Elicitation, assessment and pooling of expert judgements using possibility theory

S.A. Sandri, D. Dubois**, H.W. Kalfsbeek***[1]*

* Brazilian National Institute for Space Research (INPE), Caixa Postal 515, 12200 São José dos Campos, SP, Brazil

** Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, France

*** Joint Research Center of the CEU, Ispra, Italy

Abstract

The problem of modelling expert knowledge about numerical parameters in the field of reliability is reconsidered in the framework of possibility theory. Usually expert opinions about quantities such as failure rates are modelled, assessed and pooled in the setting of probability theory. This approach does not seem to always be natural since probabilistic information looks too rich to be currently supplied by individuals. Indeed, information supplied by individuals is often incomplete, imprecise rather than tainted with randomness. Moreover the probabilistic framework looks somewhat restrictive to express the variety of possible pooling modes.

In this paper, we formulate a model of expert opinion by means of possibility distributions that are thought to better reflect the imprecision pervading expert judgements. They are weak substitutes to unreachable subjective probabilities. Assessment evaluation is carried out in terms of calibration and level of precision, respectively measured by membership grades and fuzzy cardinality indices. Lastly, drawing from previous works on data fusion using possibility theory, we present various pooling modes with their formal model under various assumptions concerning the experts. A comparative experiment between two computerized systems for expert opinion analysis has been carried out, and its results are presented in this paper.

[1] Dr. Kalfsbeek is currently with the Directorate General for Environment, Nuclear Safety and Civil Protection, Commission of the European Union, Brussels.

1 - Introduction

The use of information originating from human experts in the field of reliability and safety analysis of newly designed installations or regarding processes on which no experimental observations are possible becomes more and more accepted by the scientific community. In particular, it is clear that failure rates and failure probabilities of equipment for which no operating experience is available, as well as probabilities of occurrence of rare events and of unexplored physical/chemical processes, are the subject matter of expert judgement, simply because there exist no other data sources. The typical piece of information that has to be assessed in such situations is the number of times that a certain type of event occurs within a certain time span or during a certain number of trials, or the number of hours it takes to repair equipment or to restore faulty situations, etc. Several methods for elicitation, assessment and pooling of this type of information have been proposed and are being applied, including classical, Bayesian and psychologic scaling approaches [4]. Nonetheless, the area is still developing, in particular with respect to the assessment of the experts and the way in which information originating from different experts and/or "objective" data sources can be combined in order to arrive at the best possible result (pooling of expert judgement). The present paper addresses both issues, basing its approach on possibility and fuzzy set theory for modelling the uncertainty [11].

The uncertainty model plays a central role in the use of expert judgements, because no human being would claim that he is absolutely sure about his judgements or advice. Hence, it is necessary to incorporate into any model the individual expert's uncertainty about his advice, the decision maker's uncertainty about the quality of the expert(s), and how these two kinds of uncertainty interact and impact on the credibility of the final results. The classical and Bayesian approaches use the concept of probability for modelling the uncertainty (respectively, the "frequentist" and "subjectivist" way of looking at probability), whereas we take possibility/necessity as the basic framework. The main reason for adopting such a framework is that possibility theory offers a simple theory of uncertainty that explicitly takes into account the lack of precision of the expert knowledge. As such, the possibilistic framework is weaker than the probabilistic one because partial ignorance may be represented in an unbiased way, including even the extreme case of complete ignorance which occurs when the expert cannot supply informative data. A probability distribution never accounts for a lack of precision in the data, and so the possibilistic model may turn out to be more faithful to the available data supplied by experts.

In order to get useful information from the experts, several problems must be solved. The first one is a proper modeling of the pieces of data supplied by a single expert about a given parameter. This type of data is almost never precise and reliable because the expert possesses only a rough idea of the value of quantitative parameters, due to the limited precision of human assessments and to the variability of such values (e.g., failure rates). In most studies the expert's response is represented by a probability distribution because probability theory is often considered to be the only well-established framework for modelling uncertainty. In this paper we argue that a pure probabilistic model of expert knowledge is not so satisfactory, and that possibility theory is a more natural framework.

The second task to be solved is the assessment of the quality of the expert, namely his calibration and the precision of his response. In the case of probabilistic modeling, scoring rules have been devised for this purpose (see Cooke [4] for a survey). We shall suggest a rating index which may act as a scoring rule for the case of possibilistic modeling. Lastly when several expert responses are available, they must be combined so as to yield a unique, hopefully better response. This problem is also addressed here.

In this paper we compare the possibilistic approach to expert judgement with other approaches found in the literature. The next section addresses the representation and elicitation of expert knowledge. Section 3 describes numerical indices for expert response assessment. Section 4 deals with the pooling of uncertain pieces of information. Section 5 describes two operational expert judgement analysis systems (the probabilistic system EXCALIBR [4] and the possibilistic system PEAPS), and outlines an experiment that aims at comparing these systems in relation to the assessment and pooling of expert opinion. Section 6 describes the use of this experiment on real world data, and Section 7 contains conclusion remarks. This paper is based on a study whose preliminary results appear in Kalfsbeek [22], Dubois and Kalfsbeek [11], Sandri et al. [27] and Sandri [32, 28].

2 - Expert Knowledge Elicitation

2.1 - The Probabilistic Approach

In the probabilistic approach, experts are typically asked about the numerical value of some parameter v (typically, the failure rate of a device) by specifying quantiles of a probability distribution function (pdf) on an interval containing v. Let v be a random variable on an interval X. The smallest number $x \in X$, such that $P(v \leq x) = k/100$, is called the k% quantile, and denoted $q_{k\%}$. In this approach the experts are often asked to supply the 5%, 50% and the 95% quantiles. In other words, an expert supplies x' and x" such that $P(v \leq x') = 0.05$ and $P(v \leq x") = 0.95$ respectively. Beside these quantiles, some information about the mode, the mean, or the median of the distribution is often requested. Based on these values, and on the choice of a parameterized family of distribution functions, (for instance a beta-distribution), a given distribution function is choosen that supposedly best represents the available information. Experts may also be asked to choose between two alternatives in a series of experiments, and a probability distribution is derived based upon their choices (this elicitation method is called dichotomic in [4]). Note that in some approaches, experts are only asked for point values of x, as in some Bayesian methods [26, 35].

The meaning of such degrees of probability is not so obvious: they can be understood as pure subjective quantities assessing degrees of belief in the various possible values of the parameter. They can also be construed as subjective estimations of frequencies, the probability distribution then represents the frequency distribution of the parameter over a class of similar devices. As a consequence, the nature of the probabilities appearing during the elicitation is sometimes controversial. Some authors call them "subjective probabilities", the term being ambiguous insofar as it may mean the numerical estimate of a feeling of certainty or of a subjectively assessed objective frequency. In reliability applications the second interpretation would appear to be more relevant.

2.2 - Expert Knowledge as Possibility Distributions

The choice of a particular probability distribution for the modelling of an expert opinion may appear debatable because, in the process of fitting a given parametrized distribution to a pair of quantiles and a median, the imprecision pervading the data is lost for the sake of computational convenience. There are indeed many probability distributions that have prescribed 5% and the 95% quantiles and, say, mode. Faithfulness to the expert's opinion, therefore, dictates incorporation of sets of probability distributions into the model; this approach, however, would be very cumbersome even if exact. Another idea is to use an approximate representation of the data which captures both uncertainty and imprecision. The simplest model of a family of probability distributions is offered by possibility theory [13], and this paper investigates how far we can go with this simple model.

A possibility distribution π_v [39] attached to parameter v can be viewed as the membership function of the fuzzy set of possible values of a variable v. The possible values as described by π_v are assumed to be mutually exclusive, since v takes on only one value (its true value) from a set X taken here to be a closed, bounded real interval $[x_l,x_u]$. Moreover, since one of the elements of X is the true value of v, $\pi_v(x) = 1$ for at least one value $x \in X$. Possibility distributions can be rigorously related to probability distributions, in which case $\pi_v(x)$ is taken to be an upper probability bound [13].

The simplest form of a possibility distribution on X is the characteristic function of a sub-interval $[s_l,s_u]$ of X, i.e., $\pi_v(x) = 1$ if $x \in [s_l,s_u]$, 0 otherwise. This type of possibility distribution results when experts claim that "v lies between s_l and s_u". Note that $\pi_v(x) = 1$ has a weaker meaning than in probability theory, it only means that x is a completely *possible* value for v. This way of expressing knowledge is more natural than giving a point-value, say x*, for v right away, because it allows for some imprecision: the true value of v is more likely to lie between s_l and s_u than to be equal to x*. Clearly, allowing for imprecision reduces the uncertainty of the assessment. Indeed imprecise statements are always safer than precise ones.

However, this representation is not entirely satisfactory. Namely, claiming that $\pi_v(x) = 0$ for some x means that v = x is impossible, a very strong statement. This is too strong for the expert who is then tempted to give wide, uninformative intervals (e.g., $s_l = x_l$, $s_u = x_u$). It is more satisfactory, in this connection, to obtain from the expert several nested intervals with various levels of confidence, and to admit that even the widest, safest intervals contain some residual uncertainty, here denoted ε. These nested intervals will lead to membership functions of fuzzy intervals [9, 11].

A fuzzy interval can be viewed as a finite set of nested ("focal") subsets $\{A_1, A_2,\ldots, A_m\}$ as long as the set of possibility values $\{\pi_v(x) \mid x \in X\}$ is finite. In this case, there is a set of weights $p_1, p_2,\ldots, p_m$ summing to one, such that [11]

$$\forall x,\ \pi_v(x) = \Sigma_{x\in A_i}\, p_i \tag{2.1}$$

Namely it can be proved that if the set of possibility values is $\{\alpha_1 = 1, \alpha_2 \geq \alpha_3 \geq\ldots \geq \alpha_m\}$, and letting $\alpha_{m+1} = 0$ we have

$$A_i = \{x \mid \pi_v(x) \geq \alpha_i\}\ \ ;\ \ p_i = \alpha_i - \alpha_{i+1},\ 1 \leq i \leq m.$$

From a mathematical point of view, this definition extends to the infinite case, changing the set of weights p_i into a pdf on the unit interval and turning (2.1) into an integral (see, e.g., [9]). This view of possibility distributions corresponds to the natural embedding of possibility theory in belief function theory [30] and hence in the framework of upper and lower probabilities of which belief function theory is formally a special case.

Knowing a possibility distribution, the likelihood of events can be described by means of two set-functions: the possibility measure ($\prod$) and the necessity measure (N) [11]. When π is the membership function of a crisp set A given as the evidence, an event B is said to be possible if and only if $A \cap B \neq \emptyset$, and certain if and only if $A \subseteq B$; by definition we let $\prod(B) = 1$ and $N(B) = 1$ in these respective situations. In the general case where π_v is the membership function of a fuzzy set, the possibility and necessity measures are defined as follows. Letting Π_i and N_i be the {0,1}-valued possibility and necessity measure induced by the set A_i, we define:

$$\Pi(B) = \Sigma_{i=1,m}\, p_i\, \Pi_i(B) = \sup_{x\in A} \pi_v(x) \qquad (2.2)$$

$$N(B) = \Sigma_{i=1,m}\, p_i\, N_i(B) = \inf_{x\notin A} 1 - \pi_v(x) \qquad (2.3)$$

$$= 1 - \Pi(\bar{B})$$

where $\bar{B}$ is the complement of B with respect to X. This duality expresses the fact that B tends toward certainty as $\bar{B}$ tends toward impossibility. The above expressions emphasize the fact that possibility and necessity degrees are special cases of plausibility and belief degrees in the sense of Shafer (see also [40], for instance).

The expert is supposed to be capable of supplying several intervals $A_1, \ldots, A_m$ directly, corresponding to prescribed levels of confidence $\lambda_1,\ldots, \lambda_m$. The level of confidence λ_1 can be conveniently interpreted as the smallest probability that the true value of v hits A_i (e.g., from the point of view of the experts, the proportion of cases where $v \in A_i$ from his experience). The interval A_i is the smallest one whose probability of being hit is at least λ_i. In practice, only three intervals have been kept: A_1 with $\lambda_1 = 0.05$, A_2 with $\lambda_2 = 0.5$, and A_3 with $\lambda_3 = 0.95$. A_1 corresponds to the "usual values" of v, and $A_3 = [s_l,s_u]$ corresponds to the interval which leaves a 0.05 probability (= ε) that v misses A_3, i.e., the residual uncertainty of the conservative evaluation. The links between the λ_i's and the degrees of possibility are defined by $\lambda_i = 1 - \alpha_{i+1}$ for i = 1,m, i.e., the degree of possibility α_{i+1} is related to the degree of certainty (λ_i) that x lies in A_i; this degree of certainty being interpreted as a lower bound on the probability $P(A_i)$; in the terminology of possibility theory, $\lambda_i = N(A_i)$ the degree of necessity of A_i [11]. Finally, the focal subset $A_m = A_4$ is always X itself, due to the residual uncertainty. The following Table 2.1 summarizes the data supplied by one expert.

From a mathematical point of view, the information can be viewed as a nested random set $\{(A_i,p_i), i = 1,m\}$, which allows for imprecision (the size of the A_i's) and uncertainty (the p_i's). The intuitive meaning of p_i (= $\lambda_i - \lambda_{i-1}$) is the probability that the interval A_i represents the actual knowledge of the expert and not the probability that x hits A_i. The latter probability is lower-bounded by λ_i and gathers p_j's for all A_j that contain A_i. The first three lines of Table 2.1 correspond to specific questions asked to experts (see [22]). Although intervals $[c_l,c_u]$, $[m_l,m_u]$, $[s_l,s_u]$ are not used in the probabilistic approaches, these intervals can be interpreted in terms of quantiles of a probability distribution, e.g., $[s_l,s_u]$ corresponds to the range between the 2.5% and the 97.5% quantiles. In terms of fuzzy sets, $[c_l,c_u]$ corresponds to the core of the fuzzy set since $\forall\, x \in [c_l,c_u]$, $\pi_v(x) = 1$. The obtained possibility distribution is pictured in Figure 2.1.

A_1	$[c_l, c_u]$	0.05	1	0.05
A_2	$[m_l, m_u]$	0.5	0.95	0.45
A_3	$[s_l, s_u]$	0.95	0.5	0.45
A_4	X	1	0.05	0.05
	selected intervals	levels of confidence λ_i	degrees of possibility α_i	weights p_i

Table 2.1: Data supplied by experts (s_l, s_u, m_l, m_u, c_l, c_u)
(in the bold-face rectangle)

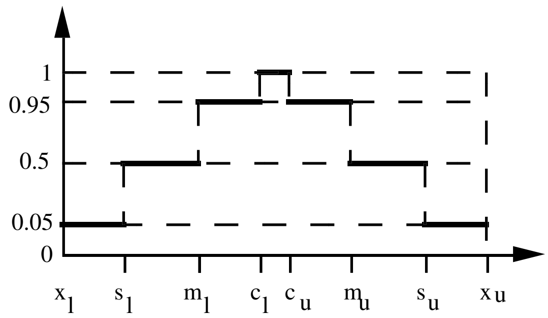

Figure 2.1: Expert-originated possibility distribution

Note that by adopting a possibilistic model of the expert opinion we are not rejecting probability theory as an underlying framework, rather, we enlarge it so as to leave room for the representation of imprecision in the supplied data. The nestedness property of the supplied intervals presupposes that the expert, although having imprecise knowledge, give coherent answers to the various questions.

3 - Evaluation of Expert Judgement

Once the possibility distributions of the uncertain variables under consideration have been determined, the next question is how "good" this information is before embarking on the final processing step, i.e., the pooling of the pieces of information obtained from several experts. In order to build a meaningful rating system, one must first identify the type of deficiencies experts may be prone to, and then define indices that enable the true answer and the expert answer to be compared so as to take these deficiencies into account. Experts can be deficient with regard to three aspects:

- inaccuracy: values given by the expert are inconsistent with the real values of the parameters, for instance, underestimated. The expert is then said to be miscalibrated.
- imprecision: the expert, although not miscalibrated, is too cautious. So, the intervals he supplies are too large to be informative. Such an expert is said to be underconfident.
- exagerated precision: the value of the parameter is not precisely known but the expert supplies intervals that are too narrow (or even point-values). Such an expert is said to be overconfident.

Most expert judgement systems try to detect and treat these deficiencies. In this case two basic approaches are used; either the analyst knows the experts' deficiencies and is able to furnish coefficients which will modify the experts' estimates, or the experts are submitted to a battery of tests. In these tests, the experts are asked questions whose answers are known and are rated on the basis of the results. The questions pertain to the true values of a series $v_1, v_2, \ldots, v_n$ of "seed" variables; the values of these parameters are either known by the analyst and not known by the experts, or more often can be determined afterwards by means of physical experiments or other means. It is considered here, contrary to usual practice, that the true value of a seed variable may not be precisely known, either because the state-of-the-art in the field does not allow for an exact evaluation or because the available information is available only in histogram form. In the following we discuss how the deficiencies cited above are treated in the probabilistic and possibilistic frameworks.

3.1 - The Probabilistic Approach

Cooke [4] suggests that the quality of an expert can be determined by the product of his *informativeness* and his *calibration*. The proposed informativeness measure is Shannon's entropy, and the calibration measure is a particular implementation of the original concept of calibration ex-ante given by Winkler [34]:

> "... If I expect rain to occur on 10 percent of the days for which the probability is 0.10, on 20 percent of the days for which the probability is 0.20, and so on, then I view the forecaster as perfectly calibrated ex-ante."

In the following we discuss the evaluation in this paradigm for the case where a variable domain has more than two elements (called the non-dichotomic case).

Let X be a discrete domain with B + 1 possibilities and let p = $\{p_1,..., p_B, p_{B+1})$ and r = $\{r_1,..., r_B, r_{B+1})$ be probability distributions on X. The relative information between the distributions r and p is defined as follows:

$$I(r,p) = \sum_{j=1}^{B+1} r_j * \ln\left(\frac{r_j}{p_j}\right) \tag{3.1}$$

I(r,p) evaluates a distance between r and p. It is the index commonly taken to measure the amount of information learned if one initially believes that p is correct and subsequently learns that r is correct; it goes to 0 only when r = p. The counterpart of (3.1) in a continuous domain $X = [x_l,x_u]$ is:

$$I(r,p) = \int_{x_l}^{x_u} r(v) \ln\left(\frac{r(v)}{p(v)}\right) dv. \tag{3.2}$$

If p is a uniform distribution u on Ω then (3.2) becomes

$$I(r,u) = \ln(x_u - x_l) + \int_{x_l}^{x_u} r(v) \ln r(v)\, dv \tag{3.3}$$

The informativeness of an expert e_i with respect to a seed variable v_j taking its values in X_j, is calculated as follows. First of all, the expert is asked to supply a fixed number of quantiles, representing his knowledge about the true value of variable v_j. A probability distribution $r_j = p(e_i,v_j)$ is then derived from these quantiles, using the hypothesis that the distribution of probability in each inter-quantile interval is uniform. Distribution r_j is the least informative probability distribution that is compatible with the quantiles supplied by the expert. At this point, equation (3.3) is used to calculate of $I(r_j,u_j)$. This index verifies how much the expert's assessment for variable v_j differs from the completely uninformed (i.e., uniform) distribution on X_j.

The global informativeness of expert e_i is given by the average of the individual informativeness index on all seed variables:

$$M(e_i) = \frac{1}{m}\sum_{j=1}^{m} I(r_j,u_j) \tag{3.4}$$

where m denotes the number of seed variables.

The calibration of an expert e_i is calculated as follows. Let B be the (fixed) number of quantiles $q_{k\%}$ that describes each expert's assessment in the experiment. Let us take the values k% in increasing order and associate an index b to each one of them, 1 ≤ b ≤ B. For instance, if the assessment of a seed variable is given by the quantiles $q_{5\%}$, $q_{50\%}$ and $q_{95\%}$, then B = 3 and $q_1 = q_{5\%}$, $q_2 = q_{50\%}$, $q_3 = q_{95\%}$. For a given a variable v defined on interval $[x_l,x_u]$, an assessment consisting of B quantiles can thus be characterized by B + 1 inter-quantile intervals $i_b = [q_{b-1}, q_b]$, $1 \leq b \leq B + 1$, where $q_0 = q_{0\%} = x_l$ and $q_{B+1} = q_{100\%} = x_u$. Let J_b be the set of variables which had their realizations occurring inside interval i_b. The frequency of realizations of an expert in an interval i_b is given by $r_b = |J_b| / m$, i.e., the ratio between the number of variables with realizations in i_b and the total number of seed variables m. Vector r, having the r_b's as elements, represents thus the frequency of success of expert e_i. The *ideal* frequency is given by a vector p, where each p_b is calculated as the total amount of mass assigned between quantiles q_{b-1} and q_b. For instance, for B = 3, the ideal frequency vector p is calculated as $p_1 = 5\%$, $p_2 = 45\%$, $p_3 = 45\%$, $p_4 = 5\%$.

The calibration measure relates each element of the expert's success frequency vector r to the corresponding element in the ideal vector p. An expert is considered to be perfectly calibrated if r = p, i.e., if 5% of the total number of realizations of variables ocurrs between x_l and $q_{5\%}$, 45% between quantiles $q_{5\%}$ and $q_{50\%}$, etc. The difference between the expert's frequencies and the ideal frequencies is calculated using (3.1). The final calibration measure is derived through a hypothesis test given by:

$$C(e_i) = 1 - \mathrm{Chi}^2{}_B(2 \cdot m \cdot I(r(e_i), p) \cdot \omega) \tag{3.5}$$

where $\omega \in [0.1, 1.0]$ is the calibration power of the hypothesis test (usually set to 1) and Chi^2 is the well-known Chi-square function in statistics.

The global measure of an expert's performance is then calculated using the product

$$W(e_i) = C(e_i) \cdot M(e_i). \tag{3.6}$$

Note that this product is in accordance with a fuzzy conjunction of criteria.

In this framework, the concept of an individual calibration measure for each variable does not exist. As a result, no individual quality measure can be obtained. This lack of individual measures may lead to distortions, and represents the major inconvenient of this method. We illustrate one such distortion in the following example. Let us suppose that we ask a group of economists to estimate the value of a share of a given company for each day of a 10 day period. The price of the share on day j is modelled by a variable v_j which varies in the interval [0, 10]. The expert's estimate of the value of a given variable v_j is given in the form of 3 quantiles ($q_{5\%}$, $q_{50\%}$, $q_{95\%}$). An economist will be considered as perfectly calibrated if 5%

of the variables have their realizations below $q_{5\%}$, 5% above $q_{95\%}$, 45% in the interval $[q_{5\%}, q_{50\%}]$ and 45% in the interval $[q_{50\%}, q_{95\%}]$.

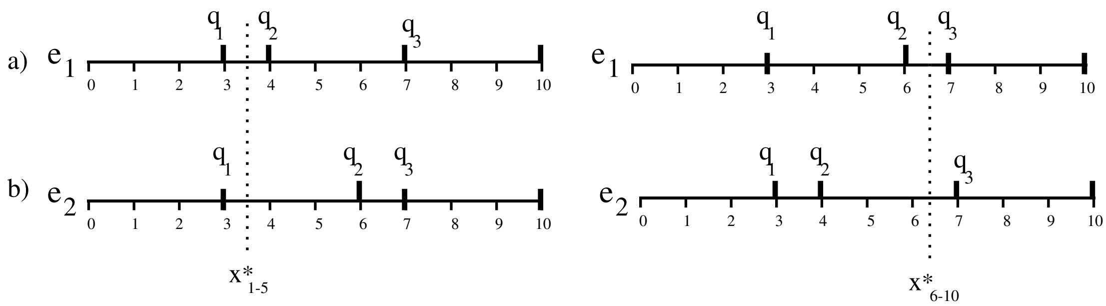

Figure 3.1

a) Quantiles corresponding to the estimations of economist e_1 for variables v_1 to v_{10}

b) Quantiles corresponding to the estimations of economist e_2 for variables v_1 to v_{10}

The true value of v_j (the actual price of the share) is denoted by x_j*. Let us suppose that some time after the estimations were obtained, we verify the following values from actual market data: $x_j^* = 3.5$, $1 \leq j \leq 5$, and $x_j^* = 6.5$, $6 \leq j \leq 10$. In other words, we verify that the price of the share was 3.5 during the first 5 days and 6.5 in the remaining 5 days.

Let us suppose that economist e_1 has supplied the estimation q' = (3, 4, 7) for each variable from x_1 to x_5, and the estimation q" = (3, 6, 7) for each variable from x_6 to x_{10} (see Figure 3.1a). Let us also suppose that economist e_2 has supplied estimation q" for variables x_1 to x_5 and estimation q' for x_6 to x_{10} (see Figure 3.1b). The hypothesis of a uniform mass distribution within each inter-quantile interval is used on all the estimations, in order to derive the corresponding pdf's (see Figures 3.2 and 3.3). From these pdf's we can see that economist e_1 believes that the share will have most likely a value between 3 and 4 units until the fifth day of the period, and will then rise to a value between 6 and 7 units, remaining unchanged for the next 5 days. We can also see that economist e_2 believes that the price of the share will go down exactly when e_1 believes it will go up, and vice-versa.

An examination of the realizations of the variables (see Figure 3.2), allows us to say that economist e_1 is very *accurate*, since the realization of each variable falls in the interval that he considers the most likely to contain the real value of the share. Economist e_2 is clearly less accurate than e_1 (see Figure 3.3), and since they are equally precise, we can say that the opinions of e_1 are more *useful* that those of e_2. However, the utilization of the formulae above yields measures $C(e_1) = C(e_2) = .55$, $M(e_1) = M(e_2) = .68$, and $W(e_1) = W(e_2) = .37$. In this formulation the two economists are considered to be equally calibrated and precise, and the quality function does not discriminate between them. We can see through this simple example that the calibration function is questionable, since it uses so little information from the estimations. We can also see that it does not succeed well at capturing the concept of plausibility, since it does not reward e_1 for considering the intervals where the realizations actually occur as the most plausible ones.

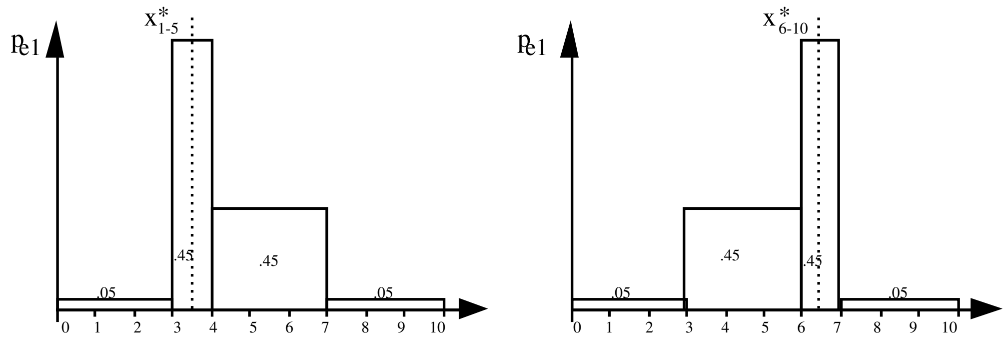

Figure 3.2: Pdf's derived from the quantiles supplied by economist e_1 for variables x_1 to x_{10}

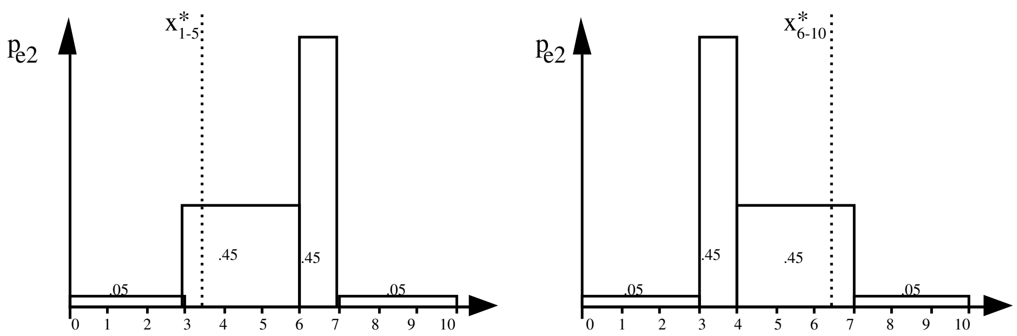

Figure 3.3: Pdf's derived from the quantiles supplied by economist e_2 for variables x_1 to x_{10}

Let us now consider the case of a third economist who supplied the estimation (4, 7, 8) for variable v_1, q'' for variables v_2 to v_5, q' for v_6 to v_9, and (2, 3, 6) for v_{10}. The pdf's derived from these estimations are depicted in Figure 3.4.

We see here that e_3 is even less accurate than e_2; both estimations for variables v_2 to v_9 coincide, but the estimations of e_2 for v_1 and v_{10} are much more accurate than those of e_3. However, $C(e_3) = .82 > C(e_2)$, i.e., e_3 is better calibrated than e_2. This result is consistent with respect to calibration, since the realizations of e_3 are more distributed within the intervals. It is, however, not consistent with a reasonable concept of accuracy, since for economist e_3 there are realizations in the intervals that he considered the least plausible.

We might expect that the informativeness function would somehow be able to compensate for the disagreement between calibration and accuracy. However, $M(e_3) = .7$, which means that e_3 is also more informative than e_1 and e_2. We see here that although e_1 and e_2 clearly give better estimations than e_3, economist e_3 is considered better than e_1 and e_2 in relation to the informativeness/calibration paradigm. This distortion comes partly from the fact that in this paradigm there exists no *individual* quality index. As a consequence, it may happen that a source which gives precise information only when it is inaccurate, and accurate information only when it is imprecise, is considered to be "good". It is important to note that problems such as those described here may be absent in a particular experiment; should they occur, however, they will not be detected, when the method is applied blindly..

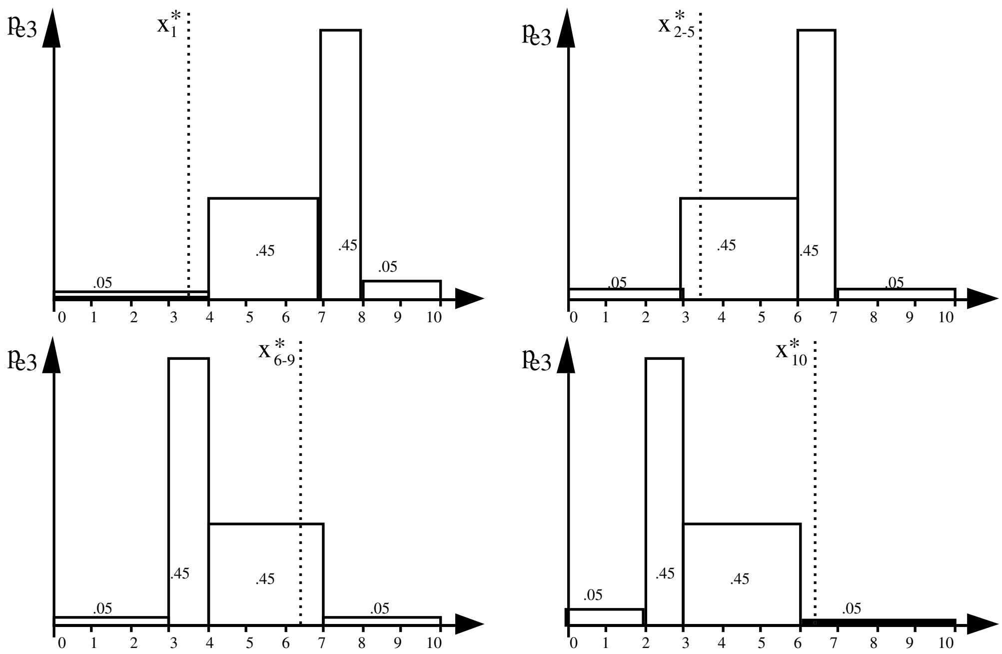

Figure 3.4: Estimations of economist e_3 for variables v_1 to v_{10}

3.2 - The Possibilistic Approach

In order to build scoring indices that reflect these issues in the possibilistic framework, let us first consider a seed variable v whose value x* is precisely known, and let E be the fuzzy set supplied by an expert e, to describe his knowledge about v. Let μ_E be the membership function of E (so that $\mu_E = \pi_v$). In this situation overconfidence cannot arise. It is easy to see that

– the greater $\mu_E(x^*)$, the more accurate is the expert. Indeed if $\mu_E(x^*) = 0$, E totally misses x* while if $\mu_E(x^*) = 1$, x* is acknowledged as a usual value of v. Hence a natural measure of accuracy is given by

$$A(e,v) = \mu_E(x^*) \tag{3.7}$$

– if E is a crisp interval [a,b] the wider E, the more imprecise (hence under-confident) the expert. The width of E is then $|E| = b - a$. When E is fuzzy the width of E is generalized by

$$|E| = \int_X \mu_E(v)\, dv \tag{3.8}$$

This is a generalized fuzzy cardinality (where cardinality becomes the Lebesgue measure). Other extended cardinalities exist to evaluate imprecision (see [23]). In our situation, where E is a finite nested random set, $|E| = \sum_{i=1,m} |A_i| p_i$ (cf. equation 2.1) can be established [5]. This evaluation must be rescaled so as to account for the residual uncertainty ε and so that it yields 1 when $s_l = s_u$ (precise response, for which $|E| = \varepsilon \cdot |X|$ and $|X|$ when $s_l = x_l$, $s_u = x_u$ (empty response). A reasonable specificity index is then

$$Sp(e,v) = f(|E|) = \frac{|X| - |E|}{(1-\varepsilon) \cdot |X|}. \tag{3.9}$$

On the whole, the overall rating of the expert with respect to a single seed variable can be defined as

$$Q(e,v) = A(e,v) \cdot Sp(e,v) \tag{3.10}$$

which requires him to be both accurate and informative in order to score high. By convention, one may take $Q(e,v) = 0$ if $x^* \notin [s_l,s_u]$ instead of $Q(e,v) = \varepsilon \cdot Sp(e,v)$, if we do not want to allow for residual uncertainty in the accuracy index.

When the seed variable is not precisely known, the index $Q(e,v)$ can be extended as follows:

- if the actual value of a seed variable value is described by a histogram leading to a probability distribution P then

$$Q(e,v) = P(E) \cdot Sp(e,v) \tag{3.11}$$

  where P(E) is the probability of the fuzzy event E [38], i.e., $P(E) = \int_X \mu_E(v) d\, P(v)$.

- if the actual value of a seed variable is described by a possibility distribution $\pi^*_v = \mu_F$ then

$$Q(e,v) = \Pi^*(E) \cdot f(|E \,\Delta\, F|) \tag{3.12}$$

  where Π^* is the possibility measure attached to π^*_v and Δ is the symmetric difference of fuzzy sets. More specifically, $\Pi^*(E) = \sup_x \min(\mu_F(x), \mu_E(x))$ is the possibility of the fuzzy event E [39], and $\mu_{E\Delta F}(x) = |\mu_E(x) - \mu_F(x)|$ (see [11]). $\Pi^*(E)$ evaluates the extent to which the expert's response is consistent with the available information about v, and $f(|E \,\Delta\, F|)$ (see (3.9)) penalizes both underconfidence and overconfidence on the expert's part. When the possibility (or the probability) distribution of v reduces to deterministic information ($v = x^*$) then the above indices collapse into the first definition (3.10) up to the scaling factor in (3.9) that can be added if needed.

Global measures of accuracy, precision and quality to an expert e can be obtained using the simple arithmetic mean over the individual scores. It m is the total number of seed variables, then

$$A(e) = {}^{1}/_{m} \cdot \Sigma_{j=1,m}\, A(e,v_j), \tag{3.13}$$

$$Sp(e) = {}^{1}/_{m} \cdot \Sigma_{j=1,m}\, Sp(e,v_j), \tag{3.14}$$

$$Q(e) = {}^{1}/_{m} \cdot \Sigma_{j=1,m}\, Q(e,v_j), \tag{3.15}$$

It is important to note that generally $Q(e) \neq A(e) \cdot Sp(e)$.

Thus an expert e is rated by the set $\{Q(e,v_j) \mid j = 1,m\}$ of evaluations. Ranking of experts can be obtained based on the average rating of each expert. The standard deviation is also useful to check the significance of the gaps between average ratings of experts. Based on these evaluations a set K of experts can be divided into groups of equal reliability. Moreover, the fuzzy set R of reliable experts can be defined by the membership function

$$\mu_R(e_i) = Q(e_i),\ i = 1, \ldots, k \tag{3.16}$$

if there are k experts. The cardinality of R, say

$$|R| = \Sigma_{i=1,k}\, \mu_R(e_i) \tag{3.17}$$

gives a good idea of the number of reliable experts in the group. The reader can check that the drawbacks of calibration, pointed out in Section 3.1 disappear in the market share example, if the possibility distributions supplied by the three experts have shapes similar to the ones in Figures 3.2, 3.3 and 3.4. In particular $A(e_1) > A(e_2) > A(e_3)$.

4 - The Pooling of Expert Judgements

4.1 - The Probabilistic Approach

When several experts supply probability distributions, their responses are pooled so as to derive a single distribution that reflects the opinion of the group. It is clear, however, that the opinion of reliable experts should be more important than those of unreliable ones. There are two main approaches to the pooling of probability distributions, the direct aggregation of distributions such as the consensus method justified by Wagner and Lehrer [32], and recently used by Cooke [1], and the Bayesian approach, exemplified by the works of Winkler [33], Morris [25], and Apostolakis and Mosleh [26] (see French [19] and Cooke [4] for critical surveys). Most direct aggregation methods use a generalized mean operation acting on distributions, followed by a remormalization. These methods often suffer from defects such as sensitivity with respect to marginalization and the lack of independence preservation. Only the weighted arithmetic mean possesses the so-called "strong setwise function property" [24, 32]; it prescribes that for any event the aggregated probability should depend only on the expert probabilities of that event, which ensures insensitivity with respect to marginalization.

In the consensus method each expert e_i supplies a pdf p_i, and the resulting distribution is a weighted average $p = \sum_i w_i p_i$ where the weights w_i reflect the reliability of experts. No renormalization is needed, and the basic problem is to find the weights. Cooke [1] has developed a theory of weights that derive from proper scoring rules which tend to force experts to be calibrated and informative. In this approach, additional methods are obtained if a significance level is taken into account and used to discard experts with low scores.

In the Bayesian methods, the a priori opinion of the analyst about the true value of v is updated on the basis of expert opinions, expressed either as point-values or distributions, depending upon the specific method. In the work of Apostolakis and Mosleh [26], the credibility of experts, from the standpoint of the analyst, is modelled by conditional probabilities of what an expert will claim the true value of v is, given this true value. Once the expert point-values are known, the a priori probability distribution of v, as given by the analyst, is updated through Bayes theorem. The model tries to account for the dependence between experts via a correlation coefficient.

The probabilistic pooling approaches can be criticized for several reasons [15]:

– The consensus method has a basic flaw in the context of reliability: it is a voting-like procedure. Indeed, if two reliable experts have conflicting opinions about the value of v such that one gives a small value to v, and the other gives a high value to v, the mean of the distribution obtained by the consensus method will be an intermediate value, i.e., a value which both expert agree is not the true one. When such conflicts occur, different expert opinions are usually not combined in realworld applications. The pieces of information are propagated separately when conflicting. Hence expert judgement approaches usually do not solve the problem of conflicting experts and try to avoid such situations during the selection of experts. What is needed is a method which, in the best case guesses the true value and discards the wrong expert, or, in the worst case, proposes a cautious response that fits the available data (e.g., v is either small or large, but certainly not medium). The weighted average method sounds more natural when expert opinions express preference and the preference of a group must be estimated. It does not seem to be useful when a true answer is to be determined instead of a preferred one.

– The weighted average method may affect the variance in the sense that the variance of the result may become smaller than the one of any input distributions. This phenomenon is acceptable it the experts are independent. However experts often share a great deal of technical background, and the expert independence assumption is highly questionable.

– The main defect of the Bayesian method seems to be, as usual, the need for a priori knowledge about the value of v. In other words, the analyst who looks for expert advice must be an expert himself. However, in many cases, the analyst has no idea about the value of v and he may learn little more than the extent to which the experts are reliable. Techniques as the one by Cooke [1], nonetheless, have inspired the method described in this paper. The Bayesian methods, being unable to update from the state of complete ignorance, require an a priori probability. The Bayesian approach, as construed by Mosleh and Apostolakis [26] has the merit of handling dependency among experts via correlation coefficients. However, as pointed out by Cooke [4], this Bayesian method places heavy assessment burden on the analyst. Moreover, in the case of several conflicting experts, voting-like effects resulting in values that no expert supplies can be observed with the Bayesian method (see [15] for a more detailed analysis of the Bayesian method).

The possibilistic approach that is proposed in this paper tries to cope with most of the difficulties faced by the probabilistic approach. Its main features are faithfulness of the representation of subjective data, no need for a priori knowledge, and a variety of pooling methods whose choice depends about the reliability of experts and the level of conflict among their opinions.

4.2 - The Possibilistic Approach

The basic principle of the possibilistic approach to the pooling of expert judgements is that there is no unique mode of combination that fits all situations: the choice of the combination mode depends on an assumption about the reliability of experts, as formulated by the analyst. This point of view strikingly differs from the one adopted in probabilistic approaches, and applies beyond the possibilistic setting. No a priori knowledge about the variable under study is needed, and the experts are viewed as a set of parallel sources to be combined in a symmetric way only if all experts are equally reliable. There are basically two extreme modes of symmetric combination, the conjunctive mode when all experts agree and are reliable, and the disjunctive mode when experts disagree and at least one of them is considered to be reliable. These modes are implemented respectively as a fuzzy set intersection and a fuzzy set union. A third mode of symmetric combination is averaging, which considers the expert opinions in a more statistical way. This set-theoretic view on combination of uncertain pieces of information, introduced in [8, 10], has been applied to multiple source interrogation systems [29]. In the case of expert knowledge, the pooling mode depends upon the results of the assessment step and the extent to which expert responses on the inquired variable agree with one another (see [14] for discussion of a variety of combination rules in possibility theory, some of which have been used in the present study).

Conjunctive Mode

Let π_i be the possibility distribution supplied by expert i, for $i \in K$. If all the experts are considered to be reliable (e.g., all the ratings $\mu_R(i)$ are high) then the response of the group of experts is defined by

$$\pi_C(x) = \min_{i \in K} \pi_i(x) \tag{4.1}$$

This mode makes sense if all the π_i overlap significantly, for instance if $\exists x$, $\pi_C(x) = 1$, expressing that there is a value of v that all experts consider as having a high degree of possibility. If $\pi_C(x)$ is significantly smaller than 1 this mode of combination makes no sense since in that case one of the experts may be wrong. Note that when all experts agree perfectly ($\pi_i = \pi_C$, $\forall i$), there is no reinforcement effect. Generally, agreement between experts is due to common background, and the idempotence of min deals with this kind of redundancy. If the experts can be considered as independent, the minimum can be replaced by product. This pooling method is sensitive to marginalization and can be criticized on this basis.

Example (adapted from [4])

Two experts judge the state of a flashlight that was kept unprotected. It either works (w) or not (¬w). The two experts agree that it is very unlikely that it works ($\pi_1(w) = \pi_2(w) =.2$; $\pi_1(\neg w) = \pi_2(\neg w) = 1$). Hence $\pi_C(w) =.2$, $\pi_2(\neg w) = 1$. But suppose ¬w means either dead battery (db) or corroded contacts (cc) but not both. Experts may disagree on the failure cause e.g., $\pi_1(db) = 1$, $\pi_1(cc) = .1$ and $\pi_2(db) = .1$, $\pi_2(cc) = 1$. Using the min rule we get $\pi_{C'}(db) = \pi_{C'}(cc) = .1$ from which we can conclude only that $\pi_{C'}(\neg w) = .1$ instead of the 1 derived previously.

One may be tempted to discard the rule as being self-contradictory. A closer examination of the situation, however, reveals that the reason for the inconsistency is a superficial agreement on the failure of the flashlight which hides a deeper disagreement. Only an aggregation on the set {w, db, cc} highlights this disagreement. The example shows that it is better to pool the distributions than the set-functions. Moreover $\pi_{C'}$ is strongly subnormalized, in such a case the min rule does not apply since one of the experts is wrong. In fact, each time the sensitivity to marginalization manifests itself, it corresponds to a conflict between experts and the min rule should not be applied.

When the resulting possibility distribution is subnormalized but the assumption of reliable experts is taken for granted, it makes sense to renormalize the distribution, by dividing π_C by its height $h(\pi_C) = \sup \pi_C$. However the drawback of renormalization is that it obliterates the conflict between experts. A more faithful normalization technique is to use $1 - h(\pi_C)$ as a residual uncertainty and compute $\pi_{C'} = \pi_C + 1 - h(\pi_C)$.

Disjunctive Mode

A rather cautious optimistic assumption about a group of experts is that one expert is right, but it is not known which. This assumption corresponds to the following aggregation

$$\pi_D(x) = \max_{i \in K} \pi_i(x) \tag{4.2}$$

This is a very conservative pooling mode that allows for contradiction among experts but may not lead to an informative result, although not necessarily a vacuous one either. Note that if the reliability of experts is unknown, and that it is not even certain that one of them is right, then the only pooling method that remains is to look for consensus among experts and to discard outliers.

Averaging Mode

This mode correspond to viewing experts as random sources and hence potentially unreliable. Values of the parameters that experts agree are possible are considered more plausible than values that most expert reject. The following averaging rule is then applied

$$\pi_A(x) = \frac{1}{K}\Sigma_{i\in K}\ \pi_i(x) \tag{4.3}$$

Note that this value is normalized only if the conjunctive rule gives a normalized result. The lack of normalization indicates the all experts may be wrong. The two modes of renormalization still apply, if this option is ruled out. Generally, in the case of disagreement among experts, a multimodal possibility distribution is obtained as with the disjunctive mode.

Pooling Based on Numerical Quantifiers

Another intermediary mode of pooling the π_i's consists in assuming that j experts out of K are reliable, selecting a subset $J \subseteq K$ of experts such that $|J| = j$, assuming that they are the reliable ones and combining their opinions conjunctively. Assuming finally that at least one of these subsets J contain reliable experts, the intermediary results can be combined disjunctively from the following formula is obtained [17]

$$\pi_{(j)}(x) = \max_{J\subseteq K,\ |J|=j} \min_{i\in J} \pi_i(x) \tag{4.4}$$

The choice of j can be guided by the value |R| obtained from the assessment step. Indeed |R| gives a rough idea of the number of reliable experts in K, so that j could be chosen close to |R|. Clearly, $\pi_{(k)} = \pi_C$ and $\pi_{(1)} = \pi_D$, i.e., this mode of aggregation subsumes the two previous ones. The above combination rule is equivalent to certain rules proposed by Yager [36] in the past, and can be easily calculated, as follows:

i) rank-order the $\pi_i(x)$ such that $\pi_{i_1}(x) \geq \pi_{i_2}(x) \geq \ldots \geq \pi_{i_k}(x)$
ii) then $\pi_{(j)}(x) = \pi_{i_j}(x)$.

This scheme can be extended to fuzzy quantifiers, in order to model assumptions such as "most experts are reliable", "approximately j experts are reliable", etc. [17, 36]. Again, the choice of the fuzzy quantifier can derive from the fuzzy cardinality of the fuzzy set R of reliable experts.

Consistency-Based Trade-Offs

Another way to trade-off between the conjunctive and disjunctive modes of pooling is to use a measure c of conflict between two experts and to define

$$\pi_T(x) = c\max(\pi_1,\pi_2) + (1 - c)\min(\pi_1,\pi_2).$$

This index gives the conjunctive (disjunctive) mode if c = 0 (c = 1). It is easy to define conflict measures between π_1 and π_2 namely [14]

•) $c = 1 - \mathrm{cons}(\pi_1,\pi_2)$, where $\mathrm{cons}(\pi_1,\pi_2) = \sup_x \min(\pi_1(x),\pi_2(x))$ is the level of consistency between π_1 and π_2 [39]

•) $c = 1 - \mathcal{J}(\pi_1,\pi_2)$ where $\mathcal{J}(\pi_1,\pi_2)$ is the Jacquard index defined by a quotient of fuzzy cardinalities

$$\mathcal{J}(\pi_1,\pi_2) = \frac{|F_1 \cap F_2|}{|F_1 \cup F_2|}$$

where $\mu_{F_1} = \pi_1$ and $\mu_{F_2} = \pi_2$. The extension of this index to n sets $F_1, \ldots, F_n$ is obvious, changing $F_1 \cap F_2$ into $F_1 \cap \ldots \cap F_n$ and $F_1 \cup F_2$ into $F_1 \cup \ldots \cup F_n$ in the expression of $\mathcal{J}(\pi_1,\pi_2)$, so as to form $\mathcal{J}(\pi_1, \pi_2, \ldots, \pi_n)$. An alternative consistency-dependent rule is described in [14], and an extension to more than two sources is proposed in [16].

Discounting Experts

If the degree of certainty that a given expert is reliable is known, say w_i then it is possible to account for this information by changing π_i to $\pi'_i = \max(\pi_i, 1 - w_i)$ (see, e.g., [11]). When $w_i = 1$ (reliable expert), $\pi'_i = \pi_i$ and when $w_i = 0$ (unreliable expert), then $\pi'_i = 1$. Note that $w_i = 0$ does not mean that the expert lies, but that it is impossible to know whether his advice is good or not. Once discounted, expert opinions can be combined conjunctively. It is difficult, however, to quantitatively relate w_i to the rating $\mu_R(i)$ except, of course in the fact that a higher $\mu_R(i)$ corresponds to a higher w_i. Moreover, the result of a conjunctive combination of discounted possibility distributions is rather difficult to interpret in the case of conflicting opinions of equally reliable experts, since the resulting possibility distribution can be overwhelmed by uncertainty levels.

Priority Aggregation of Expert Opinions

As pointed out earlier, the fuzzy set R of reliable experts is useful to partition the set K of experts into classes K_1, K_2,…, K_q of equally reliable ones, where K_j corresponds to a higher reliability level than K_{j+1}, for j = 1,…, q. In this case, the symmetric aggregation schemes discussed above can be applied to each class K_j. The combination between results obtained from the K_j's can be performed using the following principle: the response of K_2 is used to refine the response of K_1 insofar as it is consistent with it. If π_1 is obtained from K_1 and π_2 from K_2, the degree of consistency of π_1 and π_2 is $\text{cons}(\pi_1,\pi_2) = \sup_x \min(\pi_1(x),\pi_2(x))$ and the following combination rule has been proposed [12, 37]:

$$\pi_{1-2} = \min(\pi_1, \max(\pi_2, 1 - \text{cons}(\pi_1,\pi_2)) \quad (4.5)$$

Note that when $\text{cons}(\pi_1,\pi_2) = 0$, K_2 contradicts K_1 and only the opinion of K_1 is retained ($\pi_{1-2} = \pi_1$), while if $\text{cons}(\pi_1,\pi_2) = 1$ then $\pi_{1-2} = \min(\pi_1,\pi_2)$. π_{1-2} can be similarly combined with π_3, $\pi_{(1-2)-3}$ with π_4 and so on. A similar result can be obtained if, instead of $\text{cons}(\pi_1,\pi_2)$, we use a Jacquard index $\mathfrak{J}(\pi_1,\pi_2)$.

Remark

One may wonder about the consistency between the basic aggregation rules (4.1) and (4.2) and the probabilistic interpretations of the possibility distributions as upper probability bounds. Clearly, the results are only approximations. For instance, the min-rule (4.1) correspond to performing the intersection of the sets of probabilities corresponding to and and considering the possibility distribution that is the best inner approximation to the result of the intersection (see, e.g., [13]). Other probabilistic justifications for fuzzy set connectives can be found in [20, 21].

5 - Comparison of Expert Judgement Systems: Principles and Set-Up

A procedure for processing expert-supplied information, in a given uncertainty model, can be divided into three parts. In the first part of the process the domain of expertise is established, with the determination of the experts and the seed variables that will be used to calibrate them. The second part of the process consists of the calibration of the experts based on the principles of the uncertainty model adopted. The basic phases of this part are:

i) elicitation of variables values by experts

ii) assessment using the Q(e,v) indices on seed variables
iii) pooling of the experts assessments
iv) comparison of the performance of pooling methods on seed variables.

At this point in the process the analyst will use the best method as determined by step iv to produce the values for the variables of interest.

An expert judgement system devised on this procedure, called EXCALIBR, has been implemented in the probabilistic framework [4]. Based on the ideas presented in this paper (see also [22]), a possibilistic expert judgement system has also been implemented [27]. An ideal way of comparing the probabilistic and possibilistic approach is to set up an experiment in which the experts give possibilistic and probabilistic data for the same set of variables. The results in one framework can then be transformed into the second framework and compared with the results obtained directly in the second framework.

A slightly different experiment might consist of using the same set of data in both the probabilistic and possibilistic frameworks. Of course, the initial data would have to be transformed to fit one of the the frameworks. Such an experiment is not capable of analysing human preference with respect to the different elicitation paradigms, it can, however, give us an important insight into the relationship between the probabilistic and possibilistic treatment of data. Figure 5.1 illustrates this experiment using probabilistic data as input.

An experiment such as the one depicted in Figure 5.1 has been set up to evaluate the ideas presented in this paper and conducted by IRIT and the Joint Research Center of the C.E.C. Two sets of data, DSM [2] and ESTEC [3], were analyzed by the possibilistic expert judgement system and the probabilistic system EXCALIBR. The experts and the combination methods proposed by each system calibrated in both systems. Transformations between possibilistic and probabilistic data were provided by the possibilistic system. Note that in both systems the combination and model assessment module can be used iteratively, and the analyst may experiment with various combinations methods before making his final choice.

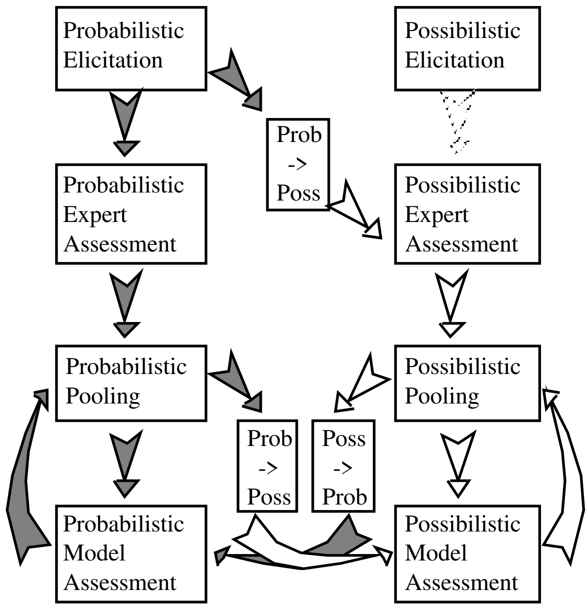

Figure 5.1: Cross-reference experiment between the Probabilistic and Possibilistic Expert Judgement Models.

In order to make the features of each system easier to grasp we demonstrate the experiment with a very simple set of data. Such a toy example has been chosen to make calculations easy to follow, and should not be taken as a basis for real comparison between the systems. Moreover, it only illustrates part of the features of each of the systems above, the full capacities of each system can be found in the referenced literature. We shall present first the methods used in the transformations between the probabilistic and possibilistic framework followed by the basic notation used in the remainder of the text, and our simple example. Its application in both the possibilistic and probabilistic system is demonstrated in the following sections.

5.1 - Transformations Between Possibility and Probability

Let p be a unimodal pdf, and let x_0 be the mode of p. A possibility distribution can be derived from p by applying the transformation T1 [18]:

$$\text{T1: } \pi(x) = \pi(x') = \int_{x_l}^{x} p(v)dv + \int_{x'}^{x_u} p(v)dv$$

where x' is such that $p(x') < p(x_0)$, and there is no y such that $x < y < x'$, and $p(y) < p(x)$. The possibility distribution π is the most specific one among those which dominate p (i.e., $\prod(A) \geq$

P(A) for all events A), by virtue of the possibility-probability consistency principle (see [39] and [7]).

Conversely the transformation T2 can be used to transform a possibility distribution into a pdf, where T2 is given by [18]:

$$\text{T2: } p(x) = \int_0^{\pi(x)} \frac{d\alpha}{|A_\alpha|}$$

where $A_\alpha = \{x \,/\, \pi(x) \geq \alpha\}$. The characteristics of our data allow us to use the discrete equivalent of T2:

$$p(x) = \sum_{i=1}^{n} \frac{\alpha_i - \alpha_{i+1}}{|A_i|} \mu_{A_i}(x)$$

where $A_1, \ldots, A_n$ correspond to $\alpha_1 = 1 > \alpha_2 > \ldots > \alpha_n > \alpha_{n+1} = 0$, and function $\mu_{A_i}(x)$ is such that $\mu_{A_i}(x) = 1$ when $x \in A_i$ and 0 otherwise. This transformation obeys Laplace's principle of indifference applied to all level-cuts of π. It corresponds to picking at random a value in [0,1], and at random an element in the corresponding level cut. It is identical to what Smets [31] calls the pignistic transformation of a belief function into a probability measure, when restricted to a possibility measure. Note that T2 is not the converse of T1 because different informational principles govern each transformation [18].

5.2 - Basic Notation for the Experiments

Suppose that a given experiment involves n experts and m variables. We will denote an expert i by e_i, and a group of experts by g_k. The true value and the range of seed variable v_j are respectively denoted by $x^*(v_j)$, and $[x_l,x_u](v_j)$. Let $q(e_i,v_j)$ represent a group of quantiles of a subjective probability distribution yielded by expert e_i for variable v_j. Then $p(e_i,v_j)$ denotes the pdf obtained from $q(e_i,v_j)$ using the uniform mass distribution hypothesis, and $\pi(e_i,v_j)$ denotes possibility distribution derived from $p(e_i,v_j)$ using transformation T1. $\varepsilon(e_i,v_j)$ denotes the residual uncertainty factor supplied by the expert e_i in relation to variable v_j. The argument v_i will be dropped when no confusion is possible. The pooling methods will be denoted by type/method/group; a probabilistic pooling method m applied to a group of experts g_k on a given variable will be denoted by $p/m/g_k$. Analogously, the result of the application of a possibilistic method n will be denoted by $\pi/n/g_k$.

5.3 - The Example

Before describing a real experiment, we describe a simple example in order to facilitate the assimilation of the notation used here, and to clarify the characteristics of each system. We have 10 seed variables, and two experts who give assessments in the form of 5%, 50%, and 95% quantiles of a subjective distribution, i.e., $q(e_i,v_j) = (q_{5\%}, q_{50\%}, q_{95\%})$. The true value of each seed variable is given in the form of a real number. The data of the simple example is summarized as follows:

Number of experts: $n = 2$,

Number of test variables: $m = 10$

Variable domain: $[x_l,x_u](v_j) = [0, 10]$, $1 \leq j \leq 10$

Real value of variables:
$x^*(v_1) = 2.5$
$x^*(v_2) = x^*(v_3) = x^*(v_4) = x^*(v_5) = 3.5$
$x^*(v_6) = x^*(v_7) = x^*(v_8) = x^*(v_9) = 4.5$
$x^*(v_{10}) = 7.5$

Input expert e_1: $q(e_1,v_j) = (1, 4, 8)$, $1 \leq j \leq 10$
Input expert e_2: $q(e_2,v_j) = (3, 4, 7)$, $1 \leq j \leq 10$

Group of experts used in the experiment: $g_1 = \{e_1,e_2\}$

In Figure 5.2 we illustrate in a condensed manner the estimations given by the experts e_1 and e_2 for variables v_1 to v_{10}, as well as the variables true values.

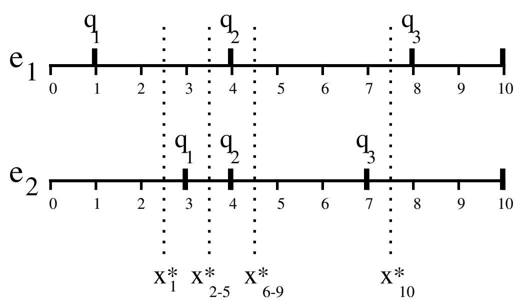

Figure 5.2: Experts' estimations and true values of variables

6 - The Possibilistic System Peaps

6.1 - The elicitation step

In the elicitation step we will distinguish two subphases: system knowledge elicitation and expert knowledge elicitation. In the first subphase we will acquire the necessary variables and their basic attributes, such as range, their role in the procedure (seed, target variable or both) and the representation model in which each seed variable is encoded (real number, pdf or possibility distribution), and the value itself. In the second subphase we will acquire the expert's estimate for each variable, the model he is using to represent it, and his confidence in it. The system provides the means of transforming a pdf into a possibility distribution when the pdf is given in the form of a discrete distribution, a set of quantiles, or a linear by parts continuous function.

In our simple example the variable ranges were given, in practice they can be estimated from the information supplied by experts using conservative extrapolations. From the quantiles given by each expert we derive a pdf taking as hypothesis that the distribution is uniform (see Figure 6.1). Then we transform the pdf into a possibility distribution using transformation T1 as illustrated in Figure 6.2.

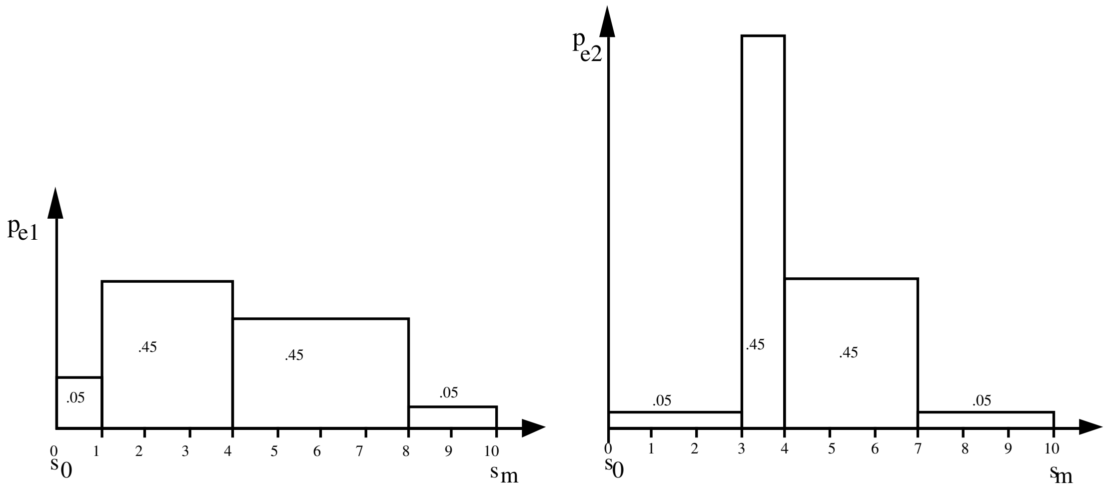

Figure 6.1: Pdf constructed with the 5%, 50% and 95% quantiles

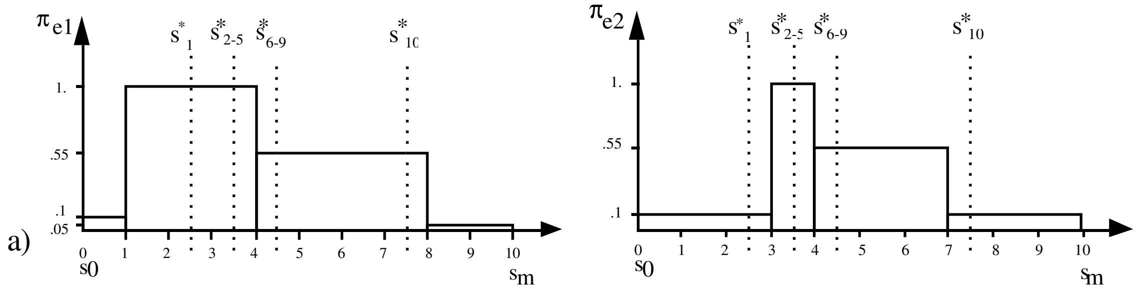

Figure 6.2: Possibility distributions constructed from pdf derived from the .05, .5 and .95 quantiles.

It is important to note that the uniform mass distribution hypothesis is completely arbitrary; there are an infinite number of pdf's compatible with a given group of quantiles.

6.2 - The assessment step

In this step the experts' assessments are compared to the true values of the seed variables as described in Section 2. For each expert e_i and seed variable v_j we calculate the precision measure $Sp(e_i,v_j)$, the accuracy measure $A(e_i,v_j)$, and the quality measure $Q(e_i,v_j)$. We then calculate the mean for the individual accuracy, precision and quality measures (i.e., the global measures $A(e_i)$, $Sp(e_i)$ and $Q(e_i)$), and the corresponding standard deviations, denoted respectively by $\sigma_A(e_i)$, $\sigma_{Sp}(e_i)$, $\sigma_Q(e_i)$. For each measure we then establish a ranking on the experts.

As an example, let us verify the assessment of expert e_1 in relation to variable v_6: his accuracy is $A(e_1,v_6) = \pi_{e_1,v_6}(4.5) = .55$, his precision is $Sp(e_1,v_6) = 1 - (.1 + 3 * 1 + 4 * .55 + 2 * .05) / 10 = .46$, and his quality is $Q(e_1,v_6) = .55 * .46 = .253$. In Table 6.1 we depict the mean and standard deviation for each expert's assessments over the total set of seed variables.

	Possibilistic Evaluation		
Experts	$A(e_i)$	$Sp(e_i)$	$Q(e_i)$
e_2	.64	.675	.432
e_1	.775	.46	.3565

Table 6.1.a: Mean of expert's assessments (global measures)

$\sigma_A(e_i)$	$\sigma_{Sp}(e_i)$	$\sigma_Q(e_i)$
e_2 .3828	e_2 0	e_2 .3349
e_1 .225	e_1 0	e_1 .1012

Table 6.1.b: Standard deviation of expert's assessments

We can see that, considering the whole set of seed variables, even if expert e_1 is more accurate than expert e_2, his low performance on precision leads us to consider him less "good" than expert e_2. Note that here $Q(e_i) \neq A(e_i) \cdot Sp(e_i)$.

Using the possibilistic criteria cr (A for accuracy, Sp for precision, and Q for quality) a series of orderings $\mathbf{order_{cr}}$ on a given group are determined. In our example, the orderings are

$$order_A(g_1) = (e_1,e_2),\ order_{Sp}(g_1) = (e_1,e_2),\ order_Q(g_1) = (e_1,e_2),$$

On a given group of experts we can identify subsets of homogeneous experts. Let $order_{cr}(g_k)$ be the ordering induced on the experts in group g_k by criterion cr (A, Sp or Q), such that e_i is classified better than e_{i+1}. Experts e_i and e_{i+1} are in the same homogeneous group in relation to a given discrimination factor ρ, if

$$cr(e_i) - cr(e_{i+1}) \leq \rho \cdot (\sigma_{cr}(e_i) + \sigma_{cr}(e_{i+1})),$$

We establish the homogeneous groups by comparing first of all the two best classified experts with respect to criterion cr, i.e., the first two experts in $order_{cr}$. We also compare the second expert with the third, and so on. By increasing factor ρ we obtain a coarsening of $order_{cr}$ denoted by $order_{cr\text{-}\rho}$. In $order_{cr\text{-}\rho}(g_k)$ we will then have L homogeneous subsets K_l, $1 \leq l \leq L$, where K_l is superior to K_{l+1} with respect to criterion cr and factor ρ.

For instance, taking $\rho = 0$ and the accuracy index, we have $order_{A\text{-}0}(g_1) = (K_1,K_2) = (\{e_1\}, \{e_2\})$, i.e., we obtain two homogeneous subsets, each one consisting of a single expert. When $\rho = 1$ we have $order_{A\text{-}1}(g_1) = (K_1) = (\{e_1,e_2\})$, i.e., with this factor the group is considered homogeneous. For group g_1, we have:

$$order_{A\text{-}0}(g_1) = (\{e_1,e_2\})$$
$$order_{A\text{-}1}(g_1) = order_{Q\text{-}1}(g_1) = (\{e_1,e_2\})$$
$$order_{Sp\text{-}0}(g_1) = order_{Sp\text{-}1}(g_1) = order_{Q\text{-}0}(g_1) = (\{e_2\},\{e_1\}),$$

Note that $order_{cr}(g_k)$ represents an order induced on g_k, and $order_{cr\text{-}\rho}(g_k)$ represents an order induced on a partition of g_k; this partition itself being determined by $order_{cr}(g_k)$.

Let g be a set of experts and v_j be a variable. The conflict inside g with respect to v_j can be measured by $\kappa_j(g) = 1 - \mathfrak{J}_j(g)$, where $\mathfrak{J}_j(g)$ is the Jacquard index on g for variable v_j (see Section 4.2). The global conflict index is denoted by $\kappa(g)$, and is the mean of the individual indices $\kappa_j(g)$. In our example, the conflict on group g_1 with respect to each variable is $\kappa_j(g_1) = 1 - (3.15 / 5.5) = .4273$, and consequently $\kappa(g_1) = .4273$, which represents a rather large conflict.

6.3 - The Pooling Step

In this step the analyst may choose to combine the opinions of a group of experts regarding the value of a variable. He is presented with a set of choices that tries to model the possible situations he may be faced with according to the reliability he grants to the experts. He may regard the reliability of the experts as known, either based on his own experience or on one of the measures yielded by the assessment step, or he may consider that information concerning the experts does not allow him to have them ranked. In any case the system uses combination methods that should optimally reflect what the analyst could possibly learn about the unknown variable by consulting the experts. The choice of combination methods offered by the system is based on the development of an interface for combining pieces of information derived from distinct data banks [29], but other operations can easily be incorporated.

In the present experiment two groups of combination methods have been used. In the first group we have the symmetric operations min, max, and trade-off. These methods are denoted by $\pi/m/g_k$, where m stands for the chosen method. Using any of these methods the opinions of all the experts in a group are taken into account equally. The basis of the second group of pooling methods is the asymetric operation (4.5) that weights more heavily the opinions of experts which are better ranked in relation to an evaluation criterion cr and a factor ρ. The methods in this group are denoted by $\pi/cr\text{-}\rho/g_k$, where $cr \in \{A, Sp, Q\}$. In the table below we summarize the various pooling methods and the normalization operation associated with each symmetric method.

PEAPS Symetric Aggregation Methods: $\boldsymbol{\pi/min/g_k}$, $\boldsymbol{\pi/max/g_k}$, $\boldsymbol{\pi/trade/g_k}$
Let π_i represent the distribution given by expert e_i for a given variable v defined on X. We have: $\pi/min/g_k$: $\forall\ x \in X, \pi'(x) = \min_{e_i \in g_k} \pi_i(x)$, $\forall\ x \in X, \pi(x) = \pi(x) / h(\pi')$ $\pi/max/g_k$: $\forall\ x \in X, \pi(x) = \max_{e_i \in g_k} \pi_i(x)$ $\pi/trade/g_k$: $\forall\ x \in X, \pi'(x) = [(1 - \kappa(g_k)) \cdot \min_{e_i \in g_k} \pi_i(x)] + [\kappa(g_k) \cdot \max_{e_i \in g_k} \pi_i(x)]$ $\forall\ x \in X, \pi(x) = \pi'(x) + (1 - h(\pi'))$

Table 6.2

Let π_i represent the distribution given by expert e_i for a given variable v defined on X. Let $order_{cr\text{-}\rho}(g_k)$ represent the ordering of the experts with respect to criterion cr and factor ρ. The asymmetric pooling process is divided in two phases (see Table 6.3), symmetric aggregation inside homogeneous subgroups K_k, and asymmetric combination of the obtained partial results. Let K_k be a homogeneous set of experts and π_i the possibility distribution. Distributions π_i supplied by expert $e_i \in K_k$ for a given variable v, are combined using the trade-off method π/trade, without the normalization step. In the second step, the resulting set of distributions is aggregated applying a pairwise top-down procedure that favors the best ranked set of experts. Let π_{Kk} be obtained by π/trade/K_k. The combination of two homogeneous groups K_k and K_{k+1} is made by (4.5). The mechanism for determining π' is illustrated in Figure 6.3, for a group of 5 experts partitioned in 5 singletons $K_k = \{e_k\}$ (i.e., each homogeneous group contains a single expert).

PEAPS Asymmetric Aggregation Methods: **π/A-ρ/g_k**, **π/Sp-ρ/g_k**, **π/Q-ρ/g_k**
1) π/trade/K_l: $\forall x \in X, \pi(x) = [(1 - \kappa(K_l)) \cdot \min_{e_i \in K_l} \pi_i(x)] + [\kappa(K_l) \cdot \max_{e_i \in K_l} \pi_i(x)]$ 2) $\forall x \in X, \pi'(K_l, K_{l+1})(x) = \min(\pi_{K_l}(x), \max(\pi_{K_{l+1}}(x), k(\{K_1,K_2\})]$

Table 6.3

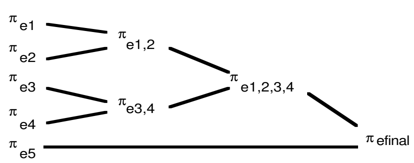

Figure 6.3: Pairwise top-down combinations of unequally ranked distributions

The first step of the mechanism implementing an asymmetric method yields a distribution that summarises the collective opinion of a group of n closely ranked experts, none of whom is regarded as completely reliable. The second step comes down to discounting the information of the less reliable distribution (here π_2) by the degree of conflict with the more reliable one. We note here that, instead of the Jacquard index $\kappa_{1,2}$, we could also use $\sup_x \min(\pi_1(\omega),\pi_2(\omega))$ as a discounting factor.

In our example, the pooling methods π/Sp-0/g_1, π/Sp-1/g_1 and π/Q-0/g_1 yield the same distributions of expert e_2 alone. Pooling methods π/A-1/g_1 and π/Q-1/g_1 yield the same distributions as π/trade/g_1. Figure 6.4 illustrates the application of rules π/min/g_1, π/max/g_1, π/trade/g_1 and π/A-0/g_1:

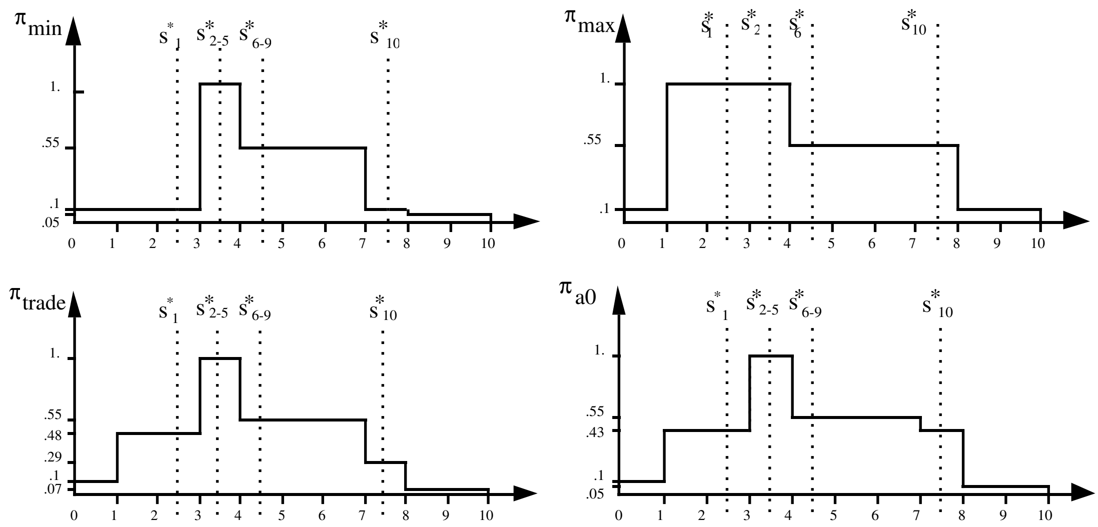

Figure 6.4: Application of rules π/min/g_1, π/max/g_1, π/trade/g_1 and π/A-0/g_1 to the distributions furnished by e_1 and e_2

6.4 - The Model Assessment Step

When a series of estimations on test variables is available, the procedure allows for a quality assessment of the results of any of the described combination mechanisms by treating the aggregated result as a 'virtual' expert, and comparing it with the observed true values. The three average scores of the 'model' are computed and may be compared with those of the participating experts, in order to see whether the aggregation results behave better globally than any of experts taken individually. By varying the composition of the expert pool and/or the combination mechanism, the analyst may search for some optimal processing, yielding results that are more reliable than the individual expert input data.

In our example the model assessment yields the following measures:

	Possibilistic Evaluation		
Possibilistic Methods	A	Sp	Q
π/min/g_1	.64	.685	.4384
π/Sp-0/g_1, π/Q-0/g_1, π/Sp-1/g_1	.64	.675	.432
π/A-0/g_1	.7054	.5868	.4139
π/trade/g_1, π/A-1/g_1, π/Q-1/g_1	.6976	.5845	.4078
π/max/g_1	.775	.45	.3487

Table 6.4: Assessment of Possibilistic methods by PEAPS

We see that the methods presented in Table 6.4 behave well in practice. Some even perform better than individual experts (compare with Table 6.1). Since the normalization step is not effective in this example, the accuracy coefficient increases from $\pi/\min/g_1$ to $\pi/\max/g_1$, and the specificity coefficients vary in the inverse order. The experts agree most of the time, which ensures that $\pi/\min/g_1$ has the best overall performance.

7 - The Probabilistic System EXCALIBR

7.1 - The Elicitation Step

In this system, experts' knowledge is obtained by means of either dichotomic tests (qualitative method) [4], or by the elicitation of a set of quantiles (quantitative method). The first elicitation method is outside the scope of this paper. Here we report an experiment using the second method with the elicitation of 3 quantiles: $(q_{5\%}, q_{50\%}, q_{95\%})$. Before the elicitation the analyst determines for each test-variable the choice between two scales: uniform or loguniform. The bounds of the variables depend on the particular subset of experts being analysed. Since the bounds of each variable may differ, once the bounds are determined the system uses a transformation that standardizes all the distributions to the [0,1] interval.

7.2 - The Assessment Step

EXCALIBR uses two basic functions to evaluate an expert: the *mean relative information measure* M(e) and the *calibration measure* C(e). The product of these two measures produces the global measure called the *unormalized global weight*; this measure also serves to determine the weight the expert will have in the combination step. These functions were analysed in Section 3.1, and correspond respectively to formulae (3.4), (3.5) and (3.6). In our example, each expert e_i supplied the set of 3 quantiles $q_{e_i} = (q_{5\%}, q_{50\%}, q_{95\%})$ as his estimation for each variable v_j. Thus, B = 3 and (q_1, q_2, q_3) denotes $(q_{5\%}, q_{50\%}, q_{95\%})$. We have 4 inter-quantile intervals $[q_{b-1}, q_b]$, $1 \leq b \leq 4 = B + 1$, where $q_0 = x_l$, $q_{B+1} = x_m$. To each interval $[q_{b-1}, q_b]$, $1 \leq b \leq 4$ we associate a probability mass p_b, where $p_1 = .05$, $p_2 = .45$, $p_3 = .45$ and $p_4 = .05$.

It is important to note that the informativeness index M depends on the size of the interval $[x_l,x_u]$. In EXCALIBR the domain of the variables changes for each group of experts, which means that the informativeness $M(e_i)$ may change depending on the groups in which e_i takes part. In order to deal with this problem, we have introduced in all experiments a dummy virtual expert, so imprecise as to guarantee that the bounds of each variable are determined by the assessments of the dummy expert. In our example, we created a dummy expert that forced the bounds of all variables to be $[x_l,x_u](v_j) = [.3, 8.7]$, $1 \le j \le 10$. The informativeness of expert e_1 using (3.3) and (3.4) is then

$$M(e_1) = \frac{1}{10} * 10\left(\ln 8.4 + \int_{.3}^{1} \frac{.05}{.7} \ln \frac{.05}{.7} + \int_{1}^{4} \frac{.45}{3} \ln \frac{.45}{3} + \int_{4}^{8} \frac{.45}{4} \ln \frac{.45}{4} + \int_{8}^{8.7} \frac{.05}{.7} \ln \frac{.45}{.7}\right) = .027 \ .$$

Each r_b in (3.5) stands for the distribution of the expert's successes on each quantile interval $[q_{b-1}, q_b]$. Since expert e_1 had 5 out of 10 variables with true values occurring in interval $[q_{5\%}, q_{50\%}] = [q_1, q_2]$ then $r(e_1) = .5$. The experts' frequencies $r(e_1)$ and $r(e_2)$ and the ideal frequence p_b are depicted in Figure 7.1.

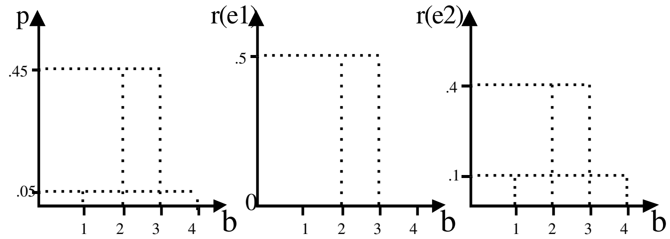

Figure 7.1: Illustration of realization frequences produced by EXCALIBR

The calibration index of expert e_1, with P = 1, is $C(e_1) = 1 - Chi^2{}_3(2 \cdot 10 \cdot (2 \cdot .5 \cdot \ln(.5 / .45)) \cdot 1) = 1 - Chi^2{}_3(20 \cdot .1053) = .55$. The quality of the expert e_1 is thus calculated using (3.6) which yields $W(e_1) = .027 \cdot .55 = .015$. In Table 7.1 we show the experts' evaluation in our example.

	Probabilistic Evaluation		
Experts	C	M	W
e_2	.82	.539	.442
e_1	.55	.027	.015

Table 7.1: Experts' Assessments in EXCALIBR

Note that in the probabilistic case expert e_2 is considered to be better calibrated than expert e_1, whereas in the possibilistic assessment, e_2 is considered to be less accurate than e_1 (see Table 6.1). These differences have not, however, interfered with the quality assessment; in both models expert e_2 is considered globally better than e_1.

7.3 - The Pooling Step

The analyst may choose between Bayesian updating or the classical model. In Bayesian updating a non-informative prior is used with a multinomial likelihood function; this method is not considered in the present paper. Opinion aggregation in the classical model is performed via weighted average. If F_i is the cumulative distribution function given by expert e_i for a given item v_j, then the Decision Maker's (DM) cummulative distribution for that item is defined by

$$F_{DM} = \frac{\sum \tau_i F_i}{\sum \tau_i}$$

where $\tau_i \geq 0$, $\sum_i \tau_i = 1$. It is also possible to weight individual items, but this approach will not be treated here.

Weights τ_i can either be calculated by the system or given by the user; the latter case will not be discussed here. The analyst can choose a group g_k of experts to which weights will be assigned, and then select one of the following weighting methods p/m/g_k given in the table below. In the first method,the system distributes equal weights among a group of experts specified by the analyst. The global weights method uses the expert's quality index and a significance level α acting as a selection threshold. This method p/opt/g_k leaves to the system the burden of determining α such that the calibration index C(DM),corresponding to the virtual expert DM, is maximal. The last method is based on the global weights method, but here the signification level is chosen by the analyst.

EXCALIBR Aggregation Methods: ***p/equ/g***

EXCALIBR Aggregation Methods: ***p/equ/g_k***, ***p/opt/g_k***, ***p/unn/g_k***
Equal Weights: ***p/equ/g_k***
$\tau_i = 1 / \lvert g_k \rvert$.
Global Weights with Optimization: ***p/opt/g_k***
$\tau_i' = W(e_i)$, if $C(e_i) > \alpha$ $= 0$ otherwise $t_i = \dfrac{\tau_i'}{\sum_{i\,:\,C(e_i) > \alpha} \tau_i'}$
Global Weights without Optimization: ***p/unn/g_k:***
Same as above with $0 \leq \alpha \leq 1$ supplied by the user .

Table 7.2

In our example, method p/opt/g_1 takes only expert e_2 into account (expert e_1 is rejected), and thus $\tau(e_2) = 1$. Using method p/unn/g_1 with $\alpha = 1$, the weight of each expert e_i corresponds to his normalized quality evaluation, i.e.. $\tau(e_i) = W(e_i) / \Sigma_i W(e_i)$. In method p/equ/$g_1$, each expert receives weight $\tau(e_i) = 1/2$. Table 7.3 summarizes the weights given to the experts by each pooling method. The results of the application of these methods to the estimations given by experts e_1 and e_2 are shown in Figure 7.2.

Normalized weights τ_i considering the experts in g_1		
p/equ	p/unn	p/opt
e_1 0.5	e_1 0.033	e_1 0.0
e_2 0.5	e_2 0.967	e_2 1.0

Table 7.3: Weights used by the probabilistic methods

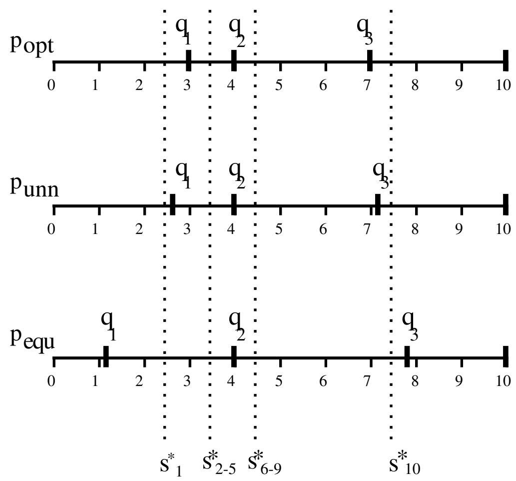

Figure 7.2: Results produced by the application of probabilistic pooling methods

7.4 - The Model Assessment Step

In EXCALIBR, the user chooses a group of experts and a pooling method and the system derives the weight of each expert, and the weight of the method itself, considered as a virtual expert[1]. This makes the evaluation of a group of pooling methods a bit cumbersome, since each result has to be collected separately.

---

[1] From the literature we have not been able to determine if the informativity of a pooling method m is calculated directly from the pdf yielded by the application of m, or if this pdf is first transformed into quantiles, which are then retransformed into a new pdf.

The evaluation of methods p/opt, p/unn and p/equ for group $g_1 = \{e_1, e_2\}$ is shown in Table 7.4. Here the best method is p/opt/g_1, which corresponds to taking only the assessments of expert e_2 into account.

	Probabilistic Evaluation		
Probabilistic Methods	C	M	W
p/opt/g_1	.82	.539	.442
p/unn/g_1	.82	.373	.305
p/equ/g_1	.55	.059	.032

Table 7.4: Assessment of probabilistic methods

7.5 - Cross Comparison of Expert Judgement Systems

In order to compare the possibilistic results in the probabilistic system, we used the multimodal equivalent of transformation T2 described in Section 5 and we extracted the quantiles ($q_{5\%}$, $q_{50\%}$, $q_{95\%}$) from the results of the conversion. Figure 7.3 shows the quantiles resulting from the possibilistic combination methods applied to the pieces of information given by experts e_1 and e_2. The evaluations of all experts —real or virtual— in the probabilistic framework are shown in Table 7.5.

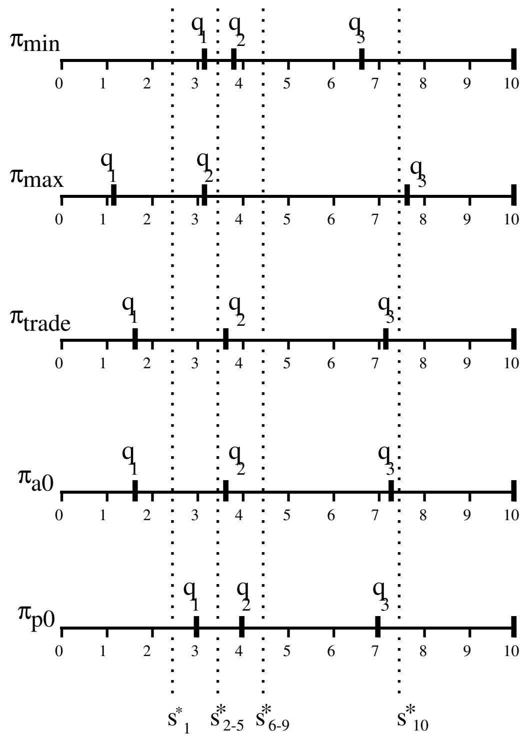

Figure 7.3: Quantiles derived from possibilistic pooling methods

Note that even if π/Sp-0, π/Q-0 and π/Sp-1 have the same possibility distribution as the one given by expert e_2, their transformation into probability distributions do not produce the same quantiles. This is due to the fact that the transformations probability —> possibility and possibility —> probability do not stand in inverse relation to one another. That's why the probabilistic evaluation of methods π/Sp-0, π/Q-0 and π/Sp-1 differs from the one of expert e_2.

Figure 7.4 shows the result of the conversion probability —> possibility related to the probabilistic pooling methods p/unn and p/equ. Method p/opt corresponds to using the pieces of information given by expert e_2 exclusively. The evaluation of all experts —real or virtual— in the possibilistic framework is shown in Table 7.6.

We can see that in general the possibilistic methods evaluate better than the probabilistic ones with respect to both systems. Note that with respect to the possibilistic evaluation, π/max is more imprecise than p/equ, whereas π/max is considered to be more informative than p/equ in the probabilistic framework. This inversion problem occurs because the conversion of π/max generates a very "thin" pdf which compares badly with the original p/equ, which did not suffer any "thinning" conversion.

	Probabilistic Evaluation		
All experts (real and virtual)	C	M	W
π/min/g_1	.82	.649	.531
π/Sp-0/g_1, π/Q-0/g_1, π/Sp-1/g_1	.82	.640	.524
e_2, p/opt/g_1	.82	.539	.442
p/unn/g_1	.82	.373	.305
π/trade/g_1, π/A-1/g_1, π/Q-1/g_1	.68	.189	.128
π/A-0/g_1	.68	.182	.123
p/equ/g_1	.55	.059	.032
e_1	.55	.027	.015
π/max/g_1	.02	.130	.002

Table 7.5: EXCALIBR's evaluation of all experts (real and virtual)

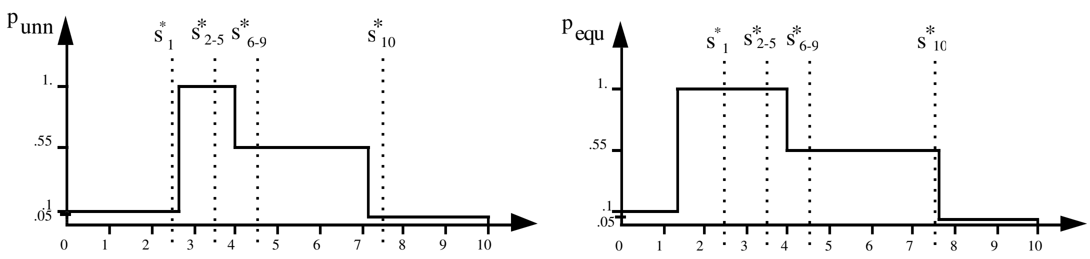

Figure 7.4: Possibility distributions derived from the conversion of probabilistic results

It is also important to note that the probability —> possibility transformation generates a very imprecise distribution. We are tempted to think that this conversion therefore, would cause the probabilistic methods to perform poorly in the possibilistic evaluation. However, if the conversion produces a decrease in precision, it also forces the accuracy to increase, counterbalancing the final evaluation. The conversions are also responsible for some inversions of evaluations between methods defined inside the same framework. This, is for instance the case with the possibilistic methods π/A-0 and π/trade. In the possibilistic evaluation π/A-0 is slightly more precise than π/trade, whereas it is a bit less informative than π/trade in the probabilistic evaluation. This can be explained as follows. Let us suppose that we have two possibility distributions π_1 and π_2. In order to evaluate them in the probabilistic framework, we convert them into pdf's p_1 and p_2. Then, from p_1 (p_2) we extract a set of quantiles q_1 (q_2). In order to evaluate the informativeness of q_1 and q_2 we generate new pdf's p_1' and p_2' by distributing the mass in a uniform way inside the inter-quantile intervals. The inversion is a consequence of the loss of infomation produced by the quantile extraction. This problem could be solved if we could evaluate the informativeness directly from p_1 and p_2, or if it a larger number of quantiles was extracted.

	Possibilistic Evaluation		
All experts (real or virtual)	A	Sp	Q
π/min/g_1	.64	.685	.4384
e_2, π/Sp-0/g_1, π/Q-0/g_1, π/Sp/1/g_1, p/opt/g_1	.64	.675	.432
π/A-0/g_1	.7054	.5868	.4139
p/unn/g_1	.635	.6452	.4097
π/trade/g_1, π/A-1/g_1, π/Q-1/g_1	.6976	.5845	.4078
p/equ/g_1	.775	.4901	.3798
e_1	.775	.675	.3565
π/max/g_1	.775	.45	.3487

Table 7.6: Evaluation of all experts (real and virtual) by PEAPS

8 - Real World Experiment

The DSM and ESTEC experiments are part of the project "Expert opinions in safety studies" developed by the Safety Science Group from the University of Technology of Delft (Netherlands), and the Industrial Safety Department from the Dutch Organization for the Applied Scientific Research TNO [2, 4]. We discuss here only the DSM experiment.

8.1 - The DSM Data

The DSM experiment represents an implementation of the classical combination model in probability theory, concerning the utilization of expert judgement in the determination of the causes of irregularities in flanged connections in a Dutch chemical plant. For this experiment 10 experts answered 14 questions; 8 questions had known values and were used to evaluate of the experts. The experts supplied, for each question x, a set of three quantiles ($q_{5\%}$, $q_{50\%}$, $q_{95\%}$), which summarised a subjective probability distribution concerning their opinions with respect to question x. The details of this experiment are described in [2].

The information given by the experts concerning the 8 test variables were used in a cross-evaluation experiment between the probabilistic (EXCALIBR) and possibilistic (PEARS) approaches to expert judgement. Table 8.1 depicts the original data for the 8 test variables.

In EXCALIBR the interval $[x_l,x_u]$, corresponding to the domain of a given variable, is always calculated as a function of the limit values found in the estimations given by the group of experts. Since the informativeness measure depends on this interval, the results of the evaluation cannot be compared on an equal basis if we are not able to compare all the experts at the same time. Due to screen visualisation problems, only a restricted number of experts (or methods) can be evaluated at the same time in EXCALIBR.

	v_1	v_2	v_3	v_4
	($q_{5\%}$,$q_{50\%}$,$q_{95\%}$)	($q_{5\%}$,$q_{50\%}$,$q_{95\%}$)	($q_{5\%}$,$q_{50\%}$,$q_{95\%}$)	($q_{5\%}$,$q_{50\%}$,$q_{95\%}$)
e_1	(10., 25., 75.)	(3., 15., 50.)	(50., 70., 80.)	(100., 4500., 9000.)
e_2	(20., 35. 50.)	(2., 5., 10.)	(75., 85., 95.)	(6500., 10000., 13500.)
e_3	(10., 20., 25.)	(5., 10., 15.)	(80., 88., 92.)	(1000., 5000., 15000.)
e_4	(5., 20., 50.)	(5., 10., 50.)	(70., 85., 100.)	(5000., 10000., 20000.)
e_5	(10., 30., 50.)	(5., 10., 20.)	(87., 92., 98.)	(1000., 1500., 2500.)
e_6	(10., 15., 20.)	(2., 4., 8.)	(70., 80., 90.)	(20000., 25000., 50000.)
e_7	(50., 60., 80.)	(4., 5., 8.)	(85., 90., 91.)	(3000., 5000., 8000.)
e_8	(50., 75., 100.)	(2., 5., 10.)	(84., 87., 90.)	(2500., 4000., 16000.)
e_9	(15., 20., 30.)	(1., 3., 6.)	(90., 94., 96.)	(1500., 2800., 3500.)
e_{10}	(15., 20., 30.)	(3., 7., 10.)	(70., 80., 90.)	(1000., 5000., 10000.)

Table 8.1.a: Original data: quantiles supplied by the experts for variables v_1 to v_4

	v_5	v_6	v_7	v_8
	$(q_{5\%},q_{50\%},q_{95\%})$	$(q_{5\%},q_{50\%},q_{95\%})$	$(q_{5\%},q_{50\%},q_{95\%})$	$(q_{5\%},q_{50\%},q_{95\%})$
e_1	(40000., 75000., 80000.)	(3., 15., 40.)	(10., 60., 100.)	(5., 14., 25.)
e_2	(1000000., 1500000., 2000000.)	(10., 20., 35.)	(25., 125., 250.)	(50., 70., 90.)
e_3	(65000., 75000. 85800.)	(10., 15., 25.)	(25., 125., 200.)	(90., 95., 99.)
e_4	(40000., 50000., 70000.)	(4., 8., 20.)	(63., 500., 1000.)	(30., 80., 95.)
e_5	(1100000., 1400000., 1700000.)	(5., 15., 30.)	(25., 125., 400.)	(80., 88., 92.)
e_6	(1500000., 1680000., 1750000.)	(15., 20., 25.)	(50., 100., 200.)	(83., 85., 87.)
e_7	(1200000., 1300000., 1600000.)	(15., 20., 35.)	(10., 100., 250.)	(86., 88., 92.)
e_8	(1200000., 1600000., 2000000.)	(5. 15. 40.)	(10., 50., 100.)	(75., 82., 90.)
e_9	(800000., 980000., 1000000.)	(15., 25., 40.)	(50., 100., 250.)	(82., 84., 87.)
e_{10}	(720000., 816000., 864000.)	(15., 20., 30.)	(3., 4., 10.)	(80., 93., 97.)

Table 8.1.b: Original data: quantiles supplied by the experts for variables v_5 to v_8.

In DSM each method in each mathematical model was tested with 3 groups of experts, chosen from the original set. Table 8.2 shows the real value of each variable, as well as the interval $[x_l,x_u]$, used by each system. In PEAPS, we fixed an interval $[x_l,x_u]$ for each variable, and in EXCALIBR we introduced a dummy virtual expert whose assessments determined the bounds of the variables. The use of different $[x_l,x_u]$ by each system does not influence the final values, in fact, we are interested only in the ranking produced in each system by the performance measures, and not in the numerical values of these measures. Figure 8.1 illustrates the quantiles supplied by the experts for the test-variables matched against the true value of these variables.

		$[x_l,x_u]$	
v_j	x_j*	EXCALIBR	PEAPS
v_1	51.	[1.e-13, 1143,95]	[0, 109.5]
v_2	3.	[1.e-13, 549.9]	[0, 54.9]
v_3	86.6	[1.e-13, 1121.52]	[45, 105.]
v_4	7000.	[1.e-13, 574639.]	[0, 54990.]
v_5	1500000.	[1.e-13, 2294200.]	[0, 2196000.]
v_6	15.	[1.e-13, 439.97]	[0, 43.7]
v_7	6.3	[1.e-13, 10999.99]	[0, 1099.7]
v_8	81.4	[1.e-13, 1132.995]	[0, 108.4]

Table 8.2: True values and domains of variables

The experiment was implemented with 3 groups of experts:

$g_1 = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8, e_9, e_{10}\}$
$g_2 = \{e_2, e_8\}$
$g_3 = \{e_2, e_8, e_7\}$

The mean conflict in each group was calculated taking the mean of the individual conflict for each variable. The individual conflict $\kappa_j = 1 - \mathfrak{J}_j$ is the complement of the degree of coherence in the group with respect to variable v_j, given by the Jacquard index. The mean conflict for groups g_1, g_2 and g_3 are:

$$\kappa(g_1) = .921, \qquad \kappa(g_2) = .369, \qquad \kappa(g_3) = .538$$

8.2 - DSM Treatment by EXCALIBR

Table 8.3 shows the experts evaluation (in decreasing order, for each index) according to EXCALIBR.

$C(e_i)$	$M(e_i)$	$W(e_i)$
e_2 .54	e_9 4.077	e_2 1.69800
e_8 .14	e_{10} 4.066	e_8 0.47838
e_{10} .02	e_3 3.915	e_{10} 0.08131
e_4 .02	e_7 3.817	e_4 0.06265
e_3 .001	e_6 3.737	e_3 0.00392
e_7 .001	e_5 3.457	e_7 0.00382
e_5 .001	e_1 3.452	e_5 0.00346
e_1 .001	e_8 3.417	e_1 0.00345
e_9 .0001	e_2 3.144	e_9 0.00041
e_6 .0001	e_4 3.132	e_6 0.00037

Table 8.3: Experts evaluation according to EXCALIBR

The relative order of the experts in each group with respect to the probabilistic criteria of calibration ($order_C$), mean informativeness ($order_M$), and expertise ($order_W$) are:

$order_C(g_1) = (e_2, e_8, e_7, \{e_4, e_{10}\}, \{e_1, e_3, e_5, e_6\}, e_9)$
$order_M(g_1) = (e_9, e_{10}, e_3, e_7, e_6, e_5, e_1, e_8, e_2, e_4)$
$order_W(g_1) = (e_2, e_8, e_{10}, e_4, e_3, e_7, e_5, e_1, e_9, e_6)$

$order_C(g_2) = (e_2, e_8)$
$order_M(g_2) = (e_8, e_2)$

$order_W(g_2) = (e_2, e_8)$

$order_C(g_3) = (e_2, e_8, e_7)$
$order_M(g_3) = (e_7, e_8, e_2)$
$order_W(g_3) = (e_2, e_8, e_7)$

The probabilistic pooling methods p/equ/g_k, p/unn/g_k and p/opt/g_k, $1 \leq k \leq 3$, were applied to the data in Table 8.1; the normalized weights determined by the application of these methods are shown in Table 8.4.

Normalized weights τ_i considering the experts in g_1		
p/equ	p/unn	p/opt
e_1 0.1	e_2 0.716	e_2 0.7464
e_2 0.1	e_8 0.201	e_8 0.2535
e_3 0.1	e_{10} 0.034	e_1 0.0
e_4 0.1	e_4 0.026	e_3 0.0
e_5 0.1	e_7 0.016	e_4 0.0
e_6 0.1	e_3 0.00165	e_5 0.0
e_7 0.1	e_5 0.001459	e_6 0.0
e_8 0.1	e_1 0.001455	e_7 0.0
e_9 0.1	e_9 0.000172	e_9 0.0
e_{10} 0.1	e_6 0.000156	e_{10} 0.0

Normalized weights τ_i considering the experts in g_2		
p/equ	p/unn	p/opt
e_2 0.5	e_2 0.716	e_2 1.0
e_8 0.5	e_8 0.201	e_8 0.0

Normalized weights τ_i considering the experts in g_3		
p/equ	p/unn	p/opt
e_2 0.33	e_2 0.7674	e_2 1.0
e_8 0.33	e_8 0.2154	e_8 0.0
e_7 0.33	e_7 0.01714	e_7 0.0

Table 8.4: Normalized weights of experts used by the probabilistic fusion methods

8.3 - Treatment of the DSM data by PEAPS

The values in Table 8.1 have been transformed into possibility distributions which have then been evaluated in the possibilistic model. Table 6.5 shows the experts evaluation according to PEAPS. The standard deviation relative to the performance measures are found in Table 8.6.

$A(e_i)$	$Sp(e_i)$	$Q(e_i)$
e_8 .8312	e_9 .8409	e_8 .5782
e_2 .7687	e_7 .8274	e_2 .5085
e_7 .4875	e_3 .8264	e_7 .3984
e_4 .475	e_{10} .8208	e_{10} .3373
e_{10} .425	e_6 .7920	e_6 .3132
e_6 .4187	e_5 .7446	e_4 .2946
e_5 .4187	e_8 .7243	e_5 .2892
e_1 .4125	e_1 .6905	e_3 .2313
e_3 .3	e_2 .6800	e_1 .2223
e_9 .1812	e_4 .6249	e_9 .1561

Table 8.5: Experts evaluation according to PEAPS

$\sigma_A(e_i)$	$\sigma_{Sp}(e_i)$	$\sigma_Q(e_i)$
e_8 .3348	e_9 .1181	e_8 .2656
e_2 .4284	e_7 .0857	e_2 .3039
e_7 .3833	e_3 .0706	e_7 .3026
e_4 .4652	e_{10} .1043	e_{10} .2412
e_{10} .3338	e_6 .1169	e_6 .3588
e_6 .4817	e_5 .1296	e_4 .2971
e_5 .4174	e_8 .1774	e_5 .2687
e_1 .4155	e_1 .2104	e_3 .2554
e_3 .3505	e_2 .1314	e_1 .1854
e_9 .3315	e_4 .1907	e_9 .2957

Table 8.6: Standard deviation of the precision, accuracy and quality indices

The experts relative order with respect to the criteria of global precision ($order_{Sp}$), global accuracy ($order_A$), and global quality ($order_Q$) are :

$order_{Sp}(g_1) = (e_9, e_7, e_3, e_{10}, e_6, e_5, e_8, e_1, e_2, e_4)$
$order_A(g_1) = (e_8, e_2, e_7, e_4, e_{10}, e_6, e_5, e_1, e_3, e_9)$

$order_Q(g_1) = (e_8, e_2, e_7, e_{10}, e_6, e_4, e_5, e_3, e_1, e_9)$

$order_{Sp}(g_2) = (e_8, e_2)$
$order_A(g_2) = (e_8, e_2)$
$order_Q(g_2) = (e_8, e_2)$

$order_{Sp}(g_3) = (e_7, e_8, e_2)$
$order_A(g_3) = (e_8, e_2, e_7)$
$order_Q(g_3) = (e_8, e_2, e_7)$

Using the information given in Tables 8.5 and 8.6 and taking 0 and 1 as values of ρ, the experts have been classified in each group g_k into homogeneous subsets:

$order_{Sp\text{-}0}(g_1) = (\{e_9\}, \{e_7\}, \{e_3\}, \{e_{10}\}, \{e_6\}, \{e_5\}, \{e_8\}, \{e_1\}, \{e_2\}, \{e_4\})$
$order_{A\text{-}0}(g_1) = (\{e_8\}, \{e_2\}, \{e_7\}, \{e_4\}, \{e_{10}\}, \{e_6, e_5\}, \{e_1\}, \{e_3\}, \{e_9\})$
$order_{Q\text{-}0}(g_1) = (\{e_8\}, \{e_2\}, \{e_7\}, \{e_{10}\}, \{e_6\}, \{e_4\}, \{e_5\}, \{e_3\}, \{e_1\}, \{e_9\})$
$order_{A\text{-}1}(g_1) = order_{Sp\text{-}1}(g_1) = order_{Q\text{-}1}(g_1) = (\{e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8, e_9, e_{10}\})$

$order_{A\text{-}0}(g_2) = order_{Sp\text{-}0}(g_2) = order_{Q\text{-}0}(g_2) = (\{e_8\}, \{e_2\})$
$order_{A\text{-}1}(g_2) = order_{Sp\text{-}1}(g_2) = order_{Q\text{-}1}(g_2) = (\{e_2, e_8\})$

$order_{Sp\text{-}0}(g_3) = (\{e_7\}, \{e_8\}, \{e_2\})$
$order_{A\text{-}0}(g_3) = order_{Q\text{-}0}(g_3) = (\{e_8\}, \{e_2\}, \{e_7\})$
$order_{A\text{-}1}(g_3) = order_{Sp\text{-}1}(g_3) = order_{Q\text{-}1}(g_3) = (\{e_2, e_7, e_8\})$

The possibility distributions have been pooled using the methods $\pi/min/g_k$, $\pi/max/g_k$, $\pi/trade/g_k$, $\pi/A\text{-}\sigma/g_k$, $\pi/Sp\text{-}\sigma/g_k$, $\pi/Q\text{-}\sigma/g_k$, with the standard deviation values $\rho = 0$ and $\rho = 1$.

8.4 - Cross Evaluation on DSM

The results of the application of probabilistic and possibilistic pooling methods were evaluated in both systems, after all necessary conversions. Table 8.7 shows the evaluation of all experts and methods according to EXCALIBR. Table 8.8 shows the evaluation according to PEAPS of all the experts and methods, except the probabilistic method p/unn. Since the results are so numerous, we illustrate here only part of the pooling results; complete data can be found in [28]. Figure 8 shows the quantiles given by all the experts for variables v_1 to v_8 and also the results, in terms of quantiles, of the pooling methods $\pi/min/g_2$, $\pi/A\text{-}0/g_2$, $\pi/max/g_3$, and $p/equ/g_2$.

	Probabilistic Evaluation

All experts (real and virtual)	C	M	W
1 π/min/g_2	.54	3.953	2.1344
2 π/A-1/g_2, π/Sp-1/g_2, π/Q-1/g_2	.66	3.093	2.0410
3 π/A-1/g_3, π/Sp-1/g_3, π/Q-1/g_3	.66	3.057	2.0173
4 π/max/g_2	.66	3.005	1.9831
5 π/max/g_3	.66	3.002	1.9811
6 p/equ/g_2	.66	2.999	1.9790
7 p/unn/g_2	.66	2.994	1.9763
8 p/unn/g_3	.66	2.994	1.9761
9 π/trade/g_2	.66	2.989	1.9728
10 p/opt/g_1	.66	2.981	1.9677
11 π/trade/g_3	.66	2.948	1.9458
12 p/unn/g_1	.64	2.788	1.7840
13 e_2, p/opt/g_2, p/opt/g_3	.54	3.144	1.6980
14 π/A-1/g_1, π/Sp-1/g_1, π/Q-1/g_1	.66	2.296	1.5153
15 π/max/g_1	.66	2.294	1.5140
16 π/trade/g_1	.66	2.274	1.5007
17 p/equ/g_1	.53	2.579	1.3667
18 p/equ/g_3	.42	3.043	1.2781
19 π/min/g_1	.23	2.786	0.6408
20 π/A-0/g_2, π/Sp-0/g_2, π/Q-0/g_2	.17	3.427	0.5826
21 e_8	.14	3.417	0.4783
22 π/A-0/g_3, π/Q-0/g_3	.06	3.494	0.2096
23 π/A-0/g_1	.06	3.019	0.1811
24 π/Sp-0/g_3	.03	3.624	0.1087
25 e_{10}	.02	4.066	0.0813
26 π/Sp-0/g_1	.02	3.262	0.0652
27 e_4	.02	3.132	0.0626
28 π/min/g_3	.01	4.071	0.0407
29 π/Q-0/g_1	.01	3.032	0.0303
30 e_3	.001	3.915	0.0039
31 e_7	.001	3.817	0.0038
32 e_5	.001	3.457	0.0034
33 e_1	.001	3.452	0.0034
34 e_9	.0001	4.077	0.0004
35 e_6	.0001	3.737	0.0003

Table 8.7: Global evaluation according to EXCALIBR

	Possibilistic Evaluation		
All experts (real and virtual)	A	F	Q
1 π/A-0/g_2, π/Sp-0/g_2, π/Q-0/g_2	.8872	.7245	.6233
2 π/max/g_3	1.	.5921	.5921
3 π/min/g_2	.7687	.7952	.5822
4 e_8	.8312	.7243	.5782
5 p/equ/g_2	.8312	.6807	.5719
6 π/A-1/g_2, π/Sp-1/g_2, π/Q-1/g_2	.8775	.6598	.5617
7 π/trade/g_2	.8775	.6387	.5407
8 π/A-0/g_1	.7483	.7272	.5402
9 p/opt/g_1	.775	.6807	.5268
10 e_2, p/opt/g_3	.7687	.6800	.5085
11 π/max/g_2	.8875	.5952	.5059
12 π/Q-0/g_1	.7212	.7062	.4934
13 π/A-1/g_3, π/Sp-1/g_3, π/Q-1/g_3	.7640	.6593	.4893
14 p/equ/g_3	.7187	.6902	.4787
15 π/A-0/g_3, π/Q-0/g_3	.6390	.7698	.4734
16 π/trade/g_3	.7651	.6410	.4723
17 e_7	.4875	.8274	.3984
18 p/equ/g_1	.7562	.5285	.3939
19 π/Sp-0/g_3	.4522	.8248	.3732
20 e_{10}	.425	.8208	.3373
21 e_6	.4187	.7920	.3132
22 π/A-1/g_1, π/Sp-1/g_1, π/Q-1/g_1	.9846	.3105	.3049
23 π/max/g_1	1.	.3018	.3018
24 e_4	.475	.6249	.2946
25 π/min/g_3	.3356	.8603	.2903
26 e_5	.4187	.7446	.2892
27 π/trade/g_1	.9855	.2903	.2853
28 e_3	.4125	.8264	.2313
29 e_1	.3	.6905	.2223
30 π/Sp-0/g_1	.2003	.8027	.1622
31 e_9	.1812	.8409	.1561
32 π/min/g_1	.7840	.2741	.0903

Table 8.8: Global evaluation according to PEAPS

From Tables 8.7 and 8.8 we can see that, given the initial data, the pooling methods yield better overall performance than those based individual experts. These results justify the use of expert judgement pooling techniques in the DSM experiment. In the following we analyze the results in more detail, and formulate a conclusion for the experiment.

8.5 - Discussion

1) The expert rankings with respect to the *informativeness* criterion in EXCALIBR ($order_M$), and the *precision* criterion in PEAPS ($order_{Sp}$) are very similar. Indeed, we can see that the five most precise experts according to PEAPS are the same as those considered to be the most informative ones according to EXCALIBR. In particular, the expert considered the most precise according to PEAPS is also considered the most informative by EXCALIBR, and the two most imprecise experts according to PEAPS are also the least informative ones according to EXCALIBR. This result demonstrates in practice that the possibilistic precision index and the probabilistic informativeness index are similar.

2) The expert rankings with respect to the *calibration* criterion in EXCALIBR ($order_C$), and the *accuracy* criterion in PEAPS ($order_A$), are very similar in most cases. This shows that the DSM data does not contain such strange cases as these pointed out in Section 3.1, where extremely inaccurate experts are considered to be well-calibrated (and vice-versa). This means that the well-calibrated experts have most of their realizations occurring in the central inter-quantile intervals in a balanced way (the experts are not biased). We note in particular that the least accurate expert is also the least calibrated one, and that the three most accurate experts are also the best calibrated ones. Moreover, the calibration criteron would not appear to discriminate as well as the accuracy one. It is possible, for instance, to distinguish experts e_1 and e_4 with respect to the accuracy criterion (as well as experts e_1, e_3, e_5 and e_6), but not with respect to the calibration criterion.

3) The expert rankings with respect to the *expertise* criterion in EXCALIBR ($order_W$), and the *quality* criterion in PEAPS ($order_Q$), do not present many differences. In particular the two best experts according to PEAPS are the also the two best ones according to EXCALIBR; expert e_9 is considered to be the worst one according to both systems. An examination of Figure 8.1 verifies that expert e_9 is indeed inferior in relation to other experts. This shows that both the accuracy and the calibration indices are capable of detecting extremely bad cases. The most noticeable difference between the rankings $order_W$ and $order_Q$, has to do with expert e_3, who is precise mostly when he is inaccurate. EXCALIBR is not capable of recognizing this phenomenon because the expertise criterion is based on the global evaluations of informativeness and calibration, and not on the individual evaluations. This shows in practice

that we can have an expert who is considered to be informative and well-calibrated, but who is not precise and accurate in the same instances.

4) With respect to the pooling methods, we verify again that the precision and informativeness orderings present a strong similarity. Indeed, if we divide the "virtual" experts into two groups according to their precision, we can see that the group which has the best performance in terms of this criterion contain the most informative experts according to EXCALIBR. The differences in each group result partly from the loss of information induced by the transformations probability —> quantiles —> probability.

5) The most imprecise and the least informative virtual experts are those corresponding to the application of pooling methods on group g_1, the complete set of experts. This reflects the large conflict that appears with respect to most variables due to the large size of the group. In the probabilistic case, the weighted mean operations treat the conflict by generating "flat" distributions, which are penalised by EXCALIBR. In the possibilistic case, on the other hand, the conflict is treated by a normalization operation. This may considerably augment the imprecision of a "precise" method where conflict is produced, as compared to a less "imprecise" method where no conflict is produced. For instance, the application of the maximum rule on group g_1 yields results which are globally slightly more precise than those produced by the minimum rule applied on the same group (see Table 8.8). A formal classification of some methods according to their performance can be established. For instance, considering any given group of estimations, the minimum method is more precise and less accurate than the maximum method if a normalization step is not necessary. We can also formally characterize the performance of nested subsets of a group of estimations: a group will produce more precise and less accurate estimations than a subset of this group if a normalization step is not necessary.

Method π/max has no need for normalization but is still very imprecise for group g_1, since for most variables the support of the estimation produced by this method is very close to $[x_l,x_u]$. The only method which produces reasonably precise results for g_1 is method π/Sp-0; this is explained by the fact that in this type of method, the final distribution is very much influenced by the first elements in the corresponding order. Since here the precision order is used, it is the most precise experts who influence the final result, thereby generating very precise estimations. Note that in contrast, since the most precise experts are also the least accurate, π/Sp-0/g_1 is considered very inaccurate and its evaluation with respect to the quality criterion is one of the lowest.

6) In this experiment, the possibilistic fusion methods look superior to the probabilistic ones, according to both the possibilistic and the probabilistic quality evaluation. However, the

orderings given by the systems are not the same. In the first place, we note that the best methods (with respect to the expertise index) according to EXCALIBR are also considered good by PEAPS (with respect to the quality index). This is especially true for groups g_2 and g_3, which contain only highly-evaluated experts. For these groups the phenomenon described above, where a sometimes imprecise and otherwise inaccurate expert could be considered as informative and well-calibrated, does not occur. The best methods according to PEAPS are in general also ranked high by EXCALIBR (with the noticeable exception of π/A-0/g_2, which will be examined in detail below). This is particularly true when we consider the groups composed exclusively of those experts who are undoubtedly trustworthy.

7) The most startling result in the whole experiment is that the best expert according to PEAPS —π/A-0/g_2— is very poorly evaluated by EXCALIBR. However, when we examine Figure 8.1, comparing the quantiles corresponding to π/A-0/g_2 with those supplied by the best expert according to EXCALIBR (π/min/g_2), we see that the results of these methods are practically the same relative to calibration in EXCALIBR. For variables v_2, v_3, v_5 and v_8, the results of π/A-0/g_2 and π/min/g_2 are exactly the same, while for variables v_1, v_4, v_6 and v_7, π/A-0/g_2 yields more imprecise results and is duly penalized by the informativeness index (see Table 8.7). This is consistent with the possibilistic evaluation. The most important difference in performance between these methods lies in the calibration. Table 8.7 shows that π/min/g_2 is considered to be three times better calibrated than π/A-0/g_2, yet, as can be seen in Figure 8.1 the only difference between these methods from the calibration point of view concerns variable v_1. Indeed, the values of variables v_2 to v_8 lie in the same inter-quantile intervals for each method: 1 value in $[x_l,q_{5\%}]$, 4 values in $[q_{5\%},q_{50\%}]$, two values in $[q_{50\%},q_{95\%}]$, and no values in $[q_{95\%},x_u]$. Variable v_1 has its value in the interval $[x_l,q_{5\%}]$ with respect to π/min/g_2, and in $[q_{95\%},x_u]$ with respect to π/A-0/g_2. As a consequence, π/min/g_2 has 1 value in $[x_l,q_{5\%}]$ and 1 value in $[q_{95\%},x_u]$, whereas π/A-0/g_2 has three values in $[x_l,q_{5\%}]$, and no value in $[q_{95\%},x_u]$. The calibration will thus reward π/min/g_2 and penalize π/A-0/g_2. However, neither π/min/g_2 nor π/A-0/g_2 have been very accurrate (see Figure 8.1). It is thus very disturbing that π/min/g_2 should be considered better than numerous others while π/A-0/g_2 should be considered worse. This case illustrates the sensitivity of the calibration index with respect to small variations in the data in the probabilistic method. We see here that a single variable may completely alter the ranking of two experts (real or virtual). Moreover, due to a single variable, a very accurate and precise method may be ranked worse than some methods which are undoubtedly inferior to it.

8) We note that possibilistic methods yielding different results are sometimes considered as equally calibrated according to EXCALIBR. This problem is a consequence of the fact that the calibration index takes only part of the information into account, but also of the fact that information is lost in the transformations possibility —> probability —> quantiles.

9) The accuracy of method π/max/g_k calculated by PEAPS, with respect to a variable v_i is equal to 1 each time that the realization of v_i is found to be in the core of at least one of the distributions furnished by the experts in group g_k. The global evaluation is maximal, however, only when the individual evaluations are maximal for all test variables. The examination of Table 8.8 shows that methods π/max/g_1 and π/max/g_3 obtain maximal values in global accuracy.

10) It is interesting to investigate why π/A-0/g_2 is considered to be better than π/min/g_2 by the possibilistic evaluation. Let us recall that our asymmetric methods (for instance π/A-0), take the minimum of the estimations when there is no conflict, and favor the most reliable source (according to a given ordering) otherwise. Method π/A-0 is practically the same as π/min for group g_2, since e_2 and e_8 yield non conflicting distributions; π/A-0/g_2 is only slightly less precise than π/min/g_2. In this example the difference in classification between these methods is caused by the evaluation of variable v_1 for which method π/min/g_2 has a very low accuracy. Method π/A-0/g_2 yields a distribution for v_1 which is more accurate than π/min/g_2 because the conflict in group g_2 is very significant and because method π/A-0/g_2 favors expert e_8's estimations (e_8 is the most accurate expert) up to the conflict level. This shows that dissymetric methods, which are optimistic in case of agreement and conservative in case of disagreement, can be efficient in practice. Moreover, it is important to note that, contrary to the probabilistic approach, methods π/A-0/g_2 and π/min/g_2 are both well-classified relative to the possibilistic quality index.

Based on the results shown here and the discussion above, we can see that taking the possibilistic pooling methods as good virtual experts is justified. In particular, methods π/A-0/g_2, and π/min/g_2 should be preferred to π/max/g_3, since the latter method exhibits good performance due to particularities in the probability/possibility transformations. Our choices are also justified when we analyze Table 8.5. We can see there that e_8 and e_2 are the only experts who have an evaluation higher than .5 for both the precision and accuracy criteria. Moreover, from Figure 8.1 we can see that if we considered an answer such that $q_{5\%} < x^* < q_{95\%}$ as *correct*, then expert e_8 gave wrong assessments only once (variable v_7), and e_2 only twice (variables v_1 and v_7), whereas all the other experts are wrong either 4 or 5 times. Therefore, only experts e_8 and e_2 should be retained for subsequent assessments in the field.

Conclusion

A procedure for processing human-originated information has been devised, based on possibility theory and implemented on a computer [27]. This procedure performs the elicitation of seed variables by experts and evaluates their performance, applies several pooling methods whose utilities depend on the results of the expert evaluation, compares the performances of the pooling methods on seed variables, and finally applies the best one found to the analysis of unknown parameter values supplied by the experts. The study of a real-world experiment allowed us to verify in practice the applicability of the possibilistic approach in the expert judgement domain relative to the evaluation and the pooling methods. The possibilistic evaluation methods appears to be more discriminating than those used in the probabilistic approach. This difference has already been verified at a conceptual level, and the experience with the DSM data seems to confirm it in practice. The possibilistic evaluation does not overload the system analyst, can easily be checked by the experts, and does not lead to incoherences.

Relative to pooling, the possibilistic approach is richer than the probabilistic one, and presents less difficulties relative to a possible dependencies among the sources. We have seen that, in practice, with only a small set of methods we can find satisfying ways of pooling expert opinions. In the particular experiment reported here, this was confirmed by both the possibilistic and the probabilistic evaluations. The only aspect involved in the use of expert judgement which was not treated here is the elicitation of the experts' knowledge. We believe, however, that this aspect represents the strongest advantage of the use possibility theory, it lets the expert express his knowledge in a more natural way than those used in the probabilistic approach. Moreover, it allows the data to be used untouched in all phases of the expert judgement process, contrary to the probabilistic approach studied here, where the data is constantly transformed into pdf's and quantiles. The possibilistic approach seems to behaves well in all other aspects of the process (pooling and evaluation), and so it is thus expected that possibilistic elicitation in practical experiments will confirm the utility of possibility theory in expert judgement pooling systems. Since the experiments were carried out, new pooling methods have been found. Especially, a trade-off rule not based on weighted arithmetic average is described in [16]. This rule is adaptive (depending on the amount of conflict) and applies to more than two sources. It should be added to PEAPS in the future.

Of course the above results are in some sense preliminary and partial and should not be taken to cast any discredit on the probabilistic approach whose results are already satisfactory. Actually, most of the concepts developed in the possibilistic approach were directly inspired by Cooke's methodology which appears more convincing that most Bayesian techniques. What we suggest here is that this methodology can be applied outside the probabilistic framework with equal success. Our main claim is that the possibilistic framework is more flexible than a pure probabilistic one for expressing expert opinions and pooling them. Given that possibility theory is less developed than probability theory, some of the pooling methods used here remain somewhat ad hoc, and much work remain to be done before a rigorous justification of the possibilistic method, similar to the one that already exists in the probabilistic setting, is fully developed.

Acknowledgements

The authors wish to thank Roger Cooke whose works have inspired the PEAPS approach, for many discussions on this paper. The referees also deserve many thanks for carefully checking the manuscript.

References

1. R.M. Cooke. "Uncertainty in risk assessment: A probabilist's manifesto," *Reliability Eng. and Syst. Safety* 23, 1988, 277-283.
2. R.M. Cooke, *Experts Opinions in Safety Studies — Vol. 5, Case 1: DSM Case Study*, Department of Mathematics, TU Delft, The Netherland, 1989.
3. R.M. Cooke, *Experts Opinions in Safety Studies — Vol. 5, Case 2: ESTEC Case Study*, Department of Mathematics, TU Delft, The Netherland, 1989.
4. R.M. Cooke, *Experts in Uncertainty*, Oxford University Press, 1991.
5. D. Dubois and M.C. Jaulent, "A general approach to parameter evaluation in fuzzy digital pictures," *Pattern Recognition Letters* 6, 1987, 251-259.
6. D. Dubois and H.W. Kalfsbeek, "Elicitation, assessment and pooling of expert judgement using possibility theory," in *Proc. of the 8th Inter. Congress of Cybernetics and Systems* (C.N. Manikopoulos, ed.), New Jersey Institute of Technology Press, Newark, NJ, 1990, 360-367.
7. D. Dubois and H. Prade, *Fuzzy Sets and Systems: Theory and Applications*, Academic Press, New York, 1980.
8. D. Dubois and H. Prade, "Une approche ensembliste de la combination d'informations incertaines," *Revue d'Intelligence Artificielle* 1(4), 1987, 23-42.
9. D. Dubois and H. Prade, "The mean value of a fuzzy number," *Fuzzy Sets and Systems* 24, 1987, 279-300.
10. D. Dubois and H. Prade, "Representation and combination of uncertainty with belief functions and possibility measures," *Computational Intelligence* 4, 1988, 244-264.
11. D. Dubois and H. Prade, *Possibility Theory*, Plenum Press, New York, 1988.

12. D. Dubois and H. Prade, "Default reasoning and possibility theory," *Artificial Intelligence* 35, 1988, 243-257.
13. D. Dubois and H. Prade, "When upper probabilities are possibility measures," *Fuzzy Sets and Systems* 49, 1992, 65-74.
14. D. Dubois and H. Prade, "Combination of fuzzy information in the framework of possibility theory," in *Data Fusion in Robotics and Machine Intelligence* (M. Abidi, R. Gonzalez, eds.), Academic Press, New York, 1992, 481-505.
15. D. Dubois and H. Prade, "On the relevance of non-standard theories of uncertainty in modeling and pooling expert opinions," *Reliability Eng. and Syst. Safety* 36, 1992, 95-107.
16. D. Dubois and H. Prade, "Possibility theory and data fusion in poorly informed environments," *Control Eng. Practice* 2(5), 1994, 811-823.
17. D. Dubois, H. Prade and C. Testemale, "Weighted fuzzy pattern matching," *Fuzzy Sets and Systems* 28, 1988, 313-331.
18. D. Dubois, H. Prade and S. Sandri, "On possibility/probability transformations," in *Fuzzy Logic: State of the Art* (R. Lowen, M. Roubens, eds.), Kluwer Academic Publ., Dordrecht, 1993, 103-112.
19. S. French, "Group consensus probability distributions: A critical survey," in *Bayesian Statistics 2* (J.M. Bernardo et al., eds.), Elsevier, 1985, 183-202.
20. J. Gebhart and R. Kruse, "A possibilistic interpretation of fuzzy sets by the context model," in *Proc. of the 1st Inter. Conf. on Fuzzy Systems (FUZZ-IEEE'92)*, San Diego, CA, March 1992, 1089-1096.
21. I.R. Goodman and H.T. Nguyen, *Uncertainty Models for Knowledge Based Systems*, North-Holland, Amsterdam, 1985.
22. H.W. Kalfsbeek, "Elicitation, assessment and pooling of expert judgement using possibility theory," Working Paper PER1829/89 (revised version), Joint Research Center of the EEC, Ispra, Italy, February 1990.
23. G. Klir and T. Folger, *Fuzzy Sets, Uncertainty and Information*, Prentice Hall, Englewood Cliffs, NJ, 1988.
24. K.J. McConway, "Marginalization and linear opinion pools," *J. Ann. Stat. Ass.* 76, 1981, 410-418.
25. P.A. Morris, "Combining expert judgements, a Bayesian approach," *Management Science* 23, 1977, 679-692.
26. A. Mosleh and G. Apostolakis, "Models for the use of expert opinions," in *Low Probability / High Consequence Risk Analysis* (R.A. Waller, V.T. Covello, eds.), Plenum Press, New York, 1984.
27. S. Sandri "Fuzzy sets and possibility theory for RPES development," Final Report, Contract n° 3309-87-12, Joint Research Center of the EEC, Ispra, Italy and IRIT, Université P. Sabatier, Toulouse, France, 1990.

28. S. Sandri, "La Combinaison de l'information incertaine et ses aspects algorithmiques," Doctoral Dissertation, Université Paul Sabatier, Toulouse, France, 1991.

29. S. Sandri, A. Besi, D. Dubois, G. Mancini, H. Prade and C. Testemale, "Data fusion problems in an intelligent data base interface," in *Reliability Data Collection and Use in Risk and Availability Assessment* (U. Colombari, ed.), Springer Verlag, Berlin, 1989, 655-670.

30. G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, NJ, 1976.

31. Ph. Smets, "Constructing the pignistic probability function in a context of uncertainty," in *Uncertainty in AI, Vol. 5* (M. Henrion et al., eds.), North-Holland, Amsterdam, 1990, 29-40.

32. C. Wagner and K. Lehrer, *Rational Consensus in Science and Society*, D. Reidel, Dordercht, 1981.

33. R.L. Winkler, "The consensus of subjective probability distributions," *Management Science* 15, 1968, B61-B75.

34. R.L. Winkler, "On "good probability appraisers," in *Bayesian Inference and Decision Techniques* (P. Goel, A. Zellner, eds.), Elsevier Science Publ., 1986, 265-290.

35. J.S. Wu, G.E. Apostolakis and D. Okrent, "Uncertainty in system analysis: Probabilistic versus non-probabilistic theories," *Reliability Eng. and Syst. Safety* 30, 1990, 163-181.

36. R.R. Yager, "Aggregating evidence using quantified statements," *Information Sciences* 36, 1985, 179-206.

37. R.R. Yager, "Using approximate reasoning to represent default knowledge," *Artificial Intelligence* 31, 1987, 99-112.

38. L.A. Zadeh, "Probability measures of fuzzy events," *J. Math. Anal. Appl.* 23, 1968, 421-427.

39. L.A. Zadeh, "Fuzzy sets as a basis for a theory of possibility," *Fuzzy Sets and Systems* 1, 1978, 3-28.

40. L.A. Zadeh, "Fuzzy sets and information granularity," in *Advances in Fuzzy Set Theory and Applications* (M.M. Gupta et al., eds.), North-Holland, Amsterdam, 1979, 3-18.