WPS 3201

A Literature Review on the Use of Expert Opinion in Probabilistic Risk Analysis

Fumika Ouchi

World Bank Policy Research Working Paper 3201, February 2004

A Literature Review on the Use of Expert Opinion in Probabilistic Risk Analysis

Introduction

Risk assessment is part of the decision-making process in many fields of discipline, such as engineering, public health, environment, program management, regulatory policy, and finance. There has been considerable debate over the philosophical and methodological treatment of risk in the past few decades; and yet the discussion continues today on many basic issues ranging from the definition and classification of risk, to approaches to its assessment and management, to societal risk acceptance (Klinke and Renn 2002; Hatfield and Hipel 2002; Jenni and Fischhoff 1997; The U.S. Presidential/Congressional Commission on Risk Assessment and Risk Management 1997; Skjong and Eknes 2002).

Bedford and Cooke (1996) characterize risk with two particular elements: hazard (a source of danger) and uncertainty (quantified by probability). Probabilistic risk analysis (PRA) specifically deals with events represented by low probabilities of occurring with a high level of consequence. Expert opinion is one of the key research areas in PRA. The use of expert judgment is critical, and often inevitable, when there are no empirical data or information available on the variables of interest. In reliability analysis, for example, expert opinion has been extensively examined by Singpurwalla and others (Campodonico and Singpurwalla 1995; van Noortwijk et al. 1992; Singpurwalla 1988; Singpurwalla and Song 1988; Lindley and Singpurwalla 1986).

Because of the complex, subjective nature of expert opinion, there has been no formally established methodology for treating expert judgment. In recent years, there has been increasing effort in establishing a more systematic approach to eliciting expert opinion (Winkler et al. 1992; von Winterfeld 1989). Cooke and Goossens (2000) provide formal protocols, comprehensive procedures and guidelines on the elicitation process and handling of such data in uncertainty analysis.

This paper presents a review of the literature on the use of expert opinion, specifically, the approaches to aggregating different experts' probability assessments. There are numerous studies on how to reconcile multiple experts' assessments, many of which are extensively reviewed by Cooke (1991), Bedford and Cooke (1996), and Clemen and Winkler (1999). Genest and Zidek (1986) also provide a useful annotated bibliography of more than 90 studies on this subject. Drawing results from these reviews, the present paper attempts to provide a short summary of key aggregation methods and discussions. The paper also makes reference to new approaches proposed in recent articles on the subject.

Methods for Eliciting and Aggregating Expert Opinion

Clemen and Winkler (1999) classify the elicitation and aggregation processes of expert assessments in two groups: mathematical and behavioral approaches. In mathematical approaches, experts' individual assessments on an uncertain quantity are expressed as subjective

probabilities. They are combined through various mathematical methods by the decision-maker after their elicitation. Behavioral approaches aim at producing some type of group consensus among experts, who are typically encouraged to interact with one another and share their assessments.

One of the most well-known behavioral approaches is the Delphi technique, which was developed in the 1950s. In this method, experts are asked to anonymously judge the assessments made by other experts in a panel. Each of the experts is then given a chance to reassess his/her initial judgment based on the others' review. Typically, the process is repeated several rounds until a smaller spread of experts' opinions is achieved. The Delphi method later incorporated a self-rating mechanism, allowing experts to rate their expertise (Cooke 1991; Parenté and Parenté 1987; Sackman 1975). The Nominal Group method is another well-known behavioral method, in which experts are allowed to discuss their estimates directly with one another in a controlled environment (Delbecq et al. 1975). This method is considered more favorably than other group methods, particularly the Delphi method (Gustafson et al. 1973).

While a group consensus method may help identify experts' errors and misunderstandings during the process, there are no formal rules to apply to reconcile differences when the consensus is difficult to achieve among different experts. Conformity induced by the group interaction is a major concern with such an approach. Mosleh et al. (1988) note that the group interactive process can suffer from, for example, the tendency for less confident experts to limit their participation, the influence of dominant personalities, and a tendency to reach speedy conclusions. Genest and Zidek (1986) warn of potential "strategic manipulation, bluffing, intimidating tactics and threats" if unrestricted dialogue is permitted. Cooke (1991) points out that the group interaction tends to produce more extreme probability estimates, potentially making the participants overconfident. Issues on group polarization are discussed by Plous (1993), Sniezek (1992), and Phillips (1987). There is a wide range of literature on the method for estimating a group consensus (Zahedi 1986; Goicoechea et al. 1982; Eliashberg and Winkler 1981).

It is generally agreed that mathematical approaches yield more accurate results than do behavioral approaches in aggregating expert opinions (Clemen and Winkler 1999; Mosleh, Bier and Apostolakis 1988). In the next section, three well-established mathematical modeling approaches to aggregating expert opinion are discussed: non-Bayesian axiomatic models, Bayesian models, and psychological scaling models (paired comparisons). The general framework of the present review is based on Cooke (1991), with reference to additional reviews and discussions from other relevant literature.

Modeling Approach 1: Non-Bayesian Axiomatic Models

An axiomatic-based approach to combining expert opinion is discussed in many of the earlier studies (Genest and Zidek 1986). In this approach, axioms or certain properties and regularity conditions for combining probability distributions are established, based on which the

form of combination rules is derived[1]. The approach requires that the relationship between experts' opinions and the aggregated opinion must satisfy a certain set of axioms. Most axiomatic models proceed to identify some form of 'weights' as parameters to be estimated[2].

General mathematical principles

Bedford and Cooke (1996) provide a fundamental mathematical concept behind the process of combining probability distributions. General principles are reproduced here. Using their notations, suppose we have experts 1, …, e and let P be a set of probability measures over some fixed but unspecified probability space ($P_1, \ldots, P_e \in P$). Then a combination rule that depends only on $P_1, \ldots, P_e$ is a function G: $P^e \rightarrow P$, i.e. G($P_1, \ldots, P_e$) = p. A set of axioms or constraints on G determines the form of G. Assuming that the probability space contains a finite number of events $a_1, \ldots a_n$, let p_{ij} be the expert i's probability for event a_j. Then a vector of probabilities given by expert i is $P_i = (p_{i1}, \ldots, p_{in})$ where $p_{ij} \geq 0$ and $\Sigma_j\ p_{ij} = 1$ for $j = 1, \ldots, n$ and $i = 1, \ldots, e$. Let w_i be a set of non-negative weights, $w_i \geq 0$ and $\Sigma_i\ w_i = 1$. Then for any real number $r \in R$, the elementary r-norm weighted mean probability of a_j is $M_r(j) = (\Sigma_i\ w_i\ p_{ij}^r)^{1/r}$, $i = 1, \ldots e$, and the r-norm probability of a_j is $P_r(j) = M_r(j) / \Sigma_k\ M_r(k)$, $k = 1, \ldots, n$. (Note: for r=1, P_r is the weighted arithmetic mean of P_i; for $r = 0$, P_r is the weighted geometric mean, and for $r = -1$, P_r is the weighted harmonic mean of P_i.) The interpretation of M_r is the following (Bedford and Cooke 1996, Chapter 10):

i) $M_r(j) \rightarrow \Pi_i\ p_{ij}^{wi}$ as $r \rightarrow 0$ ($i = 1, \ldots, e$)
ii) $M_r(j) \rightarrow \max_{i=1,\ldots,e}(p_{ij})$ as $r \rightarrow \infty$
iii) $M_r(j) \rightarrow \min_{i=1,\ldots,e}(p_{ij})$ as $r \rightarrow -\infty$
iv) If $r < s$ then $M_r(j) \leq M_s(j)$ with equality if and only if $p_{ij} = p_{kj}$ for $1 \leq i, k \leq e$
v) Define $M_r(j+k) = (\Sigma_i\ w_i\ (p_{ij} + p_{ik})^r)^{1/r}$, $i = 1, \ldots, e$, and assume that p_{ij}/p_{ik} is not constant in i. Then the following (in)equalities hold:

$r > 1 \Rightarrow M_r(j+k) < M_r(j) + M_r(k)$
$r = 1 \Rightarrow M_r(j+k) = M_r(j) + M_r(k)$
$r < 1 \Rightarrow M_r(j+k) > M_r(j) + M_r(k)$

The value for r should be chosen so that the combination function G($P_1, \ldots, P_e$) satisfies the following three properties.

i) *Strong set-wise function property*: For every subset $A \subseteq \{a_1, \ldots a_n\}$, the decision maker's probability depends only on the experts' judgments of A, so if $Q_i = (P_i(A), 1 - P_i(A))$, then $G(P_1, \ldots, P_e)(A) = G(Q_1, \ldots, Q_e)(A)$.
ii) *Marginalization property*: The probabilities are unaffected by refinements of the partition of alternatives $a_1, \ldots a_n$.
iii) *Independence preservation*: For all $A, B \subseteq \{a_1, \ldots a_n\}$ such that $P_i(A \cap B) = P_i(A) * P_i(B)$, $i = 1, \ldots, e$, $G(P_1, \ldots, P_e)(A \cap B) = G(P_1, \ldots, P_e)(A) * G(P_1, \ldots, P_e)(B)$.

[1] For example, Morris (1983, 1986) provides an axiomatic approach to expert aggregation, which created a series of reactions and debate by Lindley (1986), French (1986), Schervish (1986), and Winkler (1986).
[2] A departure from the traditional axiom approach is given by Myung et al. (1996), who use the principle of maximum entropy in their aggregation method.

A notion of opinion pools

Decisions on many political and public policy issues rely on the opinions of experts. An opinion pool is a method of combining a number of different opinions about some unknown quantity θ to generate a single pooled opinion about θ. The two most widely used opinion pool methods are linear opinion pools and logarithmic opinion pools. Suppose we have n experts, and let $p_i(\theta)$ represent expert i's probability distribution for unknown quantity θ, i =1, …, n, and w_i be expert i's weight. Then the combined probability distribution $p(\theta)$ is a weighted linear combination of the experts' probabilities (weighted arithmetic mean model) in a linear opinion pool, whereas $p(\theta)$ in a logarithmic opinion pool is expressed as multiplicative averaging (weighted geometric mean model):

Linear opinion pool	$p(\theta) = \sum_i w_i * p_i(\theta)$	
Logarithmic opinion pool	$p(\theta) = k \prod_i p_i(\theta)^{wi}$	(k is a normalizing constant)

The problem of opinion pools generally reduces down to determining the optimal weights w_i for experts. Various methods for finding the optimal models are explored, for example, in DeGroot and Mortera (1991), Bordley (1982), and DeGroot (1974). Apparently, the simplest choice of weights is assigning all experts an equal weight, $w_i = 1 / n$. A simple arithmetic averaging of experts' assessments is used in many studies, including a U.S. Nuclear Regulatory Commission study on the frequency of accidents at nuclear reactors (NRC 1989). Cook (1991) discusses that while there are some efforts in compensating such a simplistic method by improving the elicitation procedure itself, such as those witnessed for the NRC document (Honano et al. 1990; Wheeler et al. 1989), this type of method is less than optimal as it lacks any attempt to evaluate the quality of experts' estimates.

Cooke's classical performance-based weight model

The method proposed by Cooke (1991) is a performance-based weighted averaging model using properties of scoring rules, known as the classical model[3]. He emphasizes that the fundamental goal of science is to build rational consensus and, therefore, the process of collecting expert assessments must be subjected to the following five basic principles (the first and second principles are later combined as a scrutability/accountability principle (Cooke and Goossens 2000)):

1. *Reproducibility*: All results must be reproducible, with calculation models and data being clearly specified and made available.
2. *Accountability*: The source of data (name and institution) must be identified, and data must correspond to the exact source from which the data are elicited.
3. *Empirical Control*: Experts' assessments must be, in principle, physically observable.
4. *Neutrality*: The elicitation process must ensure that the actual beliefs of experts be collected (e.g. no punishment or rewords through a self-rating system).
5. *Fairness*: All experts must be regarded equally before the aggregation process.

[3] The term classical comes from the calibration measure's close association with classical hypothesis testing.

The classical model is designed to satisfy all these principles of rational consensus. In case of continuous variables, the model requires experts to provide a set of fixed quantiles q_r, $r = 1, ..., R$, for some unknown variables (seed variables) $X_1, ..., X_N$. The decision-maker then determines the intrinsic range (lower and upper bound, $[q_l, q_h]$) of each variable for each expert. The weights for the experts are established by two measures of performance: calibration and information scores. Calibration of expert e, $C(e)$, is the statistical likelihood that an expert's quantile assessment corresponds to a set of experimental results.

Let $p = p_1, ..., p_n$ be the hypothesized probability distribution of a variable over a set of alternatives $\{1, \ldots, n\}$, and $s = s_1, ..., s_n$ be the empirical distribution from N independent samples (i.e. N is the number of seed variables) from the distribution p. The discrepancy between p and s can be measured by the relative information of s with respect to p, $I(s, p) = \Sigma_i s_i \ln(s_i/p_i), i=1, ..., n,$ where we note that $P(2N\, I(s, p) \leq x) \rightarrow \chi^2_{n-1}(x)$ as $N \rightarrow \infty$. Then $C(e)$, which is the probability of obtaining a relative information score worse than what has been observed, is obtained as $C(e) = 1 - \chi^2_{n-1}(2N I(s, p))$.

The information score, or informativeness, measures the degree to which an expert's distribution is concentrated, relative to some background measure (usually of uniform or log-uniform distribution over an intrinsic range for each variable). The relative information for expert e on a given variable is obtained as $I(e) = \Sigma_i p_i \ln(p_i/r_i), i=1, ..., n,$ where r_i is the background measure for interval i. The overall information score of each expert is the average of the information scores over all variables.

The weights for the experts (which satisfy the asymptotically strictly proper scoring rules) are then determined as: $w_e = w_e' / W$, where $w_e' = C(e) \times I(e) \times 1_\alpha(C(e))$, and $W = \Sigma\, w_e'$ (note: for the level of significance α, $1_\alpha(C(e)) = 1$ if $C(e) \geq \alpha$, and $= 0$ if $C(e) < \alpha$, $\alpha \in (0, 1)$). Given w_e and each expert's cumulative distribution function F_e, the combined expert distribution is $\Sigma_e\, w_e\, F_e$.

Modeling Approach 2: Bayesian Models

Perhaps the most robust technique in combining expert opinion is the Bayesian method[4]. In this method, the decision maker uses experts' probability assessments as data to update his own prior belief about the distribution of an unknown quantity of interest, according to Bayes' Theorem. The early framework of the Bayesian aggregation method was provided by Winkler (1968) and Morris (1974, 1977). Winker (1968) discusses his natural conjugate theory, in which prior and posterior distributions belong to the same parametric family of distribution.

Let $P(x)$ be the decision-maker's prior probability distribution for some unknown quantity x, and $P(D|x)$ be the likelihood of some observational data D given x. Then the decision-maker's posterior distribution is $P(x|D) = [P(D|x)*P(x)] / P(D)$ via Bayesian update.

[4] Winkler (1986) notes, for example, that all contributors to Morris' (1983, 1986) axiomatic approach, i.e. Lindley, French, and Schervish, would in principle agree that the modeling approach with Bayesian principles is the most appropriate way to combine experts' assessments.

Suppose the decision-maker uses the beta distribution to model the prior uncertainty about x and then updates that beta prior on the basis of an observation, i.e. information from each successive expert, then the posterior distribution of x also will be a beta distribution. Knowledge of D and parameters of the prior distribution will immediately lead the decision-maker to his posterior distribution.

Morris' work (1974, 1977) fully establishes the foundations for the Bayesian paradigm in the analysis of expert judgment. He presents a straightforward aggregation method for the single expert case and the multi-expert case. In general, suppose the decision-maker wishes to make an inference about an unknown quantity X, and experts $1, \ldots, e$ provide their estimates of X, $x_1, \ldots x_e$. Let $p(x)$ be the decision-maker's prior probability density for X, and $p(x_1, \ldots x_e \mid x)$ be his beliefs about the experts' assessments as estimates of X, i.e. the likelihood function. The primary goal of the decision-maker is to find his posterior density, $p(x \mid x_1, \ldots x_e)$:

$$p(x \mid x_1, \ldots x_e) \propto p(x) * p(x_1, \ldots x_e \mid x)$$

Morris assumed that the experts are independent, and their density functions are represented by a normal distribution with the mean and variance. When independence among the experts is assumed, the likelihood term is simply the product of $p(x_i \mid x)$, $i=1, \ldots, e$:

$$p(x_1, \ldots x_e \mid x) = \Pi\, p(x_i \mid x).$$

Building on Morris' method, Mosleh and Apostolakis (1986) suggest two practical models for determining the above product.

Additive error and multiplicative error models

Mosleh and Apostolakis (1986) express each of the experts' assessments using the true value of unknown quantity and an error term.

1. Additive error model. The expert i's assessment, x_i, is expressed as the sum of the true value of X, x, and an error ξ_i : $x_i = x + \xi_i$ where $\xi_i \sim N(\mu_i, \delta_i^2)$ and independent (μ_i and δ_i^2 are the mean and variance determined by the decision-maker to reflect his belief about the expert's bias and accuracy). The likelihood $p(x_i \mid x)$ of obtaining x_i when the true value is x is the value of the normal density with mean $x + \mu_i$, and variance δ_i^2. Suppose the decision-maker acts as the e+1st expert and his prior distribution $p(x) \sim N(x_{e+1}, \delta_{e+1}^2)$. Then the decision-maker's posterior $p(x \mid x_1, \ldots x_e)$ is normal with mean and variance given as follows:

$$E(x \mid x_1, \ldots x_e) = \sum_{i=1}^{e+1} w_i (x_i - \mu_i) \quad \text{where } w_i = \delta_i^{-2} \Big/ \sum_{j=1}^{e+1} \delta_j^{-2} \quad \text{and} \quad \mu_{e+1} = 0$$

$$Var(x \mid x_1, \ldots x_e) = 1 \Big/ \sum_{i=1}^{e+1} \delta_i^{-2}$$

2. Multiplicative error model. The expert i's assessment, x_i, is expressed as the product of the true value x and an error ξ_i : $x_i = x * \xi_i$ where $\ln \xi_i \sim N(\mu_i, \delta_i^2)$ and independent.

Thus, $ln(x_i) = ln(x) + ln(\xi_i) \sim N((ln(x) + \mu_i, \delta_i^2)$ i.e. x_i is lognormally distributed. Given the decision-maker's prior distribution $p(x) \sim Lognormal(x_{e+1}, \delta_{e+1}^2)$,

$$E(ln(x) \mid x_1, \ldots x_e) = \sum_{i=1}^{e+1} w_i (ln(x_i) - \mu_i) \quad \text{and} \quad Var(x \mid x_1, \ldots x_e) = 1 / \sum_{i=1}^{e+1} \delta_i^{-2}$$

Note: $X \sim Lognormal(\mu, \delta^2) \Rightarrow$ $E(X) = e\{exp(\mu + \sigma^2/2)\}$ and

$$Var(X) = e\{exp(2\mu + 2\sigma^2)\} - e\{exp(2\mu + \sigma^2)\}$$
$$= [e\{exp(\mu + \sigma^2/2)\}]^2 * [e\{exp(\sigma^2)\} - 1]$$

Thus the decision-maker's posterior expectation and variance are given as:

$$E(x \mid x_1, \ldots x_e) = e\{exp(\sum_{i=1}^{e+1} w_i (ln(x_i) - \mu_i) + (1 / \sum_{i=1}^{e+1} \delta_i^{-2})/2)\}$$

$$= \prod_{i=1}^{e+1} (x_i / e^{\mu i})^{w_i} * e\{exp(1/2 \sum_{i=1}^{e+1} \delta_i^{-2})\}$$

$$Var(x \mid x_1, \ldots x_e) = [\prod_{i=1}^{e+1} (x_i / e^{\mu i})^{w_i} * e\{exp(1/2 \sum_{i=1}^{e+1} \delta_i^{-2})\}]^2 * [e\{exp(1/\sum_{i=1}^{e+1} \sigma^2)\} - 1]$$

Issues on stochastic dependence

Winkler (1981) stresses the importance of taking into account the possibility of stochastic dependence in modeling expert opinion. Dependence among experts' assessments arises when, for example, the experts selected by the decision-maker have worked in the same field or shared similar training and techniques. Dependent opinions are not only redundant, but they can cause a significant impact on the decision-maker's assessment. The sensitivity of the decision-maker's posterior density distribution in the presence of high correlations among experts is discussed by Winkler and Clemen (1992) and Clemen and Winkler (1985). Chhibber and Apostolakis (1993) also provide a sensitivity analysis on the use of dependent information sources.

Because of its practicality, an independent normal distribution is assumed for experts' assessments in many studies that followed Morris' (1974, 1977) Bayesian approach (French 1980, 1981; Winkler 1981; Lindley 1983, 1985). Identifying a joint likelihood distribution for experts' probability assessments is considered as one of the difficulties in using a Bayesian method. Some of the recent studies indicate numerous attempts to tackle the issue. Clemen and Winkler (1993) provide a procedure for constructing the joint likelihood function as a product of marginal and conditional likelihood functions. Lipscomb et al. (1998) propose a hierarchical approach to incorporate dependencies among experts, in which experts' variation is assumed to follow a normal distribution and a sample of experts can be drawn from the second-order distribution. Saaty and Vargas (1998) use the analytic hierarchy process to address dependence symptoms.

The use of copulas is another approach to specifying a likelihood function that seems to have gained some attention in the recent literature (Jouini and Clemen 1996; Yi and Bier 1998; Clemen and Reilly 1999; and Clemen et al. 2000). A copula is a function that is used to join a set of particular marginal distributions, thereby generating a multivariate distribution that carries those marginals. The approach discussed by Jouini and Clemen (1996), for example, shows the following: Let $f_i(\theta)$ be the continuous density for unknown θ provided by expert i, $F_i(\theta)$ be the corresponding cumulative distribution function, and C be the copula density function. Then under some conditions, the decision-maker's posterior distribution is shown as

$$P(\theta \mid f_i, \ldots, f_n) \propto C\,[1 - F_1(\theta), \ldots, 1 - F_n(\theta)]\,\Pi_i\, f_i(\theta) \qquad i=1, \ldots, n.$$

Mendel and Sheridan (1989) use the Bayesian method to recalibrate and combine experts' assessments. Their model does not require experts to follow any particular distribution, and thus offers more flexibility and an ease of implementation. In this model, each expert provides R quantiles for a variable of interest, which defines $R+1$ cells that could be hit by the true value. Given n experts, an $(R+1)^n$ array is formed, which is considered as a random variable. The decision-maker uses the theory of exchangeability to update his probability distribution for this random variable, before receiving the experts' advice on the current variable. Information from past assessments and outcomes provides the likelihood of hit in each cell.

Modeling Approach 3: Psychological Scaling Models (Paired Comparisons)

The psychological scaling models assume that every expert has some internal value associated with a variable of interest and he/she can only provide his or her qualitative input (no numerical estimates). The decision-maker asks experts to state their preference or views on pair-wise comparisons. This approach originated from the study of estimating intensities of physical stimuli, which later developed into the study of estimating relative intensities of psychological stimuli among experts[5]. Using simulation, experts' assessments lead to a consensus with confidence bounds. Their inputs are measured for their consistency and concordance (Cooke 1991).

Suppose there are n experts, and each expert is asked to express his/her preference for one of two events. Let $A(1), \ldots, A(t)$ be events to be compared, $V(1), \ldots, V(t)$ be the true probabilities of events. Let $V(i, e)$ denote the internal value of expert e for event i. If expert e prefers event $A(i)$ to event $A(j)$, (i.e. $A(i) > A(j)$), then $A(i)$ is judged more probable than $A(j)$ by e, and $V(i, e) > V(j, e)$. Three models are presented below.[6]

Thurstone model (1927)

Assumptions: Normal distribution of internal values over the population of experts

$V(i, e) \sim N(\mu_i, \delta_i)$

[5] An important contribution to the study is the assessment of human error probabilities. See Kirwan (1994), Humphreys (1988), USNRC (1983), and Swain and Guttmann (1983).

[6] In addition to the three models shown here, see Pulkkinen (1994) for his various paired comparison methods.

$\mu_i = V(i)$
$\delta_i = \delta$ (δ_i does not depend on i)
$V(i, e) - V(j, e) \sim N(\mu_{ij}, \delta_{ij})$ where $\mu_{ij} = \mu_i - \mu_j$ and $\delta_{ij} = (\delta_i^2 + \delta_j^2 - 2\rho_{ij} * \delta_i \delta_j)^{1/2}$
$= (\delta^2 + \delta^2)^{1/2}$

Then the probability that expert e prefers $A(i)$ to $A(j)$ is:
$P((V(i, e)-V(j, e))>0) = P\{ (V(i, e) - V(j, e)- \mu_{ij})/\delta_{ij} > - \mu_{ij}/\delta_{ij} \}$
$= P(X > - \mu_{ij}/\delta_{ij}) = \Phi (\mu_{ij}/\delta_{ij})$

Note: By setting $x_{ij} = \Phi^{-1}$(the percentage of experts who prefer $A(i)$ to $A(j)$) $\approx \mu_{ij} / \delta_{ij}$, and $\delta_{ij} = 1 \Rightarrow \mu_i - \mu_j \approx x_{ij}$. Using the least-squares method, the estimate of μ_i is obtained as:

$$\mu_i' = 1/t * \Sigma_j x_{ij} \quad j = 1, \ldots, t.$$

Bradley-Terry model (1952, 1953)

Assumptions: Each event $A(i)$ is associated with a true scale value $V(i)$, and the probability that event $A(i)$ is preferred to event $A(j)$, $r(ij)$, is given by:

$$r(ij) = V(i) / [V(i) + V(j)]$$

Experts' judgments are treated as independent coin-toss trials with the probability of head = $r(ij)$. The proportion of experts preferring $A(i)$ over $A(j)$ is used as an estimate of $r(ij)$.

Using the maximum likelihood method (David 1963), $V(i)$, $i = 1, \ldots, t$, is given as:

$$V(i) = a(i) / \Sigma_j n[V(i) + V(j)]^{-1} \quad (\text{for } j \neq i)$$

where $a(i)$ is the number of times event $A(i)$ is preferred to other events by experts. The value of $V(i)$ is obtained by iteration, which converges to a unique solution.

Negative exponential lifetime (NEL) model

This model is used specifically for estimating constant failure rates. For each pair of independently-operating components, the decision-maker asks experts which one of the two components will fail first. Component $A(i)$ is assumed to have an exponential life distribution with failure rate $r(i)$, and all components are assumed as good as new at time t =0. Thus, the probability that $A(i)$ fails before time T is:

$$P(A(i) < T) = \int_t r(i) e^{-r(i)t} dt \quad (t = 0 \text{ to } T)$$

And the probability that $A(i)$ fails before $A(j)$ is $P(A(i) < A(j)) = r(i) / [r(i)+r(j)]$. The solution is obtained in a similar manner as in the Bradley-Terry Model.

The paired comparison models are appealing in that experts are not required to be familiar with numerical assessments and the overall elicitation process is relatively simple. They have, however, disadvantages such as requiring a large number of experts and forcing stringent

assumptions about experts' underlying assessment mechanisms. Cooke (1991) notes that the paired comparison approach has a difficulty in satisfying the principle of empirical control. The goodness-of-fit tests used in the paired comparison models examine the fitness of experts' assessments against modeling assumptions, but not against the empirical values. The confidence intervals obtained through the models reflect uncertainty due to the choice of experts, and do not reflect uncertainty due to modeling assumptions.

Conclusions

The three modeling approaches presented in the paper have all been used in some real-life situations. Each approach clearly has advantages and disadvantages, for example, as seen in the case of the Bayesian modeling approach that has mathematical sophistication but suffers from limitation in its practical application. A general agreement appears to be that there is no single all-purpose aggregation method for expert opinion.

In their critical review of several case studies involving the use of expert opinion, Mosleh et al. (1988) note that while there is evidence that expert opinion can be highly beneficial in probabilistic risk analysis, little attention has been paid to structuring the elicitation process. Although empirical evidence indicates that mathematical methods of aggregation generally yield better results than behavioral methods, the latter methods are often perceived appealing, particularly when experts have knowledge in different areas and the synthesis of their expertise is needed. Based on a case study that successfully assessed seismic hazard rates using expert opinion, Mosleh et al. (1988) show the use of a multiple-team approach with mathematical aggregation as one of the most promising methods in dealing with such a problem.

As decision-makers in general tend to use the most convenient aggregation methods of their choice (and not necessarily the most appropriate), Cooke and others call for formalizing the elicitation process and using expert opinion by ensuring the basic principles of rational consensus (i.e. satisfying reproducibility, accountability, empirical control, neutrality, and fairness in resulting assessments), which seems truly timely and appropriate.

In order to support the practical application of expert analysis, two relevant software programs have been developed by the Department of Mathematics at Delft University of Technology. *EXCALIBUR* (from Expert CALIBRation) is a Windows program for expert judgment analysis. It allows the user to input experts' quantile assessments and parameters, and combine their assessments based on equal weights, user weights, and expert performance-based weights (Cook 2001, 1991). *UNICORN* is a software package for uncertainty analysis. It is designed for dependence modeling with high dimensional distributions, including graphic features, such as cobwebs (Cooke 1995). Both programs are designed to facilitate the formal procedure of expert analysis, which is based on the principles of rational consensus.

References on Expert Opinion (*Cited in the paper)

Bedford T. and Cooke, R.T. (2001), "Probabilistic Risk Analysis: Foundations and Methods," Cambridge University Press.*

Bier, V.M. and Yi, W. (1995), "A Bayesian Method for Analyzing Dependencies in Precursor Data," *International Journal of Forecasting*, Volume 11, Issue 1, pp.25-41.*

Bordley, R.F. (1982), "A Multiplicative Formula for Aggregating Probability Assessments," *Management Science*, Volume 28, Issue 10, pp.1137-1148.*

Campodonico, S. and Singpurwalla, N.D. (1995), "Inference and Predictions from Poisson Point Processes Incorporating Expert Knowledge," *Journal of the American Statistical Association*, Volume 90, Issue 429, pp.220-226.*

Chande, P.K. and Tokekar, S. V. (1998), "Expert-Based Maintenance: A Study of Its Effectiveness," *IEEE Transactions on Reliability*, Volume 46, No. 1, pp.53-58.

Chhibber, S. and Apostolakis, G. (1993), "Some Approximations Useful to the Use of Dependent Information Sources," *Reliability Engineering and System Safety*, Volume 42, Issue 1, pp.67-86.*

Clemen, R.T. (1986), "Calibration and the Aggregation of Probabilities," *Management Science*, Volume 32, Issue 3, pp.312-314.

Clemen, R.T., Fischer, G.W., and Winkler, R.L. (2000), "Assessing Dependence: Some Experimental Results," *Management Science*, Volume 46, Issue 8, pp.1100-1115.*

Clemen, R.T. and Winkler, R.L. (1999), "Combining Probability Distributions from Experts in Risk Analysis," *Risk Analysis*, Volume 19, Issue 2, pp.187-203.*

Clemen, R.T. and Reily, T. (1999), "Correlations and Copulas for Decision and Risk Analysis," *Management Science*, Volume 45, Issue 2, pp. 208-224.*

Clemen, R.T. and Winker, R.L. (1993), "Aggregating Point Estimates: A Flexible Modeling Approach," *Management Science,* Issue 39, No. 4, pp. 501-515.*

Clemen, R.T. and Winker, R.L. (1985), "Limits for the Precision and Values of Information from Dependent Sources," *Operations Research,* Volume 33, pp. 427-442.*

Cooke, R.M. and Slijkhuis, K.A. (2003), "Expert Judgment in the Uncertain Analysis of Dike Ring Failure Frequency," Appearing in Case Studies in Reliability and Maintenance, pp.331-350.

Cooke, R.M. (2001), "EXCALIBUR – Windows version of EXCALIBR: Software for performance based combination of expert judgments," Department of Mathematics, Delft University of Technology.*

Cooke, R.M. and Goossens, L.H.J. (2000), "A Procedures Guide for Structured Expert Judgment," EUR 18820, European Commission Report.*

Cooke, R.M. and Goossens, L.H.J. (2000), "Procedures Guide for Structured Expert Judgment in Accident Consequence Modeling," Radiation Protection and Dosimetry, Vol. 90, No. 3, pp.303-309.*

Cooke, R.M., and Goossens, L.H.J. (2000), "Expert Judgment Elicitation in Risk Assessment," Delft University of Technology ("Abstract NATO workshop, Lisbon, 1-4 October 2000)*

Cooke, R.M. (1995), "UNICORN: Methods and Code for Uncertainty Analysis," AEA Technologies.*

Cooke, R. M. (1991), "Experts in Uncertainty: Opinion and Subjective Probability in Science," Oxford University Press, Oxford.*

David, H.A. (1963), "The Method of Paired Comparisons," Charles Griffin, London

DeGroot, M.H. (1974), "Reaching a Consensus," *Journal of the American Statistical Association*, Volume 69, pp. 118-121.*

DeGroot, M.H. and Mortera, J. (1991), "Optimal Linear Opinion Pools," *Management Science*, Volume 37, Issue 5, pp 546-558.*

Delbecq, A., Van de Ven, A., and Gusstafson, D. (1975), "Group Techniques for Program Planning," Glenview, III, Scott-Foresman.*

Eliashberg, J. and Winkler R.L. (1981), "Risk Sharing and Group Decision Making," *Management Science*, Volume 27, Issue 11, pp.1221-1235.*

French, S. (1986), "Calibration and the Expert Problem," *Management Science*, Volume 32, Issue 3, pp 315-321.*

French, S. (1981), "Consensus of Opinion," *European Journal of Operations Research*, 7, pp. 332-340.*

French, S. (1980), "Updating of Belief in the Light of Someone Else's Opinion," *Journal of the Royal Statistical Society: Series A*, Volume 143, pp. 43-48*

Genest, C. and Zidek, J.V. (1986), "Combining Probability Distributions: A Critique and an Annotated Bibliography, *Statistical Science*, Volume 1, Issue 1, pp. 114-135.*

Goicoechea, A., Hansen, D.R. and Duckstein, L. (1982), "Multiobjective Decision Analysis with Engineering and Business Applications," Wiley.*

Goossens, L.H.J., Harper, F.T., Kraan, B.C.P. and Metivier, H. (2000), "Expert Judgment for a Probabilistic Accident Consequence Uncertainty Analysis," Radiation Protection and Dosimetry, Vol. 90, No. 3, pp.295-303.

Goossens, L.H.J., Cooke, R.M. and Kraan, B.C.P. (1998), "Evaluation of Weighting Schemes for Expert Judgment Studies," *Probabilistic Safety Assessment and Management* (Proceedings of PSAMA 4), pp.2389-2396.

Gustafson, D., Shulka, R., Delbecq, A., and Walster, A. (1973), "A Comparative Study of Differences in Subjective Likelihood Estimates Made by Individuals, Interacting Groups, Delphi Groups, and Nominal Groups," *Organizational Behaviour and Human Performance,* Volume 9, pp. 280-291.*

Hatfield, A.J. and Hipel, K.W. (2002), "Risk and Systems Theory," *Risk Analysis*, Volume 22, No. 6, pp.1043-1057.*

Honano, E.J., Hora, S.C., Keeney, R.L., and von Winterfeldt, D. (1990), "Elicitation and Use of Expert Judgment in Performance Assessment for High-level Radioactive Waste Repositories," NUREG/CR-5411.*

Humphreys, R. (1988), "Human Reliability Assessors Guide," Safety and Reliability Directorate, United Kingdom Atomic Energy Authority*

Jenni, K. and Fischhoff, B. (1997), "Attributes for Risk Evaluation," 1997 Annual Meeting of Society for Risk Analysis.*

Jouini M. N., and Clemen, R. T. (1996), "Copula Models for Aggregating Expert Opinions," *Operations Research*, Volume 44, Issue 3, pp. 444-457.*

Kirwan, B. (1994), "A Guide to Practical Human Reliability Assessment," Taylor and Francis, London.*

Kline, A. and Renn, O. (2002), "A New Approach to Risk Evaluation and Management: Risk-Based, Precaution-Based, and Discourse-Based Strategies," *Risk Analysis*, Volume 22, No. 6, pp.1071-1094.*

Kraan, B.C.P. and Cooke, R.M. (2000), "Processing Expert Judgment in Accident Consequence Modeling," Appearing in Radiation Protection Dosimetry (Expert Judgment and Accident Consequence Uncertainty Analysis; special issue), Vol. 90, No. 3, pp.311-315.

Lindley, D.V., and Singpurwalla, N.D. (1986), "Reliability (and Fault Tree) Analysis Using Expert Opinions," *Journal of the American Statistical Association*, Volume 81, Issue 393, pp. 87-90.*

Lindley, D.V. (1987), "Using Expert Advice on a Skew Judgmental Distribution," *Operations Research*, Volume 35, Issue 5, pp 716-721.

Lindley, D.V. (1986), "Another Look at an Axiomatic Approach to Expert Resolution," *Management Science*, Volume 32, Issue 3, pp 303-306.*

Lindley, D.V. (1985), "Reconciliation of Discrete Probability Distributions," *Bayesian Statistics* 2, pp.375-387.*

Lindley, D.V. (1983), "Reconciliation of Probability Distributions," *Operations Research,* Volume 31, Issue 5, pp.866-660.*

Lipscomb, J., Parmigiani, G., and Hasselblad, V. (1998), "Combining Expert Judgment by Hierarchical Modeling: An Application to Physician Staffing," *Management Science*, Volume 44, Issue 2, pp 149-161.*

Mendel, M. and Sheridan, T. (1989), "Filtering Information from Human Experts," *IEEE Transaction Systems, Man and Cybernetics*, Volume 36, pp. 6-16. *

Morris, P.A. (1986), "Observations on Expert Aggregation," *Management Science*, Volume 32, Issue 3, pp 321-328.*

Morris, P.A. (1983), "An Axiomatic Approach to Expert Resolution," *Management Science*, Volume 29, Issue 1, pp 24-32.*

Morris, P.A. (1977), "Combining Expert Judgments: A Bayesian Approach," *Management Science*, Volume 23, Issue 7, pp 679-693.*

Morris, P.A. (1974), "Decision Analysis Expert Use," *Management Science*, Volume 20, Issue 9, Theory Series, pp 1233-1241.*

Mosleh, A., Bier, V.M., and Apostolakis, G. (1988), "A Critique of Current Practice for the Use of Expert Opinions in Probabilistic Risk Assessment," *Reliability Engineering and System Safety*, Volume 20, pp. 63-85.*

Mosleh, A., Bier, V.M., and Apostolakis, G. (1987), "Methods for the Elicitation and Use of Expert Opinion in Risk Assessment," NUREG/CR-4962, PLG-0533, US Nuclear Regulatory Commission (micro fiche)

Mosleh, A. and Apostolakis, G. (1986), "The Assessment of Probability Distributions from Expert Opinions with an Application to Seismic Fragility Curves," *Risk Analysis*, Volume 6, No. 4, pp 447-461.*

Myung, I.J., Ramamoorti, S., and Bailey, A.D.Jr. (1996), "Maximum Entropy Aggregation of Expert Predictions," *Management Science*, Volume 42, Issue 10, pp 1420-1436.*

Parenté, F.J. and Anderson-Parenté, J.K. (1987), "Delphi Inquiry Systems," *Judgmental Forecasting*.*

Phillips, L.D. (1987), "On the Adequacy of Judgmental Forecasts," *Judgmental Forecasting*, pp. 11-30.*

Plous, S. (1993), "The Psychology of Judgment and Decision Making." New York, McGraw-Hill.*

Pulkkinen, U. (1994), "Gaussian Paired Comparison Models," *Reliability Engineering and System Safety*, Volume 44, Issue 2, pp. 207-217.*

Pulkkinen, U. (1994), "Bayesian Analysis of Consistent Paired Comparisons," *Reliability Engineering and System Safety*, Volume 43, Issue 1, pp. 1-16.*

Pulkkinen, U. (1993), "Methods for Combination of Expert Judgments," *Reliability Engineering and System Safety*, Volume 40, Issue 2, pp. 111-118.

Saaty, T.L. and Vargas, L.G. (1998), "Diagnosis with Dependent Symptoms: Bayes Theorem and the Analytic Hierarchy Process," *Operations Research*, Volume 46, Issue 4, pp 491-502.*

Sackman, H. (1975), "Delphi Critique: Expert Opinion, Forecasting and Group Processes," Lexington, MA, Lexington Books.*

Schervish, M.J. (1986), "Comments on Some Axioms for Combining Expert Judgments," *Management Science*, Volume 32, Issue 3, pp 306-312.*

Singpurwalla, N.D. (1988). "An Interactive PC-Based Procedure for Reliability Assessment Incorporating Expert Opinion and Survival Data," *Journal of the American Statistical Association*, Volume 83, Issue 401, pp. 43-51*

Singpurwalla, N.D., and Song, M.S. (1988), "Reliability Analysis using Weibull Lifetime Data and Expert Opinion*," IEEE Transactions on Reliability*, Volume 37, No. 3, pp.340-347.*

Skjong, R. and Eknes, M.L. (2002), "Societal Risk and Societal Benefits," *Risk Decision and Policy*, Volume 7, pp. 57-67.*

Sniezek, J. (1992), "Groups under uncertainty: An Examination of Confidence in Group Decision Making," *Organizational Behavior and Human Decision Processes*, 52, pp.124-155.*

Swain, A.D., and Guttmann, H.E., (1983), "Handbook of Human Reliability Analysis with Emphasis on Nuclear Power Plant Applications," NUREG/CR-1278.*

The U.S. Presidential/Congressional Commission on Risk Assessment and Risk Management (1997), "Risk Assessment and Risk Management in Regulatory Decision-Making," Final Report, Volume 2.*

Thurstone, L. (1927), "A Law of Comparative Judgment," *Psychological Review*, Volume 34, pp.273-286.*

U.S. Environmental Protection Agency (2003), "Framework for Cummulative Risk Assessment," Risk Assessment Forum, EPA (EPA/630/P-02/001F, April 2002).

U.S. Nuclear Regulatory Commission (1989), "Severe Accident Risks: An Assessment for Five US Nuclear Power Plants," U.S. NRC, NUREG/CR-1150*

U.S. Nuclear Regulatory Commission (1983), "PRA Procedure Guide," U.S. NRC, NUREG/CR-2300*

Van Noortwijk, J.M., Dekker, R., Cooke, R. and Mazzuchi, T.A. (1992), "Expert Judgment in Maintenance Optimization," *IEEE Transactions on Reliability*, Volume 41, No. 3, pp.427-432.*

Von Winterfeld, D. (1989), "Eliciting and Communicating Expert Judgments: Methodology and Application to Nuclear Safety," Joint Research Centre, Commission of the European Communities.*

West, M. (1988), "Modeling Expert Opinion," *Bayesian Statistics* 3, Oxford University Press, pp.493-508.

Winkler, R.L. (1986), "Expert Resolution," *Management Science*, Volume 32, Issue 3, pp.298-303.*

Winkler, R.L. (1981), "Combining Probability Distributions from Dependent Information Sources," *Management Science*, Volume 27, Issue 4, pp.479-488.*

Winkler, R.L. (1968), "The Consensus of Subjective Probability Distributions," *Management Science*, Volume 15, Issue 2, pp. B61-B75.*

Winkler, R.L. and Clemen, R.T. (1992), "Sensitivity of Weights in Combining Forecasts," *Operations Research,* Volume 40, pp.609-614.*

Winker, R.L., Hora, S.C., Baca, R.G. (1992), "The Quality of Experts' Probabilities Obtained Through Formal Elicitation Techniques," Center for Nuclear Waste Regulatory Analyses CNWRA, CNWRA T. Rep.*

Wheeler, T.A., Hora, S.C., Cramond, W.R., and Unwin, S.D. (1989), "Analysis of Core Damage Frequency from Internal Events: Expert Judgment Elicitation," NUREG/CR-4550, Volume 2, Sandia National Laboratories. *

Yi, W. and Bier, V.M. (1998), "An Application of Copulas to Accident Precursor Analysis," *Management Science*, Volume 44, Issue 12, Part 2 of 2, S257-S270.*

Zahedi, F. (1986), "Group Consensus Function Estimation when Preferences are Uncertain," *Operations Research*, Volume 34, Issue 6, pp. 883-894.*