HAL open science

Validation of Smets' hypothesis in the crowdsourcing environment

Constance Thierry, Arnaud Martin, Jean-Christophe Dubois, Yolande Le Gall

To cite this version:

Constance Thierry, Arnaud Martin, Jean-Christophe Dubois, Yolande Le Gall. Validation of Smets' hypothesis in the crowdsourcing environment. 6th International Conference on Belief Functions, Oct 2021, Shanghai, China. hal-03348663

HAL Id: hal-03348663

https://hal.science/hal-03348663v1

Submitted on 19 Sep 2021

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.

Validation of Smets' hypothesis in the crowdsourcing environment*

Constance Thierry, Arnaud Martin, Jean-Christophe Dubois, and Yolande Le Gall

Univ. Rennes, CNRS, IRISA, DRUID, France
name.surname@irisa.fr
http://www-druid.irisa.fr

Abstract. In the late 1990s, Philippe Smets hypothesizes that the more imprecise humans are, the more certain they are. The modeling of human responses by belief functions has been little discussed. In this context, it is essential to validate the hypothesis of Ph. Smets. This paper focuses on the experimental validation of this hypothesis in the context of crowdsourcing. Crowdsourcing is the outsourcing of tasks to users of dedicated platforms. Two crowdsourcing campaigns have been carried out. For the first one, the user could be imprecise in his answer, for the second one he had to be precise. For both experiments, the user had to indicate his certainty in his answer. The results show that by being imprecise, users are more certain of their answers.

Keywords: Uncertainty · Imprecision · Belief functions · Crowdsourcing.

1 Introduction

The theory of belief functions is well known for information fusion and for catching uncertainty and imprecision in machine learning, tracking and data association. A simple mass function allows to represent uncertainty, imprecision and ignorance at the same time. Of course, when asking someone a question, he or she does not answer with a belief function.

Few works consider the modeling of uncertain and imprecise responses to questionnaires. Some works have focused on questionnaires allowing probabilistic [2, 13, 7, 9] or fuzzy [12] answers, but very few allow belief answers. Diaz *et al.* [5] have proposed a questionnaire to directly build a mass function, but it remains very unintuitive for any user. Other works [1, 3] have considered directly generated mass functions without worrying about their construction.

In this work, we are interested in modeling the responses of users of a crowdsourcing platform, allowing them to express their uncertainty and imprecision. These responses will then be modeled by mass functions. However, it is necessary

* This work is supported by the general council of the Côtes d'Armor and the ANR project HEADWORK

to understand the links between uncertain and imprecise information provided by humans.

Ph. Smets [11] presents in his paper different types of data imperfection and methods to model them. In particular, he presents imprecision as an element relative to an assertion, and uncertainty as the relationship between the information provided by the assertion and the knowledge that the human has of the subject. Ph. Smets then proposes that the more imprecise a person is, the more certain he is, and conversely the more precise, the less certain. This assertion is decomposed here into two hypotheses:

- H1: The more imprecise a person is, the more certain he is.
- H2: The more precise a person is, the less certain he is.

According to Dubois *et al.* [6], there are two types of uncertainty related to human perception. The first one on the realization of an action presenting a risk, the second on the truth of an assertion due to a lack of knowledge. In this article, uncertainty is the consequence of a lack of knowledge that does not allow man to define the truth of an assertion.

H1 and H2 are not the same hypotheses, H2 is the reciprocal of H1. Since the reciprocal of a hypothesis is not always true we work in this paper on the independent validation of H1 and H2. To our knowledge, there is no work to validate these assumptions. In order to perform the experimental validation of the two hypotheses, two crowdsourcing campaigns have been carried out. Crowdsourcing consists in outsourcing tasks on dedicated platforms. The users of the platform perform the tasks for a micro payment. The tasks are very diverse, but generally do not require expertise, so the user profiles are very varied. In this study, the user's task consists of photo annotation through multiple choice questionnaires (MCQs). In the first crowdsourcing campaign the user has to give a precise answer, in the second one he can be imprecise and choose several answers of the MCQ. For both campaigns, the certainty of the user in his answer is required.

The plan of the paper is as follows. Section 2 reviews the theory of belief functions used for data modeling. Section 3 presents the defined crowdsourcing campaigns, the results obtained for the validation of H1 and H2 and the answer modeling. Section 4 concludes the paper.

2 Theory of belief functions

The theory of belief functions, introduced by Dempster [4] and formalized by Shafer [10], models the imprecision and uncertainty of imperfect sources. The user u of a platform is a source of information. Considering a question q asked to this user u, the finite set of possible answers to q composes the frame of discernment Ω. A mass function $m_{uq}^{\Omega} : 2^{\Omega} \rightarrow [0, 1]$ is defined such that:

$$\sum_{X \in 2^{\Omega}} m_{uq}^{\Omega}(X) = 1,$$

with 2^{Ω} the set of the disjunctions of Ω.

Let $X \in 2^{\Omega}$, the mass $m_{uq}^{\Omega}(X)$ characterizes the belief of the user u in the answer X to the question q. When $m_{uq}^{\Omega}(X) > 0$, X is called the focal element. A function $m_{uq}^{\Omega}(X) = 1$, $X \in 2^{\Omega}$, is a categorical mass function, the user is absolutely certain of this answer which may be imprecise if X is an union of elements of Ω. The set $\Omega \in 2^{\Omega}$ symbolizes ignorance, if $m_{uq}^{\Omega}(\Omega) = 1$ then the user is totally unaware of what the right answer is. The element $\emptyset \in 2^{\Omega}$, under the open world hypothesis, symbolizes a value outside Ω, in the case of normalized mass functions $m_{uq}^{\Omega}(\emptyset) = 0$.

The response X of u, can be modeled by a simple mass function (X^w):

$$\begin{cases} m_{uq}^{\Omega}(X) = \omega \text{ with } X \in 2^{\Omega} \setminus \Omega \\ m_{uq}^{\Omega}(\Omega) = 1 - \omega \end{cases} \tag{1}$$

This mass function allows to model: the uncertainty ω on the answer, the imprecision of u by the cardinality of X, and the remaining ignorance on the answer. That is the simplest way to model uncertainty and imprecision of the answer of the users. Consonant mass functions that can model different levels of imprecision, have all the focal elements nested. A consonant mass function is a possibility distribution.

In case of doubt about the reliability of a source, a weakening coefficient $\alpha \in [0, 1]$ modeling this reliability can be introduced:

$$\begin{array}{l} m_{uq}^{\Omega,\alpha}(X) = \alpha * m_{uq}^{\Omega}(X), \forall X \in 2^{\Omega} \setminus \Omega \\ m_{uq}^{\Omega,\alpha}(\Omega) = 1 - \alpha * (1 - m_{uq}^{\Omega}(\Omega)) \end{array} \tag{2}$$

If the user u is absolutely unreliable then $\alpha = 0$ and the whole mass is assigned to Ω concluding to total ignorance. Weakening allows to reduce the conflicts occurring during the combination.

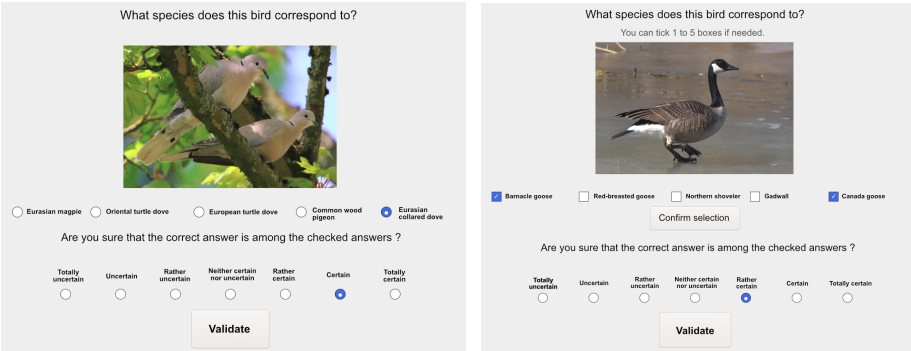
The main goal of crowdsourcing platforms is to fuse the answers of the users and decide the best answer to the questions. In the theory of belief function, many combination [8] operators exist. But the most used is the conjunctive rule of combination, given by equation (3), which requires that the sources be reliable, distinct and independent.

$$m_{Conj}^{\Omega}(X) = \sum_{Y_1 \bigcap \ldots \bigcap Y_N = X} \prod_{u=1}^{N} m_u^{\Omega}(Y_u) \tag{3}$$

This operator reduces the imprecision on the focal elements and increases the belief on the concordant ones. It can generate a non-zero mass on the empty set, so, in order to stay in a closed world, the normalized Dempster operator is preferred:

$$m_D^{\Omega}(X) = \frac{1}{1-k} m_{Conj}^{\Omega}(X) \tag{4}$$

with $k = m_{Conj}^{\Omega}(\emptyset)$ the global conflict from the sum of the partial conflicts. Once the combination of information has been achieved, it is necessary to return to a probabilistic framework to make a decision. To do this, it is possible to calculate the pignistic probability [1, 3] on the elements of Ω.



(a) Experiment 1 (b) Experiment 2

Fig. 1: Interfaces used for crowdsourcing campaigns

3 Experimental validation of hypotheses H1 and H2

This section presents the crowdsourcing campaigns conducted for data collection. Data is then analyzed to validate hypotheses H1 and H2, and the responses are modeled by the theory of belief functions.

3.1 Crowdsourcing campaigns

Two crowdsourcing campaigns on the annotation of 50 bird photos have been conducted. For both experiments, and each question, a photo is presented to the user with five bird names as possible answers. The photos to be annotated are the same for both campaigns, as well as the corresponding names in response. All the proposed names are real and the 50 photos are only birds which can be met in natural place in Metropolitan France.

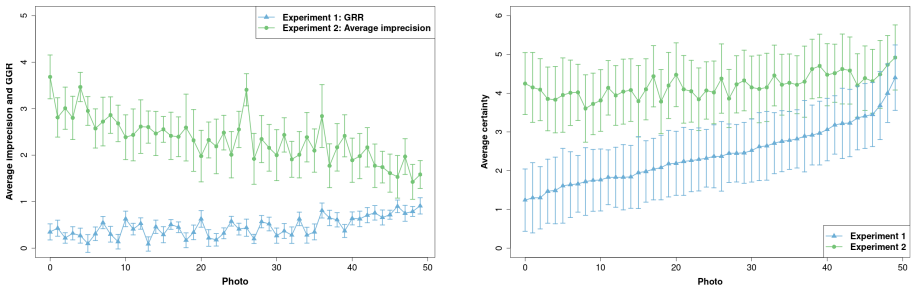
The interfaces of the campaigns are given in Figure 1. For both campaigns, users specify their degree of certainty in their answer. The proposed degrees of certainty are summarized in Table 1. The user is notified that it is not penalizing to be uncertain in his answers. In the first campaign, corresponding to Experiment 1 in figure 1, the user is forced to choose a single answer by checking a radio button. In the second campaign, corresponding to Experiment 2 on figure 1, the user can be imprecise if he feels the need, by selecting 1 to 5 bird names by checking a checkbox button. At the beginning of this campaign, the user is informed that there is no penalty for selecting multiple names.

The campaigns were both carried out by 100 different users on the Crowdpanel platform[1], and a user who did the first campaign cannot do the second one. This makes a total of 50 photos $\times$ 100 users = 5000 data for each crowdsourcing campaign. The following section analyzes the collected data in order to validate Ph. Smets' hypothesis.

[1] https://crowdpanel.io/

Table 1: Numerical values associated to the certainty scale proposed to the user

Totally uncertain	Uncertain	Rather uncertain	Neither certain nor uncertain	Rather certain	Certain	Totaly certain
0	1	2	3	4	5	6



(a) GRR (Experiment 1) and average imprecision (Experiment 2) per photo.

(b) Average certainty per photo for both experiments.

Fig. 2: Comparison of experiments per photo - GRR, imprecision and certainty.

3.2 Analysis of the results

In order to validate H1 and H2, we recall that we conducted two experiments: in Experiment 1, the user is required to be precise, whereas in Experiment 2 the user can be imprecise. For both experiments, we associate to each answer of a user u to a question q, a numerical certainty value c_{uq}. The values of c_{uq} are taken in the interval $[0, 6]$, with 0 corresponding to total user uncertainty and 6 to total certainty, as shown in Table 1. Moreover, for Experiment 2, we associate an imprecision degree i_{uq} to each answer, that takes its values in the interval $[1, 5]$ according to the number of selected bird names.

Task difficulty. Certainty and imprecision of the user's answer may depend on the difficulty of the question according to the proposed bird names. In both crowdsourcing campaigns, the questions are of varying difficulty relative to the user's knowledge of the domain. The more difficult the question is, the more difficult it will be for the contributor to answer it. In order to evaluate the difficulty of each question for Experiment 1, one can calculate a good recognition rate (GRR), given by the average of the photos correctly annotated by the users. To compare the perceived difficulty of the questions by the users of the two campaigns, the GRR of Experiment 1 and the average imprecision of Experiment 2 are calculated for each question with 95% confidence intervals and presented in Figure 2a. On Figure 2b, the average user certainty for each question c_q is presented with a 95% confidence intervals. On both figures, bird photos are ordered according to the average certainty of the Experiment 1. The value c_q is

increasing for Experiment 1 confirming a variable difficulty between questions. Comparing the two blue curves, we note a link between difficulty and certainty. Experiment 1 users have a GRR between 9% and 91%. The higher the GRR, the simpler the question.

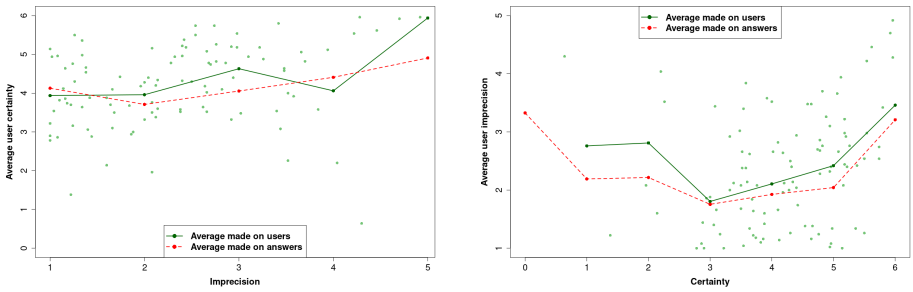
Users who participated in Experiment 2 made good use of the opportunity to be imprecise since the average imprecision varies between 1.4 and 3.7 on Figure 2a (with a minimum of 1 response to a maximum of 5 and an average of about 2 bird names selected). The more imprecise the user is, the more difficult the question is.

Figure 2a shows that as GRR increases, the average imprecision of users decreases, which means that users in both campaigns had difficulties with the same questions. The users of both experiments therefore have varying levels of knowledge about birds, which is usual in crowdsourcing platforms where the profiles are diverse. Since in both experiments users have varying degrees of knowledge, user confidence in their response will also vary.

Comparison of certainty between precise and imprecise answers. In Figure 2b, the average user certainty is higher for users who had the opportunity to be imprecise in Experiment 2 compared to Experiment 1. The difference between the two curves is between 8.68% and 50.13% and on average 29.68% which makes a significant increase in certainty. The value c_q varies between 3.6 and 4.9 for Experiment 2, which is a certainty gap of 1.3, and between 1.2 and 4.4 for Experiment 1, making a difference of 3.2. The difference between the average certainty values is smaller for Experiment 2. Users are therefore on average always positively certain ($c_q > 3$) of their answers for Experiment 2 contrary to the Experiment 1 where they sometimes uncertain. By having the possibility to be imprecise users are more certain of their answer and this certainty is more constant according to the difficulty. The users to Experiment 1 who were required to give a specific answer are less certain, which confirms hypothesis H2.

Analysis of the use of certainty and imprecision for Experiment 2 users. In order to understand the relations between certainty and imprecision of the users of Experiment 2, we plot the average user certainty as a function of user imprecision in Figure 3a, and on the contrary in Figure 3b, we plot the average user imprecision as a function of user certainty. In these figures a point represents a user positioned according to c_u and i_u.

In Figure 3a, the values i_u are discretized and the average of c_u is realized for the degrees of imprecision 1 to 5. The red curve is a reference value, it presents the averages, on the answers, of the c_{uq} values for the degrees of imprecision 1 to 5. The average certainty made on the answers is increasing for an imprecision degree ranging from 2 to 5 selected answers. For precise answers, with only one name chosen, the certainty is on average higher than for an imprecision degree of 2, but remains lower or equal to the imprecision degrees 3 to 5. This slightly higher certainty value for accurate responses is explained by the fact that some users in the crowd have more knowledge about birds. These qualified



(a) Average certainty as a function of the degree of imprecision.

(b) Average imprecision as a function of the certainty.

Fig. 3: Relation between certainty and imprecision for Experiment 2.

users manage to give accurate answers, while being certain. For the user curve, the certainty increases with the imprecision, except for a degree of imprecision of 4 where the certainty is lower than the reference value. This is due to three users who have a very low certainty c_u while they are on average very imprecise.

As in Figure 3a, the c_u values are discretized and the averages of i_c is performed for each certainty value to obtain the green curve Figure 3b. The red curve is a reference calculated by averaging the imprecision i_{uq} of the answers for the certainty values in Table 1. For positive certainty values, the average imprecision of the data is increasing. Symmetrically, the average imprecision is also increasing for negative certainty values, whereas one would expect the imprecision to decrease. Nevertheless, the certainty values 0 to 2 are represented by only 16% of the answers, so the variations in this part of the graph are less relevant than for the certainty values 3 to 6. Likewise, for the majority of users $c_u \geq 3$, for these values, the average imprecision of users increases with the certainty.

The obtained results show that allowing the user to be imprecise makes him more certain of his answers and requiring him to be precise less certain. It is notably observed that the more imprecise the user is, the more certain he is, which validates hypothesis H1. When the user chooses a single answer, if he has domain knowledge, he is certain, but if he does not have domain knowledge his answer will be uncertain. Therefore, in the absence of a qualification for the task, the more precise the user is the less certain he is, which validates H2. On the basis of the validated hypotheses H1 and H2, one can model the imprecision and uncertainty of the answers by belief functions.

3.3 Modeling and aggregation of responses

Traditionally, certainty of user's answer is not required in crowdsourcing platforms. Also, it is generally not possible for the user to be imprecise. The answers are aggregated by majority voting, which consist of selecting the answer given by the majority of the crowd. However, as it has been shown in the section above,

when they can be imprecise, users are more certain. It is therefore interesting for the employer to introduce the notions of imprecision and certainty in the tasks in order to improve the quality of the answers obtained by crowdsourcing campaigns. Indeed, in crowdsourcing campaigns, sometimes the user is in a situation of indecision but has to give a precise answer while hesitating between different choices. The user is required to select an answer among all his hesitations. This answer is then not very certain or even random which is not desirable for the employer. On the contrary, by offering the user the possibility to select the set of choices that he considers correct, the certainty of the latter is higher. The collected responses can be modeled and aggregated using the theory of belief functions that takes into account the imprecision and the certainty of the answer.

For the conducted experiments, the user associates to his answer X to the question q a certainty value c_{uq}. In the case where the answer is imprecise (such as in Experiment 2), the imprecision correspond to the selected bird names. We therefore propose to model the answers thanks to belief functions by a simple mass function $X^{\omega_{uq}}$ with $\omega_{uq} = \frac{c_{uq}}{c_{MAX}}$, in this paper $c_{MAX} = 6$ and with X the subset of selected bird names. Hence, $|X|$ is the number of selected bird names and is equal to i_{cq}. With this mass function, the more certain the user is of his answer, the higher the value of ω_{uq}. The mass function is then weakened by a coefficient $\alpha = 0.8$ and aggregated by questions by the Dempster conjunctive combination operator. Then, the decision on the answer is made by the pignistic probability on the answer.

After the decision phase, a correct answer rate of 84% is obtained for Experiment 1, while for Experiment 2, this value is 90%. This 6% increase in the correct answer rate between the two experiments is interesting for the employer because it shows an improvement in the quality of the data collected. Moreover, when the data is aggregated by majority voting, the correct answer rate is 70% for Experiment 1 and 83% for Experiment 2. Even with majority voting it is interesting to allow the contributor to be imprecise in case of indecision. The modeling and aggregation of the answers is even more interesting with belief functions, because it offers better results than the majority vote. By allowing users to be imprecise, the employer can be more confident that the data collected is of good quality.

4 Conclusion

Ph. Smets hypothesizes that the more imprecise one is, the more certain one is and reciprocally, the more certain one is, the less precise one is. We carried out two crowdsourcing campaigns, one where the user is required to be precise (Experiment 1) and the other where he could be imprecise (Experiment 2). For both campaigns, the certainty of the contributor in his answer is required.

The users in Experiment 2 made good use of the opportunity to be imprecise. The analysis of the collected data shows that the average certainty per question is quite stable and higher for users who were able to be imprecise compared to those

who had to be precise. In addition, when we plot the average of users' certainty according to the average of users' imprecision, we find that certainty increases with imprecision. The experimental analysis thus validates the hypothesis of Ph. Smets.

In order to model the uncertainty and imprecision of responses, the theory of belief functions is used. Currently we use simple support mass functions to represent the user's response. A mass equivalent to the user's certainty is associated to his answer. This gives more weight to answers that users are certain of when aggregating the data. Modeling and aggregating responses using belief function theory offers better results than the majority voting commonly used in crowdsourcing platforms.

In future work we would like to offer the contributor the possibility to select several sets of answers with different degrees of certainty in order to consider more than two focal elements. Consistent mass functions can then be used to model the responses.

References

1. Abassi, L., Boukhris, I.: A worker clustering-based approach of label aggregation under the belief function theory. Applied Intelligence pp. 1–10 (2018)
2. Ambroise, C., Denœux, T., Govaert, G., Smets, P.: Learning from an imprecise teacher: probabilistic and evidential approaches. Applied Stochastic Models and Data Analysis **1**, 100–105 (2001)
3. Ben Rjab, A., Kharoune, M., Miklos, Z., Martin, A.: Characterization of experts in crowdsourcing platforms. Belief Functions: Theory and Applications **9861** (2016)
4. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. The Annals of Mathematical Statistics **38**, 325–339 (1967)
5. Diaz, J., Rifqi, M., Bouchon-Meunier, B., Jhean-Larose, S., Denhiére, G.: Imperfect answers in multiple choice questionnaires. In: European Conference on Technology Enhanced Learning. pp. 144–154. Springer (2008)
6. Dubois, D., Prade, H., Smets, P.: Representing partial ignorance. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans **26**(3), 361–377 (May 1996)
7. Ipeirotis, P.G., Provost, F., Wang, J.: Quality management on amazon mechanical turk. KDD-HCOMP'10 (2010)
8. Martin, A.: Conflict management in information fusion with belief functions. In: Bossé, E., Rogova, G.L. (eds.) Information quality in information fusion and decision making, pp. 79–97. Information Fusion and Data Science (2019)
9. Raykar, V.C., Yu, S.: Annotation models for crowdsourced ordinal data. Journal of Machine Learning Research (2012)
10. Shafer, G.: A mathematical theory of evidence. Princeton university press (1976)
11. Smets, P.: Imperfect information: Imprecision and uncertainty. In: Motro, A., Smets, P. (eds.) Uncertainty Management in Information Systems: From Needs to Solutions. pp. 225–254. Springer US, Boston, MA (1997)
12. Wagner, C., Anderson, D.T.: Extracting meta-measures from data for fuzzy aggregation of crowd sourced information. In: 2012 IEEE International Conference on Fuzzy Systems. pp. 1–8. IEEE (2012)
13. Wang, J., Ipeirotis, P.G., Provost, F.: Managing crowdsourcing workers. In: The 2011 winter conference on business intelligence. pp. 10–12 (2011)