D

Dantzig-Wolfe Decomposition Algorithm

A variant of the simplex method designed to solve block-angular linear programs in which the blocks define subproblems. The problem is transformed into one that finds a solution in terms of convex combinations of the extreme points of the subproblems.

See

- ► Block-Angular System
- ► Decomposition Algorithms

References

Dantzig, G. (1963). *Linear programming and extensions*. Princeton, NJ: Princeton University Press.

Dantzig, G., & Thapa, M. (2003). *Linear programming 2: Theory and extensions*. New York: Springer.

Dantzig, G., & Wolfe, P. (1960). Decomposition principle for linear programs. *Operations Research, 8*(1), 101–111.

Data Envelopment Analysis

William W. Cooper
The University of Texas at Austin, Austin, TX, USA

Introduction

DEA (Data Envelopment Analysis) is a data oriented approach for evaluating the performance of a collection of entities called DMUs (Decision Making Units) which are regarded as responsible for converting inputs into outputs. Examples of its uses have included hospitals and U.S. Air Force Wings, or their subdivisions, such as surgical units and squadrons. The definition of a DMU is generic and flexible. The objective is to identify sources and to estimate amounts of inefficiency in each input and output for every DMU included in a study. Uses that have been accommodated include: (i) discrete periods of production in a plant producing semiconductors in order to identify when inefficiency occurred; and (ii) marketing regions to which advertising and other sales activities have been directed in order to identify where inefficiency occurred. Inputs as well as outputs may be multiple and each may be measured in different units.

A variety of models have been developed for implementing the concepts of DEA, for example, the following dual pair of linear programming models:

$$\begin{aligned} \min\ h_0 &= \theta_0 - \varepsilon\left(\sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+\right) \\ \text{subject to } 0 &= \theta_0 x_{i0} - \sum_{j=1}^{n} x_{ij}\lambda_j - s_i^- \\ y_{r0} &= \sum_{j=1}^{n} y_{rj}\lambda_j - s_r^+ \\ 0 &\le \lambda_j,\ s_r^+,\ s_i^- \end{aligned} \tag{1a}$$

and

$$\begin{aligned} \max\ y_0 &= \sum_{r=1}^{s} \mu_r y_{r0} \\ \text{subject to } 1 &= \sum_{i=1}^{m} v_i x_{i0} \\ 0 &\ge \sum_{r=1}^{s} \mu_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \\ \varepsilon &\le \mu_r,\ v_i \end{aligned} \tag{1b}$$

S.I. Gass, M.C. Fu (eds.), *Encyclopedia of Operations Research and Management Science*,
DOI 10.1007/978-1-4419-1153-7, © Springer Science+Business Media New York 2013

where x_{ij} = observed amount of input i used by DMU$_j$ and y_{rj} = observed amount of output r produced by DMU$_j$, with $i = 1, \ldots, m$; $r = 1, \ldots, s$; $j = 1, \ldots, n$. All inputs and outputs are assumed to be positive. (This condition may be relaxed (Charnes et al. 1991).

Efficiency

The orientation of linear programming has changed here from ex-ante uses, for planning, and apply it to choices already made ex-post, for purposes of evaluation and control. To evaluate the performance of any DMU, (1) is applied to the input–output data for all DMUs in order to evaluate the performance of *each* DMU in accordance with the following definition:

> *Efficiency — Extended Pareto-Koopmans Definition* : Full (100%) efficiency is attained by any DMU if and only if none of its inputs or outputs can be improved without worsening some of its other inputs or outputs.

This definition has the advantage of avoiding the need for assigning a priori weights or other measures of relative importance to any input or output. In most management or social science applications, the theoretically possible levels of efficiency will not be known. For empirical use, the preceding definition is therefore replaced by the following:

> *Relative Efficiency*: A DMU is to be rated as fully (100%) efficient if and only if the performances of other DMUs do not show that some of its inputs or outputs can be improved without worsening some of its other inputs or outputs.

To implement this definition, it is necessary only to designate any DMU$_j$ as DMU$_0$ with inputs x_{i0} and outputs y_{r0} and then apply (1) to the input and output data recorded for the collection of DMU$_j$, $j = 1, \ldots, n$. Leaving this DMU$_j$ = DMU$_0$ in the constraints insures that solutions will always exist with an optimal $\theta_0 = \theta_0 * \leq 1$. The above definition applied to (1) then gives

> *DEA Efficiency*: The performance of DMU$_0$ is fully (100%) efficient if and only if, at an optimum, both (i) $\theta_0 * = 1$, and (ii) all slacks = 0 in (1a) or, equivalently, $\sum_{r=1}^{s} \mu_r^* \; y_{r0} = 1$ in (1b), where * represents an optimal value.

A value $\theta_0^* < 1$ shows (from the data) that a non-negative combination of other DMUs could have achieved DMU$_0$'s outputs at the same or higher levels while reducing *all* of its inputs. Non-zero slacks similarly show where input reductions or output augmentations can be made in DMU$_0$'s performance without altering other inputs or outputs. These non-zero slacks show where changes in *mixes* could have improved performance in each of DMU$_0$'s inputs or outputs, while a $\theta_0^* < 1$ shows "technical inefficiency" in which *all* inputs could have been reduced in the same proportion. (This is a so-called input-oriented model. An output-oriented model can be similarly formulated by associating a variable φ_0 with all outputs to be maximized DMU$_0$. The measures are reciprocal, i.e., $\varphi_0^* \; \theta_0^* = 1$, so this topic is not developed here.)

Many applications to many different kinds of entities engaged in complex activities with no clearly defined bottom line have been reported in many publications by many different authors in many different countries. Examples include applications to schools (including universities), police forces, military units, and country performances (including United Nations evaluations of country performances). See, for example, Emrouznejad et al. (2008) who list more than 1,600 published papers by more than 2,500 different authors in more than 40 different countries. Also see Berber et al. (2011) and Cooper et al. (2009).

Farrell Measure

The scalar θ_0^* is sometimes referred to as the Farrell measure after M.J. Farrell (1957). Notice, however, that a value of $\theta_0^* = 1$ does not completely satisfy the above definition of Relative Efficiency if any of the associated slacks, s_i^{+*} or s_r^{+*}, in (1) are positive — because any such non-zero slack provides an opportunity for improvement which may be used without affecting any other variable, as should be clear from the primal problem which is shown in (1a).

There is a need to insure that an optimum with $\theta_0^* = 1$ and all slacks zero is not interpreted to mean that full (100%) efficiency has been attained when an alternate solution with $\theta_0^* = 1$ and some slacks positive is also available. To see how this is dealt with, attention is called to the fact that the slack variables s_i^- and s_r^+ in the objective of the primal (minimization) problem, (1a), are each multiplied by

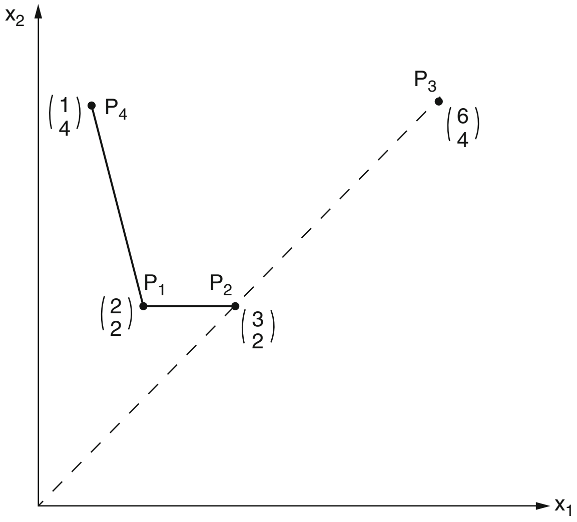
$\varepsilon > 0$ which is a non-Archimedean infinitesimal — the reciprocal of the "big M" associated with the artificial variables in ordinary linear programming — so that choices of slack values *cannot* compensate for any increase they might cause in θ_0. This accords pre-emptive status to the minimization of θ_0, and DEA computer codes generally handle optimizations in a two-stage manner which avoids the need for specifying ε explicitly. Formally, this amounts to minimizing the value of θ_0 in stage 1. Then one proceeds in a second stage to maximize the sum of the slacks with the condition $\theta_0 = \theta_0^*$ fixed for the primal in (1a). Since the sum of the slacks is maximized, one can be sure that a solution with all slacks at zero in the second stage means that DMU$_0$ is fully efficient if the first stage yielded $\theta_0^* = 1$.

N.B. Weak efficiency is another term used instead of Farrell efficiency when attention is restricted to (i) in DEA Efficiency above. It is also referred to as a measure of technical efficiency. However, when (1a) is used, this might be referred to as purely technical efficiency in order to distinguish these inefficiencies from the mix inefficiencies associated with changes in the proportions used that are then associated with non-zero slack. The term technical efficiency can then be used to comprehend both purely technical and mix inefficiencies as determined by reference to technical conditions without recourse to prices, costs, and/or subjective evaluations.

Example

Figure 1 is a geometric portrayal of four DMUs interpreted as points $P_1,\ldots, P_4$, with coordinate values corresponding to the amounts of two inputs which each DMU used to produce the same amount of a single output. P_3 is evidently inefficient compared to P_2 because it used more of both inputs to achieve the same output. In fact, its Farrell measure of inefficiency relative to P_2 can be determined via the formula

$$\theta_0 = \frac{d(0,\ P_2)}{d(0,\ P_3)} = \frac{\sqrt{3^2+2^2}}{\sqrt{6^2+4^2}} = \frac{1}{2},$$



Data Envelopment Analysis, Fig. 1 DEA efficiencies

where $d(.,.)$ refers to the Euclidean, or l_2, measure of distance.

Referred to as a radial measure of efficiency in the DEA literature, θ_0 is really a ratio of two distance measures, namely, the distance along the ray from the origin to the point being evaluated relative to the distance from the origin to the frontier measured along this same ray. This same value of θ_0 is obtained, and hence this same radial measure, by omitting the slacks and rewriting the primal problem in (1a) in the following inequality form,

$$\begin{aligned}
&\text{minimize } \theta_0\\
&\text{subject to}\\
&6\theta_0 \ \geq \ 2\lambda_1 + 3\lambda_2 + 6\lambda_3 + 1\lambda_4\\
&4\theta_0 \ \geq \ 2\lambda_1 + 2\lambda_2 + 4\lambda_3 + 4\lambda_4\\
&1 \ \leq \ 1\lambda_1 + 1\lambda_2 + 1\lambda_3 + 1\lambda_4\\
&0 \ \leq \ \lambda_1,\ \ldots,\ \lambda_4,
\end{aligned} \tag{2}$$

where the third constraint reflects the output $y = 1$ which was produced by each of these DMUs.

An optimum is achieved with $\theta_0^* = 1/2$, $\lambda_2^* = 1$ and this designates P_2 for the evaluation of P_3. However, it is also needed to take account of the slack possibilities. This is accomplished without specifying $\varepsilon > 0$ explicitly by proceeding to a second stage by using the thus obtained value of θ_0^* to form the following problem:

$$\begin{aligned}
&\text{maximize}\, s_1^- + s_2^- + s^+\\
&\text{subject to}\\
&0 = -6\theta_0 + 2\lambda_1 + 3\lambda_2 + 6\lambda_3 + 1\lambda_4 + s_1^-\\
&0 = -4\theta_0 + 2\lambda_1 + 2\lambda_2 + 4\lambda_3 + 4\lambda_4 + s_2^-\\
&-1 = -1\lambda_1 - 1\lambda_2 - 1\lambda_3 - 1\lambda_4 + s^+\\
&0.5 = \theta_0\\
&0 \leq \lambda_1, \ldots, \lambda_4, s_1^-, s_2^-, s^+
\end{aligned} \tag{3}$$

Following through in this second stage, with $\theta_0^* = 0.5$, it can be found that $\lambda_2^* = 1$ and $s_1^{-*} = 1$, with all other variables zero. This solution is interpreted to mean that the evidence from other DMUs (as exhibited by P_1's performance) shows that P_3 should have been able (a) to reduce both inputs to one-half their observed values, as given by the value of θ_0, and should also have been able (b) to reduce the first input by the additional amount given by $s_1^{-*} = 1$.

This slack, $s_1^{-*} = 1$, represents the excess amount of the first input used by P_2, and it, too, must be accounted for if the above definition of relative efficiency is to be satisfied. In fact, using the primal in (1a) to evaluate P_2, it will be found that it is also inefficient with $\theta_1{}^* = 1$ and $\lambda^* = s_1^{-*} = 1$. The use of (1a) to determine whether the conditions (i) and (ii) for relative efficiency are satisfied has a further consequence in that it insures that only efficient DMUs enter into the solutions with positive coefficients in the basis sets that are used to effect efficiency evaluations. Computer codes that have been developed for DEA generally use this property to reduce the number of computations by identifying all such members of an optimal basis as efficient and, hence, not in need of further evaluation.

As can be seen from Fig. 1, P_1 dominates P_2 and hence also dominates P_3. Only P_1 and P_4 are not dominated and hence can be regarded as efficient when DEA is restricted to dominance, as in Bardhan et al. (1996). However, if an assumption of continuity is added, then the entire line segment connecting P_1 and P_4 becomes available for use in effecting efficiency evaluations. This line segment is referred to as the efficiency frontier. The term efficient frontier is appropriate because it is not possible to move from one point to another on the line connecting P_1 and P_4 without worsening one input to improve the other input.

Given the assumption of continuity, points not on the efficiency frontier are referred to it for evaluation. Even when not dominated by actually observed performances, the nonnegative combinations of λ_j^* and slack values will locate points on the frontier which can be used for effecting efficiency evaluations of any DMU in the observation set.

The following formulas, called the CCR projection formulas, may be used to move points up to the efficiency frontier:

$$\begin{cases} \hat{x}_{i0} = \theta_0^* \hat{x}_{i0} - s_i^{-*} \leq \hat{x}_{i0}, & i = 1, \ldots, m \\ \hat{y}_{r0} = y_{r0} + s_r^{+*} \geq y_{r0}, & r = 1, \ldots, s \end{cases} \tag{4}$$

where each $(\hat{x}_{i0}, \hat{y}_{i0})$ represents a point on the efficiency frontier obtained from (x_{i0}, y_{r0}), DMU$_0$'s observed values. The point on the efficiency frontier thus obtained from these CCR projections is the point used to evaluate (x_{i0}, y_{r0}), $i = 1, \ldots, m$; $r = 1, \ldots, s$, for any DMU$_0$.

Ratio Form Models

The name Data Envelopment Analysis is derived from the primal (minimization) problem (1a) by virtue of the following considerations. The objective is to obtain as tight a fit as possible to the input–output vector for DMU$_0$ by enveloping its observed inputs from below and its observed outputs from above. As can be seen from (1a), an optimal envelopment will always involve a touching of the envelopment constraints to at least one of DMU$_0$'s inputs and one of its outputs.

The primal problem, (1a), is said to be in envelopment form. The dual problem, (1b), is said to be in multiplier form by reference to the values of μ and v as dual multipliers. The objective is to maximize y_0, which is called the virtual output. This maximization is subject to the condition that the corresponding virtual input is unity, that is, $\sum_{i=1}^m v_i x_{i0} = 1$, as given in the first constraint. The other constraints require that the virtual output cannot exceed virtual input for any of the DMU$_j$, $j = 1, \ldots, n$, that is,

$$\sum_{r=1}^{s} \mu_r y_{rj} \leq \sum_{i=1}^{m} v_i x_{ij} \quad j = 1, \ldots, n.$$

Finally, the conditions $\mu_r, v_i \geq \varepsilon > 0$ mean that every input and every output is to be assigned "some" positive value in this "multiplier" form, where as previously noted, the value of ε need not be specified explicitly.

To add interpretive power for the use in DEA, all of the variables in (1b) are multiplied, the (dual) problem of (1a), by $t > 0$ and then introduce new variables defined in the following manner:

$$\begin{aligned} \mu_r = t\mu_r \ &\geq t\varepsilon, \ v_i = tv_i \ \geq t\varepsilon, \\ t = \sum_{i=1}^{m} &\ tv_i \ x_{i0}. \end{aligned} \tag{5}$$

Multiplying and dividing the objective of the dual problem in (1b) by $t > 0$ and then multiplying all constraints by t gives the following model, which accords a ratio form to the DEA evaluations:

$$\begin{aligned} &\max \ \frac{\sum_{r=1}^{s} u_r \, y_{r\,0}}{\sum_{i=1}^{m} v_i \, x_{i0}} \\ &\text{subject to} \ \frac{\sum_{r=1}^{s} \mathrm{u_r y_{rj}}}{\sum_{i=1}^{m} \mathrm{v_i \, x_{ij}}} \leq 1, \quad j = 1, \ldots, n \\ &\frac{u_r}{\sum_{i=1}^{m} v_i x_{i0}} \geq \varepsilon, \quad r = 1, \ \ldots, \ s \\ &\frac{v_i}{\sum_{i=1}^{m} v_i \, x_{i0}} \geq \varepsilon, \quad i = 1, \ \ldots, \ m. \end{aligned} \tag{6}$$

An immediate corollary from this development is

$$\begin{aligned} 0 \leq \frac{\sum_{r=1}^{s} u_r^* y_{r0}}{\sum_{i=1}^{m} v_i^* x_{i\,0}} = \sum_{r=1}^{s} u_r^* y_{r0} = \theta_0^* \\ - \sum_{i=1}^{m} s_i^{-*} + \sum_{r=1}^{s} s_r^{+*} \leq 1, \end{aligned} \tag{7}$$

where "*" designates an optimal value. Thus, in accordance with the theory of fractional programming, as given in Charnes and Cooper (1962), the optimal values in (6) and (1b) are equal.

The formulation (6) has certain advantages. For instance, Charnes and Cooper (1985) used it to show that the optimal ratio value in (6) is invariant to the units of measure used in any input and any output and, hence, this property carries over to (1b). Equation 6 also add interpretive power and provide a basis for unifying definitions of efficiency that stretch across various disciplines. For instance, as shown in Charnes et al. (1978), the usual single-output to single-input efficiency definitions used in science and engineering are derivable from (6). It follows that these definitions contain an implicit optimality criterion. The relation of (6) to (4), established via fractional programming, also relates these optimality conditions to the definitions of efficiency used in economics. (See the above discussion of Pareto-Koopmans efficiency.) This accords a ratio form (as well as a linear programming form) to the DEA evaluations.

As (6) makes clear, DEA also introduces a new principle for determining weights. In particular the weights are not assigned a priori, but are determined directly from the data. A best set of weights is determined for each of the $j, \ldots, n$ DMUs to be evaluated. Given this set of best weights the test of inefficiency for any DMU$_0$ is whether any other DMU$_j$ achieved a higher ratio value than DMU$_0$ using the latter's best weights [Care needs to be exercised in interpreting these weights, since (a) their values will in general be determined by reference to different collections of DMUs and (b) when determined via (1), allowance needs to be made for non-zero slacks. See the discussion in Charnes et al. (1989), where dollar equivalents are used to obtain a complete ordering to guide the use of efficiency audits by the Texas Public Utility Commissions].

DEA also introduces new principles for making inferences from empirical data. This flows from its use of n optimizations — to come as close as possible to *each* of n observations — in place of other approaches, as in statistics, for instance, which uses a single optimization to come as close as possible to all of these points. In DEA, it is also not necessary to specify the functional forms explicitly. These forms may be nonlinear and they may be multiple (differing, perhaps, for each DMU) provided they satisfy the mathematical property of isotonicity (Charnes et al. 1985).

Other Models

The models in (1) and (6) are a subset of several DEA models that are now available. Thus, DEA may be regarded as a body of concepts, and methods which unite these models and their uses to each other. These concepts, models and methods comprehend extensions to identify scale, and allocative and other inefficiencies.

By virtue of the already described relations between (6) and (1) the models are referred to as the CCR ratio model. Other models include the additive model, namely,

$$
\begin{aligned}
&\max \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+ \\
&\text{subject to} \\
&0 = \hat{x}_{i0} - \sum_{j=1}^{n} \hat{x}_{ij}\lambda_j - s_i^- \\
&\hat{y}_{r0} = \sum_{j=1}^{n} \hat{y}_{rj}\lambda_j - s_r^+ \\
&0 \leq \lambda_j,\ s_r^+,\ s_i^-;\ \ \forall i,\ j,\ r
\end{aligned}
\tag{8}
$$

for which the conditions for efficiency are given by Additive Model Efficiency: DMU$_0$ is fully (100%) efficient if and only if all slacks are zero — namely, $s_i^{-*}, s_r^{+*} = 0, \forall\, i, r$ in (8).

With the constraint $\sum_{j=1^n} \lambda_j = 1$ adjoined, the model (8) becomes “translation invariant.” That is, as shown by Ali and Seiford (1990), the solution to (8) is not altered if the original data $(\hat{x}_{ij},\hat{y}_{rj})$ are replaced by new data

$$
\begin{aligned}
\hat{x}'_{ij} &= \hat{x}'_{ij} + d_i, \quad i = 1, \ldots, m \\
\hat{y}'_{rj} &= \hat{y}'_{rj} + c_r, \quad r = 1, \ldots, s
\end{aligned}
\tag{9}
$$

where the d_i and c_r are arbitrarily constants. This property can be of value in treating negative data since most theorems in DEA assume that the data are positive or at least semi-positive. See Pastor (1996) for examples and extensions of the Ali-Seiford theorems. Theorems like the following from Ahn et al. (1988) relate the additive models to their CCR counterparts.

Theorem: A DMU$_0$ will be evaluated as fully (100%) efficient by the CCR model if and only if it is rated as fully (100%) efficient by the corresponding additive model.

Note, however, that the CCR and additive models use different metrics, so they need not identify the same sources and amounts of inefficiency in an inefficient DMU.

The additive model (8) can also be related to another class, called multiplicative models (Charnes et al. 1982). An easy way is to assume that the $(\hat{x}_{ij},\hat{y}_{rj})$ are stated in logarithmic units. Taking antilogs then gives

$$
\begin{aligned}
x_{i0} &= a_i^* \prod_{j=1}^{n} x_{ij}^{\lambda_j^*}, \quad i = 1, \ldots, m, \\
y_{r0} &= b_r^* \prod_{j=1}^{n} y_{rj}^{\lambda_j^*}, \quad r = 1, \ldots, s,
\end{aligned}
\tag{10}
$$

where $a_i^* = e^{s^{-*i}}$, $b_r^* = e^{s^{+*r}}$, and the (x_{ij}, y_{rj}) are stated in natural units. Each x_{i0}, y_{r0} is thus generated by a Cobb-Douglas process with estimated parameters given by the starred values of the variables.

To relate these results to a ratio form for efficiency evaluation, the dual to (8) is written as

$$
\begin{aligned}
&\min \sum_{i=1}^{m} v_i\hat{x}_{i0} - \sum_{r=1}^{s} \mu_r\hat{y}_{r0} \\
&\text{subject to} \\
&\sum_{i=1}^{m} v_i\hat{x}_{ij} - \sum_{r=1}^{s} \mu_r\hat{y}_{rj} \geq 0, \quad j = 1, \ldots, n \\
&v_i,\ \mu_r \geq 1, \quad i = 1, \ldots, m; \quad r = 1, \ldots, s,
\end{aligned}
\tag{11}
$$

where the $(\hat{x}_{ij}, \hat{y}_{rj})$ are stated in logarithmic units. Recourse to antilogarithms then produces

$$
\begin{aligned}
&\max \prod_{r=1}^{s} \hat{y}_{r0}^{\mu r} \bigg/ \prod_{i=1}^{m} \hat{x}_{i0}^{v_i} \\
&\text{subject to} \\
&\prod_{r=1}^{s} \hat{y}_{rj}^{\mu r} \bigg/ \prod_{i=1}^{m} \hat{x}_{i0}^{v_i} \leq 1, \quad j = 1, \ldots, n \\
&v_i, \mu_r \geq 1, \quad i = 1, \ldots, m; \quad r = 1, \ldots, s,
\end{aligned}
\tag{12}
$$

and we once again make contact with a ratio form for effecting efficiency evaluations.

To obtain conditions for efficiency, antilogs to (8) are applied and (10) is used to obtain

$$\max \frac{\prod_{r=1}^{s} e^{s_r^{+*}}}{\prod_{i=1}^{m} e^{-s_i^{-*}}} = \frac{\prod_{r=1}^{s}\prod_{j=1}^{n} y_{rj}^{\lambda_j^*}/y_{r0}}{\prod_{i=1}^{m}\prod_{j=1}^{n} x_{ij}^{\lambda_j^*}/x_{i0}} \geq 1. \qquad (13)$$

The lower bound on the right is obtainable if and only if all slacks are zero. Thus the efficiency conditions for the multiplicative model are the same as for the additive model.

An interpretation of (13) can be secured by noting that

$$\left(\prod_{j=1}^{n} y_{rj}^{\lambda_j^*}\right)^{1\Big/\sum_{j=1}^{n}\lambda_j^*}, \qquad \left(\prod_{j=1}^{n} x_{ij}^{\lambda_j^*}\right)^{1\Big/\sum_{j=1}^{n}\lambda_j^*}$$

represent weighted geometric means of outputs and inputs, respectively. Thus (13) is a ratio of the product of weighted geometric totals relative to the outputs and inputs which each of these expressions is evaluating.

It is necessary to note that the results in (13) are not units invariant (i.e., they are not dimension free in the sense of dimensional analysis) except in the case of constant returns to scale (see Thrall, 1996). This property, when wanted, can be secured by adjoining $\sum_{j=1^n}\lambda_j = 1$ to (8). See also Charnes et al. (1983). To conclude this discussion it is noted that the expression on the left of (13) is simpler and easier to interpret and the computations from (8) are straightforward.

The class of multiplicative models has not been much used, possibly because other models are easier to comprehend. Even allowing for this, however, they have potentials for use either on their own or in combination with other DEA models as when, for instance, returns to scale characterization are needed that differ from those which are available from other types of DEA models. See Banker and Maindiratta (1986) for further discussion of such uses.

Extensions and Uses of Dea Models

1. *Returns to Scale* — There is an extensive literature on returns to scale and their uses in DEA which reflects two different approaches. One approach, due to Färe et al. (1985, 1994) proceeds in an axiomatic manner and employs only radial measures. The other approach is based on mathematical programming. Conceptualized by Banker et al. (1984), it was subsequently ex-tended (and made wholly rigorous) by Banker and Thrall (1992). As might be expected, equivalences between the two approaches have been established in (among other places) Banker et al. (1996). See also Banker et al. (1998).
2. *Returns to Scope* — Partly because of difficulty in assembling data in pertinent forms, the literature on returns to scope is relatively sparse in DEA. Indeed, a bare beginning has been made in Chapter 10 of Färe et al. (1994).
3. *Assurance Regions and Allocative Inefficiency* — Many other developments have occurred and continue to occur. Thompson, Dharmapala and Thrall and their associate introduced the now widely used concept of assurance regions (Thompson et al. 1986; Dyson and Thanassoulis, 1988). This approach uses a priori knowledge to set upper and lower bounds on the values of the multiplier variables in DEA models like (1b). This can alleviate problems encountered in treating allocative or price efficiency either because (i) exact data on prices, costs, etc., are not available, or (ii) because the presence of wide variations in these data make the use of exact value a questionable undertaking. See Schaffnit et al. (1997), where limiting arguments are used to establish an exact relation between allocative efficiency and the bounds used in assurance region approaches.
4. *Cone Ratio Envelopments* — In a similar spirit, but in a different manner, Charnes et al. (1990) and their associates developed what they refer to as a cone-ratio envelopment approach. In contrast to the assurance region treatments of bounds on the variables, these cone-ratio approaches utilize a priori information to adjust the data. This makes it possible to take account of complex (multiple) considerations that might otherwise be difficult to articulate. See Brockett et al. (1997), who show how to implement the Basle Agreement, which was recently adopted by U.S. bank regulators to treat multiple risk factors in banking by adjusting the data reported in the FDIC call reports. These regulations are rigid and ill-fitting, so Brockett

et al. (1997) provide an alternative Cone-ratio envelopment approach which uses results from excellent banks (that are also found to be efficient) to adjust the call-report data for other banks in a use of DEA to effect such risk-adjusted evaluations.

5. *Exogenous and Categorical Variables* — Other important developments include methods for treating input or output values which are exogenously fixed for some, or all, DMUs. Developed by Banker and Morey (1986a) for treating demographic variables as important inputs in different locations for a chain of fast food outlets, these methods have found widespread use in many other applications. Similar remarks apply to the Banker and Morey (1986b) introduction of methods for treating categorical (classificatory) variables in work which has since been modified and extended by other authors; see Neralić and Wendell (2000).
6. *Statistical Treatments* — Various attempts have recently been made to join statistical and probabilistic characterizations to the deterministic models and methods of inference in DEA. For instance, using relatively mild postulates, Banker (1993) has shown that (i) DEA estimators of θ_0^* are statistically consistent; (ii) DEA estimates maximize the likelihood of obtaining the corresponding true values; and (iii) these properties hold under fairly general structures that do not require assumptions about the parametric forms of the probability density functions. See pages 272–275 in Banker and Cooper (1994) for a succinct discussion. See also Korostelev et al. (1995), who show that the rates of convergence are slow.

   Simar and Wilson (1998) utilize bootstrap procedures to study sampling properties of the efficiency measures in DEA. Unlike Banker, who restricts his analysis to the single output case, this bootstrap approach accommodates multiple outputs as well as multiple inputs. Omitted, however, is any treatment of nonzero slacks. Brockett and Golany (1996) also approach the topic of statistical characterizations by means of Mann–Whitney rank order statistics, but do not note that need for explicitly stating a ranking principle. This is needed because (as noted above) the DEA efficiency scores are generally determined relative to different reference sets (or peer groups) of efficient DMUs. (For a discussion of how this problem is treated for the efficiency audits conducted by Texas Public Utility Commission, see Charnes et al. 1989).
7. Probabilistic Models — Alternate approaches via chance constrained programming were initiated by Land et al. (1994) and have been ex-tended by others to include the use of joint chance constraints in addition to the conditional chance constraints used by Land, Lovell and Thore (Olesen and Petersen 1995; Cooper et al. 1998). Of special interest is the use of chance constraints to obtain a satisficing approach for efficiency evaluation, as in Cooper et al. (1996), where the term satisficing is used in the sense of H.A. Simon's (1957) behavioral characterizations in terms of (i) achievement of a satisfactory level of efficiency, and (ii) a satisfactory probability (=chance) of achieving this level. Finally, allowance is also made for situations in which these levels or probabilities may need to be revised because the data show that they are not possible of attainment.

   Unlike the statistical characterizations described in item 6, these chance constrained programs generally require knowledge of the parameters as well as the forms of the probability functions so that here, too, there is more work to be done. See Jagannathan (1985) for a start.
8. *Cross-Checking* — As noted in the earlier discussions, the inference principles in DEA differ from those in statistics. This suggests additional possibilities for their joint use. One such possibility is to use the two approaches as cross checks on each other to help avoid what is referred to as methodological bias in Charnes et al. (1988). See also Ferrier and Lovell (1990).
9. *Complementary Uses* — Another possibility is to use statistics and DEA in a complementary manner. An example is provided by Arnold et al. (1996), who applied this strategy in a two-stage manner to a study of Texas public schools as follows. At stage 1, DEA is used to identify efficient schools; then, at stage 2, these results are incorporated as dummy variables in an OLS (Ordinary Least Squares) regression. This yielded very satisfactory results on data which had previously yielded unsatisfactory results with an OLS regression. A subsequent simulation study by Bardhan et al. (1998)

compares this approach not only to OLS but also to stochastic frontier regressions (i.e., regressions which apply statistical principles to obtain frontier estimates for efficiency evaluations). Using observations that reflected mixtures of efficient and inefficient performances the OLS and SF approaches always failed to provide correct estimates whereas, with only one minor exception, the complementary two-stage use of DEA and statistics always yielded estimates that did not differ significantly from the true parameter values.

Sources and References

As the above discussions suggest, many important developments have been effected in DEA since its initiation by Charnes et al. (1978). These developments have occurred pari passu with numerous and widely varied applications of DEA which are being reported from many different parts of the world. See the bibliography by Seiford (1994). For a comprehensive text, see Cooper et al. (1999).

See

- ▶ Dual Linear-Programming Problem
- ▶ Fractional Programming
- ▶ Linear Programming

References

Ahn, T., Charnes, A., & Cooper, W. W. (1988). A note of the efficiency characterizations obtained in different DEA models. *Socio-Economic Planning Sciences, 22*, 253–257.

Ali, A. I., & Seiford, L. M. (1990). Translation invariance in data envelopment analysis. *Operations Re-search Letters, 9*, 403–405.

Arnold, V., Bardhan, I., & Cooper, W. W. (1993). *DEA models for evaluating efficiency and excellence in Texas Secondary Schools* (Working Paper, IC2) Austin: Institute of the University of Texas.

Arnold, V., Bardhan, I., Cooper, W. W, & Gallegos, A. (1984). Primal and dual optimality in ideas (integrated data envelopment analysis systems) and related computer codes. *Proceeding of a Conference in Honor of G.L. Thompson*, Quorum Books

Arnold, V., Bardhan, I., Cooper, W. W., & Gallegos, A. (1998). Primal and dual optimality in IDEAS (Integrated Data Envelopment Analysis Systems) and related computer codes, operations research: Methods, models and applications. *Proceedings of a Conference in Honor of G.L. Thompson*, Westport, CT: Quorum Books.

Arnold, V., Bardhan, I., Cooper, W. W., & Kumbhakar, S. (1996). New uses of DEA and statistical regressions for efficiency evaluation and estimation–with an illustrative application to Public Secondary Schools in Texas. *Annals Operations Research, 66*, 255–278.

Banker, R. D. (1993). Maximum likelihood, consistency and data envelopment analysis: A statistical foundation. *Management Science, 39*, 1265–1273.

Banker, R. D., Chang, H. S., & Cooper, W. W. (1996). Equivalence and implementation of alternative methods for determining returns to scale in data envelopment analysis. *European Journal Operational Research, 89*, 473–481.

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science, 30*, 1078–1092.

Banker, R. D., & Cooper, W. W. (1994). Validation and generalization of DEA and its uses. *TOP (Sociedad de Estadistica e Investigación Operativa), 2*, 249–297. (with discussions by E. Grifell-Tatje, J. T. Pastor, P. W. Wilson, E. Ley and C. A. K. Lovell).

Banker, R.D., Cooper, W.W., & Thrall, R. M. (1998). *Finished and unfinished business for returns to scale in DEA*. Research Report, Graduate School of Business, University of Texas at Austin.

Banker, R. D., & Maindiratta, A. (1986). Piecewise loglinear estimation of efficient production surfaces. *Management Science, 32*, 385–390.

Banker, R. D., & Morey, R. C. (1986a). Efficiency analysis for exogenously fixed inputs and outputs. *Operations Research, 34*, 513–521.

Banker, R. D., & Morey, R. C. (1986b). Data envelopment analysis with categorical inputs and outputs. *Management Science, 32*, 1613–1627.

Banker, R. D., & Thrall, R. M. (1992). Estimation of returns to scale using data envelopment analysis. *European Journal Operational Research, 62*, 74–84.

Bardhan, I., Bowlin, W. F., Cooper, W. W., & Sueyoshi, T. (1996). Models and measures for efficiency dominance in DEA. Part I: Additive models and med measures. Part II: Free disposal hulls and Russell measures. *Journal of the Operations Research Society Japan, 39*, 322–344.

Bardhan, I. R., Cooper, W. W., & Kumbhakar, S. C. (1998). A simulation study of joint uses of DEA and statistical regression for production function estimation and efficiency evaluation. *Journal Productivity Analysis, 9*, 249–278.

Berber, P., et al. (2011). Efficiency in fundraising and distributions to cause-related social profit enterprises. *Socio-Economic Planning Sciences, 45*, 1–9.

Bowlin, W. F., Brennan, J., Cooper W. W., & Sueyoshi, T. (1984). *A DEA model for evaluating efficiency dominance*, Research Report. Texas: Center for Cybernetic Studies, Austin, (submitted for publication).

Brockett, P. L., Charnes, A., Cooper, W. W., Huang, Z. M., & Sun, D. B. (1997). Data transformations in DEA cone ratio envelopment approaches for monitoring bank performances. *European Journal of Operational Research, 95*, 250–268.

Charnes, A., & Cooper, W. W. (1962). Programming with linear fractional functionals. *Naval Research Logistics Quarterly, 9*, 181–186.

Charnes, A., & Cooper, W. W. (1985). Preface to topics in data envelopment analysis. In R. Thompson & R. M. Thrall (Eds.), *Annals operations research* (Vol. 2, pp. 59–94).

Charnes, A., Cooper, W. W., Divine, D., Ruefli, T. W., & Thomas, D. (1989). Comparisons of DEA and existing ratio and regression systems for efficiency evaluations of regulated electric cooperatives in Texas. *Research in Governmental and Nonprofit Accounting, 5*, 187–210.

Charnes, A., Cooper, W. W., Golany, B., Seiford, L., & Stutz, J. (1985). Foundations of data envelopment analysis and Pareto-Koopmans efficient empirical production functions. *Journal of Econometrics, 30*, 91–107.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring efficiency of decision making units. *European Journal of Operational Research, 1*, 429–444.

Charnes, A., Cooper, W. W., Seiford, L., & Stutz, J. (1982). A multiplicative model for efficiency analysis. *Socio-Economic Planning Sciences, 16*, 223–224.

Charnes, A., Cooper, W. W., Seiford, L., & Stutz, J. (1983). Invariant multiplicative efficiency and piecewise Cobb-Douglas envelopments. *Operations Research Letters, 2*, 101–103.

Charnes, A., Cooper, W. W., & Sueyoshi, T. (1988). Goal programming-constrained regression review of the bell system breakup. *Management Science, 34*, 1–26.

Charnes, A., Cooper, W. W., Sun, D. B., & Huang, Z. M. (1990). Polyhedral cone-ratio DEA models with an illustrative application to large commercial banks. *Econometrics Journal, 46*, 73–91.

Charnes, A., Cooper, W. W., & Thrall, R. M. (1991). A structure for classifying and characterizing efficiency and inefficiency in data envelopment analysis. *Journal of Productivity Analysis, 2*, 197–237.

Cooper, W. W., Huang, Z. M., & Li, S. (1996). Satisficing DEA models under chance constraints. *Annals Operations Research, 6*, 279–295.

Cooper, W. W., Huang, Z., Lelas, V., Li, X. S., & Olesen, O. B. (1998). Chance constrained programming formulations for stochastic characterizations of efficiency and dominance in DEA. *Journal of Productivity Analysis, 9*, 53–79.

Cooper, W. W., Seiford, L. M., & Tone, K. (1999). *Data envelopment analysis*. Boston, MA: Kluwer Academic Publishers.

Cooper, W. W., Seiford, L. M., & Zhu, J. (2011). *Handbook on data envelopment analysis*. New York: Springer.

Cooper, W. W., Thore, S., & Traverdyan, R. (2009). A utility function approach for evaluating country performances — The twin goals of decent work and affair globalization. In R. R. Hockley (Eds.), *Global operations management*. NOVA Science Publishers.

Dyson, R. G., & Thanassoulis, E. (1988). Reducing weight flexibility in data envelopment analysis. *Journal of Operational Research Society, 39*, 563–576.

Emrouznejad, A., Parker, B. R., & Tavares, G. (2008). Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in DEA. *Socio-Economic Planning Sciences, 42*, 151–157.

Färe, R., Grosskopf, S., & Lovell, C. A. K. (1994). *Production frontiers*. Cambridge, UK: Cambridge University Press.

Färe, R., Grosskopf, S., & Lovell, C. A. K. (1995). *The measurement of efficiency of production*. Norwell, MA: Kluwer.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of Royal Statistical Society*, Series A, 253–290.

Ferrier, G. D., & Lovell, C. A. K. (1990). Measuring cost efficiency in banking: Econometric and linear programming evidence. *Journal of Econometrics, 46*, 229–245.

Jagannathan, R. (1985). Use of sample information in stochastic recourse and chance constrained programming models. *Management Science, 31*, 96–108.

Kamakura, W. A. (1988). A note on the use of categorical variables in data envelopment analysis. *Management Science, 34*, 1273–1276.

Korostelev, A., Simar, L., & Tsybakov, A. (1995). Efficient estimation of monotone boundaries. *Annals Statistics, 23*, 476–489.

Land, K., Lovell, C. A. K., & Thore, S. (1994). Chance constrained data envelopment analysis. *Managerial and Decision Economics, 14*, 541–554.

Neralić, L., & Wendell, R. (2000). A generalized additive, categorical model in data envelopment analysis. *TOP: Journal of the Spanish Society of Statistics and Operations Research, 8*, 235–263.

Olesen, O. B., & Petersen, N. C. (1995). Chance constrained efficiency evaluation. *Management Science, 41*, 442–457.

Pastor, J. T. (1996). Translation invariance in data envelopment analysis. *Annals Operations Research, 66*, 93–102.

Rousseau, J. J., & Semple, J. H. (1993). Categorical outputs in data envelopment analysis. *Management Science, 39*, 384–386.

Schaffnit, C., Rosen, D., & Paradi, J. C. (1997). Best practice analysis of bank branches: An application of DEA in a large canadian bank. *European Journal of Operational Research, 98*, 269–289.

Seiford, L. M. (1994). A bibliography of data envelopment analysis. In A. Charnes, W. W. Cooper, A. Y. Lewin, & L. M. Seiford (Eds.), *Data envelopment analysis: Theory, methodology and applications*. Norwell, MA: Kluwer.

Seiford, L. M., & Thrall, R. M. (1990). Recent development in DEA. In A. Y. Lewin, & C. A. Knox Lovell (Eds.), *Frontier analysis, parametric and nonparametric approaches. Journal of Econometrics*, *46*, 7–38.

Simar, L., & Wilson, P. W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science, 44*, 49–61.

Simon, H. A. (1957). *Models of man*. New York: John Wiley.

Thompson, R., Singleton, F., Thrall, R. M., & Smith, B. (1986). Comparative site evaluations for locating a high energy physics laboratory in Texas. *Inter-faces, 16*, 35–49.

Thrall, R. M. (1996). Duality, classification and slacks in DEA. *Annals Operations Research, 66*, 109–138.

Data Mining

Syam Menon[1] and Ramesh Sharda[2]
[1]The University of Texas at Dallas, Richardson, TX, USA
[2]Oklahoma State University, Stillwater, OK, USA

Introduction

When Wal-Mart installed their 24 terabyte data warehouse, it was among the largest in the world. Just a few years later, they were adding over a billion rows of data a day (Babcock 2006), and operating a 5 petabyte database (Lai 2008). An even more striking example is eBay, which started with a 14 terabyte database in 2002. It has since been adding over 40 terabytes of auction and purchase data every day into a data warehouse that is expected to exceed 20 petabytes by 2011. Clearly, as the cost of capturing data has decreased and easier-to-use data capture tools have become available, the volumes of data being accumulated have grown at a very rapid pace. Technological developments, with the evolution of the Internet playing a fundamental role, have enabled an increase in the volume of traditional data being recorded. Further, such developments have made possible the capture of information in far greater detail than ever before (based on barcodes or RFID, for example) and often of information that was not easily recordable before, such as eye or mouse movements.

What is Data Mining?

The availability of large data repositories has resulted in significant developments in the methodologies to analyze them, both in terms of the technology available for analysis, and in terms of its mainstream acceptance. From what was a relatively esoteric technology at the close of the 20th century, data mining – defined succinctly as “the science of extracting useful information from large data sets” (Hand et al. 2001) – has developed into a powerful set of tools indispensable to most organizations. In fact, it is gradually morphing into a key component of the merger of quantitative techniques into a new label called business analytics.

Many of the techniques used in data mining have their roots in traditional statistics, artificial intelligence, and machine learning. Developments in data mining techniques went hand-in-hand with developments in data warehousing and online analytical processing (OLAP). From the early 1990s when data mining started being viewed as a viable business solution, the cost of computing has dropped steadily, while processing power has increased. This made the benefits of data mining apparent, and triggered many companies to start using it regularly.

Commercial applications of data mining abound. A 2010 poll of data miners (conducted by KDNuggets) listed customer relations management, banking, healthcare, and fraud detection as the top four fields where data mining is applied. It is also commonly used in finance, direct marketing, insurance, and manufacturing. In fact, it has become common practice in almost every industry to discern new knowledge from data; only the extent of penetration varies across industries.

This is, of course, in addition to the vast quantities of data collected in the non-business world. It has found application in disciplines as varied as astronomy, genetics, healthcare, and education, just to name a few. The U.S. Department of Homeland Security applies data mining for a variety of purposes, including the comparison of “traveler, cargo, and conveyance information against intelligence and other enforcement data by incorporating risk-based targeting scenarios and assessments,” and “to improve the collection, use, analysis, and dissemination of information that is gathered for the primary purpose of targeting, identifying, and preventing potential terrorists and terrorist weapons from entering the United States” (DHS 2009).

The availability of new types of data has opened up additional opportunities for selective extraction of useful information. Data originating from the Web can be mined based on content, network structure, or usage (e.g., when was a page used and by whom). There has been considerable interest in the mining of text from a variety of perspectives – to filter e-mail, to gain intelligence about competitors, to analyze the opinions of movie viewers to better understand movie reviews, as well as the mining of social network data both in terms of user behaviors and networks, including text mining of comments. The analysis of audio and video files is another difficult but promising avenue for data mining. Speech recognition technologies have improved significantly. But, audio mining goes much further by providing users the ability to search and index the digitized audio content in a variety of contexts like news and webcasts, recorded telephone conversations, office meetings, and archives in libraries and museums.

How Does Data Mining Work?

Most of the general ideas applicable to modeling of any kind hold true for data mining as well. To work effectively, data mining requires clearly stated objectives and evaluation criteria. The process (often referred to as the Knowledge Discovery in Databases – or KDD – process) entails various critical steps. All data need to be cleaned to eliminate noise and correct errors. As data usually come from multiple, heterogeneous sources, there has to be a logical process of data integration. Once an objective has been identified for analysis, all appropriate data needs to be retrieved from the storage warehouse(s). If necessary, extracted data may need to be transformed into a form amenable for mining. Once all these preprocessing steps are completed, relevant data mining techniques can be applied. As with any analysis technique, the output from the mining process usually needs to be interpreted by the analyst after imposing as much domain knowledge as possible to intelligently glean useful information. Any model that is built should be tested and validated before putting to full use. Additionally, the KDD process has to be iterative for it to be beneficial. The knowledge discovered through mining can be used to obtain feedback from the user which in turn can be used to improve the mining process.

Data mining tasks fall into two main groups – descriptive tasks that characterize properties of the data being analyzed, and predictive tasks which make predictions about new data points based on inferences made from existing data. Data mining algorithms traditionally fall into one of three categories — classification and prediction, clustering, and association discovery. Other functionalities like data characterization and outlier analysis are also common, as are applications that form key components of recommender systems. Data visualization plays an important role in many of these techniques by guiding the users in the right direction. Some of these techniques are described briefly below.

Classification. Classification, or supervised induction, is perhaps the most common of all data mining activities. The objective of classification is to analyze the historical data stored in a database and to automatically generate a model that can predict future behavior. This induced model consists of generalizations over the records of a training data set, which help distinguish predefined classes. The hope is that this model can then be used to predict the classes of other unclassified records. When the output variable of interest is categorical, the models are referred to as classifiers, while models where the output variable is numerical are called prediction models.

Tools commonly used for classification include neural networks, decision trees, and if-then-else rules that need not have a tree structure. Statistical tools like logistic regression are also commonly used. Neural networks involve the development of mathematical structures with the ability to learn. They tend to be most effective where the number of variables involved is large and the relationships between them too complex and imprecise. It can easily be implemented in a parallel environment, with each node of the network doing its calculations on a different processor. There are disadvantages as well. It is usually very difficult to provide a good rationale for the predictions made by a neural network. Also, training time on neural networks tends to be considerable. Further, the time needed for training tends to increase as the volume of data increases, and in general, such training cannot be done on very large databases. These and other factors have limited the acceptability of neural networks for data mining.

Decision trees (DTs) classify data into a finite number of classes, based on the values of the variables. DTs are comprised of essentially a hierarchy of if-then statements and are thus significantly faster than neural nets. Logistic regression models are used for binary classification, with multinomial logistic models being used if there are more than two output categories.

Clustering. Most clustering algorithms partition the records of a database into segments where members of a segment share similar qualities. In fact, clustering is sometimes referred to as unsupervised classification. Unlike in classification, however, the clusters are unknown when the algorithm starts. Consequently, before the results of clustering techniques are put to actual use, it might be necessary for an expert to interpret and potentially modify the suggested clusters. Once reasonable clusters have been identified, they could be used to classify new data. Not surprisingly, clustering techniques include optimization; we want to create groups, which have maximum similarity among members within each group and minimum similarity among members across the groups. Another common application is market basket analysis.

Association Discovery. A special case of association rule mining looks at sequences in the data. Sequence discovery has many applications, and is a significant sub-field in itself. It can be to conduct temporal analysis to identify customer behavior over time, to identify interesting genetic sequences, for website re-design, and even for intrusion detection.

Visualization. The insights to be gained from visualizing the data cannot be over-emphasized. This holds true for most data analysis techniques, but is of special relevance to data mining. Given the sheer volume of data in the databases being considered, visualization in general is a difficult endeavor. It can be used, however, in conjunction with data mining to gain a clearer understanding of many underlying relationships.

Recommender Systems. Many companies claim that a substantial portion of their revenues are a result of effective recommendations. Among the better known examples are Amazon.com, which was one of the earlier proponents of recommender systems, and Netflix, which claims that “roughly two-thirds of the films rented were recommended to subscribers by the site” (Flynn 2006). The impact and importance of a well implemented recommendation system is exemplified by the fact that Netflix offered a million-dollar prize for anyone who could improve their recommendation accuracy by at least 10%. A variety of techniques exist for making recommendations, with user and item based collaborative filtering being the most common.

Other Relevant Aspects

Software. There are many large vendors of data mining software. Some of the key commercial packages include SAS Enterprise Miner, IBM SPSS Modeler (Formerly SPSS Clementine), Oracle, DigiMine, Microsoft SQL Server, SAP Business Objects. Weka is a well reputed freeware out of The University of Waikato in New Zealand. Another open source data mining software is Rapid Miner.

Privacy. Data mining has been restricted in its impact due to privacy concerns. In particular, in privacy concerns when applying data mining to healthcare data. A contested court case concerns the mining of physicians’ prescription history to increase drug sales; some states are trying to limit access to this information (Field 2010). The fundamental issue underlying these concerns relate to the intent behind data collection. For example, while consumers explicitly agree to the use of data collected for bill payment for that specific purpose, they may not know or want to agree to the use of their data for mining – that would go beyond the original intent for which the data were acquired.

Another area of data mining privacy concerns counterterrorist information Claburn (2008). A report dealing with the balance between privacy and security by the National Research Council recommends that the U.S. government rethink its approach to counterterrorism in light of the privacy risks posed by data mining.

Although some work has been done to incorporate privacy concerns explicitly into the mining process, this is still a developing field. In all likelihood, the matter of privacy in the context of data mining will be an issue for some time. A simple solution is unlikely. These issues will probably be resolved only through a blend of legislation and additional research into privacy preserving data mining.

The Role of Operations Research

Data mining algorithms are a heterogeneous group, loosely tied together by the common goal of generating better information. Operations research is concerned with making the best use of available information. By selecting the appropriate definition of information, operations research has been playing a significant role on both sides of the data mining engine. Formulations for clustering and classification were introduced in the 1960s and 70s (Ólafsson 2006). Nonlinear programming solution techniques have been adapted for faster training in neural network applications. Scalability, the ability to deal with large amounts of data, is a difficult and important issue in data mining, one in which OR could play a significant role.

The lack of reliable data (or of the data itself) is a common problem faced by operations researchers trying to get a good model to work in the real world. This problem becomes more acute when data needs to be deciphered from terabytes of stored information. Data mining tools make accessing and processing the data easier and may provide more reliable data to the OR modeler. There are opportunities for operations research to be applied at a more fundamental level as well. Ultimately, as with any analysis tool, the outputs of data mining models are only as good as the inferences the analyst can make from them. OR techniques can be of assistance in making the best use of the outputs obtained. For example, research has been conducted to improve recommendations by combining information from multiple association rules, and to provide the best set of recommendations to maximize the likelihood of purchase. Similarly, combining information on prior purchase histories and revenue optimization models enables a new blend of practical business decision making. As noted, this integration of data mining and optimization has been labeled business analytics. IBM and other major vendors are developing new business groups focused on analytics that arise from combinations of organizations in optimization and data mining (Turban et al. 2010, pp. 78).

Concluding Remarks

By detecting patterns hitherto unknown, data mining techniques could suggest new modes to pursue old objectives. They could even allow the formulation of better, more sophisticated models in the wake of new information. In general, the gains to be made from exploiting newly discovered information are significantly higher than the marginal improvements that can be made by improving existing solution procedures. As the volume and types of data being collected increase, so will the need for better tools to analyze the data. Consequently, the future of data mining seems to be full of possibilities. The enthusiasm for discovering new information, however, needs to be tempered with the need to address privacy concerns, as not doing so could have long term repercussions on the parties involved.

See

- ▶ Artificial Intelligence
- ▶ Cluster Analysis
- ▶ Computer Science and Operations Research Interfaces
- ▶ Decision Trees
- ▶ Neural Networks
- ▶ Nonlinear Programming
- ▶ Visualization

References

Babcock, C. (2006, January 9). Data, data, everywhere. *InformationWeek*.

Claburn, T. (2008, October 7). Counterterrorist data mining needs privacy protection. *InformationWeek*.

Department of Homeland Security. (2009, December). *DHS Privacy Office: 2009 data mining report to Congress*.

Field, A. (2010, April 3). Legal briefing: Will drugmakers' prescription data mining be undermined? *Daily Finance*.

Flynn, L. (2006, January 23). Like this? You'll hate that. (not all web recommendations are welcome.). *The New York Times*.

Hand, D., Mannila, M., & Smyth, P. (2001). *Principles of data mining*. Cambridge, MA: MIT Press.

Lai, E. (2008, October 14). Teradata creates elite club for petabyte-plus data warehouse customers. *Computerworld*.

Ólafsson, S. (2006, November 2006). Editorial: introduction to operations research and data mining. *Computers and Operations Research, 33*(11).

Turban, E., Sharda, R., & Delen, D. (2010). *Decision support and business intelligence systems* (9th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Data Warehousing

Paul Gray
Claremont Graduate University, Claremont, CA, USA

Introduction

The data warehouse is one of the key information infrastructure resources for Operations Researchers. Its difference from the conventional transactional database, which is used to keep track of individual events, is shown in Table 1.

The typical transaction database contains details about individual transactions such as the purchase of merchandise or individual invoices sent or paid.

Data Warehousing, Table 1 Data warehouse vs. transaction database

Data Warehouse	Subject oriented	Integrated	Time-variant	Non-volatile
Transaction Database	Transaction oriented	Un-integrated	Current status	Changes as trans- actions occur

Transactional databases are concerned with operations while data warehouses are organized by subject. For example, operational data in a bank focuses on transactions involving loans, savings, credit cards, and trust accounts, while the data warehouse is organized around customer, vendor, product, and activity history.

The continually changing transactional data is not in the form needed for planning, managing, and analyzing. That is where the data warehouse comes in.

The classic data warehouse is defined as “a subject oriented, integrated, non-volatile, time variant, collection of data to support management’s decisions” (Inmon 1992, p. 29).

The characteristics of the data warehouse that were summarized in Table 1 are given in more detail in Table 2.

In addition, the characteristics of the data itself are different, as shown in Table 3.

Data warehouses are really databases that provide both aggregated and detailed data for decision making. They are usually physically separated from both the organization’s transaction databases and its operational systems.

Note that data normalization, which is used in transactional databases, makes sure that an individual data point appears once and only once. Normalization is not required conceptually in data warehouses. Some data warehouse designs, however, do normalize their data.

Flow of Data

The flow of data into and out of the data warehouse follows these steps:

1. Obtain inputs
2. Clean inputs
3. Store in the warehouse
4. Provide output for analysis

Inputs to the data warehouse are the first step in what is called the extract, transform, and load process (ETL). Data sources, often from what are called legacy systems, push data to the warehouse rather than the warehouse pulling data from the sources. The sources send updates to the data warehouses at pre-specified intervals. This operation is performed on a fixed schedule where the interval between updates can range from nearly real time to once a day or longer, depending on the source.

Each source may have its own convention for what to call things and may even use different names and/or different metrics. For example, different transactional databases may store gender as (m, f), (1, 0), (x, y), (male, female) or may have different names for the same person (e.g., S. Smith, Sam Smith, and S. E. Smith). To overcome inconsistencies and to make

Data Warehousing, Table 2 Data warehouse characteristics

Subject orientation	Data are organized by how users refer to it, not by client
Data Integration	Data are organized around a common identifier, consistent names, and the same values throughout. Inconsistencies are removed.
Time	Data provide time series and focus on history, rather than current status.
Non-volatile	Data can be changed only by the upload process, not by the user.

Data Warehousing, Table 3 Characteristics of data in the warehouse

Summarized	In addition to current operational data when needed, data summaries used for decision making are also stored.
Larger database	Time series implies much more data is included.
Not normalized	Data can be redundant.
Metadata	Includes data about how the data is organized and what it means.
Sources of input data	Data comes from operational systems

sure that users see only one version of the truth, data cleansing is performed by the warehouse on the input.

Data cleansing involves changing the input data so that it meets the warehouse's standards. Specialized software (usually referred to as ETL) makes the input data extracted from the sources consistent (e.g., in format, scaling, and naming) with the way data is stored in the warehouse. For example, the warehouse standardizes on one of the formats for gender and translates all other versions to the standard. Transformation uses metadata (i.e., data about the data) to accomplish this. The data are loaded (i.e., stored) in the warehouse only after they are cleansed. The goal is to establish a single value of the truth within the warehouse.

The data warehouse is used for analytics and routine reporting. Both create information useful to managers and professionals. Analytics refers to using models and performing computations on the data. Routine reporting refers to creating, documents, tables, and graphics, usually on a repetitive schedule. Routine outputs include dashboards (which mostly present status), scorecards (which show how well goals are being met), and alerts (which notify managers when current values are outside prescribed limits).

What is in the Data Warehouse

The data warehouse contains not only the current detail data that was transferred from the legacy systems, but also lightly summarized or highly summarized data, as well as old detail data. Metadata are usually also stored in the data warehouse.

The current detail data reflects the most recent happenings and is usually stored on disk. Detail data is voluminous and is stored at higher levels of granularity. Granularity refers to the level of detail provided in the data warehouse. The more detail provided, the higher the level of granularity. The highest level is transaction data such as is required for data mining. For decision support, analysis, and planning, the level of granularity can be much lower. Granularity is an important trade-off because the higher the level of granularity, the more data must be stored, the greater the level of detail available, and the more computing needs to be done, even for problems that do not use that level of granularity. For example, if a gasoline company records every motorist's stop at its stations, it can use the credit transaction to understand its customers detailed buying patterns. For total sales by station, that level of granularity is not needed.

Lightly summarized data is generally used at the analyst level, whereas highly summarized data (which is compact and easily accessible) is used by senior managers. The choice of summarization level involves tradeoffs because the more highly summarized the data, the more the data is actually accessed and used, the quicker it is to retrieve, but the less detail is available for understanding it. One way to speed query response time is to pre-calculate aggregates which are referred to often, such as annual sales data.

To keep storage requirements within reason, older data are moved to lower cost storage with much slower data retrieval. An aging process within the data warehouse is used to decide when to move data to mass storage.

Metadata contains two types of information:

1. What the user needs to know to be able to access the data in the warehouse. It tells the user what is stored in the warehouse and where to find it.
2. What information systems personnel need to know about how data is mapped from operational form to warehouse form, i.e., what transformations occurred during input and the rules used for summarization.

Metadata keeps track of changes made converting, filtering, and summarizing data, as well as changes made in the warehouse over time, e.g., data added, data no longer collected, and format changes.

Warehouse Data Retrieval and Analysis

The data stored in the data warehouse are optimized for speedy retrieval through on-line analytical processing (OLAP). The retrieval methods depend on the data format. The three most common are:

- Relational OLAP (ROLAP), which works with relational databases
- Multidimensional OLAP (MOLAP) for data stored in multi-dimensional arrays
- Hybrid OLAP (HOLAP) which works with both relational and multidimensional databases.

OLAP involves answering multidimensional questions such as the number of units of Product

A sold in California at a discount to resellers in November (i.e., product, state, terms of sale, customer class, time).

To enable relational databases (that store data in two dimensions) to deal with multidimensionality, two types of tables are introduced: fact tables that contain numerical facts, or dimension tables that contain pointers to the fact tables and show where the information can be found. A separate dimension table is provided for each dimension (e.g., market, product, time). Fact tables tend to be long and thin and the dimension tables tend to be small, short, and wide. Because a single fact table is pointed to by several dimension tables, the visualization of this arrangement looks like a star and hence is called a star schema. A variant, used when the number of dimensions is large and multiple fact tables share some of the same dimension tables, is called a snowflake schema.

Multidimensionality allows analysts to slice and dice the data, i.e., to systematically reduce a body of data into smaller parts or views that yield more information. Slice and dice is also used to refer to the presentation of warehouse information in a variety of different and useful ways.

Why a Separate Warehouse?

A fundamental tenet of data warehouses is that their data are separate from operational data. The reasons for this separation are:

Performance. Requests for data for analysis are not uniform. At some times, for example, when a proposal is being written or a new product is being considered, huge amounts of data are required. At other times, the demand may be small. The demand peaks create havoc with conventional on-line transaction systems because they slow them down considerably, keeping users (and often customers) waiting.

Data Access. Analysis requires data from multiple sources. These sources are captured and integrated by the warehouse.

Data Formats. The data warehouse contains summary and time-based data as well as transaction data. Because the data are integrated, the information in the warehouse is kept in a single, standard format.

Data Quality. The data cleansing process of ETL creates a single version of the truth.

Other Forms of Data Warehouses

As organizations found new ways of using the warehouse, they created specialized forms for specific uses. Among these are:

- Data marts
- Operational data stores
- Real-time warehouses
- Data warehouse appliances
- Data warehouses in the cloud
- Separate data warehouses for casual and power users

Data marts are a small-scale version of a data warehouse that include all the characteristics of an enterprise data warehouse, but are much smaller in size and cost. Data marts can be independent or dependent.

- Independent data marts are typically stand-alone units used by departments or small strategic business units that often support only specific subject areas. A data mart is appropriate if it is the only data warehouse for a small or medium sized firm. Multiple independent data marts become a problem rather than a solution if they differ from department to department. Integrating them so that there is only a single value of the truth throughout the organization is difficult, particularly if a comprehensive data warehouse is later attempted.
- Dependent data marts, such as those used by analytics groups, contain a subset of the warehouse data needed by a particular set of users. To maintain a single value of the truth, care is taken that the dependent data mart does not change the data from the warehouse.

An Operational Data Store (ODS) is a data warehouse for transaction data. It is a form of data warehouse for operational use. The ODS is used where some decisions need to be made in near real-time and require the characteristics of a warehouse (e.g., clean data). The ODS is subject oriented and integrated like the warehouse but, unlike the data warehouse, information in an ODS can be changed and updated rather than retained forever. Thus, an ODS contains current and near-current information, but not much historical data.

When data moves from legacy systems to the ODS, the data are re-created in the same form as in the warehouse. Thus, the ODS converts data, selects among sources, may contain simple summaries of the current situation for management use, alters the key structures and the physical structure of the data, as well as its internal representation. Loading data into a data warehouse from an ODS is easier than loading from individual legacy systems, because most of the work on the data has been performed. It contains much less data than a data warehouse but also includes some that is not stored in the data warehouse. The ODS is usually loaded more frequently by data sources than the warehouse to keep it much more current. For example, the Walmart ODS receives information every 15 minutes.

The real-time data warehouse is used to support ongoing analysis and actions. A form of operational data store, real time data warehouses are closely tied to operational systems. They hold detailed, current data and try to use even shorter times between successive loadings than operational data stores. With these data warehouses, enterprises can respond to customer interactions and changing conditions in real time. For example, credit card companies use it to detect and stop fraud as it happens, a transportation company uses it to reroute its vehicles, and online retailers use it to communicate special offers based on a customer's Web surfing or mobile phone behavior. The real-time data warehouse is an integral part of both short-term (tactical) and long-term (strategic) decisions.

The real-time data warehouse changes the decision support paradigm, which has long been associated with strategic decision making. It supplies support for operational decision making such as customer-facing (direct interactions or communications with customers) and supply chain applications.

A data warehouse appliance is similar in concept to an all-in-one PC, i.e., it integrates the physical components of a data warehouse (servers, storage, operating system) with a database management system and software optimized for the data warehouse. These low-cost appliances are designed to provide terabyte to petabyte capacity warehouses.

Cloud computing refers to using the networked, on-demand, shared resources available through the Internet for virtual computing. Typically, rather than each firm owning its own warehouse, a third-party vendor provides a centralized service to multiple clients based on hardware and software usage. Although, as of 2010 - no data warehouse in the cloud exists, some inferences can be drawn. Agosta (2008) argues that in cloud computing the data in a warehouse will have to be location independent and transparent rather than being a centralized, non-volatile repository. Furthermore, the focus will be on distributed data marts and analytics rather than large data stores because of the problems and costs in moving the huge amounts of data in a warehouse to the cloud.

Data warehouses attract two types of users (Eckerson 2010):

- Casual users. These users are executives and other knowledge workers who consume information but do not usually create it. Their use is mostly static. They check dashboards, monitor regular reports, respond to alerts, and only occasionally dig deeper into the warehouse to create ad hoc reports.
- Power users. These users explore the data and build models. Conventional reports are insufficient for their needs. They model data in unique ways and supplement warehouse contents with data obtained from other sources.

In most organizations, the conventional data warehouse is used by both types of users despite their different needs. Some organizations, however, are moving to separate warehouses, one for each type of user. The conventional data warehouse feeds its data to the one for the power users, so that there is still only one version of the truth. In these organizations, conventional data warehouses continue to serve casual users whose requirements are mostly static. The idea is that performance gains are achieved by creating a separate warehouse customized to power users. Over the years, the special warehouses for power users have operated under a variety of names such as exploration data warehouse (for number crunching) (Inmon 1998), prototype data warehouse (for new approaches to warehouse design), and data warehouse sandbox. Eckerson (2010) describes to three types of sandbox architectures for analytics: physical, virtual, and desktop.

The physical sandbox is built around a data warehouse appliance or a specialized database with rapid access (e.g., columnar or massively parallel processing) that contains a copy of the data in the warehouse. Complex queries from the data warehouse are offloaded and used, together with data not stored in the warehouse. The result is that runaway queries (so large that they overload the warehouse) do not slow the warehouse and analysts can safely and easily explore large amounts of data.

The virtual sandbox is created inside the warehouse by using workload management utilities. Again, data can be added to that available in the warehouse. The advantage is that warehouse data does not need to be replicated. The disadvantage is that care must be taken to keep processing for casual and power users separate.

In desktop sandboxes, analysts are provided with powerful in-memory desktop databases that can be downloaded from the warehouse. Analysts gain local control and fast performance but much less data scalability than in physical or virtual sandboxes.

Applications

Data warehousing is central to data mining and business intelligence. Other applications include:

- Customer churn prediction
- Decision support
- Financial forecasting
- Insurance fraud analysis
- Logistics and inventory management
- Trend analysis

See

▶ Business Intelligence
▶ Data Mining
▶ Decision Support Systems (DSS)
▶ Information Systems and Database Design in OR/MS
▶ Visualization

References

Agosta, L. (2008). Data warehousing in the clouds: Making sense of the cloud computing market. *Beye Network*, 9 October 2008.

Eckerson, W. W. (2010). Dual BI architectures: The time has come. *The Data Warehousing Institute*, 18 Nov 2010.

Gray, P., & Watson, H. J. (1998). *Decision support in the data warehouse*. Upper Saddle River, NJ: Prentice-Hall.

Inmon, W. H. (1992). *Building the data warehouse*. New York: Wiley.

Inmon, W. H. (1998). *The exploration warehouse*. DM Review, June 1998.

Inmon, W. H. (2005). *Building the data warehouse* (4th ed.). Indianapolis, IN: Wiley.

Kimball, R., et al. (2009). *Kimball's data warehouse toolkit classics: The data warehouse toolkit, 2nd Edn; The data warehouse lifecycle, 2nd Edn; The data warehouse ETL toolkit*. New York: Wiley.

Sprague, R. H., & Carlson, E. D. (1982). *Building effective decision support systems*. Englewood Cliffs, NJ: Prentice Hall.

Database Design

▶ Information Systems and Database Design in OR/MS

DEA

▶ Data Envelopment Analysis

Decision Analysis

David A. Schum
George Mason University, Fairfax, VA, USA

Introduction

The term decision analysis identifies a collection of technologies for assisting individuals and organizations in the performance of difficult inferences and decisions. Probabilistic inference is a natural element of any choice made in the face of uncertainty. No single discipline can lay claim to all advancements made in support of these technologies. Operations research, probability theory, statistics, economics, psychology, artificial intelligence, and other disciplines have contributed valuable ideas now being exploited in various ways by individuals in many governmental, industrial, and military organizations. As the term decision analysis suggests, complex inference and choice tasks are decomposed into smaller and presumably more manageable elements, some of which are probabilistic and others preferential or value-related. The basic strategy employed in decision analysis is divide and conquer. The presumption is that individuals or groups find it more difficult to make holistic or global judgments required in undecomposed inferences and decisions than to make specific judgments about identified elements of these tasks. In many cases we may easily suppose that decision makers are quite unaware of all of the ingredients that can be identified in the choices they face. Indeed, one reason why a choice may be perceived as difficult is that the person or group charged with making this choice may be quite uncertain about the kind and number of judgments this choice entails. One major task in decision analysis is to identify what are believed to be the necessary ingredients of particular decision tasks.

The label decision analysis does not in fact provide a complete description of the activities of persons who employ various methods for assisting others in the performance of inference and choice tasks. This term suggests that the only thing accomplished is the decomposition of an inference or a choice into smaller elements requiring specific judgments or information. It is, of course, necessary to have some process by which these elements can be reassembled or aggregated so that a conclusion or a choice can be made. In other words, we require some method of synthesis of the decomposed elements of inference and choice. A more precise term for describing the emerging technologies for assistance in inference and choice would be the term decision analysis and synthesis. This fact has been noted in an account of progress in the field of decision analysis (Watson and Buede 1987). As it happens, the same formal methods that suggest how to decompose an inference or choice into more specific elements can also suggest how to reassemble these elements in drawing a conclusion or selecting an action.

Processes and Stages of Decision Analysis

Human inference and choice are very rich intellectual activities that resist easy categorization. Human inferences made in natural settings (as opposed to contrived classroom examples) involve various mixtures of the three forms of reasoning that have been identified: (1) deduction (showing that some conclusion is necessary), (2) induction (showing that some conclusion is probable), and (3) abduction (showing that something is possibly or plausibly true). There are many varieties of choice situations that can be discerned. Some involve the selection of an action or option such as where to locate a nuclear power plant or a toxic waste disposal site. Quite often one choice immediately entails the need for another and so we must consider entire sequences of decisions. It is frequently difficult to specify when a decision task actually terminates. Other decisions involve determining how limited resources may best be allocated among various demands for these resources. Some human choice situations involve episodes of bargaining or negotiation in which there are individuals or groups in some competitive or adversarial posture. Given the richness of inference and choice, analytic and synthetic methods differ from one situation to another as observed in several surveys of the field of decision analysis (von Winterfeldt and Edwards 1986; Watson and Buede 1987; Clemen 1991; Shanteau et al. 1999). Some general decision analytic processes can, however, be identified.

Most decision analyses begin with careful attempts to define and structure an inference and/or decision problem. This will typically involve consideration of the nature of the decision problem and the individual or group objectives to be served by the required decision(s). A thorough assessment of objectives is required since it is not possible to assist a person or group in making a wise choice in the absence of information about what objectives are to be served. It has been argued that the two central problems in decision analysis concern uncertainty and multiple conflicting objectives (von Winterfeldt and Edwards 1986, pp. 4–6). A major complication arises when, as usually observed, a person or a group will assert objectives that are in conflict. Decisions in many situations involve multiple stakeholders and it is natural to expect that their stated objectives will often be in conflict. Conflicting objectives signal the need for various tradeoffs that can be identified. Problem structuring also involves the generation of options, actions, or possible choices. Assuming that there is some element of uncertainty, it is also necessary to generate hypotheses representing relevant alternative states of the world that act to produce possibly different consequences of each option being considered. The result is that when an action is selected we are not certain about which consequence or outcome will occur.

Another important structuring task involves the identification of decision consequences and their attributes. The attributes of a consequence are measurable characteristics of a consequence that are related to a decision maker's asserted objectives. Identified attributes of a consequence allow us to express how well a consequence measures up to the objectives asserted in some decision task. Stated in other words, attributes form value dimensions in terms of which the relative preferability of consequences can be assessed. There are various procedures for generating attributes of consequences from stated objectives (e.g., Keeney and Raiffa 1976, pp. 31–65). Particularly challenging are situations in which we have multiattribute or vector consequences. Any conflict involving objectives is reflected in conflicts among attributes and signals the need for examining possible tradeoffs. Suppose, for some action A_i and hypothesis H_j, vector consequence $\boldsymbol{C}\mathrm{v}_{ij}$ has attributes $\{A_1, A_2, \ldots, A_r, \ldots, A_s, \ldots, A_t\}$. The decision maker may have to judge how much of A_r to give up in order to get more of A_s; various procedures facilitate such judgments. Additional structuring is necessary regarding the inferential element of choice under uncertainty. Given some exhaustive set of mutually exclusive hypotheses or action-relevant states of the world, the decision maker will ordinarily use any evidence that can be discovered that is relevant in determining how probable are each of these hypotheses at the time a choice is required. No evidence comes with already-established relevance, credibility, and inferential force credentials, these credentials have to be established by argument. The structuring of complex probabilistic arguments is a task that has received considerable attention (e.g., see Pearl 1988; Neapolitan 1990; Schum 1990, 1994).

At the structural stage just discussed, the process of decomposing a decision is initiated. On some occasions such decomposition proceeds according to formal theories of probability and value taken to be normative. It may even happen that the decision of interest can be represented in terms of some existing mathematical programming or other formal technique common in operations research. In some cases the construction of a model for a decision problem proceeds in an iterative fashion until the decision maker is satisfied that all ingredients necessary for a decision have been identified. When no new problem ingredients can be identified the model that results is said to be a requisite model (Phillips 1982, 1984). During the process of decomposing the probability and value dimensions of a decision problem it may easily happen that the number of identified elements quickly outstrips a decision maker's time and inclination to provide judgments or other information regarding each of these elements. The question is: how far should the process of divide and conquer be carried out? In situations in which there is not unlimited time to identify all conceivable elements of a decision problem, simpler or approximate decompositions at coarser levels of granularity have to be adopted.

In most decision analyses there is a need for a variety of subjective judgments on the part of persons involved in the decision whose knowledge and experience entitles them to make such judgments. Some judgments concern probabilities and some concern the value of consequences in terms of identified attributes. Other judgments may involve assessment of the relative importance of consequence attributes. The study of methods for obtaining dependable quantitative judgments from people represents one of the most important contributions of psychology to decision analysis (for a survey of these judgmental contributions, see von Winterfeldt and Edwards 1986). After a decision has been structured and subjective ingredients elicited, the synthetic process in decision analysis is then exercised in order to identify the best conclusion and/or choice. In many cases such synthesis is accomplished by an algorithmic process taken as appropriate to the situation at hand. Modern computer facilities allow decision makers to use these algorithms to test the consequences of various possible patterns of their subjective beliefs by means of sensitivity analyses. The means for defending the wisdom of conclusions or choices made by such algorithmic methods re-quires consideration of the formal tools used for decision analysis and synthesis.

Theories of Analysis and Synthesis

Two major pillars upon which most of modern decision analysis rests are theories of probabilistic reasoning and theories of value or preference. A very informative summary of the roots of decision theory has been provided by Fishburn (1999). It is safe to say that the conventional view of probability, in which Bayes' rule appears as a canon for coherent or rational probabilistic inference, dominates current decision analysis. For some body of evidence $\boldsymbol{E}$v, Bayes' rule is employed in determining a distribution of posterior probabilities $P(H_k|\boldsymbol{E}\mathrm{v})$, for each hypothesis H_k in an exhaustive collection of mutually exclusive decision-relevant hypotheses. The ingredients Bayes' rule requires, prior probabilities (or prior odds) and likelihoods (or likelihood ratios), are in most cases assumed to be assessed subjectively by knowledgeable persons. In some situations, however, appropriate relative frequencies may be available. The subjectivist view of probability, stemming from the work of Ramsey and de Finetti, has had a very sympathetic hearing in decision analysis (see Mellor 1990, and de Finetti 1972, for collections of the works of Ramsey and de Finetti).

Theories of coherent or rational expression of values or preferences stem from the work of von Neumann and Morgenstern (1947). In this work appears the first attempt to put the task of stating preferences on an axiomatic footing. Adherence to the von Neumann and Morgenstern axioms places judgments of value on a cardinal or equal-interval scale and are often then called judgments of utility. These axioms also suggest methods for eliciting utility judgments and they imply that a coherent synthesis of utilities and probabilities in reaching a decision consists of applying the principle of expected utility maximization. This idea was extended in the later work of Savage (1954), who adopted the view that the requisite probabilities are subjective in nature. The canon for rational choice emerging from the work of Savage is that the decision maker should choose from among alternative actions by determining which one has the highest subjective expected utility (SEU). Required aggregation of probabilities is assumed to be performed according to Bayes' rule. In some works, this view of action-selection is called Bayesian decision theory (Winkler 1972; Smith 1988).

Early works by Edwards (1954, 1961) stimulated interest among psychologists in developing methods for probability and utility elicitation; these works also led to many behavioral assessments of the adequacy of SEU as a description of actual human choice mechanisms. In a later work, Edwards (1962) proposed the first system for providing computer assistance in the performance of complex probabilistic inference tasks. Interest in the very difficult problems associated with assessing the utility of multiattribute consequences stems from the work of Raiffa (1968). But credit for announcing the existence of the applied discipline now called decision analysis belongs to Howard (1966, 1968).

Decision Analytic Strategies

There are now many individuals and organizations employed in the business of decision analysis. The inference and decision problems they encounter are many and varied. A strategy successful in one context may not be so successful in another. In most decision-analytic encounters, an analyst plays the role of a facilitator, also termed high priests (von Winterfeldt and Edwards 1986, p. 573). The essential task for the facilitator is to draw out the experience and wisdom of decision makers while guiding the analytic process toward some form of synthesis. In spite of the diversity of decision contexts and decision analysts, Watson and Buede (1987, pp. 123–162) were able to identify the following five general decision analytic strategies in current use. They make no claim that these strategies are mutually exclusive.

1. Modeling. In some instances decision analysts will focus upon efforts to construct a conceptual model of the process underlying the decision problem at hand. In such a strategy, the decision maker(s) being served not only provide the probability and value ingredients their decision requires but are also asked to participate in constructing a model of the context in which this decision is embedded. In the process of constructing these often-complex models, important value and uncertainty variables are identified.
2. Introspection. In some decision analytic encounters, a role played by the facilitator is one of assisting decision makers in careful introspective efforts to determine relevant preference and probability assessments necessary for a synthesis in terms of subjective expected utility maximization. Such a process places great emphasis upon the reasonableness and consistency of the often large number of value and probability ingredients of action selection.
3. Rating. In some situations, especially those involving multiple stakeholders and multiattribute consequences, any full-scale task decomposition would be paralytic or, in any case, would not provide the timely decisions so often required.

In order to facilitate decision making under such circumstances, models involving simpler probability and value assessments are often introduced by the analyst. In some forms of decision analysis, many of the difficult multiattribute utility assessments are made simpler through the use of various rating techniques and by the assumption of independence of the attributes involved.

4. Conferencing. In a decision conference the role of the decision analyst as facilitator (or high priest) assumes special importance. In such encounters, often involving a group of persons participating to various degrees in a decision, the analyst promotes a structured dialogue and debate among participants in the generation of decision ingredients such as options, hypotheses and their probability, and consequences and their relative value. The analyst further assists in the process of synthesis of these ingredients in the choice of an action. The subject matter of a decision conference can involve action selection, resource allocation, or negotiation.
5. Developing. In some instances, the role of the decision analyst is to assist in the development of strategies for recurrent choices or resource allocations. These strategies will usually involve computer-based decision support systems or some other computer-assisted facility whose development is justified by the recurrent nature of the choices. The study and development of decision support systems has itself achieved the status of a discipline (Sage 1991). An active and exciting developmental effort concerns computer-implemented influence diagrams stemming from the work of Howard and Matheson (1981). Influence diagram systems can be used to structure and assist in the performance of inference and/or decision problems and have built-in algorithms necessary for the synthesis of probability and value ingredients (e.g., Shachter 1986; Shachter and Heckerman 1987; Breese and Heckerman 1999). Such systems are equally suitable for recurrent and nonrecurrent inference and choice tasks.

Controversies

As an applied discipline, decision analysis inherits any controversies associated with theories upon which it is based. There is now a substantial literature challenging the view that the canon for probabilistic inference is Bayes' rule (e.g., Cohen 1977, 1989; Shafer 1976). Regarding preference axioms, Shafer (1986) has argued that no normative theories of preference have in fact been established and that existing theories rest upon an incomplete set of assumptions about basic human judgmental capabilities. Others have argued that the probabilistic and value-related ingredients required in Bayesian decision theory often reflect a degree of precision that cannot be taken seriously given the imprecise or fuzzy nature of the evidence and other information upon which such judgments are based (Watson et al. 1979). Philosophers have recently been critical of contemporary decision analysis. Agreeing with Cohen and Shafer, Tocher (1977) argued against the presumed normative status of Bayes' rule. Rescher (1988) argued that decision analysis can easily show people how to decide in ways that are entirely consistent with objectives that turn out not to be in their best interests. Keeney's work (1992) took some of the sting out of this criticism. Others (e.g., Dreyfus 1984) question whether or not decomposed inference and choice is always to be preferred over holistic inference and choice; this same concern is reflected in other contexts such as law (Twining 1990, pp. 238–242). So, the probabilistic and value-related bases of modern decision analysis involve matters about which there will be continuing dialogue and, perhaps, no final resolution. This acknowledged, decision makers in many contexts continue to employ the emerging technologies of decision analysis and find, in the process, that very complex inferences and choices can be made tractable and far less intimidating.

See

- ▶ Choice Theory
- ▶ Decision Analysis in Practice
- ▶ Decision Making and Decision Analysis
- ▶ Decision Support Systems (DSS)
- ▶ Decision Trees
- ▶ Group Decision Making
- ▶ Influence Diagrams
- ▶ Multi-attribute Utility Theory
- ▶ Utility Theory

References

Breese, J., & Heckerman, D. (1999). Decision-theoretic troubleshooting: A framework for repair and experiment. In J. Shanteau, B. Mellers, & D. Schum (Eds.), *Decision science and technology: Reflections on the contributions of ward Edwards* (pp. 271–287). Boston: Kluwer Academic.

Clemen, R. T. (1991). *Making hard decisions: An introduction to decision analysis*. Boston: PWS-Kent.

Cohen, L. J. (1977). *The probable and the provable*. Oxford: Clarendon Press.

Cohen, L. J. (1989). *An introduction to the philosophy of induction and probability*. Oxford: Clarendon Press.

De Finetti, B. (1972). *Probability, induction, and statistics: The art of guessing*. New York: Wiley.

Dreyfus, S. (1984). The risks ! and benefits ? Of risk-benefit analysis. *Omega, 12*, 335–340.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin, 41*, 380–417.

Edwards, W. (1961). Behavioral decision theory. *Annual Review of Psychology, 12*, 473–498.

Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Human Factors, 4*, 59–73.

Fishburn, P. (1999). The making of decision theory. In J. Shanteau, B. Mellers, & D. Schum (Eds.), *Decision science and technology: Reflections on the contributions of Ward Edwards* (pp. 369–388). Boston: Kluwer Academic.

Hammond, J., Keeney, R., & Raiffa, H. (2002). *Smart choices: A practical guide to making better decisions*. New York: Random House.

Howard, R. (1966). Decision analysis: Applied decision theory. In D. B. Hertz & J. Melese (Eds.), *Proceedings fourth international conference on operational research*. New York: Wiley-Interscience.

Howard, R. (1968). The foundations of decision analysis. *IEEE Transactions on Systems Science and Cybernetics, SSC-4*, 211–219.

Howard, R., & Matheson, J. (1981). Influence diagrams. In R. Howard & J. Matheson (Eds.), *The principles and applications of decision analysis* (Vol. 2). Menlo Park, CA: Strategic Decisions Group, 1984.

Keeney, R. (1992). *Value-focused thinking*. Cambridge, MA: Harvard University Press.

Keeney, R., & Raiffa, H. (1976). *Decision with multiple objectives: Preferences and value tradeoffs*. New York: Wiley.

Mellor, D. H. (1990). *F.P. Ramsey: Philosophical papers*. Cambridge, UK: Cambridge University Press.

Neapolitan, R. (1990). *Probabilistic reasoning in expert systems: Theory and algorithms*. New York: Wiley.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible reasoning*. San Mateo, CA: Morgan Kaufmann.

Phillips, L. (1982). Requisite decision modeling: A case study. *Journal of the Operational Research Society, 33*, 303–311.

Phillips, L. (1984). A theory of requisite decision models. *Acta Psychologica, 56*, 29–48.

Raiffa, H. (1968). *Decision analysis: Introductory lectures on choices under uncertainty*. Reading, MA: Addison-Wesley.

Rescher, N. (1988). *Rationality: A philosophical inquiry into the nature and rationale of reason*. Oxford: Clarendon.

Sage, A. (1991). *Decision support systems engineering*. New York: Wiley.

Savage, L. J. (1954). *The foundations of statistics*. New York: Wiley.

Schum, D. (1990). Inference networks and their many subtle properties. *Information and Decision Technologies, 16*, 69–98.

Schum, D. (1994). *Evidential foundations of probabilistic reasoning*. New York: Wiley.

Shachter, R. (1986). Evaluating influence diagrams. *Operations Research, 34*, 871–882.

Shachter, R., & Heckerman, D. (1987). Thinking backward for knowledge acquisition. *AI Magazine, Fall, 8*, 55–61.

Shafer, G. (1976). *A mathematical theory of evidence*. Princeton, NJ: Princeton University Press.

Shafer, G. (1986). *"Savage revisited," statistical science* (Vol. 1, pp. 463–501). Hayward, CA: Institute of Mathematical Statistics (with comments).

Shanteau, J., Mellers, B., & Schum, D. (1999). *Decision science and technology: Reflections on the contributions of Ward Edwards*. Boston: Kluwer Academic.

Smith, J. Q. (1988). *Decision analysis: A Bayesian approach*. London: Chapman and Hall.

Tocher, K. (1977). Planning systems. *Philosophical Transactions of the Royal Society of London, A287*, 425–441.

Twining, W. (1990). *Rethinking evidence: Exploratory essays*. Oxford: Basil Blackwell.

von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge, UK: Cambridge University Press.

Watson, S. R., & Buede, D. (1987). *Decision synthesis: The principles and practice of decision analysis*. Cambridge, UK: Cambridge University Press.

Watson, S. R., Weiss, J. J., & Donnell, M. L. (1979). Fuzzy decision analysis. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-9*(1), 1–9.

Winkler, R. L. (1972). *Introduction to Bayesian inference and decision*. New York: Holt, Rinehart, and Winston.

Decision Analysis in Practice

James E. Matheson
SmartOrg, Inc., Menlo Park, CA, USA

Introduction

Decision analysis (DA) is all about practice, as the title of Ronald Howard's defining paper (Howard 1966; presented in 1965) was "Decision Analysis: Applied Decision Theory." He went on to elaborate: "Decision analysis is a logical procedure for the balancing of the factors that influence a decision. The procedure incorporates uncertainties, values, and preferences in a basic structure that models a decision. Typically it includes technical, marketing, competitive, and environmental factors. The essence of the procedure is the construction of a structural model of the decision in a form suitable for computation and manipulation; the realization of this model is often a set of computer programs."

In about 1968, a program of DA was begun at Stanford Research Institute. This group rapidly grew into a major department called the Decision Analysis Group dedicated to helping decision makers in organizations, both industry and government, reach good decisions, while also consolidating these experiences and doing research on DA methodology (Howard and Matheson 1983). This group was the most intensive DA consulting group through the early 1980s. One of the powerful new methodological tools invented by this group was the Influence Diagram (see entry). DA practice has always developed new tools and approaches based on the challenges of real problems.

At the end of the next decade, with this experience behind him, Professor Howard goes on to say (Howard 1980), "Decision Analysis, as I have described it, is, as a formalism, a logical procedure for decision making. When Decision Analysis is practiced as an applied art the formalism interacts with the intuitive and creative facilities to provide understanding of the nature of the decision problem and therefore guidance in selecting a desirable course of action. I know of no other formal-artistic approach that has been so effective in guiding decision-makers."

In this sense there is no real theory of DA. Its philosophy is grounded in decision theory and systems engineering, with more recent contributions from psychology, but in the end it is an applied art. This Decision Engineering approach is discussed in depth in an INFORMS tutorial (Matheson 2005). This article describes some of the keys to good application and the kinds of positive changes DA promotes in the organizations that adopt it.

A Decision: The Defining Element

A decision is defined as an irrevocable allocation of resources. Exactly what is meant by irrevocable depends on the context. If a single individual—the decision maker (DM)—makes and executes a decision, then the decision and the irrevocable action are one – the individual might decide to take one path versus another along a road. Traveling down the new path is an irrevocable decision in the sense that changing the decision would require going back to the junction and taking the second path, but at a later time. However, when an organizational DM takes a big strategic decision, the DM asks many other people to take later irrevocable actions, which might not even be fully specified at that time of the original decision (for example, asking someone to find an appropriate company and acquire it). In these settings, a decision is often defined as a commitment to allocate resources, which opens new questions of possible execution failure and nested or sequential decisions. In any case, the decisions at hand provide the focus for DA, which distinguishes DA from all kinds of studies and statistical analyses that are not directly serving decisions. This means that, once the decision arena has been defined, the DM can guide all subsequent activity, such as modeling and information gathering, on its ability to inform better decisions. Issues that might make a great deal of difference to the outcome, but do not have the potential to change the decision taken, are unimportant, while issues of less impact but that do inform the decision are of greater importance. The DM uses this sort of decision sensitivity to intuitively and analytically guide the whole process, and to do what is most important to making a decision in the limited time and resources available to make it.

Framing: The Perceived Situation

Perhaps the biggest decision failure is a careful analysis of the wrong problem. Often a decision arises in an organization as just another tactical decision, when actually new strategies are called for – but strategy is not the prerogative or in the comfort zone of those considering the decisions. Thus, old products and whole companies are displaced by competitors who perceived the situation differently, and who were able to act in new ways. Also, executives spend most of their time and energy operating efficiently and find it difficult to "waste time" on strategy or to get into a strategic mind set. The beginning of a DA should review the decision frame, possibly bringing in outside perspectives and new team members, often expanding the frame, and then reviewing that frame at key points during the process. When a DA process gets stuck, reframing maybe in order (Matheson 1990).

Outcomes: What are the Results

In the face of uncertainty, the decision maker (DM) is forced to distinguish between decisions – what can be done, outcomes – what happens, and preferences – what is wanted. The DM wants good outcomes, but can only control the quality of the decisions, not the outcomes. For example, the DM may invest $10,000 in a venture having only a 10% chance of returning $10,000,000, and considers that a good investment. Quite likely, however, the bad outcome may occur. Clearly, the quality of this decision cannot be judged by its outcome; a bad outcome should not dissuade the DM from looking for similar good investments later. Given this distinction between decision quality and outcome quality, there is a need for a definition of a good decision – DA itself is that definition!

In many organizational cultures, champions are asked to claim that investment proposals are sure things and guarantee that they will succeed. On course, many of these investments fail, but inconsistency does not stop this irrational culture from persisting. However, organizations that can overcome a culture of hiding from uncertainty and instead actually search for the hidden uncertainties in their investments often outperform those that do not. Good DA vets these uncertainties, assesses their probabilities and impacts, and determines what to do about them, such as information gathering and hedging, or even creating new alternatives, before proceeding to recommend the primary decision – a principle called embracing uncertainty (Matheson and Matheson 1998).

There are well established procedures for assessing uncertainties and avoiding well-known biases, such as the work on probability assessment processes by Spetzler and Staël von Holstein (1975). Most practical decision analyses, however, do not require such careful assessment; three points, say 10-50-90 percentiles, are so much better than one single and often biased point. It is essential that those three points not be biased. Most of the de-biasing techniques of Spetzler and Staël von Holstein (1975) are useful preparation before assessing even a three-point distribution. Perhaps the most useful technique is backcasting, as it simultaneously eliminates all sorts of biases.

Preferences: What is Wanted

Because only one thing can be maximized, a good or optimal decision cannot be defined without being clear about value trade-offs that create a single measure to maximize. In most commercial decision analyses, it is best to reduce all values to monetary ones. In fact, seeking a monetary value scale is always a good practice, because money can often be spent to create better alternatives or seek better information, and, without a monetary scale, the DM cannot evaluate those efforts. There is a story about a Swedish executive who had promised the residents of a town that he would never close their factory, but, under hard times, he was facing heavy losses by keeping it open. He was asked by a decision analyst if he would close it if he were losing a million dollars a year, to which he quickly answered, “of course not – this is Sweden where we owe that much to the community.” He was then asked if he would close the plant if it were losing a hundred million dollars a year, to which he replied, “it would be our duty to close it as the country and our company cannot sustain such heavy losses.” After haggling over the price, he realized that the high monetary value he had just made explicit allowed him to visualize new alternatives, where he would close the plant, pay some additional closing costs to the community and guarantee workers jobs in other factories. He ultimately took these actions and saved his company from financial ruin. Being forced to make a monetary value tradeoff enabled him to invent to better alternatives. He was not valuing things like higher employment on an absolute scale. He was only assessing a tradeoff value in the context of his specific decision – this value is personal and subjective, just like probability, in this case not his own, but one he expresses as a fiduciary of the company he represents. Converting values to monetary equivalents is an excellent practice, because it establishes how much money could be afforded to build new alternatives – money is a common denominator to translate disparate values.

Decision Analysis in Practice, Fig. 1 Risk tolerance in millions of dollars as measured from *top* executives of three publically traded companies, A, B, and C

Size Measure	A	B	C	Approximate Ratio to Risk Tolerance
Net Sales	2,300	16,000	31,000	15:1
Net Income	120	700	1,900	1:1
Equity	1,000	6,500	12,000	6:1
Market Value	940	4,600	9,900	5:1
Risk Tolerance	150	1,000	2,000	

What about value over time? In a simple case, a highly rated company regularly adjusts or rebalances its financial capital at a weighted cost of capital of R%. If the company has opportunities (or preferences) that imply a value other than R%, the company should rearrange its investments using its banking relationships until its needs are exactly in line with the financial rate of R%. At that point, the company's own time preferences are exactly the same as the financial rate. Because of this harmonization process, this cost of capital becomes the company's own time value of money. Another way to state this observation is that the company should invest to maximize net present value (NPV) at its cost of capital, and then spread that NPV over time optimally using financial transactions at the same rate, separating investment funding and usage decisions.

How should preferences under uncertainty be treated? Assuming that each uncertainty has been characterized satisfactorily in the form of probability distributions over NPV, which investment should be picked? If the company is large enough to undertake many investments of this size during each year, then maximizing the expected value is a reasonable way to maximize long-term economic-value creation. However, if the range of the uncertainties could impact the financial structure or soundness of the company, it would be wise for it to be risk averse. Some financial pundits argue that companies traded on the stock market should not be risk averse as the shareholders can diversify. There are many arguments against this position, including the actual behavior of most companies, the cost of bankruptcy or other financial distress, the inability of the shareholder to gain information and change positions quickly (lack of liquidity), but, perhaps most significantly, are the availability of risk hedging options to the company that are not available to shareholders. The risk attitude of the company is assessed by asking series of questions about which of several hypothetical investments they would undertake or reject. This attitude is almost always captured as the risk tolerance, say expressed in millions of dollars, which is the parameter of an exponential utility function:

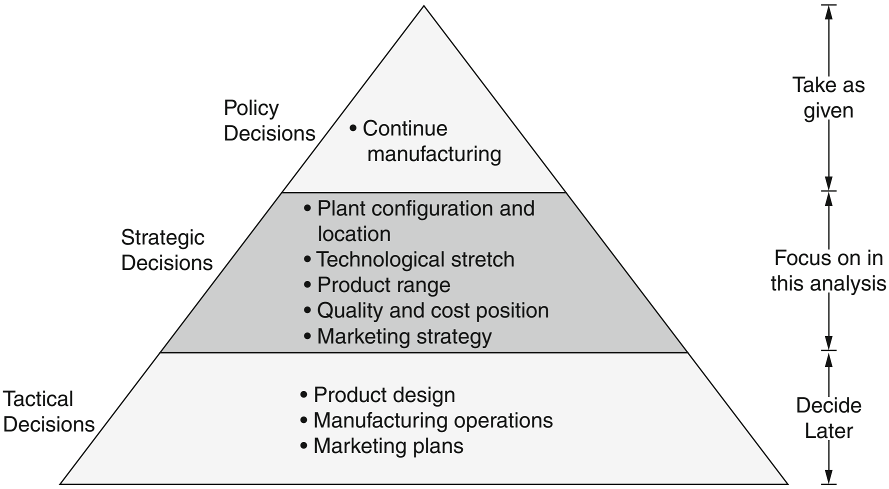
$$U(x) = -ae^{(-x/\rho)} \text{ where } a > 0 \text{ and } \rho = \text{risk tolerance}$$

One test question to determine the risk tolerance is considering a hypothetical but typical investment, in terms of complexity and time duration, where there is a 0.5 probability of winning the risk tolerance and a 0.5 probability of losing one-half that amount. The risk tolerance is then adjusted until the DM is indifferent between taking and rejecting this investment.

There are good arguments that risk tolerance should be set for the total organization and not for a division or a project. One advantage of being a division of a large organization is to be able to use the corporate risk tolerance, which a similar stand-alone organization could not do. Figure 1 compares the measured risk tolerances of three large corporations, which were all engaged in a joint venture. This chart can be used to get an initial approximation for other public companies, commonly by estimating risk tolerance as 1/6 of shareholders' equity or 1/5 of the market value of outstanding shares of stock.

Investments with a range of outcomes on the order of the risk tolerance need explicit treatment using utility theory. Investments with a range of outcomes less than of 10% of the risk tolerance should usually be evaluated using expected values, and investments with a range of outcomes larger than the risk tolerance should be avoided, partnered, or treated by a very

Decision Analysis in Practice, Fig. 2 Decision hierarchy for a plant modernization decision



experienced decision analyst. The author has seen one such case in a lifetime of professional practice. If the exponential utility will not suffice, the analysis is in very deep water indeed! In dealing with uncertainties large enough to require risk aversion, there is a need to beware of dependencies among uncertainties in other investments or the background cash flow of the organization. Hedging and diversification impacts are likely to overshadow other considerations.

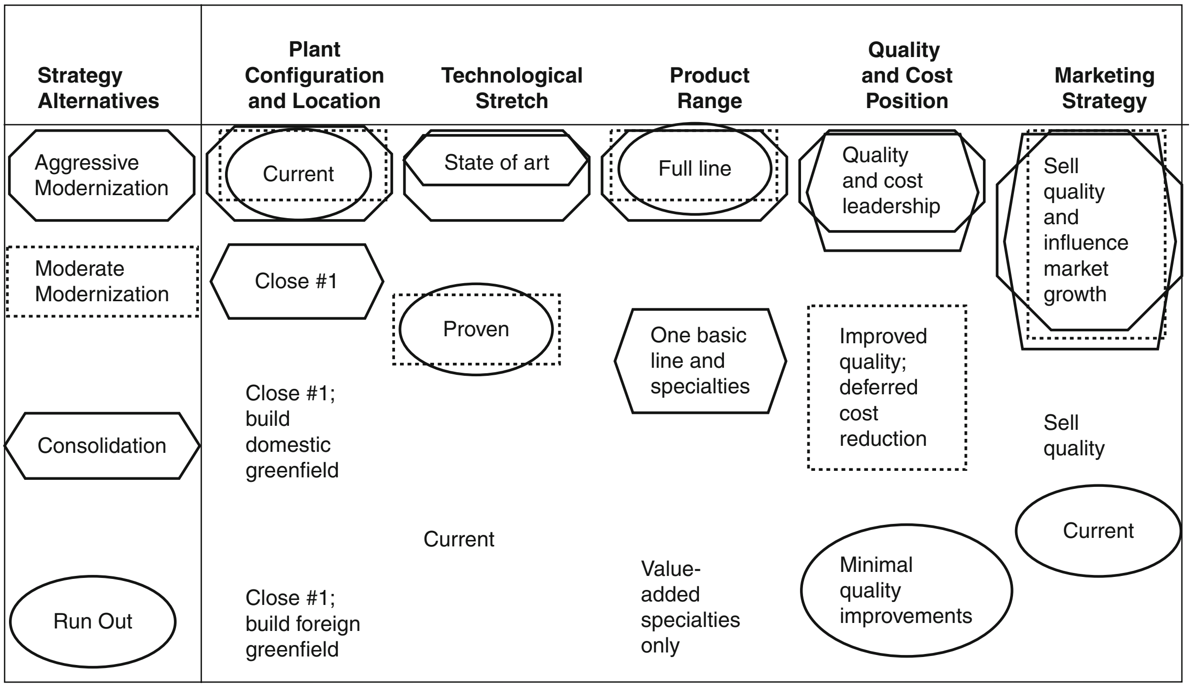
Alternatives: What Can be Done

In simple decisions problems, such as classroom examples, a limited number of well-specified alternatives are given. In most real situations, however, new alternatives can and should be created to uncover more valuable ones. Part of the natural reluctance of organizations to generate and consider new alternatives is that the decision problems arise out of situations where natural alternatives are evident. In addition, those product or investment champions and others who have made an emotional investment by picking winners prematurely, see alternative generation as a waste of time or even a direct threat. There are many ways to create new alternatives, but a simple one is to use the project team itself in a session with a ground rule that at least five new significantly different alternatives must be developed. There are many tools to stimulate creativity, most requiring that a wealth of information and new possibilities be put on the table before evaluating them; such as examples of what others have been done, what competitors are saying, what consumers are asking for. After the analysis enters the financial modeling stage; sensitivity analysis should also be used to drive the discussion of alternatives that minimize risk (hedge or diversify) or take advantage of uncertainties.

For situations with complex multidimensional alternatives, decision hierarchies and strategy tables are extremely useful. The decision hierarchy for a plant modernization decision (Fig. 2) identifies the strategic decisions under consideration, the policy decisions that are not currently being questioned, and the tactical or implementation decisions which will be made or optimized after the strategy is selected. The list of identified strategic decisions are further specified in the columns of the strategy table, illustrated in Fig. 3. The columns list specific mutually-exclusive alternatives for each strategy variable. Thus, a selection of one item from each column constitutes a well-specified strategic alternative. The special column at the left gives names and symbols for each alternative, which is read by following its symbol across the columns. Further descriptions of these tools can be found in Matheson and Matheson (1998) and McNamee and Celona (2007).

Decision Modeling: Analyzing as Simply as Possible

The process of DA uses the decision to be made as a guide to cut through many complex modeling issues. Often details, such as numerous market segments or



Strategy Alternatives	Plant Configuration and Location	Technological Stretch	Product Range	Quality and Cost Position	Marketing Strategy
Aggressive Modernization	Current	State of art	Full line	Quality and cost leadership	Sell quality and influence market growth
Moderate Modernization	Close #1	Proven	One basic line and specialties	Improved quality; deferred cost reduction	Sell quality
Consolidation	Close #1; build domestic greenfield	Current	Value-added specialties only	Minimal quality improvements	Current
Run Out	Close #1; build foreign greenfield				

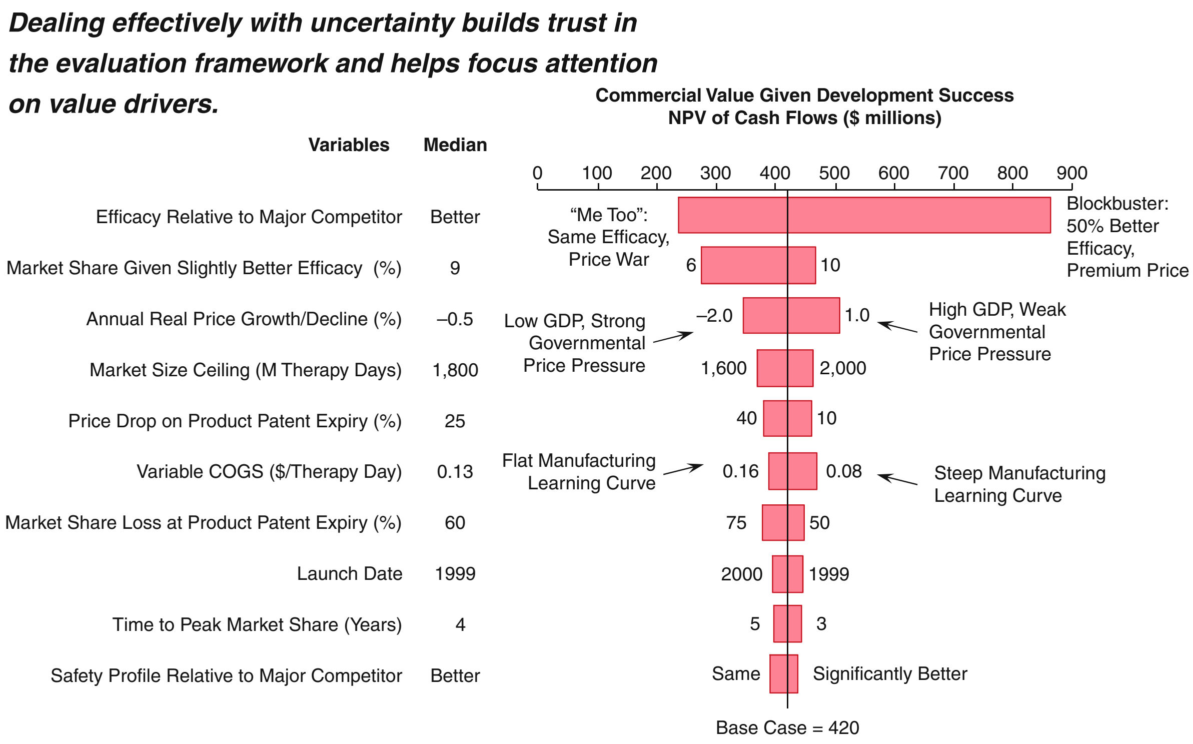
Decision Analysis in Practice, Fig. 3 Strategy table for a plant modernization decision

multiple product generations, can be treated with multipliers, followed by sensitivity analysis to the value of those multipliers, to determine if something important was missed. Verisimilitude is unimportant, only the impact on gaining clarity of action. Good modeling for decision making is an important professional task, see McNamee and Celona, (2007).

A special kind of sensitivity analysis called a tornado chart (Fig. 4) is a key tool for checking the model and gaining new insights. Each uncertain variable is varied one at a time over the range of the low (10 percentile) and high (90 percentile) assessments, to determine the range of (deterministic) NPV resulting from different runs of the model, usually while holding the other values at their medians. Notice that output ranges of each variable correspond to the same range of uncertainty on their inputs, so if the results are arranged in a decreasing order of the output ranges, they are also in order of the impact of each uncertainty on value, as in Fig. 4. Since for independent variables, the uncertainty ranges should add as the square root of the sum of the squares, only the first several results are big contributors, which often produces insight into which factors are driving risk, as well as ideas for how to reduce that risk. More sophisticated tornado diagrams overlay results for multiple alternatives to give insight into which uncertainties could actually cause a decision switch, as these would be the most critical to learn more about.

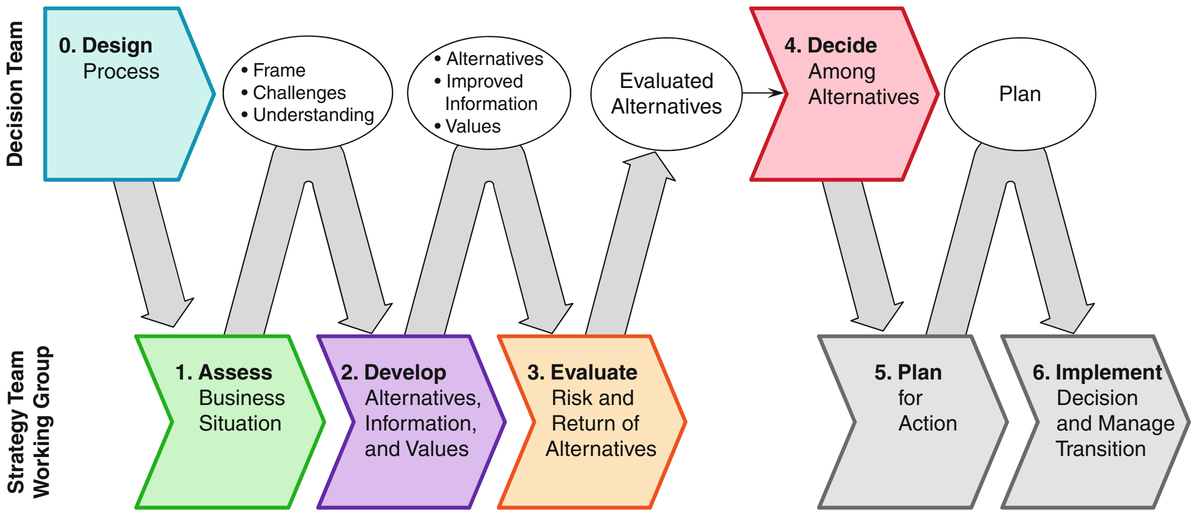
Commitment to Action: Getting It Done

The author has decided to diet many times, without actually following through. And that is only dealing with himself! It is much more difficult to align an organization to carry out the chosen action. A good analysis sets the stage for implementation success at the beginning by the choice of individuals involved in reaching the decision. It is natural not to put the potential naysayers on the decision making or the decision analyzing team, but if they are not chosen, they will often veto the result, overtly if they have the power and covertly if not. It is best to put any skeptical person with veto power on the team, even if only in a review board role, and require that they raise their issues during the analysis process rather than objecting later – speak up or forever hold your peace. In this way they have the opportunity to inform the team of their



Dealing effectively with uncertainty builds trust in the evaluation framework and helps focus attention on value drivers.

Decision Analysis in Practice, Fig. 4 Tornado chart



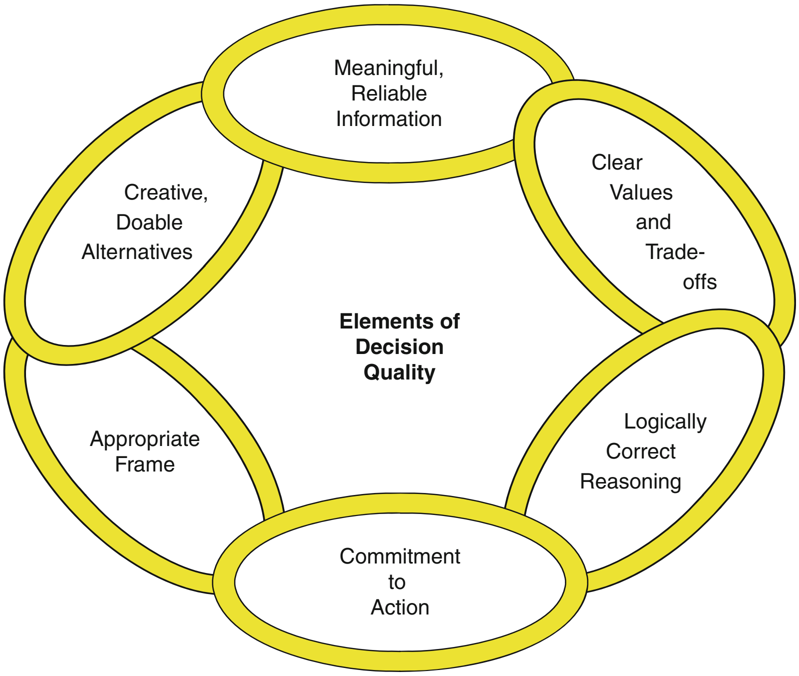
Decision Team
Strategy Team Working Group
0. Design Process
• Frame
• Challenges
• Understanding
• Alternatives
• Improved Information
• Values
Evaluated Alternatives
4. Decide Among Alternatives
Plan
1. Assess Business Situation
2. Develop Alternatives, Information, and Values
3. Evaluate Risk and Return of Alternatives
5. Plan for Action
6. Implement Decision and Manage Transition

Decision Analysis in Practice, Fig. 5 Dialog decision process

important issues, which can be taken into account during the analysis, and they acquire a deeper understanding of the decision situation by participating, giving them a much better chance of ultimately buying in to the conclusions. It gives them needed psychological time and space to reconsider and revise long held convictions. Also, put key implementers on the team so they understand and buy

Decision Analysis in Practice, Fig. 6 Decision quality chain

A high-quality decision produces personal or organizational commitment to the best prospects for creating value.


Meaningful, Reliable Information
Clear Values and Trade-offs
Logically Correct Reasoning
Commitment to Action
Appropriate Frame
Creative, Doable Alternatives
Elements of Decision Quality

These links also specify good design principles for each decision.

into what they are asked to implement. The Dialog Decision Process (Fig. 5) was devised to organize all of these actors into a highly workable project structure.

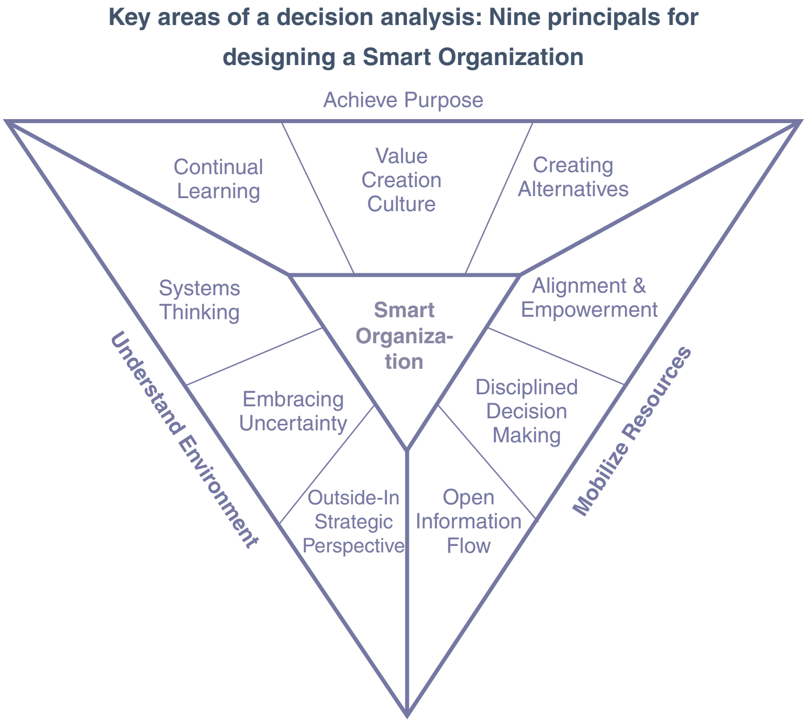
The Decision Quality Chain

The key elements described above are often arranged in a decision quality chain (Fig. 6), originally proposed by Carl Spetzler (Keelin and Spetzler 1992). The metaphor of a chain is used to express that the chain is only as good as its weakest link – that is the most important one; the weakest link changes as the DA proceeds. Decision analysts sometimes use a spider diagram to score progress at each team review (Keelin et al. 2009).

Embedding Good Decision-Making Skills into Organizations

The book, *The Smart Organization*, (Matheson and Matheson 1998), describes "Nine Principles of a Smart Organization" that characterizes a set of habits and a mindset conducive to good decisions, Fig. 7. This book also presents an organizational IQ test to measure compliance with these norms. These tests have been administered to thousands of organizations. The payoff for being a smart organization was striking – organizations in the top quartile of IQ were over five times more likely to be in the top quartile of financial performance, as reported in Matheson and Matheson (2001). Organizations with high scores have patterns of behavior that enable them to spontaneously see the need for decisions, request and frame appropriate decision analyses, and conduct and participate in decision analyses more efficiently and effectively. A few organizations are leading the way by integrating DA into their organizational DNA. Among them, most notably, has been Chevron, which won the annual Decision Analysis Society's Practice Award (2010) for "The implementation of Decision Analysis Practice at Chevron: 20 years of building a DA culture." Matheson and Matheson (2007) discuss how DA principles can become the basis of the Decision Organization.

Decision Analysis in Practice, Fig. 7 Nine principals (Matheson and Matheson 1998)



Concluding Remarks

DA has evolved from specialized high-level consulting to changing culture and embedding processes into organizational routines. The various roles that a DA professional might be called upon to play include:

1. Decision Analyst - responsible for processing numbers,
2. Decision Facilitator - responsible for meetings,
3. Decision Consultant - responsible for attaining commitment,
4. Decision Engineer - responsible for process, systems and organizational design,
5. Decision Change Agent - responsible for personal, organizational, and cultural change necessary for routine, high quality decision making.

See

▶ Decision Analysis
▶ Decision Making and Decision Analysis
▶ Decision Trees
▶ Influence Diagrams

References

Howard, R. A. (1966). Decision analysis: Applied decision theory. In Hertz, D. B., & Melese, J. (Eds.), *Proceedings of the Fourth International Conference on Operational Research* (pp. 55–71).

Howard, R. A. (1980). An assessment of decision analysis. *Operations Research, 28*(1), 4–27.

Howard, R. A., & Matheson, J. E. (Eds.). (1983). *Readings on the principles and applications of decision analysis*. Menlo Park: Strategic Decisions Group.

Keelin, T., Schoemaker, P., & Spetzler, C. (2009). *Decision quality – The fundamentals of making good decisions*. Palo Alto, CA: Decision Education Foundation.

Keelin, T., & Spetzler, C. (1992). *Decision quality: Opportunity for leadership in total quality management*. Palo Alto, CA: Strategic Decision Group.

Matheson, D. (1990). *When should you reexamine your frame?* Ph.D. dissertation, Stanford University.

Matheson, J. (2005). *Decision analysis = decision engineering*, Ch.7 (pp. 195–212). Tutorials in Operations Research INFORMS 2005.

Matheson, D. & Matheson, J. (1998). *The smart organization, creating value through strategic R&D*. Harvard Business School Press.

Matheson, D., & Matheson, J. (2001). *Smart organizations perform better*, *Research-Technology Management*, Industrial Research Institute, July-August.

Matheson, D., & Matheson, J. (2007). From decision analysis to the decision organization. In W. Edwards, R. Miles Jr., & D. von Winterfeldt (Eds.), *Advances in decision analysis:*

From foundations to applications. Cambridge: Cambridge University Press.

McNamee, P., & Celona, J. (2007). *Decision analysis for the professional*. Menlo Park, CA: SmartOrg.

Spetzler, C., & Staël von Holstein, C.-A. (1975). Probability encoding in decision analysis. *Management Science, 22*, 340–358.

Decision Maker (DM)

An individual (or group) who is dissatisfied with some existing situation or with the prospect of a future situation and who possesses the desire and authority to initiate actions designed to alter the situation. In the literature, the letters DM are often used to denote decision maker.

See

- ▶ Decision Analysis
- ▶ Decision Analysis in Practice
- ▶ Decision Making and Decision Analysis

Decision Making and Decision Analysis

Dennis M. Buede
Innovative Decisions, Inc., Vienna, VA, USA

Introduction

Decision making is a process undertaken by an individual or organization. The intent of this process is to improve the future position of the individual or organization, relative to current projections of that future position, in terms of one or more criteria. Most scholars of decision making define this process as one that culminates in an irrevocable allocation of resources to affect some chosen change or the continuance of the status quo. The most commonly allocated resource is money, but other scarce resources are goods and services, and the time and energy of talented people.

Once the concept of making a decision is accepted as a human action, an immediate question is "what is the difference between a good and a bad decision?" The common tendency is to attribute good decisions to situations in which good outcomes were obtained. This approach, however, implies that good decisions cannot be recognized when they are made, but only after the outcomes are observed (which may be seconds or decades later). This common tendency also implies that good decisions have nothing to do with the decision-making process; throwing a dart at a chart of alternatives may lead, on occasion, to good outcomes, while long, hard thought about values and uncertainties does not always yield good outcomes. So leaders in the decision analysis field have defined a good decision as one that is consistent with the values and uncertainties of the decision maker (DM) after considering as many relevant alternatives as possible within the appropriate time frame and with the available information. The belief is that better outcomes will be more likely, but are not guaranteed, with a sound decision making process than throwing darts at a chart of alternatives.

Three primary decision modes have been identified by Watson and Buede (1987): (1) choosing one option from a list, (2) allocating a scarce resource(s) among competing projects, and (3) negotiating an agreement with one or more adversaries. Decision analysis is the common analytical approach for the first mode, optimization using decision analysis concepts of value objectives for the second, and a host of techniques have been applied to negotiation decisions.

The three major elements of a decision that cause decision making to be troublesome are the creative generation of options; the identification and quantification of multiple conflicting criteria, as well as time and risk preference; and the assessment and analysis of uncertainty associated with the causal linkage between alternatives and objectives. To claim to have made a good decision, the DM must be able to defend how these three elements were addressed.

Many DMs claim to be troubled by the feeling that there is an, as yet unidentified, alternative that must surely be better than those so far considered. The development of techniques for identifying such alternatives has received considerable attention (Keller and Ho 1988; Keeney 1992). Additional research has been undertaken to identify the pitfalls in assessing probability distributions that represent the uncertainty of a DM (Edwards et al. 2007). Research has also focused on the identification of the most appropriate preference assessment techniques (Edwards et al. 2007). Keeney (1992) has advanced concepts for the development and structuring of a value hierarchy for key decisions. Very little research has been done on the issue of causal linkages between alternatives and the objectives.

The making of a good decision requires a sound decision making process. However, doing research on competing decision processes, with sound validation using ground truth, is not possible. It is not possible to create multiple versions of reality so that the DM can try the preferred alternative from competing decision processes to identify which would have turned out best. Researchers have proposed multi-phased processes for decision making, e.g., (Howard 1968; Witte 1972; Mintzberg et al. 1976). The common phases include: intelligence or problem definition, design or analysis, choice, and implementation. A weakness in one phase in the decision making process often cannot be compensated for by strengths in the other phases. In general, the decision-making process must address the development of a reasoned set of objectives and associated preference structure; decision alternatives; and the facts, data, opinions, and judgments needed to relate the alternatives to the value model. Then, of course, the logic of evaluating the alternatives in light of the value structure must be sound.

Decision Analysis

The field of decision analysis involves both analysis and synthesis. Analysis is a process for dividing a problem into parts and performing some quantitative assessment of those parts. Synthesis then combines those assessments into a macro assessment. Decision analysis provides an integrating framework for doing this assessment, as well as the theory and techniques for doing the analyses of the parts. These parts are traditionally values (objectives for improving the DM's situation), alternatives (resources the DM can expend to change the world), and the linkage between the alternatives and the values (the facts and uncertainties within the DM's world). Nonetheless, experienced decision analysts often ask the DM for a holistic assessment of the alternatives prior to showing the analysis results (as part of the synthesis process) so that the analysis results can be compared to this holistic standard and the differences noted and examined. Often this comparison to the holistic assessment identifies some issues that were missed in the analysis.

Decision analysis has its roots in many fields. Some of the most obvious are operations research, engineering, business, psychology, probability and statistics, and logic. Fishburn (1999) provides a well-documented summary of these roots of decision analysis. Von Neumann and Morgenstern (1947) provided the first axiomatic structure for decision making, incorporating both probabilistic and value preferences into a principle of expected utility maximization. Savage (1954) recognized the need for subjective probabilities to be combined with subjective utility judgments, leading to subjective expected utility (SEU). Since decision making involves trying to predict how the future world will evolve, the subjectivist approach to uncertainty is the primary perspective adopted in decision analysis. De Finetti (1972) provides a detailed review of the subjectivist approach. Bayes' rule is often required in the computation of expected utility, i.e., Bayesian decision theory is used to describe the decision analysis process (Smith 1988). Interestingly, Bayesian probability and subjectivist probability are used interchangeably. Howard (1966, 1968), Raiffa (1968), and Edwards (1962) all made important contributions in transforming an academic theory into a practical discipline to guide DMs through the difficulties of real world decision making.

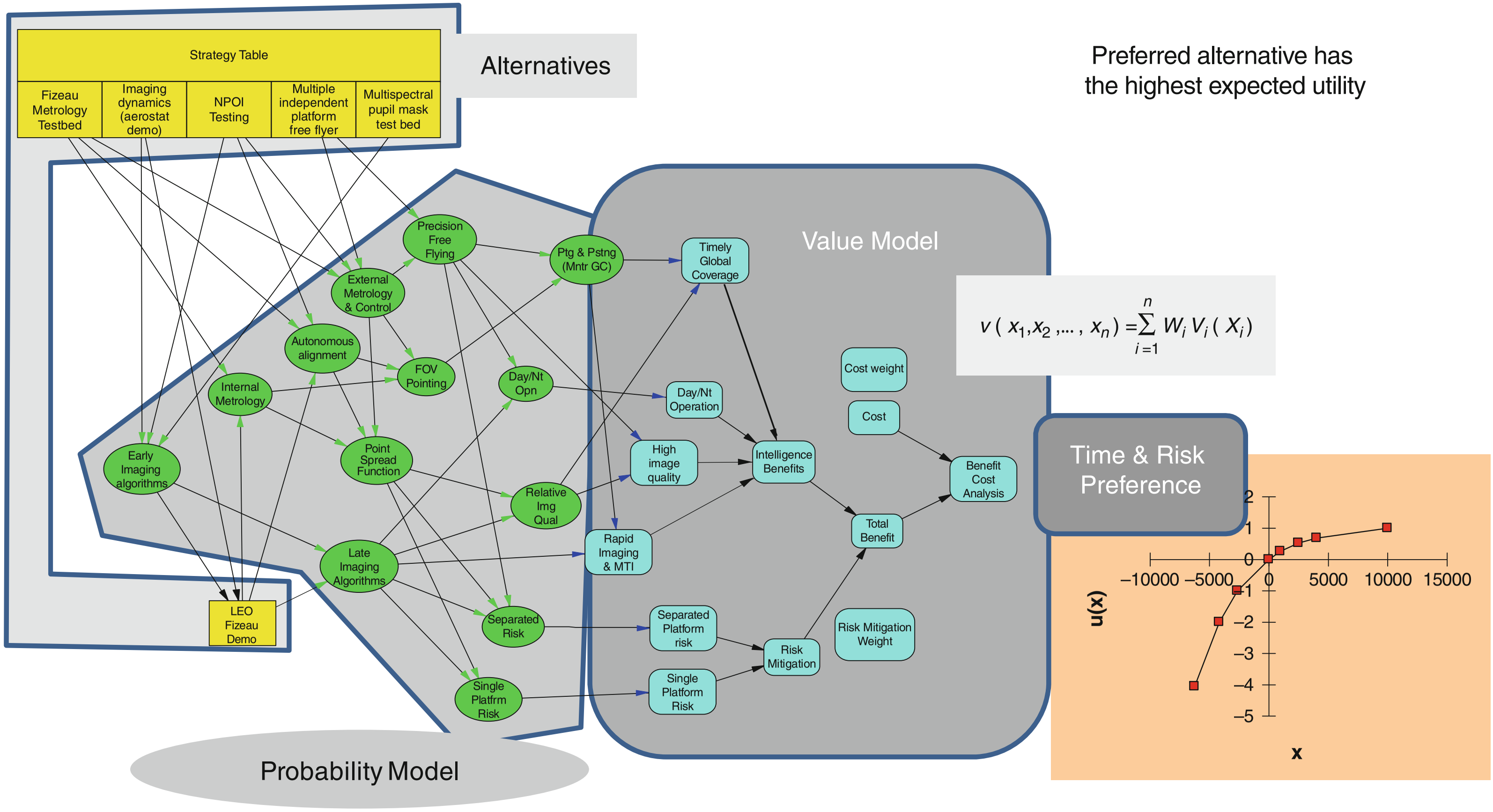
Values represent what the DM wants to improve in the future. As an example, when considering the purchase of a new car, the DM may be weighing reduced cost in the future against improved safety, comfort, prestige, and performance. The context of this decision and, therefore, the values, is the likely uses of a car for commuting, long distance travel, errands, etc. Keeney (1992) provides a structure for thinking about how to separate out the ends (or fundamental) objectives from the means objectives. Several authors have defined the mathematics behind the quantification of a value structure for the analysis of alternatives, see Keeney and Raiffa (1976), French (1986), and Kirkwood (1997). In general, the quantification of preferences must deal with tradeoffs among objectives, risk preference introduced by uncertainty, and time preference introduced by achieving payoffs across the objectives at different points in time. Besides having complex issues to quantify, the DM must deal with subjective judgments, because there can be no source of preference information other than human judgment. Those approaches that attempt to avoid human judgment are throwing the proverbial baby out with the bath water.

Alternatives are the actions (expenditures of resources) that the DM can take now and into the future. In general, the set of alternatives also includes what are termed options or delayed actions that the DM can decide to take in the future if certain events occur between now and the time associated with the option. The space of alternatives is commonly defined over a discrete set, though there is nothing in the theory of decision analysis that precludes a continuous selection set. Various processes have been used to define this set of alternatives, including brainstorming activities. The most commonly discussed approach is called a strategy table or morphological box (Buede 2009). The strategy table divides the alternative space (including any options) into a discrete number of elements or components. For each element, multiple possible selections are defined. The combination of elements and choices within each element are analogous to a buffet dinner during which each diner selects zero, one or more choices from each element and places them onto a plate. If we require each diner to take one and only one selection from each of N elements of the dinner, there are (n_1 x n_2 x . . . x n_N) possible dinners that could be selected. When the choice process is broaden to include no selection or several selections from each element, the number of possible dinners grows. (Note: it is also possible that some of these combinations are impossible or very negatively valued.) Typically, members of the decision-making team are asked to pick five to fifteen representative and interesting selections from the large number of possible selections for the analysis to consider. Often, the evaluation of the initial selection of alternatives from the strategy table will be followed by a second selection of alternatives from the strategy table, with a second round of analysis for this new set. The second set (and possibly a third set) would examine alternatives more like those that did well in the first evaluation and less like those that did poorly.

The linkage between the alternatives and values (both certain and uncertain) is the third element of analytic decomposition of decision analysis. Some parts of this linkage may be well known and deterministic, such as a specific cost of a car, a defined amount of money to purchase. Other parts of this linkage may not be well known, thus requiring the development of a probability distribution; for example, the same car with a known purchase price may not have such a well-known operating cost over the next five to ten years. In some cases, we can develop a probability distribution for this intermediate variable which has a known relationship to a measure for the relevant objective. In other cases, the relationship to one of the objectives may also be probabilistic, requiring the development of an influence diagram with chance nodes separating some or all of the alternatives from the objectives, see Fig. 1.

Once the analytical structure has been built by decomposing the decision problem into such constructs as alternatives, value models, and uncertainties, there is a need to compute (or synthesize) the expected utility of each possible alternative, and to answer additional questions that the DM may have. Examples of common questions are: there is some disagreement about what the risk preference (or time preference or value trade-offs or probabilities) should be, does this make any difference?; alternatives 1 and 2 are much better than the rest, but are very close in expected utility, what are the major differences between these two alternatives?; if one cannot be sure about some parameter's value in the model, will changing it by x% change the order of the alternatives in terms of expected utility? This whole process of computing the results and posing/answering questions regarding the meaning of the analysis and the robustness of the parameters in the analysis is called synthesis. This is exactly why a quantitative model is so much more helpful than a qualitative model. A qualitative model cannot provide these answers without a great deal of fuzziness, leading to continued discussion and argument.

A common criticism of decision analysis is that those involved cannot provide the preference and probabilistic numbers reliably and consistently. Many years of research has demonstrated this conclusively (Edwards et al. 2007; von Winterfeldt and Edwards 1986; Watson and Buede 1987). The real question, however, is not whether humans can provide these judgments accurately, but whether inaccurate judgments for a specified quantitative model leads to a better conversation about the decision being made than does a meandering, fuzzy conversation that starts and stops many times without having such a model or



Decision Making and Decision Analysis, Fig. 1 Representative Influence Diagram any other anchor guiding it. Those who have participated in such meandering, fuzzy conversations have been often left with an empty feeling that there is no real agreement or understanding about the implications of the decision. As long as the key DMs have been involved in the quantitative modeling and understand the results of the synthesis, it is possible to argue that the quantitative analysis, with all of it flaws, has produced useful insights into the decision and provides an accurate audit trail about what was known and not known at the time of decision. The quantitative model is, however, a model and thus subject to the famous quote: "Essentially, all models are wrong, but some are useful" (Box and Draper 1987, p 424).

Decison Analytic Strategies

Many individuals and consulting companies have aided DMs and their organizations to arrive at better decisions. Watson and Buede (1987, pp. 123-159) identified five strategies: (1) modeling, (2) introspection, (3) rating, (4) conferencing, and (5) developing. A sixth strategy that is added here is aggregating mathematically.

1. **Modeling**. The modeling strategy involves building complex representations (models) that link the selection of specific options or alternatives to the values of the DM so that the expected utility across time of each option can be calculated. These models may be decision trees, influence diagrams (Shachter 1986) or simulation models. This approach runs the risk that the DM cannot understand the modeling and, therefore, does not gain the important insights from the model nor trust the results.
2. **Introspection**. The introspection strategy requires deep thought about (i.) the multiple-objective utility function across competing objectives, and (ii.) the joint probability distribution that relates the alternatives to these objectives. This approach is characterized by a question and answer process involving the decision analyst and a single DM (Keeney 1977). This approach does not benefit from additional opinions and expertise beyond the single DM.
3. **Rating**. The rating strategy is the simplest and most used. This strategy typically involves the assumption of an additive value model across multiple objectives, while ignoring time and risk preference, and a deterministic relationship between each alternative, the set of objectives, and their measures. Edwards (1971) introduced this approach under the acronym SMART, but later changed it to SMARTS to reflect the importance of using swing weights rather than importance weights. This approach ignores the complexities of value issues and uncertainty relating the alternatives to the objectives, and uses an ad hoc approach towards gathering information from other participants and experts.
4. **Conferencing**. The conferencing strategy employs simple models as used in Rating with a carefully constructed group (Phillips 2007). The advantage of the simple model is that it is transparent enough to the group to be trusted, and can then focus group discussions across the spectrum of concerns characterized by the objectives, allowing the appropriate experts to weigh in on their topics of expertise. This approach assumes the complexity of the problem is being addressed by the collection of individuals in their reasoning processes, but always runs the risks that the collective reasoning process has interpreted the complexity incorrectly. This alternative reasoning process is difficult to document and scrutinize. Other conferencing approaches exist that utilize computer technology extensively (Nunamaker et al. 1993). These technological approaches to conferencing emphasize giving every participant a chance to enter their inputs via keypads, often limiting discussion. The critical issue is information transfer via open discussion versus group domination by a few individuals. The collective reasoning process is even harder to assess when individuals are communicating via key pads.
5. **Developing.** The developing strategy involves the development of a decision support system that will be used by an individual or collection of individuals for a specific class of decisions over time. This approach usually adopts either a modeling or rating approach to be embed inside the decision support system, along with access to a changing database (see Sauter (1997) for a summary). There continues to be a wide variety of software implementations that serve as a basis for these decision support systems.
6. **Aggregating mathematically**. There are a number of academics and some practitioners who believe a group is best supported by analyzing the decision

from each individual's perspective, and then creating a mathematical aggregation of those individual perspectives. These approaches have been categorized as: social choice theory, group utility analysis, group consensus, and game theory.

See

- ▶ Computational Organization Theory
- ▶ Corporate Strategy
- ▶ Decision Analysis
- ▶ Decision Analysis in Practice
- ▶ Decision Support Systems (DSS)
- ▶ Influence Diagrams
- ▶ Multi-attribute Utility Theory
- ▶ Multiple Criteria Decision Making
- ▶ Simulation of Stochastic Discrete-Event Systems
- ▶ Utility Theory

References

Box, G., & Draper, N. (1987). *Empirical model-building and response surfaces*. New York: John Wiley.

Buede, D. M. (2009). *The engineering design of systems: Models and methods*. New York: John Wiley.

De Finetti, B. (1972). *Probability induction, and statistics: The art of guessing*. New York: John Wiley.

Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Human Factors, 4*, 59–73.

Edwards, W. (1971). Social utilities. *The Engineering Economist, 6*, 119–129.

Edwards, W., Miles, R. F., Jr., & von Winterfeldt, D. (Eds.). (2007). *Advances in decision analysis: From foundations to applications*. New York: Cambridge University Press.

Fishburn, P. (1999). The making of decision theory. In J. Shanteau, B. Mellers, & D. Schum (Eds.), *Decision science and technology: Reflections on the contributions of Ward Edwards* (pp. 369–388). Boston, MA: Kluwer.

French, S. (1986). *Decision theory: An introduction to the mathematics of rationality*. Chichester, UK: John Wiley.

Hammond, F. S., Keeney, R. L., & Raiffa, H. (1999). *Smart choices: A practical guide to making better decisions*. Cambridge, MA: Harvard Business School.

Howard, R. (1966). Decision analysis: Applied decision theory. In Hertz, D.B., & Melese, J. (eds), *Proceedings fourth international conference on operational research*. New York: Wiley-Interscience.

Howard, R. (1968). The foundations of decision analysis. *IEEE Transactions on Systems, Science, and Cybernetics, SSC-4*, 211–219.

Keeney, R. L. (1977). The art of assessing multiattribute utility functions. *Organizational Behavior and Human Performance, 19*, 267–310.

Keeney, R. (1992). *Value-focused thinking*. Boston: Harvard University Press.

Keeney, R. A., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. New York: John Wiley.

Keller, L., & Ho, J. (1988). Decision problem structuring: Generating options. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-15*, 715–728.

Kirkwood, C. W. (1997). *Strategic decision making: Multiple objective decision analysis with spreadsheets*. Belmont, CA: Duxbury Press.

Mintzberg, H., Raisinghani, D., & Theoret, A. (1976). The structure of 'unstructured' decision processes. *Administrative Sciences Quarterly, 21*, 246–275.

Nunamaker, J., Dennis, A., Valacich, J., Vogel, D., & George, J. (1993). Group support systems research: Experience from the lab and field. In L. Jessup & J. Valacich (Eds.), *Group support systems*. New York: Macmillan.

Phillips, L. D. (2007). Decision conferencing. In W. Edwards et al. (Eds.), *Advances in decision analysis*. New York: Cambridge University Press.

Raiffa, H. (1968). *Decision analysis: Introductory lectures on choices under uncertainty*. Reading, MA: Addison-Wesley.

Sauter, V. L. (1997). *Decision support systems: An applied managerial approach*. New York: John Wiley.

Savage, L. J. (1954). *The foundations of statistics*. New York: John Wiley.

Shachter, R. D. (1986). Evaluating influence diagrams. *Operations Research, 34*, 871–882.

Smith, J. Q. (1988). *Decision analysis: A Bayesian approach*. London: Chapman and Hall.

von Neumann, J., & Morgenstern, O. (1947). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. New York: Cambridge University Press.

Watson, S., & Buede, D. (1987). *Decision synthesis: The principles and practice of decision analysis*. Chichester, UK: Cambridge University Press.

Witte, E. (1972). Field research on complex decision-making processes–The phase theorem. *International Studies of Management and Organization*, 156–182.

Decision Problem

The basic decision problem is as follows: Given a set of r alternative actions $A = \{a_1,\ldots, a_r\}$, a set of q states of nature $S = \{s_1,\ldots, s_q\}$, a set of rq outcomes $O = \{o_1, \ldots, o_{rq}\}$, a corresponding set of rq payoffs $P = \{p_1,\ldots, p_{rq}\}$, and a decision criterion to be optimized, $f(a_j)$, where f is a real-valued function defined on A, choose an alternative action a_j that optimizes the decision criterion $f(a_j)$.

See

- ▶ Decision Analysis
- ▶ Decision Analysis in Practice
- ▶ Decision Making and Decision Analysis
- ▶ Group Decision Making
- ▶ Multi-Criteria Decision Making (MCDM)
- ▶ Utility Theory

Decision Support Systems (DSS)

Andrew Vazsonyi
University of San Francisco, San Francisco, CA, USA

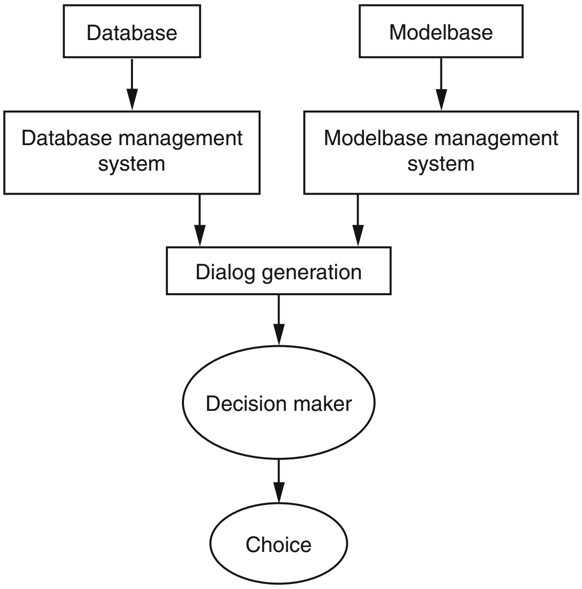
Introduction

Throughout history there has been a deeply embedded conviction that, under the proper conditions, some people are capable of helping others come to grips with problems in daily life. Such professional helpers are called counselors, psychiatrists, psychologists, social workers, and the like. In addition to these professional helpers, there are less formal helpers, such as ministers, lawyers, teachers, or even bartenders, hairdressers, and cab drivers.

The proposition that science and quantitative methods, such as those used in OR/MS, may help people is relatively new, and is still received by many with deep skepticism. There are some disciplines overlapping and augmenting OR/MS. One important one is called decision support systems (DSS).

Before discussion of DSS, it is to be stressed that the expression is used in a different manner by different people, and there is no general agreement of what DSS really is. Moreover, the benefits claimed by DSS are in no way different from the benefits claimed by OR/MS. To appreciate DSS, a pluralistic view must be taken of the various disciplines offered to help managerial decision making.

Features of Decision Support Systems



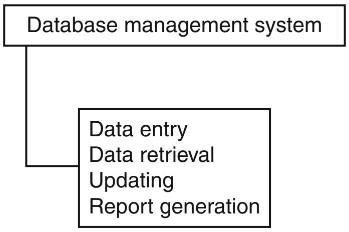
Decision Support Systems (DSS), Fig. 1 Components of a DSS

During the early 1970s, under the impact of new developments in computer systems, a new perspective about decision making appeared. Keen and Morton (1973) coined the expression decision support systems, to designate their approach to the solution of managerial problems. They postulated a number of distinctive characteristics of DSS, especially the five listed below:

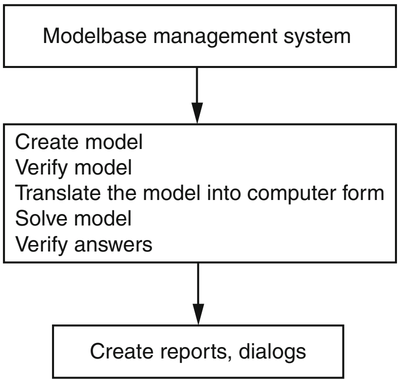
- A DSS is designed for specific decision makers and their decision tasks,
- A DSS is developed by cycling between design and implementation,
- A DSS is developed with a high degree of user involvement,
- A DSS includes both data and models, and
- Design of the user-machine interface is a critical task in the development of a DSS.

Figure 1 shows the structure and major components of a DSS. The database holds all the relevant facts of the problem, whether they pertain to the firm or to the environment. The database management system (Fig. 2) takes care of the entry, retrieval, updating, and deletion of data. It also responds to inquiries and generates reports.

The modelbase holds all the models required to work the problem. The modelbase management system (Fig. 3) assists in creating the mathematical model, and in translating the human prepared mathematical model into computer understandable form. The critical process of the modelbase management system is finding the solution to the mathematical model. The system also generates



Decision Support Systems (DSS), Fig. 2 Database management system



Decision Support Systems (DSS), Fig. 3 Modelbase management system

reports and assists in the preparation of computer-human dialogs.

While OR/MS stresses the model, DSS stresses the computer-based database. DSS emphasizes the importance of the user-machine interface, and the design of dialog generation and management software.

Advocates of DSS assert that by combining the power of the human mind and the computer, DSS is capable of enhancing decision making, and that DSS can grapple with problems not subject to the traditional approach of OR/MS.

Note that DSS stresses the role of humans in decision making, and explicitly factors human capabilities into decision making. A decision support system accepts the human as an essential subsystem. DSS does not usually try to optimize in a mathematical sense, and bounded rationality and satisficing provide guidance to the designers of DSS.

Designing Decision Support Systems

The design phases of DSS are quite similar to the phases of the design, implementation, and testing of other systems. It is customary to distinguish six phases, although not all six phases are required for every DSS.

1. During the systems analysis and design phase, existing systems are reviewed and analyzed with the objective of establishing requirements and needs of the new system. Then it is established whether meeting the specifications is feasible from the technical, economical, psychological, and social points of view. Is it possible to overcome the difficulties, and are opportunities commensurate with costs? If the answers are affirmative, meetings with management are held to obtain support. This phase produces a conceptual design and master plan.
2. During the design phase, input, processing, and output requirements are developed and a logical (not physical) design of the system is prepared. After the logical design is completed and found to be acceptable, the design of the hardware and software is undertaken.
3. During the construction and testing phase, the software is completed and tested on the hardware system. Testing includes user participation to assure that the system will be acceptable both from the points of view of the user and management, if they are different.
4. During the implementation phase, the system is retested, debugged, and put into use. To assure final user acceptance, no effort is spared in training and educating users. Management is kept up-to-date on the progress of the project.
5. Operation and maintenance is a continued effort during the life of the DSS. User satisfaction is monitored, errors are uncovered and corrected, and the method of operating the system is fine-tuned.
6. Evaluation and control is a continued effort to assure the viability of the system and the maintenance of management support.

A Forecasting System

Connoisseur Foods is a diversified food company with several autonomous subdivisions and subsidiaries (adapted from Alter 1980, and Turban 1990). Several of the division managers were old-line managers relying on experience and judgment to make major decisions. Top management installed a DSS to provide quantitative help to establish and monitor levels of such marketing efforts as advertising, pricing, and promotion. The DSS model was based on S-shaped response functions of marketing conditions to such decision functions as advertising. The curves were derived by using both historical data and marketing experts. The databases for the farm products division contained about 20 million data items on sales both in dollars and number of units for 400 items sold in 300 branches.

The DSS assisted management in developing better marketing strategies and more competitive positions. Top management, however, stated that the real benefit of the DSS was not so much the installation of isolated systems and models, but the assimilation of new approaches in corporate decision making.

A Portfolio Management System

The trust division of Great Eastern Bank employed 50 portfolio managers in several departments (adapted from Alter 1980 and Turban 1990). The portfolio managers controlled many small accounts, large pension funds, and provided advice to investors in large accounts. The on-line DSS portfolio management system provided information to the portfolio managers.

The DSS includes lists of stocks from which the portfolio managers could buy stocks, information, and analysis on particular industries. It is basically a data retrieval system that could display portfolios, as well as specific information on securities.

The heart of the system is the database that allowed portfolio managers to generate reports with the following functions:

- Directory by accounts,
- Table to scan accounts,
- Graphic display of breakdown by industry and security for an account,
- Tabular listing of all securities within an account,
- Scatter diagrams between data items,
- Summaries of accounts,
- Distribution of data on securities,
- Evaluation of hypothetical portfolios,
- Performance monitoring of portfolios,
- Warnings if deviations from guidelines occur; and
- Tax implications.

The benefits of the systems were better investment performance, improved information, improved presentation formats, less clerical work, better communication, improved bank image, and enhanced marketing capability.

Concluding Remarks

Advocates of DSS claim that DSS deals with unstructured or semistructured problems, while OR/MS is restricted to structured problems. Few workers in OR/MS would agree.

At the onset, it is frequently the case that a particular business situation is confusing, and, to straighten it out, a problem must be instituted and the problem must be structured. Thus, whether OR/ MS or DSS or both are involved, attempts will be made to structure as much of the situation as possible.

The problem will be structured by OR/MS or DSS to the point that some part of the problem can be taken care of by quantitative methods and computers, and some others are left to human judgment, intuition, and opinion. There may be a degree of difference between OR/MS and DSS: OR/MS may stress optimization, the model base; DSS the database.

Attempts to draw the line between DSS and OR/MS are counterproductive. Those who are dedicated to help management in solving hard problems need to be concerned with any and all theories, practices, and principles that can help. To counsel management in the most productive manner requires that no holds be barred when a task is undertaken.

The principles of DSS are often used without mention in simulation programs. Moreover, as in the spirit of DSS, the user-machine interface is often visual, given the animation capability of modern computers. Thus, managerial decisions may be influenced not only by using traditional quantitative measures, but also by judging customer perceptions.

See

- ▶ Bounded Rationality
- ▶ Choice Theory
- ▶ Decision Analysis
- ▶ Decision Analysis in Practice
- ▶ Decision Problem
- ▶ Information Systems and Database Design in OR/MS
- ▶ Satisficing
- ▶ Soft Systems Methodology

References

Alter, S. L. (1980). *Decision support systems: Current practice and continuing challenges*. Reading, MA: Addison-Wesley.

Bennett, J. L. (1983). *Building decision support systems*. Reading, MA: Addison-Wesley.

Burstein, F., & Holsapple, C. (2008). *Handbook on decision support systems 2: Variations*. New York: Springer.

Holsapple, C., & Whinston, A. (1996). *Decision support systesms: A knowledge-based approach*. Eagan, MN: West Publishing.

Keen, P. G. W., & Morton, S. (1973). *Decision support systems*. Reading, MA: Addison-Wesley.

Pritsker, A. A. B. (1996). Life & death decisions. *OR/ MS Today, 25*(4), 22–28.

Simon, H. A. (1992). Methods and bounds of economics. In Praxiologies and the philosophy of economics. New Brunswick and London: Transaction Publishers.

Turban, E. (1990). *Decision support and expert systems* (2nd ed.). New York: Macmillan.

Decision Trees

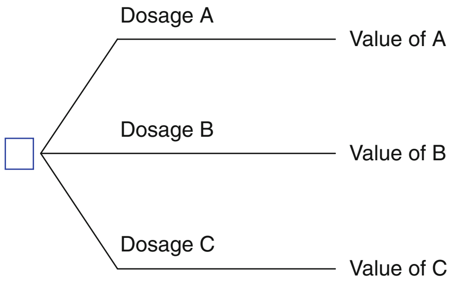
Stuart Eriksen[1], Candice H. Huynh[2] and L. Robin Keller[2]
[1]Santa Ana, CA, USA
[2]University of California, Irvine, CA, USA

Introduction

A decision tree is a pictorial description of a well-defined decision problem. It is a graphical representation consisting of nodes (where decisions are made or chance events occur) and arcs (which connect nodes). Decision trees are useful because they provide a clear, documentable, and discussible model of either how the decision was made or how it will be made.

The tree provides a framework for the calculation of the expected value of each available alternative. The alternative with the maximum expected value is the best choice path based on the information and mind-set of the decision makers at the time the decision is made. This best choice path indicates the best overall alternative, including the best subsidiary decisions at future decision steps, when uncertainties have been resolved.

The decision tree should be arranged, for convenience, from left to right in the temporal order in which the events and decisions will occur. Therefore, the steps on the left occur earlier in time than those on the right.



Decision Trees, Fig. 1 The choice of drug dosage

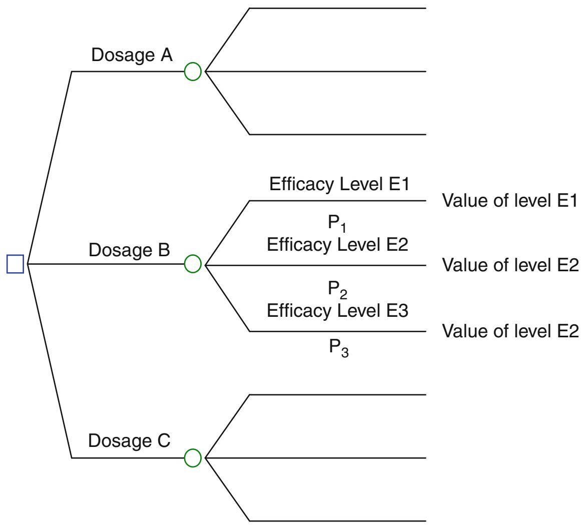
Decision Nodes

Steps in the decision process involving decisions between several choice alternatives are indicated by decision nodes, drawn as square boxes. Each available choice is shown as one arc (or path) leading away from its decision node toward the right. When a planned decision has been made at such a node, the result of that decision is recorded by drawing an arrow in the box pointing toward the chosen option. As an example of the process, consider a pharmaceutical company president's choice of which drug dosage to market. The basic dosage choice decision tree is shown in Fig. 1. Note that the values of the eventual outcomes (on the far right) will be expressed as some measure of value to the eventual user (for example, the patient or the physician).

Chance Nodes

Steps in the process which involve uncertainties are indicated by circles (called chance nodes), and the possible outcomes of these probabilistic events are again shown as arcs or paths leading away from the node toward the right. The results of these uncertain factors are out of the hands of the decision maker; chance or some other group or person (uncontrolled by the decision maker) will determine the outcome of this node. Each of the potential outcomes of a chance node is labeled with its probability of occurrence.

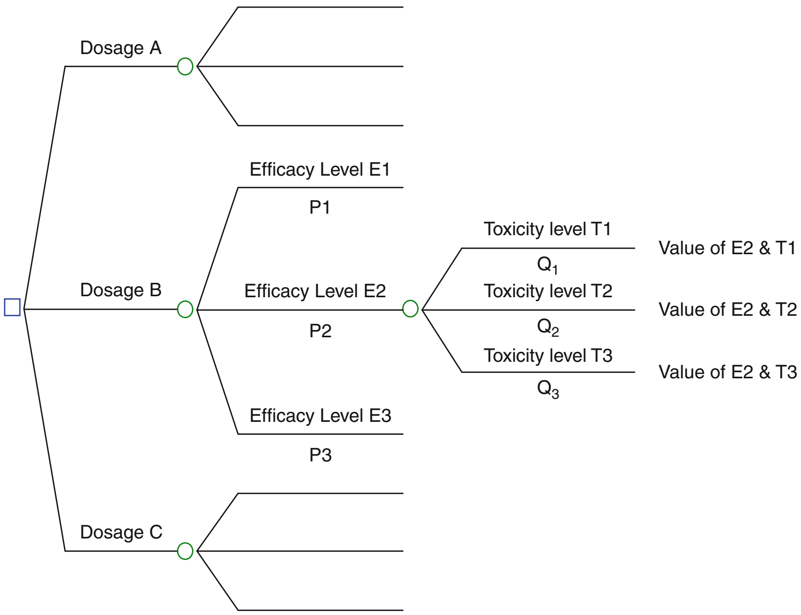
All possible outcomes must be indicated, so the sum of the potential outcome probabilities of a chance node must equal 1.0. Using the drug dose selection problem noted above, the best choice of dose depends on at least one probabilistic event: the level of performance of the drug in clinical trials, which is a proxy measure of the efficacy of the drug. A simplified decision tree for that part of the firm's decision is shown in Fig. 2. Note that each dosage choice has a subsequent efficacy chance node similar to the one shown, so the expanded tree would have nine outcomes. The probabilities (p_1, p_2, and p_3) associated with the outcomes are expected to differ for each dosage.



Decision Trees, Fig. 2 The choice of drug dosage based on efficacy outcome

There are often several nodes in a decision tree; in the case of the drug dosage decision, the decision will also depend on the toxicity as demonstrated by both animal study data and human toxicity study data, as well as on the efficacy data. The basic structure of this more complex decision is shown in Fig. 3. The completely expanded tree has 27 eventual outcomes and associated values. Notice that although not always the case, here the probabilities (q_1, q_2, and q_3) of the toxicity levels are independent of the efficacy level.

One use of a decision tree is to clearly display the factors and assumptions involved in a decision. If the decision outcomes are quantified and the probabilities of chance events are specified, the tree can also be analyzed by calculating the expected value of each alternative. If several decisions are involved in the problem being considered, the strategy best suited to each specific set of chance outcomes can be planned in advance.



Decision Trees, Fig. 3 The choice of dosage based on uncertain efficacy and toxicity

Probabilities

Estimates of the probabilities for each of the outcomes of the chance nodes must be made. In the simplified case of the drug dose decision above, the later chance node outcome probabilities are modeled as being independent of the earlier chance nodes. While not intuitively obvious, careful thought should show that the physiological factors involved in clinical efficacy must be different from those involved in toxicity, even if the drug is being used to treat that toxicity. Therefore, with most drugs, the probability of high human toxicity is likely independent of the level of human efficacy. In the more general non-drug situations, however, for sequential steps, the latter probabilities are often dependent conditional probabilities, since their value depends on the earlier chance outcomes.

For example, consider the problem in Fig. 4, where the outcome being used for the drug dose decision is based on the eventual sales of it. The values of the eventual outcomes now are expressed as sales for the firm.

The probability of high sales depends on the efficacy as well as on the toxicity, so the dependent conditional probability of high sales is the probability of high sales given that the efficacy is level 2 and toxicity is level 2, which can be written as $p_{(\mathrm{High\ Sales}|\mathrm{E2\&T2})}$.

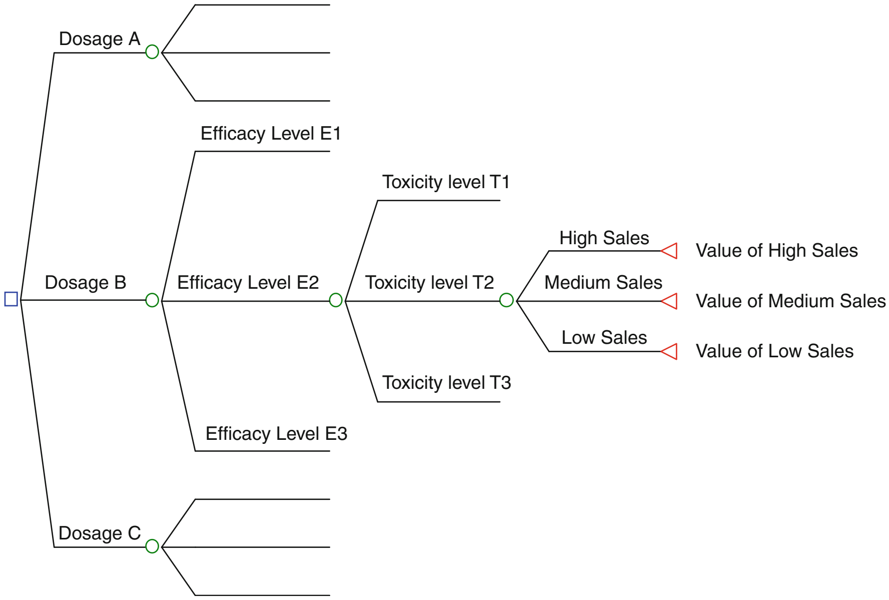
Outcome Measures

At the far right of the tree, the possible outcomes are listed at the end of each branch. To calculate numerical expected values for alternative choices, outcomes must be measured numerically and often monetary measures will be used. More generally, the utility of the outcomes can be calculated. Single or multiple attribute utility functions have been elicited in many decision situations to represent decision makers' preferences for different outcomes on a numerical (although not monetary) scale.

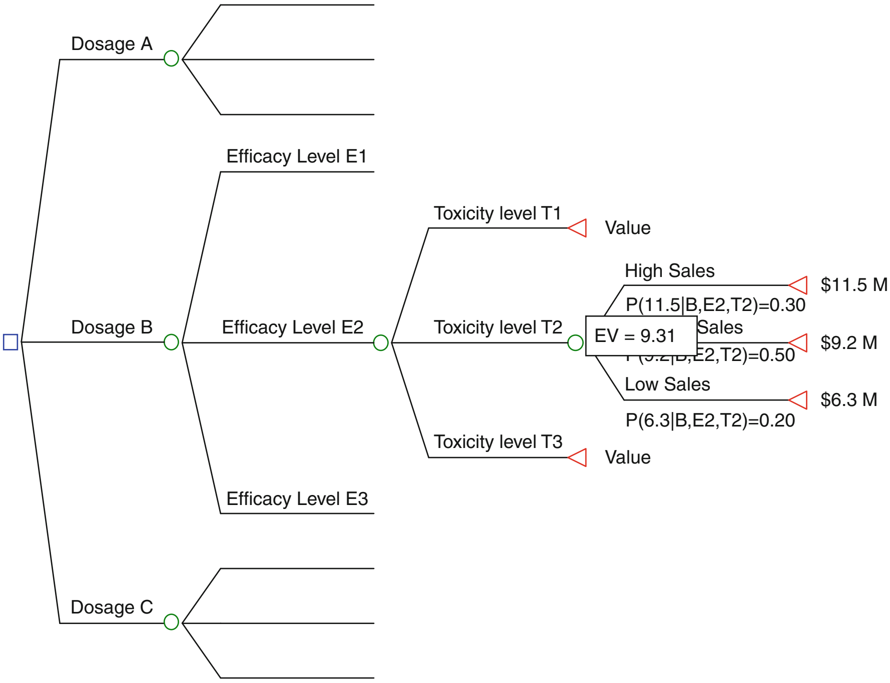
The Tree as an Aid in Decision Making

The decision tree analysis method is called fold-back and prune. Beginning at a far right chance node of the tree, the expected value of the outcome measure is calculated and recorded for each chance node by summing, over all the outcomes, the product of the probability of the outcome times the measured value of the outcome. Figure 5 shows this calculation for the first step in the analysis of the drug-dose decision tree.

This step is called folding back the tree since the branches emanating from the chance node are folded

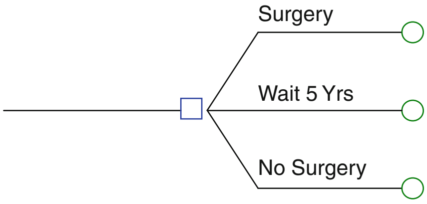


Decision Trees, Fig. 4 The choice of dosage based on efficacy and toxicity and their eventual effect on sales



Decision Trees, Fig. 5 The first step, calculating the expected value of the chance node for sales: EV = 0.3(11.5) + 0.5(9.2) + 0.2(6.3) = 9.31

up or collapsed, so that the chance node is now represented by its expected value. This is continued until all the chance nodes on the far right have been evaluated. These expected values then become the values for the outcomes of the chance or decision nodes further to the left in the diagram. At a decision node, the best of the alternatives is the one with the maximum expected value, which is then recorded by drawing an arrow towards that choice in the decision node box and writing down the expected value associated with the chosen option. This is referred to as pruning the tree, as the less valuable choices are eliminated from further consideration. The process continues from right to left, by calculating the expected value at each chance node and pruning at each decision node. Finally the best choice for the overall decision is found when the last decision node at the far left has been evaluated.

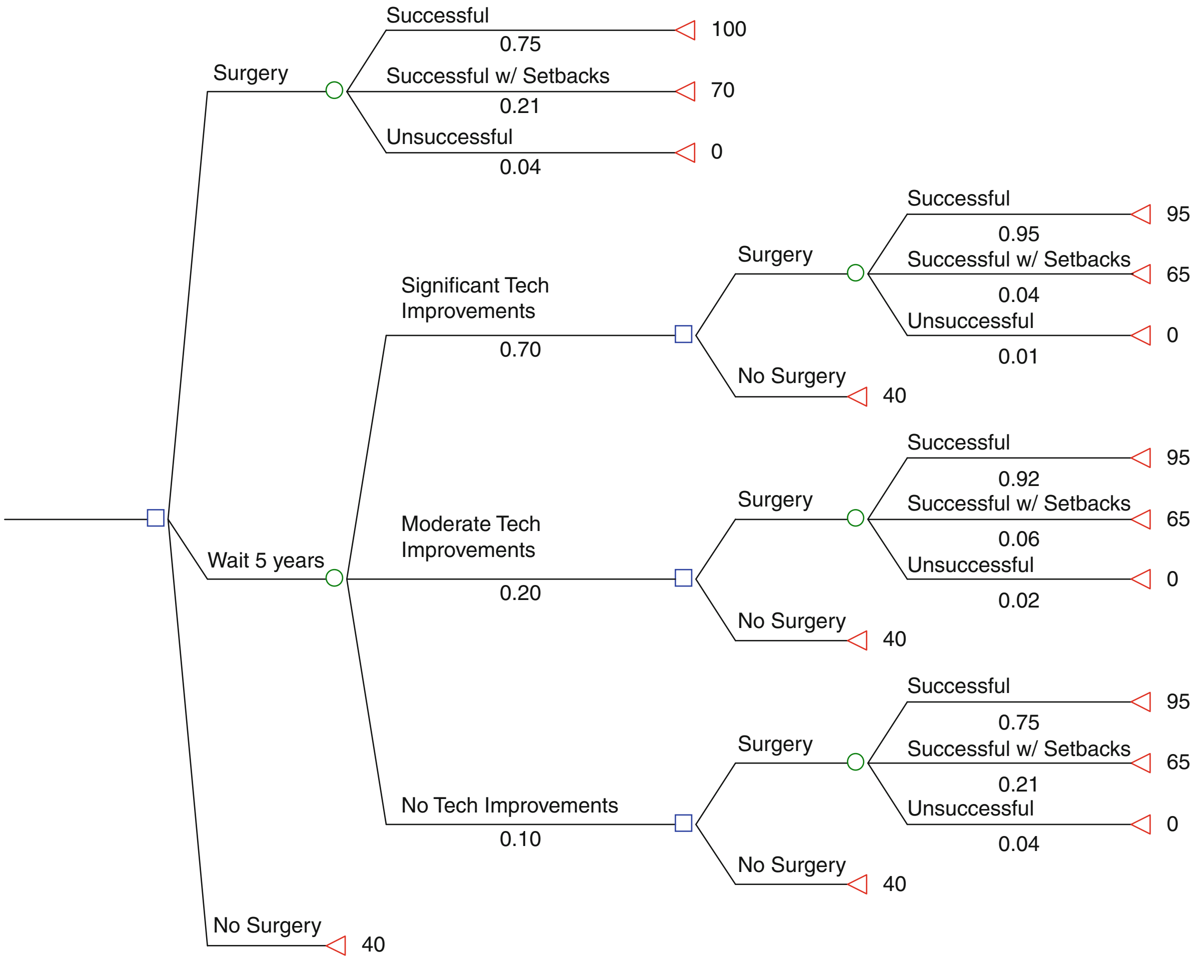


Decision Trees, Fig. 6 The initial decision point

Example

In this example, a decision faced by a patient who is considering laser eye surgery to improve her vision will be considered. The basic decision process is shown in Fig. 6. The initial decision a patient encounters is whether to: have the surgery, wait for more technological advances, or not have the surgery at all.

Suppose that if a patient chooses to wait at the first decision node, she will observe the outcome of possible technological advances at the first chance node, and then will have to make the decision of whether to have the surgery or not. Figure 7 shows a detailed decision tree of this patient's decision process. The entries at the end of the branches can be seen as a measure of health utility to the patient, on a 0-100 scale, where 100 is the best level of health utility. Other patients can customize this tree to their personal circumstances using a combination of chance and decision nodes.



Decision Trees, Fig. 7 Complete mapping of the decision process of whether or not to have lasik surgery

Following the method of folding back the tree, the expected health utility of having the surgery immediately is 89.70, waiting 5 years is 91.74, and not having the surgery at all is 40.00, where the calculation of each chance node is the expected health utility. And so waiting 5 years is the optimal decision for the patient in this example.

See

- ▶ Bayesian Decision Theory, Subjective Probability, and Utility
- ▶ Decision Analysis
- ▶ Decision Analysis in Practice
- ▶ Decision Making and Decision Analysis
- ▶ Multi-attribute Utility Theory
- ▶ Preference Theory
- ▶ Utility Theory

References

Clemen, R., & Reilly, T. (2004). *Making hard decisions with decision tools*. Belmont, CA: Duxbury Press.

Eriksen, S. P., & Keller, L. R. (1993). A multi-attribute approach to weighing the risks and benefits of pharmaceutical agents. *Medical Decision Making, 13*, 118–125.

Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value tradeoffs*. Wiley, New York.

Kirkwood, C. (1997). *Strategic decision making: Multiobjective decision analysis with spreadsheets*. Belmont, CA: Duxbury Press.

Raiffa, H. (1968). *Decision analysis*. Reading, MA: Addison-Wesley.

Decision Variables

The variables in a given model that are subject to manipulation by the specified decision rule.

See

- ▶ Controllable Variables

Decomposition Algorithms

- ▶ Benders Decomposition Method
- ▶ Block-Angular System
- ▶ Dantzig-Wolfe Decomposition Algorithm
- ▶ Large-Scale Systems

References

Dantzig, G. B., & Thapa, M. N. (2003). *Linear programming 2: Theory and extensions*. New York: Springer.

Deep Uncertainty

Warren E. Walker[1], Robert J. Lempert[2] and
Jan H. Kwakkel[1]
[1]Delft University of Technology, Delft,
The Netherlands
[2]RAND Corporation, Santa Monica, CA, USA

Introduction

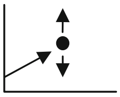
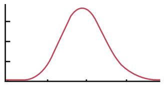
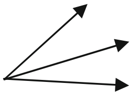
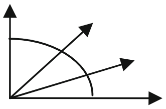
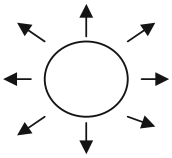
The notion of uncertainty has taken different meanings and emphases in various fields, including the physical sciences, engineering, statistics, economics, finance, insurance, philosophy, and psychology. Analyzing the notion in each discipline can provide a specific historical context and scope in terms of problem domain, relevant theory, methods, and tools for handling uncertainty. Such analyses are given by Agusdinata (2008), van Asselt (2000), Morgan and Henrion (1990), and Smithson (1989).

In general, uncertainty can be defined as limited knowledge about future, past, or current events. With respect to policy making, the extent of uncertainty clearly involves subjectivity, since it is related to the satisfaction with existing knowledge, which is colored by the underlying values and perspectives of the policymaker and the various actors involved in the policy-making process, and the decision options available to them.

Shannon (1948) formalized the relationship between the uncertainty about an event and information in "A Mathematical Theory of Communication." He defined a concept he called entropy as a measure of the average information content associated with a random outcome. Roughly speaking, the concept of entropy in information theory describes how much information there is in a signal or event and relates this to the degree of uncertainty about a given event having some probability distribution.

Uncertainty is not simply the absence of knowledge. Funtowicz and Ravetz (1990) describe uncertainty as a situation of inadequate information, which can be of three sorts: inexactness, unreliability, and border with ignorance. However, uncertainty can prevail in situations in which ample information is available (Van Asselt and Rotmans 2002). Furthermore, new information can either decrease or increase uncertainty. New knowledge on complex processes may reveal the presence of uncertainties that were previously unknown or were understated. In this way, more knowledge illuminates that one's understanding is more limited or that the processes are more complex than previously thought (van der Sluijs 1997).

Uncertainty as inadequacy of knowledge has a very long history, dating back to philosophical questions debated among the ancient Greeks about the certainty of knowledge, and perhaps even further. Its modern history begins around 1921, when Knight made a distinction between risk and uncertainty (Knight 1921). According to Knight, risk denotes the calculable and thus controllable part of all that is unknowable. The remainder is the uncertain – incalculable and uncontrollable. Luce and Raiffa (1957) adopted these labels to distinguish between decision making under risk and decision making under uncertainty. Similarly, Quade (1989) makes a distinction between stochastic uncertainty and real uncertainty. According to Quade, stochastic

			LEVEL					
			Level 1	Level 2	Level 3	Level 4	Level 5	
LOCATION	Context	Complete Certainty	A clear enough future 	Alternate futures (with probabilities) 	Alternate futures with ranking 	A multiplicity of plausible futures 	An unknown future 	Total ignorance
	System model		A single (deterministic) system model	A single (stochastic) system model	Several system models, one of which is most likely	Several system models, with different structures	Unknown system model; know we don't know	
	System outcomes		A point estimate for each outcome	A confidence interval for each outcome	Several sets of point estimates, ranked according to their perceived likelihood	A known range of outcomes	Unknown outcomes; know we don't know	
	Weights on outcomes		A single set of weights	Several sets of weights, with a probability attached to each set	Several sets of weights, ranked according to their perceived likelihood	A known range of weights	Unknown weights; know we don't know	

Deep Uncertainty, Fig. 1 The progressive transition of levels of uncertainty from complete certainty to total ignorance

uncertainty includes frequency-based probabilities and subjective (Bayesian) probabilities. Real uncertainty covers the future state of the world and the uncertainty resulting from the strategic behavior of other actors. Often, attempts to express the degree of certainty and uncertainty have been linked to whether or not to use probabilities, as exemplified by Morgan and Henrion (1990), who make a distinction between uncertainties that can be treated through probabilities and uncertainties that cannot. Uncertainties that cannot be treated probabilistically include model structure uncertainty and situations in which experts cannot agree upon the probabilities. These are the more important and hardest to handle types of uncertainties (Morgan 2003). As Quade (1989, p. 160) wrote: "Stochastic uncertainties are therefore among the least of our worries; their effects are swamped by uncertainties about the state of the world and human factors for which we know absolutely nothing about probability distributions and little more about the possible outcomes." These kinds of uncertainties are now referred to as deep uncertainty (Lempert et al. 2003), or severe uncertainty (Ben-Haim 2006).

Levels of Uncertainty

Walker et al. (2003) define uncertainty to be "any departure from the (unachievable) ideal of complete determinism."

For purposes of determining ways of dealing with uncertainty in developing public policies or business strategies, one can distinguish two extreme levels of uncertainty—complete certainty and total ignorance—and five intermediate levels (e.g. Courtney 2001; Walker et al. 2003; Makridakis et al. 2009; Kwakkel et al. 2010d). In Fig. 1, the five levels are defined with respect to the knowledge assumed about the various aspects of a policy problem: (a) the future world, (b) the model of the relevant system for that future world, (c) the outcomes from the system, and (d) the weights that the various stakeholders will put on the outcomes. The levels of uncertainty are briefly discussed below.

Complete certainty is the situation in which everything is known precisely. It is not attainable, but acts as a limiting characteristic at one end of the spectrum.

Level 1 uncertainty (A clear enough future) represents the situation in which one admits that one is not absolutely certain, but one is not willing or able to measure the degree of uncertainty in any explicit way (Hillier and Lieberman 2001, p. 43). Level 1 uncertainty is often treated through a simple sensitivity analysis of model parameters, where the impacts of small perturbations of model input parameters on the outcomes of a model are assessed.

Level 2 uncertainty (Alternate futures with probabilities) is any uncertainty that can be described adequately in statistical terms. In the case of uncertainty about the future, Level 2 uncertainty is often captured in the form of either a (single) forecast (usually trend based) with a confidence interval or multiple forecasts (scenarios) with associated probabilities.

Level 3 uncertainty (Alternate futures with ranking) represents the situation in which one is able to enumerate multiple alternatives and is able to rank the alternatives in terms of perceived likelihood. That is, in light of the available knowledge and information there are several different parameterizations of the system model, alternative sets of outcomes, and/or different conceivable sets of weights. These possibilities can be ranked according to their perceived likelihood (e.g. virtually certain, very likely, likely, etc.). In the case of uncertainty about the future, Level 3 uncertainty about the future world is often captured in the form of a few trend-based scenarios based on alternative assumptions about the driving forces (e.g., three trend-based scenarios for air transport demand, based on three different assumptions about GDP growth). The scenarios are then ranked according to their perceived likelihood, but no probabilities are assigned, see Patt and Schrag (2003) and Patt and Dessai (2004).

Level 4 uncertainty (Multiplicity of futures) represents the situation in which one is able to enumerate multiple plausible alternatives without being able to rank the alternatives in terms of perceived likelihood. This inability can be due to a lack of knowledge or data about the mechanism or functional relationships being studied; but this inability can also arise due to the fact that the decision makers cannot agree on the rankings. As a result, analysts struggle to specify the appropriate models to describe interactions among the system's variables, to select the probability distributions to represent uncertainty about key parameters in the models, and/or how to value the desirability of alternative outcomes (Lempert et al. 2003).

Level 5 uncertainty (Unknown future) represents the deepest level of recognized uncertainty; in this case, what is known is only that we do not know. This ignorance is recognized. Recognized ignorance is increasingly becoming a common feature of life, because catastrophic, unpredicted, surprising, but painful events seem to be occurring more often. Taleb (2007) calls these events "Black Swans." He defines a Black Swan event as one that lies outside the realm of regular expectations (i.e., "nothing in the past can convincingly point to its possibility"), carries an extreme impact, and is explainable only after the fact (i.e., through retrospective, not prospective, predictability). One of the most dramatic recent Black Swans is the concatenation of events following the 2007 subprime mortgage crisis in the U.S. The mortgage crisis (which some had forecast) led to a credit crunch, which led to bank failures, which led to a deep global recession in 2009, which was outside the realm of most expectations. Another recent Black Swan was the level 9.0 earthquake in Japan in 2011, which led to a tsunami and a nuclear catastrophe, which led to supply chain disruptions (e.g., for automobile parts) around the world.

Total ignorance is the other extreme on the scale of uncertainty. As with complete certainty, total ignorance acts as a limiting case.

Lempert et al. (2003) have defined deep uncertainty as "the condition in which analysts do not know or the parties to a decision cannot agree upon (1) the appropriate models to describe interactions among a system's variables, (2) the probability distributions to represent uncertainty about key parameters in the models, and/or (3) how to value the desirability of alternative outcomes. They use the language 'do not know' and 'do not agree upon' to refer to individual and group decision making, respectively. This article includes both individual and group decision making in all five of the levels, referring to Level 4 and Level 5 uncertainties as 'deep uncertainty', and assigning the 'do not know' portion of the definition to Level 5 uncertainties and the 'cannot agree upon' portion of the definition to Level 4 uncertainties.

Decision Making Under Deep Uncertainty

There are many quantitative analytical approaches to deal with Level 1, Level 2, and Level 3 uncertainties. In fact, most of the traditional applied scientific work in the engineering, social, and natural sciences has been built upon the supposition that the uncertainties result from either a lack of information, which "has led to an emphasis on uncertainty reduction through ever-increasing information seeking and processing" (McDaniel and Driebe 2005), or from random variation, which has concentrated efforts on stochastic processes and statistical analysis. However, most of the important policy problems faced by policymakers are characterized by the higher levels of uncertainty, which cannot be dealt with through the use of probabilities and cannot be reduced by gathering more information, but are basically unknowable and unpredictable at the present time. And these high levels of uncertainty can involve uncertainties about all aspects of a policy problem — external or internal developments, the appropriate (future) system model, the parameterization of the model, the model outcomes, and the valuation of the outcomes by (future) stakeholders.

For centuries, people have used many methods to grapple with the uncertainty shrouding the long-term future, each with its own particular strengths. Literary narratives, generally created by one or a few individuals, have an unparalleled ability to capture people's imagination. More recently, group processes, such as the Delphi technique (Quade 1989), have helped large groups of experts combine their expertise into narratives of the future. Statistical and computer simulation modeling helps capture quantitative information about the extrapolation of current trends and the implications of new driving forces. Formal decision analysis helps to systematically assess the consequences of such information. Scenario-based planning helps individuals and groups accept the fundamental uncertainty surrounding the long-term future and consider a range of potential paths, including those that may be inconvenient or disturbing for organizational, ideological, or political reasons.

Despite this rich legacy, these traditional methods all founder on the same shoals: an inability to grapple with the long term's multiplicity of plausible futures. Any single guess about the future will likely prove wrong. Policies optimized for a most likely future may fail in the face of surprise. Even analyzing a well-crafted handful of scenarios will miss most of the future's richness and provides no systematic means to examine their implications. This is particularly true for methods based on detailed models. Such models that look sufficiently far into the future should raise troubling questions in the minds of both the model builders and the consumers of model output. Yet the root of the problem lies not in the models themselves, but in the way in which models are used. Too often, analysts ask what will happen, thus trapping themselves in a losing game of prediction, instead of the question they really would like to have answered: Given that one cannot predict, which actions available today are likely to serve best in the future?

Broadly speaking, although there are differences in definitions, and ambiguities in meanings, the literature offers four (overlapping, not mutually exclusive) ways for dealing with deep uncertainty in making policies, see van Drunen et al. (2009).

Resistance: plan for the worst conceivable case or future situation,

- *Resilience*: whatever happens in the future, make sure that you have a policy that will result in the system recovering quickly,
- *Static robustness*: implement a (static) policy that will perform reasonably well in practically all conceivable situations,
- *Adaptive robustness*: prepare to change the policy, in case conditions change.

The first approach is likely to be very costly and might not produce a policy that works well because of Black Swans. The second approach accepts short-term pain (negative system performance), but focuses on recovery.

The third and fourth approaches do not use models to produce forecasts. Instead of determining the best predictive model and solving for the policy that is optimal (but fragilely dependent on assumptions), in the face of deep uncertainty it may be wiser to seek among the alternatives those actions that are most robust — that achieve a given level of goodness across the myriad models and assumptions consistent with known facts (Rosenhead and Mingers 2001). This is the heart of any robust decision method. A robust policy is defined to be one that yields outcomes that are deemed to be satisfactory according to some selected assessment criteria across a wide range of future plausible states of the world. This is in contrast to an optimal policy that may achieve the best results among all possible plans but carries no guarantee of doing so beyond a narrowly defined set of circumstances. An analytical policy based on the concept of robustness is also closer to the actual policy reasoning process employed by senior planners and executive decision makers. As shown by Lempert and Collins (2007), analytic approaches that seek robust strategies are often appropriate both when uncertainty is deep and a rich array of options is available to decision makers.

Identifying static robust policies requires reversing the usual approach to uncertainty. Rather than seeking to characterize uncertainties in terms of probabilities, a task rendered impossible by definition for Level 4 and Level 5 uncertainties, one can instead explore how different assumptions about the future values of these uncertain variables would affect the decisions actually being faced. Scenario planning is one approach to identifying static robust policies, see van der Heijden (1996). This approach assumes that, although the likelihood of the future worlds is unknown, a range of plausible futures can be specified well enough to identify a (static) policy that will produce acceptable outcomes in most of them. It works best when dealing with Level 4 uncertainties. Another approach is to ask what one would need to believe was true to discard one possible policy in favor of another. This is the essence of Exploratory Modeling and Analysis (EMA).

Long-term robust policies for dealing with Level 5 uncertainties will generally be dynamic adaptive policies—policies that can adapt to changing conditions over time. A dynamic adaptive policy is developed with an awareness of the range of plausible futures that lie ahead, is designed to be changed over time as new information becomes available, and leverages autonomous response to surprise. Eriksson and Weber (2008) call this approach to dealing with deep uncertainty Adaptive Foresight. Walker et al. (2001) have specified a generic, structured approach for developing dynamic adaptive policies for practically any policy domain. This approach allows implementation to begin prior to the resolution of all major uncertainties, with the policy being adapted over time based on new knowledge. It is a way to proceed with the implementation of long-term policies despite the presence of uncertainties. The adaptive policy approach makes dynamic adaptation explicit at the outset of policy formulation. Thus, the inevitable policy changes become part of a larger, recognized process and are not forced to be made repeatedly on an ad hoc basis. Under this approach, significant changes in the system would be based on an analytic and deliberative effort that first clarifies system goals, and then identifies policies designed to achieve those goals and ways of modifying those policies as conditions change. Within the adaptive policy framework, individual actors would carry out their activities as they would under normal policy conditions. But policymakers and stakeholders, through monitoring and corrective actions, would try to keep the system headed toward the original goals. McCray et al. (2010) describe it succinctly as keeping policy “yoked to an evolving knowledge base.” Lempert et al. (2003, 2006) propose an approach called Robust Decision Making (RDM), which conducts a vulnerability and response option analysis using EMA to identify and compare (static or dynamic) robust policies. Walker et al. (2001) propose a similar approach for developing adaptive policies, called Dynamic Adaptive Policymaking (DAP).

Some Applications of Robust Decision Making (RDM) and Dynamic Adaptive Policymaking (DAP)

RDM has been applied in a wide range of decision applications, including the development of both static and adaptive policies. The study of Dixon et al. (2007) evaluated alternative (static) policies considered by the U.S. Congress while debating reauthorization of the Terrorism Risk Insurance Act (TRIA). TRIA provides a federal guarantee to compensate insurers for losses due to very large terrorist attacks in return for insurers providing insurance against attacks of all sizes. Congress was particularly interested in the cost to taxpayers of alternative versions of the program. The RDM analysis used a simulation model to project these costs for various TRIA options for each of several thousand cases, each representing a different combination of 17 deeply uncertain assumptions about the type of terrorist attack, the factors influencing the pre-attack distribution of insurance coverage, and any post-attack compensation decisions by the U.S. Federal government. The RDM analysis demonstrated that the expected cost to taxpayers of the existing TRIA program would prove the same or less than any of the proposed alternatives except under two conditions: the probability of a large terrorist attack (greater than $40 billion in losses) significantly exceeded current estimates and future Congresses did not compensate uninsured property owners in the aftermath of any such attack. This RDM analysis appeared to help resolve a divisive Congressional debate by suggesting that the existing (static) TRIA program was robust over a wide range of assumptions, except for a combination that many policymakers regarded as unlikely. The analysis demonstrates two important features of RDM: (1) its ability to systematically include imprecise probabilistic information (in this case, estimates of the likelihood of a large terrorist attack) in a formal decision analysis, and (2) its ability to incorporate very different types of uncertain information (in this case, quantitative estimates of attack likelihood and qualitative judgments about the propensity of future Congresses to compensate the uninsured).

RDM has also been used to develop adaptive policies, including policies to address climate change (Lempert et al. 1996), economic policy (Seong et al. 2005), complex systems (Lempert 2002), and health policy (Lakdawalla et al. 2009). An example that illustrates RDM's ability to support practical adaptive policy making is discussed in Groves et al. (2008) and Lempert and Groves (2010). In 2005, Southern California's Inland Empire Utilities Agency (IEUA), that supplies water to a fast growing population in an arid region, completed a legally mandated (static) plan for ensuring reliable water supplies for the next twenty-five years. This plan did not, however, consider the potential impacts of future climate change. An RDM analysis used a simulation model to project the present value cost of implementing IEUA's current plans, including any penalties for future shortages, in several hundred cases contingent on a wide range of assumptions about six parameters representing climate impacts, IEUA's ability to implement its plan, and the availability of imported water. A scenario discovery analysis identified three key factors — an 8% or larger decrease in precipitation, any drop larger than 4% in the rain captured as groundwater, and meeting or missing the plan's specific goals for recycled waste water — that, if they occurred simultaneously, would cause IEUA's overall plan to fail (defined as producing costs exceeding by 20% or more those envisioned in the baseline plan). Having identified this vulnerability of IEUA's current plan, the RDM analysis allowed the agency managers to identify and evaluate alternative adaptive plans, each of which combined near-term actions, monitoring of key supply and demand indicators in the region, and taking specific additional actions if certain indicators were observed. The analysis suggested that IEUA could eliminate most of its vulnerabilities by committing to updating its plan over time and by making relative low-cost near-term enhancements in two current programs. Overall, the analysis allowed IEUA's managers, constituents, and elected officials, who did not all agree on the likelihood of climate impacts, to understand in detail vulnerabilities to their original plan and to identify and reach consensus on adaptive plans that could ameliorate those vulnerabilities.

An example of DAP comes from the field of airport strategic planning. Airports increasingly operate in a privatized and liberalized environment. Moreover, this change in regulations has changed the public's perception of the air transport sector. As a result of this privatization and liberalization, the air transport industry has undergone unprecedented changes, exemplified by the rise of airline alliances and low cost carriers, an increasing environmental awareness, and, since 9/11, increased safety and security concerns. These developments pose a major challenge for airports. They have to make investment decisions that will shape the future of the airport for many years to come, taking into consideration the many uncertainties that are present. DAP has been put forward as a way to plan the long-term development of an airport under these conditions (Kwakkel et al. 2010a). As an illustration, a case based on the current challenges of Amsterdam Airport Schiphol has been pursued. Using a simulation model that calculates key airport performance metrics such as capacity, noise, and external safety, the performance of an adaptive policy and a competing traditional policy across a wide range of uncertainties was explored. This comparison revealed that the traditional plan would have preferable performance only in the narrow bandwidth of future developments for which it was optimized. Outside this bandwidth, the adaptive policy had superior performance. The analysis further revealed that the range of expected outcomes for the adaptive policy is significantly smaller than for the traditional policy. That is, an adaptive policy will reduce the uncertainty about the expected outcomes, despite various deep uncertainties about the future. This analysis strongly suggested that airports operating in an ever increasing uncertain environment could significantly improve the adequacy of their long-term development if they planned for adaptation (Kwakkel et al. 2010b, 2010c).

Another policy area to which DAP has been applied is the expansion of the port of Rotterdam. This expansion is very costly and the additional land and facilities need to match well with market demand as it evolves over the coming 30 years or more. DAP was used to modify the existing plan so that it can cope with a wide range of uncertainties. To do so, adaptive policy making was combined with Assumption-Based Planning (Dewar 2002). This combination resulted in the identification of the most important assumptions underlying the current plan. Through the adaptive policy making framework, these assumptions were categorized and actions for improving the likelihood that the assumptions will hold were specified (Taneja et al. 2010).

Various other areas of application of DAP have also been explored, including flood risk management in the Netherlands in light of climate change (Rahman et al. 2008), policies with respect to the implementation of innovative urban transport infrastructures (Marchau et al. 2008), congestion road pricing (Marchau et al. 2010), intelligent speed adaptation (Agusdinata et al. 2007), and magnetically levitated (Maglev) rail transport (Marchau et al. 2010).

See

▶ Exploratory Modeling and Analysis

References

Agusdinata, D. B. (2008). *Exploratory modeling and analysis: A promising method to deal with deep uncertainty*. Ph.D. dissertation, Delft University of Technology, The Netherlands.

Agusdinata, D. B., Marchau, V. A. W. J., & Walker, W. E. (2007). Adaptive policy approach to implementing intelligent speed adaptation. *IET Intelligent Transport Systems (ITS), 1*(3), 186–198.

Ben-Haim, Y. (2006). *Information-gap decision theory: Decisions under severe uncertainty* (2nd ed.). New York: Wiley.

Courtney, H. (2001). *20/20 foresight: Crafting strategy in an uncertain world*. Boston: Harvard Business School Press.

Dewar, J. A. (2002). *Assumption-based planning: A tool for reducing avoidable surprises*. Cambridge, UK: Cambridge University Press.

Dixon, L., Lempert, R.J., LaTourrette, T., & Reville, R.T. (2007). *The Federal role in terrorism insurance: Evaluating alternatives in an uncertain world (MG-679-CTRMP)*. Santa Monica, CA: RAND.

Eriksson, E. A., & Weber, K. M. (2008). Adaptive foresight: Navigating the complex landscape of policy strategies. *Technological Forecasting and Social Change, 75*, 462–482.

Funtowicz, S. O., & Ravetz, J. R. (1990). *Uncertainty and quality in science for policy*. Dordrecht, The Netherlands: Kluwer.

Groves, D. G., Davis, M., Wilkinson, R., Lempert, R. (2008). Planning for climate change in the inland empire: Southern California. *Water Resources IMPACT*, July 2008.

Hillier, F. S., & Lieberman, G. J. (2001). *Introduction to operations research*. New York: McGraw Hill.

Knight, F. H. (1921). *Risk, uncertainty and profit*. New York: Houghton Mifflin Company (republished in 2006 by Dover Publications, Mineola, NY).

Kwakkel, J. H., Walker, W. E., & Marchau, V. A. W. J. (2010a). Adaptive airport strategic planning. *European Journal of Transport and Infrastructure Research, 10*(3), 249–273.

Kwakkel, J. H., Walker, W. E., & Marchau, V. A. W. J. (2010b). From predictive modeling to exploratory modeling: How to use non-predictive models for decision-making under deep uncertainty. *25th Mini-EURO Conference on Uncertainty and Robustness in Planning and Decision Making (URPDM 2010)*, Coimbra, Portugal, 15–17 April 2010.

Kwakkel, J. H., Walker, W. E., & Marchau, V. A. W. J. (2010c). Assessing the efficacy of adaptive airport strategic planning: Results from computational experiments. *World Conference on Transport Research*, Porto, Portugal, 11–15 July 2010.

Kwakkel, J. H., Walker, W. E., & Marchau, V. A. W. J. (2010d). Classifying and communicating uncertainties in model-based policy analysis. *International Journal of Technology, Policy and Management, 10*(4), 299–315.

Lakdawalla, D. N., Goldman, D. P., Michaud, P.-C., Sood, N., Lempert, R., Cong, Z., de Vries, H., & Gutlerrez, I. (2009). US pharmaceutical policy in a global marketplace. *Health Affairs, 28*, 138–150.

Lempert, R. J. (2002, May 14). A new decision sciences for complex systems. *Proceedings of the National Academy of Sciences*, *99*(Suppl. 3), 7309–7313.

Lempert, R. J., & Collins, M. T. (2007). Managing the risk of uncertain threshold response: Comparison of robust, optimum, and precautionary approaches. *Risk Analysis, 27*(4), 1009–1026.

Lempert, R. J., & Groves, D. G. (2010). Identifying and evaluating robust adaptive policy responses to climate change for water management agencies in the American west. *Technological Forecasting and Social Change, 77*, 960–974.

Lempert, R. J., Groves, D. G., Popper, S. W., & Bankes, S. C. (2006). A general, analytic method for generating robust strategies and narrative scenarios. *Management Science, 52*(4), 514–528.

Lempert, R. J., Popper, S. W., & Bankes, S. C. (2003). *Shaping the next one hundred years: New methods for quantitative long-term strategy analysis (MR-1626-RPC)*. Santa Monica, CA: The RAND Pardee Center.

Lempert, R. J., Schlesinger, M. E., & Bankes, S. C. (1996). When we don't know the costs or the benefits: Adaptive strategies for abating climate change. *Climatic Change, 33*, 235–274.

Luce, R. D., & Raiffa, H. (1957). *Games and decisions*. New York: Wiley.

Makridakis, S., Hogarth, R. M., & Gaba, A. (2009). Forecasting and uncertainty in the economic and business world. *International Journal of Forecasting, 25*, 794–812.

Marchau, V., Walker, W., & van Duin, R. (2008). An adaptive approach to implementing innovative urban transport solutions. *Transport Policy, 15*(6), 405–412.

Marchau, V. A. W. J., Walker, W. E., & van Wee, G. P. (2010). Dynamic adaptive transport policies for handling deep uncertainty. *Technological Forecasting and Social Change, 77*(6), 940–950.

McCray, L. E., Oye, K. A., & Petersen, A. C. (2010). Planned adaptation in risk regulation: An initial survey of US environmental, health, and safety regulation. *Technological Forecasting and Social Change, 77*, 951–959.

McDaniel, R. R., & Driebe, D. J. (Eds.). (2005). *Uncertainty and surprise in complex systems: Questions on working with the unexpected*. Springer.

Morgan, M. G. (2003). Characterizing and dealing with uncertainty: Insights from the integrated assessment of climate change. *The Integrated Assessment Journal, 4*(1), 46–55.

Morgan, M. G., & Henrion, M. (1990). *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis*. Cambridge, UK: Cambridge University Press.

Patt, A. G., & Dessai, S. (2004). Communicating uncertainty: Lessons learned and suggestions for climate change assessment. *Comptes Rendu Geosciences, 337*, 425–441.

Patt, A. G., & Schrag, D. (2003). Using specific language to describe risk and probability. *Climatic Change, 61*, 17–30.

Popper, S. W., Griffin, J., Berrebi, C., Light, T., & Min, E. Y. (2009). *Natural gas and Israel's energy future: A strategic analysis under conditions of deep uncertainty* (TR-747-YSNFF). Santa Monica, CA: RAND.

Quade, E. S. (1989). *Analysis for public decisions* (3rd ed.). New York: Elsevier Science.

Rahman, S. A., Walker, W. E., & Marchau, V. (2008). *Coping with uncertainties about climate change in infrastructure planning – An adaptive policymaking approach*. ECORYS Nederland BV, P.O. Box 4175, 3006 AD, Rotterdam, The Netherlands.

Rosenhead, J., & Mingers, J. (Eds.). (2001). *Rational analysis for a problematic world revisited: Problem structuring methods for complexity, uncertainty, and conflict*. Chichester, UK: Wiley.

Seong, S., Popper, S. W., & Zheng, K. (2005). *Strategic choices in science and technology Korea in the era of a rising China* (MG-320-KISTEP). Santa Monica, CA: RAND.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423. 623–656, July, October.

Smithson, M. (1989). *Ignorance and uncertainty: Emerging paradigms*. New York: Springer.

Taleb, N. N. (2007). *The black swan: The impact of the highly improbable*. New York: Random House.

Taneja, P., Walker, W. E., Ligteringen, H., Van Schuylenburg, M., & van der Plas, R. (2010). Implications of an uncertain future for port planning. *Maritime Policy & Management, 37*(3), 221–245.

van Asselt, M. B. A. (2000). *Perspectives on uncertainty and risk*. Dordrecht, The Netherlands: Kluwer.

van Asselt, M. B. A., & Rotmans, J. (2002). Uncertainty in integrated assessment modelling: From positivism to pluralism. *Climatic Change, 54*, 75.

van der Heijden, K. (1996). *Scenarios: The art of strategic conversation*. Chichester, UK: Wiley.

van der Sluijs, J. P. (1997). *Anchoring amid uncertainty: On the management of uncertainties in risk assessment of anthropogenic climate change*. Ph.D. dissertation, University of Utrecht, The Netherlands.

van Drunen, M., Leusink, A., Lasage, R. (2009). Towards a climate-proof Netherlands. In A. K. Biswas, C. Tortajade, & R. Izquierdo (Eds.), *Water management in 2020 and beyond*. Springer.

Walker, W. E., Harremoës, P., Rotmans, J., van der Sluijs, J. P., van Asselt, M. B. A., Janssen, P., & Krayer von Krauss, M. P. (2003). Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support. *Integrated Assessment, 4*(1), 5–17.

Walker, W. E., Rahman, S. A., & Cave, J. (2001). Adaptive policies, policy analysis, and policymaking. *European Journal of Operational Research, 128*(2), 282–289.

Degeneracy

The situation in which a linear-programming problem has a basic feasible solution with at least one basic variable equal to zero. If the problem is degenerate, then an extreme point of the convex set of solutions may correspond to several feasible bases. As a result, the simplex method may move through a sequence of bases with no improvement in the value of the objective function. In rare cases, the algorithm may cycle repeatedly through the same sequence of bases and never converge to an optimal solution. Anticycling rules, and perturbation and lexicographic techniques prevent this risk, but usually at some computational expense.

See

- ▶ Anticycling Rules
- ▶ Bland's Anticycling Rules
- ▶ Cycling
- ▶ Linear Programming
- ▶ Simplex Method (Algorithm)

Degeneracy Graphs

Tomas Gal
Fern Universität in Hagen, Hagen, Germany

Introduction

For a given linear-programming problem, primal degeneracy means that a basic feasible solution has at least one basic variable equal to zero. The problem is dual degenerate if a nonbasic variable has its reduced cost equal to zero (the condition for a multiple optimal solution to exist). Primal degeneracy may arise when there are some (weakly) redundant constraints (Karwan et al. 1983) or the structure of the corresponding convex polyhedral feasible set causes an extreme point to become overdetermined.

In nonlinear programming, such points are sometimes called singularities (Guddat et al. 1990). Here, constraint redundancy is equivalent to the failure of the linear independence constraint qualification of the binding constraint gradients, which, in general, leads to the nonuniqueness of optimal Lagrange multipliers (Fiacco and Liu 1993).

We focus here on primal degeneracy in the linear case: it is associated with multiple optimal bases and it allows for basis cycling to occur, that is, the nonconvergence of the simplex method due to the repeating of a sequence of nonoptimal feasible bases.

Let σ, called the degeneracy degree, be the number of zeros in a basic feasible solution. Also, let $U_{\min}$ and $U_{\max}$ be the minimal and the maximal number of possible bases associated with a degenerate vertex, respectively (Kruse 1986). To illustrate how many bases can be associated with a degenerate vertex, Table 1 shows, for some values for n, the number of (decision) variables, the associated values of σ, $U_{\min}$ and $U_{\max}$.

Historical Background

Soon after the simplex method had been invented by George Dantzig, he recognized that degeneracy in the primal problem could cause a cycle of bases to occur. In fact, Dantzig's original convergence proof of the simplex method assumed that all basic feasible solutions were nondegenerate. In the Fall of 1950, Dantzig made the first suggestion of a nondegeneracy procedure in a lecture on linear programming (LP) (Dantzig 1963). Charnes (1952) proposed a so-called perturbation method to prevent cycling. Since then, many variants of nondegeneracy and anticycling methods have been developed. For a review of degeneracy and its influence on computation, see Gal (1993).

In the end of the 1970s, a unifying approach to the analysis of degeneracy problems was proposed in terms of degeneracy graphs (Gal 1985). These graphs are used to define the connections among the bases associated with a degenerate vertex. From Table 1, it obvious that for real-world problems, with large numbers of constraints and variables, such systems of connections might have quite complex structures. It was felt that the language of graph theory could be applied to good advantage in explaining the relationships between degenerate bases.

Since they were first proposed, degeneracy graphs have become an important topic of research (Geue 1993; Kruse 1986; Niggemeier 1993; Zörnig 1993). In these works, the general theory of degeneracy graphs has been developed, the possibilities for their application to transportation, integer programming and other problems have been studied, and algorithmic aspects to solve various degeneracy problems have been investigated.

The main problem that led to the idea of using a graph theoretical representation was the so called

Degeneracy Graphs, Table 1 Values for σ, $U_{\min}$, $U_{\max}$

n	σ	$U_{\min}$	$U_{\max}$
5	3	16	56
10	5	12	3003
50	5	752	3.48×10^{6}
50	40	6.59×10^{12}	5.99×10^{25}
100	30	3.865×10^{10}	2.61×10^{39}
100	50	2.93×10^{16}	2.01×10^{40}
100	80	1.33×10^{25}	3×10^{52}

neighboring problem: Given a vertex of a convex polytope, find all neighboring vertices. This is not a problem if the given vertex is nondegenerate. It becomes a problem (Table 1) when the given vertex is degenerate.

Degeneracy Graphs

Given a σ-degenerate vertex $\boldsymbol{x}^{\mathrm{o}}$; to this vertex the set

$$B^{\mathrm{o}} = \{B|B\,\text{feasible basis of}\,x^{\mathrm{o}}\}$$

is assigned. Denote

$$\begin{aligned}&\text{by}\text{``}\leftarrow + \rightarrow\text{''}\,\text{a pivot} - \text{step with a positive}\\&\quad\text{pivot}\ (\text{a positive pivot} - \text{step})\\&\text{by}\text{``}\leftarrow - \rightarrow\text{''}\,\text{a pivot} - \text{step with a negative}\\&\quad\text{pivot}\ (\text{a negative pivot} - \text{step})\\&\text{by}\text{``}\leftarrow\rightarrow\text{''}\,\text{a pivot} - \text{step if any nonzero}\\&\quad\text{pivot can be used}\ (\text{pivot} - \text{step}).\end{aligned}$$

The graph of a polytope X is the undirected graph

$$G(X) := G = (V, E),$$

where

$$V = \{B|B\ \text{is a feasible basis of the corresponding system of equations}\}$$

and

$$E = \{\{B, B'\} \subseteq V|B \leftarrow + \rightarrow B'\}.$$

The degeneracy graph (DG) that is used to study various degeneracy problems with respect to a degenerate vertex is defined as follows. Let $\boldsymbol{x}^{\mathrm{o}} \in X \subset \mathfrak{R}^n$ be a σ-degenerate vertex. Then the (undirected) graph

$$G(x^{\mathrm{o}}) := G^{\mathrm{o}} = (B^{\mathrm{o}}, E^{\mathrm{o}})$$

where

$$\begin{aligned}E^{\mathrm{o}} = \{\{B_u, B_v\} \subseteq B^{\mathrm{o}}|B_u \longleftrightarrow B_v\}, u, v\\ \in \{1, \ldots, U\}, U_{\min} \leq U \leq U_{\max}\end{aligned} \qquad (1)$$

and U, the degeneracy power of $\boldsymbol{x}^{\mathrm{o}}$, is called the general $\sigma \times n - G$ of $\mathbf{x}^{\mathrm{o}}$. If, in (1), the operator is $\leftarrow + \rightarrow$ or $\leftarrow - \rightarrow$, then the corresponding graph is called the positive or negative DG of $\boldsymbol{x}^{\mathrm{o}}$, respectively.

These notions have been used to develop a theory of the DG. For example: the diameter, d, of a general DG satisfies $d \leq \min\{\sigma, n\}$; a general DG is always connected; a formula to determine the number of nodes of a DG has been developed; the connectivity of a DG is ≥2; every pair of nodes in any DG lies on a cycle (Zörnig 1993).

An interesting consequence of this theory is that every degenerate vertex can be exited in at most d (diameter) steps. Other theoretical properties of DGs help in explaining problems in, for example, sensitivity analysis with respect to a degenerate vertex (Gal 1997; Kruse 1993). Also, this theory helps to work out algorithms to solve the neighborhood problem and to determine all vertices of a convex polytope (Gal and Geue 1992; Geue 1993; Kruse 1986). With respect to a degenerate optimal vertex of an LP-problem, algorithms to perform sensitivity analysis and parametric programming have been developed (Gal 1995). Also, the connection between weakly redundant constraints, degeneracy and sensitivity analysis has been studied (Gal 1992).

Concluding Remarks

Degeneracy graphs have been applied to help solve the neighborhood problem, to explain why cycling in LP occurs, to develop algorithms to determine two-sided shadow prices, to determine all vertices of a (degenerate) convex polyhedron, and to perform sensitivity analysis under (primal) degeneracy. DGs can be used in any mathematical-programming problem that uses some version of the simplex method or, more generally, in any vertex searching method.

See

▶ Degeneracy
▶ Graph Theory
▶ Linear Programming

▶ Parametric Programming
▶ Redundant Constraint
▶ Sensitivity Analysis

References

Charnes, A. (1952). Optimality and degeneracy in linear programming. *Econometrica, 20*, 160–170.

Dantzig, G. B. (1963). *Linear programming and extensions*. Princeton, New Jersey: Princeton University Press.

Fiacco, A. V., & Liu, J. (1993). Degeneracy in NLP and the development of results motivated by its presence. In T. Gal (Ed.), *Degeneracy in optimization problems. Annals of OR, 46/47*, 61–80

Gal, T. (1985). On the structure of the set bases of a degenerate point. *Journal of Optimization Theory and Applications, 45*, 577–589.

Gal, T. (1986). Shadow prices and sensitivity analysis in LP under degeneracy — state-of-the-art survey. *OR-Spektrum, 8*, 59–71.

Gal, T. (1992). Weakly redundant constraints and their impact on postoptimal analysis in LP. *European Journal of Operational Research, 60*, 315–336.

Gal, T. (1993). Selected bibliography on degeneracy. In: T. Gal (Ed.), *Degeneracy in optimization problems. Annals of OR, 46/47*, 1–7.

Gal, T. (1995). *Postoptimal analyses, parametric programming, and related topics*. Berlin, New York: W. de Gruyter.

Gal, T. (1997). Linear programming 2: Degeneracy graphs. In T. Gal & H. J. Greenberg (Eds.), *Advances in sensitivity analysis and parametric programming*. Dordrecht: Kluwer Academic Publishers.

Gal, T., & Geue, F. (1992). A new pivoting rule for solving various degeneracy problems. *Operations Research Letters, 11*, 23–32.

Geue, F. (1993). An improved N-tree algorithm for the enumeration of all neighbors of a degenerate vertex. In: T. Gal (Ed.), *Degeneracy in optimization problems. Annals of OR, 46/47*, 361–392.

Guddat, J. F., Guerra Vasquez, F. & Jongen, Th. H. (1991). *Parametric Optimization: Singularities, Path Following and Jumps*. New York: R.G. Teubner and J. Wiley.

Karwan, M.H., Lotfi, F., Telgen, J.& Zionts, S. (Eds), (1983). *Redundancy in mathematical programming: A state-of-the-art survey*. Lecture Notes in Econ. and Math. Systems 206. Berlin: Springer Verlag.

Kruse, H. J. (1986). *Degeneracy graphs and the neighborhood problem*. Lecture Notes in Econ. and Math. Systems 260. Berlin: Springer Verlag.

Kruse, H. J. (1993). On some properties of σ-degeneracy graphs. In: T. Gal (Ed.), *Degeneracy in optimization problems. Annals of OR, 46/47*, 393–408.

Niggemeier, M. (1993). Degeneracy in integer linear optimization problems: A selected bibliography. In: T. Gal (Ed.), *Degeneracy in optimization problems. Annals of OR, 46/47*, 195–202.

Zörnig, P. (1993). A theory of degeneracy graphs. In: T. Gal (Ed.), *Degeneracy in optimization problems. Annals of OR, 46/47*, 541–556.

Degenerate Solution

A basic (feasible) solution in which some basic variables are zero.

See

▶ Anticycling Rules
▶ Cycling
▶ Degeneracy
▶ Degeneracy Graphs

Degree

The number of edges incident with a given node in a graph.

See

▶ Graph Theory

Delaunay Triangulation

▶ Computational Geometry
▶ Voronoi Constructs

Delay

The time spent by a customer in queue waiting to start service.

See

▶ Queueing Theory
▶ Waiting Time

Delphi Method

James A. Dewar and John A. Friel
RAND Corporation, Santa Monica, CA, USA

Introduction

The Delphi method was developed at the RAND Corporation from studies on decision making that began in 1948. The seminal work, "An Experimental Application of the Delphi Method to the Use of Experts," was written by Dalkey and Helmer (1963).

The primary rationale for the technique is the age-old adage "two heads are better than one," particularly when the issue is one where exact knowledge is not available. It was developed as an alternative to the traditional method of obtaining group opinions — face-to-face discussions. Experimental studies had demonstrated several serious difficulties with such discussions. Among them were: (1) influence of the dominant individual (the group is highly influenced by the person who talks the most or has most authority); (2) noise (studies found that much communication in such groups had to do with individual and group interests rather than problem solving); and (3) group pressure for conformity (studies demonstrated the distortions of individual judgment that can occur from group pressure).

The Delphi method was specifically developed to avoid these difficulties. In its original formulation it had three basic features: (1) anonymous response — opinions of the members of the group are obtained by formal questionnaire; (2) iteration and controlled feedback — interaction is effected by a systematic exercise conducted in several iterations, with carefully controlled feedback between rounds; and (3) statistical group response — the group opinion is defined as an appropriate aggregate of individual opinions on the final round.

Procedurally, the Delphi method begins by having a group of experts answer questionnaires on a subject of interest. Their responses are tabulated and fed back to the entire group in a way that protects the anonymity of their responses. They are asked to revise their own answers and comment on the group's responses. This constitutes a second round of the Delphi. Its results are tabulated and fed back to the group in a similar manner and the process continues until convergence of opinion, or a point of diminishing returns, is reached. The results are then compiled into a final statistical group response to assure that the opinion of every member of the group is represented.

In its earliest experiments, Delphi was used for technological forecasts. Expert judgments were obtained numerically (e.g., the date that a technological advance would be made), and in that case it is easy to show that the mean or median of such judgments is at least as close to the true answer as half of the group's individual answers. From this, the early proponents were able to demonstrate that the Delphi method produced generally better estimates than those from face-to-face discussions.

One of the surprising results of experiments with the technique was how quickly in the successive Delphi rounds that convergence or diminishing returns is achieved. This helped make the Delphi technique a fast, relatively efficient, and inexpensive tool for capturing expert opinion. It was also easy to understand and quite versatile in its variations. By 1975, there were several hundred applications of the Delphi method reported on in the literature. Many of these were applications of Delphi in a wide variety of judgmental settings, but there was also a growing academic interest in Delphi and its effectiveness.

Critique

Sackman (1975), also of the RAND Corporation, published the first serious critique of the Delphi method. His book, *Delphi Critique*, was very critical of the technique — particularly its numerical aspects — and ultimately recommended (p. 74) "that … Delphi be dropped from institutional, corporate, and government use until its principles, methods, and fundamental applications can be experimentally established as scientifically tenable."

Sackman's critique spurred both the development of new techniques for obtaining group judgments and a variety of studies comparing Delphi with other such techniques. The primary alternatives can be categorized as statistical group methods (where the answers of the group are tabulated statistically without any interaction); unstructured, direct interaction (another name for traditional, face-to-face discussions); and structured, direct interaction (such as the Nominal Group Technique of Gustafson et al. 1973). In his comprehensive review, Woudenberg (1991) found no clear evidence in studies done for the superiority of any of the four methods over the others. Even after discounting several of the studies for methodological difficulties, he concludes that the original formulation of the quantitative Delphi is in no way superior to other (simpler, faster, and cheaper) judgment methods.

Another comprehensive evaluation of Delphi (Rowe et al. 1991) comes to much the same conclusion that Sackman and Woudenberg did, but puts much of the blame on studies that stray from the original precepts. Most of the negative studies use non-experts with similar backgrounds (usually undergraduate or graduate students) in simple tests involving almanac-type questions or short-range forecasts. Rowe et al. (1991) point out that these are poor tests of the effects that occur when a variety of experts from different disciplines iterate and feed back their expertise to each other. They conclude that Delphi does have potential in its original intent as a judgment-aiding technique, but that improvements are needed and those improvements require a better understanding of the mechanics of judgment change within groups and of the factors that influence the validity of statistical and nominal groups.

Applications

In the meantime, it is generally conceded that Delphi is extremely efficient in achieving consensus and it is in this direction that many subsequent Delphi evaluations have been used. Variations of the Delphi method, such as the policy Delphi and the decision Delphi, generally retain the anonymity of participants and iteration of responses. Many retain specific feedback as well, but these more qualitative variations generally drop the statistical group response. Delphi has been used in a wide variety of applications from its original purpose of technology forecasting (one report says that Delphi has been adopted in approximately 90% of the technological forecasts and studies of technological development strategy in China) to studying the future of medicine, examining possible shortages of strategic materials, regional planning of water and natural resources, analyzing national drug abuse policies, and identifying corporate business opportunities.

In addition, variations of Delphi continue to be developed to accommodate the growing understanding of its shortcomings. For example, a local area network (LAN) was constructed, composed of lap-top computers connected to a more capable workstation. Each participant had a dedicated spreadsheet available on a lap-top computer. The summary spreadsheet maintained by the workstation was displayed using a large-screen projector, and included the mean, media, standard deviation, and histogram of all the participants scores. In real-time, the issues were discussed, the various participants presented their interpretation of the situation, presented their analytic arguments for the scores they believed to be appropriate, and changed their scoring as the discussion developed. Each participant knew their scores, but not those of the other participants. When someone was convinced by the discussions to change a score they could do so anonymously. The score was transmitted to the workstation where a new mean, median, standard deviation, and histogram were computed and then displayed using a large screen projector. This technique retained all the dimensions of the traditional Delphi method and at the same time facilitated group discussion and real-time change substantially shortening the time typically required to complete a Delphi round.

See

- ▶ Decision Analysis
- ▶ Group Decision Computer Technology
- ▶ Group Decision Making

References

Dalkey, N., & Helmer, O. (1963). An experimental application of the delphi method to the use of experts. *Management Science*, *9*, 458–467.

Gustafson, D. H., Shukla, R. K., Delbecq, A., & Walster, G. W. (1973). A comparison study of differences in subjective likelihood estimates made by individuals, interacting groups, delphi groups, and nominal groups. *Organizational Behavior and Human Performance*, *9*, 280–291.

Keeney, S., & McKenna, H. (2011). *The delphi method in nursing and health research*. West Sussex, UK: John Wiley & Sons.

Rowe, G., Wright, G., & Bolger, F. (1991). Delphi: A reevaluation of research and theory. *Technological Forecasting and Social Change, 39*, 235–251.

Sackman, H. (1975). *Delphi critique*. Lexington, MA: Lexington Books.

Woudenberg, F. (1991). An evaluation of delphi. *Technological Forecasting and Social Change, 40*, 131–150.

Density

The proportion of the coefficients of a constraint matrix that are nonzero. For a given $(m \times n)$ matrix $\boldsymbol{A} = (a_{ij})$, if k is the number of nonzero a_{ij}, then the density is given by $k/(m \times n)$. Most large-scale linear-programming problems have a low density of the order of 0.01.

See

- ▶ Sparse Matrix
- ▶ Super-Sparsity

Density Function

When the derivative $f(x)$ of a cumulative probability distribution function $F(x)$ exists, it is called the density or probability density function (PDF).

See

- ▶ Probability Density Function (PDF)

Departure Process

Usually refers to the random sequence of customers leaving a queueing service center. More generally, it is the random point process or marked point process with marks representing aspects of the departure stream and/or the service center or node from which they are leaving. For example, the marked point process $(\boldsymbol{X}^d, \boldsymbol{T}^d)$ for departures from an M/G/1 queue takes $\boldsymbol{X}^d$ as the Markov process for the queue length process immediately after the departure time and $\boldsymbol{T}^d$ is the actual time of departure.

See

- ▶ Markov Chains
- ▶ Markov Processes
- ▶ Queueing Theory

Descriptive Model

A model that attempts to describe the actual relationships and behavior of a man/machine system. For a decision problem, such a model attempts to describe how individuals make decisions.

See

- ▶ Decision Problem
- ▶ Expert Systems
- ▶ Mathematical Model
- ▶ Model
- ▶ Normative Model
- ▶ Prescriptive Model

Design and Control

For a queueing system, design deals with the permanent, optimal setting of system parameters (such as service rate and/or number of servers), while control deals with adjusting system parameters as the system evolves to ensure certain performance levels are met. A typical example of a control rule is that a server is to be added when the queue size is greater than a certain number (say N_1) and when the queue size drops down to $N_2 < N_1$, the server goes to other duties.

See

- ▶ Dynamic Programming
- ▶ Markov Decision Processes
- ▶ Queueing Theory

Detailed Balance Equations

A set of equations balancing the expected, steady-state flow rates or probability flux between each pair of states or entities of a stochastic process (most typically a Markov chain or queueing problem), for example written as:

$$\pi_j q(j,k) = \pi_k q(k,j)$$

where π_m is the probability that the state is m and $q(m,\ n)$ is the flow rate from states m to n. The states may be broadly interpreted to be multi-dimensional, as in a network of queues, and the entities might be individual service centers or nodes. Contrast this with global balance equations, where the average flow into a single state is equated with the flow out.

See

▶ Markov Chains
▶ Networks of Queues
▶ Queueing Theory

Determinant

▶ Matrices and Matrix Algebra

Deterministic Model

A mathematical model in which it is assumed that all input data and parameters are known with certainty.

See

▶ Descriptive Model
▶ Mathematical Model
▶ Model
▶ Normative Model
▶ Prescriptive Model
▶ Stochastic Model

Developing Countries

Roberto Diéguez Galvão[1] and Graham K. Rand[2]
[1]Federal University of Rio de Janeiro, Brazil
[2]Lancaster University, Lancaster, UK

Introduction

OR started to establish itself in the developing countries in the 1950s, approximately one decade after its post-war inception in Great Britain and the United States. The main organizational basis of OR in the developing world are the national OR societies. These are in some cases well established, in other cases incipient. A number of them are members of the International Federation of Operational Research Societies (IFORS) and belong to regional groups within IFORS. In particular, ALIO, the Association of Latin American OR Societies, has the majority of its member societies belonging to developing countries. APORS, the Association of Asian-Pacific OR Societies within IFORS, also represents OR societies from developing countries. In 1989 a Developing Countries Committee was established as part of the organizational structure of IFORS, with the objective of coordinating OR activities in the developing countries and promoting OR in these countries.

The Social, Political, and Technological Environment

To speak of developing countries in general may lead to erroneous conclusions, since the conditions vary enormously from one country to another. First of all, how to characterize a developing country? Which countries may be classified as developing? The United Nations has, for some years now, started to distinguish between more and less developed countries in the developing world. It has adopted the term “less developed countries” (LDCs) to address those developing countries that fall below some threshold levels measured by social and economic indicators. But these questions are clearly well beyond the scope here.

The view here is that developing countries are those in which large strata of the population live at or below the subsistence level, where social services are practically nonexistent for the majority of the population, where the educational and cultural levels are in general very low. The political consequence of this state of affairs is a high degree of instability for the institutions of these countries, at all levels.

The economy is generally very dependent on the industrialized nations. Bureaucracy, economic dependence and serious problems of infrastructure conspire against economic growth. In the technical sphere there is again a high level of dependency on the industrialized world, with very little technological innovation produced locally. It is against this difficult background that one must consider the role OR can play and how OR can be used as a tool for development.

The Use of OR

Here the existence of three different emphases in the development of OR is considered: (i) development of theory, which takes place mostly in the universities; (ii) development of methods for specific problems, which occurs both in the universities and in the practical world; (iii) applications, which occur mostly in the practical world. The problems of OR are therefore a continuum, and both developing and industrialized nations share in all these three aspects of the continuum. The more important aspect for the developing countries tends, however, to be applications due to the nature of problems these nations have to face and their social, political, and technological environment discussed above. According to Rosenhead (1995), another important aspect is that existing theory and methods, grown in the developed world, are in many cases a poor fit for the problems facing the developing countries. Work on novel applications will be likely to throw up new methods and techniques of general interest.

The use of OR in the developing world is often seen as disconnected from the socio-economic needs of the respective countries, see Galvão (1988). Valuable theoretical contributions originate in these countries, but little is seen in terms of new theory and methods developed for the problems facing them.

A common situation in developing countries is a highly uncertain environment, which leads to the notion of wicked problems. These are, for example, problems for which there is often little or no data available, or where the accuracy of data is very poor. Complex decisions must nevertheless be made, against a background of competing interests and decision makers. There are not many tools available for solving these wicked problems, which are quite common in developing countries.

One of the main characteristics of applied OR projects in developing countries is that a large majority of them have not been implemented, see Löss (1981). This is due to a high degree of instability in institutions in these countries, to a lack of management education in OR, and to a tendency by OR analysts to attempt to use sophisticated OR techniques without paying due attention to the local environment and to the human factor in applied OR projects. These issues arise both in developed and developing countries, but experience indicates that they are more often overlooked in the latter.

A Special Issue of the European Journal of Operational Research (Bornstein et al. 1990) was dedicated to OR in Developing Countries. A review paper (White et al. 2011) provides an overall picture of the state of OR in the developing countries. In particular, it examines coverage in terms of countries and methods and highlights the contribution which OR is making towards the theme of poverty, the reduction of which is regarded as the key focus of development policy interventions as reflected in the Millennium Development Goals. Jaiswal (1985) and Rosenhead and Tripathy (1996) contain important contributions to the subject of OR in developing countries.

ICORD ’92: The Ahmedabad Conference

Since the 1950s, there has been a controversy on the role of OR in developing countries. The central issue in this controversy is the following: Is there a separate OR for developing countries? If so, what makes it different from traditional OR? What steps could be taken to further OR in developing countries?

This issue has been discussed in different venues and several published papers have addressed it, see, for example, Bornstein and Rosenhead (1990). At one end of the scale there are those who think that there is nothing special about OR in developing countries, perhaps only less resources are available in these countries to conduct theoretical/applied work. They argue that the problem should resolve itself when each country reaches appropriate levels of development, and not much time should be dedicated to this issue. At the other end there are those who think that because of a different material basis and due to problems of infrastructure, OR does have a different role to play in these countries. In the latter case, steps should be taken to ensure that OR plays a positive role in the development of their economies and societies.

Much changed in the latter part of the 1990s with the demise of communism in Europe and the emphasis on the globalization of the economy. The viewpoint that there is a separate OR for developing countries lost strength as a consequence. It had its high moment during ICORD '92, the first International Conference on Operational Research for Development, which took place in December 1992, at the Indian Institute of Management (IIM) in Ahmedabad. It was supported by IFORS, The British OR Society and the OR Society of India. It was partly funded by IIM itself, The Tata Iron and Steel Company (India) and (indirectly) by the Commonwealth Secretariat. Participants at the Conference numbered more than 60 and countries represented included Australia, Brazil, Eire, Great Britain, Greece, India, Kenya, Malaysia, Mexico, Nigeria, Peru, South Africa, Sri Lanka, United States, and Venezuela. Some 40 contributed papers were delivered and plenary speakers included the President of IFORS, Professor Brian Haley, Professor Kirit Parikh, Director of the Indira Ghandi Institute for Development in Bombay, and Dr. Francisco Sagasti of Peru, who had just spent five years in senior positions at the World Bank (Rosenhead 1993).

A series of plenary sessions were held, which resulted in a statement which has come to be known as the Ahmedabad Declaration, a political document drafted with the intention of strengthening the OR for Development movement, that called for a range of actions from IFORS to support and strengthen OR in developing countries, including a call for more space for discussion of OR for Development issues in OR departments in developed countries, for IFORS support for successor conferences to ICORD '92, and for IFORS increased economic support of OR activities in developing countries. It relied mainly on IFORS for its implementation. Despite IFORS' continued support of some OR activities in the developing countries, few of the main recommendations of the declaration were implemented. ICORD '96, the second Conference in the series, which took place in Rio de Janeiro, Brazil, in August 1996, was a disappointing sequel to the Ahmedabad Conference and signaled the decline of the movement.

Despite the perceived lack of commitment on the part of IFORS to implement these proposals (Rosenhead 1998), IFORS support of development related OR activities have continued, including the support of successor ICORDs, held in Manila, The Philippines (1997), Berg-en-Dal, South Africa (2001), Jamshedpur, India (2005), Fortaleza, Brazil (2007) and Djerba Island, Tunisia (2012). The IFORS Prize for OR in Development (known as the Third World Prize until 1993) competition has been held at every triennial conference since 1987. The Prize recognizes exemplary work in the application of OR to address issues of development. More recently, a particular focus has been encouraging the development of an OR infrastructure in Africa, and, with EURO, IFORS has sponsored conferences and scholarships in the African continent.

A fuller account of IFORS initiatives in promoting the use of OR for development is described in by Rand (2000). See also del Rosario and Rand (2010).

Is it safe to conclude, therefore, that those who advocate that there is nothing special about OR in developing countries had the better insight on the controversy? The hard facts of life show that little has changed in the social, political and technological environment in the developing countries. The decline of the OR for Development movement is a consequence of the new balance of power in global affairs since the Soviet Union ceased to exist. This decline did not occur because conditions in the developing world improved, or because OR has failed to contribute to the development of the respective economies and societies.

See

▶ IFORS
▶ Practice of Operations Research and Management Science
▶ Wicked Problems

References

Bornstein, C. T., & Rosenhead, J. (1990). The role of operational research in less developed countries: A critical approach. *European Journal of Operational Research, 49*, 156–178.

Bornstein, C. T., Rosenhead, J., & Vidal, R. V. V. (Eds.). (1990). Operational research in developing countries. *European Journal of Operational Research*, *49*(2), 155–294

del Rosario, E. A., & Rand, G. K. (2010). IFORS: 50 at 50. *Boletín de Estadística e Investigación Operativa, 26*(1), 84–96.

Galvão, R. D. (1988). Operational research in latin America: Historical background and future perspectives. In G. K. Rand (Ed.), *Operational research '87* (pp. 19–31). Amsterdam: North-Holland.

Jaiswal, N.K. (Ed.). (1985). *OR for developing countries*. Operational Research Society of India.

Löss, Z. E. (1981). O Desenvolvimento da Pesquisa Operacional no Brasil (The Development of OR in Brazil), M. Sc. Thesis, COPPE/Federal University of Rio de Janeiro.

Rand, G. K. (2000). IFORS and developing countries. In A. Tuson (Ed.), *Young OR 11: Tutorial & keynote papers* (pp. 75–86). Birmingham: Operational Research Society.

Rosenhead, J. (1993). ICORD '92–International Conference on operational research for development. *OR for Developing Countries Newsletter, 3*(3), 1–4.

Rosenhead, J. (1998). Ahmedabad 6 years on – has IFORS delivered? *OR for Developing Countries Newsletter, 6*(2), 5–8.

Rosenhead, J. (1995). Private communication.

Rosenhead, J., & Tripathy, A. (Eds.). (1996). *Operational research for development*. New Delhi: New Age International Limited.

White, L., Smith, H., & Currie, C. (2011). OR in developing countries: A review. *European Journal of Operational Research, 208*, 1–11.

Development Tool

Software used to facilitate the development of expert systems. The three types of tools are programming languages, shells, and integrated environments.

See

▶ Expert Systems

Devex Pricing

A criterion for selecting the variable entering the basis in the simplex method. Devex pricing chooses the incoming variable with the largest gradient in the space of the initial nonbasic variables. This is contrasted with the usual simplex method entering variable criterion that chooses the incoming variable based on the largest gradient in the space of the current nonbasic variables. The Devex criterion tends to reduce greatly the total number of simplex iteration on large problems.

See

▶ Linear Programming
▶ Simplex Method (Algorithm)

Deviation Variables

Variables used in goal programming models to represent deviation from desired goals or resource target levels.

See

▶ Goal Programming

DFR

Decreasing failure rate.

See

▶ Reliability of Stochastic Systems

Diameter

The maximum distance between any two nodes in a graph.

See

▶ Graph
▶ Graph Theory

Diet Problem

A linear program that determines a diet satisfying specified recommended daily allowance (RDAs) requirements at minimum cost. Stigler's diet problem was one of the first linear-programming problems solved by the simplex method.

See

- ▶ Linear Programming
- ▶ Simplex Method (Algorithm)
- ▶ Stigler's Diet Problem

References

Gass, S. I., & Garille, S. (2001). Stigler's diet problem revisited. *Operations Research, 49*(1), 1–13.

Stigler, G. J. (1945). The cost of subsistence. *Journal of Farm Economics, 27*(2), 303–314.

Differential Games

Gary M. Erickson
University of Washington, Seattle, WA, USA

Introduction

Differential games offer a valuable modeling approach for problems in operations research (OR) and management science (MS). Differential game models are useful because they combine the key aspects of dynamic optimization and game theory. As such, differential game modeling allows the analysis of a broad set of problems that involve decisions by multiple players over a time horizon. After a discussion of the essential concepts of differential games, applications from the literature are reviewed as examples of how differential game methodology has been used to study problems of interest to OR and MS.

Discussion

A differential game is a game with continuous-time dynamics. Two types of variables are involved, state variables and control variables, both of which vary with time. Control variables are managed by the players. State variables are subject to the dynamic influence of the control variables, and evolve according to differential equations. Each player has an objective function that consists of a stream of instantaneous payoffs integrated over a horizon, plus, perhaps, a salvage value if the horizon is finite. The decision problem for each player is to determine a continuous path of control variable values that maximizes the player's objective function, while taking into account what the player knows or anticipates about the decisions of the other players in the game.

Complete information is assumed in a differential game, so that player outcomes given different combinations of player strategies are known to all players, and each player is able to infer correctly the best strategies for the other players. Also, an assumption is typically made that the players are unable to agree to cooperate, and so are engaged in a noncooperative differential game. Further, if the players choose their strategies simultaneously, the appropriate way to determine what strategies the players are likely to adopt is to identify a Nash equilibrium. A Nash equilibrium is a set of player strategies such that each player is unable to improve their outcome, given the strategies of the remaining players. In a Nash equilibrium, no individual player has an incentive to deviate to another strategy.

There are two types of Nash equilibrium that can be derived: open-loop and feedback. Alternative terms for feedback are closed-loop and Markovian (Dockner et al. 2000, p. 59). The two equilibrium types differ in terms of what information is used to develop the players' strategies. In an open-loop Nash equilibrium, the players' strategies are a function of time only, while feedback Nash equilibrium strategies depend on levels of the state variables as well as time. Further, for a differential game with an infinite horizon, and in which time is an explicit factor in the objective functions only through discount factors, it is appropriate to focus on stationary feedback strategies, which depend on levels of the state variables only (Jørgensen and Zaccour 2004, pp. 7–8).

Different methods are typically used to derive the different Nash equilibrium concepts. The maximum principle of optimal control, with Hamiltonians and costate variables, is used to determine open-loop Nash equilibria (Kamien and Schwartz 1991, p. 274). To derive an open-loop equilibrium, a Hamiltonian is created for each player, and necessary conditions produce a system of differential equations that can be solved numerically as a two-point boundary value problem.

In theory, a feedback Nash equilibrium can also be determined using optimal control methods, but the maximum principle is difficult to apply for feedback strategies, since the solution requires that the strategies of the players be known even as they need to be derived. An alternative way to develop feedback Nash equilibrium strategies is through a dynamic programming approach with value functions and Hamilton-Jacobi-Bellman equations (Kamien and Schwartz 1991, p. 276). The Hamilton-Jacobi-Bellman equations form a system of partial differential equations, which for many problems are inherently impossible to solve. For certain problems, though, it is possible to discern an appropriate functional form for the value functions that allows a solution. In particular, for infinite horizon games, it is often possible to derive stationary feedback equilibrium strategies analytically as closed-form functions of the state variables.

An alternative to simultaneous play of strategies is that of Stackelberg games (Dockner et al. 2000, ch.5; Jørgensen and Zaccour 2004, pp. 17–22). Stackelberg games have an alternative information structure, one in which one player takes on a leadership role and makes their strategy choice known before other players choose their strategies. Such a structure can be appropriate for certain problems, such as supply chain management, where coordination may be achieved to benefit of the supply chain overall through one of the members of the supply chain taking a leadership role.

As for Nash equilibria in games with simultaneous play, there are open-loop and feedback Stackelberg equilibria that can be derived. In an open-loop Stackelberg equilibrium with two players (Dockner et al. 2000, pp. 113–134; Jørgensen and Zaccour 2004, pp. 17–20), the Stackelberg leader announces a control path, and, if the Stackelberg follower believes that the leader will stay with the announced control path, the follower will determine their best response control path by solving an optimal control problem with the leader's control path as given. The leader then solves an optimal control problem that incorporates the follower's best response.

For a feedback Stackelberg equilibrium, Basar and Olsder (1995, pp. 416–420) present a feedback Stackelberg solution, which involves instantaneous stagewise Stackelberg leadership, where a stage is an arbitrary combination of time and state variable values. In the development of the feedback Stackelberg solution, stagewise Hamilton-Jacobi-Bellman equations are formed for the leader and the follower, the equation for the follower defining an optimal response and that for the leader incorporating the optimal response of the follower.

The open-loop and feedback equilibrium concepts for both Nash and Stackelberg games can be further examined on the basis of important credibility-related criteria. Dockner et al. (2000, pp. 98–105) and Jørgensen and Zaccour (2004, pp. 15-16) discuss two such criteria, time consistency and subgame perfectness.

A Nash equilibrium is time consistent if at some intermediate point in a differential game, the players choose not to depart from their equilibrium strategies. Dockner et al. (2000, p. 99) and Jørgensen and Zaccour (2004, p. 15) define a subgame that begins at an intermediate time point in the game, and has particular values for the state variables at the time. An equilibrium for the original game "…is time consistent if it is also an equilibrium for any subgame that starts out on the equilibrium state trajectory…" (Jørgensen and Zaccour 2004, p. 15). Both open-loop and feedback Nash equilibria are time consistent. The open-loop Stackelberg equilibrium is not always time consistent, however. As Dockner et al. (2000, pp. 113–134) discuss, an open-loop Stackelberg equilibrium fails to be time consistent in games in which the leader finds it to their benefit to reset its control path at a some point in time after the game has begun.

Subgame perfectness is a stronger condition than time consistency, requiring that an equilibrium also be an equilibrium for any possible subgame, "…not only along the equilibrium state trajectory, but also in any (feasible) position…off this trajectory." (Jørgensen and Zaccour 2004, p. 16). A feedback Nash equilibrium that satisfies the Hamilton-Jacobi-Bellman equations, is by construction subgame perfect. Also, the feedback Stackelberg solution is, according to Basar and Olsder (1995, p. 417), "…strongly time consistent (by definition)", and strong time consistency coincides, at least essentially, with subgame perfectness (Dockner et al. 2000, pp. 106–107).

Differential Game Applications

The differential game framework is designed to model the decisions of multiple decision makers in a continuous-time dynamic context. This framework can be applied to a variety of problem areas of interest and relevance to OR and MS. Furthermore, modeling the passage of time as continuous, rather than discrete, allows the possibility of mathematical, and therefore generalizable, conclusions. This section discusses applications in advertising, pricing, production, and supply chain management.

Advertising

Competitive advertising in the context of dynamics has been especially a popular area of study. Erickson (2003) provides a review. Two particular models of demand evolution have acted as foundations for differential-game applications to advertising. Kimball (1957, pp. 201–202) presents four versions of Lanchester's formulation of the problem of combat, one of which, Model 4,

$$dn_1/dt = k_1 n_2 - k_2 n_1, dn_2/dt = k_2 n_1 - k_1 n_2$$

has become the foundation for what is known as the Lanchester model. Kimball (1957, p. 203) offers the following interpretation of Model 4: "The n_1 and n_2 are then to be interpreted as the numbers of customers for two similar products, while k_1 and k_2 are in essence the amounts of advertising." The Lanchester model in application is generally interpreted in terms of market shares rather than numbers of customers (Erickson 2003, p. 10), so that advertising for a competitor works to attract market share from the competitor's rival.

Vidale and Wolfe (1957) introduce a model of sales evolution for a monopolistic company

$$dS/dt = \beta A(t)(M - S)/M - \lambda S$$

in which $A(t)$ is the advertising rate, S the sales rate, M the maximum sales potential, β an advertising effectiveness coefficient, and λ a sales decay parameter. In the Vidale-Wolfe model, advertising attracts demand from the untapped sales potential, and the sales attracted are subject to decay. Although the Vidale-Wolfe model is defined for a monopolist, it has been adapted for the study of advertising competition.

Many differential-game applications using the Lanchester and Vidale-Wolfe models study open-loop Nash equilibria, since the two models do not readily allow the derivation of subgame-perfect feedback Nash equilibria. Sorger (1989) offers a modification of the Lanchester model that does allow a feedback equilibrium to be derived for duopolistic competitors. Sorger (1989, p. 58) develops a differential game with market-share dynamics

$$\dot{x}(t) = u_1(t)\sqrt{1 - x(t)} - u_2(t)\sqrt{x(t)}, x(0) = x_0.$$

where $\dot{x}(t) = dx/dt$, $x(t)$ is competitor 1's market share, and $u_1(t)$ and $u_2(t)$ are advertising rates for firm's 1 and 2, respectively. The square-root form in the market share equation in the model allows value functions that are linear in the market share state variable, which allows a solution to the Hamilton-Jacobi-Bellman equations for the differential game. Sorger derives both open-loop and feedback equilibria, and finds that the feedback and open-loop equilibria do not coincide.

The Sorger (1989) modification of the Lanchester model allows subgame-perfect feedback Nash equilibria for a duopoly. Feedback equilibria, however, are not achievable in an extension of the Lanchester model to a general oligopoly, in which the number of competitors may exceed two. For an oligopoly, Erickson (2009a, b) provides a modification of the Vidale-Wolfe model that allows the derivation of feedback equilibria. Erickson's (2009a) model has sales dynamics for each oligopolistic competitor i of $n > 2$ total competitors,

$$\dot{s}_i = \beta_i a_i \sqrt{N - \sum_{j=1}^{n} s_j} - \rho_i s_i.$$

In the model, a_i is the advertising rate, s_i the sales rate, N the maximum sales potential, β_i an advertising effectiveness parameter, and ρ_i a sales decay parameter. The expression under the square-root sign represents untapped potential, that is, the maximum sales potential minus the total sales for all n competitors, including competitor i. An instantaneous change in the sales rate for a competitor comes from two sources: (1) the competitor's advertising attracting sales from the untapped potential in square-root form, (2) a decay from the competitor's current sales rate. Erickson (2009b) extends the model to allow multiple brands for each competitor. As for the Sorger (1989) model, the square-root form in the model allows value functions linear in the state variables, so that the Hamilton-Jacobi-Bellman equations can be solved. Both the Sorger (1989) and Erickson (2009a, b) models are related to a monopolistic modification of the Vidale-Wolfe model suggested by Sethi (1983). Erickson (2009a) uses the derived expressions for feedback Nash equilibrium advertising strategies in an empirical study of the U.S. beer market, and Erickson (2009b) empirically applies the multiple-brand model extension to the carbonated soft drink market.

Pricing

Pricing is a primary and challenging task for management. Prices are the source of revenue for the firm, but also affect demand for the firm's products, especially in a competitive setting. The challenge is compounded when dynamics are involved, and prices are expected not to stay at the same levels. This is the case for new products, in particular new durable products, for which demand tends to develop through a diffusion process that is influenced by the price strategies of competing firms.

Bass (1969) provides a diffusion model of first-time adoption of a new durable product that combines innovation and imitation on the part of customers

$$S(T) = (p + qY(T)/m)(m - Y(T)),$$

where $S(T)$ represents current sales at time T and $Y(T)$ cumulative sales, so that $S(T) = dY(T) / dT$. Further, p is an innovation coefficient, q is an imitation coefficient, and m is the total number of customers who will eventually adopt the new product. The Bass (1969) model has been accepted by much of the OR and MS literature as the primary model of new durable product diffusion.

The Bass (1969) model is for a single firm, and does not consider price explicitly. Dockner and Jørgensen (1988) develop a more general framework for new product diffusion, one that includes competition and prices, which they use to study new-product pricing strategies through differential-game analysis. Dockner and Jørgensen (1988, p. 320) offer the general diffusion model specification

$$\dot{x}_i = f^i(x_1, \ldots, x_M, p_1, \ldots, p_M), x_i(0) = x_{i0} \geq 0.$$

In the model, x_i is the cumulative sales volume of competitor $i = 1, 2, \ldots, M$, and the prices $p_1, \ldots, p_M$ of the competitors are assumed to vary with time. To determine their dynamic price strategies, each competitor is assumed to seek to maximize its objective function

$$J^i = \int_0^T e^{-r_i t}(p_i - c_i)f^i dt$$

where unit cost c_i is a nonincreasing function of cumulative sales x_i, as is often the case with new durable products, that unit cost declines with experience. For mathematical tractability reasons, Dockner and Jørgensen (1988) study open-loop Nash equilibria.

Dockner and Jørgensen (1988) derive the necessary conditions for an open-loop Nash equilibrium for their differential game involving the general diffusion model; for further insights, they analyze more specific functional forms. They consider three special cases, competition with price effects only, multiplicative separable price and adoption effects, and adoption effects only with a multiplicative own-price effect.

Production

The management of production quantities and timing is a critical operations function. Dynamics are involved, since production plans may imply that production does not equal customer demand at particular times. This can result in inventories, which need to be carried at a cost, or backlogs, which involve delay in delivery to customers, presumably at a cost to the firm.

Production management can be studied in a competitive context. Eliashberg and Steinberg (1991) consider the dynamic price and production strategies of two competing firms with asymmetric cost structures. As Eliashberg and Steinberg (1991, p. 1453) explain: “The objective of this paper is to gain insight into the dynamic nature of the competitive aspects of the various policies of two firms, one operating at or near capacity, facing a convex production cost, and the other operating significantly below capacity, facing a linear cost structure. The firms are assumed to face a demand surge condition. We will refer to the firm operating at or near capacity as the ‘Production-smoother’ and the firm operating below capacity as the ‘Order-taker.’ ”

Eliashberg and Steinberg (1991) define a differential game in which production levels and prices are control variables for the two competing firms, and pursue an open-loop Nash equilibrium. They derive several propositions regarding the equilibrium policies of the two competitors. A particular finding is that the Production-smoother follows a strategy of first building up inventory, then drawing the inventory down, and finishing a seasonal period by engaging in a policy of carrying zero inventory for a positive interval.

Supply Chain Management

A supply chain involves various independent players—e.g., supplier, manufacturer, wholesaler, retailer—as raw materials become products that are distributed to retail locations where final customers are able to buy them. All players have an economic stake in their position in the supply chain that is affected by the decisions of the other players. The interest of supply chain management is in coordination of the decisions of the players, given the players’ strategic interdependence.

When dynamics are involved, the interdependence of the players in a supply chain can be interpreted as a differential game. A cooperative differential game would produce full coordination. However, since binding agreements among the supply chain players are difficult to establish and maintain, an alternative focus is to consider noncooperative games with coordinating mechanisms.

One mechanism for achieving coordination is through one of the players in the chain becoming the leader. If there are two players in a supply chain, the differential game becomes a leader-follower game in which a Stackelberg equilibrium provides the coordinating solution. A study that considers this approach is Jørgensen et al. (2001), who analyze the advertising and pricing strategies of two players in a marketing channel, a manufacturer and a retailer.

With the differential game that they develop, Jørgensen et al. (2001) derive four different equilibrium solutions: Markovian (feedback) Nash, feedback Stackelberg with the retailer as the Stackelberg leader, feedback Stackelberg with the manufacturer as the leader, and a coordinated channel solution. They give a detailed comparison of the outcomes for the four solutions.

Concluding Remarks

This article outlines the basic concepts of differential games, along with brief descriptions of relevant applications. More in-depth coverage is given in Dockner et al. (2000) and Jørgensen and Zaccour (2004). Differential games provide a powerful modeling framework for the study of the interaction of multiple decision makers in dynamic settings. As the applications illustrate, the understanding of dynamic and game-theoretic OR and MS problems has been advanced through the analysis of differential-game models.

See

- ▶ Advertising
- ▶ Decision Analysis
- ▶ Dynamic Programming
- ▶ Game Theory
- ▶ Marketing
- ▶ Production Management
- ▶ Supply Chain Management

References

Basar, T., & Olsder, G. J. (1995). *Dynamic noncooperative game theory* (2nd ed.). London: Academic Press.

Bass, F. M. (1969). A new product growth model for consumer durables. *Management Science, 15*, 215–227.

Dockner, E., & Jørgensen, S. (1988). Optimal pricing strategies for new products in dynamic oligopolies. *Marketing Science, 7*, 315–334.

Dockner, E., Jørgensen, S., Long, N. V., & Sorger, G. (2000). *Differential games in economics and management science*. Cambridge, UK: Cambridge University Press.

Eliashberg, J., & Steinberg, R. (1991). Competitive strategies for two firms with asymmetric production cost structures. *Management Science, 37*, 1452–1473.

Erickson, G. M. (2003). *Dynamic models of advertising competition* (2nd ed.). Boston/Dordrecht/London: Kluwer Academic Publisher.

Erickson, G. M. (2009a). An oligopoly model of dynamic advertising competition. *European Journal of Operational Research, 197*, 374–388.

Erickson, G. M. (2009b). Advertising competition in a dynamic oligopoly with multiple brands. *Operations Research, 57*, 1106–1113.

Jørgensen, S., Sigué, S.-P., & Zaccour, G. (2001). Stackelberg leadership in a marketing channel. *International Game Theory Review, 3*, 13–26.

Jørgensen, S., & Zaccour, G. (2004). *Differential games in marketing*. Boston/Dordrecht/London: Kluwer Academic Publishers.

Kamien, M. I., & Schwartz, N. L. (1991). *Dynamic optimization: The calculus of variations and optimal control in economics and management*. Amsterdam/New York/London/Tokyo: North-Holland.

Kimball, G. E. (1957). Some industrial applications of military operations research methods. *Operations Research, 5*, 201–204.

Nerlove, M., & Arrow, K. J. (1962). Optimal advertising policy under dynamic conditions. *Economica, 39*, 129–142.

Sethi, S. P. (1983). Deterministic and stochastic optimization of a dynamic advertising model. *Optimal Control Applications and Methods, 4*, 179–184.

Sorger, G. (1989). Competitive dynamic advertising: A modification of the case game. *Journal of Economic Dynamics and Control, 13*, 55–80.

Vidale, M. L., & Wolfe, H. B. (1957). An operations research study of sales response to advertising. *Operations Research, 5*, 370–381.

Diffusion Approximation

A heavy-traffic approximation for queueing systems in which the infinitesimal mean and variance of the underlying process are used to develop a Fokker-Planck diffusion type differential equation which is then typically solved using Laplace transforms.

See

▶ Queueing Theory

Diffusion Process

A continuous-time Markov process on $\mathbb{R}$ or $\mathbb{R}^{n}$ which is analyzed similar to a continuous-time physical diffusion.

Digital Music

Elaine Chew
Queen Mary University of London, London, UK

Introduction

The advent of digital music has enabled scientific approaches to the systematic study, computational modeling, and explanation of human abilities in music perception and cognition, and in music making, which includes the activities of music performance, improvisation, and composition. The move from analog to digital music, and from music stored on a compact disc to music streamed live over the Internet, has brought new engineering challenges, innovation opportunities, and creative outlets. The pervasiveness of computing power and the Internet has changed the ways in which people interact with, and make, music. The research communities at the cusp of music science and engineering came about after the turn of the last millennium, and have been increasing exponentially since. A short list of the communities involved in scholarly pursuits in music science and engineering is provided in Chew (2008).

Impact of Digital Music Research

Science and technology has changed the face of arts and humanities scholarship. Advances in digital music technology have enabled new discoveries by harnessing the computational power of modern computers for music scholarship. For example, the Joyce Hatto scandal, documented in *The Economist* and elsewhere in 2007, in which over 100 CDs released in recent years under her name were in fact the work of other pianists, was unveiled in part because of the machinery available to automatically evaluate and compare recordings of musical works. The technology exists to begin mapping the myriad decisions involved in composing and performing music, and to start charting human creativity. The fact that mathematical models, and by extension operations research (OR) methods, are widely applied in digital music research and practice should come as no surprise, given the historical connections between music, mathematics, and computing.

The music technology industry has emerged as a major economic force. The phenomenal explosion in digital music information has led to the need for new technologies to organize, retrieve, and navigate digital music databases. Examples of major advances in the organizing and retrieval of digital music include Pandora, a personalized Internet radio service that helps people discover new music according to their tastes, and Shazam, a service that helps people identify and locate the music they are hearing. Pandora generates a playlist based on an artist or song entered by the user, and refines future recommendations based on user preference ratings of the songs in that list. Shazam identifies the song and artist, and the precise recording, from a musical excerpt supplied by the user over a device such as an iPhone. In both Pandora and Shazam, the user is offered the opportunity to purchase the song that is playing, or that has been identified, from various vendors. As of 2010, Pandora had 50 million registered users, and more than 1 billion stations, covering 52% of the Internet radio market share. In December 2010, Shazam announced that it has surpassed 100 million users in 200 countries.

Any young or young-at-heart person may be familiar with the music video game, Guitar Hero®, which allows everyone to live the dream of being a rock star in their own living room by pushing colored buttons on the guitar interface in sync with approaching knobs in the video screen. In a few short years, Guitar Hero took over a significant share of the video game market, grossing over two billion dollars by 2009 and leading to it being featured in a South Park television episode. Bands featured in the game — owned and marketed by Activision — experience significant increases in song sales, so much so that major labels vie for their music to be included in new versions of it and in its successor, Rock Bandpt® vie for their music to be included in new versions.

Music Structure

The understanding of music structure is fundamental to computer analysis of music, and a precursor to digital music processing and manipulation. Music consists of organized sounds with perceptible structures in both time and frequency domains. Often, music can be considered to comprise of a sequence of tones, or several concurrent sequences of tones. Each tone has properties such as pitch (the perceived fundamental frequency of the tone), duration, timbre, and loudness. Much of the music that is heard consists of more than a single stream of tones. When hearing multi-tone textures, the ear can segregate the collection of sounds into streams. The most prominent of these streams is often considered to be the melody of the music piece. Structures relating to individual streams as they progress over time are sometimes referred to as horizontal structures. Like language, music streams can be segmented into phrases. Salient tone patterns in music phrases form motifs, short patterns that recur and vary throughout the piece. The varying of these patterns forms the surface structure of the music piece.

Overlapping pitches in the overlay of multiple tone sequences form chords; conversely, one could say that chords consist of the synchronous sounding of two or more pitches. Chords constitute mid-level structure in music. Structures, such as chords, that relate to synchronous sounds or chunks of music over overlapping streams are sometimes referred to as vertical structures. In Western tonal music, the pitches and durations and their ordering generates the perception of pitch stability relative to one another. This pattern of perceived stability is set up as soon as the ear hears as few as only three to four tones in the sequence. The most stable pitch is the name of the key of the tone sequence. The key, in turn, implies adherence to the pitch set of the scale. The pitches in a scale have varying levels of perceived stability, the result of the physics of sound, the physiology of the ear, or exposure to music. The varying of the most stable pitch over time forms the deep structure of the piece.

The structure of a musical piece can also be conceptualized as a sequence of section labels such as AB (binary form), ABA (ternary form), ABACAC′ADA (a sample rondo form), and intro-(verse-chorus)^n-outro (a common popular music form). While some composers, when writing in a particular genre, choose to adopt a particular form for a composition, structure can also emerge from choices made in composition or improvisation to manage a listener's attention.

Sequences of durations, or sequences of inter-onset-intervals, form rhythms. Periodic onsets generate perceived beats, and accent and stress patterns in beat and in rhythm sequences. The periodic accent patterns in beat sequences, in turn, result in meter. For example, there are cyclic patterns of four beats in the march with a strongest-weak-strong-weak accent pattern, whereas each of the four beats in a tango is subdivided into two with a resulting strong-weak-weak-strong-weak-weak-strong-weak accent pattern. Conversely, the meter of a composition often implies a persistent periodic accent pattern. The beat rate charts the tempo of the music: a high beat rate results in fast music, and a low beat rate results in slow music. Like many things in art, it is deviations from the norm that form the core of artistic expression. Thus, a large part of expressive musical performance is the art of systematically varying the tempo, and deviating from an underlying time grid. For example, not playing the beats as notated is essential to playing a convincing swing rhythm. Other important parameters of variation in expressive performance include loudness and timbre.

Structure guides expressive decisions in performance, and expressive performance, in turn, influences structure. For example, a performer may choose to emphasize unusual key changes by slowing down the tempo and dramatically reducing the loudness of the sound produced at the juncture of change. Alternatively, by punctuating the playback of a tone stream with judicially placed accents and pauses, the performer can impute phrase and motivic structure on a music stream.

Music problems can be broadly categorized into the areas of analysis, performance, and composition and improvisation. When the problems are concerned with human abilities in music making and listening, they also touch upon the area of music perception and cognition. It is beyond the scope of this article to give a comprehensive survey of problem formulations and solutions in computational modeling of music. Rather, this article focuses on representative problems in each category and solutions, covering some essential background on music representation and computation.

Computational Music Analysis

The goal of computational music analysis is to automatically abstract structures, such as those described above, from digital music.

Key and Harmony

The determination of key is a problem in the detection of vertical pitch structure. Key finding (a.k.a. tonal induction) can be described as the problem of finding the note on which a music piece is expected to end. The most stable pitch in a key is also the one that is expected to end a piece of music in that key. Key finding is an important step preceding a number of music applications such as automatic music transcription, accompaniment, improvisation, and similarity assessment. While the focus here is key finding, it is worthwhile to mention chord tracking, a related problem for which the solution bears similarities to key finding. A survey of automatic chord analysis algorithms can be found in Mauch (2010).

Key Finding Using Correlation: Key is most often inferred from pitch information. Each pitch can be represented as an integer, according to pitch height. For example, in MIDI (musical instrument digital interface) notation, the pitches A, B♭, B, C in the middle range of the piano keyboard are represented as 57, 58, 59, 60. Pitches repeat on the keyboard, and the twelfth tone above C is C again, one octave higher. Sometimes only the pitch class is of interest, and pitch numbers can be collapsed into pitch classes using modulo arithmetic. If p is a pitch number, then the corresponding pitch class is $p \mod 12$.

Key-finding algorithms tend to match music data with templates representing the prototypical profile for the 24 major and minor keys. A key-finding algorithm by Krumhansl and Schmuckler (described in Krumhansl 1990) compares a vector, $\mathbf{d} = [d_i]$, summarizing total note duration for each of the twelve pitch classes, to experimentally obtained probe tone profiles for each of the major and minor keys, $\mathbf{v}_i = [v_{ij}]$ for $i = 1 \ldots 24$, by calculating their correlation coefficients, $\rho_{\mathbf{d}\mathbf{v}_i}$. Each probe tone profile is generated by playing a short sequence of chords to establish the key context, then having listeners rate (on a scale of 1 to 7) how well a probe tone that is then played fit in the context. The best match key probe tone profile is the one having the highest correlation coefficient with the query vector, i.e.

$$\arg\max_i \rho_{\mathbf{d}\mathbf{v}_i} = \arg\max_i \frac{\sigma_{\mathbf{d}\mathbf{v}_i}}{\sigma_{\mathbf{d}}\sigma_{\mathbf{v}_i}}.$$

Creating Spatial Models: Having a spatial model that mirrors the mental representation of tonal space is something that is of interest not only to cognitive scientists, but also to computational scientists who use these spaces to design algorithms for tonal induction. Kassakian and Wessel (2005) proposed a convex optimization solution for incrementally creating spatial representations of musical entities, such as key and melody, in Euclidean space in such a way as to satisfy a set of dissimilarity measures. Assuming the existing elements to be $\mathbf{r}_i \in \mathbb{R}^n$ and the vector of dissimilarity distances between the new element and existing ones to be $\mathbf{s} = [s_i] \geq 0$, where $i = 1, 2, \ldots, m$. The problem then becomes one of finding

$$\arg\min_{\mathbf{x},\gamma} \sum_{i=1}^{m} (||\mathbf{x} - \mathbf{r}_i|| - \gamma s_i)^2.$$

Using the geometric insight that each $(||\mathbf{x} - \mathbf{r}_i|| - \gamma s_i)$ is the optimal value of $\min_{\mathbf{b}_i} ||\mathbf{x} - \mathbf{b}_i||^2$ for some $\mathbf{b}_i \in \mathbb{R}^n$ inscribed on the ball of radius γs_i around the point $\mathbf{r}_i$, the problem can be re-written as:

$$\begin{aligned} \min_{\mathbf{x},\gamma,\mathbf{b}} \quad & ||\mathbf{J}\mathbf{x} - \mathbf{b}||^2 \\ \text{s.t.} \quad & ||\mathbf{r}_i - \mathbf{b}_i||^2 = \gamma^2 s_i^2, i = 1, 2, \ldots, m \\ \text{where} \quad & \mathbf{b} \equiv [b_1^T, b_2^T, \ldots, b_m^T]^T \in \mathbb{R}^{mn} \\ \text{and} \quad & \mathbf{J} \equiv [I, I, \ldots, I]^T \in \mathbb{R}^{mn \times n} \end{aligned}$$

While the primal problem is not convex, the dual obtained by Lagrangian relaxation is convex, as is the dual of the dual. The authors used a semi-definite programming solver to obtain a solution to the dual of the dual. Because the dual's dual is a relaxation of the primal, they computed a primal feasible solution from the relaxation using a randomized method reported by Goemans and Williamson, and generalized by Nesterov. The problem can also be solved using more conventional gradient descent methods.The resulting key space map generated in this fashion corresponds well to Krumhansl's map created using multi-dimensional scaling (Krumhansl 1990).

Key Finding Using Geometric Spaces: Starting from a model of tonal space that concurs with human perception can be an advantage in the design of computational algorithms for key finding. Observing that the pitch classes in a major key and in a minor key each occupy distinctly shaped compact spaces on the harmonic network or tonnetz, Longuet-Higgins, and Steedman (1971) proposed a shape matching algorithm to determine key from pitch class information.

The tonnetz is a network model for pitch classes where horizontal neighbors are pitch classes whose elements have a fundamental frequency ratio of approximately 2:3 (four major/minor scale steps apart), neighbors on the northeast diagonal have a ratio of approximately 4:5 (two major scale steps apart), and neighbors on the northwest diagonal have a ratio of approximately 5:6 (two minor scale steps apart). The dual graph of the harmonic network connects all triads (three-note chords) sharing two pitches, the transition between which has the property of smooth voice leading. Lewin (1987) lays the foundation for the theory underlying transformations on this space in his treatise on Generalized Intervals and Transformations. Callendar, Quinn, and Tymoczko (Tymoczko 2006; Callender et al. 2008) further generalized these chord transition principles to non-Euclidean space.

The tonnetz is inherently a toroid structure. By rolling up the planar network so that repeating pitch classes line up one on top of another, one gets the pitch class spiral configuration of the harmonic network. Inspired by interior point approaches, Chew (2000) proposed the spiral array model, which uses successive aggregation to generate higher level representations, inside this three-dimensional structure, from their lower level components. For example, if pitch classes were indexed by their positions on the line of fifths, then each pitch classes can be represented as:

$$\mathbf{P}_{k+1} \equiv \mathbf{R} \cdot \mathbf{P}_k + \mathbf{h},$$

$$\text{where } \mathbf{R} = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \mathbf{h} = \begin{bmatrix} 0 \\ 0 \\ h \end{bmatrix}, k \in \mathbb{Z}.$$

The positions of major and minor chords are computed as convex combinations of their component pitches:

$$\mathbf{C}_{M,k} \equiv \omega_1 \cdot \mathbf{P}_k + \omega_2 \cdot \mathbf{P}_{k+1} + \omega_3 \cdot \mathbf{P}_{k+4},$$

and

$$\mathbf{C}_{m,k} \equiv u_1 \cdot \mathbf{P}_k + u_2 \cdot \mathbf{P}_{k+1} + u_3 \cdot \mathbf{P}_{k-3},$$

respectively, where $\omega_1 \geq \omega_2 \geq \omega_3 > 0$, $u_1 \geq u_2 \geq u_3 > 0$, $\sum_{i=1}^{3} \omega_i = 1$, and $\sum_{i=1}^{3} u_i = 1$. Major and minor keys are generated from the weighted average of their defining chords:

$$\begin{aligned}\mathbf{T}_{M,k} &\equiv \omega_1 \cdot \mathbf{C}_{M,k} + \omega_2 \cdot \mathbf{C}_{M,k+1} + \omega_3 \cdot \mathbf{C}_{M,k-1},\\ \mathbf{T}_{m,k} &\equiv \upsilon_1 \cdot \mathbf{C}_{M,k} + \upsilon_2 \cdot [\alpha \cdot \mathbf{C}_{M,k+1} + (1-\alpha) \cdot \mathbf{C}_{m,k+1}]\\ &\quad + \upsilon_3 \cdot [\beta \cdot \mathbf{C}_{m,k-1} + (1-\beta) \cdot \mathbf{C}_{M,k-1}],\end{aligned}$$

where $\omega_1 \geq \omega_2 \geq \omega_3 > 0$, $\upsilon_1 \geq \upsilon_2 \geq \upsilon_3 > 0$, $\sum_{i=1}^{3} \omega_i = 1$, $\sum_{i=1}^{3} \upsilon_i = 1$, and $0 \geq \alpha \geq 1$, $0 \geq \beta \geq 1$. The calibration of the spiral array, finding solutions to the variables that satisfy perceived properties of pitch relations, is a nonlinear constraint satisfaction problem for which the author found near-feasible solutions using a gradient-inspired heuristic.

Given a music sequence of pitches that map to the pitch representations $\{\mathbf{P}_i\}$, with corresponding durations, $\mathbf{d} = [d_i]$, where $i = 1, \ldots, m$, the center of effect of the sequence, $\mathrm{CE} \equiv \sum_{i=1}^{m} d_i \cdot \mathbf{P}_i$. The most plausible key for the sequence is given by the key representation nearest to the center of effect of the sequence:

$$\arg \min_{\mu \in \{M,m\},k} ||\mathrm{CE} - \mathrm{T}_{\mu,k}||.$$

Extensions: The descriptions of key-finding algorithms have focussed on discrete information. It is possible to apply probabilistic approaches using the same representations. For example, Temperley (2007) explores a Bayesian approach to the Krumhansl key-finding framework.

Both Krumhansl's probe tone profile method and Chew's spiral array center of effect generator algorithm have been extended from symbolic to audio key finding. The underlying methodology remains the same. However, when starting from audio, some pre-processing of the signal needs to be done to convert it to pitch class information. Similarly, the key templates may have to be adapted for audio input. Common techniques for extracting frequency information from the signal include the Fast Fourier Transform and the Constant-Q Transform. This step is followed by the mapping of spectral information to pitch class bins, then the key finding algorithm is applied accordingly. While signal-based information tends to be more noisy than discrete symbolic information, much of the noise results from the harmonics of the fundamental frequency of each tone, which tend to be frequencies in the key, and help reinforce and stabilize key identity.

Meter and Rhythm

While historically the modeling of meter and rhythm has not received as much attention as that of key and harmony, the feeling of pulse, and the grouping of events embedded in that pulse, are some of the most visceral responses humans have to music. An overview of symbolic and literal (signal) representations of rhythm can be found in Sethares (2007) and Smith and Honing (2008). In symbolic music, tone onsets are encoded explicitly in the representation. When analyzing audio, a pre-processing step of extracting onset information must first be performed. An overview of onset detection methods is given in Bello et al. (2003).

Meter Induction: The determining of meter can be described as the finding of the periodic accent patterns in the underlying pulse of music. Meter induction, like key finding, is an important step for numerous music applications such as automatic music transcription, generation, and accompaniment. Most algorithms for finding meter apply autocorrelation to find periodicity in the signal, see for example, Gouyon and Dixon (2006). A different computational model for extracting meter from onset information is described in Mazzola's extensive volume on mathematical music theory (Mazzola 2002), and expanded by Volk (2008) to investigate local versus global meters.

The solution method is restated here in a slightly different format. Suppose $\mathbb{N}$ indexes the smallest grid possible to capture all event onsets in a score. And suppose we are interested in pulse layers at onset times of all possible periodicities, $g \in \mathbb{N}$, and offsets, $f = 0, \ldots, i-1$, then a pulse layer might be indexed by $y = \frac{1}{2} \cdot g(g-1) + 1 + f$ and be represented as a vector $\mathbf{p}_y = [p_{yi}]$, where

$$p_{yi} = \begin{cases} 1 \text{ if } i \in \{gk - f : k \in \mathbb{N}\}, \\ 0 \text{ otherwise.} \end{cases}$$

Suppose the onsets in the music are represented as a vector, $\mathbf{o} = [o_i]$, where

$$o_i = \begin{cases} 1 \text{ if an onset occurs on that grid marking, and} \\ 0 \text{ otherwise,} \end{cases}$$

$$p^o_{yi} = \begin{cases} 1 \text{ if } (p_{yi} = 1) \cap (o_i = 1), \text{ and} \\ \text{otherwise.} \end{cases}$$

Effectively, $\mathbf{p}^o_y$ serves as an indicator function for when an onset in the music coincides with a pulse at layer y. Introducing one more variable, let ℓ_{yi} be the span of the longest chain of ones surrounding p^o_{yi}. ℓ_{yi} can be defined recursively as follows:

$$\ell_{yi} = \ell^R_{yi} + \ell^L_{yi},$$

$$\text{where } \ell^R_{yi} = \begin{cases} 0 & \text{if} \quad p^o_{yi} = 0, \\ 1 + \ell_{yi+1} & \text{if} \quad p^o_{yi+1} = 1, \end{cases}$$

$$\ell^L_{yi} = \begin{cases} 0 & \text{if} \quad p^o_{yi} = 0, \\ 1 + \ell_{yi-1} & \text{if} \quad p^o_{yi-1} = 1. \end{cases}$$

The metric weight of an onset at time i is then given by

$$w_i = \sum_y \ell^a_{yi}.$$

The resulting vector, $\mathbf{w}$ gives a profile of the accents and reveals the periodicity in the rhythm. Recall that $y = \frac{1}{2} \cdot g(g-1) + 1 + f$. A variation on this technique (Nestke and Noll 2001) assigns the weight ℓ_{yi} to all points on pulse layer y, i.e. $\forall i = gk - f, k \in \mathbb{N}$.

Genre Classification using Metric Patterns: Periodicity patterns are one of the defining characteristics of dance music, and this feature has been used to classify music into different genres such as tango, rumba, and cha cha (Dixon et al. 2003; Chew et al. 2005). Dixon et al. (2003) uses a set of rules, which can be implemented using decision trees, to categorize the music using tempo and periodicity features. Similar to the key-finding methods, (Chew et al. 2005) uses correlation to compare the metric weight profiles derived from the data to templates for each dance category.

Segmentation in Time

Few pieces of music stay entirely in one key or one rhythmic pattern. Composers generate interest by varying the tonal and rhythmic content of the music over time. Thus, it would be unrealistic to compute only one key or one meter based on available information. A common adaptation of key-finding or meter induction algorithms to allow for changing key or metric identity is to use a sliding window (Shmulevich and Yli-Harja 2000), or an exponential decay function (Chew and François 2005).

The determining of section boundaries is important in music structure analysis, the applications for which include music summarization. Using the key and meter representation frameworks introduced above, it is possible to create a dynamic programming formulation, with an appropriate penalty function for change between two adjacent windows, for assigning boundaries in a piece of music, for example for key as discussed in Temperley (2007). Another method for determining key change is described in Chew (2002), which borrows ideas from statistical quality control. In large structure analysis, it is often useful to be able to label sections (for example, as chorus or verse in popular songs). Toward this end, Levy and Sandler (2008) have applied a number of clustering techniques to audio features extracted from music signal.

Melody

Melody represents the horizontal structure of music. Apart from the straightforward event string representation of melody, melody can also be decomposed into building blocks and represented as grammar trees, as prescribed by Lerdahl and Jackendoff (1983).

Similarity Assessment: Quantifying the similarity between two melodies is important for music information retrieval. Typke et al. (2003) describe the use of the Earth Mover's Distance (EMD) to quantify melodic similarity. Represent each melody as weighted points in pitch-time space, for example, melody $\mathrm{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\}$ and melody $\mathrm{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ with respective weights, $\omega_i, u_j \in \mathbb{R}^+ \cup \{0\}$, where $i = 1, \ldots m$ and $j = 1, \ldots n$. The similarity measure between the two melodies is the EMD, the minimum cost flow to transform one melody into another by moving weight from one point in A to one point in B, where the cost is the weight moved times the distance traveled.

Suppose $W = \sum_{i=1}^{m} w_i$ and $U = \sum_{i=1}^{n}$, and f_{ij} is the flow of weight from a_i to b_j over the distance d_{ij}. The problem can thus be stated as:

$$
\begin{aligned}
\min \quad & \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} d_{ij} \\
\text{s.t.} \quad & \sum_{j=1}^{n} f_{ij} \le w_i, i = 1, \ldots, m, \\
& \sum_{i=1}^{m} f_{ij} \le u_j, j = 1, \ldots, n, \\
& \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min(W, U), \\
& f_{ij} \ge 0, \; i = 1, \ldots, m, \; j = 1, \ldots, n,
\end{aligned}
$$

which can be solved using linear programming, and

$$
\mathrm{EMD}(\mathrm{A}, \mathrm{B}) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}^{*} d_{ij}}{\min(W, U)}.
$$

Stream Segregation: A number of approaches have been proposed to tackle the problem of automatically separating a polyphonic (multi-line) music texture into its component voices. An example might be to separate a fugue by Johann Sebastian Bach into its four parts. A randomized local search method to optimize a parametric cost function that penalizes undesirable traits in a voice-separated solution was proposed by Kilian and Hoos (2002). Chew and Wu (2004) proposed a contig-mapping approach to first break a piece of music into contigs with overlapping fragments of music. Then, exploiting perceptual principles such as voices tend not to cross in maximal voice contigs, the algorithm re-connects the fragments in neighboring contigs using distance minimization.

Composition and Improvisation

The use of mathematical models in music composition has become an active area for musical innovation since Xenakis (2001), who used stochastic processes, probabilistic models, and game theory to guide his compositions. With widespread access to computing to help solve music composition mathematical problems, computer-assisted composition has emerged as a useful tool to help composers create new music, as well as an important area of digital music research.

Constraints

A number of music composition problems can be naturally described as constraint satisfaction problems (CSPs). Solution methods for CSPs include combinatorial optimization and local search techniques such as Tabu search, simulated annealing, and genetic algorithms.

Truchet and Codognet (2004) list fourteen examples of musical CSPs and propose to apply a heuristic adaptive search technique to solve the CSPs. An example of a compositional CSP is as follows: Given a sequence of chords, suppose the composer is interested in finding an ordering of the sequence such that two adjacent chords have the maximal number of common tones. If the chords were represented as nodes, and the distance between any two nodes is the number of common tones, then the problem of interest takes the form of the Traveling Salesman Problem. Every chord must be visited once, and the desired solution must minimize $(-1) \times$ distance.

Related to this is the classic problem of providing harmonization for a given melody. The most widely used solution method for generating a score from a melody is via constraints, and a variety of approaches and results are reviewed in Pachet and Roy (2004).

Markov Chains and Other Network Models

The use of Markov chains (MCs) forms another solution method that is commonly used in the generating of music. In the case of MCs, the probabilities are estimated from existing data, and used to generate music in the style of the training data set. Farbood and Schoner (2001) use MCs to generate music in the style of Palestrina. Using the tonnetz as scaffolding to reduce the search space, Chuan and Chew (2007a) use MCs to generate style-specific accompaniment for melodies given only a few examples. MCs are excellent models for imitating local structure, but lack high level structure knowledge to guide the shaping of a composition. To remedy this deficiency, researchers have considered computer systems that create the local surface structure while relegating higher level structural control to humans.

In Pachet's Continuator, the system builds prefix trees from music data, weights each possible continuation with a probability estimated from the data, and uses the resulting MC to generate music in dialog with a human musician. Extensions of the basic MC model consider hierarchical representations and ways of imputing rhythmic structure to the resulting music. Assayag and Dubnov (2004) describe an alternate approach using factor oracles. The suffix links in the resulting network model is assigned transition probabilities that causes the original music material to be recombined smoothly. Using the same factor oracle approach, François et al. (2010) created Mimi4x, an installation that allows users to make high-level structural improvisation decisions while the computer manages surface details on four improvising systems.

Expressive Music Performance

Music is rarely performed as notated. The score is an incomplete description of the experience of a music piece, and leaves much to interpretation by a performer. In expressive music performance, a performer manipulates parameters such as tempo, loudness, and articulation for expressive or interpretive ends, and to guide the listener's perception of groupings and meter. The expressive devices in the performance of music is sometimes called musical prosody. See Palmer and Hutchins (2006) for a definition and review of research on musical prosody. The extraction of performance parameters can be viewed as the continuous monitoring of expressive features such as tempo and loudness over time.

Representation

Tempo and loudness are two important features of music performance. Suppose the list of onsets in the performed music are $O = \{o_0, o_1, \ldots, o_n\}$. Then the inter-onset-interval at time i is $IOI_i = o_i - o_{i-1}$. If a listener sat and tapped along to the beat of the music, then the list of beat onsets might be $B = \{b_0, b_1, \ldots, b_n\}$. The interbeat-interval would be $IBI_i = b_i - b_{i-1}$, and the instantaneous tempo would be $T_i = \frac{1}{IBI_i}$. Often, some smoothing is desired, and one would report a moving average for the smoothed tempo. Sometimes, the acceleration is desired, where $a_i = \Delta T_i = T_i - T_{i-1}$. A number of models for deriving loudness from the signal exist, many of which have been implemented in Matlab. Timmers (2005) surveys some ways of measuring tempo and loudness in musical performance and of comparing them across performances.

Using the tempo-loudness representation proposed by Langner and Goebl, Dixon et al. (2002) created a computer system for for real-time visualization of performance parameters in the Performance Worm. The exploration of Langner's tempo-loudness space for performance analysis led to its use for performance synthesis in the Air Worm (Dixon et al. 2005).

In the spirit of annotations of speech prosody, Raphael proposed a series of markup symbols for expressing musical flow (Raphael 2009). The symbols consist of

$$\{l^-, l^\times, l^+, l^\rightarrow, l^\leftarrow, l^*\}.$$

$\{l^-, l^\times, l^+\}$ denote a sense of arrival, where l^- is a direct and assertive stress, $l^\times$ is a soft landing that relaxes upon arrival, and l^+ is an arrival whose momentum continues to carry forward into the future. $\{l^\rightarrow, l^*\}$ mark notes that continue to move forward toward a future goal, $l^\rightarrow$ is a passing tone and l^* is a passing stress, and $\{l^\leftarrow\}$ denotes a pulling back movement. Because it is hard to determine the exact sets of tempo and loudness parameters, and more locally, the exact amounts of delay or anticipating of an onset, that lead to these flow sensations, Raphael uses a hidden Markov model (HMM) to estimate the most likely hidden variables to have given rise to the observed prosodic annotation.

Phrases

In expressive performance, performers indicate phrase groupings by varying tempo (accelerate and decelerate at beginnings and ends of phrases) and/or loudness (crescendo and decrescendo at beginnings and ends of phrases). As a result, this aspect of a performer's interpretations can be directly inferred from tempo and loudness data. For example, Chuan and Chew (2007b) propose a dynamic programming (DP) method for automatic extraction of phrases. The authors test a model that fits a series of quadric curves (first modeled by polynomials of degree two, then by a series of quadratic splines) to the tempo time series. The best fit curve is found using quadratic programming, and the phrase boundaries are determined using DP.

Alignment

A common use of DP in music processing is in the alignment of music sequences that may be in the same or different format. Arifi et al. (2004) reviews the state of the art, and describes an algorithm for aligning music sequences in two of three possible formats – score, Musical Instrument Digital Interface (MIDI), and pulse-code modulation (PCM) audio format.

Assuming the two sequences are the score, $\mathbf{s} = [s_i]$, and a PCM representation of the audio performance, $\mathbf{p} = [p_j]$. The first task is to generate a cost matrix for aligning any point, s_i, in the score with any point, p_j, in the PCM audio. In Arifi et al. (2004), the distance minimization step is embedded in the cost matrix. Suppose the cost matrix is represented by $\mathbf{C} = [c_{ij}]$, each element of which expresses the cost minimizing SP-match for $[s_1, s_2, \ldots, s_i]$ and $[p_1, p_2, \ldots, p_j]$, i.e.

$$c_ij = \min\left\{c_{i,j-1}, c_{i-1,j}, c_{i-1,j-1}, d_{ij}^{SP}\right\}.$$

Then, the algorithm for synchronizing the two streams is as follows:

```
SCORE-PCM-SYNCHRONIZATION($\mathbf{C}$, $\mathbf{s}$, $\mathbf{p}$)
   $i$ = length(s), $j = p$, SP-Match = 0
   while ($i > 0$) and ($j > 0$)
       do if $c[i, j] = c[i, j-1]$
                 then $j = j - 1$
             else if $c[i, j] = c[i-1, j]$
                 then $i = i - 1$
           else SP-Match = SP-Match ∪
                  $\{(i, j)\}$, $i = i-1$,  $j = j-1$
   return SP-Match
```

Dixon and Widmer (2005) introduced MATCH, a tool chest for efficient alignment of two time series using variations on the classic dynamic time warping (DTW) algorithm. Niedermayer and Widmer (2010) proposed a multi-pass algorithm that uses anchor notes (notes for which the alignment confidence is high) to correct inexact matches.

Concluding Remarks

Digital music research has rapidly evolved with computing advances and the increasing possibilities for connections between music and computing. The latest advances in the field are reported in the annual *Proceedings of the International Conference on Music Information Retrieval*, *Proceedings of the Sound and Music Computing Conference*, and the *Proceedings of the International Symposium on Computer Music Modeling and Retrieval*, the biennial *Proceedings of the International Conference on Mathematics and Computation in Music*, and the occasional *Proceedings of the International Conference on Music and Artificial Intelligence*. They can also be found in the traditional conferences of the multimedia, databases, human computer interaction, and audio signal processing communities. The archival journals include the *Computer Music Journal*, the *Journal of New Music Research*, and the *Journal of Mathematics and Music*.

There exist close ties between digital music research and the fields of music perception and cognition and computer music (which places greater emphasis on the creating of music), and the community of researchers interested in interfaces for musical expression. Work that overlaps with these other areas can be found in the biennial *Proceedings of the International Conference on Music Perception and Cognition*, and the annual *Proceedings of the International Computer Music Conference* and *Proceedings of the International Conference on New Interfaces for Musical Expression*.

See

- ▶ Constraint Programming
- ▶ Dynamic Programming
- ▶ Linear Programming
- ▶ Markov Chains
- ▶ Mathematical Programming

References

Arifi, V., Clausen, M., Kurth, F., & Müller, M. (2004). Score-to-PCM music synchronization based on extracted score parameters. In *Proceedings of the international symposium on computer music modeling and retrieval*, Esbjerg, Denmark, pp. 193–210.

Assayag, G., & Dubnov, S. (2004). Using factor oracles for machine improvisation. *Soft Computing, 8*(9), 604–610.

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davis, M., & Sandler, M. B. (2003). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing, 13*(5), 1035–1047.

Callender, C., Quinn, I., & Tymoczko, D. (2008). Generalized voice-leading spaces. *Science, 320*(5874), 346–348.

Chew, E. (2000). *Towards a mathematical model of tonality* (Ph.D. dissertation, MIT Press, Cambridge, MA).

Chew, E. (2002). The spiral array: An algorithm for determining key boundaries. In *Music and artificial intelligence – Second international conference*. Springer LNCS/LNAI, Vol. 2445, pp. 18–31.

Chew, E. (2008). Math & music – The perfect match. *OR/MS Today, 35*(3), 26–31.

Chew, E., & François, A. R. J. (2005). Interactive multi-scale visualizations of tonal evolution in MuSA.RT Opus 2. *ACM Computers in Entertainment, 3*(4), 1–16.

Chew, E., & Wu, X. (2004). Separating voices in polyphonic music: A contig mapping approach. In *Proceedings of the international symposium on computer music modeling and retrieval*, Vol. 2, Esbjerg, Denmark, pp. 1–20.

Chew, E., Volk, A., & Lee, C.-Y. (2005). Dance music classification using inner metric analysis: A computational approach and case study using 101 Latin American Dances and National Anthems. In B. L. Golden, S. Raghavan, & E. A. Wasil (Eds.), *The next wave in computing, optimization, and decision technologies: Operations research computer science interfaces series*, New York: Springer, (Vol. 29, pp. 355–370).

Chuan, C. -H., & Chew, E. (2007a). A hybrid system for automatic generation of style-specific accompaniment. In *Proceedings of the international joint workshop on computer creativity*, London, UK, p. 4.

Chuan, C. -H., & Chew, E. (2007b). A dynamic programming approach to the extraction of phrase boundaries from tempo variations in expressive performances. In *Proceedings of the international conference on music information retrieval*, Vienna, Austria, p. 8.

Dixon, S., & Widmer, G. (2005). MATCH: A music alignment tool chest. In *Proceedings of the international conference on music information retrieval*, London, UK, pp. 492–497.

Dixon, S., Goebl, W., & Widmer, G. (2002). The performance worm: Real time visualisation of expression based on Langner's tempo-loudness animation. In *Proceedings of the international computer music conference*, Göteborg, Sweden, pp. 361–364.

Dixon, S., Goebl, W., & Widmer, G. (2005). The 'air worm': An interface for real-time manipulation of expressive music performance. In *Proceedings of the international computer music conference*, Barcelona, Spain, pp. 614–617.

Dixon, S., Pampalk, E., & Widmer, G. (2003). Classification of dance music by periodicity patterns. In *Proceedings of the international conference on music information retrieval*, Baltimore, Maryland.

Farbood, M., & Schoner, B. (2001). Analysis and synthesis of Palestrina-style counterpoint using Markov chains. In *Proceedings of the international computer music conference* Havana, Cuba, p. 27.

François, A. R. J., Schankler, I., & Chew, E. (2010). Mimi4x: An interactive audio-visual installation for high-level structural improvisation. In *Proceedings of the international conference on multimedia and expo*, Singapore, pp. 1618–1623.

Gouyon, F., Dixon, S. Computational Rhythm Description. 2006. Tutorial on computational rhythm description. *In International Conference of Music Information Retrieval*.

Kassakian, P., & Wessel, D. (2005). Optimal positioning in low-dimensional control spaces using convex optimization. In *Proceedings of the international computer music conference*, Barcelona, Spain, Vol. 31, pp. 379–382.

Kilian, J., & Hoos, H. (2002). Voice separation: A local optimisation approach. In *Proceedings of the international conference on music information retrieval*, Paris, France, Vol. 3, pp. 39–46.

Krumhansl, C. L. (1990). *Cognitive foundations of musical pitch*. New York: Oxford University Press.

Lerdahl, F., & Jackendoff, R. A. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

Levy, M., & Sandler, M. (2008). Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech, and Language Processing, 16*(2), 318–326.

Lewin, D. (1987). *Generalized musical intervals and transformations*. New Haven, CT: Yale University Press.

Longuet-Higgins, H. C., & Steedman, M. J. (1971). On interpreting Bach. *Machine Intelligence, 6*, 221–241.

Mauch, M. (2010). *Automatic chord transcription from audio using computational models of musical context* (Ph.D. dissertation, Queen Mary University of London, UK).

Mazzola, G. (2002). *The topos of music, geometric logic of concepts, theory, and performance*. Basel: Birkhäuser.

Nestke, A., & Noll, T. (2001). Inner metric analysis. In J. Haluska (Ed.), *Music and mathematics* (pp. 91–111). Bratislava: Tatra Mountains Mathematical Publications.

Niedermayer, B., & Widmer, G. (2010). A multi-pass algorithm for accurate audio-to-score alignment. In *Proceedings of the international conference on music information retrieval,* Utrecht, Netherlands.

Pachet, F. (2003). The continuator: Musical interaction with style. *Journal of New Music Research, 32*(3), 333–341.

Pachet, F., & Roy, P. (2004). Musical harmonization with constraints: A survey. *Constraints, 6*(1), 7–19.

Palmer, C., & Hutchins, S. (2006). What is musical prosody? In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 46, pp. 245–278). Amsterdam: Elsevier Press.

Raphael, C. (2009). Representation and synthesis of melodic expression. In *Proceedings of the international joint conference on AI*, Pasadena, California, Vol. 21, pp. 1475–1480.

Sethares, W. A. (2007). *Rhythm and transforms*. London: Springer-Verlag.

Shmulevich, I., & Yli-Harja, O. (2000). Localized key finding: Algorithms and applications. *Music Perception, 17*(4), 531–544.

Smith, L. M., & Honing, H. (2008). Time-frequency representation of musical rhythm by continuous wavelets. *Journal of Mathematics and Music, 2*(2), 81–97.

Temperley, D. (2001). *The cognition of basic musical structures*. Cambridge, MA: MIT Press.

Temperley, D. (2007). *Music and probability*. Cambridge, MA: MIT Press.

Timmers, R. (2005). Predicting the similarity between expressive performances of music from measurements of tempo and dynamics. *Journal of the Acoustical Society of America, 117*(1), 391–399.

Truchet, C., & Codognet, P. (2004). Musical constraint satisfaction problems solved with adaptive search. *Soft Computing, 8*, 633–640.

Tymoczko, D. (2006). The geometry of musical chord. *Science, 313*(5783), 72–74.

Typke, R., Giannopoulos, P., Veltkamp, R. C., Wiering, F., & van Oostrum, R. (2003). Using transportation distances for measuring melodic similarity. In *Proceedings of the international symposium on computer music modeling and retrieval*, Montpellier, France, pp. 107–114.

Volk, A. (2008). Persistence and change: Local and global components of Metre induction using inner metric analysis. *Journal of Mathematics and Music, 2*(2), 99–115.

Xenakis, I. (2001). *Formalized music: Thought and mathematics in music*. Hillsdale, NY: Pendragon Press.

Digraph

A graph all of whose edges have a designated one-way direction.

See

▶ Graph Theory

Dijkstra's Algorithm

A method for finding shortest paths (routes) in a network. The algorithm is a node labeling, greedy algorithm. It assumes that the distance c_{ij} between any pair of nodes i and j is nonnegative. The labels have two components $\{d(i), p\}$, where $d(i)$ is an upper bound on the shortest path length from the source (home) node s to node i, and p is the node preceding node i in the shortest path to node i. The algorithmic steps for finding the shortest paths from s to all other nodes in the network are as follows:

Step 1. Assign a number $d(i)$ to each node i to denote the tentative (upper bound) length of the shortest path from s to i that uses only labeled nodes as intermediate nodes. Initially, set $d(s) = 0$ and $d(i) = \infty$ for all $i \neq s$. Let y denote the last node labeled. Give node s the label $\{0, -\}$ and let $y = s$.

Step 2. For each unlabeled node i, redefine $d(i)$ as follows:

$d(i) = \min\{d(i), d(y) + c_{yi})\}$. If $d(i) = \infty$ for all unlabeled vertices i, then stop, as no path exists from s to any unlabeled node i with the smallest value of $d(i)$. Also, in the label, let p denote the node from which the arc that determined the minimum $d(i)$ came from. Let $y = i$.

Step 3. If all nodes have been labeled, stop, as the unique path of labels $\{d(i), p\}$ from s to i is a shortest path from s to i for all vertices i. Otherwise, return to *Step 2*.

See

▶ Greedy Algorithm
▶ Minimum-Cost Network-Flow Problem
▶ Network Optimization
▶ Vehicle Routing

Directed Graph

▶ Digraph

Direction of a Set

A vector $\boldsymbol{d}$ is a direction of a convex set if for every point $\boldsymbol{x}$ of the set, the ray $(\boldsymbol{x} + \lambda \boldsymbol{d})$, $\lambda \geq 0$, belongs to the set. If the set is bounded, it has no directions.

See

▶ Convex Set

Directional Derivative

A rate of change at a given point in a given direction of the value function of a optimization problem as a function of problem parameters.

See

▶ Nonlinear Programming

Disaster Management: Planning and Logistics

Gina M. Galindo Pacheco[1,2] and Rajan Batta[1]
[1]University at Buffalo, The State University of New York, Buffalo, NY, USA
[2]Universidad del Norte, Barranquilla, Colombia

Introduction

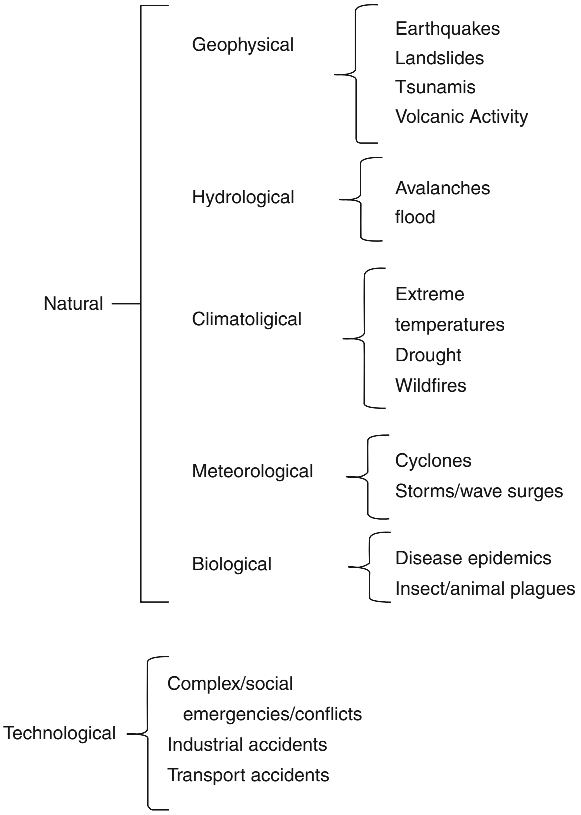
Due to significant losses of life, as well as extremely high economic costs, the prevention and improvement of disaster response has been a continuing area of research. OR analysts have been in the forefront of such work and have made significant contributions that have helped to mitigate the impact of disasters. This article reviews some of the basic concepts related to disaster management (DM) and summarizes many of the topics that have been addressed.

The presentation is as follows: section one reviews disaster definitions and types; section two focuses on the role of DM, the concepts associated, and the stages that are traditionally identified within DM; section three discusses the role of the planning process; section four reviews the related logistics issues; section five discusses DM topics based on a sample of work from the period 2005-2010; and the last section presents a summary and concluding remarks.

Definition of Disaster

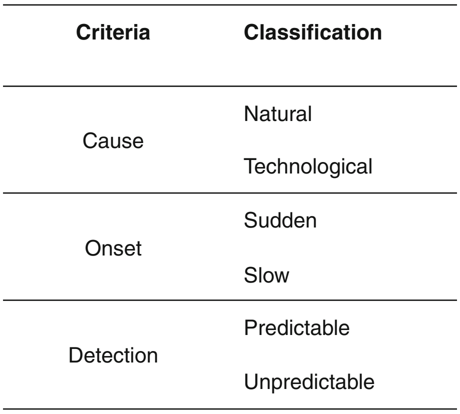
According to the International Federation of Red Cross and Red Crescent Societies (IFRC), a disaster is a sudden event that causes disruption of the normal functioning of a community; causes human, economic, and environmental losses; and generates requirements that exceeds the capacity of response using available resources.

Losses due to disasters may be of the order of thousands of lives and billions of dollars. Kunkel, Pielke, and Changnon (1999) give some statistics about human and economic losses due to weather and climate extremes in the U.S. They estimate that between 1986 and 1995 there was an annual mean loss of 96 lives due to floods and 20 due to hurricanes. In the same period, the annual mean of economic losses was $6.2 billion for hurricanes. In 2005, the National Hurricane Center estimated that hurricane Katrina left a total of 1,200 reported casualties, with a total damage cost of $81 billion. Man-made disasters can also have drastic consequences if they are purposely planned. For example, according to the National Commission on Terrorist Attacks upon United States, more than 2,981 people died in the attacks of 9/11. Even though environmental disasters typically do not involve many human casualties, they do cause great ecological damages, e.g., the Gulf of Mexico oil spill that affected thousands of turtles, birds, and mammals, as reported by the International Disaster Database Web site (in addition to the considerable monetary loss for British Petroleum). The types of natural and man-made disasters are listed in Fig. 1.



Disaster Management: Planning and Logistics, Fig. 1 Types of disasters

This classification derives partly from IFRC, Alexander (2002), and Van Wassenhove (2006).



Criteria	Classification
Cause	Natural
	Technological
Onset	Sudden
	Slow
Detection	Predictable
	Unpredictable

Disaster Management: Planning and Logistics, Fig. 2 Disaster classification

Natural disasters may be grouped into predictable ones, such as hurricanes, and unpredictable events, such as earthquakes. Data about predictable disasters are not deterministic, but some information about the time and place of such disasters is available. Such disasters can also be classified with respect to their time of onset. Tornadoes happen suddenly and last for a short period of time, while events such as pandemics may go from a few days to several months. These classifications become important at the time of planning and responding: for predictable disasters actions like evacuation or prepositioning of supplies are possible, while for unpredictable ones, such actions are not possible alternatives; for very short-term disasters it is easier to estimate the amount of resources needed to overcome the situation, where for long-term disasters this is a more difficult task. Figure 2 summarizes these classifications.

Role of Disaster Management

According to the IFRC, the management of resources and responsibilities to respond to humanitarian needs after an emergency is known as Disaster Management (DM).

DM can be viewed as including the strategic, tactical, and operational activities, as well as the personnel and technologies involved at various stages of a disaster situation for the purpose of mitigating its possible consequences (Lettieri et al. 2009).

The different stages involved in DM are classified as mitigation, preparedness, response, and recovery (McLoughlin 1985). Miller, Engemann, and Yager (2006) provide a detailed explanation of the four DM stages. Each of these stages is briefly discussed below with respect to a flood disaster.

Mitigation consists of those activities that help to reduce the long-term risk of the occurrence of a disaster or its consequences. For a flood scenario, mitigation would involve not building on low lands, and creating barriers along rivers or ponds. Preparedness refers to planning operational activities to respond to a disaster—creating shelters, prepositioning supplies, and evacuating people from most dangerous locations is a way in which preparedness may be applied for a flood setting. The response stage includes actions that correspond to those performed upon the occurrence of the disaster to help affected people to overcome their needs of essential resources or getting them out from danger e.g., delivering supplies and rescuing people. The recovery phase involves short and long-term activities to restore normal functioning of the community, as well as repairing roads and buildings.

The recovery phase should be designed in such a way that it contributes to mitigation efforts. For the flood example, rebuilt houses should not be located in lands known to be highly exposed to floods. This is how DM could be viewed as a cycle created by the link of mitigation and recovery activities. In general, the different stages of DM require a previous planning process to coordinate all the ulterior actions that would be performed. In addition, a logistic process is involved mainly, but not exclusively, for the preparedness and response phases.

Disaster Management and Planning Process

The Oxford English Dictionary defines the verb “to plan” as meaning “to devise, contrive, or formulate (something to be done, or some action or proceeding to be carried out.)” For DM, Alexander (2002) distinguishes emergency planning in terms of long and short-term. The former gives the context for the latter. It involves forecasting, warning, educating, and training people for the event of a disaster. It includes the study of patterns to predict the possible time and place at which a disaster could occur. Seasonal natural disasters, such as tropical storms in the Caribbean, are examples. The concept of long-term planning is related to the definition of emergency planning given by Perry and Lindell (2003) for whom emergency planning focuses on the two objectives of hazard assessment and risk reduction. The purpose of short-term planning is to guarantee the prompt deployment of resources where and when needed.

Alexander (2002) describes an outline of the methodological components of an emergency plan and includes a generic emergency planning model. The planning process may be summarized as gathering information, managing and analyzing it, extracting some conclusions and actions to be developed, and communicating the resulting plan to the staff involved.

Disaster Management and Logistics

Several definitions are used for the term logistics. Van Wassenhove (2006) gives a brief and illustrative review of some of these definitions as applied to business, military, and humanitarian DM logistics. In summary, logistics, when applied to DM, is referred to as the storage and deployment of resources and information, as well as the mobilization of people in an effective way to reduce the impact of the disaster. Kovács and Spens (2007) and Van Wassenhove (2006) reflect upon the comparison between business and humanitarian DM logistics. However, despite the differences, business and humanitarian logistics are intrinsically related and they both refer to a process that includes planning, distribution and transportation, storage, location and supply chain management (SCM).

In what follows, some common problems related to planning and logistics in DM and OR are discussed.

OR and DM

A survey of OR research related to DM since 2005 was conducted. A total of 222 items in journals, books, book chapters, and conference papers were reviewed. A finding was that topics of planning and logistics in DM attracted most of the attention. For planning, the most common topics were evacuation and risk analysis. General humanitarian logistics was a topic addressed in terms of (i) transportation, (ii) inventory, (iii) location analysis, and (iv) humanitarian logistics (in general). Material from (i) to (iii) are referred to as specific activities inside the concept of logistics, while that from (iv) considers logistics as a whole or that combines different aspects of humanitarian logistics. Other topics of logistics are reviewed separately because they constitute a widely studied topic as is the case for transportation that includes research on routing, traffic and network management.

Even though there were many other topics of OR interest in the reviewed research such as demand forecast, business continuity, and hospital capacity, the topics mentioned earlier represent the main streams that were studied. In the following sections, the topics will be discussed separately focusing on the relationship to DM phases, methodologies, objectives, and real-life applications.

Evacuation: The major way for reducing the potential population affected by a disaster is evacuation. An evacuation typically involves mobilizing people from endangered zones to safer ones, which includes routing strategies and preparation of shelters, among other activities. This process is mostly associated with the preparedness phase of DM, and, therefore, to the planning processes. However, some related work for real-time decisions may be linked to the response phase (Chiu and Zheng 2007). For predictable disasters, it is possible to develop evacuation plans to be performed before the disaster strikes; no pre-disaster-evacuation planning is possible for unpredictable disasters.

The most common objective in evacuation research was minimizing the evacuation time of the total affected population (Chen and Zhan 2008). Other objectives included maximizing the total number of evacuees during a given evacuation time (Miller-Hooks and Sorrel 2008), maximizing the minimum probability of reaching an exit for any evacuee (Opasanon and Miller-Hooks 2009), and minimizing total system travel time (Chiu et al. 2007). Some studies considered multiple objectives. In Saadatseresht, Mansourian, and Taleai (2009) the objectives were to minimize travel distance, evacuation time, and overload capacity of safe areas. Stepanov and Smith (2009) provide a critique of performance measures for evacuation that include clearance time, total traveled distance, and blocking probabilities.

Simulation was the most used method to solve evacuation problems. Bonabeau, (2002) and

Chen and Zhan, (2008) used agent-based simulation—the process in which entities termed autonomous agents assess their situations and make decisions according to a set of rules(say something about validation). Other studies developed multi-level models (Liu et al. 2006), queue analysis (Stepanov and Smith 2009), mixed integer linear programming (Sayyady and Eksioglu 2010); others used Cell Transmission Models (Chiu et al. 2007), and genetic algorithms (Miller-Hooks and Sorrel 2008).

Most of the studies employed real data to validate their results. For example, Chen, Meaker and Zhan (2006) developed a simulation model for evacuating the Florida Keys under a hurricane setting. They considered two questions: one related to the time for evacuating the total population, while the other considered how many residents would need to be accommodated if evacuation routes were impassable. The authors used a previous study as a reference for comparing the results of their model. However, no validation based on real evacuation times is reported.

Risk Analysis: DM **r**isk analysis is mainly concerned with quantifying the risk of the occurrence of an undesirable event, as well as developing measures to diminish the impact of a disaster. Risk analysis is mainly a planning tool related to the mitigation. The objectives of the DM risk analysis studies were forecasting, infrastructure planning and design, vulnerability, and analysis of uncertainty, as discussed next.

In relation to forecasting, Hu (2010) uses a Bayesian approach to analyze flood frequencies. Infrastructure planning and design based on risk analysis refers in some cases to making the infrastructure (buildings, networks, supply chains, etc.) more resistant to disaster damages and disruptions, and to building physical barriers or diversions to diminish the impact of a disaster on an endangered community. Snyder et al. (2006) reviewed several models for designing supply chains resilient to disruptions. These models considered costs from the business point of view, with objectives, in most of the cases, being the minimization of the expected or the worst case cost. Li, Huang and Nie (2007) used a model for flood diversion planning under uncertainty where, among the objectives considered, was the minimization of risk of system disruption. Vulnerability relates to the way in which current systems are affected by damages. Matisziw and Murray (2009) maximized system flow for a disrupted network. Barker and Haimes (2009) focused on a sensitivity analysis of extreme consequences due to uncertainties on the parameters, and Xu, Booij and Tong (2010) analyzed the sources of uncertainty in statistical modeling.

Probability and statistics were the main methods used to analyze risk analysis. In the case of Li, Huang, and Nie (2007) the authors used a methodology that combines fuzzy sets and stochastic programming. Another example in which fuzzy sets have been incorporated into risk analysis is given by Huang and Ruan (2008). In this DM area, even though some researchers used real data to develop numerical examples, complete case studies were rare.

Transportation: Transportation problems typically deal with routing, vehicle schedule, traffic, and network management. The problems may be to transport goods to provide relief supplies, evacuate people from endangered areas, or movement of resources such as medical staff to areas where their services are required.

For transportation analyses, as applied to DM, there are a wide variety of objectives related to the efficiency of delivery times. Campbell, Vandenbussche, and Hermann (2008) considered two objectives for minimizing the arrival times of relief to demand points. Similarly, Yuan and Wang (2008) minimized the total travel time through a path selection methodology, while Jin and Ekşioğlu (2008) minimized vehicle delay.

Methods used included mathematical programming and its derivates, such as stochastic and integer programming, Campbell, Vandenbussche, and Hermann (2008) and Yuan and Wang (2009). Jotshi, Gong and Batta (2009) used the HAZUS program to develop a post-earthquake scenario in Los Angeles. [HAZUS is a computer-based system created and distributed via the Web by the Federal Emergency Management Agency (FEMA) for estimating potential losses caused by earthquakes, floods and hurricanes].

Inventory: Traditionally, in the commercial area, inventory analyses address a number of areas: materials, components, work-in-process, and finished goods (Nahmias 2009). But, businesses may use inventory theory to pre-analyze forecasted disasters, e.g., Taskin and Lodree (2011) developed an inventory model for a manufacturing facility whose demand could be impacted by a potential storm. This might also be appropriate for DM in the case of items such as canned food, lamps, and coolers. In general, humanitarian logistics inventory concerns are mostly related to the prepositioning or early acquisition of relief goods. Decisions related to inventory problems fit better in the preparedness phase of DM, but they may affect directly the effectiveness of the response phase if a shortage of inventory occurs.

Most of the inventory-oriented papers shared one common objective: minimize expected cost. This cost may be expressed as a loss function (Taskin and Lodree 2011) or may be a composition of traditional inventory costs including the cost per order, holding inventory cost, and back-order cost (Beamon and Kotebla 2006). Salmerón and Apte (2010) developed a two-stage model for a humanitarian logistics for optimally allocating a budget for acquiring and positioning relief assets. Two objectives were pursued: minimization of the expected number of casualties, and minimization of the expected amount of unmet transfer population. Here, casualties were the result of seriously injured people who were not served promptly by medical staff, and people needing relief supplies who do not get them on time. On the other hand, transfer population represent people who are not in a critical condition, but still need to be evacuated to relief centers. Unmet transfer population applies when these people are not promptly evacuated.

DM inventory problems were analyzed using stochastic optimization combined with statistical tools such as Bayesian methods. Taskin and Lodree, (2011) present some numerical examples with simulated data, while other research used hypothetical data from previous studies.

Location: In general, location analysis deals with problems of siting facilities in a given area (ReVelle and Eiselt 2005). Such problems are commonly classified by businesses as strategic, i.e., a type of decision whose effects are expected to last for a long period due to the fixed cost of opening a facility, and/or changing the location of a facility may be a very expensive. In humanitarian logistics, however, location analysis may be best defined as a tactical decision, as most often it considers locating temporary shelters and warehouses where relief assets may be kept safe. These facilities generally consist of existing sites suitable, such as schools, stadiums, or churches.

Depending on the objectives pursued, results from location analysis may set the framework for ulterior decision problems such as: where to store prepositioned supplies; given the location of such relief supplies, how they would be distributed; where the evacuees will be directed to; and where to locate emergency vehicles or provisional health centers. Location analysis may be more accurately relate to the preparedness phase of DM. But, it could also be associated to the mitigation phase for locating facilities in low-risk areas, or, based on the disaster, in the response phase to improvise additional shelters or medical centers other than those that were planned.

Facility location applied in the preparedness phase is discussed by Balcik and Beamon (2008) who sought to locate distribution centers and determine the amount of supply to preposition at such centers to maximize the total expected demand covered. Lee et al. (2009) studied multiple dispensing points to service a large population searching for prophylaxis, with the objective to minimize the maximum expected traveled distance.

For the mitigation phase, Berman et al. (2009) analyzed where to locate p facilities to maximize coverage on a network whose links could be destroyed. Beraldi and Bruni (2009) studied the location of emergency vehicles under congested settings with the objective of minimizing cost.

Most of the DM location analysis research used mixed integer programming (MIP) and, in some cases, applied heuristic methods to help determine the solution of large problems (Berman et al. 2009). Other studies used stochastic programming models (Beraldi and Bruni 2009), or simulation to generate potential scenarios so as to compare the model results to actual data form a case study (Afshartous et al. 2009).

Logistics Models Overview

DM logistics involves several activities that include planning, warehousing, location, and distribution, among other elements. Some studies combined one or more of these activities, with others focused on an integrated and general concept of logistics.

Kovács and Spents (2007) and Van Wassenhoven (2006) describe humanitarian logistics as a whole. They sought a better understanding of planning and carrying out of logistics in disaster relief through a literature review. Van Wassenhoven presents a parallel between private and humanitarian logistics, and also proposes some guidelines for developing a better preparedness strategy for the latter.

Yi and Özdamar (2007) define an integrated capacitated location-routing model. Their model was designed to coordinate the distribution of relief material and the transportation of evacuees to emergency units selected through location analysis. The objective was to minimize the relationship between the weighted sum of unsatisfied demand and the weighted sum of wounded people at temporary and permanent emergency units using a two stage MIP model.

Chang, Tseng and Chen (2007) analyze a combination of location and transportation: the coordination activities related to rescue logistics efforts under a flood setting in an urban area. They consider the location of rescue resource inventory, allocation and distribution of rescue resources, and the structure of rescue organizations. Using two models, they first classified the rescue areas according to levels of emergency with the objective of minimizing the shipping cost of rescue equipments; the second model was a two stage stochastic-programming model that minimized set-up cost of storehouses and rescue equipment costs.

Yan and Shih (2009) developed a model for roadway repair scheduling and subsequent distribution of relief supplies. The objective was the minimizing the total expected time for repair and distribution using a MIP model. A related study in which a distribution system is modeled as a supply chain where the echelons are the relief suppliers, relief distribution centers, and relief demanding areas is described in Sheu (2007). Here, the objective was to minimize the expected cost of relief distribution during the three days following the onset of the disaster using a hybrid fuzzy-clustering method.

Balcik, Beamon and Smilowitz (2008) studied what is termed the last mile relief distribution, i.e., the distribution of relief assets from distribution centers to final demand. Their model dealt with the allocation of relief supplies to local distribution centers, and the delivery of schedules and routes for distributing vehicles. Their MIP model minimized the expected cost of distribution that included routing costs and a penalty for unmet demand.

Concluding Remarks

This article presented an overview of DM focused on planning and logistics. It is clear that planning and logistics are inseparable, intrinsically related, and both present in different phases of DM. These phases should be performed in a cyclic fashion so that the recovery efforts should also pursue mitigation objectives. Related research showed that many OR/MS-based studies have been directed at improving the effectiveness and efficiency of DM. The impetus for this is probably due to the catastrophic events of the Twin Towers attack in 2001, the 2004 tsunami in the Indian Ocean, and hurricane Katrina in 2005. These events have contributed to generating an increasing concern of reducing both the risk of such disasters happening and diminishing their consequences. A comparison between humanitarian and business logistics highlighted both their differences as well as their commonalities.

The main topics found from the review of OR/MS research, as related to DM, appear to be evacuation, risk analysis, and logistics. The following remarks with respect to these main topics are based on a review of a fraction of the available literature in this area; it is felt, however, that they do represent an accurate view of the state of the art in this growing field, circa 2011.

In general, the evacuation problems showed that the main concern was the minimization of evacuation time. Some researchers stated that one of the important limitations of such studies was predicting the behavior of evacuees—many variables would have to be considered, as well as social context of the evacuated population. Peacock, Morrow, and Gladwin (1997) analyzed how some people may not respond to evacuation measures before a disaster strikes as a function of their ethnic origin or their socio-economic level. The authors' main conclusion dealt with the perception the evacuee population may have about authorities who may stop them from following pre-disaster evacuation orders.

Risk analysis has proved to be a useful concept when planning for disasters, especially during the mitigation phase. A problem is the difficulty of enumerating the possible risk scenarios. Moreover, many studies are based on statistical analyses to historical data, but in some occasions, the events being studied are so infrequent that no reliable analysis can be achieved.

For humanitarian logistics research, a distinction was made between transportation, location analysis, inventory, and humanitarian logistics, in general. A limitation that may arise in a transportation study is the inability to incorporate the presence of congestion, even though some studies do, see for example Beraldi and Bruni (2009). Inventory theory has been used by both business and humanitarian logistics to better prepare for disasters, including, as well, location analysis problems from business being applied in humanitarian location settings.

The research papers reviewed referred mainly to the preparedness phase of DM, followed by response and mitigation phases; no work was found related to the recovery phase. Altay and Green, (2006) noted the lack of OR studies related to recovery efforts in comparison to the other phases. Another aspect in which the findings obtained here agree with the ones presented by Altay and Green (2006) is that most of the studies reviewed consists of the development of models, rather than theoretical studies or application tools such as software. For the disasters most commonly studied, there was not a clear reference to man-made disasters such as terrorist attacks; the case studies always dealt with natural disasters.

For DM, an important challenge for the OR/MS community "is to develop a science of disaster logistics that builds upon, among others, private sector logistics and to transfer to private business the specific core capabilities of humanitarian logistics," (Van Wassenhove 2006).

See

- ▶ Inventory Modeling
- ▶ Linear Programming
- ▶ Logistics and Supply Chain Management
- ▶ Risk Assessment
- ▶ Scheduling and Sequencing
- ▶ Simulation of Stochastic Discrete-Event Systems
- ▶ Vehicle Routing

References

Afshartous, D., Guan Y., & Mehrotra, A. (2009). US Coast Guard air station location with respect to distress calls: A spatial statistics and optimization based methodology, *European Journal of Operational Research, 196*(3), 1086–1096.

Alexander, D. E. (2002). *Principles of disaster planning and management*. Oxford University Press.

Altay, N., & Green, W. G. (2006). OR/MS research in disaster operations management. *European Journal of Operational Research, 175*, 475–493.

Balcik, B., & Beamon, B. M. (2008). Facility location in humanitarian relief, *International Journal of Logistics Research and Applications: A Leading Journal of Supply Chain Management, 11*(2), 101–121.

Balcik, B., Beamon, B. M., & Smilowitz, K. (2008). Last mile distribution in humanitarian relief. *Journal of Intelligent Transportation Systems, 12*(2), 51–63.

Barker, K., & Haimes, Y. Y. (2009). Assessing uncertainty in extreme events: Applications to risk-based decision making in interdependent infrastructure sectors. *Reliability Engineering and System Safety, 94*, 819–829.

Beamon, B. M., & Kotebla, S. A. (2006). Inventory model for complex emergencies in humanitarian relief operations. *International Journal of Logistics: Research and Applications, 9*(1), 1–18.

Beraldi, P., & Bruni, M. E. (2009) A probabilistic model applied to emergency service vehicle location. *European Journal of Operational Research, 196*(1), 323–331.

Berman, O., Drezner, T., Drezner, Z., & Wesolowsky, G. O. (2009). A defensive maximal covering problem on a network. *International Transactions in Operational Research, 16*(1), 69–86.

Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences, 99*(3), 7280–7287.

Campbell, A. M., Vandenbussche, D., & Hermann, W. (2008). Routing for relief efforts. *Transportation Science, 42*(2), 127–145.

Chang, M.-S., Tseng, Y.-L., & Chen, J.-W. (2007). A scenario planning approach for the flood emergency logistics preparation problem under uncertainty. *Transportation Research Part E: Logistics and Transportation Review, 43*(6), 737–754.

Chen, X., Meaker, J. W., & Zhan, F. B. (2006). Agent-based modeling and analysis of hurricane evacuation procedures for the Florida keys. *Natural Hazards, 38*(3), 321–338.

Chen, X., & Zhan, F. B. (2008). Agent-based modelling and simulation of urban evacuation: Relative effectiveness of simultaneous and staged evacuation strategies. *Journal of the Operational Research Society, 59*, 25–33.

Chiu, Y.-C., & Zheng, H. (2007). Real-time mobilization decisions for multi-priority emergency response resources and evacuation groups: Model formulation and solution. *Transportation Research Part E: Logistics and Transportation Review, 43*(6), 710–736.

Chiu, Y.-C., Zheng, H., Villalobos, J., & Gautam, B. (2007). Modeling no-notice mass evacuation using a dynamic traffic flow optimization model. *IIE Transactions, 39*(1), 83–94.

Hu, Z.-H. (2010). Relief demand forecasting in emergency logistics based on tolerance model. *2010 Third International Joint Conference on Computer Science and Optimization (CSO 2010)*, Vol. 1, pp. 451–455.

Huang, C., & Ruan, D. (2008). Fuzzy risks and an updating algorithm with new observations. *Risk Analysis, 28*(3), 681–694.

Jin, M., & Ekşioğlu, B. (2008). Optimal routing of vehicles with communication capabilities in disasters. *Computational Management Science, 7*(2), 121–137.

Jotshi, A., Gong, Q., & Batta, R. (2009). Dispatching and routing of emergency vehicles in disaster mitigation using data fusion. *Socio-Economic Planning Sciences, 43*(1), 1–24.

Kovács, G., & Spents, K. M. (2007). Humanitarian logistics in disaster relief operations. *International Journal of Physical Distribution and Logistics Management, 37*(2), 99–114.

Kunkel, K. E., Pielke, Jr., R. A., & Changnon, S. A. (1999). Temporal fluctuations in weather and climate extremes that cause economic and human health impacts: A review. *Bulletin of the American Meteorological Society, 80*, 1077–1098.

Lee, E. K., Smalley, H. K., Zhang, Y., Pietz, F., & Benecke, B. (2009). Facility location and multi-modality mass dispensing strategies and emergency response for biodefence and infectious disease outbreaks. *International Journal on Risk Assessment and Management, 12*(2–4), 311–351.

Lettieri, E., Masella, C., & Radaelli, G. (2009). Disaster management: Findings from a systematic review. *Disaster Prevention and Management, 18*(2), 117–136.

Li, Y. P., Huang, G. H., & Nie, S. L. (2007). Mixed interval-fuzzy two-stage integer programming and its application to flood-diversion planning. *Engineering Optimization, 39*(2), 163–183.

Liu., Y., Lai., X., & Chang., G. (2006). Two-level integrated optimization system for planning of emergency evacuation. *Journal of Transportation Engineering*, 800–807.

Matisziw, T. C., & Murray, A. T. (2009). Modeling s–t path availability to support disaster vulnerability assessment of network infrastructure. *Computers and Operations Research, 36*(1), 16–26.

McLoughlin, D. (1985). A framework for integrated emergency management. *Public Administration Review, 45*, 165–172.

Miller, H. E., Engemann, K. J., & Yager, R. R. (2006). Disaster planning and management. *Communications of the IIMA, 6*(2), 25–36.

Miller-Hooks, E., & Sorrel, G. (2008). Maximal dynamic expected flows problem for emergency evacuation planning. *Transportation Research Record: Journal of the Transportation Research Board, 2089*, 26–34.

Nahmias, S. (2009). *Production and operation analysis* (6th ed., pp. 201–202). McGraw-Hill.

Opasanon, S., & Miller-Hooks, E. (2009). The safest escape problem. *Journal of the Operational Research Society, 60*, 1749–1758.

Peacock, W. G., Morrow, B. H., & Gladwin, H. (1997). *Hurricane Andrew: Ethnicity, gender and the sociology of disasters* (p. 278). New York: Routledge.

Perry, R. W., & Lindell, M. K. (2003). Preparedness for emergency response: Guidelines for the emergency planning process. *Disasters, 27*(4), 336–350.

ReVelle, C. S., & Eiselt, H. A. (2005). Location analysis: A synthesis and survey. *European Journal of Operational Research, 165*(1), 1–19.

Saadatseresht, M., Mansourian, A., & Taleai, M. (2009). Evacuation planning using multiobjective evolutionary optimization approach. *European Journal of Operational Research, 198*(1), 305–314.

Salmerón, J., & Apte, A. (2010). Stochastic optimization for natural disaster asset preposition. *Production and Operations Management, 19*(5), 561–574.

Sayyady, F., & Eksioglu, S. D. (2010). Optimizing the use of public transit system during no-notice evacuation of urban areas. *Computers and Industrial Engineering, 59*(4), 488–495.

Sheu, J.-B. (2007). An emergency logistics distribution approach for quick response to urgent relief demand in disasters. *Transportation Research Part E: Logistics and Transportation Review, 43*(6), 687–709.

Snyder, L. V., Scaparra, M. P., Daskin, M. S., & Church, R. L. (2006). *Planning for disruptions in supply chain networks*. Tutorials in Operations Research INFORMS.

Stepanov, A., & Smith, J. M. (2009). Multi-objective evacuation routing in transportation networks. *European Journal of Operational Research, 198*, 435–446.

Taskin, S., & Lodree, E. J. (2011). A Bayesian decision model with hurricane forecast updates for emergency supplies inventory management. *Journal of the Operational Research Society, 62*, 1098–1108. Published online 19 May 2010.

Van Wassenhove, L. N. (2006). Blackett memorial lecture – humanitarian aid logistics: Supply chain management in high gear. *Journal of the Operational Research Society, 57*(5), 475–489.

Xu, Y.-P., Booij, M. J., & Tong, Y.-B. (2010). Uncertainty analysis in statistical modeling of extreme hydrological events. *Stochastic Environmental Research and Risk Assessment, 24*(5), 567–578.

Yan, S., & Shih, Y.-L. (2009). Optimal scheduling of emergency roadway repair and subsequent relief distribution. *Computers and Operations Research, 36*(6), 2049–2065.

Yi, W., & Özdamar, L. (2007). A dynamic logistics coordination model for evacuation and support in disaster response activities. *European Journal of Operational Research, 179*(3), 1177–1193.

Yuan, Y., & Wang, D. (2009). Path selection model and algorithm for emergency logistics management. *Computers and Industrial Engineering, 56*(3), 1081–1094.

Discrete-Programming Problem

► Integer and Combinatorial Optimization

Discrete-Time Markov Chain (DTMC)

A discrete-time, countable-state Markov process. It is often just called a Markov chain.

See

- ▶ Markov Chains
- ▶ Markov Processes

Disease Prevention, Detection, and Treatment

Jingyu Zhang[1], Jennifer E. Mason[2], Brian T. Denton[3] and William P. Pierskalla[4]
[1]Philips Research North America, Briarcliff Manor, NY, USA
[2]University of Virginia, Charlottesville, VA, USA
[3]University of Michigan, Ann Arbor, MI, USA
[4]University of California, Los Angeles, CA, USA

Introduction

Advances in medical treatment have resulted in a patient population that is more complex, often with multiple diseases, competing risks of complications, and medication conflicts, rendering medical decisions harder because what helps one patient or condition may harm another. The use of Operations Research (OR) methods for the study of healthcare has a long history. Furthermore, there is a growing literature on emerging applications in this area. This article provides examples of contributions of OR methods, including mathematical programming, dynamic programming, and simulation, to the prevention, detection, and treatment of diseases. More extensive surveys of OR studies of health care delivery, including medical decision making, can be found in Pierskalla and Brailer (1994), Brandeau et al. (2004), and Rais and Viana (2010).

Advances in medical treatment have extended the average lifespan of individuals, and transformed many diseases from life threatening in the near term to chronic conditions in need of longterm management. Many new applications of OR are emerging as treatment options and population health evolve over time. For example, new treatments have become available for various forms of cancer, HIV, and heart disease. In some cases, patients are living decades with diseases that previously had low short-term survival rates. As a result, more patients are living with co-morbid conditions, and competing risks, creating challenging decisions that must balance the downside of treatment (e.g., medication side effects and long-term complications) with the benefits of treatment (e.g., longer life expectancy and better quality of life).

Diabetes is a good example of a chronic disease for which medical treatment is complex. With nearly 8% of the U.S. population estimated to have diabetes, it is recognized as a leading cause of mortality and morbidity. It is associated with long-term complications that affect almost every part of the body, including coronary heart disease (CHD), stroke, blindness, kidney failure, and neurological disorders. For many patients, diabetes might be prevented through improved diet and exercise. However, due to the slow development of symptoms in many patients, diabetes can go undetected for years. For patients that are diagnosed with diabetes, risk models exist to predict the probability of complications, but alone these models do not provide optimal treatment decisions. Rather, they provide raw data that can be used in OR models to make optimal treatment decisions. This general situation is true of many chronic diseases. As a result, there are many emerging opportunities for applications of OR to disease prevention, detection, and management.

This article is organized as follows. The section on Disease Prevention and Screening describes important contributions of OR to disease prevention, including vaccination and screening methods for detecting disease in a population of potentially infected people. The section on Treatment Choices focuses on applications to long-term management of chronic diseases, including selection among multiple treatment choices, and decisions about timing and dosage of treatment. The section on Emerging Applications reviews some emerging applications to real-time decision making at the point of care and patient decision aids. Finally, research opportunities are discussed in the Conclusions section.

Disease Prevention and Screening

Prevention and screening are important factors in determining overall population health. OR has been applied to help inform decisions related to prevention and screening for decades. Two major topics in this area, that are prominent in the OR literature, are vaccination and disease screening. Vaccination emphasizes the prevention of infectious diseases, while disease screening is common for both non-infectious and infectious diseases. Each of these topics will be discussed in detail in this section.

Vaccination

The biological and genetic sciences have greatly increased the knowledge of how viruses and bacteria operate within the body to create disease. This has led to the discovery of many new vaccines. However, the myriad interactions as well as controversy about their effects on individuals, and an overall population, have drawn considerable public attention. These interactions and effects present several challenges in the utilization of the vaccines for disease control. First, there are a large number of diseases for which effective vaccines are available. Some have specific requirements, such as multiple doses that must be administered within a minimum or maximum time window. Also, some have conflicts with other vaccines. Second, many new vaccines are coming on the market, including combination (multi-valent) vaccines that can cover multiple diseases. Third, for some diseases there is uncertainty about the future evolution of epidemic strains, leading to questions about optimal design of vaccines. Finally, there are challenges in the vaccine manufacturing process including uncertain yields, quality control, supply chain logistics, and the optimal storage location of vaccine supplies. OR models have been applied to address many of these challenges.

Pediatric Vaccination

Pediatric or childhood vaccination is the most common means of mass vaccination. OR researchers have developed models to aid in the selection of a vaccine formulary, pricing of vaccines, and design of vaccination schedules. Jacobson et al. (1999) proposed integer-programming models to determine the price of combination vaccines for childhood immunization. Their models considered all available vaccine products at their market prices and constraints based on the U.S. national recommended childhood immunization schedule. Their objective was to find the vaccine formularies with the lowest overall cost from the patient, provider, and societal perspectives. Their integer-programming models considered the first five years of the recommended childhood immunization schedule against six diseases. They used binary decision variables to denote whether a vaccine is scheduled for a particular month's visit.

In a later study, Jacobson et al. (2006) investigated a pediatric vaccine supply shortage problem to assess the impact of pediatric vaccine stockpile levels on vaccination coverage rates of the guidelines during supply interruption. Their model was similar to inventory models that consider stock-outs, as well as lot sizing problems with machine breakdowns. Objectives of their model included optimizing service level and minimizing a standard loss function. Using their model, they concluded that the guidelines are only sufficient to mitigate a vaccine production interruption of eight months.

Hall et al. (2008) considered a childhood vaccination formulary problem that allows for combination vaccines. They proposed an integer-programming model to minimize the cost of fully immunizing a child under the constraints of a recommended schedule. They proved their proposed model is NP-hard. They proposed exact algorithms using dynamic programming and heuristics for approximating near optimal solutions to their model. Engineer et al. (2009) further investigated an extension that involves catch-up scheduling for childhood vaccination. They provided details of a successful implementation of their model as a decision support system.

Flu Vaccination

Some diseases evolve rapidly over time, necessitating frequent vaccination on a regular basis. For example, the composition of seasonal flu vaccine changes every year. Wu et al. (2005) proposed a model for flu vaccine design. They used a continuous-state discrete-time dynamic-programming model to find the optimal vaccine-strain selection policy. In their dynamic program, the state was represented by the antigenic history, including previous vaccine and epidemic strains. The decision variable (action) was the vaccine strain to be selected, and the reward is the cross-reactivity representing the efficacy of the vaccine. The objective was to maximize the expected discounted reward. Approximate solutions were obtained by state-space aggregation and compared to an easy to-implement myopic policy based on approximating the multi-stage problem by a series of single period problems. They compare policies suggested by their model to theWorld Health Organization (WHO) recommended policy. Based on their results, the authors suggested that the WHO policy is reasonably effective and should be continued.

Vaccination for Bio-defense

OR researchers have contributed to problems related to vaccination strategy for bio-defense. For instance, Kaplan et al. (2003) analyzed bio-terror response logistics using smallpox as an example. The authors proposed a trace vaccination model using a system of ordinary differential equations (ODEs) incorporating scarce vaccination resources and queueing of people for vaccination. An approximate analysis of the ODEs yields closed-form estimates of numbers of deaths and maximum queue length. They also obtained approximate closed-form expressions for the total number of deaths under mass vaccination. Using these results, approximate thresholds for controlling an epidemic were derived.

Kress (2006) also considered the problem of optimizing vaccination strategy in response to potential bio-terror events. The author developed a flexible, large-scale analytic model with discrete-time decisions. The author used a set of difference equations to describe the transition of the number of people at each epidemic stage and proposed a vaccination policy, which is a mixture of mass and trace vaccination policies.

Other Vaccination Related Problems

Several other vaccine-related problems have been investigated by OR researchers. For example, vaccine allocation problems must consider criteria and constraints related to vaccine manufacturing and supply chain logistics. Becker and Starczak (1997) formulated the optimal allocation of vaccine as a linear-programming problem. Their objective was to prevent epidemics with the minimum required vaccine coverage. Their linear-programming model considered heterogeneity among individuals and minimized the initial reproduction number for a given vaccination coverage. The optimal vaccine allocation strategy suggested more individuals need to be vaccinated in larger households.

Disease Screening

Disease screening is important in extending life expectancy and improving people's quality of life. Effective screening can also reduce costs to the healthcare system by avoiding the high costs associated with treatment of late-stage disease. However, when and how to screen for a specific disease is a complex decision. For instance, model formulation is often difficult due to unclear pathology and risk factors, uncertainty in disease staging and the relationship to symptoms and test results, and the trade-off between the benefit of early detection and the side effects and costs of screening and treatment. The types of OR methods employed depend on whether the disease is non-infectious or infectious. Following are several examples from each category of diseases.

Non-infectious Disease Screening

Modeling disease progression among different stages throughout a patient's lifetime, as well as the trade-off between pros (e.g., longer life expectancy and better quality of life) and cons (e.g., side effects and costs of over-diagnosis and over-treatment) of disease screening are central to non-infectious diseases. Shwartz (1978) proposed one of the first models for breast cancer screening to evaluate and compare alternative screening strategies. Their stochastic model consisted of a discrete set of breast cancer disease states and criteria including life expectancy and the probability of diagnosis. A significant amount of research on breast cancer screening has developed; see Mandelblatt et al. (2009) for a review of breast cancer screening models.

Eddy (1983) presented a general model of monitoring patients with repeated and imperfect medical tests. The model considered clinical and economic outcomes such as the probability of detecting a disease, the method and timing of detection, the stage at which the disease is detected, costs, and the benefit of screening based on the willingness to pay. The model incorporated disease incidence, the natural history of disease progression, the effectiveness of tests and subsequent treatments, and the order and frequency of tests. The model was illustrated using a hypothetical example. The model had subsequently been applied in clinical practice to several cancer screening problems.

To capture uncertainty in identifying disease states, OR techniques such as partially observable Markov decision process (POMDP) have been applied. For example, Zhang et al. (2012) developed a POMDP model for prostate cancer screening. Due to the slow growing nature of prostate cancer, the imperfect nature of diagnostic tests, and the quality of life impact of treatment, whether and when to refer a patient for biopsy is controversial. The objective of their model was to maximize the quality adjusted life expectancy and minimize the costs of screening and treatments. They assumed that cancer states are not directly observable, but the probability a patient has cancer can be estimated from their PSA test history. A control-limit type policy of biopsy referral and the existence of stopping time of prostate cancer screening were proven. The authors compared policies suggested by their model, to commonly recommended screening policies, and concluded there may be substantial benefits from using prostate cancer risk to make screening decisions.

Screening for disease is greatly influenced by the diagnostic accuracy of the tests. An example of work done in this area is given by Rubin et al. (2004) in which the authors used a Bayesian network to assist mammography interpretation. Interpreting mammographic images and making correct diagnoses are challenging even to experienced radiologists. False-negative interpretations can cause delay in cancer treatment and lead to higher morbidity and mortality. False positives, on the other hand, result in unnecessary biopsy causing anxiety and increased medical costs. The American College of Radiology developed BI-RADS which is a lexicon of mammogram findings and the distinctions that describe them. The authors showed that their Bayesian network model may help to reduce variability and improve overall interpretive performance in mammography.

Many other diagnostic areas have been addressed including gastrointestinal diseases, neurological diseases, and others.

Infectious Disease Screening

In infectious diseases screening, one of the goals is to prevent an epidemic outbreak. Therefore, disease progression and communication throughout a population is an important consideration. Lee and Pierskalla (1988) proposed a mathematical-programming model for contagious diseases with little or no latent periods. The objective of their model was to minimize the average number of infected people in the population. Their model was converted to a knapsack problem. They considered both perfect and imperfect reliability of tests and showed the optimal screening policy has equally spaced screening intervals when the tests have perfect reliability.

Disease screening problems often involve multiple criteria, stemming from the patient, provider, and societal perspectives. For example, Brandeau et al. (1993) provided a cost-benefit analysis of HIV screening for women of childbearing age based on a dynamic compartmental model incorporating disease transmission and progression over time. The model was formulated as a set of simultaneous nonlinear differential equations. The authors found the primary benefit of screening is to prevent the infection of their adult contacts, and that screening of the medium to high risk groups may be cost-beneficial, but it is not likely to be cost-beneficial for low risk women.

Blood screening tests have been used to improve the quality of the blood supply. An early example to improve the performance of testing strategies in the 1980s was provided by Schwartz et al. (1990) for screening blood for the HIV antibody, and making decisions affecting blood donor acceptance. At the time the work was done, limited knowledge was available about the biology, epidemiology, and early blood manifestations of HIV. Furthermore, the initial and conditional sensitivities and specificities of enzyme immunoassays and Western blot tests had wide ranges of errors. A decision tree, with the decisions probabilistically based on which screening test to use, and in what sequence, was used to minimize the number of HIV infected units of blood and blood products entering the nation’s blood supply subject to a budget constraint. The model was used at a meeting of an expert panel of the U.S. National Heart Lung and Blood Institute to inform the panelists who were deciding which blood screening protocol to recommend. The model provided outputs including: expected number of infected units entering the blood supply per unit time, expected number of uninfected units discarded per unit time, expected number of uninfected donors falsely notified, and the incremental cost among screening regimens.

Efficiency of screening can be a defining factor in the success or failure of proposed screening methods. Wein and Zenios (1996) proposed models for pooled testing of blood products for HIV screening. Optimization of pooled testing involves decisions such as transfusion, discarding of samples in the pool, and division of the pool into sub-pools. Several models were proposed to minimize the expected costs. The outcome of an HIV test was measured by an optical density (OD) reading, a continuous measurement which is determined by the concentration of the antibodies. The states of the system were the previous history of the OD readings. A dynamic-programming model with a discretized state space and a heuristic solution algorithm were introduced to obtain near optimal solutions. The policy obtained by the heuristic algorithm was proposed as a cost-effective, accurate, and relatively simple alternative to the implemented HIV screening policies.

Treatment Choices

The following section focuses on treatment decisions for patients with chronic diseases such as diabetes, HIV, cancer, and end-stage renal disease. Treatment of patients with chronic diseases is often complex due to the long-term nature of the illness and the future uncertainty in patient health. Complicating matters, these patients may have other comorbidities that need to be taken into account when treatment decisions are made. In the following section, two areas related to choice of treatment are presented where OR is used to address challenges related to drug treatment decisions and organ transplantation for patients with chronic conditions.

Drug Treatment Decisions

Many diseases involve complex drug treatment decisions, particularly for chronic conditions. Decisions about which medications to initiate, when to initiate treatment, and the appropriate dosage are of primary importance. Additional challenges arise from the fact that there is uncertainty about the future health of the patient, adherence to treatment, and the efficacy of drugs for a particular patient. Treatment decisions must also take into account the often irreversible nature of treatment decisions. Many treatment optimization models employ the use of a natural history model of the disease and all-cause mortality, incorporating the influence of competing risks into the treatment decision.

Choice of Treatment

When there are multiple candidate treatments available, the choice of treatment may be unclear. OR techniques have been used to select treatments. For example, Pignone et al. (2006) presented a Markov model to select among aspirin, statins, and combination treatment, for the prevention of coronary heart disease (CHD). The model simulated the progression of middle-aged males with no history of CHD. The model was used to estimate cost per quality-adjusted life year (QALY) gained. The authors found that aspirin dominates no treatment when a patient's ten-year risk of CHD is at least 7.5%. If a patient's risk is greater than 10%, combination treatment is recommended.

Hazen (2004) used dynamic influence diagrams to analyze a chain of decisions as to whether a patient should proceed to total hip replacement surgery or not. The objective in making this decision was to calculate the optimal expected costs and QALYs under each choice. The use of QALYs for the objective was important because an older person undergoing hip replacement may not have more expected years of life relative to not doing surgery, but the quality of life improvement can be considerable and, quite possibly, worth the cost.

Timing of Treatment

With chronic conditions that can span many years, the optimal time to initiate particular treatments may be unknown. There have been several studies that researched the optimal timing of treatment. Two models relate to the optimal timing of HIV treatment. This question is of particular interest since patients that begin HIV treatment will only be able to use the drug for a limited amount of time, as the virus builds up resistance to the drug. Shechter et al. (2008) used a Markov decision process (MDP) model to find the optimal time to initiate HIV therapy, while maximizing the patient's quality of life. At monthly decision epochs, the decision was made to initiate therapy or wait until the next month to decide. The health states were based on the number of CD4 white blood cells, the primary target of HIV, and the reward was the expected remaining lifetime in months. They assumed a stationary infinite horizon model and found that if it is optimal to initiate treatment at a given CD4 count, it is also optimal to initiate treatment for patients with higher CD4 counts. The model supported earlier treatment, despite trends toward later treatment. Braithwaite et al. (2008) analyzed the timing of initiation based on CD4 counts for varying viral loads. They used a simulation to compare different CD4 count treatment thresholds for initiation of therapy. The model compared life expectancy and QALYs for the different strategies of initiation. In agreement with Shechter et al.'s finding, the simulation suggested that the use of earlier initiation of treatment (higher CD4 count thresholds) results in greater life years and QALYs.

Agur et al. (2006) developed a method to create treatment schedules for chemotherapy patients using local search heuristics. The model simulated cell growth over time and finds two categories of drug protocols: one-time intensive treatment and a series of nonintensive treatments. Chemotherapy schedules were evaluated based on a patient's state at the end of a given time period, number of cancer and host cells, and the time to cure. Simulated annealing, threshold acceptance, and old bachelor acceptance—a variant of threshold acceptance in which the trial length is set by users—were used to obtain better treatment schedules. The authors reported good results with all three techniques, but they showed simulated annealing resulted in the greatest computational effort.

Denton et al. (2009) investigated the optimal timing of statin therapy for patients with type 2 diabetes. This problem was formulated as a discrete time, finite horizon, discounted MDP in which patients transition through health states corresponding to varying risks of future complications, their history of complications, and death from other causes unrelated to diabetes. The objective was to maximize reward for QALYs minus costs of treatment. The optimal timing of treatment for patients was determined using three published risk models for predicting cardiovascular risk. The earliest time to start statins was age 40 for men, regardless of which risk model was used. However, for female patients, the earliest optimal start time varied by 10 years, depending on the risk model. Mason et al. (2012) extended this work to account for poor medication adherence. The authors used a Markov model to represent uncertain future adherence after medication was initiated. They observed that the optimal timing of statins should be up to 11 years later for patients with uncertain future adherence. However, they also found that improving adherence has a much larger effect on QALYs than delaying the timing of initiation.

Paltiel et al. (2004) constructed a simulation model to treat asthma. The model forecasted asthma-related symptoms, acute exacerbations, quality adjusted life expectancy, health-care costs, and cost-effectiveness. Their intent was to reduce asthma manifestations, improve life quality, and reduce costs of care. The authors pointed out that similar models could be constructed for the control of other subpopulation-wide diseases such as obesity, smoking, and diabetes.

A great deal of work has also been done on modeling CHD interventions. Cooper et al. (2006) provided an excellent review of many models used for this disease. Most of the models reviewed by the authors are decision trees, Markov processes, or simulation models. Decisions included when and what types of interventions, and what types of drugs to employ, at various stages of disease.

Dosage of Treatment

Given a particular treatment has been selected, the appropriate dosage must be determined. He et al. (2010) provided a discrete-state MDP model for determining gonadotropin dosages for patients undergoing in vitro fertilization-embryo transfer therapy. This work focused on patients with the chronic condition of polycystic ovaries syndrome that tend to be more sensitive to the gonadotropin treatment. The resulting policies from the MDP model were evaluated through simulation to determine the impact of misclassifying patients. In general, the use of OR techniques can be used to provide a better starting dosage with less fine tuning needed after initiation of treatment.

Dosage decisions are also important in radiation treatment planning. Several studies have focused on radiotherapy for cancer using mathematical optimization techniques. Although the vast majority of these treatment plans are designed by clinicians through intelligent trial and error, it is becoming essential to use optimization for extremely complicated and complex plans. Holder (2004) used linear programming for intensity modulated radiotherapy treatment (IMRT). Ferris et al. (2004) discussed various optimization tools for radiation treatment planning. In both of these papers, the objective was to deliver a specified dose to the target area (above a minimum and below a maximum level of dosage) and spare or minimize damage to surrounding healthy tissue and nearby critical body structures and organs.

Organ Transplants

End-stage liver disease (ESLD) and end-stage renal disease (ESRD) have received a great deal of study in the OR literature. They are chronic conditions that can result in patients eventually needing liver or kidney transplants, respectively. Chronic liver disease or liver failure can result from many causes, including liver cancer and chronic hepatitis. Often, initial treatment of liver failure attempts to manage the underlying cause, followed by intensive care and management of complications such as bleeding problems. If patients continue to deteriorate to ESLD, liver transplantation may be the only option. Patients with chronic kidney disease have a continuing loss of renal function, leading to ESRD. Once a patient has ESRD, renal replacement therapy in the form of dialysis or kidney transplantation is necessary.

While organ transplants are the best long-term solution for patients with chronic liver or kidney disease, there is a shortage of organs for transplant and a growing waiting list of patients. OR techniques have been applied to optimize the allocation of organs and timing of transplants for increasing quality and length of life of the recipients. The allocation of kidneys and livers for transplantation is challenging because both living and cadaveric donors are possible. With living donors, there is more flexibility in the timing of the transplant, allowing for the transplant timing decision to be optimized. For both kidney and liver transplantation, there are challenging decisions about whether to use a living or cadaveric donor (if both are available), and when the transplant should occur. OR techniques have also aided in finding the greatest number of donor-recipient matches, considering the challenges of blood and tissue type compatibilities.

Alagoz et al. (2004) studied the question of the optimal timing of liver transplantation. They developed an MDP model to find the optimal timing for a patient to have a transplant from a living donor. The patients transitioned through health states defined by a scoring system for ESLD. With the donor assumed to be available at any time, the MDP maximized the patient's quality adjusted lifetime—striking a balance between having the transplant before the patient becomes too sick and waiting long enough due to the limited amount of time a patient can live after a transplant.

Su and Zenios (2004) presented an M/M/1 queueing model to determine if incorporating patient choice into allocation will improve efficiency and reduce waste of organs offered to patients but not accepted. Their model incorporated uncertain arrival of patients and organs, with the service process being the kidney transplant. Since organs cannot be stored, the service time was given by the interarrival time of organs. In addition to the traditional M/M/1 assumptions, each organ had a reward corresponding to its quality, and patients may reject an organ they believe has poor quality. The authors found that a first-come-first-serve policy can lead patients to refuse organs of lesser quality, leading to waste of up to 15% of organs. They also found that last-come-first-serve (LCFS) allocation lowers the wasteful effect of patient preference. While LCFS was not a feasible rule to implement, their results highlighted the need for adjustment of incentives associated with patient choice to prevent wasting organs.

A common way for patients to find organ donors is to ask willing family members or friends to be tested for compatibility. Another area, where OR has contributed, considers patients with willing donors that are not matches. Segev et al. (2005) considered the problem of paired kidney donation, matching two incompatible pairs with each other resulting in two successful transplants. The study considered a graph theory representation of a large pool of incompatible patient-donor pairs where each pair was represented with a node and two compatible pairs were linked with an edge. An algorithm based on the Edmonds matching algorithm (Edmonds 1965) was used to find all feasible matching solutions, and the best solution was chosen based on some predefined criteria, including the number of total matches and the number of transplant patients alive five years after the operation.

This matching strategy was compared to the first-accept scheme, which only finds one feasible solution, that is used in practice. The authors found that their algorithm could increase the total number of matches and take into account patient priorities.

Emerging Applications

Rapid advances in medicine are driving new OR research opportunities. As evidence of this, over the period from 2000–2010 the total number of health care related presentations at the Institute for Operations Research and Management Science (INFORMS) annual meeting has grown from 35 in 2000 to 281 in 2009 (Denton and Verter 2010). This section provides some specific examples of emerging areas of research.

Personalized Medicine

With the sequencing of the human genome and many advances in biomarkers for certain diseases, the idea of personalized medicine has received a great deal of attention. There are some examples of successful applications of personalized medicine, such as breast cancer treatment. However, for most diseases even basic risk factors are not yet considered as part of the standard guidelines. For example, gender is a well known risk factor for heart disease and stroke. While this has been known for decades, in many countries, including the U.S., the published treatment guidelines for control of risk factors such as cholesterol and blood pressure are the same for men and women. These examples point to opportunities to improve the design of screening and treatment guidelines through consideration of individual patient risk factors.

Decision Aids

The use of OR techniques in the development of decision aids is not as wide as in other areas of treatment choices. This is an area of research that must expand if OR models are to be translated into practice. Researchers have attempted to use artificial intelligence and computer science/information systems to provide decision support to the physician and/or patient. However, many clinicians still hesitate to use models for diagnosis or treatment. There are many possible reasons for the slow diffusion into practice. An important goal is the study of the clinician-model interface. In spite of adoption difficulties, there are examples of where OR has contributed significantly to treatment decisions. Several examples follow.

White et al. (1982) developed a quantitative model for diagnosing medical complaints in an ambulatory setting with the goal of reducing costs and improving quality of diagnoses. The model structure was influenced by three methods: decision analysis, partially observed semi-Markov decision process models, and multi-objective optimization therapy (MOOT). The authors used Bayesian-based modeling of disease progression and heuristics (a single-stage decision tree that reduces the amount of computation time and storage space per patient) to consider individual patient and physician preferences. For the MOOT heuristic, suggested by White et al. (1982), the list of possible diagnosis tests were provided, highlighting nondominated tests. The authors described a detailed example of the decision aid to treat a patient in an ambulatory setting.

Policies related to health information exchanges assume patients want to explicitly decide who can have access to their medical records. Marquard and Brennan (2009) tested this assumption by questioning 31 patients from a neurology clinic about their willingness to share information about their medication with a primary care physician, a neurologist, and an emergency room physician. Almost all patients decided to share their current medication usage with all three doctors citing the potential clinical care benefits. However, not all patients understood the possible effects of sharing this information. The use of realistic decision scenarios and structured conversations used in this study are likely to reveal more true patient preferences than abstract opinion surveys that are commonly used in practice. In addition to correctly identifying patient preferences, it is important to assess patient understanding of the consequences of their choices. Understanding the true willingness of patients to share health information is an important step in the development of decision aids and the inclusion of patient choices in medical decisions.

Using multi-attribute utility theory, Simon (2009) considered the choice of treatments for prostate cancer including surgery, external beam radiation, brachytherapy, and no treatment. The model used data collected from the medical literature to compute probabilities regarding the likelihood of death and other side effects for each of the choices. The model also incorporated the patient's individual preferences regarding length of life and quality of life in view of the possible side effects (impotence, incontinence, and toxicity). The model evaluated each treatment alternative and compared the results for the particular patient.

Real Time Decision Making

Many medical treatment decisions must be made in real time. Depending on the particular application, the definition of real time could be anything from a few seconds to several minutes. Such applications can be highly demanding, often trading off the need for high quality decisions with available time.

One area in which OR has contributed to real time decision making is blood glucose control in patients with diabetes. Patients with type 1 diabetes are insulin dependent, and careful control of blood glucose within defined physiological limits is necessary to avoid a potentially life threatening occurrence of hypoglycemia (very low blood glucose that can lead to coma and/or death if not treated immediately). Blood glucose levels can change significantly over very short periods of time (seconds) depending on a variety of factors, such as caloric intake. The most common treatment for patients with type 1 diabetes is to inject insulin. However, the need for regular injection has a serious impact on a patient's quality of life. Research has been conducted on the design of closed loop control algorithms that could enable an implantable device to optimize insulin delivery (Parker et al. 2001).

Outpatient procedures can also pose a series of challenging decisions that must be made in real time (minutes). For instance, radiation treatment for cancer patients involves a series of complex decisions that can influence the effectiveness of treatment. One example is brachytherapy for prostate cancer treatment, that involves the implantation of radioactive seeds in close proximity to a tumor. The method of brachytherapy is to place seeds in and around a tumor such that dual goals of maximizing dose to the tumor and minimizing dose to healthy tissue are balanced. Due to changes that occur in tumor size and shape and the physical movement of healthy tissue and organs in proximity to the tumor over short time periods, such decisions must be made in real time at the point of placement. This real time analysis selects the actual placements of the seeds in the prostate from the thousands of possible locations, millimeters apart. Lee and Zaider (2008) presented a nonlinear mathematical-programming model to make location decisions using real time imaging information. They demonstrated a practical application in which the clinical goals of reduced complications (e.g., impotence and incontinence) and reduced costs ($5,600 per patient) were achieved.

Concluding Remarks

The use of OR for the study of disease treatment and screening decisions has a long history. Furthermore, advances in medicine are creating new challenges which are in turn resulting in new applications of OR and new methods. This article surveyed some of the significant contributions of OR methods, including mathematical programming, dynamic programming, and simulation. Contributions of OR to disease prevention and screening, long term management of chronic conditions, and several emerging application areas for OR were discussed.

Many examples of successful OR applications were described, as well as many challenges. For example, the availability of data for analyzing medical decisions is often more complex compared to other real-world decision situations. This is true for a variety of reasons including confidentiality concerns, the fragmented nature of health care delivery, and the lack of the requisite information systems. There are also challenges related to the fundamental difficulty of measuring criteria related to medical decision making, such as the cost to the patient as a result of a burdensome treatment plan. Finally, there are significant challenges in the translation of OR models from theory to practice.

References

Agur, Z., Hassin, R., & Levy, S. (2006). Optimizing chemotherapy scheduling using local search heuristics. *Operations Research, 54*(5), 826–846.

Alagoz, O., Maillart, L. M., Schaefer, A. J., & Roberts, M. S. (2004). The optimal timing of living-donor liver transplantation. *Management Science, 50*(10), 1420–1430.

Becker, N. G., & Starczak, D. N. (1997). Optimal vaccination strategies for a community of households. *Mathematical Biosciences, 139*(2), 117–132.

Braithwaite, R. S., Roberts, M. S., Chang, C., Goetz, M. B., Gibert, C. L., Rodriguez-Barradas, M. C., Shechter, S., Schaefer, A., Nuclfora, K., Koppenhaver, R., & Justice, A. C. (2008). Influence of alternative thresholds for initiating HIV treatment on quality-adjusted life expectancy: A decision model. *Annals of Internal Medicine, 148*, 178–185.

Brandeau, M. L., Owens, D. K., Sox, C. H., & Wachter, R. M. (1993). Screening women of childbearing age for human-immunodeficiency-virus — a model-based policy analysis. *Management Science, 39*(1), 72–92.

Brandeau, M. L., Sainfort, F., & Pierskalla, W. P. (2004). *Operations research and health care*. Boston, MA: Kluwer Academic Publishers.

Cooper, K., Brailsford, S. C., Davies, R., & Raftery, J. (2006). A review of health care models for coronary heart disease interventions. *Health Care Management Science, 9*(4), 311–324.

Denton, B. T., & Verter, V. (2010). Health care O.R. *ORMS Today, 37*(5).

Denton, B. T., Kurt, M., Shah, N. D., Bryant, S. C., & Smith, S. A. (2009). Optimizing the start time of statin therapy for patients with diabetes. *Medical Decision Making, 29*, 351–367.

Eddy, D. M. (1983). A mathematical model for timing repeated medical tests. *Medical Decision Making, 3*(1), 45–62.

Edmonds, J. (1965). Paths, trees, and flowers. *Canadian Journal of Mathematics, 17*, 449–467.

Engineer, F. G., Keskinocak, P., & Pickering, L. K. (2009). OR practice—catch-up scheduling for childhood vaccination. *Operations Research, 57*(6), 1307–1319.

Ferris, M. J., & Shepard, L. D. (2004). Optimization tools for radiation treatment mining in matlab. In M. L. Brandeau, F. Sainfort, & W. P. Pierskalla (Eds.), *Operations research and healthcare: A handbook of methods and applications* (pp. 775–806). Boston, MA: Kluwer Academic Publishers. chap. 30.

Hall, S. N., Jacobson, S. H., & Sewell, E. C. (2008). An analysis of pediatric vaccine formulary selection problems. *Operations Research, 56*(6), 1348–1365.

Hazen, G. B. (2004). Dynamic influence diagrams: Applications to medical decision modeling. In M. L. Brandeau, F. Sainfort, & W. P. Pierskalla (Eds.), *Operations research and healthcare: A handbook of methods and applications* (pp. 613–638). Boston, MA: Kluwer Academic Publishers. chap. 24.

He, M., Zhao, L., & Powell, W. B. (2010). Optimal control of dosage decisions in controlled ovarian hyperstimulation. *Annals of Operations Research, 178*(1), 223–245.

Holder, A. (2004). Radiotherapy treatment design and linear programming. In M. L. Brandeau, F. Sainfort, & W. P. Pierskalla (Eds.), *Operations research and healthcare: A handbook of methods and applications* (pp. 741–774). Boston, MA: Kluwer Academic Publishers. chap. 29.

Jacobson, S. H., Sewell, E. C., Deuson, R., & Weniger, B. G. (1999). An integer programming model for vaccine procurement and delivery for childhood immunization: A pilot study. *Health Care Management Science, 2*(1), 1–9.

Jacobson, S. H., Sewell, E. C., & Proano, R. A. (2006). An analysis of the pediatric vaccine supply shortage problem. *Health Care Management Science, 9*(4), 371–389.

Kaplan, E. H., Craft, D. L., & Wein, L. M. (2003). Analyzing bioterror response logistics: The case of smallpox. *Mathematical Biosciences, 185*(1), 33–72.

Kress, M. (2006). Policies for biodefense revisited: The prioritized vaccination process for smallpox. *Annals of Operations Research, 148*, 5–23.

Lee, H. L., & Pierskalla, W. P. (1988). Mass-screening models for contagious-diseases with no latent period. *Operations Research, 36*(6), 917–928.

Lee, E. K., & Zaider, M. (2008). Operations research advances cancer therapeutics. *Interfaces, 38*(1), 5–25.

Mandelblatt, J. S., Cronin, K., Bailey, S., Berry, D., de Koning, H., Draisma, G., Huang, H., Lee, S., Munsell, M., Plevritis, S., Ravdin, P., Schechter, C., Sigal, B., Stoto, M., Stout, N., van Ravesteyn, N., Venier, J., Zelen, M., & Feuer, E. J. (2009). Effects of mammography screening under different screening schedules: Model estimates of potential benefits and harms. *Annals of Internal Medicine, 151*, 738–747.

Marquard, J. L., & Brennan, P. F. (2009). Crying wolf: Consumers may be more willing to share medication information than policymakers think. *Journal of Health Information Management, 23*(2), 26–32.

Mason, J. E., England, D. A., Denton, B. T., Smith, S. A., Kurt, M., & Shah, N. D. (2012). Optimizing statin treatment decisions for diabetes patients in the presence of uncertain future adherence. *Medical Decision Making, 32*(1), 154–166.

Paltiel, A., Kuntz, K., Weiss, S., & Fuhlbrigge, A. (2004). Asthma policy model. In M. L. Brandeau, F. Sainfort, & W. P. Pierskalla (Eds.), *Operations research and healthcare: A handbook of methods and applications* (pp. 659–694). Boston, MA: Kluwer Academic Publishers. chap. 26.

Parker, R., Doyle, F., & Peppas, N. (2001). The intravenous route to blood glucose control. *IEEE Engineering in Medicine and Biology, 20*(1), 65–73.

Pierskalla, W. P, Brailer, D. J. (1994). *Applications of Operations Research in Health Care Delivery. 6*. North Holland.

Pignone, M., Earnshaw, S., Tice, J. A., & Pletcher, M. J. (2006). Aspirin, statins, or both drugs for the primary prevention of coronary heart disease events in men: A cost-utility analysis. *Annals of Internal Medicine, 144*, 326–336.

Rais, A., & Viana, A. (2010). Operations research in healthcare: A survey. *International Transactions in Operational Research, 18*, 1–31.

Rubin, D., Burnside, E., & Shachter, R. (2004). A Bayesian network to assist mammography interpretation. In M. L. Brandeau, F. Sainfort, & W. P. Pierskalla (Eds.), *Operations research and healthcare: A handbook of methods and applications* (pp. 695–720). Boston, MA: Kluwer Academic Publishers. chap. 27.

Schwartz, J. S., Kinosian, B. P., Pierskalla, W. P., & Lee, H. (1990). Strategies for screening blood for humanimmunodeficiency- virus antibody — use of a decision support system. *Journal of the American Medical Association, 264*(13), 1704–1710.

Segev, D. L., Gentry, S. E., Warren, D. S., Reeb, B., & Montgomery, R. A. (2005). Kidney paired donation and optimizing the use of live donor organs. *JAMA — Journal of the American Medical Association, 293*(15), 1883–1890.

Shechter, S. M., Bailey, M. D., Schaefer, A. J., & Roberts, M. S. (2008). The optimal time to initiate HIV therapy under ordered health states. *Operations Research, 56*(1), 20–33.

Shwartz, M. (1978). Mathematical-model used to analyze breast-cancer screening strategies. *Operations Research, 26*(6), 937–955.

Simon, J. (2009). Decision making with prostate cancer: A multiple-objective model with uncertainty. *Interfaces, 39*(3), 218–227.

Su, X., & Zenios, S. (2004). Patient choice in kidney allocation: The role of the queueing discipline. *Manufacturing and Service Operations Management, 6*(4), 280–301.

Wein, L. M., & Zenios, S. A. (1996). Pooled testing for HIV screening: Capturing the dilution effect. *Operations Research, 44*(4), 543–569.

White, C. C., III, Wilson, E. C., & Weaver, A. C. (1982). Decision aid development for use in ambulatory health care settings. *Operations Research, 30*(3), 446–463.

Wu, J. T., Wein, L. M., & Perelson, A. S. (2005). Optimization of influenza vaccine selection. *Operations Research, 53*(3), 456–476.

Zhang, J., Denton, B. T., Balasubramanian, H., Shah, N. D., & Inman, B. A. (2012). Optimization of PSA screening policies: A comparison of the patient and societal perspectives. *Medical Decision Making, 32*(2), 337–349.

Distribution Selection for Stochastic Modeling

Donald Gross
George Mason University, Fairfax, VA, USA

Introduction

The choice of appropriate probability distributions is the most important step in any complete stochastic system analysis and hinges upon knowing as much as possible about the characteristics of the potential distribution and the physics of the situation to be modeled. Generally, the first thing that has to be decided is which probability distributions are appropriate to use for the relevant random phenomena describing the model. For example, the exponential distribution has the Markovian (memoryless) property. Is this a reasonable condition for the particular physical situation under study? Assume the problem is to describe the repair mechanism of a complex maintained system. If the service for all customers is fairly repetitive, then an assumption might be that the longer a failed item is in service for repair, the greater the probability that its service will be completed in the next interval of time (non-memoryless). In this case, the exponential distribution would not be a reasonable candidate for consideration. On the other hand, if the service is mostly diagnostic in nature (the trouble must be found and fixed), or there is a wide variation of service required from customer to customer so that the probability of service completion in the next instant of time is independent of how long the customer has been in service, the exponential with its memoryless property might indeed suffice.

The actual shape of the density function also gives quite a bit of information, as do its moments. One particularly useful measure is the ratio of the standard deviation to the mean, called the coefficient of variation (CV). The exponential distribution has a CV $= 1$, while the Erlang or convolution of exponentials has a CV < 1, and the hyperexponential or mixture of exponentials has a CV > 1. Hence, choosing the appropriate distribution is a combination of knowing as much as possible about distribution characteristics, the physics of the situation to be modeled, and statistical analyses when data are available.

Hazard Rate

An important concept that helps in characterizing probability distributions that is strongly associated with reliability modeling is the hazard-rate (also termed the failure-rate) function. This concept, however, can be useful in general when trying to decide upon the proper probability distribution to select. In the discussion that follows, the hazard rate will be related to the Markov property for the exponential distribution, and its use as a way to gain insight about probability distributions will be discussed.

Suppose it is desired to choose a probability distribution to describe a continuous lifetime random variable T with a cumulative distribution function (CDF) of $F(t)$. The density function, $f(t) = df(t)/dt$, can be interpreted as the approximate probability that the random time to failure will be in a neighborhood about a value t. The CDF is, of course, the probability that the time will be less than or equal to the value t. Then the hazard rate $h(t)$ is defined as the conditional probability that the lifetime will be in a neighborhood about the value t, given that the time is already at least t. That is, if the situation deals with failure times, $h(t)dt$ is the approximate probability that the device fails in the interval $(t, t + dt)$, given it is working at time t.

From the laws of conditional probability, it can be shown that

$$h(t) = \frac{f(t)}{1 - F(t)}.$$

This hazard or failure-rate function can be increasing in t (called an increasing failure rate, or IFR), decreasing in t (called a decreasing failure rate, or DFR), constant (considered to be both IFR and DFR), or a combination. The constant case implies the memoryless or ageless property, and this holds for the exponential distribution, as will be shown. If, however, it is believed that the device ages and that the longer it has been operating the more likely it is that the device will fail in the next dt, then it is desired to have an $f(t)$ for which $h(t)$ is increasing in t; that is, an IFR distribution. This concept can be utilized for any stochastic modeling situation. For example, if instead of modeling lifetime of a device, the concern is with describing the service time of a customer at a bank, then, if service is fairly routine for each customer, then an IFR distribution would be desired. But if customers required a variety of needs (say a queue where both business and personal transactions were allowed), then a DFR or perhaps a CFR exponential might be the best choice.

Reversing the algebraic calculations, a unique $F(t)$ can be obtained from $h(t)$ by solving a simple linear, first-order differential equation, i.e.,

$$F(t) = -\exp\left(-\int_0^t h(u)\,du\right).$$

The hazard rate is another important information source (as is the shape of $f(t)$ itself) for obtaining knowledge concerning candidate probability distributions.

Consider the exponential distribution

$$f(t) = \theta \exp(-\theta t).$$

From the discussion above, it is easily shown that $h(t) = \theta$. Thus, the exponential distribution has a constant failure (hazard) rate and is memoryless. Suppose, for a particular situation, there is a need for an IFR distribution for describing some random times. It turns out that the Erlang has this property. The density function is

$$f(t) = \theta^k t^{k-1} \exp(-\theta t)/(k-1)!$$

(a special form for the gamma), with its CDF determined in terms of the incomplete gamma function or equivalently as a Poisson sum. From these, it is not too difficult to calculate the Erlang's hazard rate, that also has a Poisson sum term, but is somewhat complicated to ascertain the direction of $h(t)$ with t without doing some numerical work. It does turn out, however, that $h(t)$ increases with t and at a decelerating rate.

Suppose the opposite IFR condition is desired, that is, an accelerating rate of increase with t. The Weibull distribution can obtain this condition. In fact, depending on how the shape parameter of the Weibull is chosen, an IFR can be obtained with decreasing acceleration, constant acceleration (linear with t), or increasing acceleration, as well as even obtaining a DFR or the constant failure rate exponential. The CDF of the Weibull is given by

$$F(t) = 1 - \exp(-at^b)$$

and its hazard rate turns out to be the simple monomial $h(t) = abt^{b\text{-}1}$, with shape determined by the value of b (called the shape parameter).

As a further example in the process of choosing an appropriate candidate distribution for modeling, suppose, for an IFR that has a deceleration effect, such as the Erlang, there is a believe that the CV might be greater than one. This latter condition eliminates the Erlang from consideration. But, it is known that a mixture of (k) exponentials (often denoted by H_k) does have a CV > 1. It is also known that any mixture of exponentials is DFR. In fact, it can be shown that all IFR distributions have CV < 1, while all DFR distributions have CV > 1 (Barlow and Proschan 1975). Thus, if there is convincing evidence that the model requires an IFR, CV < 1 must be accepted. Intuitively, this can be explained as follows. Situations that have CV > 1

often are cases where the random variables are mixtures (say, of exponentials). Thus, for example, if a customer has been in service a long time, chances are that it is of a type requiring a lot of service, so the probability of completion in the next infinitesimal interval of width *dt* diminishes over time. Situations that have an IFR condition indicate a more consistent pattern among items, thus yielding a CV < 1.

Range of the Random Variable

Knowledge of the range of the random variable under study can also help narrow the possible choices in selecting an appropriate distribution. In many cases, there is a minimum value that the random variable can assume. For example, suppose the analysis concerns the interarrival times between subway trains, and it is given that there is a minimum time for safety of γ. The distributions discussed thus far (and, indeed, many distributions) have zero as their minimum value. Any such distribution, however, can be made to have a minimum other than zero by adding a location parameter, say γ. This is done by subtracting γ from the random variable in the density function expression. Suppose the exponential distribution is to be used, but we have a minimum value of γ. The density function would then become $f(t) = \theta \exp(-\theta[t - \gamma])$. It is not quite so easy to build in a maximum value if this should be the case. For this situation, a distribution with a finite range would have to be chosen, such as the uniform, the triangular or the more general beta distribution (Law and Kelton 1991).

Data

While much information can be gained from knowledge of the physical processes associated with the stochastic system under study, it is very advantageous to obtain data, if at all possible. For existing systems, data may already exist or can be obtained by observing the system. These data can then be used to gain further insight on the best distributions to choose for modeling the system. For example, the sample standard deviation and mean can be calculated, and it can be observed whether the sample CV is less than, greater than, or approximately equal to one. This would give an idea as to whether an IFR, DFR or the exponential distribution would be the more appropriate.

If enough data exist, just plotting a histogram can often provide a good idea of possible distributions from which to choose, since theoretical probability distributions have distinctive shapes (although some do closely resemble each other). The exponential shape of the exponential distribution is far different, for example, than the bell-shaped curve of the normal distribution.

There are rigorous statistical goodness of fit procedures to indicate if it is reasonable to assume that the data could come from a potential candidate distribution. These do, however, require a considerable amount of data and computation to yield satisfactory results. But, there are statistical packages, for example, Expert Fit (Law and Vincent 1995), which will analyze sets of data and recommend the theoretical distributions that are the most likely to yield the kind of data being studied.

Distribution selection (or input modeling, as it is sometimes called) is not a trivial procedure. But this is a most important aspect of stochastic analysis, since inaccuracies in the input can make the output meaningless. Fitting data to standard statistical distributions, which are mostly two-parameter distributions, limits focus on the first two moments only. There is evidence to suggest that this is not always sufficient (see Gross and Juttijudata 1997).

Finally, for emphasis, the point is made again that choosing an appropriate probability model is a combination of knowing as much as possible about the characteristics of the probability distribution being considered and as much as possible about the physical situation being modeled.

See

▶ Failure-Rate Function
▶ Hazard Rate
▶ Markov Chains
▶ Markov Processes
▶ Reliability of Stochastic Systems
▶ Simulation of Stochastic Discrete-Event Systems
▶ Stochastic Input Model Selection

References

Barlow, R. E., & Proschan, F. (1975). *Statistical theory of reliability and life testing*. New York: Holt, Rinehart and Winston.

Forbes, C., Evans, M., Hastings, N., & Peacock, B. (2011). *Statistical distributions* (4th ed.). Hoboken, NJ: Wiley.

Gross, D., & Juttijudata, M. (1997). Sensitivity of output performance measures to input distributions in queueing simulation modeling. In S. Andradottir, K. J. Healy, D. H. Withers, & B. L. Nelson (Eds.), *Proceedings of the 1997 winter simulation conference*. Piscataway, NJ: IEEE.

Law, A. M., & Kelton, W. D. (1991). *Simulation modeling and analysis* (2nd ed.). New York: McGraw-Hill.

Law, A. M., & Vincent, S. (1995). *Expert fit user's guide*. Tucson, AZ: Averill M. Law and Associates.

DMU

Decision making unit.

See

▶ Data Envelopment Analysis

Documentation

Saul I. Gass
University of Maryland, College Park, MD, USA

Introduction

As many operations research studies involve a mathematical decision model that is quite complex in its form, it is incumbent upon those who developed the model and conducted the analysis to furnish documentation that describes the essentials of the model, its use, and its results. Of especial concern are those computer-based models that are represented by a computer program and its input data files. The most serious weakness in the majority of OR model applications, both those that are successful and those that fail, is the lack of documents that satisfy the minimal requirements of good documentation practices (Gass et al. 1981; Gass 1984). The reasons for requiring documentation are many-fold and include, among others, "to enable system analysts and programmers, other than the originators, to use the model and program," "to facilitate auditing and verification of the model and the program operations," and "to enable potential users to determine whether the model and programs will serve their needs" (Gass 1984).

The most acceptable view of model documentation is that which calls for documents that record and describe all aspects of the model development life-cycle. The life-cycle model documentation approach given in Gass (1979) calls for the production of 13 major documents. However, it is recognized that in terms of the basic needs of model users and analysts, these documents can be rewritten and combined into the following four manuals: *Analyst's Manual*, *User's Manual*, *Programmer's Manual, and Manager's Manual*. Brief descriptions of the contents of these manuals are given below; detailed tables of contents for each are given in Gass (1984).

Analyst's Manual

The analyst's manual combines information from the other project documents and is a source document for analysts who have been and will be involved in the development, revisions, and maintenance of the model. It should include those technical aspects that are essential for practical understanding and application of the model, such as a functional description, data requirements, verification and validation tests, and algorithmic descriptions.

User's Manual

The purpose of the user's manual is to provide (nonprogramming) users with an understanding of the model's purposes, capabilities, and limitations so they may use it accurately and effectively. This manual should enable a user to understand the overall structure and logic of the model, input requirements, output formats, and the interpretation and use of the results. This manual should also enable technicians to prepare the data and to set up and run the model.

Programmer's Manual

The purpose of the programmer's manual is to provide the current and future programming staff with the information necessary to maintain and modify the model's program. This manual should provide all the details necessary for a programmer to understand the operation of the software, to trace through it for debugging and error correction, for making modifications, and for determining if and how the programs can be transferred to other computer systems or other user installations.

Manager's Manual

The manager's manual is essential for computer-based models used in a decision environment. It is directed at executives of the organization who will have to interpret and use the results of the model, and support its continued use and maintenance. This manual should include a description of the problem setting and origins of the project; a general description of the model, including its purpose, objectives, capabilities, and limitations; the nature, interpretation, use, and restrictions of the results that are produced by the model; costs and benefits to be expected in using the model; the role of the computer-based model in the organization and decision structure; resources required; data needs; operational and transfer concerns; and basic explanatory material.

See

- ▶ Implementation
- ▶ Model Evaluation
- ▶ Model Management
- ▶ Practice of Operations Research and Management Science

References

Brewer, G. D. (1976). Documentation: An overview and design strategy. *Simulation & Games, 7*, 261–280.

Gass, S. I. (1979). *Computer model documentation: A review and an approach, National Bureau of Standards Special Publication 500–39, U.S. GPO Stock No. 033-003-02020-6*, Washington, DC.

Gass, S. I. (1984). Documenting a computer-based model. *Interfaces, 14*, 84–93.

Gass, S. I., Hoffman, K. L., Jackson, R. H. F., Joel, L. S., & Sanders, P. B. (1981). Documentation for a model: A hierarchical approach. *ACM Communications, 24*, 728–733.

NBS. (1976). *Guidelines for documentation of computer programs and automated data systems, FIPS PUB 38*. Washington, DC: U.S. Government Printing Office.

NBS. (1980). *Computer model documentation guide, NBS special publication 500-73*. Washington, DC: U.S. Government Printing Office.

Domain Knowledge

The knowledge that an expert has about a given subject area.

See

- ▶ Artificial Intelligence
- ▶ Forecasting

DP

- ▶ Dynamic Programming

DSS

- ▶ Decision Support Systems (DSS)

Dual Linear-Programming Problem

A companion problem defined by a linear-programming problem. Every linear-programming problem has an associated dual-programming program. When the linear-programming problem has the form

$$\begin{aligned} \text{Minimize} \quad & \boldsymbol{c}^T\boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b} \\ & \boldsymbol{x} \geq \boldsymbol{0} \end{aligned}$$

then its dual problem is also a linear-programming problem with the form

$$\begin{aligned} \text{Maximize} \quad & \boldsymbol{b}^T \boldsymbol{y} \\ \text{subject to} \quad & \boldsymbol{A}^T \boldsymbol{y} \leq \boldsymbol{c} \\ & \boldsymbol{y} \geq \boldsymbol{0} \end{aligned}$$

The original problem is called the primal problem. If the primal minimization problem is given as equations in nonnegative variables, then its dual is a maximization problem with less than or equal to constraints whose variables are unrestricted (free). The optimal solutions to primal and dual problems are strongly interrelated.

See

- ▶ Complementary Slackness Theorem
- ▶ Duality Theorem
- ▶ Symmetric Primal-Dual Problems
- ▶ Unsymmetric Primal-Dual Problems

References

Dantzig, G. B. (1963). *Linear programming and extensions*. Princeton, NJ: Princeton University Press.

Gass, S. I. (1984). *Linear programming* (5th ed.). New York: McGraw-Hill.

Duality Theorem

A theorem concerning the relationship between the solutions of primal and dual linear-programming problems. One form of the theorem is as follows: If either the primal or the dual has a finite optimal solution, then the other problem has a finite optimal solution, and the optimal values of their objective functions are equal. From this it can be shown that for any pair of primal and dual linear programs, the objective value of any feasible solution to the minimization problem is greater than or equal to the objective value of any feasible solution to the dual maximization problem. This implies that if one of the problems is feasible and unbounded, then the other problem is infeasible. Examples exist for which the primal and its dual are both infeasible. Another form of the theorem states: if both problems have feasible solutions, then both have finite optimal solutions, with the optimal values of their objective functions equal.

See

- ▶ Dual Linear-Programming Problem
- ▶ Strong Duality Theorem

Dualplex Method

A procedure for decomposing and solving a weakly-coupled linear-programming problem.

See

- ▶ Block-Angular System

Dual-Simplex Method

An algorithm that solves a linear-programming problem by solving its dual problem. The algorithm starts with a dual feasible but primal infeasible solution, and iteratively attempts to improve the dual objective function while maintaining dual feasibility.

See

- ▶ Dual Linear-Programming Problem
- ▶ Feasible Solution
- ▶ Primal-Dual Algorithm
- ▶ Simplex Method (Algorithm)

Dummy Arrow

A dashed arrow used in a project network diagram to show relationships among project items, a logical dummy, or to give a unique designation to an activity, thus called a uniqueness dummy. A dummy or dummy arrow represents no time or resources.

See

▶ Network Planning

Dynamic Programming

Chelsea C. White III
Georgia Institute of Technology, Atlanta, GA, USA

Introduction

Dynamic programming (DP) is both an approach to problem solving and a decomposition technique that can be effectively applied to mathematically describable problems having a sequence of interrelated decisions. Such decision-making problems are pervasive. Determining a route from an origin (e.g., home) to a destination (e.g., school) on a network of roads requires a sequence of turns. Managing a retail store (e.g., that sells, say, television sets) requires a sequence of wholesale purchasing decisions.

Such problems share important characteristics. Each is associated with a criterion to be optimized: choosing the shortest or most scenic route from home to school, and the buying and selling of television sets by the retail store manager to maximize expected profit. Also, each problem has a structure such that a currently determined decision has impact on the future decision-making environment. In going from home to school, the turn currently selected will determine the geographical location of the next turn decision; in managing the retail store, the number of items ordered today will affect the level of inventory next week.

Roots and Key References

In his 1957 book, Richard Bellman described the concept of DP and its broad potential for application. See Bellman's earlier publications that describe his initial developments of DP (Bellman 1954a, b); also see Bertsekas (1987); Denardo (1982); Heyman and Sobel (1984); Hillier and Lieberman (2004, Chapter 10), and Ross (1983) for in depth descriptions and applications of DP.

Central to the philosophy and methodology of DP is the Principle of Optimality, as related to the following multistage decision problem (Bellman 1957). Let $\{q_1, q_2, \ldots q_n\}$ be a sequence of allowable decisions called a policy; specifically, an n-stage policy. A policy that yields the maximum value of the related criterion function is called an optimal policy. Decisions are based on the state of the process, that is, the information available to make a decision. The basic property of optimal policies is expressed by the following:

> *Principle of Optimality*: An optimal policy has the property that whatever the initial state and the initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision (Bellman 1957).

The Principle of Optimality can be expressed as an optimization problem over the set of possible decisions by a recursive relationship, the application of which yields the optimal policy. This is illustrated next by two examples.

1. An itinerary selection problem. The problem is to find the shortest path from home to school. A map of the area describes the network of streets that includes home and school locations, intermediate intersections, connecting streets, and the distance from one intersection to any other intersection that is directly connected by a street. The DP model of this problem is as follows. Let N be the set composed of home, school, and all intersections. An element of N is termed a node. For simplicity, assume all of the streets are one-way. A street is described as an ordered pair of nodes; that is, (n, n') is the street going from node n to node n' (n' is an immediate successor of node n). Let $m(n, n')$ be the distance from node n to node n'; that is, $m(n, n')$ represents the length of street (n, n').

   The problem is examined recursively as follows. Let $f(n)$ equal the shortest distance from the node n to the goal node school. The objective is to find f (home), the minimum distance from home to school, and a path from home to school that has a distance equal to f (home), a minimum distance path.

Note that $f(n) \leq m(n, n') + f(n')$ for any node n' that is an immediate successor of node n. Assume that an immediate successor n'' of n such that $f(n) = m(n, n'') + f(n'')$ has been found. Then, if at node n, it seems reasonable that the street that takes us to node n'' is traversed. Thus, the evaluation of all of the values $f(n)$ determine both f(home) and a minimum distance path from home to school. Formally, determination of these values can proceed recursively from the equation $f(n) = \min \{m(n, n') + f(n')\}$, where the minimum is taken over all nodes n' that are immediate successors of node n and where f(school) $= 0$ is the initial condition.

2. An inventory problem. Let $x(t)$ be the number of items in stock at the end of week t, $d(t + 1)$ the number of customers wishing to make a purchase during week $t + 1$, and $u(t)$ the number of items ordered at the end of week t and delivered at the beginning of week $t + 1$. Although it is unlikely that $d(t)$ is known precisely, assume the probability that $d(t) = n$ is known for all $n = 0, 1, \ldots$. Keeping backorders, then $x(t + 1) = x(t) - d(t + 1) + u(t)$. A reasonable objective is to minimize the expected cost accrued over the period from $t = 0$ to $t = T$ $(T > 0)$ by choice of $u(0), \ldots, u(T - 1)$, assuming that ordering decisions are made on the basis of the current inventory level, that is, the mechanism that determines $u(t)$ (e.g., the store manager) is aware of $x(t)$, for all $t = 0, \ldots, T - 1$. Costs might include a shortage cost (a penalty if there is an insufficient amount of inventory in stock), a storage cost (a penalty if there is too much inventory in stock), an ordering cost (reflecting the cost necessary to purchase items wholesale), and a selling price (reflecting the income received when an item is sold; a negative cost). Let $c(x, u)$ represent the expected total cost to be accrued from the end of week t till the end of week $t + 1$, given that $x(t) = x$ and $u(t) = u$. Then the criterion to be minimized is

$$E\{c[x(0), u(0)] + \ldots + c[x(T-1), u(T-1)]\},$$

where E is the expectation operator associated with the random variables $d(1), \ldots, d(T)$.

This problem can be examined recursively. Let $f(x, t)$ be the minimum expected cost to be accrued from time t to time T, assuming that $x(t) = x$. Clearly, $f(x, T) = 0$. Note also that

$$f[x(t), t] \leq c[x(t), u(t)] + E\{f[x(t) - d(t+1) + u(t), t+1]\}$$

for any available $u(t)$. As was true for Example 1, an order number u'' which is such that

$$f[x(t), t] = c[x(t), u''] + E\{f[x(t) - d(t+1) + u'', t+1]\}$$

is an order to place at time t when the current inventory is $x(t)$. Thus, the recursive equation determines both $f(x, 0)$ for all x and the order number as a function of current inventory level.

Common Characteristics

Two key aspects of DP are the notion of a state and recursive equations. The state of the DP problem is the information that is currently available to the decision maker on which to base the current decision. For example, in the itinerary selection problem, the state is the current node; in the inventory problem, the state is the current number of items in stock. In both examples, how the system arrived at its current state is inconsequential from the perspective of decision making. For the itinerary selection problem, all that is needed is the current node and not the path that lead to that node to determine the best next street to traverse. The determination of the number of items to order this week depends only on the current inventory level equations (other names include functional equations and optimality equations) that can be used to determine the minimum expected value of the criterion and an optimal sequence of decisions that depend on the current node or current inventory level. In both cases, the recursive equations essentially decompose the problem into a series of subproblems, one for each node or current state value.

See

- ▶ Approximate Dynamic Programming
- ▶ Bellman Optimality Equation
- ▶ Dijkstra’s Algorithm
- ▶ Markov Decision Processes
- ▶ Network

References

Bellman, R. E. (1954a). Some problems in the theory of dynamic programming. *Econometrica, 22*(1), 37–48.

Bellman, R. E. (1954b). Some applications of the theory of dynamic programming. *Journal of the Operations Research Society of America, 2*(3), 275–288.

Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Bertsekas, D. P. (1987). *Dynamic programming: Deterministic and stochastic models*. Englewood Cliffs, NJ: Prentice-Hall.

Denardo, E. V. (1982). *Dynamic programming: Models and applications*. Englewood Cliffs, NJ: Prentice-Hall.

Dreyfus, S., & Law, A. (1977). *The art and theory of dynamic programming*. New York: Academic Press.

Heyman, D. P., & Sobel, M. J. (1984). *Stochastic models in operations research* (Vol. II). New York: McGraw-Hill.

Hillier, F. S., & Lieberman, G. J. (2004). *Introduction to operations research* (8th ed.). New York: McGraw-Hill.

Lew, A., & Mauch, H. (2007). *Dynamic programming: A computational tool*. New York: Springer.

Ross, S. M. (1983). *Introduction to stochastic dynamic programming*. New York: Academic Press.