granularité est fine, plus l'intervalle est réduit. Le locuteur peut exprimer le ni-

Tableau 1 – Dimensions d’un entier naturel $x = \sum_{i=0}^q a_i \cdot 10^i$, illustrées pour $x = 4750$ dans la dernière colonne. $B(x)$, utilisée dans la définition de la complexité, est définie dans la section 2.1.

Dimension	Définition formelle	Exemple $x = 4750$
Granularité	$Gran(x) = 10^{i^*}$ où $i^* = \min\{i a_i \neq 0\}$	10 1
Magnitude relative	$R_m(x) = a_{i^*}$	5
Précision	$Prec(x) = a_{i^*} \cdot 10^{i^*}$	50
Nombre de chiffres significatifs	$NSD(x) = q - i^* + 1$	3
Complexité	$Cpx(x) = NSD(x) - B(x)$	2.5

veau voulu en utilisant des approximateurs [9] : “exactement” fait référence au niveau de granularité le plus fin, “environ” au niveau le plus grossier (par ex., les milliers pour “environ 1000”), défini comme $Gran_C(x) = \sup(\{s_i \in S | x \bmod s_i = 0\})$. Si le système est décimal, $Gran_C(x) = Gran(x)$ (voir tableau 1).

SBM considère que les valeurs dénotées par une ENA x sont celles qui sont plus proches de x que de tout autre nombre au niveau $Gran_C(x)$. L’intervalle est formellement défini comme :

$$I_{SBM}(x) = \left[x - \frac{Gran_C(x)}{2}; x + \frac{Gran_C(x)}{2} \right]$$

Par exemple, $I_{SBM}(300) = [250; 350]$; $I_{SBM}(8150) = [8145; 8155]$.

Cette approche présente l’avantage de tenir compte de la granularité des ENA ; elle ne considère toutefois pas la magnitude relative : les intervalles de toutes les ENA au même niveau de granularité ont la même taille, alors que l’on pourrait s’attendre à ce que, par exemple, l’intervalle de “environ 100” soit plus réduit que celui de “environ 800”.

2.3 Modèle régressif (REGM)

Ferson et al. [4] proposent une approche empirique basée sur des données réelles pour tester la pertinence de prédicteurs de la taille des intervalles. Des ENA sémantiquement contextualisées (par ex., “A peu près 25% des Canadiens sont protestants.”) sont présentées aux participants, qui doivent estimer les bornes des intervalles correspondants. L’intervalle prédit par le

modèle s’écrit :

$$I_{REGM}(x) = \left[x - \frac{10^{L(x)}}{2}; x + \frac{10^{L(x)}}{2} \right]$$

avec $L(x) = A + B \cdot O_m(x) + C \cdot R(x) + D \cdot f(x) + E \cdot O_m(x) \cdot R(x) + F \cdot O_m(x) \cdot f(x) + G \cdot R(x) \cdot f(x) + H \cdot O_m(x) \cdot R(x) \cdot f(x)$ où A à H sont des paramètres fixés empiriquement en réalisant une régression linéaire sur les données. $O_m(x)$ est l’ordre de magnitude de l’ENA ($O_m(x) = \log_{10}(x)$), $R(x)$ sa “rondeur” ($R(x) = i^* + 1$), et $f(x)$ sa “fiveness”, définie comme $f(x) = 1$ si $a_{i^*} = 5$, $f(x) = 0$ sinon. $O_m(x)$, $R(x)$, $f(x)$ et leurs combinaisons ont été empiriquement sélectionnées comme prédicteurs de la taille des intervalles.

Ce modèle a l’avantage de permettre son adaptation à différents contextes, en apprenant les paramètres à partir de données. On peut cependant noter que le contexte sémantique n’est pas contrôlé alors que mélanger plusieurs contextes peut conduire à des interactions entre ce facteur et ceux liés à la valeur de référence de l’ENA.

3 Modèle proposé : PFM

Cette section décrit le modèle flou, basé sur un compromis entre la taille de la plage de valeurs dénotées et la saillance cognitive de ses bornes, que nous proposons pour estimer le support, le noyau et la 0,5-coupe des nombres flous correspondant aux ENA.

3.1 Principes

Nous proposons une approche floue dans laquelle interpréter une ENA consiste à estimer sa fonction d’appartenance, $f_{\tilde{x}}(y)$, où y sont les

valeurs pouvant être dénotées par “*environ x*”. Parmi les méthodes d’élicitation de fonctions d’appartenance (cf par ex. [1]), la perspective *random set* interprète le degré d’appartenance d’un candidat (par ex. 95 pour “*environ 100*”) comme la fréquence relative cumulée de participants qui indiquent qu’il appartient à l’intervalle dénoté. Par conséquent, si la moitié de la population indique que 95 est inclus dans “*environ 100*”, son degré d’appartenance est 0,5. La médiane de la distribution est donc critique pour l’estimation d’une fonction d’appartenance.

3.2 Le modèle des fronts de Pareto (PFM)

Le modèle que nous proposons est basé sur l’hypothèse que, lors de l’interprétation d’une ENA x , un compromis est réalisé entre le coût cognitif de la valeur des bornes, mesuré par la complexité Cpx , d’une part, et la plage de valeurs dénotées, mesurée par la distance entre la borne candidate et x , d’autre part. Cette hypothèse implique que, pour une plage de valeurs donnée, le coût cognitif est minimisé ; réciproquement, pour un coût cognitif donné, la plage de valeurs dénotées est minimisée. Par exemple, pour l’ENA “*environ 500*”, les participants de l’étude (cf section 4) ont tendance à donner des réponses telles que $[499; 501]$, $[490; 510]$ ou $[450; 550]$. Ces bornes sont les plus proches de l’ENA lorsque $Cpx(x) = 3, 2$ et $1,5$.

Par conséquent, le modèle que nous proposons détermine en premier lieu les bons candidats, en générant des fronts de Pareto [3] : toutes les valeurs candidates v dans $[1; x[$ pour la borne inférieure et dans $]x; +\infty[$ pour la borne supérieure de l’ENA x sont comparées sur deux critères : (i) leur distance à x : $d_x(v) = |v - x|$; (ii) leur complexité $Cpx(v)$. Les valeurs sélectionnées, constituant les fronts de Pareto, sont celles qui ne sont pas dominées par d’autres. Pour une ENA donnée, deux fronts de Pareto, ordonnés par distance à x croissante, sont considérés : $P^-(x) = [y_1^-, \dots, y_n^-]$ pour la borne inférieure, et $P^+(x) = [y_1^+, \dots, y_n^+]$ pour la supérieure. La figure 1 les illustre pour “*environ 440*” : $P^-(440) = [439, 438, 435, 430, 420, 400]$ et

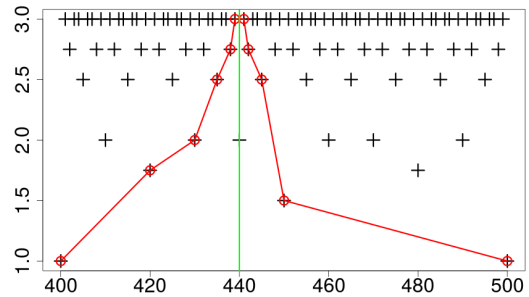


Figure 1 – Fronts de Pareto (en rouge) pour la borne inférieure (à gauche de la ligne verte) et supérieure (à droite de la ligne verte) de “*environ 440*”. Les signes + noirs représentent la complexité $Cpx(v)$ de chaque entier.

$P^+(440) = [441, 442, 445, 450, 500]$. On peut remarquer que le modèle capture naturellement l’asymétrie observée dans les données (cf section 4), due aux nombres saillants (par ex. 420, 450) présents autour de la valeur de référence.

3.3 Support, noyau et 0,5-coupe

La seconde étape du modèle proposé consiste à utiliser les valeurs des fronts de Pareto comme candidates pour être les bornes du support, du noyau et de la 0,5-coupe des nombres flous correspondant aux ENA.

Toute valeur en dehors du support, $I_S(x)$, est considérée comme non dénotée par l’ENA. Nous proposons donc de définir les valeurs des fronts de Pareto les plus éloignées de x comme bornes du support : $I_S(x) = [y_{n^-}^-; y_{n^+}^+]$.

Toute valeur à l’intérieur du noyau, $I_K(x)$, est considérée comme pleinement dénotée par l’ENA. Nous proposons donc d’utiliser les valeurs des fronts de Pareto les plus proches de x comme bornes du noyau : $I_K(x) = [y_1^-; y_1^+]$.

Les bornes de la 0,5-coupe, $I_M(x)$, sont également sélectionnées selon leur rang dans $P^-(x)$ et $P^+(x)$. Nous proposons de faire dépendre ce rang de l’ENA considérée de manière à rendre le modèle flexible. Plus précisément, nous proposons que ce rang dépende du nombre de chiffres significatifs $NSD(x)$ et de la précision $Prec(x)$ de l’ENA : une analyse exhaustive (omise dans cet article pour des raisons de place) des données collectées nous a per-

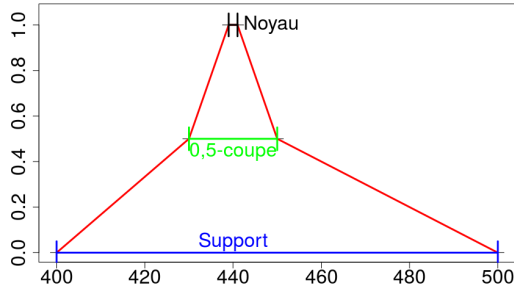


Figure 2 – Support, noyau et 0,5-coupe de la fonction d’appartenance de l’ENA *environ 440*, interpolés linéairement, à partir des points issus des fronts de Pareto.

mis de les valider comme facteurs influençant l’interprétation des ENA, à l’échelle logarithmique pour la précision et quadratique pour le nombre de chiffres significatifs. Le rang est calculé comme :

$$r_P(x) = \lfloor \log(\text{Prec}(x)) - 1 + T_{NSD(x)} \rfloor$$

La 0.5-coupe est alors $I_M(x) = [y_{r_P(x)}^-, y_{r_P(x)}^+]$. Pour $x = 440$, illustré par les figures 1 et 2, on obtient : $I_S(440) = [400; 500]$, $I_K(440) = [439; 441]$ et $I_M(440) = [430; 450]$.

4 Collecte des données

Nous avons réalisé une étude empirique pour collecter des données de manière à valider expérimentalement le modèle proposé. Cette section présente la méthodologie mise en oeuvre pour collecter et traiter ces données.

4.1 Matériel

Un questionnaire en ligne, comprenant 24 ENA décontextualisées, 15 rondes (20, 30, 40, 50, 80, 100, 200, 400, 500, 600, 800, 1000, 2000, 6000 et 8000) et 9 pseudo-rondes (110, 150, 440, 560, 1100, 1500, 4700, 4730 et 8150) a été conçu. Ces valeurs ont été sélectionnées pour couvrir différentes combinaisons de dimensions, pour éviter tout biais vers l’une d’entre elles : plusieurs magnitudes relatives à même granularité (par ex. 20/40/80), plusieurs granularités à même magnitude relative (par ex. 80/800/8000), et plusieurs nombres de chiffres significatifs à même précision (par ex. 50/150/8150). Les ENA sont présentées dans un

ordre aléatoire. La consigne donnée au participant est : “Selon vous, entre quelles valeurs (MINIMALE - MAXIMALE) se trouve “environ x ”?”. Ce questionnaire respecte les critères proposés par [1] pour éliciter des fonctions d’appartenance dans une perspective *random set*. La méthode proposée est également similaire à celle définie par [4].

146 participants ont été recrutés grâce à une annonce postée sur une liste de diffusion : 102 femmes et 44 hommes, âgés entre 20 et 70 ans ($M = 38.6$; $\sigma = 14.2$).

4.2 Traitement des données

La réponse du participant p pour l’ENA x est notée $I_p(x) = [I_p^-(x); I_p^+(x)]$. Elle est considérée comme aberrante si : (i) elle est inadéquate (par ex., $[0; \text{infini}]$), (ii) $I_p^-(x) > x$ ou $I_p^+(x) < x$, ou (iii) $I_p^-(x) < x/10$ ou $I_p^+(x) > 10x$. La moyenne et l’écart-type des bornes restantes de chaque ENA sont alors calculés. Toute borne au-delà de trois écart-types de la moyenne est considérée comme aberrante. Enfin, les participants dont plus de 70% des réponses sont manquantes ou aberrantes sont considérés comme non fiables et la totalité de leurs réponses est exclue. Les analyses sont basées sur 3177 (91%) des 3504 intervalles collectés.

4.3 Observations globales

Les données collectées montrent que les participants tendent à ne pas être d’accord sur les intervalles : on observe en moyenne 15,4, allant de 9 (“environ 20”) à 22 (“environ 8150”), réponses différentes par borne. Toutefois, 84,4% des bornes sont sur les fronts de Pareto, validant ainsi le principe sous-jacent au modèle proposé. La symétrie des intervalles dépend de l’ENA : 74,2% de l’ensemble des intervalles sont symétriques, mais les intervalles de certaines ENA, comme 440 ou 4730, sont moins souvent symétriques (63 et 50%). Cette observation valide donc la définition d’un modèle proposant des intervalles possiblement asymétriques.

5 Etude expérimentale

Cette section présente l’étude expérimentale que nous avons menée pour évaluer la qualité

des trois paramètres estimés des nombres flous correspondant aux ENA : la 0,5-coupe, le support et le noyau. Les critères de qualité définis et les résultats pour chaque paramètre sont décrits dans les sous-sections suivantes.

5.1 Evaluation de la 0,5-coupe estimée

Dans la perspective *random set* des fonctions d'appartenance [1], la 0,5-coupe correspond à la médiane des intervalles donnés par les participants. Par conséquent, nous proposons de comparer la 0,5-coupe estimée à l'intervalle médian observé.

Critères de qualité. Nous notons \mathcal{X} l'ensemble des ENA considérées et $\mathcal{P}(x)$ l'ensemble des participants dont les intervalles ne sont pas considérés comme aberrants pour $x \in \mathcal{X}$. De plus, nous notons l'estimation du modèle m $[m^-(x); m^+(x)]$, $\Delta M_m^b(x) = |m^b(x) - x|$ sa distance à x pour $b \in \{-, +\}$, et $\Delta Med^b(x)$ la médiane, sur tous les participants $p \in \mathcal{P}(x)$, des distances $\Delta P_p^b(x) = |I_p^b(x) - x|$.

Le tableau 2 donne les définitions formelles des deux critères que nous proposons.

Pour évaluer dans quelle mesure les estimations sont correctes, nous proposons, en premier lieu, d'utiliser le score de précision de la prédiction de médiane, MA , c'est-à-dire le nombre de bornes pour lesquelles la distance relative à la médiane observée est inférieure à 10%.

En second lieu, pour évaluer le degré d'erreur $MErr$, nous proposons de calculer l'équilibre entre les participants dont la réponse est au-dessus et au-dessous de la 0,5-coupe estimée, respectivement définis par : $N_+ = |\{p | \Delta P_p^b(x) > \Delta M_m^b(x)\}|$ et $N_- = |\{p | \Delta P_p^b(x) < \Delta M_m^b(x)\}|$. Le modèle m devrait aboutir à $N_+ = N_-$ pour tous les x, b .

Toutefois, les bornes données par les participants étant distribuées sur quelques valeurs, un équilibre parfait peut être impossible. Le score tient donc compte de l'équilibre de la médiane observée, c'est-à-dire, $N_+^* = |\{p | \Delta P_p^b(x) > \Delta Med^b(x)\}|$ et $N_-^* = |\{p | \Delta P_p^b(x) < \Delta Med^b(x)\}|$.

Le score du modèle dépend de la différence entre N_+ et N_+^* , et entre N_- et N_-^* , en les moyennant sur les bornes b de toutes les ENA.

Procédure expérimentale. A l'aide de ces critères de qualité, nous comparons les performances du modèle proposé PFM à celles du modèle basé sur des échelles SBM [7, 10] avec le système décimal ($S = \{1, 10, 100, \dots\}$) ainsi qu'au modèle régressif REGM [4]. En effet, ces modèles n'étant pas flous, ils ne peuvent estimer le support, le noyau et la 0,5-coupe. Nous proposons de les utiliser pour prédire la 0,5-coupe car elle constitue un indicateur central des distributions des bornes.

Une procédure de validation croisée est réalisée sur deux benchmarks pour REGM : (i) Participant (PB), où l'apprentissage est réalisé sur les intervalles donnés par 75% des participants, les 25% restants constituant le jeu de test. (ii) ENA (EB), où l'apprentissage est réalisé sur les intervalles donnés par tous les participants à 17 (66,7%) des ENA, les 7 ENA restantes sont utilisées comme jeu de test ; Chaque benchmark est composé de 1000 décompositions aléatoires en jeu d'apprentissage / test, avec la contrainte qu'ils doivent inclure un mélange d'ENA rondes et pseudo-rondes.

Pour déterminer le modèle donnant les meilleurs résultats, des analyses statistiques basées sur des tests ANOVA avec le modèle comme facteur sont réalisés. Lorsqu'un effet significatif est observé, des tests post-hoc HSD de Tukey sont réalisés. Le seuil de significativité est fixé à $p = 0,01$.

Resultats. Le tableau 3 présente les scores des modèles. Le classement des modèles est le même pour les deux benchmarks PB et EB, les performances observées sont similaires.

En premier lieu, on observe que le modèle proposé PFM obtient les meilleurs scores, aussi bien en précision de médiane (MA) qu'en erreur d'estimation ($MErr$), apportant ainsi une validation empirique du modèle.

Le comportement de REGM (MA faible, $MErr$ dans la moyenne) peut être dû au fait qu'il produit des estimations dans \mathbb{R} alors que les participants ont tendance à donner comme réponse des nombres ronds ou pseudo-ronds. Toutefois, le score $MErr$ dans la moyenne in-

Tableau 2 – Définitions formelles des critères d'évaluation de les 0,5-coupes estimées.

$$MA(m) = \frac{1}{2 \cdot |\mathcal{X}|} \sum_{x \in \mathcal{X}} \left\{ b \left| \frac{|\Delta M_m^b(x) - \Delta Med^b(x)|}{\Delta Med^b(x)} \leq 0.1 \right. \right\}$$

$$MErr(m) = \frac{1}{2 \cdot |\mathcal{X}|} \cdot \sum_{x \in \mathcal{X}} \sum_b (|N_+ - N_+^*| + |N_- - N_-^*|)$$

Tableau 3 – Moyennes et écarts-types des deux critères pour chaque modèle. Les scores en gras sont significativement les meilleurs.

Benchmark Participants		
Modèle	MA (%)	MErr
SBM	28,0 (6,9)	0,76 (0,08)
REGM	20,0 (7,2)	0,67 (0,18)
PFM	58,3 (8,9)	0,35 (0,12)
Benchmark ENA		
SBM	24,9 (14,5)	0,79 (0,18)
REGM	15,7 (13,2)	0,65 (0,14)
PFM	63,8 (14,0)	0,27 (0,16)

dique que ces estimations sont proches des médianes observées. A l'inverse, SBM obtient de meilleures performances que REGM en précision de médiane alors que les erreurs d'estimation sont plus importantes.

5.2 Estimations du support et du noyau

Critère de qualité. La présence de valeurs extrêmes ne permet pas d'évaluer la qualité du support et du noyau estimés de la même manière que la 0,5-coupe. En effet, dans la perspective *random set*, le support correspond à l'intervalle observé le plus large et le noyau au plus réduit. Par conséquent, la présence d'une seule valeur extrême conduit à un support ou un noyau extrême. La précision de l'estimation ou la distance à la valeur observée manquent donc de robustesse vis-à-vis des valeurs extrêmes.

Pour surmonter ce problème, nous proposons de générer une fonction d'appartenance $f_x^G(y)$ en interpolant linéairement le support, la 0,5-coupe et le noyau. Cette fonction d'appartenance est ensuite comparée à celle, de référence, élicitée à partir des données, $f_x^E(y)$. Cette dernière est construite dans la perspective *random set*, qui définit $f_x^E(y)$ comme la fréquence relative cu-

mulée de participants incluant y dans l'intervalle correspondant à x .

Nous proposons de comparer $f_x^G(y)$ à $f_x^E(y)$ en utilisant l'aire de leur différence, relativement à l'aire de la fonction de référence $f_x^E(y)$. Ce critère, mesurant la qualité de la fonction générée, est à minimiser et peut être formalisé comme :

$$MFQ(x) = \frac{\int_y |f_x^G(y) - f_x^E(y)|}{\int_y f_x^E(y)}$$

Resultats. La figure 3 illustre quatre exemples de fonctions générées et élicitées. Les paliers observés sur $f_x^E(y)$ sont dus aux valeurs de bornes fréquemment données par les participants.

Les fonctions générées correspondent visuellement bien à celles élicitées pour 150, 400 et 8150, correspondant à des scores MFQ valant respectivement 0,211, 0,397 et 0,618. De plus, l'asymétrie de $f_x^E(y)$ est capturée, validant notre modèle PFM.

Le score MFQ moyen est de 0,502 ($\sigma = 0,175$), allant de 0,211 ($x = 150$) à 0,950 ($x = 1100$). Avec un seuil à 0,6 pour considérer une estimation comme correcte, 17 des 24 (70,1%) fonctions d'appartenance sont validées.

Comme attendu, la présence de valeurs extrêmes (par ex. 7500 et 10000 pour $x = 8150$; 100 et 600 pour $x = 400$) réduit le score de certaines ENA. Dans le cas de $x = 1100$ (figure 3, en haut à droite), la faible correspondance visuelle et la valeur élevée de MFQ , 0,950, peuvent être expliquées par le fait que le front de Pareto supérieur finit à 2000, une valeur qui n'est pas donnée par les participants.

Il apparaît que les scores moyens obtenus pour les ENA rondes (0,488) et pseudo-rondes (0,524) sont similaires. Toutefois, les écarts-types montrent une variabilité plus importante pour les ENA pseudo-rondes (0,272) que pour

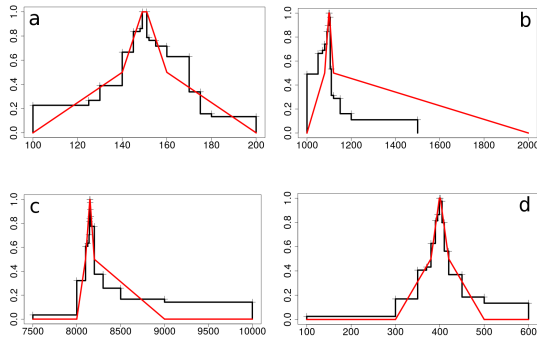


Figure 3 – Fonctions d’appartenance générées (en rouge) et élicitées (en noir) pour (a) $x = 150$, (b) $x = 1100$, (c) $x = 8150$, et (d) $x = 400$.

les rondes (0,087), ce qui indique que certaines ENA sont mieux capturées que d’autres.

6 Conclusion et travaux futurs

Nous avons proposé un modèle d’interprétation des ENA ”environ x “ sous la forme de nombres flous. Plus spécifiquement un modèle computationnel, basé sur des points critiques des fronts de Pareto et capturant la saillance cognitive des nombres, a été proposé pour caractériser le support, le noyau et la 0,5-coupe des fonctions d’appartenance correspondantes.

Nous avons conduit une étude expérimentale sur des données collectées à l’aide d’un questionnaire en ligne, qui valide le modèle proposé : il obtient de meilleurs scores que les modèles de la littérature pour l’estimation des 0,5-coupes. De plus, les fonctions d’appartenance générées en interpolant linéairement les estimations sont des approximations satisfaisantes des fonctions d’appartenance élicitées.

Les perspectives visent à étudier la pertinence d’utiliser d’autres points des fronts de Pareto pour caractériser des α -coupes spécifiques de manière à mieux s’accorder aux fonctions d’appartenance élicitées. Les perspectives visent aussi à étendre le modèle pour tenir compte du contexte d’une ENA puisque celui-ci a un effet sur son interprétation [8, 10]. En effet, ”environ 10 000 euros”, par exemple, peut être interprété différemment selon qu’elle est prononcée

par un vendeur ou un acheteur.

Remerciements

Ce travail été effectué dans le cadre du Labex SMART (ANR-11-LABX-65) et a bénéficié d’une aide de l’Etat gérée par l’Agence Nationale de la Recherche au titre du programme Investissements d’Avenir portant la référence ANR-11-IDEX-0004-02.

Références

- [1] T. Bilgiç and I. B. Türkşen. Measurement of membership functions : Theoretical and empirical work. In *Fundamentals of Fuzzy Sets*, volume 7. Springer US, 2000.
- [2] S. Dehaene and J. Mehler. Cross-linguistic regularities in the frequency of number words. *Cognition*, 43(1) :1–29, 1992.
- [3] M. Ehrgott. *Multicriteria optimization*. Springer, Berlin, 2013.
- [4] S. Ferson, J. O’Rawe, A. Antonenko, J. Siegrist, J. Mickley, C. C. Luhmann, K. Sentz, and A. M. Finkel. Natural language of uncertainty : numeric hedge words. *IJAR*, 57 :19–39, 2015.
- [5] A. González, O. Pons, and M. A. Vila. Dealing with uncertainty and imprecision by means of fuzzy numbers. *IJAR*, 21(3) :233–256, 1999.
- [6] C. J. M. Jansen and M. M. W. Pollmann. On round numbers : Pragmatic aspects of numerical expressions. *J. of Quantitative Linguistics*, 8(3) :187–201, 2001.
- [7] M. Krifka. Approximate interpretations of number words : A case for strategic communication. In *Cognitive foundations of interpretation*, pages 111–126. 2007.
- [8] P. Lasersohn. Pragmatic halos. *Language*, 75(3) :522–551, 1999.
- [9] U. Sauerland and P. Stateva. Scalar vs. epistemic vagueness : Evidence from approximators. *Proc. of SALT*, pages 228–245, 2007.
- [10] S. Solt. An alternative theory of imprecision. In *Semantics and Linguistic Theory*, volume 24, pages 514–533, 2014.