MULTI-CRITERIA DECISION ANALYSIS IN A DECISION TREE

Freerk A. Lootsma[1]
Faculty of Mathematics and Informatics
Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
F.A.Lootsma@twi.tudelft.nl

Abstract: The outcomes at the tips of a decision tree cannot always be represented by a single numerical value on a one-dimensional axis. In many decision problems the outcomes are multi-dimensional, and their performance is expressed in a mixture of verbal statements and physical or monetary values. We propose to use the scores of cardinal methods for multi-criteria decision analysis in order to represent the relative performance of the outcomes. We evaluate the chance forks in the tree via the corresponding aggregation procedure: in the Multiplicative AHP via weighted geometric means of the scores with the outcome probabilities as exponents, and in SMART via weighted arithmetic means of the scores. The procedure is based on the idea that, within a given context, the numerical values of verbal quantifiers like *somewhat more, more,...* do not depend on what we compare, whether it is relative importance or relative likelihood.

1. Introduction

A decision tree is a well-known tool to model and to evaluate a decision process which consists of an alternating sequence of actions and uncertain consequences (Keeney and Raiffa, 1976; Von Winterfeldt and Edwards, 1986). We concern ourselves here with a tree of the simplest possible form. At the initial node, the decision maker has the choice between the sure outcome A_0 on the one hand and the chance fork with the uncertain, mutually exclusive outcomes A_1 and A_2 on the other. It is customary in Decision Analysis to express the consequences in monetary values. If the value of A_0 is situated between the values of A_1 and A_2, the decision maker is supposed to estimate the probabilities p_1 and p_2 of the respective outcomes A_1 and A_2. Thereafter he/she faces the question of how to judge the chance fork, that is, how to aggregate the values of A_1 and A_2 and their associated probabilities into a single quantity that can properly be compared with a single quantity representing the sure outcome A_0. The aggregation procedure is the core of the backfolding process in larger decision trees whereby the decision maker repeatedly replaces chance forks by single quantities. During the process he/she works backwards through the tree until he/she can identify the action to be chosen at the initial node and at subsequent decision nodes.

In the aggregation procedure the decision maker first determines a range (L, H) of relevant values on the axis in question, to be used uniformly throughout the analysis. Thereafter, he/she replaces each outcome A_j ($j = 0, 1, 2$) by a lottery or gamble with the probability u_j to obtain the high endpoint H (generally high benefits) of the range and the probability $1 - u_j$ to obtain the low endpoint L (low benefits or even losses). The decision maker chooses the so-called utility u_j in such a way that he/she is indifferent between the monetary value of A_j and the gamble. Finally, he/she compares the utility u_0 of A_0 with the subjective expected utility

$$p_1 u_1 + p_2 u_2$$

of the chance fork. One usually observes that the decision maker is risk-averting, in the sense that the monetary value of A_j is lower than the expected value

$$u_j H + (1 - u_j) L$$

---

of the equivalent gamble with the probability u_j. The opposite phenomenon is also known, however: if the monetary value of A_j is higher than the expected value, the decision maker is risk-seeking. The discrepancy between the monetary value of A_j and the expected value of the equivalent gamble motivated the development of Multi-Attribute Utility Theory (MAUT).

In this paper, we start from the hypothesis that there is a particular reason for the discrepancy. Confronted with a gamble which leads either to high benefits or to low benefits or even losses, the decision maker faces a multi-criteria decision problem anyway, whether the original outcomes have one-dimensional consequences or not. He/she has to weigh, not only the benefits versus the losses with due regard to his/her financial position, but also the consequences for his/her status and prestige and for his/her relations with other people, for instance. Instead of establishing the utilities of the outcomes in a given problem, the decision maker might as well concentrate on the underlying reasons why he/she is risk-averting or risk-seeking and on the analysis of the underlying multi-criteria decision problem.

We accordingly consider here the aggregation of outcomes A_0, A_1, and A_2 with multi-dimensional consequences, under a finite number of performance criteria $C_1,...., C_m$. Our objective is to assess the chance fork with the outcomes A_1 and A_2 and their associated probabilities. To illustrate matters, we suppose that we have the following problem. A decision maker visits a number of dealers to select a car. Late in the afternoon he/she finds a car A_0 with an acceptable price and an acceptable reliability (for the time being, we ignore the other attributes of cars). The decision maker can now sign the contract, but he/she may also continue the search on the next day. On the basis of previous experience he/she estimates that there is a chance p_1 to find a cheaper, somewhat less reliable car A_1, and a chance $p_2 = 1 - p_1$ to find a somewhat more expensive and more reliable car A_2. If he/she continues the search, however, the deal with the car A_0 is over. How to decide now?

We consider two cardinal methods for multi-criteria decision analysis in order to aggregate the multi-dimensional consequences: the Multiplicative AHP '(Lootsma, 1993), a variant of 'the Analytic Hierarchy Process (Saaty, 1980), and the Simple Multi-Attribute Rating Technique (Von Winterfeldt and Edwards, 1986). In the Multiplicative AHP we are concerned with ratios of preference intensities collected via pairwise comparisons. SMART is based on distance information collected via direct rating.

The organisation of the paper is as follows. First, using the car-selection example we explain how we categorize the monetary and the physical values of the alternative cars within the context of the decision problem. This enables us to quantify comparative judgemental statements like *somewhat more, definitely more,* and *much more* expensive or reliable (section 2), and thereafter we can easily make a transition to the gradations of relative strength of preference (section 3). Next, we discuss the relative importance of the criteria and the numerical scale to quantify their relative importance (section 4). An intermezzo is given by a numerical example to prepare the ground for the proposed procedure (section 5). Finally, we aggregate the impact scores of the alternative cars A_1 and A_2 using the associated probabilities p_1 and p_2 as well as the weights of the criteria C_1 and C_2, the price and the reliability of the alternative cars respectively (section 6). The aggregated result, a weighted geometric mean of the impact scores with the outcome probabilities and the criterion weights in the exponents, is not an expected value but an expression reflecting the relative importance of the outcome probabilities (section 7).

2. Categorical judgement on a one-dimensional axis

Let us start with the subjective evaluation of the cars under the price criterion. In the Multiplicative AHP, we assume that the decision maker is only prepared to consider alternatives with prices between a desired lower bound P_{min}, the price to be paid anyway for the cars which he/she seriously has in mind, and an upper bound P_{max}, the price that he/she cannot or does not really want to exceed. In order to model the relative preference for alternative A_j respect to A_k we categorize the prices which are in principle acceptable. We first "cover" the range (P_{min}, P_{max}) by the grid with the geometric sequence of points

$$P_\mu = P_{min} + (P_{max} - P_{min}) * 2^\mu/64, \quad \mu = 0, 1,...,6.$$

In what follows we take P_μ to stand for the μ-th price category and the integer-valued parameter μ for its order of magnitude, which is given by

$$\mu = {}^2\text{log} \{64 * [P_\mu - P_{min}]/[P_{max} - P_{min}]\}. \tag{1}$$

Categorization of the prices means that each price in or slightly outside the range (P_{min}, P_{max}) is supposed to "belong" to a particular category, namely the category represented by the nearest P_μ. Of course, categorization could more appropriately be modelled via fuzzy-set theory, but for the time being we ignore this. We refer to the cars of the category P_0 as the cheap ones within the given context, and to the cars of the categories P_2, P_4, and P_6 as the *somewhat more, more*, and *much more* expensive ones. At the odd-numbered grid points P_1, P_3, and P_5 the decision maker hesitates between the two adjacent gradations of expensiveness. If necessary, we can also introduce the category P_8 of *vastly more* expensive cars which are situated beyond the range, as well as the category P_7 if the decision maker hesitates between *much more* and *vastly more* expensiveness.

The even-numbered grid points are the so-called major grid points designating the major gradations of expensiveness. They constitute a geometric sequence in the range (P_{min}, P_{max}) with progression factor 4. If we also take into account the odd-numbered grid points corresponding to hesitations, we have a geometric sequence with progression factor 2. We employed a number of examples such as the progression of historical periods and planning horizons, the classification of nations according to size, and the perception of light and sound intensities, in order to show that human beings follow the above pattern in many unrelated areas when they categorize an interval (Lootsma, 1993).

Suppose that the prices of the cars A_j and A_k belong to the categories represented by

$$P_{\mu_j} \text{ and } P_{\mu_k}$$

respectively. In the Multiplicative AHP, the relative preference for A_j with respect to A_k is expressed by the inverse ratio of the price increments above P_{min} so that it can be written as

$$[P_{\mu_k} - P_{min}]/[P_{\mu_j} - P_{min}] = 2^{\mu_k - \mu_j}. \tag{2}$$

By this definition, a car in the price category P_0 is 4 times more desirable than a car in the category P_2. The last-named car is said to be somewhat more expensive. Thus, we identify the ratio 4:1 with weak preference. Similarly, we identify the ratio 16:1 with definite preference, etc. The relative preference depends strongly on P_{min} and weakly on P_{max}. When P_{max} increases, two prices which initially belong to different price categories will tend to belong to the same one.

It is common in psychophysics, when a ratio of two stimulus intensities is discussed, to use the difference of the corresponding orders of magnitude. One records the logarithm of the ratio as a difference of values on the so-called decibel scale. We can easily follow this mode of operation via the method of direct scoring which is usually employed in SMART (Lootsma, 1993, see the figures 1 and 2). The price categories are therefore recorded on a so-called qualitative scale, such as the familiar seven-point scale with the grades 4, 5, ..., 10, between the grade 4 for poor performance that can nevertheless be compensated by a more satisfactory performance elsewhere (current practice in our schools), and the grade 10 for excellent performance. The price category around P_μ is represented by the grade $g = 10 - \mu$, so that the major gradations of comparative judgement are designated by 10 (cheap car, excellent, $\mu = 0$), 8 (somewhat more expensive car, good, $\mu = 2$), 6 (fair, $\mu = 4$), and 4 (poor, $\mu = 6$). Obviously, the relative preference for the car A_j with respect to A_k can now be expressed by the difference of grades

$${}^2\text{log} ([P_{\mu_k} - P_{min}]/[P_{\mu_j} - P_{min}]) = \mu_k - \mu_j = g_j - g_k. \tag{3}$$

253

Because we only work with differences of grades, we can replace the SMART scale 4, ..., 10 by the familiar seven-point scale 1,...,7 which is frequently used in the social sciences.

When the cars are judged under the price criterion, the desired target is at the lower limit P_{min} of the interval of possible prices. Under the reliability criterion, however, the desired target is at the upper limit. Numerical data to estimate the reliability are usually available. Consumer organizations collect information about many types and models of cars which follow the prescribed maintenance procedures, and they publish the frequencies of technical failures in the first three or five years or the proportions of users who report serious difficulties in those periods (see the Annual Auto Issue of the Consumer Reports, US Consumer Union). Let us suppose that the decision maker only considers cars with a reliability of at least R_{min}, so that he/she is restricted to the interval (R_{min}, R_{max}) with R_{max} usually set to 100 %. Following the mode of operation just described, we obtain the grid points

$$R_\mu = R_{max} - (R_{max} - R_{min}) * 2^\mu / 64, \quad \mu = 0, 1,...,6.$$

In general, the alternatives are compared with respect to the desired target. The relative performance is inversely proportional to the distance from the target. The reader can easily verify this in the two examples just given. If we take the symbols

$$R_{\mu_j} \text{ and } R_{\mu_k}$$

to denote the reliability of the alternative cars A_j and A_k respectively, then the inverse ratio

$$[R_{max} - R_{\mu_k}]/[R_{max} - R_{\mu_j}] = 2^{\mu_k - \mu_j}$$

represents the relative preference for A_j with respect to A_k under the reliability criterion. The qualifications "somewhat cheaper" and "somewhat more reliable" imply that the inverse ratio of the distances to the respective targets is 4:1. The relationship between the order of magnitude μ and the reliability category R_μ takes the explicit form

$$\mu = {}^2\log \{64 * [R_{max} - R_\mu]/[R_{max} - R_{min}]\}. \tag{4}$$

There may be criteria where the categorization starts, not from the desired target at one end of the range, but from the opposite end because the target is hazy. An example is given by the categorization of the maximum velocities. Sometimes, the target is found in the interior of the range, which usually means that the decision maker tries to satisfy a hidden objective. These issues have been discussed elsewhere (Lootsma, 1995).

The scale 1, 2, 4, 8, 16, 32, 64 represents a partitioning of a given range of acceptable performances (acceptable prices and acceptable reliabilities, for instance). This partitioning, which we found in the subjective measurement of time (historical periods, planning horizons), space (size of nations), light intensities, and sound intensities, can be achieved via extremely simple steps: bisections and duplications. One does not need a particular unit of measurement to carry out these operations. Bisections and duplications, the progression factor 2, they will certainly remind the reader of the partitioning of time in musical compositions: tone durations as bisections or duplications of the rhythm's fundamental beat, and repetitions of phrases in the melody. Bisections and duplications seem to be so deeply embedded in human behaviour that they may have a physiological basis in the neural system. It is beyond the scope of the present paper, however, to pursue this issue here.

3. The Multiplicative AHP and SMART

So far, we have been working on different dimensions such as consumer price and reliability. Judgemental statements like "somewhat more expensive" and "somewhat more reliable" cannot be aggregated, however, unless we make a transition to a new, common dimension. That is the reason why we take the expression "somewhat more reliable" to stand for "somewhat more desirable" under the criterion of reliability. Similarly, we assume that the expression "somewhat more expensive" may

stand for "somewhat less desirable" under the price criterion. We shall assume that, under each of the respective criteria, the desirability of the alternatives varies over the same interval on the new, common dimension, at least during the decision process at hand.

In the basic evaluation step of the Multiplicative AHP, two alternatives A_j and A_k are presented to the decision maker who is requested to express his/her comparative judgement between the two under a particular criterion C_i. With the results of the previous section, we have in general two different ways to collect his/her preference information.

(a). We can ask him/her to consider the axis corresponding to the criterion C_i and to specify the endpoints of the range of acceptable values. Next, we identify the judgement categories on the range and the corresponding orders of magnitude. Thereafter, we can immediately estimate his/her preference r_{ijk} for A_j with respect to A_k (the ratio of the subjective values of the two alternatives) under the criterion C_i by

$$r_{ijk} = 2^{\mu_{ik} - \mu_{ij}}, \tag{5}$$

where μ_{ij} and μ_{ik} denote the orders of magnitude of the respective alternatives under C_i. This context-related scaling is only feasible when the performance of the alternatives under criterion C_i is expressed in physical or monetary values. It is particularly attractive when the number of alternatives is large (Van Gennip, Hulshof, and Lootsma, 1996).

(b). We can ask the decision maker to express his/her graded comparative judgement in words, that is, to state whether he/she is indifferent between the two alternatives under the given criterion, or whether he/she has a weak, a definite, or a strong preference for one of the two. Thereafter, we set the numerical estimate r_{ijk} of his/her relative preference for A_j with respect to A_k under C_i to

$$r_{ijk} = 2^{\delta_{ijk}},$$

where the symbol δ_{ijk} stands for an integer-valued index designating the major gradations of his/her comparative judgement. Thus, $\delta_{ijk} = 0$ designates indifference, $\delta_{ijk} = \pm 2$ designates weak preference, etc. Obviously, even values are assigned to δ_{ijk} in order to represent major gradations, and odd values are used when the decision maker hesitates between two adjacent major gradations. This type of elicitation is always applied when the performance of the alternatives under criterion C_i cannot be measured in physical or monetary values.

Usually, we bundle the estimates r_{ijk} into pairwise-comparison matrices, one for each criterion. The impact score a_{ij} of alternative A_j under criterion C_i is given by the geometric mean of the j-th row in the corresponding pairwise-comparison matrix. Thus,

$$a_{ij} = \sqrt[n]{\prod_{k=1}^{n} r_{ijk}}. \tag{6}$$

The final score f_j, expressing the global preference for alternative A_j under all criteria simultaneously on the new common dimension of desirability, is calculated according to the geometric-mean aggregation rule

$$f_j = \prod_i a_{ij}^{c_i}, \tag{7}$$

where the symbol c_i stands for the weight of criterion C_i. Since we are working with ratio information, the impact scores and the final scores of the alternatives are not unique. They can be normalized so that they sum to 1. The ratio a_{ij}/a_{ik} which expresses the relative preference for A_j with respect to A_k under criterion C_i is unique, however, as well as the ratio f_j/f_k which globally expresses the relative preference for A_j with respect to A_k.

In SMART, the calculations are much simpler. Considering two alternatives A_j and A_k under criterion C_i, with the grades g_{ij} and g_{ik} assigned to them, we take the quantity

$$r_{ijk} = 2^{g_{ij} - g_{ik}}$$

to estimate their relative desirability

$$2^{\mu_{ik} - \mu_{ij}},$$

because g_{ij} and g_{ik} are estimates of $10 - \mu_{ij}$ and $10 - \mu_{ik}$ respectively. The final scores via SMART are accordingly given by the arithmetic-mean aggregation rule

$$s_j = \sum_{i=1}^{m} c_i g_{ij}.$$

A ratio of AHP scores and the corresponding difference of SMART grades are connected by the relationship

$$\frac{a_{ij}}{a_{ik}} = 2^{g_{ij} - g_{ik}},$$

whence

$$\frac{f_j}{f_k} = \prod_{i=1}^{m} \left(\frac{a_{ij}}{a_{ik}} \right)^{c_i} = 2^{\Delta_{jk}},$$

where

$$\Delta_{jk} = \sum_{i=1}^{m} c_i g_{ij} - \sum_{i=1}^{m} c_i g_{ik} = s_j - s_k.$$

This shows that the arithmetic-mean aggregation rule in SMART is logarithmically related to the geometric-mean aggregation rule in the Multiplicative AHP.

4. The relative importance of the criteria

With the above aggregation rules we derived scale values for the gradations of the relative importance of the criteria via an imaginary experiment. First, we established extreme values for the ratio ω of any pair of criterion weights. The imaginary decision maker was supposed to consider two real or imaginary alternatives A_j and A_k and two criteria such that his/her preference for A_j over A_k under the first criterion C_f was roughly equal to his/her preference for A_k over A_j under the second criterion C_s. These inverse preferences were estimated by

$$2^{\delta_{jk}} \text{ and } 2^{-\delta_{jk}}$$

respectively, where δ_{jk} designates the selected gradation of his/her comparative judgement. Next, assuming that these preferences do not depend on the performance of A_j and A_k under the remaining criteria, we expressed the decision maker's preference for the two alternatives under the two criteria C_f and C_s simultaneously by

$$2^{\theta_{jk}}.$$

256

It will be clear that $|\theta_{jk}| < |\delta_{jk}|$ since the preference for A_j over A_k under the two criteria simultaneously cannot be greater than the preference for A_j over A_k under one of the criteria individually. It was easy to verify that the ratio ω varies roughly between $^1/16$ and 16 when δ_{jk} varies between -8 and 8, and θ_{jk} between $-|\delta_{jk}|$ and $|\delta_{jk}|$. The rather extreme value $\omega = 16$ was obtained when $\delta_{jk} = 8$ and $\theta_{jk} = 7$, which means that the *very strong* preference for A_j over A_k under the first criterion C_f almost completely wipes out the equally strong but inverse preference under the second criterion C_s. So, a ratio of 16:1 may be taken to stand for *vastly higher* importance.

Next, we converted the gradations for the relative importance of the criteria into numerical values on the assumptions that (a) the number of gradations to express the relative importance of the criteria equals the number of gradations to express the relative preference for the alternatives, and (b) the numerical values associated with these gradations constitute a sequence with geometric progression. In the extreme case, where a *vastly higher* preference under the first criterion is practically wiped out by a *vastly higher* preference under the second criterion, we accordingly refer to the relative importance of the first criterion with respect to the second one as *vastly more*.

A simple geometric sequence of values between $^1/16$ and 16, with echelons corresponding to indifference, weak, strict, strong, and very strong preference, is the sequence $^1/16$, $^1/8$, $^1/4$, $^1/2$, 1, 2, 4, 8, 16 with progression factor 2. Hence, we obtain the following geometric scale for the major gradations in the pairwise comparison of the first criterion C_f with respect to the second criterion C_s:

16	C_f vastly more important than C_s,	weights	0.95 and 0.05,
8	C_f much more important than C_s,		0.90 and 0.10,
4	C_f more important than C_s, ·		0.80 and 0.20,
2	C_f somewhat more important than C_s,		0.66 and 0.33,
1	C_f as important C_s,		0.50 and 0.50,
$^1/16$	C_f vastly less important than C_s,		0.05 and 0.95.

Note that we have a geometric sequence with progression factor $\sqrt{2}$ if we also allow for threshold gradations to express hesitations between two adjacent qualifications. It will be obvious that the criteria can be evaluated in the same way, via pairwise comparisons, as the alternatives. In the basic experiment we present two criteria to the decision maker and we ask him/her to state whether they are *equally* important or whether one of the two is *somewhat more, more, much more,* or *vastly more* important than the other. The judgemental statements are subsequently converted into the corresponding values on the above geometric scale. Finally, we calculate and we normalize the geometric row means in the pairwise comparison matrix in order to obtain the criterion weights.

5. The illustrative example

Let us now illustrate the calculations of the sections 2 and 3 via the car-selection example which we introduced in section 1. Under the assumption that the decision maker is only prepared to consider prices in the range between Dfl 20.000 (US$ 12.000)and Dfl 30.000 (US$ 18.000), and reliabilities between 95 % and 100 % (by these ranges we model the context of the decision problem), we obtain the following grid points:

Price	Reliability	Performance	Grade
20.000	100	excellent	10
20.300	99.9		9
20.600	99.7	good	8
21.200	99.4		7
22.500	98.8	fair	6
25.000	97.5		5
30.000	95.0	poor	4

Let us now suppose that the alternative cars have the prices and the reliabilities displayed in the below table (the so-called performance tableau), with the associated SMART grades between brackets.

Criteria	Alternative A_0	Alternative A_1	Alternative A_2
C_1: Price	22.500 (6)	21.200 (7)	25.000 (5)
C_2: Reliability	98.8 (6)	97.5 (5)	99.4 (7)

The pairwise-comparison matrices under the price criterion and the reliability criterion are given by

$$\begin{pmatrix} 1 & \frac{1}{2} & 2 \\ 2 & 1 & 4 \\ \frac{1}{2} & \frac{1}{4} & 1 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 2 & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{4} \\ 2 & 4 & 1 \end{pmatrix}$$

respectively. Let us assume, for the time being, that the two criteria have equal weights. The normalized impact scores and the normalized final scores, calculated according to the procedure just described, may be found in the below table, again with the associated SMART grades between brackets.

Criteria		Alternative A_0	Alternative A_1	Alternative A_2
C_1: Price	0.5	0.29 (6)	0.57 (7)	0.14 (5)
C_2: Reliability	0.5	0.29 (6)	0.14 (5)	0.57 (7)
Final scores		0.33 (6)	0.33 (6)	0.33 (6)

In the remainder of this paper we will use the above data to illustrate the proposed procedure for multi-criteria decision analysis in a decision tree.

6. Evaluation of a chance fork

The above example is so symmetric that the decision maker must be indifferent between the sure outcome A_0 and the chance fork with the outcomes A_1 and A_2, whatever the method of analysis. When we use expected values to aggregate the outcomes, SMART does lead to such a result, both at the level of the grades assigned to the alternatives and at the level of the final scores. With $p_1 = p_2 = 0.5$ we obtain

$$7p_1 + 5p_2 = 6,$$
$$5p_1 + 7p_2 = 6,$$
$$6p_1 + 6p_2 = 6.$$

Expected values of the AHP scores do not work so nicely. At the level of the impact scores of the alternatives we obtain

$$0.57\,p_1 + 0.14\,p_2 \neq 0.29,$$
$$0.14\,p_1 + 0.57\,p_2 \neq 0.29,$$

for $p_1 = p_2 = 0.5$. Since the Multiplicative AHP is an exponential version of SMART, however, we propose to use the probabilities p_1 and p_2 as exponents in a weighted geometric mean of scores in order to aggregate a chance fork. This yields the desired result, because

$$0.57^{p_1} \cdot 0.14^{p_2} = 0.29,$$
$$0.14^{p_1} \cdot 0.57^{p_2} = 0.29,$$
$$0.33^{p_1} \cdot 0.33^{p_2} = 0.33,$$

when $p_1 = p_2 = 0.5$. Since the ability of human beings to estimate the probabilities of uncertain events is questionable, we propose to consider p_1 and p_2 as coefficients expressing the relative importance of

258

the alternatives in the chance fork. In the next section we show that this is a workable hypothesis in the analysis of decision trees in general.

Let us now briefly summarize the results of the present section. Suppose that we are concerned with a decision tree presenting the choice between the sure outcome A_0 on the one hand and the chance fork with the mutually exclusive outcomes A_j, $j = 1,...,n$, and the associated probabilities p_j, $j = 1,...,n$, on the other. Furthermore, the alternatives are to be judged under the criteria C_i, $i = 1,...,m$, with corresponding criterion weights c_i. Working with SMART, we compare the arithmetic means

$$\sum_{i=1}^{m} c_i g_{i0} \text{ and } \sum_{i=1}^{m} \sum_{j=1}^{n} c_i g_{ij} p_j, \tag{8}$$

where g_{ij} stands for the grade assigned to alternative A_j under criterion C_i, and we choose the arithmetic mean with the highest value. Obviously, the expressions in (8) have the form of an expected value (in fact expected values of logarithms). Working with the Multiplicative AHP, we compare the geometric means

$$\prod_{i=1}^{m} a_{i0}^{c_i} \text{ and } \prod_{i=1}^{m} \prod_{j=1}^{n} a_{ij}^{c_i p_j}, \tag{9}$$

where a_{ij} designates the impact score of alternative A_j under criterion C_i, and we choose the geometric mean with the highest value. The expressions in (9) cannot be interpreted as expected values. That is the starting point of the considerations to follow.

7. The relative importance of the outcome probabilities

The ability of human beings to estimate probabilities should not be over-estimated (see the last paragraph of this section). Hence, a probabilistic model for the uncertainties in an actual decision problem may be a hazardous tool, regardless of whether the decision makers are requested to estimate the probabilities numerically or in verbal terms. We propose to follow a different route, however. The formulas in (8) and (9) show that, mathematically, the outcome probabilities and the criterion weights play the same role in the aggregation procedure. We obtained a similar result in group aggregation procedures incorporating the relative power of the decision makers: the criterion weights and the power coefficients have identical positions in the structure of the aggregation formulas (Barzilai and Lootsma, 1994). Hence, we take the outcome probabilities to represent coefficients of importance. In other words, if A_1 is *somewhat more* likely to occur than A_2, we take it to be *somewhat more* important in the aggregate quantity representing the chance fork, etc. Thus, we associate the gradations of relative likelihood with the same scale values as the gradations of relative importance, so that

A_1 as likely as A_2,	odds 1,	$p_1 = 0.50$,	$p_2 = 0.50$,
A_1 somewhat more likely than A_2,	odds 2,	$p_1 = 0.66$,	$p_2 = 0.33$,
A_1 more likely than A_2,	odds 4,	$p_1 = 0.80$,	$p_2 = 0.20$,
A_1 much more likely than A_2,	odds 8,	$p_1 = 0.90$,	$p_2 = 0.10$,
A_1 vastly more likely than A_2,	odds 16,	$p_1 = 0.95$,	$p_2 = 0.05$.

The underlying idea is that comparative verbal quantifiers like *somewhat more, more, much more,* and *vastly more* have numerical values which do not depend on what we compare: the relative importance of the criteria, the relative power of the decision makers, or the relative likelihood of the outcomes. By this uniformity, comparative verbal quantifiers lubricate human communication. Imagine how irritating it would be if we always had to use numerical values instead of vague verbal terms with a quantitative connotation. The fact that comparative verbal quantifiers are usually well-understood is an argument in favour of the hypothesis that they have imprecise numerical values with a uniform validity.

With the above numerical scaling of relative likelihood we can ask the decision maker to estimate the outcome probabilities $p_1,....., p_n$ via a method of pairwise comparisons. In the basic experiment we

present a pair of possible outcomes to the decision maker, whereafter we can ask him/her to state whether they are *equally* likely or whether one of the two is *somewhat more, more, much more,* or *vastly more* likely to occur than the other. Next, we can convert the statements into elements of a pairwise-comparison matrix from which we eventually extract the outcome probabilities via the calculation of geometric row means (see also the last paragraph of section 4).

The above method has not been considered in the literature, although the significance of verbal probability estimates has extensively been studied in the social sciences (Pepper and Prytulak, 1974; Hammerton, 1976; Budescu and Wallsten, 1985; Wallsten, Filenbaum, and Cox, 1986; Brun and Teigen, 1988; Teigen, 1988; Weber and Hilton, 1990; Hendrickx, 1991; Mullet and Rivet, 1991). In the reported experiments the respondents were usually requested to assign values to probability terms within the context of an uncertain situation or to select the most suitable probability term from a given list in order to characterize the situation. The authors discovered a high degree of context-dependent variability of the meaning of these terms, as well as a weak relationship between these terms and the actual probabilities in the given situations. Particularly Teigen (1988) reports an amazing degree of innumeracy and a serious lack of probabilistic thinking among the respondents when they were requested to estimate the probabilities of a number of mutually exclusive outcomes (this is precisely what a decision maker has to do in a chance fork). On the other hand, Brun and Teigen (1988) conclude that the interpretation variability does not seem to pose a serious communication problem. Pairwise comparisons of outcome probabilities were not included in these studies, however. A method of pairwise comparisons can be used to estimate the ratios of the outcome probabilities. Thereafter, the results can be normalized to guarantee that the outcome probabilities sum to 1 or to 100%.

8. Epilogue

In the present paper, we proposed to aggregate the multi-dimensional consequences in a chance fork in such a manner that the results are compatible, whether we use the Multiplicative AHP with the geometric-mean aggregation rule, or its logarithmic counterpart SMART with the arithmetic-mean aggregation rule. The aggregation can be carried out in any order: first over the impact scores and thereafter over the criteria, or vice versa (see the double summation in (8) and the double multiplication in (9)). We also made it plausible that verbal quantifiers like *somewhat more,..., vastly more* have imprecise numerical values which do not depend on the type of comparative judgement where they are used. Hence, they constitute a vehicle for communication, provided that we use the same quantifiers throughout the comparative assessment, regardless of whether we compare performance criteria, decision makers, or outcome probabilities.

References

1. Barzilai, J., and Lootsma, F.A. (1994), "Power Relations and Group Aggregation in the Multiplicative AHP and SMART". Report 94-07, The University of Michigan, Mechanical Engineering and Applied Mechanics, Ann Arbor, Michigan, USA. To appear in the *Journal of Multi-Criteria Decision Analysis.*

2. Brun, W., and Teigen, K.H. (1988), "Verbal Probabilities: Ambiguous, Context-Dependent, or Both?". *Organizational Behavior and Human Decision Processes* 41, 390 - 404.

3. Budescu, D.V., and Wallsten, T.S. (1985), "Consistency in Interpretation of Probabilistic Phrases. *Organizational Behavior and Human Decision Processes* 36, 392 - 405.

4. Gennip, C.E.G. van, Hulshof, J.A.M., and Lootsma, F.A. (1996), "A Multi-Criteria Evaluation of Diseases in a Study for Public-Health Planning". TWI Report 96-53, Delft University of Technology, Delft, The Netherlands.

5. Hammerton, M. (1976), "How Much is a Large Part?". *Applied Ergonomics* 7, 10 -12.

6. Hendrickx, L. (1991), *"How versus How Often"*. PhD Thesis, Department of Psychology, University of Groningen, The Netherlands.

7. Keeney, R.L., and Raiffa, H. (1976), *"Decisions with Multiple Objectives: Preferences and Value-Trade Offs"*. Wiley, New York.

8. Lootsma, F.A. (1993), "Scale Sensitivity in a Multiplicative Variant of the AHP and SMART". *Journal of Multi-Criteria Decision Analysis* 2, 87 - 110.

9. Lootsma, F.A. (1995), "Fuzzy-Set Theory as a Mathematical Model for Imprecise Human Judgement and Decision Making". Lecture Notes a197, Faculty TWI, Delft University of Technology, Delft, The Netherlands. In an extended form to be published by Kluwer Academic Publishers.

10. Mullet, E., and Rivet, I. (1991), "Comprehension of Verbal Probability Expressions in Children and Adolescents". *Language and Communication* 11, 217 - 225.

11. Pepper, S., and Prytulak, L.S. (1974), "Sometimes Frequently means Seldom: Context Effects in the Interpretation of Quantitative Expressions". *Journal of Research in Personality* 8, 95 - 101.

12. Saaty, T.L. (1980), *"The Analytic Hierarchy Process, Planning, Priority Setting, Resource Allocation"*. McGraw-Hill, New York.

13. Stewart, T.J. (1992), "A Critical Survey on the Status of Multiple Criteria Decision Making Theory and Practice". *Omega* 20, 569 - 586.

14. Teigen, K.H. (1988), "When are Low-Probability Events judged to be Probable?. Effects of Outcome-Set Characteristics on Verbal Probability Estimates". *Acta Psychologica* 67, 157 - 174.

15. Wallsten, T.S., Filenbaum, S., and Cox, J.A. (1986), "Base Rate Effects on the Interpretations of Probability and Frequency Expressions". *Journal of Memory and Language* 25, 571 - 587.

16. Weber, E.U., and Hilton, D.J. (1990), "Contextual Effects in the Interpretations of Probability Words: Perceived Base Rate and Severity of Events". *Journal of Experimental Psychology: Human Perception and Performance* 16, 781 - 789.

17. Winterfeldt, D. von, and Edwards, W. (1986), *"Decision Analysis and Behavioral Research"*. Cambridge University Press, Cambridge, UK.