Multi-Expert Operational Risk Management

Giampiero E. G. Beroggi, *Senior Member, IEEE* and W. A. Wallace, *Senior Member, IEEE*

***Abstract*—Operational risk management is the process of monitoring, evaluating, and changing courses of actions with potential detrimental consequences in real time. In this paper, we extend the decision models proposed in the literature for individual risk managers to account for situations where multiple risk managers are involved. For this purpose, two dynamic and adaptive preference aggregation models for cardinal and ordinal assessments are proposed and discussed. The mechanical aspects of the models are then validated using field data collected from experienced operational risk managers in an individual-expert setting. Sensitivity analysis indicates that the models have enough flexibility to be adapted to account for behavioral considerations. The paper closes with a research agenda.**

***Index Terms*—Decision making, experts, preference aggregation, risk management.**

I. Introduction

THE NEED to assess and manage risks in real time is receiving increasing attention due to the growing complexity of large-scale operations and the advances in information and communications technologies. Prominent examples are the use of satellite communications and computing technology for logistics management, security concerns for world-wide communications networks, and information and communications technologies in support of crisis and emergency management for natural disasters such as floods, earthquakes, and fires.

The concept of operational risk management (ORM) was introduced as real-time monitoring and control of courses of action (CA's), and decision-making in the case of sudden unforeseen events, called real-time events (RTE) [6]. For example, advanced information and communications technologies are being employed by trucking companies to monitor their shipments in real time. For changes in planned shipments, such as new customer orders, traffic congestion, or truck problems, the dispatcher can assist the drivers in their decision-making process, including rerouting. Rerouting shipments (or at least the investigation of rerouting possibilities) can be motivated by many reasons, including new opportunities (e.g., picking up an unforeseen order) as well as for safety and security reasons (e.g., unforeseen bad weather or possible terrorist attack). Tests and pilot studies in the U.S. and Europe have demonstrated the potential of real-time monitoring and route guidance [2]. However, these new technologies have not yet established themselves world-wide. It is our contention, and governmental efforts especially in the European scene point in this direction [15], that safety and security issues will become an integral aspect of real-time monitoring and control of large-scale distributed operations.

Planning and modifying CA's in an ORM setting involves many different considerations, including the operators (e.g., drivers), companies (e.g., trucking companies), public interests, logistics, and emergency response issues. Assessing the impacts of unexpected threatening events is not done just by an operator but includes multiple experts, such as engineers, highway patrol, environmental experts, public officials, and emergency managers—all acting under time pressure. In this paper we extend the single-operator ORM concept as proposed by Beroggi and Wallace [6] to multi-expert operational risk management. The proposed concept is based on the continuous advance of information and communications technologies for real-time monitoring and control, including video conferencing, multimedia systems, and wireless communication.

The following section provides an overview of ORM. Section III presents the formulation of multi-expert ORM. Finally, Section IV discusses the sensitivity of the proposed models using data obtained from a quasi-field experiment with experienced dispatchers and truck drivers. The paper closes with suggestions for continued research in development of decision models and their implementation as components of decision support systems for ORM.

II. The Decision Environment for Multi-Expert ORM

ORM is the process of monitoring, evaluating, and changing CA's due to real-time events. A CA is defined as a set of chronologically ordered decisions to take specific actions. Examples of CA's are emergency response plans and transportation routes for hazardous materials. The space of feasible CA's can be represented as a graph, where the edges are the decisions to take certain actions (vertices). In the case of emergency plans (Fig. 1, left), the edges of the graph are the possible decisions and the links the response actions; in the case of transportation of hazardous materials (Fig. 1, right), the edges of the graph are the intersections and the vertices the road segments connecting the intersections.

Past research developed and assessed different models for ORM to assess the preferences of the actions on a graph structure and to compute optimal courses of actions, with two of them being the ordinal preference (OP) model and the multiattribute utility (MAU) model [5]. In their most basic form, both models use the two criteria risks and costs to express the *preferences* for taking certain actions and for making decisions for changes in case an RTE occurs.

An RTE typically affects only some of the actions; for example, a snow storm affects only parts of the road network. An

Manuscript received December 1997; revised September 23, 1999.

G. E. G. Beroggi is with the Faculty of TBM, 2600 GA Delft, The Netherlands (e-mail: beroggi@sepa.tudelft.nl).

W. A. Wallace is with DSES/CII, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: wallaw@rpi.edu).

Publisher Item Identifier S 1094-6977(00)02036-8.

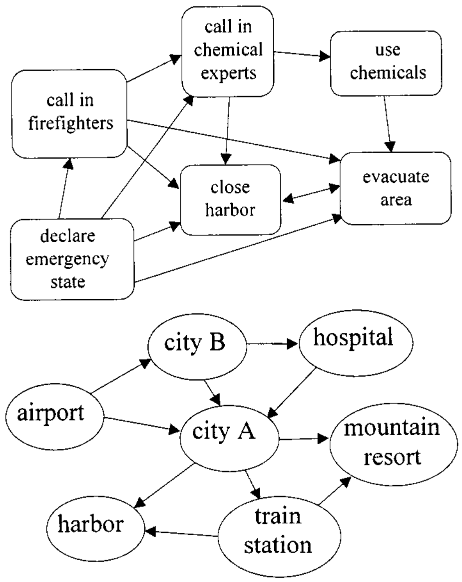


Fig. 1. Examples of emergency response (left) and transportation (right) graphs.

action (e.g., road segment) is said to be affected by an RTE if the preference to engage in this action deteriorates. An operation (e.g., vehicle) is said to be affected by an RTE if it plans to engage in at least one of affected actions (e.g., if it plans to drive through the snow storm).

In order to use these decision models in ORM conditions, all actions must first be assessed under *normal* operating conditions (i.e., for when no RTE's are present). When an RTE occurs, the operational risk manager (e.g., dispatcher, on-site emergency manager) can change the risk and cost values of the affected actions. Based on these new risk and cost values, the system will compute, possibly new, CA's for all affected operations.

A. The Multiattribute Utility (MAU) Model

The MAU model uses a cardinal scale for both risks and costs. Cost preferences are assessed in terms of some monetary unit, which stands for a monotonically decreasing preference function. Risk preferences are expressed on a pseudo-logarithmic scale which is commonly used in the practice of risk management. The scale ranges from 0 (no risk) to 100 (certain accident); a value of 90 stands for an estimated annual probability for a fatality of 10^{-1}, a value of 70 stands for 10^{-3}, and a value of 84 stands for 4×10^{-2}. Again, these values are subjective estimates made by an experienced ORM manager. The general transformation equation to transform any $r \in [0, 100]$ risk preference value into an absolute risk measure, $\lambda(x) \in [0, 1]$, is given below [5]: These absolute λ risk values are then transformed into risk-costs using estimates for life-saving costs (LSC) of 10^7 SFr to reduce one unit of λ [13].

The cost preferences in the case of hazardous material transportation are computed from the estimated driving velocity, using the cost estimate of \$1.0/km for driving at a velocity of 80 km/h. Thus, the transportation costs along a link of length l at a velocity v (km/h) are $(80/v)^*l)$. With the total cost values (i.e., the sum of risk-costs and transportation costs), the minimum cost route (i.e., CA) can be computed as follows, where $x_{ij} = 1$ if road segment l_{ij} is taken and $x_{ij} = 0$ otherwise.

$$\min: C(R) = \sum_{i,j \in I_N} \left(\frac{80}{v_{ij}} l_{ij} + \mathrm{LSC} \times \left(\sum_k \lambda_{ij} l_{ij} x_{ij} \right) \right)$$

subject to:

- R consists of connected road segments from origin to destination:

$$\sum_{i \in I_D}^{n} x_{ik} - \sum_{j \in I_D}^{n} x_{kj} = \begin{cases} 1, & \text{for } k = \text{origin} \\ 0, & \text{for all other } k \\ -1, & \text{for } k = \text{destination.} \end{cases}$$

B. The Ordinal Preference Model (OP)

The OP model uses a lexicographic preference structure to measure preferences of actions. The lexicographic classes are (where "<" stands for "less preferred"):

$$\begin{aligned} &\text{avoid action } (\alpha) \\ &\quad < \text{high risk (HR)} < \text{costs } (C) \\ &\quad < \text{low risks (LR)} < \text{negligible impact } (\omega). \end{aligned}$$

If an action is assigned an α preference, then none of the ongoing operations can engage in this action. This could mean that a road segment must be closed for hazardous material shipments or that the use of chemical dispersant to abate an oil spill is not an option. The ω preference reflects that a situation has been detected but no negative implication has been assessed yet. This could mean that possible problems have been identified for certain actions which, however, at this point in time do not affect any of the ongoing operations. The C class allows the ORM manager to express financial implications. Although they might be expressed in some monetary unit, they do not have to stand for absolute cost values or best possible estimates. Rather, they can reflect subjective estimates or preferences by the ORM manager. An action is assigned an HR preference if no financial implications are deemed too high to avoid this activity.

Consequently, if two courses of action contain HR but no α activities, then the choice between the two is based solely on risk and not on cost aspects. On the other hand, an activity is assigned an LR preference if a risk has been detected which would

$$\lambda(r) = \begin{cases} \bullet 10^{-[10 - \mathrm{INT}(r/10)]}, & \text{for } r \in Z = \{0, 10, 20, \ldots, 70, 80, 100\} \\ \bullet \left[x - 10 \times \mathrm{INT}\left(\frac{r}{10}\right)\right] \times 10^{-[10 - \mathrm{INT}(r/10)]}, & \text{for } r \in [0, 100] \backslash Z \end{cases}$$

only be avoided if it does not involve any additional financial implications. Therefore, if two courses of action contain LR but no HR or a activities, then the choice between the two is based solely on the cost and not on risk aspects.

The preference of an action is the "sum" ($\oplus$) over its preference classes; for example, an action a_1 with an HR risk preference and $2C$ cost preferences has an overall preference of: $\pi(a_1) = \mathrm{HR} \oplus 2C$. The preference of a CA is the "sum" ($\oplus$) of the preferences of its actions. For example, if CA consists of seven actions: $\pi(a_1) = \mathrm{HR} \oplus 2C$, $\pi(a_2) = 5C \oplus 5\mathrm{LR}$, $\pi(a_3) = \mathrm{HR} \oplus 3C$, $\pi(a_4) = 4C \oplus \omega$, $\pi(a_5) = \mathrm{HR} \oplus C$, $\pi(a_6) = \mathrm{HR} \oplus 2C$, and $\pi(a_7) = 3C \oplus \omega$, then, the overall preference is: $\pi(\mathrm{CA}) = 0\ \alpha \oplus 4\mathrm{HR} \oplus 20C \oplus 1\mathrm{LR} \oplus 2\omega$. A transformation of the lexicographic preference value into a numeric preference value, for a CA consisting of n actions, can be computed as follows, where a_i is the ith action in the CA:

$$\pi(\mathrm{CA}) = \sum_{i=1}^{n} \left(\sum_{k=1}^{5} M_k \pi_{ik} \right).$$

Thereby, π_{i1} is the α preference of action a_i, π_{i2} the HR preference of action a_i, π_{i3} the C preference of action a_i, π_{i4} the LR preference of action a_i, and π_{i5} the ω preference of action a_i. M_k $(k = 1, \ldots, 5)$ are numbers making the preference classes incommensurable: $M_1 \gg M_2 \gg M_3 \gg M_4 \gg M_5$ (e.g., $M_1 = 1/0$, $M_2 = 10^4$, $M_3 = 10^1$, $M_4 = 10^{-4}$, and $M_5 = 10^{-8}$). Let π_{ij} be the preference of action a_{ij} (i.e., a_{ij} is the action between the decisions i and j) and x_{ij} the decision variable for taking action a_{ij} ($x_{ij} = 1$ if action a_{ij} is taken, $x_{ij} = 0$ otherwise). Then, the most preferred CA is the one with lowest preference value $\pi(\mathrm{CA})$; it is defined by:

$$\min: \pi(\mathrm{CA}) = \sum_{ij \in I_D} \sum_{k \in I_\pi} M_k \pi_{ij} x_{ij}$$

subject to:

- flow conservation: CA starts with the currently ongoing action (e.g., using burning to abate the oil spill) and ends in the planned end state of the operation (e.g., response crews can be sent home);
- feasibility of actions: $\pi_{ij} \neq \alpha, \forall i, j$, where I_D is the set of all decisions and $I\pi = \{\alpha, \mathrm{HR}, C, \mathrm{LR}, \omega\}$ the set of the preference classes.

This means, for example, that if the overall preference of CA_1 is $3\mathrm{HR} \oplus 25C \oplus 12\mathrm{LR}$, and the overall preference of CA_2 is $5\mathrm{HR} \oplus 18C \oplus 17\mathrm{LR}$, then CA_1 is preferred to CA_2. The preference of the joint CA's, $\mathrm{CA}_1 \cup \mathrm{CA}_2$, is $\pi(\mathrm{CA}_1) \oplus \pi(\mathrm{CA}_2) = 8\mathrm{HR} \oplus 43C \oplus 29\mathrm{LR}$.

C. ORM for Multi-Expert Situations

The reason to extend the OP and MAU decision models, and not other possible models, to multi-expert ORM situations is that they outperformed traditional visual interactive (VI) and the conservative heuristics (CH) models in terms of effort and accuracy [5]. The VI model does not provide any analytic support to the ORM manager in case of an RTE, other than the visual display of the graph and the affected activities. The CH model computes without user interaction alternative CA's based on worst-case estimates.

The OP and MAU models provide decision support by *attributes* (risk and cost) while the VI and CH models provide decision support by *alternative* (i.e., the ORM decision manager compares alternative CA's in their entirety rather than based on risk and cost attributes). Our finding that decision support by attribute is superior to decision support by alternative in an ORM environment is supported by other research; see Payne *et al.*[19] for instance.

Consequently, ORM decision models should support also multiple managers in: 1) the identification of the actions affected by an RTE; 2) the assessment of the impact when engaging in those actions in terms of the attributes; 3) the identification of the affected operations (CA's); and 4) the decisions to change planned CA's.

ORM often requires the assistance of experts working at different locations. With the advent of technologies like cellular communications, the experts can access data on the RTE, discuss its ramifications for operations, and assess the current and proposed state of operations. However, different experts process data differently. Thus, the results of their analyses and the recommendations for changing CA's cannot be expected to be the same. Discrepancies among the experts occur at two stages: 1) during analysis (i.e., assessment) and 2) when making recommendations for new CA's or to modify CA's (i.e., choice).

It is our contention that experts' assessments in an ORM environment should not be judged solely in terms of right or wrong, good or bad, and reliable or unreliable. Rather, their expertise should be judged in terms of the *consistency* and *agreement* of their assessments and recommendations over multiple RTE's. Consistency is defined in terms of rank value of the assessment. For example, if one expert assesses constantly the lowest risk value, then s/he is very consistent. Agreement is defined in terms of deviations from the weighted group mean. Therefore, a trustworthy group of experts has not only high agreement but also a high level of consistency.

Typically, however, both consistency and agreement vary over time and may adapt to group norms. For example, if experts are presented the results of their assessments relative to the other experts' assessments, they might undergo an adaptation process and change their attitude toward risk and costs and, subsequently, assessment and choice.

The ways multi-expert ORM assessments and choices by n experts can be processed (i.e., aggregated) is illustrated in Fig. 2. The assessments by n experts are first checked for acceptability. If an expert's assessment is determined not to be acceptable it can be disregarded or she can be asked for a reassessment—in real time. The accepted assessments are used to derive a group assessment based on which the optimal CA's are computed.

The individually recommended changes in CA's from each of the experts whose assessment was found to be acceptable as well as the new CA's based upon the aggregated group assessment can be presented in graphical form to the decision-maker. An aggregation of the accepted CA's could also be done. We summarize our view of the ORM multi-expert assessment and choice process as follows:

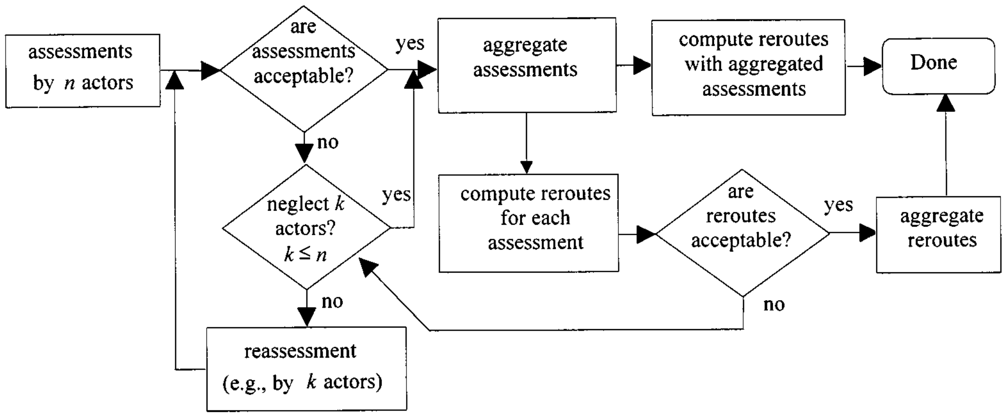


Fig. 2. Process of experts' assessment and choice.

- Each expert makes an assessment of the risk and cost preferences for the activities affected by the RTE. It is assumed that the affected activities have been identified by one expert and that they are submitted to all other experts.
- Each expert's assessment is evaluated in terms of **consistency** and **agreement** for the purpose of determining a group assessment. A weight is determined for each expert which is a function of: 1) his/her consistency in past assessments and 2) his/her degree of group membership (agreement).
- For each expert and for the group assessment, changes in CA's are computed and presented to the ORM manager for decision-making.

Preference aggregation for multiple experts has been addressed both in the context of risk management and in the context of social choices. In the latter case the issue has been subject of research for more than 200 years and resulted in well known paradoxes, such as rank reversals in the Borda count, the Condorcet paradox of non-transitivity, and Arrow's impossibility theorem [4]. Risk management methods for the aggregation of expert assessments have been addressed by Beinat *et al.*[3], Cooke [9], DeWispelare *et al.*[10], Sandri *et al.*[20], and Myung *et al.*[18].

The procedures proposed in the literature to aggregate expert assessments can be classified as **behavioral** and **mechanical** [10]. Behavioral procedures are based on either structured discussions on non-interactive approaches such as the DELPHI Method [17]. Mechanical procedures consist of mathematical formulae and algorithms, of which a large body of literature has emerged over the last two decades [8]. However, for practical purposes, simple averages of point estimates have often shown to suffice, although they have some conceptual drawbacks [10].

The multi-expert ORM environment as defined earlier is characterized by repetitive assessments of RTE's over time, subjective preference elicitation by experts, and an objective to have the experts be consistent and in agreement. Early work in the related field of command and control dates back to the 1960s [11]. More recently, models for sequential revision of belief have been proposed [12], [14]. The extensions of the MAU and OP models for multiple experts as proposed in the following section differ in the sense that long-term consistency and agreement by the experts are incorporated.

III. Multi-Expert ORM Decision Models

A. An Adaptive Model for Cardinal Preference Assessments

The multi-expert MAU model is based on cardinal preference assessments for an action in terms of costs and risks. It is assumed that n experts will have to assess RTE's as they occur in time. The process of assessing the k-th RTE is referred to as the k-th stage. Let $v_{i|k}$ be the cardinal preference value (e.g., for risk) provided by expert i $(i = 1, \ldots, n)$ for any action in stage k. Then, the aggregated group assessment for n experts is defined as the weighted average, where $w_{i|k}$ is the **weight** of expert i for the assessment in stage k:

$$m_k = \left(\sum_{i=1}^{n} w_{i|k}\right)^{-1} \times \sum_{i=1}^{n} w_{i|k} v_{i|k}. \tag{1}$$

It should be noted that the weighted mean value is invariant over multiplicative transformations of the weights. Therefore, we do not have to require that the sum of weights be equal to one. In other words, if we rescale the weights by multiplying them by a constant, the weighted mean is still the same. The weight $w_{i|k}$ accounts for consistency and agreement of expert i in stage k.

The weight of expert i in stage $k, w_{i|k} = f(c_{i|k-1}, d_{i|k-1})$, is determined iteratively and is a function of the expert's long-run consistency, $c_{i|k}$, and his/her **relative deviation** $(d_{i|k})$ from the aggregated group assessment in the previous assessment(s):

$$d_{i|k} = \left| \frac{v_{i|k-1} - m_{k-1}}{v_{\max|k-1} - m_{k-1}} \right|. \tag{2}$$

The expert whose assessment was furthest away from the aggregated group assessment in stage $k-1$ has in stage k a deviation coefficient, $d_{i|k} = 1$ (least agreement with the group). If there is an expert whose assessment was equal to m_{k-1} then the deviation coefficient is $d_{i|k} = 0$ (highest agreement with the group). The **long-run consistency**, $c_{i|k}$, is defined as the long-term average rank change:

$$c_{i|k} = \frac{1}{(n-1)(k-j)} \sum_{j=1}^{k} r_{ij|k}, \tag{3}$$

where $1 \leq j \leq k-1$, with j being the first stage to count from. The value j reflects how much of the assessment history should be accounted for. For example, j could be the stage at which the composition of the group has changed the last time.

The maximum long-run consistency (least consistent) with n experts is $c_{\max|k=1}$ and the minimum value (most consistent) is $c_{\min|k} = 0$, if the expert never changes rank. The rank change, $r_{ij|k}$, is defined as the number of positions an expert changed compared to the previous assessment. For example, let's assume that an expert provided in stage $k-1$ the second most conservative assessment and in stage k the fourth most conservative assessment; in this case, the rank change is $r_{ij|k} = 2$.

The weight for each expert can now be computed as a function of the expert's long-run consistency, $c_{i|k} \in [0$ (most consistent), 1 (least consistent)], and his/her relative deviation, $d_{i|k} \in [0$ (in highest agreement),1 (in lowest agreement)]: $w_{i|k} = f(c_{i|k-1}, d_{i|k-1})$. For the function, f, we propose an approach that has been first introduced as the Minkowski metric and then extended to fuzzy logic by Yager [21] and employed in the Swiss safety regulation for disaster scaling [7]. The **weight** for expert i in stage k is:

$$w_{i|k} = \mathrm{Round}(10 \times w^*_{i|k} + 1)$$

where

$$w^*_{i|k} = 1 - \min[1, ((c_{i|k-1})^\kappa + (d_{i|k-1})^\kappa)^{1/\kappa}] \tag{4}$$

where $\kappa = 1, 2, \ldots,$ and $w^*_{i|k} \in [0, 1]$. The characteristic of the parameter κ is that for $\kappa = \infty, w^*_{i|k} = 1 - \min[1, \max(c_{i|k-1}, d_{i|k-1})]$, and for $\kappa = 1$, $w^*_{i|k} = 1 - \min[1, c_{i|k-1} + d_{i|k-1}]$; that is, $w_{i|k} = 1$ only if $c_{i|k} = 0$ (i.e., expert never changes rank) and $d_{i|k} = 0$ (i.e., expert is in perfect agreement with the group). The weights are therefore $w_{i|k} \in [1, 11]$; the constant 1 was introduced to prevent divisions by 0.

We would like to specify a **measure of acceptability** (δ_k) for the n assessments provided in stage k, around the weighted mean value in order to determine the acceptable assessments. We propose to do this as a function of the unbiased sample variance, s_k^2, of the weighted assessments:

$$s_k^2 = \frac{\sum_{i=1}^n w_{i|k}(v_{i|k} - m_k)^2}{\sum_{i=1}^n w_{i|k} - 1} \tag{5}$$

which converges toward the variance under multiplicative transformation, that is:

$$\delta_k = \sqrt{\lim_{w\to\infty} s_k^2}$$

where

$$\begin{aligned}\lim_{w\to\infty} s_k^2 &= \frac{w\sum_{i=1}^n w_{i|k}(v_{i|k} - m_k)^2}{w\sum_{i=1}^n w_{i|k} - 1}\\ &= \frac{\sum_{i=1}^n w_{i|k}(v_{i|k} - m_k)^2}{\sum_{i=1}^n w_{i|k}}.\end{aligned} \tag{6}$$

Group consistency, C_k, which gets computed at each assessment stage is a measure of disarray of the rank-orders of the experts' assessments over time. This measure assumes that differences in the assessments are less serious if they occur because the experts have different degrees of conservatism. Assume that there are three experts and that one consistently assigns highest risk values, another always lowest, and the third lies always in-between. Then, we would say that the group as a whole is consistent and we would be more willing to accept differences between the experts than in the case where the preference-orders change very often.

The group consistency measure is therefore a relative measure of concordance of conservatism or risk attitude. Different measures have been proposed in the literature and we will use the most prevalent one: Kendall's coefficient of concordance, used to test rank correlations [16, p. 119]. Thus, the group consistency is defined as:

$$C_k = \frac{S}{S_{\max}} = \frac{12S}{n^2(k^3 - k) - n\sum U'}$$

with

$$S = \sum_{i=1}^n R_1^2 - \frac{kn^2(k+1)^2}{4}, \tag{7}$$

where the R_i^2 are the squares of deviations of the rank sums around their mean which equals $k(n+1)/2$, and $U' = \sum(n^3 - n)/12$ is to reduce the sum of square of deviations due to tied ranks. For example, assume we have two experts assessing the cost preferences as 66, three experts assessing 72, and another three experts assessing 69. Then, we get as correction factor $\sum U' = [(2^3 - 2) + 2(3^3 - 3)]/12 = 4.5$.

The group consistency is a measure of relative agreement of k rank-orders. It takes on the value $C_k = 1$ for perfect agreement, and the value $C_k = 0$ for complete disagreement. Consequently, the **acceptability range** is defined as

$$v_{i|k} \in [m_k - \gamma_k C_k \delta_k, m_k + \gamma_k C_k \delta_k] \tag{8}$$

where γ_κ is a constant. For $\gamma_k = 1$, the acceptability range is the standard deviation multiplied by C_k. A small value of γ_k causes many assessments to be rejected. This would be reasonable if time is available to ask the experts who fall outside the acceptability range to do a reassessment. Moreover, γ_k could vary in with increasing k.

Instead of using a behavioral approach where some of the experts might be asked to do a reassessment, the acceptable assessments at each stage can also be determined mechanically in an iterative way. First, with all n assessments, the acceptability range is computed for some γ_k. Assessments falling outside of this range are discarded. With the remaining assessments, a new range is determined by computing new m_k and δ_k values with the remaining assessments. Then, the unacceptable assessments are discarded. This procedure is repeated until all remaining assessments are contained within the acceptability range (i.e., until none of the remaining assessments is rejected).

However, if at the first iteration the acceptable set is empty, we propose that the assessment be considered as not valid. In such a case, the decision-maker (e.g., dispatcher or on-site emergency manager) must rely on his/her own judgment. Thus, the algorithmic approach of determining the acceptable assessments is as follows:

$F = \{v_{i|k}\}$, (set of all assessments), $x := 0$, $F'_x := \emptyset$ (the set of rejected assessments)
repeat while $F'_x \neq F'_{x+1}$
$x := x + 1$, compute $m_{k|j}$, $C_{k|j}$, and $\delta_{k|j}$,
$\{v_{i|k}\} \in F \rightarrow F'_x := F'_{x+1}$, where $v_{i|k} \notin [m_{k|j} - \gamma_k C_{k|j}\delta_{k|j}, m_{k|j} + \gamma_k C_{k|j}\delta_{k|j}]$
$F := F \setminus F'_{x+1}$
if $F = \emptyset$ and $x = 1$ then group assessment is bi-modal and unacceptable
end repeat

Note that the set of acceptable assessments could be empty for two cases: 1) if the group assessment for an RTE is bi-modal and 2) if the experts are so inconsistent over time that C_k and δ_k are very small, resulting in a range of acceptability too small to accept any assessments.

B. An Adaptive Model for Ordinal Preference Assessments

The assessment spectrum for an ordinal risk scale $\prod$ was introduced as consisting of several preference classes referring to cost and risk attributes and also to the priorities between the two (the avoidance of HR actions has higher priority than of C actions, and higher priority than of LR actions): $\prod = \prod^{(1)} < \cdots < \prod^{(n)}$. A possible assessment spectrum for the OP model was proposed as: $\alpha < \mathrm{HR} < C < \mathrm{LR} < \omega$.

Assume that the attribute-classes for the risk attribute (in this example the two classes HR and LR), as well as the classes α (action must be avoided) and ω (impact of RTE is negligible) consist of only one element, and the classes of costs of multiple elements. It is further assumed that an assessment is done properly if every activity is assigned one element reflecting every attribute. For ORM, an assessment is **proper** if every activity which gets assessed has a risk-element from $\{\alpha, \mathrm{HR}, \mathrm{LR}, \omega\}$ and a cost-element from $\{\alpha, \{\$\}, \omega\}$.

With n independent experts, the probability that all experts assign the same risk value is $4^{(1-n)}$. If their assessments were independent, the probability of having r identical assessments out of s assessments, where $r \leq s$, would be binomial distributed, where $p = 4^{(1-n)}$:

$$p_s^{(r)} = \binom{r}{s} p^s (1-p)^{r-s}.$$

For example, the probability that eight experts agree at least once in 18 assessments is about only 10^{-3}. This is an upper bound, since the experts might not be independent in their assessments. Thus, we would never expect all n experts to agree on all r assessments. On the other hand, a lower bound can be determined by considering that with four risk classes and n experts, $\mathrm{trunc}(n/4)$ experts must agree on at least one out of the four risk classes. We would therefore call a group assessment **sufficient**, if a majority of the (weighted) experts agrees on the tradeoff between risks and costs. For the assessment spectrum $\{\alpha, \mathrm{HR}, \mathrm{LR}, \omega\}$, this means that the assessment is sufficient if the majority of experts agree on one of the two sets $\{\alpha, \mathrm{HR}\}$ or $\{\mathrm{LR}, \omega\}$. The corresponding set of assessments is then referred to as the sufficient set.

We would call an assessment **efficient** if it is sufficient and if a majority of the experts choosing the sufficient set agree on one preference. The corresponding set of assessments is referred to as the efficient set. Finally, we would call an assessment **satisfactory**, if a majority of experts agrees on one preference class. The corresponding set of assessments is referred to as the satisfactory set. Evidently, a satisfactory assessment is efficient. To see this, let $\prod_< = \{\alpha, \mathrm{HR}\}$ and $\prod_> = \{\mathrm{LR}, \omega\}$. Let $\prod_i$ be the preference class which determined that the assessment is satisfactory; that is, by definition $|\prod_i| > |\prod_j|$, where $i \neq j$. Since $\prod_i \in \prod_<$ or $\prod_i \in \prod_>$, it follows that the assessment is efficient.

We shall define **acceptability** to include either efficient or satisfactory group assessments. Therefore, we may have a sufficient assessment which is not acceptable; for example, two experts assess α, two experts assess HR, one expert assesses LR, and two experts assess ω (where all experts have equal weight). If the experts have different weights, then we assume that $\mathrm{round}(100 \times w_i)$ is the number of experts agreeing on one specific assessment, where w_i is the weight of expert i.

A group assessment is proposed, equivalent to the mean group assessment for the MAU model in Section III-A. If a group assessment is satisfactory, the aggregated assessment is the preference class that is agreed upon by a majority of the experts (e.g., at least five out of eight experts). If a group assessment is efficient, the aggregated preference is the preference class that receives a majority within the two classes that are acceptable. For example, if (for equally weighted experts) two experts agree on α, three assess HR, one expert assesses LR, and two experts assess ω, we have an efficient, that is, acceptable, assessment, with HR being the aggregated group assessment. It should be noted that we do not use the class with the highest number of assessments as the group aggregated assessment. For example, in a case involving nine experts, four experts might assess LR, zero ω, three HR and two α. Then, the group aggregated assessment would not be LR, although it has the most experts (but not a majority) agreeing on this class, but HR because the assessment is efficient.

We can now define the **relative deviation**, $d_{i|k}$, to the aggregated assessment. The nonefficient values of an efficient assessment have a deviation value, $d_{i|k} = 0$. The efficient values which do not fall into the class which determined the assessment to be efficient have a deviation value, $d_{i|k} = 0.5$. Finally, the assessments that fall into the group aggregated class have a deviation value, $d_{i|k} = 1.0$. It should be noted that a non-efficient assessment does not provide an aggregated group assessment. In such cases, the decision-maker (e.g., dispatcher or on-site emergency manager) must rely on his/her own assessment.

Using these distance measures and the consistency coefficients (same definition as for the MAU model) of the experts, the weights are updated for the assessment of the next RTE in the same way as for the MAU model. The aggregation of the assessments does not, however, need to be done by an iterative procedure because the process of assessing the efficient set already takes care of the concept of eliminating outliers.

One might argue that the group aggregation procedure proposed for ordinal assessments is more severe than the one for cardinal assessments because of the different definitions of acceptability of the group as a whole. A cardinal assessment of a group is not accepted if the first iteration eliminates all assessments. An ordinal assessment is not accepted if it is not efficient, although it might be sufficient. The reason to allow this more severe definition is based on empirical evidence which shows that the ordinal assessment outperforms the cardinal assessment in ORM situations in terms of effort and accuracy [5]. Thus, we expect ordinal assessments to be more consistent than cardinal assessments, especially when one attribute is assessed on a cardinal scale (e.g., costs) and the other is assessed on an ordinal scale (e.g., risk).

C. Decision-Making

The decision-maker has the choice of maintaining the present CA or finding a new set of CA's in response to RTE's. For example, a dispatcher could identify alternative routes if the planned route is affected by an RTE (e.g., a show storm). In the case of ORM the decision-maker may be required to or may need to call upon "experts" to provide recommendations. We have modeled this process as first one of assessment (e.g., assessing the risk and cost impacts of driving through a snow storm) and then as one of choice (e.g., rerouting the vehicles to another route to avoid the snow storm).

However, in the case of changing CA's the choice phase is performed algorithmically based upon the assessments (e.g., a routing algorithm computes the alternative route). Therefore, the decision-maker has both the planned CA and a set of recommended CA's, which may include the planned CA; that is, the CA based upon the group assessment and the CA's based upon the individual expert assessments' for those experts whose assessment was found to be acceptable. These recommended CA's can be displayed graphically (e.g., on a map background). In addition, the decision-maker might want to identify other CA's that make some sort of sense.

Two CA's, i and j, are referred to as **indifferent** (preferentially equivalent) if they have the same overall preference: $\pi_i = \pi_j$. They are called **congruent** (strategically equivalent) if they consist of the same links in the same sequence: $l_k^i = l_k^j$. They are referred to as **analogous** if they are congruent but not indifferent. Finally, they are called **identical** if they are congruent and indifferent.

With any given assessment of the affected activities (e.g., the experts' assessments or the group aggregated assessment), alterations to the planned CA's can be computed. The **acceptability** of CA's can be defined in terms of how many experts come up with the same CA. In general, there are only two CA's possible based on the group aggregated assessments with decision-making by attribute—the planned CA and the CA which avoids the RTE with overall highest preference value. Thus, some experts will propose the alternative CA while the others will not. We can now compare the number of experts choosing the planned CA and the CA resulting from the aggregated assessment.

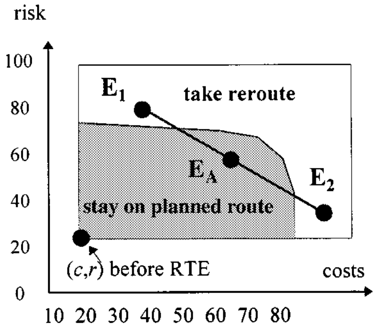


Fig. 3. Violation of the Pareto optimality axiom for high risk (or cost) RTE's.

The aggregation of multiple expert assessments for decision-making, however, could result in counterintuitive choices. Simple numerical examples can be constructed for which the linearly aggregated preferences of two experts result in a choice that contradicts both experts' recommendations. An aggregation procedure should avoid this rather awkward result. Social choice researchers have thus proposed the *Pareto optimality* axiom which says that if all experts agree on one alternative, then the aggregated assessment should result in this alternative [1]. Unfortunately, aggregation principles based on averages do not necessarily comply with this axiom.

To see this, let's assume there is an RTE (e.g., snow storm) which affects one action (e.g., route segment) of the planned CA_1 (e.g., route). A change of the planned CA_1 to an alternative CA_2 is suggested if the implications on risk and/or cost for undertaking this activity are too high. Because both risk and cost preferences are monotonically decreasing, all risk-cost pairs which result in keeping the planned CA_1 define a convex set (Fig. 3). If the risk/cost assessments of two experts E_1 and E_2 suggest taking CA_2 (e.g., to take the reroute), then a linearly aggregated assessment (E_A) will lie on the connecting line between the two assessments. This could result in either what the two experts suggested, or in a contradiction (e.g., to stay on the planned route) of their unanimous suggestion (e.g., to take the reroute).

Because of the possible violation of what is referred to as **Pareto optimality** (i.e., if all experts agree on one CA then the aggregated assessment must suggest the same CA), we propose showing all CA's based on the assessment of the "acceptable" experts, as well as the CA resulting from the aggregated assessment. The operator can then decide which of these CA's to choose. If all the experts' assessments result in the same CA, one that is different from the CA resulting from the aggregated assessment, a reasonable heuristic would be to choose the former CA. In any case, we would always consider what is referred to as **recognition**, saying that all experts must be considered in the aggregation of the assessments. In terms of the proposed models, this simply means that no expert should ever receive a weight of zero, a situation which would exclude him/her from the group of experts.

TABLE I
EXPERT ASSESSMENTS OF VELOCITY (CARDINAL) AND RISK (CARDINAL AND ORDINAL)

	E 1	E 2	E 3	E 4	E 5	E 6	E 7	E 8
RTE 1	v=0 r=95 r=HR	v=21 r=96 r=HR	v=26 r=99 r=LR	v=44 r=75 r=α	v=22 r=96 r=LR	v=39 r=88 r=HR	v=35 r=80 r=HR	v=18 r=85 r=HR
RTE 2	v=80 r=21 r=LR	v=32 r=55 r=LR	v=56 r=33 r=LR	v=46 r=52 r=LR	v=70 r=24 r=LR	v=72 r=21 r=LR	v=63 r=34 r=LR	v=65 r=49 r=LR
RTE 3	v=52 r=41 r=LR	v=39 r=62 r=LR	v=50 r=36 r=LR	v=52 r=47 r=LR	v=66 r=30 r=LR	v=61 r=49 r=LR	v=66 r=48 r=LR	v=49 r=42 r=LR
RTE 4	v=0 r=22 r=HR	v=17 r=100 r=HR	v=17 r=73 r=α	v=37 r=90 r=α	v=71 r=65 r=HR	v=39 r=81 r=α	v=60 r=74 r=HR	v=22 r=83 r=α
RTE 5	v=50 r=92 r=HR	v=17 r=93 r=HR	v=17 r=86 r=α	v=49 r=79 r=HR	v=43 r=33 r=HR	v=40 r=75 r=HR	v=44 r=77 r=HR	v=28 r=84 r=HR
RTE 6	v=58 r=21 r=LR	v=42 r=37 r=LR	v=30 r=21 r=LR	v=80 r=28 r=LR	v=58 r=21 r=LR	v=80 r=21 r=LR	v=45 r=35 r=LR	v=39 r=21 r=LR

IV. SENSITIVITY ANALYSIS

A. Preference Aggregation

The data used for this sensitivity analysis was collected in a quasi-field setting with the OP and MAU decision models applied to transportation of hazardous materials [5]. A total of 16 experienced dispatchers and truck drivers participated in the experiment, where eight were working with the MAU model and the other eight with the OP model.

The task for each expert (dispatcher) was to monitor the movement of three vehicles. After a few minutes of monitoring, two RTE's were announced which affected the planned routes of the vehicles. The dispatchers had then to assess for both RTE's the impacts on risks and travel velocity for vehicles driving through the areas which were affected by the RTE's. With these assessments, alternative routes were computed from the vehicles' current positions to their planned destinations with the MAU and OP models, respectively. The dispatcher could then decide for each vehicle whether to keep it on the planned route or to reroute it to the suggested alternative (which could be identical with the planned route if the new risk and velocity values were not too much different from original values prior to the RTE's). This task was repeated a total of three times for each expert with different RTE's and different routes for the three vehicles. As a result, each expert had to assess six RTE's and make nine rerouting decisions.

The tasks were presented in a multimedia environment were the movement of the vehicles was animated on a map background. The announcement of the RTE's was done through audio, static pictures, and text. Changes of risk and velocity preferences could be done with the computer mouse, either by moving a slide to change cardinal values or by clicking on the appropriate box to change ordinal values. These three decision situations lasted about 15 min. The six announced RTE's were taken from news accounts and described to the dispatchers as follows: very hazardous driving conditions due to ice rain (RTE1), cautious driving recommended due to heavy traffic (RTE2), cautious driving conditions due to heavy rain (RTE3), high explosion hazard due to accident with oil truck (RTE4), high danger due to falling rock (RTE5), and expected traffic delay during rush hours (RTE6).

The first two rows for each RTE in Table I show the assessments by the eight experts $(E1, \ldots, E8)$ for velocity (v) and risk (r) preferences working with the MAU model; the third row (r) shows the assessments of risk preferences by the other eight experts working with the OP model. The velocity values were transformed into cost values as discussed in Section II-A.

Because data for only six RTE's were collected in the quasi-field experiment, the assessments for the six RTE's were used three times for this sensitivity analysis, resulting in a total of 18 RTE's. The sequence for the six RTE's was generated randomly; the first (original) sequence was 1, 2, 3, 4, 5, 6; the second (random) sequence was 4, 3, 5, 6, 1, 2; and the third (random) sequence was 4, 3, 1, 2, 5, 6.Comparing the aggregated assessment in the second $(\kappa = \infty, \gamma = 2)$ and fourth column $(\kappa = 1, \gamma = 2)$, we see that the influence of κ is minor. The most extreme differences were obtained for the fifth and ninth RTE's, where $\kappa = \infty$ gave 33.6 as the aggregated value with zero (third RTE) and four (ninth RTE) experts agreeing on it, while and $\kappa = 1$ gave 17.0 as the aggregated value with two experts agreeing (both for the fifth and ninth RTE).The purpose of the sensitivity analysis is to test the mechanical aspects of the two models and not any behavioral aspects. That is, experts were not presented the assessments by the other experts, and they were not given the possibility to communicate to one another.

1) Agreement in Cardinal Assessments: The parameters to be set in the cardinal model are κ in (4) which directly affects the weights of the experts and indirectly the mean and the range of accepted assessments, and γ_k in (8) which affects the range of acceptable assessments.

TABLE II
AGGREGATED EXPERT ASSESSMENTS FOR VELOCITY WITH MAU MODEL FOR DIFFERENT PARAMETER SETTINGS AND NUMBER OF ITERATIONS

$\kappa=\infty$, $\gamma=1$ all iterations	$\kappa=\infty$, $\gamma=2$ all iterations	$\kappa=\infty$, $\gamma=2$, only 1 iteration	$\kappa=1$, $\gamma=2$ all iterations	$\kappa=\infty$, $\gamma=\infty$ all iterations
[21...21.5...22]/2/3	[-.4...25.6...51.6]/8/1	[-.4...25.6...51.6]/8/0	[-.4...25.6...51.6]/8/0	25.6
[63...63...63]/1/1	[63...63...63]/1/3	[52.6...60.5...68.4]/3/1	[63...63...63]/1/2	60.5
[61...61...61]/1/1	[50...50...50]/1/4	[48.4...56.8...65.2]/5/1	[50...50...50]/1/3	56.8
[37.8...38.3...38.8]/0/2	[17...17...17]/2/4	[15.1...35.8...56.6]/5/1	[17...17...17]/2/3	35.8
[40...40...40]/1/1	[28.7...33.6...38.5]/0/4	[21.8...33.6...45.5]/4/1	[17...17...17]/2/3	33.6
[58...58...58]/2/1	[45...45...45]/1/3	[33.9...55.9...77.9]/5/1	[39...39...39]/1/3	55.9
[37.6...38...38.5]/0/2	[17...17...17]/2/4	[19.4...39.6...59.8]/3/1	[17...17...17]/2/3	39.6
[52...52...52]/2/1	[50...50...50]/1/4	[44.8...54.5...64.2]/5/1	[49.1...49.6...50.0]/1/4	54.5
[29.3...32.6...36]/0/2	[21.6...33.6...45.7]/4/3	[22.8...36.8...50.8]/6/1	[17...17...17]/2/5	36.8
[49.4...52.8...56.2]/0/2	[34.5...46.4...58.2]/5/4	[37.9...59...80]/5/1	[39.6...40.9...42.3]/1/4	59
[22...23...24]/0/2	[21...21.5...21.9]/0/5	[13.8...24.8...35.8]/5/1	[21...21...21]/1/4	24.8
[63...63...63]/1/2	[65...65...65]/1/4	[43.4...57.8...72.2]/6/1	[65.6...67.7...69.8]/0/5	57.8
[37.7...38.1...38.6]/0/2	[22...22...22]/1/4	[13.7...37.1...60.5]/6/1	[22...22...22]/1/2	37.1
[52...52...52]/2/1	[50...50...50]/1/4	[44.4...53.56...62.7]/5/1	[49...49...49]/1/4	53.6
[24.2...24.8...25.5]/0/2	[21...21...21]/1/5	[13...27.5...41.9]/6/1	[21...21...21]/1/4	27.5
[59.7...61.3...62.9]/0/2	[65...65...65]/1/5	[44.8...58...71.2]/5/1	[65...65...65]/1/4	58
[40...40...40]/1/1	[43...43...43]/1/4	[23.1...35.6...48.2]/4/1	[43...43...43]/1/3	35.6
[58...58...58]/2/1	[58...58...58]/2/2	[39.9...58.8...77.7]/4/1	[58...58...58]/2/1	58.8

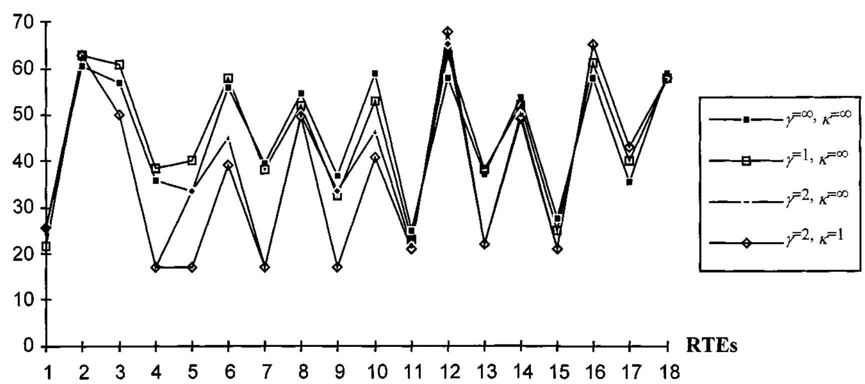


Fig. 4. Aggregated expert assessments for different parameter settings.

As previously mentioned, a small γ_k in (8) is used when time is available for the experts to repeat their assessments if they disagree to the extent that many of their assessments are rejected. This narrow band is achieved by setting $\gamma_k = 1$ for $k = 1, \ldots, 18$.

Table II shows results for the aggregated expert assessment of velocities for different values of κ and γ_k, where the experts are using the MAU model. For example, the first row in the first column, $[21 \ldots 21.5 \ldots 22]/2/3$, says that the range of acceptable assessments goes from 21 to 22, and that the aggregated assessment is 21.5. Only two experts fall within this band, and the band has been computed in three iterations. The parameter $\gamma_k = \gamma$ was held constant throughout the iterations.

A much stronger influence on the aggregated assessment has the parameter $\gamma_k = \gamma$, which was held constant throughout the interations. This can be seen when comparing columns 1 and 5 of Table II.

Fig. 4 shows the graphs for the data sets in columns 1, 2, 4, and 5 in Table II (all iterations according to algorithm in Section III-A). Obviously, the different parameter settings produce in the first two stages (RTE 1 and 2) similar group-aggregated values because the weights of the experts are assumed to be equal. Then, the group-aggregated values diverge. In the long-run, the values seem to converge again.

Fig. 5. Decline of standard deviation for values in Fig. 4 (smoothed through method of moving averages).

Fig. 5 shows that the standard deviation of the four data points and the 18 stages (RTE's) from Fig. 4, which have been smoothed through the method of moving averages, first increases monotonically up to stage seven (RTE 7) and then decreases monotonically.

The convergence of the different parameter settings in terms of aggregated group assessment, shows that in the long run one could use the weighted average model with any parameter set-

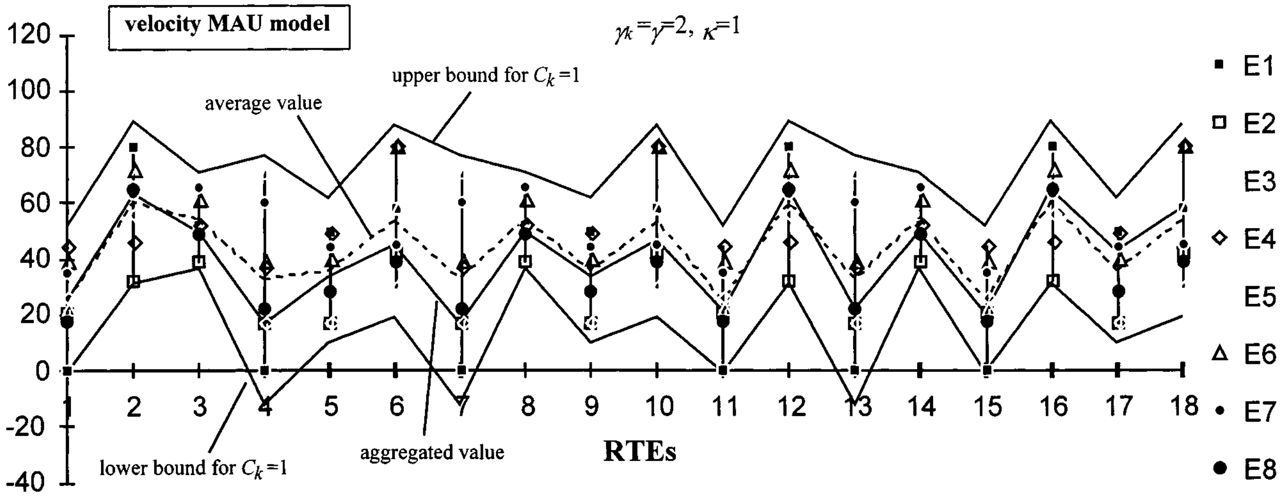


Fig. 6. Aggregation of velocity assessments made with the MAU model.

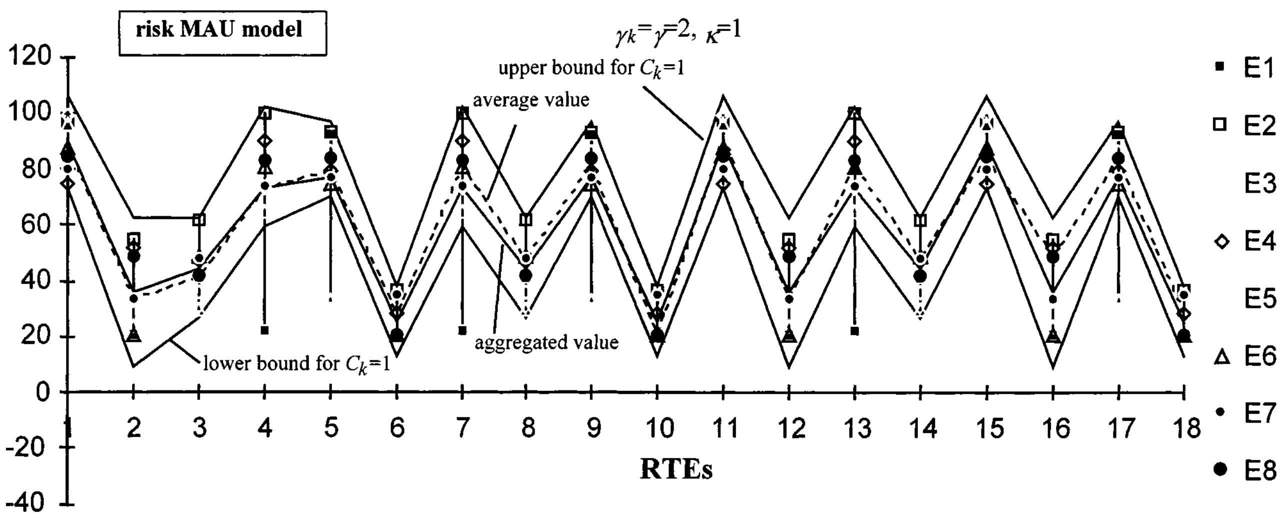


Fig. 7. Aggregation risk assessments made with the MAU model.

tings. That is, the mechanical characteristics of convergence indicate that even when the experts do not have time to discuss their differences in assessment, the model eventually converges to the weighted mean value.

2) Consistency in Cardinal Assessments: To analyze the mechanical aspects of consistency in cardinal assessments, the two extremes of most and least consistent assessments are considered. The most consistent assessments are achieved by rearranging the assessments in descending order for each RTE. In addition, the weights of the experts are held fixed. This assumption means that the long-run consistency is $c_{i|k} = 1$ for all experts i and that the group consistency is $C_k = 1$ for all RTE stages k.

Figs. 6 and 7 show the results of the group-aggregated velocity and risk values for the MAU model using data from Table I. The κ-value was chosen to be ∞ which means, according to [4], that the weight of each expert depends only on the larger of the two values $c_{i|k}$ (consistency) and $d_{i|k}$ (deviation from aggregated mean). The upper and lower bounds in Figs. 6 and 7 reflect the acceptable assessments when all experts are perfectly consistent (i.e., $C_k = 1$). It should be noted that the upper and lower bounds stand for the ranges of the double standard deviation (δ_k), since $\gamma = 2$ and $C_k = 1$ in [8].

Because of the high consistency, the assessments of velocity by all experts (Fig. 6) as well as at least seven out of eight assessments for the risk values (Fig. 7) were accepted for all 18 RTE's. Only six out of the eighteen RTE's had one risk assessment rejected (Fig. 7). This explains why the band (defined by the double standard deviation of the assessments (i.e., $\gamma = 2$) is rather large. The band of acceptable assessments depends on (1) the consistency of the group of experts, (2) the extreme values in relation to all assessments, and (3) the weights of the experts. If the weights $w_{i|k}$ were not held fixed, but computed according to (4), the band would be narrower and thus the number of acceptable assessments would be smaller.

For the other extreme case, when the experts are least consistent, the velocity values were rearranged alternatively in descending and ascending order, that is, the first expert was in the first assessment the most conservative, in the next assessment the least conservative, then again the most conservative, etc. This means that the group consistency is $C_k = 0$ for all

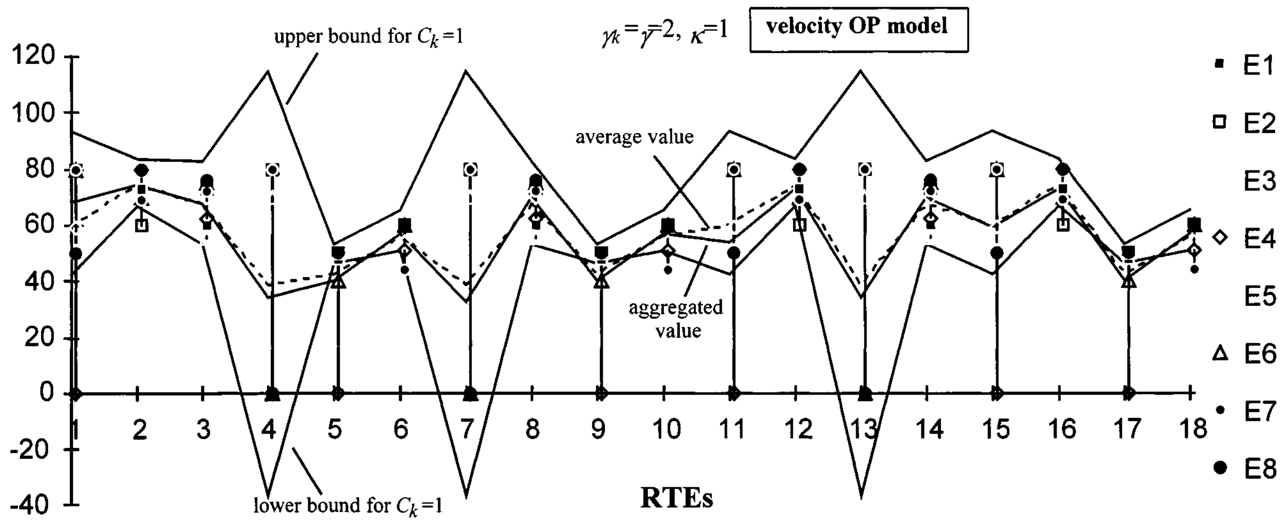


Fig. 8. Aggregation of velocity assessments made with OP model.

except the first RTE, which means that no band of acceptable assessments exists.

However, the resulting group assessments were determined by a weighted mean of all eight assessments. Consequently, the aggregated group assessments of: 1) the most consistent; 2) least consistent; and 3) nonweighted averages are almost the same.

These three special cases show that taking the simple average as the group-aggregated assessment can be misleading; that is, the assessment could be what we considered "best" (most consistent) or "worst" (least consistent). The proposed aggregation procedure discriminates for different experts' weights. For example, RTE4 in Fig. 6 shows that the group assessment can be quite lower that the simple average. The reason for this is that outliers are rejected iteratively. Thereby, an outlier is not only determined by the numeric value of the assessment but also by the expert's weight. Thus, what in numeric terms might not be an outlier, could very well be in terms of group assessment.

For the MAU velocity assessments, we computed an average group consistency for the velocity assessment, $C_k = 0.43$, and for the MAU risk assessment, $C_k = 0.39$. For the OP model, the average group consistency for the velocity assessment was very low, $C_k = 0.14$. The reason for this is that if one assesses the risks with an ordinal scale, s/he often does not assess anymore the velocity for high risk RTE's (RTE's 2, 3, and 6, as described in Section IV-A1). This is so because assessing an RTE as high risk (HR) often implicitly implies a very low velocity. In addition, if an RTE gets assigned an α risk preference, the velocity is automatically set to zero because the area affected by the RTE cannot be transited anymore.

For the velocity assessments in the MAU and OP models, and the risk assessment in the MAU model, the aggregation of cardinal values has been done similar to the case discussed in Fig. 6. The numeric risk values showed less consistency than the numeric velocity values for the MAU model.

3) Ordinal Assessments: The ordinal risks assessed in the OP model were at least sufficient; that is, a majority of the experts agreed on the tradeoff between risks and costs. For the assessment of low risk RTE's (RTE's 1, 4, and 5, as described in Section IV-A1), the assessment was always efficient; in fact the experts agreed unanimously on LR. For RTE 4, four experts assessed α preferences for the affected activities, while four assessed the affected activities as HR preferences. If only one RTE has happened, then the reroutes are the same regardless whether an α or a HR preference is assessed. However, if there are multiple RTE's, the resulting CA could be different. Thus, we would not accept such an assessment. In cases of sufficient but not efficient assessments, the group assessment is usually determined by the experts' weights. In the above example, the sum of the experts' weights determined whether the RTE must be avoided or not.

B. Choices

The CA's (routes) that resulted from the aggregated numerical values were all meaningful because a majority of the experts agreed on the CA's for all RTE's. However, because the assessments of all experts are considered, there is little meaning in debating about the quality of the reroutes resulting from the aggregated assessment. Moreover, reroutes are only suggestions both for the ORM manager (e.g., dispatcher) as well as the operator (e.g., driver) who considers the ORM manager's advice for his decision. The ORM manager always has the option to propose a different CA by "constructing" a CA which is different from the planned CA or any computationally derived CA.

In most cases, there are only two CA's based upon the proposed assessment procedure—decision-making by attribute—the planned CA (e.g., planned route) and the alternate CA (e.g., route) which avoids the actions (e.g., road segments) affected by the RTE (e.g., the most preferred reroute). This holds if the new CA is based on the aggregated assessment value. However, if for each acceptable assessment a CA is computed, some of the experts might choose to keep the planned CA while other prefer the alternate CA. If the CA based upon the aggregated assessment values is the same as the CA proposed by the majority of the experts, we recommend this CA as the solution. However, both the planned CA and the alternate CA could be displayed graphically on the graph structure (e.g., road map) with the choice based upon the group assessment highlighted for the decision-maker.

In the case of multiple RTE's, there may be more recommendations than just the planned CA and one alternate CA. We could have multiple recommended CA's—one based upon the group assessment and one based upon each expert that had an acceptable assessment. These CA's, together with the originally planned CA, could be presented to the decision-maker. Since the number of experts is small, typically less than ten (in our example it was eight), displaying the alternative CA's should not present a problem for the decision-maker. However, if we have a very complex decision situation with multiple operations (e.g., vehicles on the road) and multiple RTE's—and there is no recommended new CA (including the planned CA) that is chosen by a majority of those experts with acceptable assessments—we recommend displaying only the planned CA's and the CA chosen based upon the group assessments, but having available for display; that is, stored in the computer, the CA's chosen based upon the individual experts whose assessments were acceptable.

V. Conclusion

Operational risk management has become technologically viable and headquarters are implementing advanced technologies to improve efficiency, safety, and security of their operations. However, the decisions to change planned courses of action (such as emergency plans or transportation routes for hazardous materials) are not made by the operational manager alone. Other operational personnel such as managers and technical experts may have some responsibility and useful expertise.

In this paper, we have proposed preference aggregation and decision-making approaches both for cardinal and ordinal assessments for operational risk management. The models are adaptive due to the capability to change dynamically weights and consistency coefficients over time, and the experts' aggregated assessments are computed iteratively. At each iteration step, some of the experts are offered the possibility to reconsider their assessment. However, time pressure might cause the models to progress in a mechanical way without feedback, by eliminating step-by-step expert assessments which are inconsistent or in disagreement with the group.

The sensitivity analysis for these models using field data shows interesting insights about the influence of the parameters. Although specific parameter settings make more sense for some situations (e.g., small δ_k when time to discuss differences in assessment is available), the aggregated values seem eventually to converge. However, the use of the proposed models makes sense especially when consistency and agreement are an important consideration. Moreover, taking the simple average of the experts' assessment as the group assessment might lead to misleading results.

With the data used for the sensitivity analysis, both assessment and rerouting decisions complied with what a majority of the experts suggested. Although the sensitivity analysis was done with real field data, only mechanical aspects were subject of this research. The next step is to test the models in terms of their behavioral aspects for specific ORM problem situations. The unavailability of some experts for some RTE's, their reluctance to make or revise assessments, and the different reactions to time delay and lack of information and data will give insights in which direction the models might have to be adapted. However, the flexibility of the models, which is due to its parameters, shows that a wide range of aggregated values can be generated, all based on different assumptions. This leads us to conclude that the models encompass enough flexibility to be adapted to account for behavioral considerations. For example, the effects of feedback and learning will have to be investigated. Feedback accounts for a relative improvement of the group assessment by showing the experts the assessments and choices of the other experts. Learning, on the other hand, accounts for an absolute improvement by instructing the experts about the consequences of their assessments and decisions.

References

[1] K. J. Arrow, *Social Choice and Individual Values*, 1st, 2nd 1963 ed. New York: Wiley, 1951.

[2] T. M. Batz, "The utilization of real-time traffic information by the trucking industry," *IEEE Trans. Veh. Technol.*, vol. 40, pp. 64–67, 1991.

[3] E. Beinat, P. Nijkamp, and P. Rietveld, "Value functions for environmental pollutants: A technique for enhancing the assessment of expert judgments," *Environmental Monitoring and Assessment*, vol. 30, pp. 9–23, 1994.

[4] G. E. G. Beroggi, *Decision Modeling in Policy Management—An Introduction to the Analytic Concepts*. Boston, MA: Kluwer Academic, 1999.

[5] G. E. G. Beroggi and W. A. Wallace, "Operational control of the transportation of hazardous materials: An assessment of alternative decision models," *Management Sci.*, vol. 41, no. 12, pp. 1962–1977, 1995.

[6] G. E. G. Beroggi and W. A. Wallace, "Operational risk management: A new paradigm for risk analysis," *IEEE Trans. Syst., Man, Cybern.*, vol. 24, no. 10, pp. 1450–1457, 1994.

[7] *Handbuch zur Störfallverordnung. Bundesamt für Umwelt, Land und Landschaft*, BUWAL, Bern, Switzerland, 1991.

[8] R. T. Clemen, "Combining forecasts: A review and annotated bibliography," *Int. J. Forecasting*, vol. 5, pp. 559–583, 1989.

[9] R. M. Cooke, *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford, U.K.: Oxford Univ. Press, 1991.

[10] A. R. DeWispelare, L. T. Herren, and R. T. Clemen, "The use of probability elicitation in the high-level nuclear waste regulation program," *International Journal of Forecasting*, vol. 11, pp. 5–24, 1995.

[11] W. Edwards and L. D. Phillips, "Man as transducer for probabilities in Bayesian command and control systems," in *Human Judgments and Optimality*, M. W. Shelly, II and G. L. Bryan, Eds: Wiley, 1964, pp. 360–401.

[12] E. E. Entin and D. Serfaty, "Sequential revision of belief: An application to complex decision making situations," *IEEE Trans. Syst., Man, Cybern. A*, vol. 2, pp. 289–301, May 1997.

[13] B. Fischhoff, S. Lichtenstein, P. Slovic, L. Derby, and R. L. Keeney, *Acceptable Risk*. Cambridge, U.K.: Cambridge Univ. Press, 1981.

[14] R. M. Hogarth and H. J. Einhorn, "Order effects in belief updating: The belief adjustment models," *Cognitive Psychol.*, vol. 24, pp. 1–55, 1992.

[15] K. Keen, "European community research and technology development on advanced road transport telematics 1992–1994," *Traffic Eng. Contr.*, pp. 263–267, Apr. 1992.

[16] M. Kendall and J. D. Gibbons, *Rank Correlation Methods*, 5th ed. New York, NY: Oxford Univ. Press, 1990.

[17] H. A. Linstone and M. Turoff, *The Delphi Method, Techniques and Applications*. Reading, MA: Addison-Wesley, 1975.

[18] I. J. Myung, S. Ramamoorti, and A. D. Bailey Jr., "Maximum entropy aggregation of expert predictions," *Manage. Sci.*, vol. 42, no. 10, pp. 1420–1436, 1996.

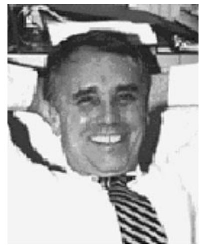
[19] J. W. Payne, J. R. Bettman, and E. J. Johnson, *The Adaptive Decision Maker*. New York: Cambridge Univ. Press, 1993.

[20] S. A. Sandri, D. Dubois, and H. W. Kalfsbeek, "Elicitation, assessment, and pooling of expert judgments using possibility theory," *IEEE Trans. Fuzzy Syst.*, vol. 3, pp. 313–335, June 1995.

[21] R. R. Yager, "On a general class of fuzzy connectives," *Fuzzy Sets Syst.*, vol. 4, pp. 235–242, 1980.

Giampiero E. G. Beroggi (S'91–M'91–SM'98) received the M.S. degree in civil engineering from ETH-Zürich, Switzerland, the M.S. degree in operations research and statistics and Ph.D. degree from Rensselaer Polytechnic Institute (RPI), Troy, NY.

He is Associate Professor of policy analysis at the School of Technology, Policy, and Management, Delft University of Technology, The Netherlands. He has taught classes in systems analysis, quantitative methods for problem solving, and risk management. His research deals with operational risk management, decision-making, and conflict resolution.



William A. Wallace (M'90–SM'96) teaches courses in the areas of information systems and decision sciences at the Department of Decision Sciences and Engineering Systems, Rensselaer Polytechnic Institute, Troy, New York. He has over 20 years experience in developing, implementing, and evaluating decision support systems. He is presently engaged in research on computer-based decision aids for emergency managers, and ethical issues in modeling.