How much is " about "? Fuzzy interpretation of approximate numerical expressions

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

How much is "about"? Fuzzy interpretation of approximate numerical expressions

Sébastien Lefort[1], Marie-Jeanne Lesot[1],
Elisabetta Zibetti[2], Charles Tijus[2], and Marcin Detyniecki[1,3]

[1] Sorbonne Universités, UPMC Univ Paris 06,
CNRS, LIP6 UMR 7606, 4 place Jussieu 75005 Paris
{sebastien.lefort,marie-jeanne.lesot,marcin.detyniecki}@lip6.fr
[2] Laboratoire CHArt-LUTIN, EA 4004, Université Paris 8,
2 rue de la liberté, 93526, Saint-Denis - Cedex 02, France
{ezibetti,tijus}@univ-paris8.fr
[3] Polish Academy of Sciences, IBS PAN, Warsaw, Poland

Abstract. Approximate Numerical Expressions (ANEs) are linguistic expressions involving numbers and referring to imprecise ranges of values, such as "*about 100*". This paper proposes to interpret ANEs as fuzzy numbers. A model, taking into account the cognitive salience of numbers and based on critical points from Pareto frontiers, is proposed to characterise the support, the kernel and the 0.5-cut of the corresponding membership functions. An experimental study, based on real data, is performed to assess the quality of these estimated parameters.

Keywords: Approximate Numerical Expression, Fuzzy number, Pareto frontier, Empirical study, Number salience

1 Introduction

Approximate numerical expressions (ANEs) are vague linguistic expressions of the general form "*about* x" where x is a number. They are used in daily life to denote imprecise ranges of values, e.g., "Berlin is located at *about 900km* from Paris"; "The patient has had fever for *about one week*". In the field of Human-Computer Interfaces, ANEs raise the issues of their interpretation, i.e., the estimation of the range of values they designate and their representation in information systems, for instance as intervals of values or as fuzzy sets.

From a linguistic perspective, Lasersohn [10] proposes to formalise vagueness in a general context, beyond the case of numerical expressions, through the use of pragmatic halos, defined as the union of the entity that is explicitly referred to by a vague expression and entities of the same semantic type that are implicitly denoted. For instance, in the proposition "*there were about 100 participants at the meeting*", the pragmatic halo of the vague expression "*about 100*" corresponds to 100 exactly and a range of possible values around 100 (e.g., $[90; 110]$). Therefore, interpreting an ANE corresponds to estimating the range of values that satisfy it, i.e., the values that are included in its pragmatic halo.

A natural approach to model the fuzziness in boundary values is to use fuzzy sets [14, 15], that lead to represent ANEs as fuzzy numbers [16], defined by their membership functions. Fuzzy numbers are classically used to represent uncertainty or imprecision in numerical data [5]. However, to the best of our knowledge, no attempt has been made to empirically characterise the membership functions of fuzzy numbers related to ANEs in natural language.

The aim of this paper is to propose a model to characterise the support, the kernel and the 0.5-cut of fuzzy numbers corresponding to ANEs of the form "*about* x", for $x \in \mathbb{N}$. More specifically, the model is based on critical points from Pareto frontiers, as a compromise between the numbers cognitive salience and their distance to the reference value x. An empirical study is conducted to collect real data and to perform an experimental validation to highlight the quality of the estimations provided by the model.

The paper is structured as follows: Section 2 describes previous works and existing models. The proposed model is presented in Section 3. The data collection procedure is described in Section 4. Section 5 presents the experimental study and its results. Finally, conlusions and future works are discussed in Section 6.

2 Related Works

This section introduces the notations and definitions of dimensions and properties of ANEs used in this work. Two models from the literature, estimating the range of denoted values, are then presented: a scale-based model [8, 13] and a regression model [4]. Finally, the fuzzy set approach to vagueness is discussed.

2.1 Definitions and Notations

The ANEs considered in this paper are of the form "*about* x", for $x \in \mathbb{N}$. In the decimal system, x can be written as $x = \sum_{i=0}^{q} a_i \cdot 10^i$, where $a_i \in [\![0, 9]\!]$. We propose four dimensions, formally defined in Tab. 1, to characterise x: granularity $Gran(x)$ is the power of ten x belongs to, relative magnitude $R_m(x)$ is the value of its last significant digit and precision $Prec(x)$ is the product of granularity and relative magnitude. These dimensions are expected to influence the interpretation of ANEs. For instance, precision is meant to reflect the expectation that the width of the interval corresponding to "*about 30.050*" is comparable to the one of "*about 150*", 50 being the common part.

From these dimensions, two classes of natural numbers can be distinguished. Round numbers are classically defined as multiples of 10 with a single significant digit (e.g., 50 or 8000). We propose to define pseudo-round numbers as multiples of 10 with at least two significant digits (e.g., 320 or 8150).

Beyond these arithmetical characteristics, we propose another one, taking into account a cognitive component. Indeed, it has been observed than some numbers occur more frequently than others in corpuses [7, 2] and complexity $Cpx(x)$ aims at capturing this salience. It appears that, firstly, the more significant digits a number has, the lower its frequency. Secondly, numbers whose last

Dimension	Formal definition	Example $x = 4750$
Granularity	$Gran(x) = 10^{i^*}$ where $i^* = \min\{i \mid a_i \neq 0\}$	10
Relative magnitude	$R_m(x) = a_{i^*}$	5
Precision	$Prec(x) = a_{i^*} \cdot 10^{i^*}$	50
Number of significant digits	$NSD(x) = q - i^* + 1$	3
Complexity	$Cpx(x) = NSD(x) - B(x)$	2.5

Table 1: Dimensions of a natural number $x = \sum_{i=0}^{q} a_i \cdot 10^i$, illustrated by $x = 4750$ in the last column. $B(x)$, used in the complexity definition, is defined in Eq. (1).

significant digit is 5 or, to a lower extent, 2, occur more frequently. For symmetry reasons around multiples of 10, we propose to process numbers with $R_m(x) = 8$ (e.g., $18 = 20 - 2$) as numbers with $R_m(x) = 2$ (e.g., $22 = 20 + 2$). Thus, we propose to formalise the complexity of a number as its number of significant digits minus a bonus to capture these specific cases, if the number of significant digits is at least 2.

The bonus function thus distinguishes three categories, depending on the value of the last significant digit $R_m(x)$ and respecting the order of frequency of appearance: $B(x_1) > B(x_2) > B(x_3)$, for $x_1, x_2, x_3 \in \mathbb{N}$ such that $R_m(x_1) = 5$, $R_m(x_2) \in \{2, 8\}$ and $R_m(x_3) \notin \{2, 5, 8\}$. We arbitrarily propose to set these values at 0.5, 0.25 and 0. The bonus function is therefore formalised as:

$$B(x) = \begin{cases} 0.5 & \text{if } R_m(x) = 5 \text{ and } NSD(x) > 1 \\ 0.25 & \text{if } R_m(x) = 2 \text{ or } R_m(x) = 8 \text{ and } NSD(x) > 1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The plus signs on Figure 1 illustrate the complexity $Cpx(x)$ for all integers x between 400 and 500.

2.2 Scale-Based Models (SBM)

The first approach in interpreting ANEs is proposed from a linguistic perspective and models the range of denoted values as an interval. Scale-based models (SBM) [8, 12, 13] rely on scale systems $S = \{s_1, \ldots, s_n\}$, where s_i are granularity levels such that $s_i < s_{i+1}$. As examples, one can mention the time scale-system, $S = \{1 \text{ min}, 5 \text{ min}, 15 \text{ min}, \ldots\}$, or the decimal one, $S = \{1, 10, 100, \ldots\}$.

The interpretation of a numerical expression can occur at any granularity level. For instance, in the decimal system, the numerical expression "*100*" can be interpreted at the 1, 10 or 100 levels. The finer the granularity, the narrower the interval. Speakers express the intended level through the use of approximators [12]: "*exactly*" refers to the finest granularity level the expression belongs to, while "*about*" refers to the coarsest one (e.g., the level of thousands for "*about 1000*"), formally defined as $Gran_C(x) = \sup(\{s_i \in S \mid x \mod s_i = 0\})$. If the scale-system S is the decimal system, $Gran_C(x) = Gran(x)$ (see Tab. 1).

SBM proposes that the values denoted by an ANE x are the ones closer to x than to any other number on $Gran_C(x)$.The interval is formally defined as:

$$I_{SBM}(x) = [x - Gran_C(x)/2; x + Gran_C(x)/2] \tag{2}$$

For instance, $I_{SBM}(300) = [250; 350]$; $I_{SBM}(8150) = [8145; 8155]$. This approach has the advantage of taking into account the ANE granularity; however, it does not address the issue of the relative magnitude: all ANEs at the same granularity level result in the same interval width, although, one may expect, for instance, that the interval of "*about 100*" would be narrower than the one of "*about 800*".

2.3 Regression Model (REGM)

Ferson et al. [4] propose an empirical approach using real data to test the relevance of predictors of the interval width. Semantically contextualised ANEs (e.g., "*Roughly 25% of Canadians are Protestant.*") were presented to participants, who were asked to estimate the boundaries of the corresponding intervals. The proposed model then estimates the interval as:

$$\begin{aligned} I_{REGM}(x) &= \left[x - \frac{10^{L(x)}}{2}; x + \frac{10^{L(x)}}{2}\right] \\ \text{where } L(x) =& A + B \cdot O_m(x) + C \cdot R(x) + D \cdot f(x) \\ &+ E \cdot O_m(x) \cdot R(x) + F \cdot O_m(x) \cdot f(x) + G \cdot R(x) \cdot f(x) \\ &+ H \cdot O_m(x) \cdot R(x) \cdot f(x) \end{aligned} \tag{3}$$

where A to H are parameters empirically set by performing a regression on the data. $O_m(x)$ is the ANE order of magnitude ($O_m(x) = \log_{10}(x)$), $R(x)$ its roundness ($R(x) = i^* + 1$), and $f(x)$ its "fiveness", defined as $f(x) = 1$ if $a_{i^*} = 5$, $f(x) = 0$ otherwise. $O_m(x)$, $R(x)$, $f(x)$ and their combinations have been empirically selected as predictors for the interval width.

This model presents the advantage of allowing the adaptation to different contexts by learning parameters on a dataset. However, it can be noted that the semantic context is not controlled in the experimental setting although mixing different contexts may result in interactions between this factor and the ones related to the ANE reference number.

2.4 Fuzzy Representation of Vagueness

From a linguistic perspective, Lakoff [9] considers that every term in natural language is, to some extent, fuzzy: category membership is not a matter of all or nothing, but rather a matter of degrees. As supported by empirical evidence [6], fuzzy logic is therefore a relevant formalisation of the vagueness inherent to natural language: any term can be modeled by a membership function.

Among the natural language terms, numerical expressions can be represented as fuzzy numbers [16], defined as fuzzy sets on the universe $\mathbb{R}$. From this point

of view, approximators are modifiers of the membership function of the fuzzy reference value [11]. For instance, the approximator *exactly* narrows the curve of the membership function whereas *approximately* widens it.

Interpreting an ANE x therefore consists in estimating its membership function, $f_{\widetilde{x}}(y)$, where y are values that can be denoted by "*about* x". Among various methods to elicit such membership functions (see, e.g. [1]), the random set view interprets the membership degree of a candidate number (e.g., 95 for "*about 100*") as the cumulative frequency of participants thinking that it belongs to the interval denoted by the ANE. Thus, if half of the population think that 95 is included in "*about 100*", the truth value of 95 is 0.5. The median of the distribution is therefore a critical point for membership functions that corresponds to the 0.5 membership degree.

3 Proposed Model

This section describes the model we propose to estimate the support, the kernel and the 0.5-cut of fuzzy numbers corresponding to ANEs.

The Pareto Frontiers Model (PFM): The model we propose is based on the assumption that, when interpreting an ANE, human beings tend to make a compromise between the cognitive cost of boundary values, which can be measured by the complexity $Cpx(x)$, on one hand, and the range of denoted values, measured by the distance between the boundaries of the interval and the ANE x, on the other hand. It implies that, for a given range of denoted values, the cognitive cost is minimised; reciprocally, for a given cognitive cost, the range of denoted values is minimised. For instance, given the ANE "*about 500*", participants of the empirical study (see Sect. 4) tend to give answers such as [499; 501], [490; 510] or [450; 550]. The boundaries of these intervals are the closest to the ANE when $Cpx(x)$ is 3, 2 and 1.5. Therefore, the values that optimise the compromise are better candidates to be the boundaries than all other values.

As a consequence, the model we propose first consists in determining these good candidates by generating Pareto frontiers [3]: all possible candidate values v in $[1; x[$ for the lower boundary, and in $]x, +\infty[$ for the upper boundary of the ANE x are compared on two criteria (i) the absolute distance from the ANE: $d_x(v) = |v - x|$; (ii) the complexity $Cpx(v)$. The selected values, constituting the Pareto frontier, are those that are not dominated by any other value. For a given ANE, two Pareto frontiers are considered: $P^-(x) = [y_1^-, \ldots, y_{n^-}^-]$ relates to the lower boundary of the interval and $P^+(x) = [y_1^+, \ldots, y_{n^+}^+]$ to the upper one, ordered by increasing distance to x, $d_x(y_i)$. Figure 1 illustrates these Pareto frontiers for the ANE "*about 440*": $P^-(440) = [439, 438, 435, 430, 420, 400]$ and $P^+(440) = [441, 442, 445, 450, 500]$. One can notice that the model naturally captures the asymmetry observed in the data (see Section 4) due to salient numbers (e.g., 420, 450) in the reference number neighborhood.

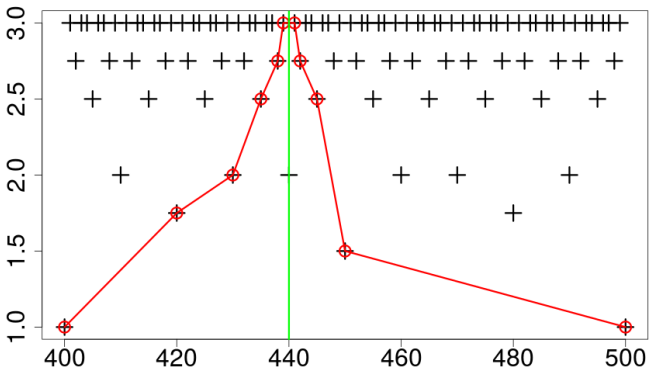

Fig. 1: Pareto frontiers (red lines) for lower (left from green line) and upper (right from green line) boundaries of the ANE "*about 440*". Black plus signs represent the complexity $Cpx(x)$ for each integer value in $[400; 500]$ (see Tab. 1).

The second step of the model we propose consists in using the values in the Pareto frontiers as candidates to be the boundaries of the support, kernel and 0.5-cut limits of fuzzy numbers corresponding to ANEs.

Support, Kernel and 0.5-cut Estimations: Any value outside the support interval, noted $I_S(x)$, is considered as not referred by the ANE. We therefore propose to define the farthest values from x of the Pareto frontiers as boundaries of this interval, formally: $I_S(x) = [y^-_{n^-}; y^+_{n^+}]$.

Any value inside the kernel interval, noted $I_K(x)$, is considered as being fully denoted by the ANE. We propose to define the nearest values from x of the Pareto frontiers as boundaries of this interval, formally: $I_K(x) = [y^-_1; y^+_1]$.

The boundaries of the 0.5-cut interval, noted $I_M(x)$, are also selected according to their rank in $P^-(x)$ and $P^+(x)$. We propose to make the chosen rank dependent on the considered ANE x, so as to make the model more flexible. More, precisely, we propose that the rank of the boundary estimation depends on the number of significant digits $NSD(x)$ and the precision $Prec(x)$ of the ANE: an exhaustive analysis of empirical data (omitted in this paper for reasons of space) has validated them as factors influencing ANE interpretation. The rank is computed as:

$$r_P(x) = round\left(\log(Prec(x)) - 1 + \sum_{k=1}^{NSD(x)} k\right) \tag{4}$$

The estimation of the 0.5-cut interval is then $I_M(x) = [y^-_{r_P(x)}; y^+_{r_P(x)}]$.

For the example $x = 440$, as illustrated in Figs. 1 and 2, one obtains: $I_S(440) = [400; 500]$, $I_K(440) = [439; 441]$ and $I_M(440) = [430; 450]$.

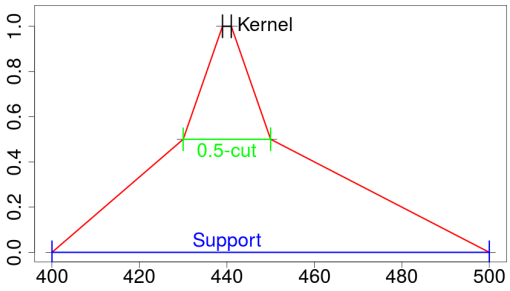

Fig. 2: Support, kernel and 0.5-cut of the membership function for ANE *about 440*, based on critical points from Pareto frontiers with piecewise linear interpolation.

4 Data collection

We conducted an empirical study to collect real intervals corresponding to ANEs so as to experimentally validate our proposed model. This section presents the methods used to collect and process the data.

Material: An online questionnaire containing 24 uncontextualised ANEs, 15 round (20, 30, 40, 50, 80, 100, 200, 400, 500, 600, 800, 1000, 2000, 6000 and 8000) and 9 pseudo-round (110, 150, 440, 560, 1100, 1500, 4700, 4730 and 8150) was designed. These values have been selected in order to cover different combinations of dimensions, to avoid biases towards any specific one: several relative magnitudes at a granularity level (e.g., 20/40/80), several granularity levels at a relative magnitude (e.g., 80/800/8000), several numbers of significant digits at the same precision (e.g., 50/150/8150). ANEs are presented in a random order. The instructions, given in French, can be translated as "*In your opinion, what are the MINIMUM and MAXIMUM values associated with "about x"?*". This questionnaire meets the criteria proposed by [1] to elicit membership functions in a random set perspective. This method is also similar to the one used by [4].

146 participants have been recruited through an announcement diffused on mailing-lists: 102 women and 44 men, aged 20 to 70 ($M = 38.6$; $\sigma = 14.2$).

Data Preprocessing: The answer to ANE x given by participant p is noted $I_p(x) = [I_p^-(x); I_p^+(x)]$. It is considered as an outlier if: (i) it is inadequate (e.g., $[0; infinity]$), (ii) $I_p^-(x) > x$ or $I_p^+(x) < x$ (e.g., $I(800) = [700; 750]$ or $I(800) = [810; 850]$), or (iii) $I_p^-(x) < x/10$ or $I_p^+(x) > 10x$ (e.g., $I(100) = [10; 1100]$). In a second step, mean and standard deviation are computed for the remaining boundaries of each ANE. Any boundary value beyond three standard deviations of the mean is considered as an outlier. Finally, participants with more than 70% missing values or outliers are considered as untrustworthy and all their answers are excluded. The analyses include 3177 (91%) of the 3504 collected intervals.

Global Observations: In the collected data, not detailed here, it can be observed that participants tend not to agree on the intervals: on average, 15.4 different answers per boundary are obtained, ranging from 9 (for "*about 20*") to 22 (for "*about 8150*"). However, 84.4% of the boundaries are located on the Pareto frontiers as defined in Section 3, which validates the principle underlying the model we propose.

When examining whether the provided intervals are symmetric around the reference value, the collected data show that symmetry depends on the ANE: 74.2% are symmetric with respect to the considered ANE, but intervals of some ANEs, such as 440 or 4730, are less often symmetric (63% and 50% respectively). This observation validates the definition of a flexible model allowing for non-symmetric observations.

5 Experimental Study

This section presents the experimental study we performed in order to assess the quality of the three estimated parameters of fuzzy numbers corresponding to ANEs: 0.5-cut, support and kernel. The used quality criteria and the results of each parameter are described in the next subsections.

5.1 Evaluation of the 0.5-cut Estimation

In the random set view of membership functions [1], 0.5-cuts correspond to the median of the intervals given by the participants. Thus, to evaluate the 0.5-cut estimation, we propose to compare it to this median interval.

As the models from the literature [4, 13] are not fuzzy, they can be used to estimate either the support, the kernel or the 0.5-cut. We propose to use them to predict the 0.5-cut as it is a central indicator of the boundary distributions.

Quality Criteria: We note X the set of considered ANEs and $P(x)$ the set of participants whose intervals are not considered as outliers for $x \in X$. Moreover, we note the prediction of model m $[m^-(x); m^+(x)]$, $\Delta M_m^b(x) = |m^b(x) - x|$ its distance from x for $b \in \{-, +\}$, and $\Delta Med^b(x)$ the median of the distances $\Delta P_p^b(x) = |I_p^b(x) - x|$ over all participants p in $P(x)$.

To assess whether the estimations are correct, we first propose to use the accuracy score of the median prediction, i.e., the number of boundary values for which the relative distance to the observed median is lower than 10%. The median accuracy, MA, to be maximised, can be formalised as:

$$MA(m) = \frac{1}{2 \cdot |X|} \sum_{x \in X} \left| \left\{ b \in \{-, +\} \,\middle|\, \frac{|\Delta M_m^b(x) - \Delta Med^b(x)|}{\Delta Med^b(x)} \leq 0.1 \right\} \right| \quad (5)$$

Secondly, to assess the degree of error, we propose to evaluate the balance between participants who are above and below the estimated 0.5-cut, formally defined as: $N_+ = |\{p \in P(x) | \Delta P_p^b(x) > \Delta M_m^b(x)\}|$ and $N_- = |\{p \in$

$P(x)|\Delta P_p^b(x) < \Delta M_m^b(x)\}|$. A correct estimation of the median interval implies that the model m should be such that $N_+ = N_-$ for all x, b.

However, since interval boundaries given by the participants are distributed on few points, a perfect balance may not be possible. Therefore, the score takes into account the balance of the actual median, i.e., $N_+^* = |\{p \in P(x)|\Delta P_p^b(x) > \Delta Med^b\}|$ and $N_-^* = |\{p \in P(x)|\Delta P_p^b(x) < \Delta Med^b\}|$.

The score of the model then depends on the difference between N_+ and N_+^* and between N_- and N_-^*. Averaging over the two boundaries $b \in \{-, +\}$ and all considered ANEs, the median error, to be minimised, can be defined as:

$$MErr(m) = \frac{1}{2 \cdot |X|} \cdot \sum_{x \in X} \sum_{b \in \{-,+\}} (|N_+ - N_+^*| + |N_- - N_-^*|) \tag{6}$$

Experimental Procedure: Using these quality criteria, we compare the performances of our proposed Pareto frontiers model PFM, the scale-based model SBM [8, 13] with the decimal system (i.e.,$S = \{1, 10, 100, \ldots\}$), and the regression model REGM [4]. The latter only provides the size of the intervals and no information about their location or symmetry around the ANE. We make the assumption that they are symmetric and centered on x.

A cross-validation procedure is performed on two benchmarks, (i) Participant (PB): REGM learning is performed on the intervals given by 75% of the participants, the remaining 25% constitute the test dataset. (ii) ANE (AB): REGM learning is performed on the intervals given by all participants on 17 (66.7%) of the ANEs. The 7 remaining ANEs are used as test dataset. Each benchmark consists in 1000 random decompositions of the learning / test datasets, with the constraint that they must include a mix of round and pseudo-round ANEs.

In order to determine which model shows the best results in each benchmark, statistical analyses using ANOVA tests with model as factor, and Tukey's HSD post-hoc tests are performed. The significance threshold is set at $p = .01$.

Results: Table 2 shows the performances of the models. Results are similar in both the Participant and the ANE benchmarks.

It can firstly be observed that our proposed model PFM shows the best performances, both in median prediction accuracy (MA) and in median estimation error ($MErr$), providing an empirical validation.

The behaviour of REGM (poor MA but an average $MErr$) can be due to the fact that it provides real-numbered boundary estimations while participants tend to give round or pseudo-round numbers, leading to erroneous predictions. However, the average $MErr$ indicates that these real-numbered estimations are close to the actual medians. On the contrary, SBM appears to perform better than REGM on prediction accuracy while the prediction errors are much more important.

Model	MA (%) - PB	$MErr$ - PB	MA (%) - AB	$MErr$ - AB
SBM	28.0 (6.9)	0.76 (0.08)	24.9 (14.5)	0.79 (0.18)
REGM	20.0 (7.2)	0.67 (0.18)	15.7 (13.2)	0.65 (0.14)
PFM	**58.3 (8.9)**	**0.35 (0.12)**	**63.8 (14.0)**	**0.27 (0.16)**

Table 2: Means and standard deviations of the two criteria for each model on the Participant (PB, left) and the ANE benchmarks (AB, right). Bold scores are the statistically best ones according to the ANOVA and post-hoc tests.

5.2 Evaluation of the Support and Kernel Estimations

Quality Criterion: Assessing the quality of the support and the kernel estimations the same way as the 0.5-cut raises the issue of the outliers. Indeed, in the random set view, the support corresponds to the largest interval, and the kernel corresponds to the narrowest one. Therefore, the presence of a single extreme answer results in aberrant support or kernel values. Prediction accuracy or distance to actual values thus lack robustness with respect to extreme values.

To overcome this issue, we propose to build a basic piecewise linear membership function, $f^G_{\widetilde{x}}(y)$, obtained by linking the generated points of support, 0.5-cut and kernel and to compare it to an elicited reference fuzzy set $f^E_{\widetilde{x}}(y)$. We build the latter in a random set view [1], defining $f^E_{\widetilde{x}}(y)$ as the cumulative relative frequency of participants including y in the interval corresponding to x.

We propose to compare $f^G_{\widetilde{x}}(y)$ to $f^E_{\widetilde{x}}(y)$ using the area of their difference, relatively to the area of the reference $f^E_{\widetilde{x}}(y)$. This criterion, measuring the membership function quality, to be minimised, can be formalised as:

$$MFQ(x) = \frac{\int_y |f^G_{\widetilde{x}}(y) - f^E_{\widetilde{x}}(y)|}{\int_y f^E_{\widetilde{x}}(y)} \tag{7}$$

Results: Figure 4 illustrates four examples of elicited and generated membership functions. The high steps observed in $f^E_{\widetilde{x}}(y)$ are due to boundary values frequently given by participants.

The generated membership functions visually fit well the elicited ones of 150, 400 and 8150, corresponding to MFQ scores 0.211, 0.397 and 0.618 respectively. Moreover, the asymmetry of the $f^E_{\widetilde{x}}(y)$ is captured, validating our PFM model.

The mean quality score is 0.502 ($\sigma = 0.175$), ranging from 0.211 ($x = 150$) to 0.950 ($x = 1100$). Setting a threshold at $MFQ = 0.6$ to consider a good estimation, 17 over 24 (70.1%) generated membership functions are correct.

As expected, the presence of outliers (i.e., 7500 and 10000 for $x = 8150$; 100 and 600 for $x = 400$) lowers the score of some ANEs. In the particular case of $x = 1100$ (Fig. 4, top right), the poor obtained fitting and score ($MFQ = 0.950$) can be explained by the fact that the upper Pareto frontier ends at 2000, a value not given by participants.

When detailing the difference between round and pseudo-round ANEs, it appears that the mean scores obtained for round (0.488) and pseudo-round numbers

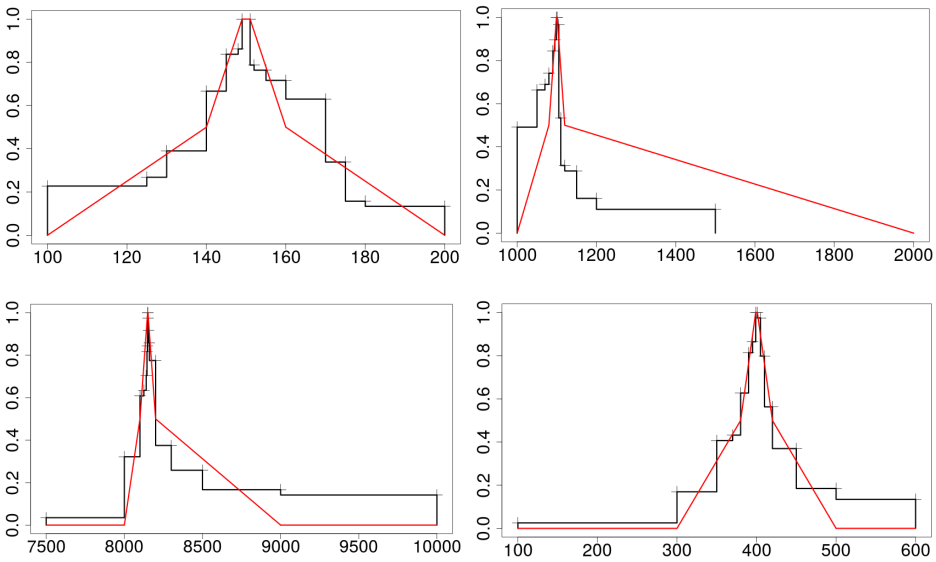

Fig. 4: Generated (red) and elicited (black) membership functions of four ANEs: $x = 150$ (top, left), $x = 1100$ (top, right), $x = 8150$ (down, left), and $x = 400$ (down, right).

(0.524) are similar. However, the standard deviation reveals a higher significantly variability for pseudo-round numbers (0.272) than for round numbers (0.087), indicating that some ANEs are well captured while some other are less. In particular, $x = 1100$ ($MFQ = 0.950$) and $x = 4730$ ($MFQ = 0.864$) result in scores far from the mean, compared to other ANEs.

6 Conclusion and Future Works

In this paper, we propose a model to interpret ANEs of the form "*about* x" as fuzzy numbers. More specifically, a computational model, based on critical points from Pareto frontiers and capturing the cognitive dimension of number salience, is proposed to characterise the support, the kernel and the 0.5-cut of the corresponding membership functions.

We conducted an experimental study on real data collected from an online questionnaire, which validates the proposed model: it shows that PFM performs better than the models from the literature in 0.5-cut estimation. Moreover, the piecewise linear membership functions generated from the estimations are close approximations of the elicited ones.

Future work will study the relevance of including other points from the Pareto frontiers as specific α-cuts to better fit the elicited membership functions. It will also focus on extension of the model to take into account the context of an ANE occurence as it has an effect on ANE interpretation [10, 13]. Indeed, "*about*

10.000 euros", for instance, may not be interpreted the same way it is said by a seller or a buyer. Extensions of the model will focus on other linguistic approximators, such as "*at least*" or "*less than*".

Finally, the proposed model will be implemented in applications such as search engines to improve the relevance of answers provided to approximate queries.

Aknowledgments. This work was performed within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements d'Avenir programme under reference ANR-11-IDEX-0004-02.

References

1. T. Bilgiç and I. B. Türkşen. Measurement of membership functions: Theoretical and empirical work. In *Fundamentals of Fuzzy Sets*, volume 7 of *The Handbooks of Fuzzy Sets Series*, pages 195–227. 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*, volume 491. Springer Science & Business Media, 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. *International Journal of Approximate Reasoning*, 57:19–39, 2015.
5. A. González, O. Pons, and M. A. Vila. Dealing with uncertainty and imprecision by means of fuzzy numbers. *International Journal of Approximate Reasoning*, 21(3):233–256, 1999.
6. H. M. Hersh and A. Caramazza. A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology: General*, 105(3):254–276, 1976.
7. C. J. M. Jansen and M. M. W. Pollmann. On round numbers: Pragmatic aspects of numerical expressions. *Journal of Quantitative Linguistics*, 8(3):187–201, 2001.
8. M. Krifka. Approximate interpretations of number words: A case for strategic communication. In *Cognitive foundations of interpretation*, pages 111–126. Amsterdam, 2007.
9. G. Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4):458–508, 1973.
10. P. Lasersohn. Pragmatic halos. *Language*, 75(3):522–551, 1999.
11. E. F. Prince, J. Frader, C. Bosk, and R. J. Dipietro. On hedging in physician-physician discourse. In *Linguistics and the professions*. Ablex, Norwood, NJ, 1982.
12. U. Sauerland and P. Stateva. Scalar vs. epistemic vagueness: Evidence from approximators. *Proceedings of SALT*, (1995):228–245, 2007.
13. S. Solt. An alternative theory of imprecision. In *Semantics and Linguistic Theory*, volume 24, pages 514–533, 2014.
14. L. A. Zadeh. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.
15. L. A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning. *Information sciences*, 8(3):199–249, 1975.
16. L. A. Zadeh. A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with applications*, 9(1):149–184, 1983.